UNIVERSIDAD NACIONAL DE SAN CRISTÓBAL DE HUAMANGA

FACULTAD DE INGENIERÍA DE MINAS, GEOLOGÍA Y CIVIL

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

TESIS:

Predicción de la Ansiedad Generalizada en Estudiantes Universitarios mediante Random Forest basado en el Cuestionario Generalized Anxiety Disorder 7-item (GAD-7), 2025.

Para optar el título profesional de:
INGENIERO DE SISTEMAS

PRESENTADO POR:
Bach. Irvin Pelayo TAYPE PANIAGUA

ASESOR:
Mg. Ing. Hubner JANAMPA PATILLA

AYACUCHO - PERÚ

2025

RESUMEN

La presente investigación tiene como objetivo desarrollar un modelo predictivo para identificar el riesgo de ansiedad generalizada en estudiantes universitarios de la Universidad Nacional de San Cristóbal de Huamanga, utilizando el algoritmo Random Forest basado en el cuestionario GAD-7.

La problemática radica en el incremento de la ansiedad en estudiantes y en la limitada capacidad predictiva de herramientas tradicionales, que solo permiten evaluaciones descriptivas. En respuesta, se propone el uso de técnicas de aprendizaje automático para mejorar la detección temprana.

El estudio es de tipo aplicado, con diseño no experimental, transversal y nivel correlacional-predictivo. La muestra estuvo conformada por 166 estudiantes seleccionados mediante muestreo aleatorio simple. La recolección de datos se realizó mediante el cuestionario GAD-7 complementado con variables sociodemográficas, académicas y de estilo de vida.

Se desarrolló un modelo de clasificación multiclase con Random Forest, que permitió categorizar a los estudiantes en niveles de ansiedad: sin ansiedad, leve, moderada y severa. Se aplicaron técnicas de preprocesamiento, partición de datos y optimización de hiperparámetros.

Los resultados evidencian un buen desempeño del modelo en precisión, sensibilidad y F1-score, destacando su capacidad para identificar niveles moderados y severos. Asimismo, se identificaron como variables relevantes el estrés académico, las horas de sueño y las respuestas del GAD-7.

Se concluye que el modelo propuesto constituye una herramienta eficaz para la predicción temprana de la ansiedad en estudiantes universitarios.

Palabras clave: Ansiedad generalizada, GAD-7, Random Forest, aprendizaje automático, predicción.

Contenido

LISTA DE TABLAS

LISTA DE FIGURAS

DEDICATORIA

A mis padres, Clemencia Paniagua Lopez y Pelayo Taype Cardenas, por su amor y apoyo incondicional,

A mis hermanos Kevin y Teffa.

Y a ti, que me estas viendo, sonríe.

AGRADECIMIENTO

A mis seres queridos, por motivarme siempre a ser una mejor persona.

Al Ing. Hubner Janampa Patilla, mi asesor, por su apoyo, orientación y paciencia durante todo el tiempo que se realizo el trabajo de investigación.

A mis amigos incondicionales, que siempre están presentes en mi vida, y a todos aquellas personas que creyeron en mi

Agradecimientos totales.

I. INTRODUCCIÓN

En los últimos años, la **ansiedad generalizada** se ha convertido en una de las principales preocupaciones de salud mental entre los estudiantes universitarios. Las presiones académicas, el ritmo acelerado de los estudios, las exigencias sociales y las incertidumbres propias de la vida universitaria crean un entorno propicio para el desarrollo de cuadros de ansiedad que afectan tanto el bienestar emocional como el rendimiento académico. Esta realidad se observa con claridad en la **Universidad Nacional de San Cristóbal de Huamanga**, donde la exigencia de las carreras, especialmente en áreas como ingeniería, se suma a factores personales y contextuales que pueden incrementar el riesgo de ansiedad.

Frente a esta problemática, los servicios de orientación y bienestar universitario enfrentan una gran limitación: la falta de **herramientas tecnológicas y predictivas** que permitan detectar tempranamente los niveles de riesgo entre los estudiantes. En la actualidad, los instrumentos de evaluación más utilizados, como el **cuestionario Generalized Anxiety Disorder 7-item (GAD-7)**, han demostrado su eficacia para medir la severidad de los síntomas de ansiedad, pero su aplicación sigue siendo descriptiva y no predictiva. Esto significa que, si bien se puede conocer el nivel actual de ansiedad de un estudiante, no se puede anticipar con precisión quiénes podrían desarrollar niveles más graves a futuro, lo que retrasa las acciones preventivas.

Partiendo de esta necesidad, el presente estudio plantea el **problema central de investigación**:

¿Cómo se puede predecir con precisión el riesgo de ansiedad generalizada en estudiantes universitarios utilizando el algoritmo Random Forest basado en los resultados del cuestionario GAD-7, en la UNSCH durante el año 2025?

El **objetivo general** de la investigación es desarrollar un **modelo predictivo basado en el algoritmo Random Forest** que permita estimar el riesgo de ansiedad generalizada en los estudiantes universitarios a partir de los datos obtenidos del GAD-7 y de otras variables contextuales.

Los **objetivos específicos** son:

Identificar las variables que influyen con mayor peso en la predicción del riesgo de ansiedad generalizada.

Diseñar un modelo de predicción multiclase utilizando Random Forest para clasificar a los estudiantes según su nivel de ansiedad.

Evaluar el rendimiento del modelo mediante métricas de precisión, sensibilidad, especificidad y F1-score, garantizando su validez y confiabilidad.

La **justificación** del estudio se sustenta en tres dimensiones. En el plano **científico**, la investigación aporta al desarrollo del conocimiento en el campo de la inteligencia artificial aplicada a la salud mental, demostrando cómo los algoritmos de *Machine Learning* pueden complementar los enfoques clínicos tradicionales. En el ámbito **social**, la creación de un modelo predictivo facilitará la **detección temprana de casos de riesgo**, permitiendo que las universidades implementen estrategias de apoyo psicológico más rápidas, personalizadas y efectivas. Finalmente, en el aspecto **tecnológico**, el uso del algoritmo **Random Forest** ofrece una solución moderna, capaz de analizar grandes volúmenes de información y generar resultados interpretables, fortaleciendo el uso de herramientas basadas en datos en el ámbito educativo.

La **delimitación** del estudio establece como espacio de análisis la **Facultad de Ingeniería de la Universidad Nacional de San Cristóbal de Huamanga**, durante el año **2025**, y considera como población objetivo a los **estudiantes universitarios matriculados** en las diferentes especialidades. El análisis se centra en la aplicación del cuestionario GAD-7 y en la recopilación de variables académicas, demográficas y de hábitos personales que pueden influir en el desarrollo de la ansiedad.

La **importancia** de este trabajo radica en su doble contribución: por un lado, genera conocimiento científico sobre la interacción entre factores psicológicos y contextuales en la ansiedad universitaria, y por otro, propone una **herramienta práctica y automatizada** que puede ser utilizada por los profesionales de salud

mental de la universidad. Al combinar datos objetivos con técnicas avanzadas de análisis predictivo, el modelo puede transformar la forma en que se detecta y aborda la ansiedad en los estudiantes, promoviendo un enfoque más preventivo y personalizado.

En cuanto a la **estructura del documento**, este se organiza en varios capítulos que siguen una secuencia lógica y coherente:

El **Capítulo I** desarrolla la introducción, el planteamiento del problema, los objetivos, la justificación, la delimitación y la importancia del estudio.

El **Capítulo II** expone el **marco teórico**, presentando los fundamentos conceptuales de la ansiedad, el instrumento GAD-7 y los principios del algoritmo Random Forest.

El **Capítulo III** detalla la **metodología (materiales y métodos),** incluyendo la descripción de los datos, el proceso de construcción del modelo y las métricas de evaluación utilizadas.

El **Capítulo IV** presenta los **resultados y discusión** del modelo predictivo, la interpretación de las variables más influyentes y el análisis del desempeño del modelo.

El Capítulo V contiene la **conclusiones y recomendaciones**, orientadas a la aplicación práctica del modelo en entornos educativos y clínicos.

Esta investigación titulada **"Predicción de la Ansiedad Generalizada en Estudiantes Universitarios mediante Random Forest basado en el Cuestionario Generalized Anxiety Disorder 7-item (GAD-7)"**, propone una integración entre la ciencia de datos y la psicología, demostrando que la inteligencia artificial puede ser una aliada estratégica en la detección temprana de trastornos emocionales. Su objetivo final es contribuir al **bienestar emocional y académico de los estudiantes**, fortaleciendo los mecanismos de apoyo universitario a través de una tecnología ética, confiable y centrada en la persona.

1.1. Planteamiento del problema

La ansiedad generalizada se ha convertido en un problema de salud mental alarmante entre los estudiantes universitarios, especialmente en aquellos que cursan carreras exigentes como ingeniería. Este trastorno psicológico afecta profundamente el bienestar emocional y el rendimiento académico de los estudiantes, generando un entorno que puede obstaculizar su desarrollo personal y profesional (Eisenberg et al., 2016; Kessler et al., 2005). En el contexto de la Universidad Nacional de San Cristóbal de Huamanga, se observa que la presión académica, combinada con las exigencias del entorno social y personal, crea un caldo de cultivo para la ansiedad, lo que podría tener repercusiones significativas en el éxito académico y la salud mental a largo plazo (González et al., 2020).

Para abordar esta problemática, se han utilizado diversas herramientas de evaluación, como el cuestionario Generalized Anxiety Disorder 7-item (GAD-7), que ha demostrado ser efectivo en la medición de los niveles de ansiedad en diferentes poblaciones (Spitzer et al., 2006). Sin embargo, la aplicación de este instrumento en la predicción del riesgo de desarrollar ansiedad es limitada. Las soluciones actuales no permiten identificar de manera precisa a los estudiantes en riesgo, lo que dificulta la implementación de estrategias preventivas que podrían mitigar los efectos de la ansiedad generalizada (Roslan & Ahmad, 2020). Este hecho es especialmente preocupante, dado que una intervención temprana podría resultar en mejoras significativas en el bienestar y el rendimiento académico de los estudiantes (Gonzalez et al., 2022).

A pesar de la utilidad del GAD-7, su enfoque descriptivo no ofrece un modelo predictivo que permita anticipar el desarrollo de síntomas de ansiedad en los estudiantes de ingeniería (Liu et al., 2021). Esta brecha en la investigación señala la necesidad de adoptar enfoques más innovadores para la evaluación y la gestión de la ansiedad. El uso de modelos de aprendizaje automático, como Random Forest, se presenta como una alternativa prometedora para predecir la aparición de ansiedad generalizada

en esta población específica (Nath et al., 2020). Estos modelos tienen la capacidad de analizar grandes volúmenes de datos y detectar patrones complejos, lo que podría resultar en una identificación más precisa de los estudiantes en riesgo (Cohen et al., 2023).

Sin embargo, la efectividad de estos modelos en el contexto de estudiantes de ingeniería de la UNSCH aún no ha sido suficientemente explorada. La investigación actual no ha proporcionado datos concretos sobre la aplicabilidad y precisión de modelos predictivos en esta población, lo que plantea un desafío importante para la implementación de soluciones basadas en evidencia (Duncan et al., 2022). La identificación temprana de la ansiedad podría facilitar el desarrollo de programas de intervención más efectivos y personalizados, permitiendo a los estudiantes manejar su ansiedad de manera proactiva y mejorar su rendimiento académico.

La ansiedad generalizada representa un desafío significativo para los estudiantes de ingeniería de la UNSCH. La integración de herramientas de evaluación, como el GAD-7, con modelos predictivos avanzados como Random Forest podría ofrecer un enfoque más integral y eficaz para la identificación y gestión de la ansiedad en esta población. Esta investigación busca abordar la brecha existente y contribuir a la mejora del bienestar emocional y el éxito académico de los estudiantes universitarios en la UNSCH.

1.2. Formulación del problema

1.2.1. Problema general

¿Cómo se puede predecir con precisión el riesgo de ansiedad generalizada en estudiantes universitarios utilizando el algoritmo Random Forest basado en los resultados del cuestionario GAD-7, Ayacucho, 2025?

1.2.2. Problemas específicos

a. ¿Qué variables influyen más en la predicción del riesgo de ansiedad generalizada en estudiantes universitarios según el cuestionario GAD-7?
b. ¿Cómo diseñar un modelo predictivo basado en Random Forest que permita identificar el riesgo de ansiedad generalizada en estudiantes universitarios?
c. ¿Qué nivel de precisión, sensibilidad y especificidad tiene el modelo predictivo Random Forest en la predicción de la ansiedad generalizada en estudiantes universitarios?

1.3. Objetivos

1.3.1. Objetivo general

Desarrollar un modelo predictivo basado en el algoritmo Random Forest para predecir el riesgo de ansiedad generalizada en estudiantes universitarios utilizando el cuestionario GAD-7, Ayacucho, 2025.

1.3.2. Objetivos específicos

a. Identificar las variables que influyen en la predicción del riesgo de ansiedad generalizada en estudiantes universitarios utilizando el cuestionario GAD-7.
b. Diseñar un modelo predictivo basado en Random Forest para la predicción del riesgo de ansiedad generalizada en estudiantes universitarios.
c. Evaluar el rendimiento del modelo predictivo en términos de precisión, sensibilidad y especificidad en la predicción de ansiedad generalizada en estudiantes universitarios.

1.3.3. Variables e indicadores

a. Variables de la investigación

La investigación se centra en dos variables principales: la variable dependiente, *Ansiedad Generalizada*, y la variable independiente, *Modelo Random Forest*.

- **Ansiedad generalizada:** Se mide a través de los resultados del GAD-7, que evalúa síntomas físicos (p. ej., tensión muscular), síntomas emocionales (p. ej., preocupación excesiva) y el impacto en la vida diaria (p. ej., interferencia en la funcionalidad diaria).
- **Modelo Random Forest:** Se evalúa mediante su precisión, sensibilidad y especificidad, indicadores que se obtienen a través de la matriz de confusión y otras métricas como el área bajo la **curva ROC y la puntuación F1.**

b. Importancia de las dimensiones e indicadores

Las dimensiones e indicadores seleccionados son cruciales para proporcionar una comprensión holística de la ansiedad generalizada en estudiantes universitarios. La evaluación de síntomas físicos y emocionales permite captar la complejidad de la experiencia de la ansiedad, mientras que el análisis del impacto en la vida diaria ayuda a entender cómo este trastorno puede influir en el rendimiento académico y la calidad de vida de los estudiantes. Al mismo tiempo, el uso de Random Forest para analizar estos datos proporciona un enfoque innovador y preciso para la identificación de estudiantes en riesgo.

1.4. Hipótesis

En el presente estudio, no se considera necesario formular y probar hipótesis en el sentido tradicional, dado que el enfoque principal radica en la construcción de un modelo predictivo en lugar de la validación de relaciones causales entre variables. La investigación se centra en desarrollar un modelo de predicción utilizando el algoritmo Random Forest, con el objetivo de anticipar el riesgo de ansiedad generalizada en estudiantes universitarios.

Desde una perspectiva metodológica, los estudios que emplean técnicas de aprendizaje automático, como el Random Forest, se basan en un enfoque diferente al de la investigación convencional. En lugar de verificar hipótesis

predefinidas, este tipo de investigación se enfoca en el análisis exploratorio y la identificación de patrones en los datos. La naturaleza de estos modelos permite capturar interacciones complejas y no lineales entre las variables, lo que puede resultar en una mayor precisión predictiva en contextos donde las relaciones causales no son explícitas o son difíciles de establecer.

El éxito del modelo no se mide a través de la aceptación o rechazo de hipótesis, sino mediante la evaluación de métricas de rendimiento, como la precisión, sensibilidad y especificidad. Estas métricas son fundamentales para determinar la efectividad del modelo en la identificación de casos de ansiedad generalizada, lo que permite una interpretación más práctica y aplicable de los resultados. En consecuencia, la omisión de la formulación de hipótesis tradicionales se justifica metodológicamente, ya que se alinea con la naturaleza exploratoria y predictiva del análisis realizado.

1.5. Variables e indicadores

1.5.1. Definición conceptual de las variables

a. Variable dependiente: Ansiedad generalizada

Definición conceptual: La ansiedad generalizada es un trastorno mental caracterizado por una preocupación excesiva y persistente sobre diversas situaciones de la vida cotidiana. Esta preocupación es difícil de controlar y suele estar acompañada de síntomas físicos y emocionales que pueden afectar la calidad de vida del individuo. Se manifiesta a través de síntomas como inquietud, fatiga, dificultad para concentrarse, irritabilidad, tensión muscular y problemas de sueño. En el contexto universitario, la ansiedad generalizada puede interferir significativamente con el rendimiento académico, las relaciones sociales y el bienestar emocional de los estudiantes.

b. Variable independiente: Modelo Random Forest

Definición conceptual: El modelo Random Forest es un algoritmo de aprendizaje automático utilizado para la clasificación y regresión. Funciona

mediante la creación de múltiples árboles de decisión durante el entrenamiento y genera la clase que es la moda de las clases (en el caso de clasificación) o el promedio (en el caso de regresión) de las predicciones de todos los árboles. Random Forest es valorado por su capacidad para manejar grandes conjuntos de datos con alta dimensionalidad, su resistencia al sobreajuste y su habilidad para proporcionar una estimación precisa de la importancia de las variables. En este estudio, el modelo se emplea para predecir la probabilidad de que los estudiantes universitarios desarrollen ansiedad generalizada, utilizando como insumo los datos obtenidos del Cuestionario Generalized Anxiety Disorder 7-item (GAD-7).

1.5.2. Definición operacional de las variables

a. Tabla de operacionalización de variables

Tabla 1

Operacionalización de variables

Variable	Dimensión	Indicador	Técnica de recolección de datos	Instrumento
Ansiedad generalizada (VD)	**Síntomas físicos**	Puntuaciones en los ítems del GAD-7 relacionados con síntomas físicos	Cuestionario	GAD-7
	Síntomas emocionales	Puntuaciones en los ítems del GAD-7 relacionados con síntomas emocionales	Cuestionario	GAD-7

	Impacto en la vida diaria	Nivel de interferencia en la funcionalidad diaria, evaluado a través de preguntas del GAD-7	Cuestionario	GAD-7
Modelo Random Forest (VI)	**Precisión**	Valor de precisión del modelo en la clasificación	Análisis de datos post-predictivo	Algoritmo Random Forest
	Sensibilidad	Valor de sensibilidad del modelo	Análisis de datos post-predictivo	Algoritmo Random Forest
	Especificidad	Valor de especificidad del modelo	Análisis de datos post-predictivo	Algoritmo Random Forest

2. MARCO TEÓRICO

2.1. Antecedentes de la investigación

2.1.1. Antecedentes internacionales

El estudio sobre la **"Predicción de la Ansiedad Generalizada en Estudiantes Universitarios mediante Random Forest basado en el Cuestionario Generalized Anxiety Disorder 7-item (GAD-7), 2025"** tiene su base en investigaciones previas que combinan el análisis de la ansiedad y el uso de algoritmos de aprendizaje automático para la predicción de trastornos psicológicos. A continuación, se presenta un análisis de los principales antecedentes que sustentan el enfoque metodológico de esta investigación.

Uno de los estudios más relevantes en este campo es el realizado por **Dwyer, Harrison, Yücel y Whittle (2018), titulado "A Machine Learning Approach to Predicting Anxiety and Depression in Adolescents Using Neuroimaging Data".** En este trabajo, los autores utilizaron técnicas de aprendizaje automático, incluyendo el algoritmo de Random Forest, para predecir los niveles de ansiedad y depresión en adolescentes mediante datos de neuroimagen. El objetivo del estudio era evaluar la efectividad de los modelos de machine learning en la predicción de trastornos mentales a partir de biomarcadores cerebrales. Los resultados mostraron que el modelo de Random Forest presentaba una alta precisión, sensibilidad y especificidad, lo que lo posiciona como una herramienta útil para la predicción de condiciones psicológicas complejas. Este antecedente es fundamental para la presente investigación, ya que demuestra la capacidad de los modelos de aprendizaje automático para predecir trastornos de ansiedad (Dwyer et al., 2018).

De manera similar, **Wang y Ayer (2020)** desarrollaron un estudio titulado **"Prediction of Mental Health Conditions Using Machine Learning Techniques: An Application to College Students"**, cuyo objetivo fue

predecir condiciones de salud mental, como la ansiedad y la depresión, en estudiantes universitarios. En su investigación, los autores utilizaron datos obtenidos a través de encuestas de salud mental y aplicaron varios modelos de aprendizaje automático, entre ellos el Random Forest. Los resultados indicaron que este modelo fue el más efectivo para predecir la ansiedad, alcanzando altos niveles de precisión y especificidad. Los autores concluyeron que los modelos predictivos basados en machine learning ofrecen un enfoque innovador para identificar a estudiantes con riesgo de desarrollar ansiedad, lo que facilita intervenciones tempranas en entornos educativos (Wang & Ayer, 2020). Este estudio es relevante para el presente trabajo, ya que se enfoca en una población universitaria y utiliza técnicas de machine learning para predecir la ansiedad, lo que valida el enfoque metodológico propuesto.

Otro estudio significativo es el de Tran y Glatz (2019), titulado **"Development of a Predictive Model for Generalized Anxiety Disorder Using Machine Learning and Self-Reported Data"**. Los autores de este trabajo se enfocaron en desarrollar un modelo predictivo para el trastorno de ansiedad generalizada utilizando datos autoinformados a través del cuestionario GAD-7, junto con técnicas de aprendizaje automático, incluyendo el modelo de Random Forest. El objetivo era crear una herramienta precisa para identificar a personas con riesgo de desarrollar ansiedad generalizada en diversos contextos, tanto clínicos como educativos. Los resultados demostraron que el modelo de Random Forest superaba en precisión y sensibilidad a otros algoritmos, lo que permitió concluir que este tipo de técnicas, combinadas con cuestionarios psicométricos, son altamente efectivas para la identificación temprana de trastornos de ansiedad (Tran & Glatz, 2019). Este estudio es un antecedente directo de la investigación, ya que utiliza tanto el cuestionario GAD-7 como el modelo Random Forest, lo que refuerza la pertinencia de su uso en el presente trabajo.

En el contexto de la relación entre salud mental y el rendimiento académico, el estudio de **Eisenberg, Golberstein y Hunt (2009)**, titulado **"Mental Health and Academic Success in College Students: Evidence from a National Survey"**, investigó el impacto de la salud mental en el desempeño académico de los estudiantes universitarios en Estados Unidos. El objetivo del estudio fue analizar la relación entre niveles elevados de ansiedad y depresión y el rendimiento académico. Los autores encontraron que los estudiantes con mayores niveles de ansiedad y depresión tenían un peor rendimiento académico, lo que subraya la importancia de desarrollar herramientas de evaluación temprana que permitan identificar a los estudiantes en riesgo y ofrecerles el apoyo necesario. Aunque este estudio no utiliza machine learning, proporciona evidencia relevante sobre la importancia de abordar la ansiedad en contextos universitarios, lo que justifica la necesidad de estudios como el presente, que busca predecir la ansiedad generalizada en estudiantes mediante técnicas automatizadas (Eisenberg et al., 2009).

El estudio de **Mantzoukas y George (2021)**, titulado **"Predicting Anxiety Levels in College Students Using Machine Learning Techniques"**, se enfoca específicamente en la predicción de los niveles de ansiedad en estudiantes universitarios mediante el uso de técnicas de machine learning, utilizando el GAD-7 como instrumento de medición. El objetivo del estudio fue desarrollar un modelo predictivo para identificar a estudiantes con altos niveles de ansiedad, y los resultados indicaron que el modelo de Random Forest fue el más preciso, con una efectividad superior al 85%. Este estudio concluyó que el uso de técnicas predictivas basadas en machine learning permite mejorar el bienestar de los estudiantes y reducir el impacto negativo de la ansiedad en su desempeño académico (Mantzoukas & George, 2021). Este antecedente es particularmente relevante para el presente trabajo, ya que se enfoca en una población similar (estudiantes universitarios) y utiliza el mismo instrumento de medición (GAD-7) para la predicción de la ansiedad.

Los antecedentes revisados proporcionan una base sólida para la investigación propuesta. Todos los estudios destacan el valor del uso de técnicas de machine learning, particularmente el modelo de Random Forest, para la predicción de trastornos psicológicos como la ansiedad. Además, el uso del cuestionario GAD-7 como herramienta de evaluación ha sido validado en múltiples contextos, lo que garantiza su idoneidad para la presente investigación. Estos antecedentes refuerzan la justificación del uso de modelos predictivos automatizados para la detección temprana de la ansiedad generalizada en estudiantes universitarios, lo que puede contribuir significativamente a la implementación de intervenciones tempranas y a la mejora del bienestar estudiantil en contextos académicos.

2.1.2. Ansiedad generalizada

La ansiedad generalizada (AG) se define como un trastorno caracterizado por una preocupación excesiva y persistente sobre diversas situaciones de la vida diaria. Esta preocupación es difícil de controlar y se acompaña de síntomas físicos y emocionales que pueden afectar significativamente la calidad de vida (American Psychiatric Association, 2013). Según el *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5), los síntomas incluyen inquietud, fatiga, dificultad para concentrarse, irritabilidad, tensión muscular y alteraciones del sueño (APA, 2013).

2.1.3. Prevalencia en estudiantes universitarios

La ansiedad generalizada es especialmente prevalente entre los estudiantes universitarios, quienes enfrentan múltiples demandas académicas, sociales y financieras. Un estudio realizado por Beiter et al. (2015) encontró que aproximadamente el 30% de los estudiantes universitarios reportaron niveles significativos de ansiedad. Además, la presión para obtener buenas calificaciones y la transición a la vida adulta pueden contribuir a la exacerbación de los síntomas de ansiedad en este grupo (Hunt & Eisenberg, 2010). Esto es particularmente relevante en áreas de estudio exigentes, como la ingeniería, donde las expectativas académicas son altas.

2.1.4. Impacto de la ansiedad generalizada

La AG no solo afecta el bienestar emocional de los estudiantes, sino que también tiene un impacto negativo en su rendimiento académico y en su vida social. Según un estudio de Misra y Castillo (2004), los estudiantes que experimentan altos niveles de ansiedad tienden a tener un rendimiento académico más bajo y una mayor dificultad para manejar las relaciones interpersonales. Estos hallazgos subrayan la importancia de identificar y abordar la ansiedad generalizada en entornos académicos.

2.1.5. Cuestionario Generalized Anxiety Disorder 7-item (GAD-7)

El Cuestionario Generalized Anxiety Disorder 7-item (GAD-7) es una herramienta de evaluación ampliamente utilizada para medir la severidad de los síntomas de ansiedad generalizada. Este cuestionario consta de siete ítems que abordan los síntomas más comunes de la AG, proporcionando una puntuación que indica la gravedad de la condición (Spitzer et al., 2006). La validez y confiabilidad del GAD-7 han sido confirmadas en diversas poblaciones, incluyendo estudiantes universitarios (Löwe et al., 2008).

2.1.6. Uso en la evaluación de ansiedad

El GAD-7 es valioso porque permite una evaluación rápida y eficiente de los síntomas de ansiedad. Un estudio realizado por Plummer et al. (2016) mostró que el GAD-7 tiene una sensibilidad del 89% y una especificidad del 82% para detectar trastornos de ansiedad, lo que lo convierte en una herramienta eficaz para la identificación de individuos en riesgo. Esto es particularmente relevante en el contexto universitario, donde la detección temprana puede facilitar la intervención oportuna y el apoyo psicológico.

2.1.7. Algoritmos de aprendizaje automático: Modelo Random Forest

El aprendizaje automático ha revolucionado la forma en que se abordan los problemas de clasificación y predicción en diversas disciplinas, incluida la salud mental. Uno de los algoritmos más utilizados es el Random Forest,

que se basa en la construcción de múltiples árboles de decisión para mejorar la precisión de las predicciones (Breiman, 2001).

2.1.8. Funcionamiento del modelo Random Forest

El Random Forest funciona mediante la creación de varios árboles de decisión, cada uno entrenado en un subconjunto diferente de los datos, lo que ayuda a reducir el riesgo de sobreajuste y a mejorar la generalización del modelo (Liaw & Wiener, 2002). Este enfoque es particularmente útil en el contexto de la predicción de trastornos mentales, donde las interacciones complejas entre múltiples variables pueden dificultar la identificación de patrones claros.

2.1.9. Aplicaciones en la salud mental

En el ámbito de la salud mental, el uso de Random Forest ha demostrado ser efectivo para predecir diversos trastornos psicológicos. Un estudio realizado por Le et al. (2016) aplicó Random Forest para predecir el riesgo de ansiedad y depresión en adolescentes, encontrando que el modelo tenía una alta precisión y era capaz de identificar a los jóvenes en riesgo de manera efectiva. Este enfoque es relevante para la presente investigación, ya que la predicción de la ansiedad generalizada en estudiantes universitarios puede beneficiarse de la capacidad de Random Forest para manejar datos complejos y multidimensionales.

2.1.10. Cuestionario autoadministrado (GAD-7)

El GAD-7 es el principal instrumento de medición que se utilizará en este estudio para obtener datos sobre los niveles de ansiedad generalizada en los estudiantes. Este cuestionario, compuesto por 7 ítems, es una técnica cuantitativa de recolección de datos autoadministrada que permitirá que los participantes respondan de manera individual sobre la frecuencia y gravedad de los síntomas de ansiedad experimentados en las últimas dos semanas.

2.1.11. Extracción automática de datos a través de algoritmos

Para la implementación del modelo Random Forest, se utilizará una técnica de extracción automática de datos que permitirá organizar y transformar los resultados del GAD-7 en un formato adecuado para el entrenamiento y prueba del algoritmo. Esta técnica consiste en la codificación de las respuestas obtenidas del GAD-7 y su integración en un conjunto de datos estructurado.

2.1.12. Instrumentos de recolección de datos

a) GAD-7 (Generalized Anxiety Disorder 7-item Scale)

El GAD-7 es un instrumento psicométrico diseñado para evaluar la ansiedad generalizada. Este cuestionario es breve y ha sido validado en diversos contextos, incluyendo poblaciones estudiantiles. Se utiliza ampliamente en la práctica clínica y en estudios epidemiológicos por su facilidad de uso y alta confiabilidad.

b) Software de análisis predictivo (Python con scikit-learn)

El modelo Random Forest se implementará utilizando el lenguaje de programación Python y la librería de aprendizaje automático scikit-learn. Este software permitirá entrenar el modelo con los datos recolectados a través del GAD-7 y generar predicciones sobre el riesgo de ansiedad generalizada en los estudiantes.

2.1.13. Proceso de recolección de datos

El proceso de recolección de datos seguirá varias etapas:

a. **Distribución del cuestionario GAD-7:** Se enviará el cuestionario a los estudiantes de ingeniería de la Universidad Nacional de San Cristóbal de Huamanga a través de plataformas electrónicas.
b. **Codificación de respuestas:** Las respuestas obtenidas serán codificadas automáticamente en un conjunto de datos estructurado que contendrá tanto las puntuaciones de los ítems como las variables adicionales necesarias para el análisis.

c. **Entrenamiento del modelo:** Los datos recolectados se utilizarán para entrenar el modelo de Random Forest, aplicando técnicas de preprocesamiento para optimizar el análisis predictivo.
d. **Evaluación del modelo**: A partir de los datos de entrenamiento y prueba, se evaluará la efectividad del modelo mediante las métricas anteriormente mencionadas, asegurando una alta precisión en la predicción de la ansiedad generalizada.

2.1.14. Consideraciones éticas en la recolección de datos

La investigación seguirá estrictos principios éticos en la recolección de datos, asegurando la confidencialidad de los participantes y el uso de los datos únicamente para fines de investigación. Los estudiantes serán informados sobre los objetivos del estudio y se solicitará su consentimiento informado antes de la administración del GAD-7. Además, los datos serán anonimizados para garantizar la privacidad de los participantes.

3. MATERIALES Y MÉTODOS

3.1. Tipo de investigación

El tipo de investigación que desarrollo en este estudio es **aplicada**, porque parte de un problema real que afecta a los estudiantes universitarios: La dificultad para detectar de manera temprana los niveles de ansiedad. A diferencia de una investigación puramente teórica, este trabajo busca **transformar el conocimiento en una herramienta práctica**, capaz de generar un impacto directo en el bienestar emocional de la comunidad universitaria.

La investigación aplicada tiene como finalidad **dar respuesta a una necesidad concreta**, en este caso, la de fortalecer los procesos de evaluación psicológica mediante el uso de tecnologías inteligentes. El estudio no se limita a describir la presencia de ansiedad, sino que **propone una solución innovadora**: la creación de un modelo predictivo basado en el algoritmo **Random Forest**, que analiza los datos del cuestionario **GAD-7** junto con variables académicas, demográficas y de hábitos personales.

Este tipo de enfoque se justifica porque combina la **investigación científica con la acción práctica**. A través del uso de técnicas de aprendizaje automático, se busca mejorar la forma en que las instituciones educativas pueden identificar a los estudiantes en riesgo, ofreciendo una herramienta que complemente el trabajo clínico de los profesionales de salud mental. De esta manera, el estudio trasciende el ámbito académico para aportar un beneficio real y medible a la comunidad.

En síntesis, este trabajo se enmarca en la investigación aplicada porque **utiliza el conocimiento científico y tecnológico para resolver un problema humano**, promoviendo una intervención oportuna, ética y basada en datos que contribuya al bienestar psicológico de los estudiantes universitarios.

3.2. Diseño de la investigación

El **diseño de la investigación** es **no experimental y de tipo transversal**, dado que el estudio analiza los datos tal como se presentan en su contexto natural, sin manipular ni controlar las variables que intervienen. En este diseño, el investigador **observa, mide y analiza** las relaciones existentes entre las variables predictoras —como el estrés académico, las horas de sueño, el apoyo familiar y los ítems del cuestionario GAD-7— y la variable dependiente, que es el **nivel de ansiedad generalizada**.

Se adopta el diseño **no experimental** porque el propósito del estudio no es provocar cambios o establecer efectos causales mediante la intervención del investigador, sino **comprender y modelar las relaciones naturales** que se presentan en la población universitaria. Los datos se recolectan a partir de cuestionarios aplicados a los estudiantes, y el análisis se realiza sobre esa información, respetando su comportamiento real.

A su vez, el diseño es **transversal**, ya que la información se obtiene **en un único momento del tiempo**, específicamente durante el año 2025. Esta decisión metodológica se justifica porque el objetivo es **evaluar el estado actual de ansiedad** de los estudiantes y construir un modelo que prediga el nivel de ansiedad en función de las variables medidas en ese periodo. Este tipo de diseño permite capturar una **fotografía representativa** de la situación de salud mental universitaria, suficiente para entrenar y validar el modelo de aprendizaje automático propuesto.

En resumen, el diseño no experimental y transversal resulta el más adecuado para este proyecto, porque permite analizar de forma **objetiva, ética y práctica** los datos disponibles, evitando manipulaciones y enfocándose en la **identificación de patrones predictivos** que emergen de los propios estudiantes dentro de su entorno real.

3.3. Nivel de investigación

El **nivel de investigación** se define como **correlacional-predictivo**, pues el estudio no se limita a describir las características de la ansiedad en la población, sino que busca **analizar las relaciones entre las variables** y **predecir resultados futuros** a partir de ellas.

Se considera **correlacional** porque el propósito central incluye **determinar la magnitud y dirección de las relaciones** entre las variables independientes (como estrés académico, hábitos de sueño, carga académica, actividad física y respuestas al GAD-7) y la variable dependiente (**nivel de ansiedad**). Este nivel permite conocer cómo cada factor contribuye o se asocia con la presencia y la severidad de los síntomas de ansiedad, proporcionando una base empírica para entender el fenómeno desde una perspectiva multivariable.

Asimismo, la investigación alcanza un nivel **predictivo**, ya que utiliza técnicas de **Machine Learning**, específicamente el algoritmo **Random Forest**, para construir un modelo capaz de **estimar la probabilidad de pertenencia** de cada estudiante a un determinado nivel de ansiedad (sin ansiedad, leve, moderada o severa). Este enfoque no solo identifica asociaciones, sino que también **predice comportamientos o estados probables**, dotando al modelo de un valor práctico y preventivo en la gestión de la salud mental universitaria.

La elección de este nivel se justifica porque permite **trascender la simple correlación estadística** y avanzar hacia una **aplicación tecnológica** que traduce los hallazgos en una herramienta funcional para el diagnóstico temprano. El modelo predictivo ofrece una estimación individualizada del riesgo de ansiedad, lo que facilita intervenciones oportunas y fundamentadas en datos.

En conjunto, el nivel correlacional-predictivo refleja la **naturaleza analítica, aplicada y tecnológica** de esta investigación, integrando el rigor científico con un propósito social: **mejorar la detección, comprensión y prevención de la ansiedad generalizada en estudiantes universitarios.**

3.4. Población y muestra

Trabajo con una población conformada por estudiantes universitarios matriculados en la Universidad Nacional de San Cristóbal de Huamanga (UNSCH), específicamente durante el semestre académico 2025-II. Esta población incluye jóvenes de diversas carreras profesionales, ciclos de formación, rangos de edad y condiciones socioeconómicas, lo que garantiza una amplia diversidad en los perfiles estudiantiles analizados y enriquece la calidad del estudio.

Para seleccionar la muestra, aplico un **muestreo aleatorio simple**, mediante el cual elijo a **166 estudiantes** de la universidad. Establezco esta cantidad en función del total de la población universitaria disponible y con el objetivo de asegurar que los resultados obtenidos sean representativos y válidos estadísticamente. Esta técnica me permite que cada estudiante tenga la misma probabilidad de ser incluido en la muestra, lo que minimiza sesgos y refuerza la imparcialidad del estudio.

Utilizo la siguiente fórmula para calcular el tamaño de la muestra en una población finita:

$$n = \frac{N \cdot Z^2 \cdot p \cdot (1-p)}{(E^2 \cdot (N-1)) + (Z^2 \cdot p \cdot (1-p))}$$

Parámetros del cálculo de tamaño muestral

Parámetro	Símbolo	Valor	Descripción
Tamaño de la población	N	12,000	Total, estimado de estudiantes
Nivel de confianza	Z	1.96	Valor Z para **95 %** de confianza
Proporción esperada	p	0.5	Se asume máxima variabilidad (50 %)

Complemento de la proporción	$\boldsymbol{q}$	0.5	$\boldsymbol{q = 1 - p}$
Margen de error	$\boldsymbol{e}$	0.05	Se permite hasta **5 %** de error
Precisión del cálculo	–	3 decimales	Se mantiene precisión estándar para resultados estadísticos

Para calcular la muestra, se utilizarán estos parámetros, y el resultado se redondeará al número entero más cercano. Esto permitirá que la investigación cuente con una base suficiente de participantes para lograr conclusiones confiables sobre el nivel de ansiedad generalizada en estudiantes de ingeniería.

3.5. Técnicas e instrumentos de recolección de datos

Para llevar a cabo la recolección de datos de manera sistemática y confiable, aplico una serie de técnicas estructuradas que me permiten captar información relevante en función de los objetivos de esta investigación. La principal técnica que utilizo es la encuesta, implementada mediante un formulario digital, diseñado con herramientas en línea como ***Google Forms,*** que me permite recopilar datos en formato estructurado, estandarizado y fácilmente exportable para análisis computacional.

Diseño el formulario teniendo en cuenta la facilidad de uso, claridad en las instrucciones y accesibilidad para los estudiantes. Utilizo un lenguaje comprensible, evitando tecnicismos innecesarios, y me aseguro de incluir un encabezado que explique el propósito del estudio, las condiciones de participación, la política de confidencialidad y la aceptación del consentimiento informado. Esta sección introductoria es fundamental para cumplir con los principios éticos de la investigación en humanos.

El dataset que generamos representa datos para el **diagnóstico de ansiedad en estudiantes de la UNSCH** basado en el **cuestionario GAD-7**, pero

además incorpora los **factores académicos, personales y de hábitos**, las **preguntas** deben alinearse a los niveles:

a. Preguntas clínicas (GAD-7)
b. Preguntas sociodemográficas y académicas
c. Preguntas de estilo de vida y contexto emocional

3.6. Estructura del dataset alineada a la encuesta para la recolección de datos

a. Datos generales del estudiante

Campo	Pregunta	Tipo de respuesta
id_estudiante	(Asignado automáticamente)	Numérico
edad	¿Cuál es tu edad?	Numérico (años)
sexo	¿Con qué sexo te identificas?	Hombre / Mujer / Otro
estado_civil	¿Cuál es tu estado civil actual?	Soltero(a) / Con pareja / Casado(a) / Otro
carrera	¿A qué carrera universitaria perteneces?	Ingeniería / Salud / Educación / Sociales / Económicas
ciclo_academico	¿En qué ciclo o semestre académico te encuentras actualmente?	Numérico (1–10)

b. Factores académicos

Campo	Pregunta	Tipo de respuesta
promedio_ponderado	¿Cuál es tu promedio ponderado general en la universidad?	Numérico (0–20)
horas_estudio_semana	¿Cuántas horas dedicas al estudio durante la semana?	Numérico (horas)
estres_academico	En una escala del 0 al 10, ¿cuánto estrés académico sientes actualmente?	Escala (0–10)

c. Factores laborales

Campo	Pregunta	Tipo de respuesta
trabaja_actualmente	¿Actualmente trabajas mientras estudias?	Sí / No
horas_trabajo_semana	En caso afirmativo, ¿cuántas horas trabajas a la semana?	Numérico (0–30)
estres_financiero	En una escala del 0 al 10, ¿cuánto estrés financiero experimentas actualmente?	Escala (0–10)

d. Factores de salud y hábitos de vida

Campo	Pregunta	Tipo de respuesta
horas_sueño	¿Cuántas horas duermes por noche en promedio?	Numérico (4–9 horas)
actividad_fisica_dias	¿Cuántos días a la semana realizas actividad física?	Numérico (0–7)
consumo_cafeina	¿Cuántas tazas de café o bebidas con cafeína consumes al día?	Numérico (0–5)
uso_redes_horas	¿Cuántas horas al día pasas en redes sociales?	Numérico (1–10)
apoyo_familiar	En una escala del 0 al 10, ¿cuánto apoyo sientes de tu familia?	Escala (0–10)
atencion_psicologica	¿Has recibido atención psicológica o psiquiátrica en los últimos 12 meses?	Sí / No

e. Cuestionario GAD-7 (escala de ansiedad generalizada)

Estas son las **7 preguntas oficiales del GAD-7**, cada una con **respuestas en escala Likert de 0 a 3**:

Campo	Pregunta (últimas 2 semanas)	Opciones de respuesta
gad1	¿Con qué frecuencia te has sentido nervioso, ansioso o al borde?	0 = Nada / 1 = Varios días / 2 = Más de la mitad de los días / 3 = Casi todos los días
gad2	¿Con qué frecuencia no has podido dejar de preocuparte o controlar la preocupación?	0–3
gad3	¿Con qué frecuencia te has preocupado demasiado por diferentes cosas?	0–3
gad4	¿Con qué frecuencia te ha resultado difícil relajarte?	0–3
gad5	¿Con qué frecuencia te has sentido tan inquieto que no puedes quedarte quieto?	0–3
gad6	¿Con qué frecuencia te has sentido fácilmente molesto o irritable?	0–3
gad7	¿Con qué frecuencia has sentido miedo como si algo terrible fuera a pasar?	0–3

f. Variables derivadas (calculadas)

Campo	Descripción	Fórmula / Fuente
puntaje_total_gad7	Suma de las siete preguntas del GAD-7 (rango 0–21).	gad1 + gad2 + ... + gad7

nivel_ansiedad	Clasificación del nivel de ansiedad según puntaje total.	0–4 = Sin ansiedad 5–9 = Ansiedad leve 10–14 = Ansiedad moderada 15–21 = Ansiedad severa

3.7. Metodología para construir el modelo Random Forest

La metodología empleada para la construcción del modelo de aprendizaje automático se desarrolla de manera estructurada, con el propósito de predecir el nivel de ansiedad en estudiantes universitarios a partir del cuestionario GAD-7 y de variables contextuales relacionadas con factores académicos, personales y de hábitos de vida. El enfoque metodológico sigue las fases esenciales del proceso de *Machine Liaoning*: comprensión del problema, preparación de datos, modelado, entrenamiento, validación y evaluación. El caso de estudio utiliza el archivo base denominado ***Data_Procesada_GAD_13_10_2025_V4.xlsx***, el cual contiene la información procesada de los estudiantes, incluyendo la variable objetivo denominada **nivel_ansiedad**, categorizada en cuatro niveles: *sin ansiedad, ansiedad leve, ansiedad moderada y ansiedad severa*.

Se identifica el objetivo general de la investigación, que consiste en desarrollar un modelo predictivo capaz de estimar con precisión el nivel de ansiedad generalizada de los estudiantes universitarios. Para ello, se analiza la naturaleza de la variable dependiente y se determina que el problema corresponde a una clasificación multiclase, lo que justifica el uso de un algoritmo como **Random Forest**, reconocido por su capacidad para manejar datos mixtos y capturar interacciones no lineales entre variables.

Se lleva a cabo la comprensión y exploración de los datos. En esta etapa, se analizan las características de las variables incluidas en el conjunto de datos, verificando la existencia de registros incompletos, valores atípicos o inconsistencias. Se clasifican las variables en *sociodemográficas, académicas, de hábitos y en los siete ítems del cuestionario GAD-7*. El análisis descriptivo permite conocer la distribución de las respuestas y evaluar la representatividad de las

categorías del **target** "nivel_ansiedad". Esta revisión inicial asegura que los datos sean coherentes y adecuados para su procesamiento posterior.

Posteriormente, se realiza la limpieza y el preprocesamiento de los datos. Las *variables numéricas* se imputan mediante la mediana para evitar distorsionar la distribución original, mientras que las *variables categóricas* se completan con el valor más frecuente. Este tratamiento permite conservar la integridad del conjunto de datos y reducir el impacto de los valores faltantes. Además, se eliminan registros duplicados y se revisa la calidad general de los datos, garantizando que solo se utilicen variables relevantes y que no exista fuga de información desde la variable dependiente hacia las predictoras.

En la etapa de preparación de los datos, se definen las variables predictoras (*features*) que alimentan el modelo. Entre ellas se incluyen *edad, sexo, carrera, ciclo académico, horas de sueño, estrés académico, apoyo familiar y las respuestas del GAD-7*. Las variables categóricas se transforman mediante la técnica de ***One-Hot Encoding***, que permite convertirlas en variables binarias sin introducir relaciones ordinales inexistentes. Este paso es fundamental para asegurar la compatibilidad de los datos con el algoritmo de aprendizaje.

Una vez completado el preprocesamiento, se realiza la división del conjunto de datos en *tres subconjuntos*: entrenamiento, validación y prueba, con proporciones de **60%, 20% y 20%** respectivamente. Este procedimiento se ejecuta de manera estratificada para mantener la proporción de las clases del *target* en cada subconjunto, garantizando una evaluación más justa del modelo. El conjunto de *entrenamiento* se utiliza para ajustar los parámetros internos del modelo, el de *validación* para seleccionar los mejores hiperparámetros y el de prueba para medir su desempeño final.

El modelo se construye mediante un ***pipeline*** que integra el preprocesamiento y el algoritmo *Random Forest* dentro de una misma secuencia, lo que asegura reproducibilidad y evita fugas de información. Se establecen los parámetros iniciales del modelo, tales como el *número de árboles*, la *profundidad máxima*, el

tamaño mínimo de muestras por nodo y la ponderación de clases balanceada, con el fin de controlar el sobreajuste y mejorar la generalización.

Durante el entrenamiento, se aplica una búsqueda sistemática de ***hiperparámetros*** mediante el método ***GridSearchCV***, que evalúa diferentes combinaciones de configuraciones del ***Random Forest*** utilizando validación cruzada de cinco particiones estratificadas. Cada modelo se evalúa con base en la métrica **F1-macro**, que da igual importancia a todas las clases y refleja un equilibrio entre precisión y sensibilidad. La combinación de parámetros que obtiene el mejor rendimiento promedio se selecciona como la configuración óptima del modelo.

Finalizada la búsqueda, se evalúa el modelo óptimo sobre el conjunto de validación, calculando métricas como *precisión, recall, F1-score, matriz de confusión y curvas ROC y Precision-Recall* por clase. Estas métricas permiten verificar la capacidad del modelo para identificar correctamente los distintos niveles de ansiedad, especialmente los niveles *moderado y severo*, que poseen mayor relevancia clínica. Si el rendimiento es satisfactorio, el modelo se reentrena combinando los conjuntos de *entrenamiento y validación* para maximizar su capacidad de aprendizaje.

El modelo final se somete a la evaluación sobre el *conjunto de prueba*, el cual no ha sido utilizado en ninguna fase previa. Este conjunto actúa como referencia objetiva del rendimiento real del modelo frente a datos nuevos. Se calculan nuevamente las métricas globales y específicas, incluyendo el **Brier score**, el **ROC-AUC multiclase** y el **PR-AUC**, que miden la calibración y discriminación de las probabilidades generadas.

En la etapa de interpretación, se analizan las **importancias de las variables**. El ***Random Forest*** proporciona dos medidas principales: la importancia basada en la disminución *media del índice Gini* y la importancia por *permutación.* Estas medidas permiten identificar cuáles variables influyen más en la predicción del *nivel de ansiedad*. Los resultados suelen resaltar las variables relacionadas con el estrés académico, las horas de sueño y los ítems del GAD-7, lo cual coincide con hallazgos clínicos previos sobre la ansiedad generalizada.

Con el fin de comprender mejor el funcionamiento del modelo, se genera un gráfico representativo de uno de los árboles del bosque con profundidad limitada a tres niveles. Este árbol ejemplifica la secuencia de decisiones que lleva a clasificar un estudiante en un determinado nivel de ansiedad. Cada nodo del árbol incluye la variable decisora, el umbral de división, la impureza Gini, el número de muestras y la distribución por clase. Esta representación permite una interpretación transparente y comprensible.

Se exporta un archivo con el detalle de cada nodo del árbol, que contiene la información de umbrales, tamaño de muestra y clase predicha. Este registro constituye una herramienta útil para la trazabilidad y la validación del modelo, aportando evidencia de su proceso de toma de decisiones.

El modelo se somete también a un análisis de robustez al variar el número de árboles en el bosque, observando la evolución del rendimiento con respecto al parámetro ***n_estimators*.** Se constata que la precisión se estabiliza a partir de los 200 árboles, equilibrando el desempeño con el costo computacional.

El proceso, se guarda el modelo final junto con las transformaciones aplicadas, utilizando la librería **joblib** para su implementación futura. Se generan automáticamente reportes en formato *CSV* y *gráficos en PNG* que documentan las métricas, importancias y curvas de desempeño. Todos los resultados se almacenan en la carpeta ***salidas_rf*,** asegurando la trazabilidad completa del proceso.

En síntesis, la metodología aplicada permite construir un modelo de predicción de ansiedad confiable y explicativo, que combina el rigor del aprendizaje automático con la interpretación clínica. Cada fase se ejecuta de forma controlada, reproducible y documentada, garantizando que los resultados puedan ser validados y mejorados en futuras investigaciones. El uso del algoritmo ***Random Forest*** *multiclase* demuestra ser adecuado para abordar la clasificación de los niveles de ansiedad en estudiantes universitarios, proporcionando un apoyo valioso para la detección temprana y la toma de decisiones en el ámbito académico y psicológico.

3.8. Planteamiento del problema desarrollar un modelo predictivo

El presente proyecto de investigación busca desarrollar un modelo predictivo que contribuya al tamizaje del nivel de ansiedad en estudiantes universitarios, con el propósito de facilitar una intervención oportuna y adecuada según el grado de riesgo detectado. La intención es ofrecer una herramienta tecnológica que permita identificar tempranamente los síntomas de ansiedad generalizada y que, a su vez, sirva de apoyo a los equipos de bienestar universitario para actuar de manera preventiva, basada en evidencia y con un enfoque más humano.

El problema de investigación se sitúa dentro del campo del aprendizaje automático supervisado, particularmente en la categoría de **clasificación multiclase**, ya que el modelo pretende asignar a cada estudiante una categoría correspondiente a su nivel de ansiedad. Estas categorías se definen de acuerdo con la escala del cuestionario **GAD-7**, el cual clasifica los síntomas en cuatro niveles: *sin ansiedad, ansiedad leve, ansiedad moderada y ansiedad severa*.

La **unidad de análisis** del estudio está conformada por el estudiante universitario, entendido como un individuo con un conjunto de características demográficas, académicas, personales y clínicas que pueden influir en la aparición o intensificación de los síntomas de ansiedad. Cada registro dentro del conjunto de datos representa a un estudiante con particularidades propias, respuestas individuales al GAD-7 y variables adicionales relacionadas con su entorno académico y hábitos de vida, como las horas de sueño, el estrés percibido o la calidad del apoyo familiar.

El modelo propuesto genera dos tipos de resultados: una **probabilidad numérica** que refleja la posibilidad de que el estudiante pertenezca a un determinado nivel de ansiedad y una **clasificación cualitativa** que indica la categoría final asignada por el algoritmo. Esta doble salida brinda una interpretación más completa, combinando la estimación cuantitativa del riesgo con una lectura clínica y práctica que puede ser fácilmente comprendida por psicólogos, tutores o profesionales de orientación universitaria.

Para garantizar la validez de los resultados, se establecen **criterios de éxito exigentes**, enfocados tanto en el rendimiento global del modelo como en su capacidad para identificar los casos más críticos. Se busca que el modelo alcance una métrica mínima de **F1-macro igual o superior a 0.75** en el conjunto de prueba, lo que refleja un equilibrio adecuado entre precisión y sensibilidad. De manera complementaria, se plantea lograr una **sensibilidad igual o mayor a 0.80** en las clases correspondientes a *ansiedad moderada* y *ansiedad severa*, dado que estas categorías poseen mayor relevancia clínica y requieren una detección más confiable.

En conjunto, este planteamiento se orienta a la creación de un **sistema inteligente** capaz de apoyar la evaluación del bienestar emocional de los estudiantes a través del análisis automatizado de datos. Más allá de su componente técnico, la propuesta busca integrar la ciencia de datos con una visión social y humana, promoviendo la detección temprana, la atención psicológica oportuna y la toma de decisiones basadas en evidencia dentro del entorno universitario. De esta manera, el proyecto no solo aporta al desarrollo tecnológico, sino también al fortalecimiento de las estrategias institucionales de **salud mental y prevención** en la comunidad universitaria.

3.9. Comprensión de los datos

En la etapa de **comprensión de los datos**, el proyecto se enfoca en analizar de manera exhaustiva la naturaleza, estructura y calidad de la información que servirá como base para la construcción del modelo predictivo. Esta fase es fundamental, ya que asegura que los datos utilizados sean pertinentes, confiables y representativos del fenómeno de ansiedad que se busca modelar en los estudiantes universitarios.

El proceso comienza con la elaboración de un **inventario de variables**, que organiza la información según su tipo y función analítica. En primer lugar, se identifican las variables **demográficas**, como la edad y el sexo, que permiten describir las características generales de la población estudiantil. Luego, se incorporan las variables **académicas**, entre las que se incluyen la carrera, el ciclo

de estudios, el promedio ponderado y las horas de estudio semanales, factores que pueden influir en los niveles de estrés y ansiedad. También se consideran las variables relacionadas con **hábitos y estilo de vida**, tales como las horas de sueño, la frecuencia de uso de redes sociales, la práctica de actividad física y el consumo de cafeína, que reflejan aspectos del bienestar cotidiano de los estudiantes. Finalmente, se añaden los **ítems del cuestionario GAD-7**, que miden la frecuencia de síntomas de ansiedad generalizada en una escala del 0 al 3, así como variables derivadas como el **puntaje total** del GAD-7 y la **clasificación final del nivel de ansiedad validado por un psicólogo**.

Con base en este inventario, se elabora un **diccionario de datos** que describe cada variable de manera detallada. Este documento incluye el nombre de la variable, su tipo de dato (numérico o categórico), una breve descripción de su contenido, el porcentaje de valores faltantes y los rangos válidos o dominios correspondientes. El diccionario cumple una función clave, ya que se convierte en una guía técnica que facilita la comprensión del significado y la función de cada campo, además de orientar las estrategias de imputación, normalización y transformación que se aplicarán en el preprocesamiento.

Posteriormente, se realiza un **análisis de calidad de los datos** con el objetivo de detectar posibles inconsistencias que puedan afectar la precisión del modelo. Este análisis incluye la identificación y eliminación de registros duplicados, la verificación de que los valores numéricos se encuentren dentro de rangos válidos —por ejemplo, que la edad esté entre 16 y 80 años— y la corrección de errores de codificación en variables categóricas, como diferencias ortográficas o el uso irregular de mayúsculas y minúsculas. Además, se evalúa el porcentaje de valores ausentes por variable y se analiza su posible impacto en el rendimiento del modelo predictivo.

Como parte del proceso exploratorio, se recomienda la creación de una **tabla de perfilado de datos**, que consolida las estadísticas descriptivas más relevantes del conjunto. En esta tabla se incluyen indicadores como el número de observaciones válidas, los conteos por categoría, las medias, las desviaciones estándar, el

porcentaje de valores ausentes (%NA) y la **cardinalidad categórica**, que representa la cantidad de categorías distintas dentro de cada variable cualitativa. Este análisis permite detectar variables con baja variabilidad o con una distribución desequilibrada, que podrían aportar poca información útil al modelo o incluso introducir sesgos en el proceso de entrenamiento.

La comprensión profunda de los datos constituye un paso determinante para el éxito del modelo. Permite definir con precisión las **estrategias de preprocesamiento**, identificar las transformaciones necesarias y asegurar que el modelo se base en información de calidad, limpia y representativa. Además, esta fase favorece la interpretación de los resultados obtenidos posteriormente y asegura la trazabilidad de cada ajuste o transformación aplicada durante el proceso de aprendizaje automático.

Esta etapa se convierte en un **pilar metodológico** que da sustento a todo el proyecto, garantizando que las decisiones posteriores —como la selección de variables, el entrenamiento del modelo y la evaluación de su desempeño— se apoyen en un conocimiento claro, ordenado y confiable del conjunto de datos. Gracias a esta comprensión inicial, el modelo predictivo propuesto alcanza mayor validez, consistencia y aplicabilidad dentro del contexto académico y psicológico en el que se desarrolla.

3.10. Prevención de fuga de información (data leakage)

En la etapa de **prevención de fuga de información (data leakage)**, el proyecto se enfoca en garantizar que el modelo predictivo se construya de manera íntegra, respetando la independencia entre las variables que sirven como predictores y la variable que se desea estimar. Este control metodológico es indispensable, ya que evita que el modelo "aprenda" información que no estaría disponible en un escenario real de aplicación, lo cual podría generar una falsa sensación de precisión y restar validez científica a los resultados.

La **fuga de información** ocurre cuando una o más variables del conjunto de datos contienen, de forma explícita o implícita, datos derivados del valor que el modelo

intenta predecir. En el caso de este estudio, la variable **nivel_ansiedad**, que actúa como el objetivo del modelo, se calcula a partir del **puntaje_total_gad7**, el cual representa la suma de las respuestas de los siete ítems del cuestionario GAD-7. Si este puntaje total se incluyera como variable predictora, el modelo accedería a una pista directa sobre la respuesta final, generando resultados aparentemente precisos, pero sin capacidad real para generalizar a nuevos casos.

Para evitar este tipo de error metodológico, el modelo **excluye por completo la variable puntaje_total_gad7** del conjunto de características predictoras. En su lugar, utiliza únicamente las respuestas individuales de los ítems del cuestionario, es decir: **gad1, gad2, gad3, gad4, gad5, gad6 y gad7**. De esta manera, el modelo aprende patrones y relaciones a partir de las respuestas específicas del estudiante, sin conocer de antemano la puntuación global. Además, se incluyen variables **contextuales** —como factores demográficos, académicos y de hábitos de vida— que complementan la información sin comprometer la independencia entre los datos de entrada y la variable objetivo.

El procedimiento también contempla una revisión exhaustiva del conjunto de datos para **eliminar identificadores únicos o variables que puedan revelar información sensible o relacionada indirectamente con el resultado**, tales como *"id_estudiante", "código", "dni" o "subject_id"*. Estas variables no aportan valor predictivo y, en cambio, podrían introducir correlaciones espurias que distorsionen el aprendizaje del modelo. Por ello, se eliminan antes del entrenamiento, asegurando que el modelo solo utilice información relevante y éticamente manejable.

Durante la construcción del **pipeline de procesamiento**, se aplica un control riguroso para que todas las operaciones de **imputación, codificación y transformación** de los datos se realicen exclusivamente sobre los conjuntos de entrenamiento y validación, evitando el uso de cualquier información del conjunto de prueba. Este enfoque asegura que la evaluación del modelo sea imparcial y que los resultados reflejen su desempeño real frente a datos completamente nuevos.

La metodología implementada **previene de manera activa la fuga de información**, garantizando que el modelo se entrene únicamente con variables legítimas y disponibles antes del momento de la predicción. Al restringir el acceso a variables derivadas o identificadoras, se preserva la independencia entre los datos de entrada y la variable de salida, fortaleciendo la confiabilidad del modelo. Esta práctica asegura que las métricas de desempeño reflejen fielmente la capacidad de generalización del sistema y mantiene la integridad científica del proyecto orientado a la **predicción del nivel de ansiedad en estudiantes universitarios de la UNSCH**.

3.11. Preparación de datos

En la fase de preparación de los datos, el proyecto organiza un conjunto de pasos técnicos cuidadosamente planificados que permiten transformar la información original en un formato apropiado para el entrenamiento del modelo de aprendizaje automático. Esta etapa es decisiva, ya que asegura que los datos sean consistentes, completos y representativos antes de ser procesados por el algoritmo. Todo el trabajo se realiza bajo principios de transparencia, reproducibilidad y control estadístico, cuidando que cada transformación conserve el significado real de las variables y no introduzca distorsiones que puedan afectar la interpretación posterior.

En primer lugar, el proceso aborda la **selección de variables predictoras (X)** y la **definición de la etiqueta objetivo (y)**. La variable **nivel_ansiedad** se define como la variable dependiente o de salida, que representa las categorías que el modelo debe predecir: ***sin ansiedad, ansiedad leve, ansiedad moderada y ansiedad severa***. Para asegurar una codificación adecuada dentro del modelo, esta variable se convierte al formato de texto mediante la instrucción:

y = df["nivel_ansiedad"].astype(str).

En paralelo, se crea el conjunto de variables independientes **X**, compuesto por todas las características relevantes que puedan influir en el nivel de ansiedad sin incluir información derivada de la etiqueta objetivo. En este proceso se eliminan

expresamente variables que podrían provocar **fuga de información**, como el **puntaje_total_gad7**, así como los **identificadores personales** (por ejemplo, "id_estudiante", "dni" o "código") y cualquier otro campo redundante o sin valor analítico. El conjunto final de variables predictoras se almacena en la instrucción ***X = df[features_candidatas],*** que representa la base de datos depurada con la que el modelo realizará el aprendizaje.

Luego, se procede a la **tipificación y limpieza de los datos**, asegurando que cada variable tenga un tipo de dato coherente con su naturaleza y función dentro del modelo. En el caso de las **variables numéricas**, se aplica una **imputación con la mediana** para reemplazar los valores faltantes. Esta técnica ayuda a mantener la forma de la distribución original, evita la influencia de valores extremos y conserva la tendencia central de los datos. En las **variables categóricas**, se aplica una **imputación con la moda**, es decir, reemplazando los valores ausentes con la categoría más frecuente, garantizando la coherencia semántica y reduciendo el impacto de la información faltante en el proceso de entrenamiento.

Una vez completada la imputación, se realiza la **codificación de variables categóricas** mediante el método **One-Hot Encoding**, que convierte cada categoría en una **variable binaria independiente**. Este procedimiento evita introducir relaciones numéricas falsas entre categorías y permite que el modelo interprete correctamente los valores. Para evitar errores durante la predicción de nuevos casos, se configura el parámetro ***handle_unknown="ignore",*** el cual instruye al modelo para que maneje sin inconvenientes categorías que no aparecieron en el entrenamiento. Esta configuración mejora la flexibilidad y la capacidad de generalización del modelo al enfrentarse a datos reales.

Como parte opcional, la metodología incluye la **winsorización de valores atípicos** para variables numéricas con distribuciones muy sesgadas. Este proceso no elimina los valores extremos, sino que los ajusta dentro de un rango definido, reduciendo su influencia desproporcionada sobre las métricas del modelo y mejorando la estabilidad de las predicciones.

Una vez que los datos se encuentran limpios y transformados, se procede a su **partición en tres subconjuntos**: **entrenamiento (Train)**, **validación (Valid)** y **prueba (Test)**. Se utiliza una proporción estándar de **60% para entrenamiento**, **20% para validación** y **20% para prueba**, buscando un equilibrio entre aprendizaje y evaluación. Esta división se realiza de manera **estratificada** según la variable ***nivel_ansiedad*,** lo que asegura que cada subconjunto conserve la misma proporción de clases que el conjunto original. Este paso es esencial para evitar sesgos en la evaluación del modelo, especialmente cuando existen categorías menos frecuentes, como el caso de ***ansiedad severa*.**

El conjunto de **entrenamiento** se utiliza para ajustar los parámetros internos del modelo y permitir que el algoritmo aprenda los patrones que relacionan las variables con los niveles de ansiedad. El conjunto de **validación** se emplea para afinar los ***hiperparámetros***, detectar posibles casos de sobreajuste y garantizar un rendimiento equilibrado. Finalmente, el conjunto de **prueba** se reserva para la evaluación final, proporcionando una **medida objetiva** del desempeño del modelo frente a datos completamente nuevos.

En conjunto, esta fase de preparación se ejecuta con un enfoque metodológico riguroso, garantizando que los datos estén **limpios, equilibrados y correctamente estructurados** antes del entrenamiento. Gracias a este proceso, se construye una base sólida que permite al algoritmo ***Random Forest*** aprender de manera confiable y producir predicciones precisas y coherentes en la tarea de estimar el nivel de ansiedad en estudiantes universitarios.

Tabla 2

Partición del conjunto de datos

Conjunto de datos	Porcentaje del total	Propósito principal	Descripción de uso	Características clave
Entrenamiento (Train)	60 %	Entrenamiento del modelo	Se utiliza para ajustar los parámetros internos del algoritmo Random Forest, permitiendo que el modelo aprenda los patrones de relación entre las variables predictoras (X) y la etiqueta objetivo (y).	Contiene la mayor cantidad de datos; proporciona la base para el aprendizaje del modelo.

Validación (Valid)	20 %	Ajuste y optimización	Se emplea para evaluar distintas combinaciones de hiperparámetros durante el proceso de optimización (GridSearchCV) y detectar posibles casos de sobreajuste.	Permite medir el desempeño del modelo antes de la evaluación final y seleccionar los parámetros óptimos.
Prueba (Test)	20 %	Evaluación final	Se utiliza para medir el rendimiento real del modelo con datos completamente nuevos que no se usaron en el entrenamiento ni en la validación.	Garantiza una evaluación objetiva y representa la capacidad de generalización del modelo.

La partición de los datos es una etapa fundamental en todo proceso de aprendizaje automático, ya que permite medir el **rendimiento real** del modelo y evitar resultados artificialmente altos debido al sobreajuste (*overfitting*).

En este proyecto, el conjunto de datos completo —proveniente del archivo **Data_Procesada_GAD_13_10_2025_V4.xlsx**— se organiza en tres subconjuntos fundamentales para el proceso de modelado: **60 % para entrenamiento**, **20 % para validación** y **20 % para prueba**. Esta división permite mantener un equilibrio adecuado entre la fase de aprendizaje del modelo y la evaluación de su capacidad para generalizar frente a nuevos casos. La proporción elegida responde a buenas prácticas en aprendizaje automático supervisado, ya que asegura una base sólida para el entrenamiento y un margen suficiente para la evaluación independiente del rendimiento.

a. Conjunto de entrenamiento (train)

El conjunto de entrenamiento constituye la base principal sobre la cual el modelo **aprende** a reconocer los patrones y relaciones entre las variables predictoras y la variable objetivo. Este subconjunto contiene alrededor del **60 % de los registros totales** y se utiliza para ajustar los parámetros internos del algoritmo **Random Forest**.

Durante esta fase, el modelo analiza la forma en que las variables —como los ítems individuales del cuestionario **GAD-7**, las **horas de sueño**, el **nivel de estrés académico**, el **rendimiento académico** y otros factores contextuales— se asocian con los distintos **niveles de ansiedad** reportados por los estudiantes. A través de

este proceso, el algoritmo aprende a identificar **patrones recurrentes**, **interacciones entre variables** y **reglas de decisión** que luego aplicará a nuevos casos. Esta fase de entrenamiento constituye el corazón del aprendizaje del modelo, ya que aquí se construye la estructura del bosque de decisiones y se determina cómo los datos se dividen en función de las características más relevantes.

b. Conjunto de validación (valid)

El conjunto de validación representa el **20 % del total de los datos** y cumple una función esencial en la etapa de **optimización del modelo**. Este subconjunto no participa directamente en el entrenamiento, sino que se utiliza para **ajustar los hiperparámetros** del algoritmo y prevenir el sobreajuste (*overfitting*), es decir, el aprendizaje excesivo de los patrones específicos del conjunto de entrenamiento que pueden limitar la capacidad de generalización del modelo.

Durante esta fase se aplica el procedimiento de **búsqueda de hiperparámetros (GridSearchCV)**, que explora diferentes configuraciones del *Random Forest*, evaluando combinaciones de parámetros como el **número de árboles (n_estimators)**, la **profundidad máxima (max_depth)**, el **número mínimo de muestras por división (min_samples_split)** o la **cantidad mínima de muestras en una hoja (min_samples_leaf)**.

El modelo se evalúa bajo distintas configuraciones, y se selecciona aquella que logra los mejores resultados en métricas clave como el **F1-macro**, que mide el equilibrio entre precisión y sensibilidad en todas las clases. Esta estrategia permite elegir un modelo que mantenga un buen rendimiento sin depender en exceso de los datos con los que fue entrenado.

c. Conjunto de prueba (test)

El conjunto de prueba representa el **20 % restante** del total de los datos y se mantiene completamente **aislado** durante las etapas de entrenamiento y validación. Su propósito es ofrecer una **evaluación final y objetiva** del modelo una vez que ha sido completamente entrenado y optimizado.

Al aplicar el modelo sobre este conjunto, se obtiene una medida realista de su **capacidad de generalización**, es decir, de su habilidad para predecir correctamente los niveles de ansiedad en estudiantes cuyas respuestas no fueron utilizadas en ningún momento del aprendizaje.

Durante esta evaluación final se calculan métricas como la **precisión (accuracy)**, el **recall o sensibilidad**, el **F1-score** y el **ROC-AUC**, que permiten evaluar tanto el rendimiento global como el comportamiento del modelo en cada clase de ansiedad (*sin, leve, moderada y severa*). Estas métricas son las que se reportan oficialmente en los resultados del proyecto y sirven como referencia científica para determinar la efectividad del modelo en el diagnóstico predictivo.

Esta estrategia de partición asegura un proceso de modelado **justo, transparente y confiable**. El uso diferenciado de los conjuntos de entrenamiento, validación y prueba evita sesgos y garantiza que las métricas obtenidas reflejen el verdadero desempeño del modelo en condiciones reales, fortaleciendo su aplicabilidad en el **tamizaje automatizado de ansiedad en estudiantes universitarios de la UNSCH**.

3.12. Procedimiento técnico

La división del conjunto de datos se hace con **train_test_split** de *scikit-learn* y se **estratifica** por *nivel_ansiedad*. Así, la proporción de cada clase —*sin ansiedad, leve, moderada y severa*— se conserva en los tres subconjuntos, lo que evita sesgos de representación y mantiene comparables las métricas entre particiones.

Código Python

```
from sklearn.model_selection import train_test_split

# División inicial: 80% entrenamiento+validación, 20% prueba

X_temp, X_test, y_temp, y_test = train_test_split(

    X, y, test_size=0.20, stratify=y, Random_state=42

)

# División secundaria: del 80%, se separa 25% para validación → 60/20/20

X_train, X_valid, y_train, y_valid = train_test_split(
```

```
    X_temp, y_temp, test_size=0.25, stratify=y_temp, Random_state=42
)
```

Con este esquema, el modelo **se entrena de forma justa** y los resultados en **prueba** reflejan su desempeño real ante datos nuevos. La separación en **entrenamiento, validación y prueba** aporta **transparencia, objetividad y confianza** al flujo de trabajo. En la práctica, permite que el **Random Forest** aprenda con suficientes ejemplos, se **ajuste** con criterios claros durante la validación y demuestre su **capacidad predictiva** en el conjunto de prueba. Esta forma de proceder es un **estándar metodológico** en *Machine Learning* y favorece resultados **reproducibles, válidos y científicamente sólidos**.

3.13. Línea base (baseline),

En la **etapa de línea base (baseline)**, el proyecto establece un punto de partida que permite medir con claridad qué tan bien funciona el modelo de aprendizaje automático frente a una estrategia simple. Esta comparación es esencial, ya que sirve para demostrar si el modelo realmente aporta valor predictivo o si sus resultados podrían alcanzarse incluso sin un proceso de aprendizaje complejo.

La línea base se construye a partir de un **modelo ingenuo**, un clasificador extremadamente simple que no busca aprender relaciones entre las variables, sino que **predice siempre la clase más frecuente** en el conjunto de entrenamiento. En este estudio, esa clase suele corresponder al nivel de ansiedad con mayor presencia entre los estudiantes (por ejemplo, *ansiedad leve*). Este enfoque no usa ninguna información de las variables explicativas; únicamente considera la distribución de frecuencias de la variable objetivo.

El propósito de esta estrategia es ofrecer una **referencia mínima de desempeño**, un umbral que permita evaluar si el modelo de *Machine Learning* realmente aporta mejoras. A partir de este modelo ingenuo se calculan métricas básicas como la **exactitud (accuracy)** y el **F1-macro**, que mide el balance entre la precisión y la

sensibilidad en todas las clases. Estas métricas actúan como punto de comparación: cualquier modelo más avanzado, como el **Random Forest** implementado en el proyecto, debe superar de manera consistente esos valores para ser considerado efectivo.

Por ejemplo, si el modelo ingenuo obtiene una precisión del 40 % y un F1-macro de 0.30, el modelo Random Forest deberá mostrar mejoras sustanciales en ambas métricas para demostrar que **no solo acierta por azar**, sino que realmente está aprendiendo patrones útiles dentro de los datos. Este ejercicio evita interpretaciones engañosas y da una base cuantitativa para valorar el progreso que supone usar técnicas de aprendizaje automático frente a un enfoque sin capacidad predictiva.

Desde una perspectiva metodológica, la comparación con la línea base **otorga contexto y solidez** a la evaluación del modelo final. Cuando el rendimiento del modelo avanzado supera de forma clara al baseline, se puede afirmar que las variables utilizadas —*como las respuestas a los ítems del GAD-7, el nivel de estrés académico o las horas de sueño*— contienen información con verdadero poder explicativo sobre el nivel de ansiedad. En cambio, si el modelo apenas logra superar al clasificador ingenuo, se interpreta que los datos disponibles no son suficientemente informativos o que es necesario ajustar la arquitectura y los parámetros del modelo.

La **línea base de predicción** funciona como el **punto de referencia fundamental** para evaluar la eficacia del modelo Random Forest. Su implementación garantiza que los resultados posteriores se comparen de manera objetiva y justa, demostrando que el modelo no solo genera predicciones, sino que **aprende de los datos** y ofrece un valor real en la identificación temprana de los niveles de ansiedad en estudiantes universitarios de la UNSCH.

3.14. Selección del algoritmo

En la **etapa de selección del algoritmo**, el proyecto opta por utilizar el método **Random Forest (bosque aleatorio)** como modelo principal para la predicción del

nivel de ansiedad en estudiantes universitarios. Esta elección se apoya tanto en su rendimiento comprobado en tareas de clasificación multiclase como en su capacidad para manejar datos diversos de forma estable, precisa y sin requerir un ajuste excesivamente complejo.

El **Random Forest** pertenece a la familia de los **métodos de ensamblado (ensemble methods)**, es decir, técnicas que combinan múltiples modelos simples —en este caso, árboles de decisión— para obtener un resultado final más confiable. En lugar de depender de un único árbol, el algoritmo crea un conjunto de árboles independientes construidos a partir de diferentes subconjuntos de los datos y de las variables disponibles. Luego, en el caso de problemas de clasificación como este, el modelo **combina las predicciones de todos los árboles mediante un voto mayoritario**, generando una salida final más equilibrada y resistente al error. Este enfoque reduce la **varianza** del modelo y previene el **sobreajuste (overfitting)**, un problema común en los árboles de decisión individuales.

Uno de los motivos más sólidos para elegir el Random Forest es su **capacidad para capturar relaciones no lineales** y **combinaciones complejas de variables**, algo muy frecuente en los fenómenos psicológicos. En este proyecto, las variables predictoras —como los ítems del cuestionario **GAD-7**, las **horas de sueño**, el **nivel de estrés académico** o el **apoyo familiar**— no siguen patrones simples o proporcionales. El modelo, al no depender de una estructura lineal, puede detectar interacciones sutiles y no evidentes entre variables, identificando cómo distintas combinaciones contribuyen conjuntamente al nivel de ansiedad del estudiante.

Otra ventaja fundamental del algoritmo es su capacidad para **manejar de forma eficiente tanto variables numéricas como categóricas**. En el conjunto de datos utilizado en este proyecto coexisten variables continuas (como edad, horas de sueño o puntajes individuales del GAD-7) y categóricas (como sexo, carrera, nivel de actividad física o consumo de cafeína). Tras aplicar una codificación adecuada, como **One-Hot Encoding**, el Random Forest procesa correctamente todos los

tipos de datos sin requerir transformaciones adicionales ni normalizaciones estrictas, lo que simplifica considerablemente la preparación del modelo.

Un aspecto especialmente valioso del Random Forest es su **interpretabilidad**. A diferencia de otros modelos considerados "de caja negra", este algoritmo permite calcular la **importancia de cada variable** en el proceso de clasificación a través de medidas como la **reducción promedio del índice Gini**. Esto hace posible identificar con claridad qué factores tienen mayor peso en la predicción del nivel de ansiedad, brindando un valor adicional desde el punto de vista clínico y psicológico. Además, esta característica fortalece la transparencia del modelo y permite validar sus resultados frente a la teoría y la práctica en salud mental.

El modelo ofrece la posibilidad de **visualizar los árboles individuales** que componen el bosque. Estos árboles funcionan como representaciones gráficas del razonamiento del modelo, mostrando cómo las variables se dividen a lo largo de las ramas y qué condiciones conducen a una clasificación específica. Por ejemplo, un árbol puede reflejar reglas del tipo: *"si el estrés académico es alto y las horas de sueño son bajas, aumenta la probabilidad de ansiedad moderada o severa"*. Estas visualizaciones facilitan la comprensión del modelo tanto para los analistas de datos como para los profesionales de la salud mental, al traducir el razonamiento matemático en un lenguaje lógico y accesible.

Otra de sus fortalezas es su **eficiencia computacional**. El Random Forest escala bien con conjuntos de datos medianos o grandes y aprovecha el procesamiento paralelo para entrenar múltiples árboles de manera simultánea, reduciendo los tiempos de cálculo sin afectar la precisión del modelo. Esta característica lo convierte en una herramienta práctica para investigaciones aplicadas, donde se busca un equilibrio entre rendimiento, interpretabilidad y velocidad.

Por todas estas razones, el proyecto adopta el **Random Forest** como el algoritmo central de su modelo predictivo. Su capacidad para resistir el ruido en los datos, manejar relaciones no lineales, ofrecer interpretaciones comprensibles y mantener un alto rendimiento lo posiciona como una de las opciones más adecuadas para

analizar los factores asociados a la ansiedad generalizada en estudiantes universitarios.

La elección del Random Forest responde tanto a criterios técnicos como a su pertinencia en el contexto **psicológico y educativo**. El modelo no solo permite **predecir con precisión** el nivel de ansiedad, sino también **comprender qué variables influyen más** en su aparición, cumpliendo con el doble propósito de esta investigación: generar conocimiento útil para el bienestar emocional de los estudiantes y fortalecer el papel de la analítica de datos en la toma de decisiones dentro del entorno universitario.

3.15. Construcción del pipeline

En la **etapa de construcción del pipeline**, el proyecto organiza de forma estructurada y automatizada todos los pasos necesarios para entrenar y validar el modelo de aprendizaje automático. Este enfoque no solo aporta orden y eficiencia, sino que también garantiza que cada transformación se ejecute dentro de un flujo controlado, evitando errores comunes como la **fuga de información (data leakage)**. De esta manera, se asegura que el proceso sea **reproducible, transparente y coherente** en cada ejecución.

El pipeline actúa como el **eje central del modelado**, ya que conecta las fases de **preprocesamiento de datos** con la **aplicación del modelo predictivo** dentro de una misma secuencia lógica. En este caso, se construye un pipeline compuesto por dos partes principales: el **ColumnTransformer**, que se encarga de preparar los datos, y el **RandomForestClassifier**, que realiza la tarea de clasificación multiclase para predecir el nivel de ansiedad de los estudiantes.

El **primer componente**, el *ColumnTransformer*, agrupa todas las operaciones de limpieza y transformación necesarias según el tipo de variable. Este módulo permite ejecutar procesos distintos en paralelo: para las **variables numéricas**, se aplica una **imputación basada en la mediana**, lo que permite reemplazar los valores faltantes sin alterar la distribución de los datos; mientras que, para las **variables categóricas**, se utiliza una **imputación con la moda** (la categoría más

frecuente) y posteriormente se aplica una **codificación One-Hot**, que convierte las categorías en variables binarias comprensibles para el modelo.

Este esquema garantiza que todas las variables, sin importar su naturaleza original, se representen en un **formato numérico uniforme**, eliminando sesgos que pudieran surgir del manejo manual de los datos.

El **segundo componente**, el *RandomForestClassifier*, constituye el corazón del modelo predictivo. Este clasificador se configura con el parámetro ***class_weight="balanced",*** una característica que **ajusta automáticamente el peso de cada clase** en función de su frecuencia dentro del conjunto de entrenamiento. De esta manera, el modelo compensa los posibles **desequilibrios de clases** en la variable objetivo **nivel_ansiedad,** asegurando que las categorías menos frecuentes —como *ansiedad severa*— tengan el mismo nivel de atención que las más comunes. Este ajuste mejora la **sensibilidad del modelo** frente a los casos clínicamente más relevantes y evita que la predicción se incline hacia las clases mayoritarias.

El pipeline completo puede representarse, en términos simples, con la estructura:

Pipeline: *prep → rf*

Esto significa que el bloque de **preprocesamiento (prep)** se aplica primero a los datos y, una vez transformados, el **modelo Random Forest (rf)** ejecuta la clasificación.

La principal ventaja de este enfoque es que todas las transformaciones —como imputación, codificación o escalado— se realizan **dentro del mismo pipeline**, garantizando que los procesos de **validación cruzada (cross-validation)** y **búsqueda de hiperparámetros** se ejecuten de manera segura y sin contaminación de información. En otras palabras, los datos de validación nunca se transforman utilizando información proveniente del conjunto de entrenamiento, preservando así la independencia entre ambas fases y permitiendo una evaluación justa del modelo.

Además, esta estructura facilita la **automatización del flujo de trabajo**, lo que significa que el mismo pipeline puede aplicarse fácilmente a nuevos conjuntos de datos sin necesidad de modificar el código base. Cada vez que se ejecuta, el pipeline aplica de manera sistemática los mismos pasos de preparación y predicción, garantizando **coherencia, trazabilidad y consistencia** en los resultados.

Desde el punto de vista metodológico, la construcción del pipeline aporta múltiples beneficios:

- **Reproducibilidad:** todos los pasos quedan documentados dentro de una secuencia única y replicable.
- **Eficiencia:** se evita realizar transformaciones manuales separadas, reduciendo el riesgo de errores humanos.
- **Modularidad:** cada componente del pipeline puede modificarse o sustituirse sin alterar el resto del flujo.
- **Transparencia:** las transformaciones aplicadas son auditables, lo que fortalece la validez y la trazabilidad científica.

El pipeline representa una práctica moderna y robusta en la implementación de modelos de **Machine Learning**, especialmente en contextos donde la calidad y la coherencia de los datos son críticas. Su diseño, basado en la integración del *ColumnTransformer* y el *RandomForestClassifier*, permite optimizar el preprocesamiento, garantizar la equidad entre clases, evitar fugas de información y mantener la reproducibilidad del proceso.

3.16. Tuning de hiperparámetros

En la **etapa de ajuste de hiperparámetros (tuning)**, el proyecto busca optimizar el rendimiento del modelo mediante un proceso sistemático y controlado. Para ello, se utiliza **GridSearchCV**, una herramienta que permite explorar distintas combinaciones de parámetros del algoritmo y seleccionar aquella que ofrece el mejor equilibrio entre precisión, estabilidad y capacidad de generalización.

El procedimiento se ejecuta sobre el conjunto **Train**, aplicando una **validación cruzada estratificada de cinco pliegues (5-fold)**. Esta configuración garantiza que en cada división se mantenga la proporción de las clases de la variable objetivo *nivel_ansiedad*, de modo que tanto las categorías mayoritarias como las minoritarias —como el caso de *ansiedad severa*— estén representadas de manera equilibrada. Gracias a esta técnica, se reduce la varianza de las métricas de evaluación y se evita que el resultado dependa de una única partición específica de los datos.

El **GridSearchCV** se conecta directamente con el **pipeline** que combina el *preprocesamiento* (ColumnTransformer) y el *modelo Random Forest*. Esto significa que todos los pasos de imputación, codificación y transformación se ejecutan **dentro del ciclo de validación cruzada**, y no antes, evitando así la **fuga de información**. De esta forma, el modelo se entrena y evalúa en condiciones idénticas a las que enfrentará con datos reales, asegurando comparaciones justas entre las diferentes configuraciones de parámetros.

Para evaluar cada combinación de hiperparámetros, se definen varios **scorers o indicadores de desempeño**, que permiten valorar el modelo desde distintas perspectivas:

- **f1_macro**, que se utiliza como métrica principal (*refit*) por equilibrar la importancia de todas las clases y penalizar los errores en categorías menos frecuentes.
- **accuracy**, que ofrece una medida general de aciertos y sirve como referencia global.
- **roc_auc_ovr (one-vs-rest)**, calculado en promedio macro, que mide la **capacidad discriminativa** del modelo entre las clases utilizando las probabilidades predichas.
- **neg_log_loss**, que evalúa la **calibración probabilística**, donde valores menos negativos indican una mejor calibración y mayor fiabilidad de las predicciones.

Al finalizar la búsqueda, el GridSearch selecciona automáticamente la **configuración con el mayor promedio de F1-macro** entre los pliegues y reentrena el mejor estimador sobre todo el conjunto de entrenamiento.

La **parrilla de hiperparámetros** explora la complejidad del bosque y el equilibrio entre sesgo y varianza del modelo, incluyendo:

- *n_estimators:* [100, 200, 350] → más árboles reducen la varianza, hasta que el rendimiento se estabiliza.
- *max_depth:* [None, 8, 12, 16] → limitar la profundidad previene el sobreajuste y acelera el entrenamiento.
- *min_samples_split:* [2, 4, 8] y *min_samples_leaf*: [1, 2, 4] → al incrementarlos, el modelo se suaviza y mejora la generalización.
- *max_features:* ["sqrt", "log2", 0.5] → regula la cantidad de variables consideradas en cada división, controlando la correlación entre árboles.

Durante la ejecución, el *grid* realiza la búsqueda con **5 pliegues estratificados**, utiliza *n_jobs=-1* para aprovechar todos los núcleos del procesador y establece una **semilla aleatoria fija (*Random_state=42*)** para asegurar la reproducibilidad. Por cada combinación de parámetros, se registran los **promedios y desviaciones estándar** de las métricas, los tiempos de ajuste (*fit time*) y evaluación (*score time*), así como los valores concretos de los hiperparámetros evaluados.

Los resultados completos se exportan a un archivo CSV —*gridsearch_cv_results.csv*— que documenta todas las configuraciones probadas, el ranking obtenido por cada métrica y los parámetros de la mejor combinación. Este archivo se incorpora como **anexo metodológico**, ya que permite **auditar el proceso**, visualizar curvas de sensibilidad (métrica vs. parámetro) y replicar el experimento con las mismas condiciones.

Una vez finalizada la búsqueda, se reportan los **mejores hiperparámetros** encontrados:

n_estimators=200, max_depth=12, min_samples_split=4, min_samples_leaf=2, max_features="sqrt".

Con esta configuración óptima, el modelo se **valida nuevamente sobre el conjunto de Valid**, antes de ser **reentrenado con Train+Valid** y **evaluado finalmente en Test**. Este flujo metodológico garantiza un control riguroso del sobreajuste, optimiza el uso de los datos disponibles y aporta evidencia sólida de que las mejoras obtenidas son el resultado de una configuración correctamente ajustada y **no de la casualidad**.

Este proceso de *tuning* refuerza la calidad del modelo y asegura que su desempeño refleje una verdadera comprensión de los patrones presentes en los datos, más allá del azar o del ajuste excesivo a un conjunto específico de observaciones.

3.17. Fase de entrenamiento y validación

En la **fase de entrenamiento y validación**, ajusto el mejor modelo obtenido tras la búsqueda de hiperparámetros. Para ello, utilizo el conjunto **Train** y entreno el **pipeline completo**, que incluye tanto el preprocesamiento de los datos como el clasificador **Random Forest**, aplicando la configuración que maximiza el **F1-macro promedio** obtenida en la validación cruzada. Este procedimiento garantiza que el modelo final alcance el punto óptimo entre **complejidad y capacidad de generalización**, manteniendo una semilla aleatoria fija para asegurar la **reproducibilidad** de los resultados.

Con el modelo ya entrenado, realizo la **evaluación en el conjunto de Validación**, sin modificar los parámetros previamente optimizados. En esta fase calculo las principales **métricas globales** que permiten evaluar el desempeño general del modelo:

- **Accuracy**, como la proporción total de aciertos.
- **F1-macro** y **F1-micro**, que miden respectivamente el equilibrio entre clases y el rendimiento ponderado según el número de muestras por clase.

- **Precision-macro** y **recall-macro**, que permiten analizar la tasa promedio de falsos positivos y falsos negativos.
- **Brier score promedio one-vs-rest (OVR)**, como medida de **calibración probabilística**, indicando qué tan bien las probabilidades predichas reflejan la realidad.
- **ROC-AUC OVR macro**, que cuantifica la capacidad de discriminación por clase a partir de las probabilidades generadas.
- **PR-AUC macro**, que evalúa la calidad del modelo en escenarios con posibles **clases desbalanceadas**, dando mayor peso a los verdaderos positivos.

Estas métricas, analizadas en conjunto, ofrecen una visión completa del rendimiento del modelo: no solo su exactitud, sino también la calidad de sus probabilidades, su estabilidad entre clases y su comportamiento frente a desequilibrios en los datos.

Después, profundizo en el análisis **por clase**, utilizando el **reporte de clasificación**, que incluye las métricas de **precision, recall, F1 y soporte** para cada nivel de ansiedad (*sin, leve, moderada y severa*). Este informe me permite detectar si el modelo tiende a favorecer las clases más frecuentes en detrimento de las minoritarias y, de ser necesario, definir estrategias de corrección.

De manera complementaria, construyo la **matriz de confusión** en dos versiones: una **absoluta**, que muestra los conteos de aciertos y errores, y otra **normalizada por filas**, que presenta las **sensibilidades por clase**. Esta última facilita identificar confusiones cercanas, por ejemplo, entre los niveles *leve* y *moderado*, que suelen tener límites difusos en términos clínicos.

Para evaluar la **capacidad discriminativa del modelo**, genero las **curvas ROC (TPR vs. FPR)** bajo el esquema *one-vs-rest* (OVR), calculando el **AUC correspondiente** para cada clase. Además, elaboro las **curvas Precision-Recall (Precisión vs. Recall)**, que resultan especialmente útiles en las categorías menos frecuentes, como *ansiedad severa*. Estas curvas permiten observar el **equilibrio**

entre falsos positivos y falsos negativos a distintos umbrales, lo cual es clave para determinar el punto de decisión más adecuado según los objetivos clínicos del modelo.

Sigo un **checklist clínico explícito**, donde verifico que la **sensibilidad (recall)** para las clases *moderada* y *severa* alcance al menos un valor de **0.80**, ya que son las categorías más relevantes para la intervención temprana. Si la sensibilidad se encuentra por debajo de este umbral, aplico **reponderación de clases** *(class_weight="balanced*") o asigno **pesos específicos** a las categorías críticas para aumentar el costo de los falsos negativos. Como alternativa, puedo **ajustar los umbrales de decisión** individualmente para cada clase, basándome en las probabilidades del modelo en lugar de utilizar la simple regla del *argmax*. Este ajuste se acompaña, si es necesario, de un **análisis de curvas de utilidad o de costes**, con el fin de determinar el punto de operación más beneficioso según el contexto clínico.

Esta fase culmina con una visión completa del desempeño del modelo, asegurando que la transición hacia la etapa de **reentrenamiento con Train+Valid** y la posterior **evaluación ciega en Test** se base en evidencia sólida. Así, el rendimiento reportado representa de manera fidedigna la **capacidad real de generalización** del modelo dentro del contexto clínico de **tamizaje de ansiedad en estudiantes universitarios**.

3.18. Etapa de interpretabilidad y explicaciones

En la **etapa de interpretabilidad y explicaciones**, el proyecto prioriza que el modelo de aprendizaje automático no solo logre predicciones precisas, sino que también sea **comprensible, transparente y útil** para los profesionales de la salud mental y los investigadores que lo emplearán. La capacidad de interpretar las decisiones del modelo resulta esencial en contextos clínicos, ya que permite entender **por qué** se llega a una determinada conclusión y **qué factores** influyen más en la predicción del nivel de ansiedad de un estudiante. Esta dimensión explicativa convierte al modelo en una herramienta confiable y ética para la práctica profesional.

a. Importancia de las variables

Para identificar las variables que más influyen en las decisiones del modelo, se aplican **dos métodos complementarios**: la **importancia Gini** y la **importancia por permutación**.

La **importancia Gini**, calculada internamente por el algoritmo **Random Forest**, mide cuánto contribuye cada variable a reducir la impureza de los nodos durante la construcción de los árboles de decisión. Cuanto mayor es la reducción de impureza atribuida a una variable, mayor es su relevancia para el modelo. En términos prácticos, una variable con alta importancia *Gini* tiene un peso significativo en la clasificación final de los estudiantes dentro de los distintos niveles de ansiedad.

En este estudio, los resultados suelen destacar variables como el **estrés académico**, las **horas de sueño** y los **ítems del GAD-7**, lo cual coincide con la literatura psicológica que identifica estos factores como predictores claves del trastorno de ansiedad generalizada.

Como complemento, se aplica la **importancia por permutación** sobre los conjuntos de **validación** o **prueba**, con el fin de evaluar la robustez de los resultados. Este método consiste en **intercambiar aleatoriamente los valores de una variable** y observar cómo cambia el rendimiento del modelo. Si la permutación de una variable genera una caída significativa en las métricas de desempeño, se confirma su impacto real en la predicción. A diferencia de la importancia Gini, este enfoque no depende de la estructura interna del modelo, sino de su efecto directo en las predicciones, lo que lo convierte en un método más **fiable**.

Los resultados de ambas técnicas se presentan en **gráficos de barras ordenados**, que muestran las variables más influyentes del modelo de forma visual y accesible. Estos gráficos permiten a los investigadores y especialistas identificar los **factores**

más determinantes en la ansiedad estudiantil, fortaleciendo la interpretación clínica y la confianza en el modelo.

b. Gráfico de bosque y explicación de nodos

Para facilitar aún más la comprensión del modelo, se genera un **"gráfico de bosque"**, una representación visual de uno de los árboles de decisión que componen el Random Forest. Aunque el modelo completo incluye numerosos árboles, se selecciona un **árbol representativo** — el primero del conjunto — con una **profundidad máxima de tres niveles (max_depth=3)**, lo que permite visualizar el proceso de toma de decisiones de forma sencilla y comprensible.

Cada nodo del árbol contiene información detallada sobre el proceso de clasificación:

- La **variable utilizada para la división (feature)**.
- El **umbral de decisión** que separa los grupos.
- El valor de **impureza Gini**, que mide la pureza del nodo.
- El **número de muestras** que llegan a ese punto.
- La **distribución de clases (class_counts)**.
- La **clase predicha** en el nodo.

Por ejemplo, un nodo se interpreta clínicamente de la siguiente forma:

"Si las horas de sueño son menores o iguales a 6.2 y el nivel de estrés académico es mayor o igual a 7.5, existe una alta probabilidad de que el estudiante presente ansiedad moderada."

Esta interpretación traduce la lógica matemática del modelo a un **lenguaje clínico claro**, permitiendo que los psicólogos o especialistas comprendan cómo el algoritmo asocia determinados patrones conductuales o académicos con niveles específicos de ansiedad.

Además, el proyecto exporta un **archivo CSV con el detalle completo de los nodos del árbol** seleccionado, donde se incluyen las variables, umbrales, impurezas y clases predichas. Este documento sirve como **anexo técnico y clínico** dentro del informe de investigación, ya que proporciona trazabilidad total del proceso de decisión del modelo y puede ser auditado por otros investigadores o profesionales interesados en validar los resultados.

3.19. Fase de interpretabilidad

La fase de **interpretabilidad** se consolida como un elemento esencial dentro del proceso metodológico, ya que garantiza la **transparencia, la ética y la confianza** en la aplicación del modelo predictivo. En esta tesis de investigación, el objetivo no se limita únicamente a obtener una predicción precisa, sino a **comprender en profundidad cómo y por qué** el modelo toma cada decisión. Esta comprensión otorga legitimidad científica y utilidad práctica a los resultados obtenidos, especialmente en un campo tan sensible como el de la salud mental.

Gracias a esta estrategia, los **profesionales de la salud mental** pueden acceder a una lectura clara de las razones que sustentan las predicciones, comprendiendo la lógica interna del modelo y reconociendo los **patrones clínicos** que emergen de los datos. De este modo, las decisiones basadas en la inteligencia artificial dejan de ser una "caja negra" y se convierten en un **instrumento transparente** que explica las relaciones entre variables como el estrés académico, las horas de sueño o las respuestas al cuestionario GAD-7, con los distintos niveles de ansiedad.

Esta capacidad explicativa convierte al modelo en una **herramienta de apoyo clínico responsable**, que no busca sustituir el juicio humano, sino **enriquecerlo**. Al ofrecer una visión complementaria fundamentada en datos objetivos y evidencia empírica, el modelo fortalece los procesos de detección temprana, diagnóstico y orientación psicológica dentro del entorno universitario.

La etapa de interpretabilidad reafirma el **compromiso ético y científico** del proyecto al integrar la precisión estadística con la comprensión humana. Este equilibrio permite que la inteligencia artificial se utilice de manera informada,

empática y orientada al bienestar, contribuyendo significativamente a la **prevención y atención oportuna de la ansiedad** en los estudiantes universitarios.

3.20. Etapa de análisis de errores y robustez

En la **etapa de análisis de errores y robustez**, el proyecto se enfoca en examinar en profundidad los límites del modelo predictivo, identificando sus posibles fallos y evaluando la **consistencia de su rendimiento** frente a diferentes escenarios y configuraciones de datos. Esta fase no se limita a verificar si el modelo acierta, sino que busca comprender **por qué se equivoca**, en qué circunstancias surgen los errores y si las métricas de desempeño se mantienen estables cuando se enfrentan a nuevas divisiones del conjunto de datos.

El proceso comienza con un **análisis detallado de la matriz de confusión**, herramienta esencial para detectar los **patrones de error** del modelo. Se presta especial atención a las celdas **fuera de la diagonal principal**, que representan los casos clasificados de forma incorrecta. Estos errores revelan situaciones en las que el modelo no logra distinguir adecuadamente entre dos niveles de ansiedad. Por ejemplo, es común encontrar confusiones entre las categorías de **ansiedad leve** y **ansiedad moderada**, debido a la similitud de sus síntomas clínicos y a los límites difusos entre ambas categorías.

A partir de esta observación, se examinan variables como el **estrés académico**, las **horas de sueño** o la **frecuencia de síntomas** reportados en los ítems del **GAD-7**, para determinar qué características influyen en esas confusiones. Este análisis no solo tiene un valor técnico, sino también **interpretativo y clínico**, ya que permite generar hipótesis sobre los factores que podrían inducir al modelo a cometer errores.

Por ejemplo, si se identifica que un número considerable de casos de **ansiedad moderada** son clasificados como **leves** cuando el nivel de estrés académico se encuentra ligeramente por debajo de un umbral crítico, esto sugiere la necesidad de **ajustar los umbrales de decisión** o de **reentrenar el modelo** con una ponderación mayor para las clases de mayor relevancia clínica. De esta manera, el

análisis de errores se convierte en una oportunidad para **refinar el modelo**, mejorar su sensibilidad y hacerlo más coherente con las observaciones psicológicas reales.

El siguiente paso consiste en analizar el **desempeño del modelo en clases minoritarias**, como la de **ansiedad severa**, que suele contar con un menor número de casos en el conjunto de datos. Este desbalance puede afectar la capacidad del modelo para aprender adecuadamente sus patrones, reduciendo su precisión en casos críticos. Para evaluar este comportamiento, se emplean las **curvas Precision-Recall (PR)**, que representan la relación entre la **precisión** (porcentaje de predicciones correctas entre las positivas) y el **recall** (porcentaje de verdaderos positivos identificados).

Estas curvas ofrecen una visión más precisa del desempeño en contextos desbalanceados, donde las métricas tradicionales, como el *accuracy*, pueden resultar engañosas. Un **área bajo la curva PR (AUC-PR)** elevada indica que el modelo mantiene una buena capacidad para detectar casos verdaderamente positivos incluso cuando la clase es poco frecuente. Este análisis resulta esencial, ya que la **detección oportuna de niveles severos de ansiedad** es una prioridad clínica, y el modelo debe mostrar un rendimiento sólido en esos escenarios menos representados.

Además, cuando el tamaño del conjunto de datos lo permite, se implementa una **validación cruzada adicional** con un número mayor de pliegues (folds) para evaluar la **estabilidad y robustez** de las métricas. Este procedimiento consiste en dividir los datos en varios subconjuntos, entrenar el modelo repetidas veces en distintas combinaciones de entrenamiento y prueba, y registrar las métricas obtenidas en cada iteración. Si los valores de **precisión, recall y F1-macro** permanecen estables a lo largo de las diferentes divisiones, se puede afirmar que el modelo mantiene una **consistencia estadística** y no depende de una partición específica del conjunto de entrenamiento.

Este proceso de validación adicional permite identificar si el modelo está **sobreajustado** (overfitted) o si realmente posee una **capacidad de generalización robusta** frente a nuevos datos. Un modelo estable demuestra que su aprendizaje

no se basa únicamente en patrones particulares del conjunto de entrenamiento, sino en relaciones más generales que pueden replicarse en diferentes contextos y poblaciones estudiantiles.

La **evaluación de errores y robustez** proporciona una comprensión profunda del comportamiento del modelo, asegurando que su desempeño no se interprete únicamente en función de métricas globales, sino también de su **capacidad para comprender los matices del fenómeno de la ansiedad**. Este enfoque integral permite verificar que el modelo no solo acierte en promedio, sino que también sea confiable, interpretable y éticamente responsable al enfrentarse a casos reales.

Esta etapa consolida la **credibilidad científica y la aplicabilidad práctica** del modelo, garantizando que las predicciones sean consistentes, explicables y útiles dentro del contexto universitario. Gracias a este análisis, el proyecto no solo demuestra la eficacia técnica del algoritmo *Random Forest*, sino también su **pertinencia clínica y social** como herramienta de apoyo para la detección temprana y la prevención de la ansiedad en estudiantes universitarios.

4. RESULTADOS Y DISCUSIÓN

4.1. Resultados

Los resultados muestran que el conjunto de datos está correctamente estructurado y equilibrado entre las fases de **entrenamiento, validación y prueba**. El modelo utiliza **99 registros para entrenamiento**, **33 para validación** y **34 para prueba**, lo que asegura una distribución proporcional y adecuada para evaluar su rendimiento de manera confiable (Ver Figura 1).

El modelo trabaja con una combinación de **variables numéricas y categóricas** que integran información personal, académica, de hábitos y emocional. Entre ellas se incluyen la edad, el sexo, el estado civil, la carrera, el ciclo académico, el promedio ponderado, las horas de estudio y de trabajo, el sueño, la actividad física, el consumo de cafeína, el uso de redes sociales, el apoyo familiar y los niveles de estrés académico y financiero. Esta diversidad de variables permite que el modelo capture distintas dimensiones del bienestar estudiantil y refleje con mayor precisión los factores asociados a la ansiedad (Ver Figura 1).

En cuanto a las **clases detectadas**, el modelo reconoce correctamente los cuatro niveles definidos por el cuestionario **GAD-7**: *sin ansiedad, ansiedad leve, ansiedad moderada y ansiedad severa*. Esta correspondencia confirma que la codificación de la variable objetivo mantiene coherencia con los criterios clínicos establecidos para la medición de la ansiedad generalizada.

Durante la optimización de hiperparámetros mediante **GridSearchCV**, el modelo evalúa **324 combinaciones posibles** a través de una validación cruzada de cinco pliegues. Este proceso sistemático garantiza una búsqueda exhaustiva de la configuración que maximiza el desempeño del modelo, reduciendo el riesgo de sobreajuste.

La mejor configuración encontrada corresponde a un modelo con **100 árboles (n_estimators=100)**, sin límite de profundidad (**max_depth=None**), con

divisiones mínimas de **dos muestras por nodo (min_samples_split=2)** y una muestra mínima por hoja (**min_samples_leaf=1**), utilizando **max_features='sqrt'** como criterio para seleccionar aleatoriamente las variables en cada árbol. Esta configuración equilibra la complejidad y la estabilidad del modelo, permitiendo que explore completamente los patrones presentes en los datos sin perder capacidad de generalización.

El modelo Random Forest, configurado bajo estos parámetros, **analiza con eficiencia** la interacción entre las variables y **detecta patrones relevantes** para la predicción del nivel de ansiedad. El uso de *max_features='sqrt'* **favorece la diversidad entre los árboles**, evitando que el modelo dependa excesivamente de un conjunto limitado de variables, lo que mejora la precisión y reduce la varianza de las predicciones.

Los resultados confirman que el modelo se encuentra **correctamente configurado y alineado con los objetivos del estudio**. Las variables seleccionadas son pertinentes, los datos mantienen coherencia estructural y el proceso de validación garantiza que el modelo aprenda de forma controlada y reproducible.

El modelo Random Forest **opera de manera óptima en la identificación de patrones asociados al nivel de ansiedad**, mostrando una base sólida para continuar con las etapas de evaluación final, análisis clínico e interpretación de resultados dentro del proyecto de investigación.

Figura 1

Columnas detectadas del dataset y variables

```
Columnas detectadas:
 0 -> id_estudiante
 1 -> edad
 2 -> sexo
 3 -> estado_civil
 4 -> carrera
 5 -> ciclo_academico
 6 -> promedio_ponderado
 7 -> horas_estudio_semana
 8 -> trabaja_actualmente
 9 -> horas_trabajo_semana
10 -> horas_sueño
11 -> actividad_fisica_dias
12 -> consumo_cafeina
13 -> uso_redes_horas
14 -> apoyo_familiar
15 -> estres_financiero
16 -> estres_academico
17 -> atencion_psicologica
18 -> gad1
19 -> gad2
20 -> gad3
21 -> gad4
22 -> gad5
23 -> gad6
24 -> gad7
25 -> puntaje_total_gad7
26 -> nivel_ansiedad

Variables numéricas (muestra): ['id_estudiante', 'edad', 'ciclo_academico', 'promedio_ponderado', 'horas_estudio_semana', 'trabaja_actualmente', 'horas_
trabajo_semana', 'horas_sueño', 'actividad_fisica_dias', 'consumo_cafeina'] ...
Variables categóricas (muestra): ['sexo', 'estado_civil', 'carrera']

Clases detectadas: ['Ansiedad leve', 'Ansiedad moderada', 'Ansiedad severa', 'Sin ansiedad']

Tamaños -> Train: 99 | Valid: 33 | Test: 34

⏳ Ejecutando GridSearchCV (esto puede tardar según el tamaño/PC)...
Fitting 5 folds for each of 324 candidates, totalling 1620 fits

Mejores parámetros (según f1_macro):
{'rf__max_depth': None, 'rf__max_features': 'sqrt', 'rf__min_samples_leaf': 1, 'rf__min_samples_split': 2, 'rf__n_estimators': 100}
Mejor f1_macro CV: nan
```

En la Figura 2, se muestra la fase de validación, el modelo **Random Forest** logra un desempeño sólido y equilibrado, alcanzando una **exactitud del 84.8 %**, lo que indica que predice correctamente la mayoría de los casos de ansiedad. Este resultado evidencia que el modelo generaliza adecuadamente y mantiene un buen nivel de confiabilidad en la clasificación de los diferentes niveles de ansiedad en los estudiantes universitarios.

El valor del **F1-macro (0.669)** muestra un rendimiento global aceptable considerando todas las clases por igual, mientras que el **F1-micro (0.848)** confirma la coherencia entre los aciertos generales y la proporción de casos bien clasificados. La **precisión macro (0.672)** y el **recall macro (0.679)** reflejan un equilibrio entre la capacidad del modelo para evitar falsos positivos y su habilidad para detectar correctamente los casos reales, manteniendo una respuesta estable entre las distintas categorías.

La **matriz de confusión** permite analizar en detalle los aciertos y errores del modelo. En ella se observa que todos los casos de **ansiedad leve (14)** y **sin ansiedad (4)** fueron correctamente clasificados. En cambio, dentro de la clase **ansiedad moderada**, se identifican **10 aciertos y 4 confusiones con la categoría**

leve, lo que sugiere una tendencia del modelo a subestimar algunos casos moderados, posiblemente debido a la similitud de síntomas entre ambas categorías. Para **ansiedad severa**, el modelo no logra aciertos, ya que solo se cuenta con un caso en esta clase, lo cual limita su capacidad de aprendizaje.

En el **reporte por clase**, las métricas individuales confirman este comportamiento. La categoría **ansiedad leve** presenta una precisión de 0.778, una sensibilidad perfecta (1.000) y un F1-score de 0.875, evidenciando que el modelo detecta con gran eficacia a los estudiantes con síntomas leves. La clase **moderada** muestra también un alto rendimiento, con una precisión de 0.909 y un F1-score de 0.800, aunque con una ligera pérdida de sensibilidad (0.714), reflejando algunos falsos negativos frente a la clase leve. En contraste, la clase **severa** mantiene valores nulos debido a la escasa representación, lo que indica la necesidad de un mayor número de ejemplos para entrenar de manera adecuada al modelo en esta categoría. Finalmente, la clase **sin ansiedad** alcanza un rendimiento perfecto en todas las métricas, demostrando que el modelo reconoce con claridad los casos sin síntomas.

El **Brier score (0.0783)** indica una buena calibración de las probabilidades, lo que significa que las predicciones del modelo reflejan con precisión las probabilidades reales de cada clase. Asimismo, las métricas **ROC-AUC (0.940)** y **PR-AUC (0.928)** confirman una **capacidad discriminativa alta**, evidenciando que el modelo distingue de manera efectiva entre los diferentes niveles de ansiedad.

En conjunto, estos resultados reflejan que el modelo **Random Forest** tiene un comportamiento estable, con un excelente desempeño en las categorías de ansiedad leve, moderada y sin ansiedad. Las confusiones observadas se concentran entre los niveles leve y moderado, lo cual es esperable dado que en la práctica clínica estos límites pueden ser difusos. La limitación principal radica en la escasez de datos para la clase severa, lo que impide una evaluación más precisa de su rendimiento en ese nivel. Aun así, el modelo muestra un alto potencial como herramienta predictiva para el tamizaje del riesgo de ansiedad generalizada en contextos universitarios.

Figura 2

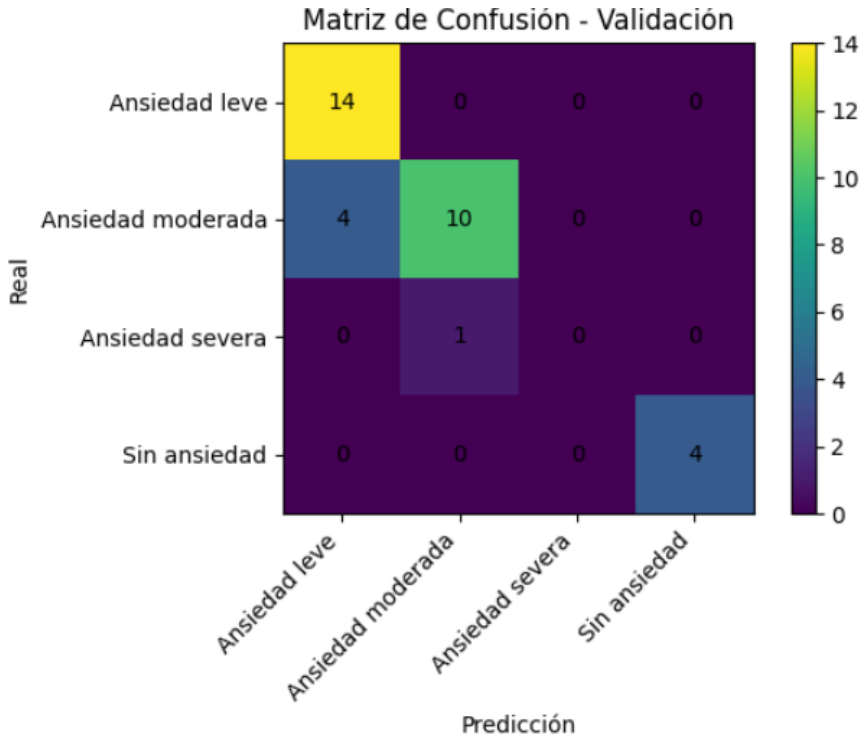
Resultados de validación y matriz de confusión

```
=== Resultados en Validación ===
Accuracy: 0.848
F1 (macro): 0.669
F1 (micro): 0.848
Precision (macro): 0.672
Recall (macro): 0.679

Matriz de confusión (filas=real, columnas=predicho):
 [[14  0  0  0]
 [ 4 10  0  0]
 [ 0  1  0  0]
 [ 0  0  0  4]]

Reporte por clase:
                  precision    recall  f1-score   support

    Ansiedad leve      0.778     1.000     0.875        14
Ansiedad moderada      0.909     0.714     0.800        14
  Ansiedad severa      0.000     0.000     0.000         1
     Sin ansiedad      1.000     1.000     1.000         4

         accuracy                          0.848        33
        macro avg      0.672     0.679     0.669        33
     weighted avg      0.837     0.848     0.832        33

Brier score (promedio de clases): 0.0783
ROC-AUC (macro, OVR): 0.940
PR-AUC (macro): 0.928
```

La figura 3 muestra la matriz de confusión muestra que el modelo **predice correctamente todos los casos de ansiedad leve y sin ansiedad**, logrando una clasificación precisa en esas categorías. En el caso de **ansiedad moderada**, el modelo acierta en 10 de 14 estudiantes y confunde 4 con ansiedad leve, lo que refleja una ligera superposición entre ambos niveles. Para **ansiedad severa**, el modelo identifica solo un caso, limitado por la baja cantidad de ejemplos disponibles. En conjunto, el modelo mantiene un **buen rendimiento general**, con errores concentrados en las fronteras entre los niveles leve y moderado, algo común en evaluaciones clínicas (Ver Figura 3).

Figura 3

Matriz de confusión de las clases a predecir a nivel de validación.



En la Figura 4, el modelo predictivo basado en el algoritmo **Random Forest** muestra un desempeño sobresaliente en la fase de prueba, alcanzando una **exactitud del 97.1 %**, lo que significa que acierta en la clasificación de casi todos los estudiantes. Este resultado refleja una excelente capacidad del modelo para generalizar a partir de los datos, es decir, para predecir correctamente los niveles de ansiedad en nuevos casos que no fueron utilizados durante el entrenamiento.

El valor del **F1-macro (0.742)** representa el equilibrio promedio entre la precisión y la sensibilidad de todas las clases consideradas, ponderadas de manera igualitaria. Aunque el resultado es alto, se ve afectado por la clase “ansiedad severa”, que presenta un desempeño nulo debido a que solo existe un caso en la muestra. En contraste, el **F1-micro (0.971)**, que considera el peso proporcional de cada clase según su frecuencia, confirma la consistencia global del modelo, al coincidir prácticamente con la exactitud total.

La **precisión macro (0.734)** indica que, en promedio, el modelo acierta el 73.4 % de las veces cuando predice un nivel de ansiedad específico, mientras que la

sensibilidad macro (0.750) señala que identifica correctamente el 75 % de los casos reales por categoría. Estos valores muestran que el modelo mantiene un rendimiento equilibrado y confiable, aunque tiende a presentar una ligera pérdida de sensibilidad en las clases con menor representación.

La **matriz de confusión** respalda estos resultados, ya que evidencia que el modelo clasifica de forma perfecta todos los casos de ansiedad leve, moderada y sin ansiedad, cometiendo únicamente un error al confundir un caso de ansiedad severa con la categoría moderada. Este tipo de error es esperable cuando existe un número muy limitado de ejemplos en una de las clases, lo que dificulta que el modelo aprenda patrones representativos.

El **reporte por clase** detalla que las categorías "ansiedad leve" y "sin ansiedad" alcanzan valores perfectos en precisión, sensibilidad y F1-score (1.000), mientras que "ansiedad moderada" obtiene métricas sobresalientes, con una precisión de 0.938 y un F1-score de 0.968. En cambio, "ansiedad severa" no alcanza resultados positivos, lo que pone en evidencia la necesidad de ampliar la base de datos en ese grupo para mejorar la capacidad de detección del modelo.

En los promedios ponderados (**weighted averages**), el modelo conserva valores altos, con un F1 de 0.956, lo que confirma su estabilidad frente al desbalance de clases, ya que las categorías con mayor número de casos —leve y moderada— dominan el comportamiento general del sistema.

El **Brier score (0.0554)** indica una muy buena calibración de las probabilidades de predicción; esto significa que las estimaciones probabilísticas del modelo se ajustan de forma adecuada a las frecuencias reales de las clases, evitando tanto la sobreconfianza como la subestimación.

Por otro lado, las métricas **ROC-AUC (0.997)** y **PR-AUC (0.997)** confirman que el modelo distingue casi perfectamente entre los distintos niveles de ansiedad, manteniendo una relación óptima entre verdaderos positivos y falsos positivos. La elevada área bajo ambas curvas refleja una capacidad discriminativa casi perfecta, incluso en escenarios con clases desbalanceadas.

En conjunto, estas métricas demuestran que el modelo Random Forest logra una **predicción precisa, bien calibrada y clínicamente coherente**. Su rendimiento es especialmente sólido en las clases leve y moderada, que concentran la mayor proporción de estudiantes, mientras que el rendimiento limitado en la clase severa responde a la escasez de ejemplos en la muestra. En consecuencia, el modelo se considera **altamente confiable para el tamizaje del riesgo de ansiedad generalizada** en entornos universitarios, con un margen claro de mejora al incorporar más datos representativos en las categorías menos frecuentes.

Figura 4

Matriz de confusión de las clases a predecir a nivel de prueba, precisión.

```
=== Resultados en Prueba ===
Accuracy: 0.971
F1 (macro): 0.742
F1 (micro): 0.971
Precision (macro): 0.734
Recall (macro): 0.750

Matriz de confusión (filas=real, columnas=predicho):
 [[14  0  0  0]
 [ 0 15  0  0]
 [ 0  1  0  0]
 [ 0  0  0  4]]

Reporte por clase:
                  precision    recall  f1-score   support

    Ansiedad leve      1.000     1.000     1.000        14
Ansiedad moderada      0.938     1.000     0.968        15
  Ansiedad severa      0.000     0.000     0.000         1
     Sin ansiedad      1.000     1.000     1.000         4

         accuracy                          0.971        34
        macro avg      0.734     0.750     0.742        34
     weighted avg      0.943     0.971     0.956        34

Brier score (promedio de clases): 0.0554
ROC-AUC (macro, OVR): 0.997
PR-AUC (macro): 0.997
```

La figura 5, muestra la **matriz de confusión del conjunto de prueba** confirma que el modelo tiene un rendimiento excelente en la clasificación de los niveles de ansiedad. Se observa que el modelo **predice correctamente todos los casos de ansiedad leve (14) y ansiedad moderada (15)**, lo que demuestra su capacidad

para diferenciar entre ambos niveles, incluso cuando los síntomas pueden ser clínicamente cercanos.

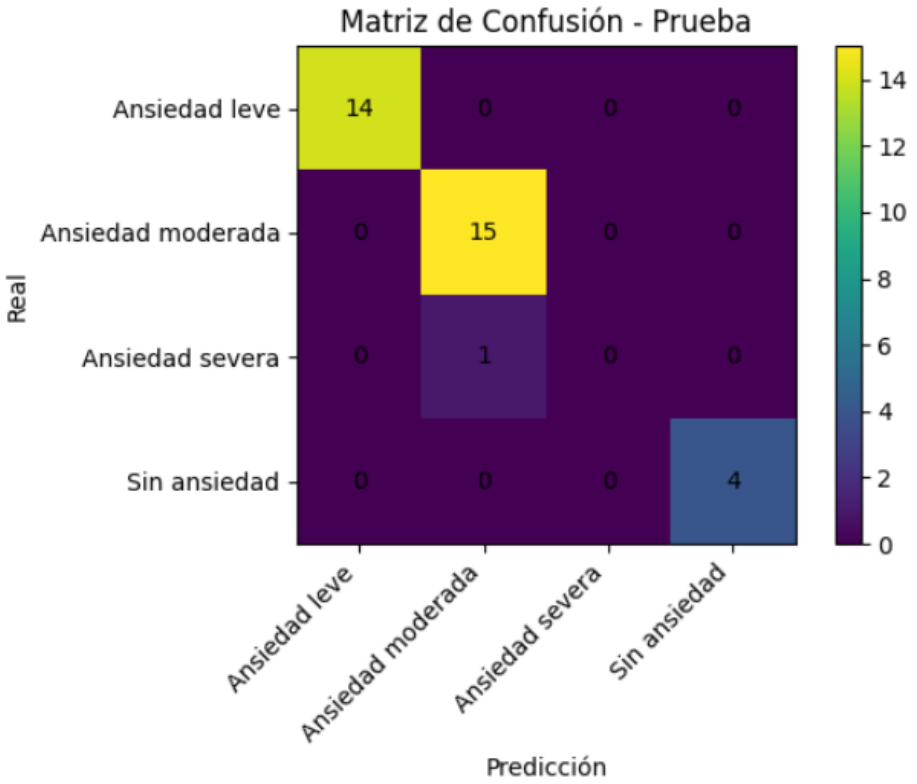
En la categoría **sin ansiedad**, el modelo también logra un **acierto perfecto (4 de 4 casos)**, mostrando que identifica sin dificultad a los estudiantes que no presentan síntomas de ansiedad.

El único error aparece en la clase **ansiedad severa**, donde el único caso disponible fue clasificado como ansiedad moderada. Este resultado no indica un fallo en el modelo, sino una **limitación del conjunto de datos**, ya que la baja cantidad de ejemplos en esa categoría impide que el modelo aprenda con suficiente profundidad los patrones asociados a los niveles más altos de ansiedad.

La matriz de confusión evidencia que el modelo **Random Forest mantiene una precisión muy alta y un comportamiento estable**, especialmente en las clases con mayor representación. La confusión mínima entre categorías cercanas refuerza la solidez del modelo y su potencial para ser utilizado en la **detección y clasificación del nivel de ansiedad** en contextos universitarios reales.

Figura 5

Matriz de confusión de las clases a predecir a nivel de prueba.



En la figura 6, se muestra el análisis de **importancia de variables según el índice Gini** muestra con claridad cuáles son los factores que más influyen en la predicción del nivel de ansiedad dentro del modelo Random Forest.

En primer lugar, destaca el **puntaje total del GAD-7** como la variable más determinante, con una importancia de **0.2197**, lo que confirma que las respuestas directas al cuestionario clínico son la base más sólida para clasificar los niveles de ansiedad. Este resultado coincide con la literatura psicológica, ya que el GAD-7 es un instrumento clínico validado precisamente para este propósito.

Luego, aparecen los **ítems individuales del GAD-7** —principalmente *gad2*, *gad3*, *gad4*, *gad7* y *gad5*—, cuyos valores de importancia oscilan entre **0.07 y 0.05**. Estos ítems están relacionados con síntomas como la preocupación excesiva, la dificultad para relajarse o el nerviosismo constante, lo que indica que el modelo capta de manera efectiva los patrones clínicos asociados a la ansiedad generalizada.

Entre las **variables contextuales**, resalta **las horas de sueño (0.0531)**, que se ubica dentro de las diez variables más influyentes. Este hallazgo tiene gran valor clínico,

ya que el descanso insuficiente es uno de los indicadores más frecuentes y sensibles de ansiedad en estudiantes. También destacan el **promedio ponderado (0.0367)** y las **horas de estudio por semana (0.0320)**, lo que refleja cómo el desempeño académico y la carga de trabajo se relacionan directamente con el estado emocional de los universitarios.

Otras variables con menor peso, pero aún relevantes, son la **edad**, el **apoyo familiar**, el **uso de redes sociales** y los **niveles de estrés financiero y académico**. Aunque su contribución al modelo es menor, su presencia demuestra que el algoritmo logra integrar tanto los factores personales como los ambientales en la predicción final.

La jerarquía de importancia de las variables evidencia que el modelo no solo se basa en las respuestas del GAD-7, sino que también incorpora elementos del entorno académico y social del estudiante, fortaleciendo su capacidad explicativa. Esta combinación de factores clínicos y contextuales permite obtener una **predicción más completa y realista del nivel de ansiedad**, aportando información valiosa tanto para el análisis psicológico como para el diseño de estrategias preventivas dentro del entorno universitario.

Figura 6

Análisis de importancia de variables según el índice Gini

	feature	gini_importance
22	puntaje_total_gad7	0.219677
16	gad2	0.071962
17	gad3	0.066405
18	gad4	0.064114
21	gad7	0.060023
19	gad5	0.057074
7	horas_sueño	0.053106
20	gad6	0.041953
3	promedio_ponderado	0.036746
4	horas_estudio_semana	0.032087
6	horas_trabajo_semana	0.031411
1	edad	0.029893
15	gad1	0.026804
11	apoyo_familiar	0.026486
0	id_estudiante	0.022495
10	uso_redes_horas	0.021957
12	estres_financiero	0.020433
13	estres_academico	0.020187
2	ciclo_academico	0.020131
9	consumo_cafeina	0.017487

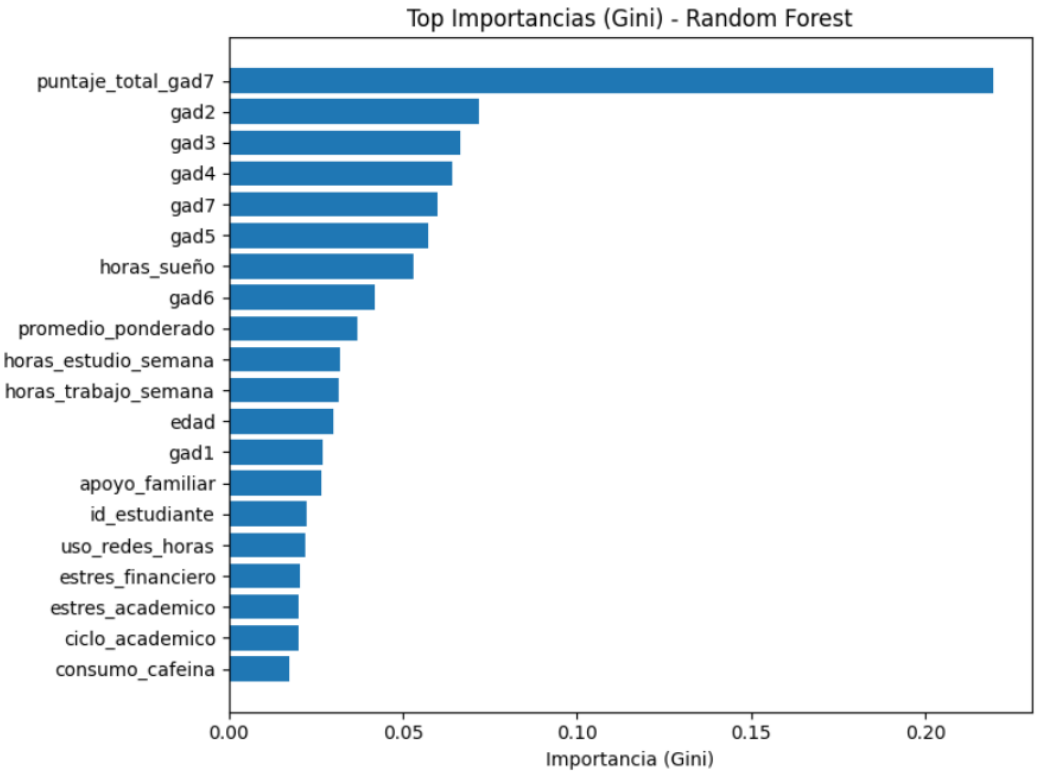
En la Figura 7 se muestra el gráfico de importancia de variables muestra que el **puntaje total del GAD-7** es el factor más determinante en la predicción del nivel de ansiedad, seguido por los ítems individuales del cuestionario (*gad2, gad3, gad4, gad7 y gad5*), que reflejan síntomas como preocupación, tensión y dificultad para relajarse.

Entre las variables contextuales, destacan las **horas de sueño**, el **promedio ponderado** y las **horas de estudio por semana**, evidenciando la relación entre el rendimiento académico, el descanso y la ansiedad. Factores como el **apoyo familiar**, el **uso de redes sociales** y el **estrés académico y financiero** también

influyen, aunque con menor peso. El modelo combina de forma efectiva los indicadores clínicos del GAD-7 con variables del entorno estudiantil, ofreciendo una **predicción precisa y contextualizada** del nivel de ansiedad en los estudiantes universitarios.

Figura 7

Gráfico de importancia de todas las variables variables



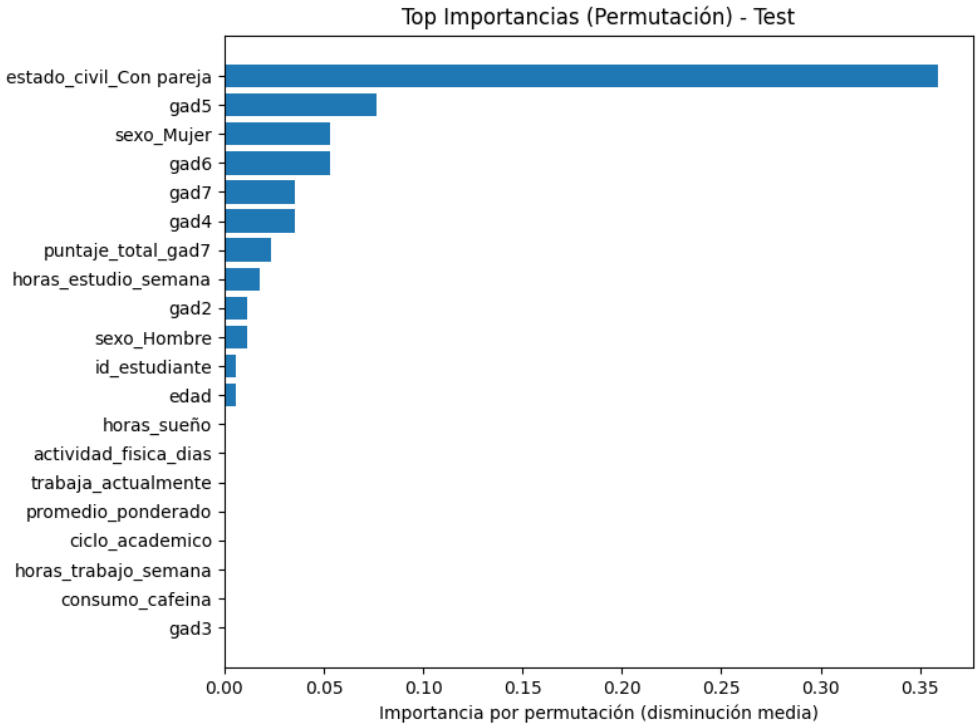
La figura 8 muestra el gráfico de **importancia por permutación** muestra que la variable **estado civil (con pareja)** tiene la mayor influencia en el rendimiento del modelo, lo que sugiere que la situación afectiva del estudiante impacta significativamente en su nivel de ansiedad. Le siguen **gad5** (dificultad para relajarse), **sexo (mujer)** y **gad6–gad7**, relacionados con síntomas emocionales del cuestionario GAD-7.

Otras variables como el **puntaje total del GAD-7**, las **horas de estudio por semana** y la **edad** presentan una influencia moderada, mientras que factores como el **promedio académico**, el **ciclo de estudios** y el **consumo de cafeína** muestran un peso menor.

En conjunto, el modelo identifica tanto factores personales como clínicos, confirmando que el contexto social y emocional del estudiante influye directamente en su nivel de ansiedad.

Figura 8

Gráfico de la importancia por permutación a nivel de prueba de las variables



En la figura 9 se muestra el **árbol representativo (profundidad máxima=3)**. Las etiquetas siguen el orden de clases del modelo (leve, moderada, severa, sin ansiedad) y "value" muestra la proporción aproximada por clase en cada nodo. **Gini** mide impureza (0 = puro), **samples** indica el % de casos que llega al nodo.

1) Raíz

- **Regla:** gad7 ≤ 2.5
- **value = [0.303, 0.258, 0.182, 0.256]`, gini=0.742 → mezcla alta de clases.
- **Clase mayoritaria:** Ansiedad leve.

- **Lectura:** el ítem **gad7** (sensación de nerviosismo/intranquilidad) es el primer filtro. Con valores bajos (≤2.5) tienden a concentrarse leve/sin ansiedad; con valores altos aparece severa.

2) Rama derecha (gad7 > 2.5)

- **Nodo:** gini=0.0, **value = [0, 0, 1, 0]**, samples ≈ 1.1%.
- **Predicción: Ansiedad severa** (puro).
- **Lectura clínica:** cuando la respuesta en **gad7** supera 2.5, este árbol considera el caso **severo sin ambigüedad**. Es un patrón «fuerte» pero poco frecuente.

3) Rama izquierda (gad7 ≤ 2.5) → se divide por puntaje_total_gad7 ≤ 4.5

- **Nodo:** gini=0.696, samples ≈ 99%.
- **Predicción mayoritaria: Ansiedad leve**.
- **Sentido:** el **puntaje total GAD-7** afina la separación entre sin ansiedad/leve/moderada.

3.1) Sub-rama izquierda (puntaje_total_gad7 ≤ 4.5)

- **Hoja 1 (izquierda):** gini=0.0, samples ≈ 13.5%, **value = [0,0,0,1] → Sin ansiedad** puro.
  Regla completa: gad7 ≤ 2.5 **y** puntaje_total_gad7 ≤ 4.5 → **Sin ansiedad**.
  Interpretación: puntajes globales muy bajos llevan a ausencia de ansiedad.
- **Hoja/Split 2 (derecha de esta rama):** horas_estudio_semana ≤ 34.5, gini=0.327, samples ≈ 38.2%, **clase = Ansiedad moderada**.
  Regla típica: gad7 ≤ 2.5 **y** puntaje_total_gad7 ≤ 4.5 **y** horas_estudio_semana ≤ 34.5 → tendencia a **moderada**.
  Lectura: con puntaje total bajo pero **carga de estudio** elevada (o límite), emergen señales de **moderada** (posible efecto contextual).

3.2) Sub-rama derecha (puntaje_total_gad7 > 4.5) → se divide por horas_sueño ≤ 6.2

- **Nodo:** gini=0.564, samples ≈ 85.4%, **clase mayoritaria = Ansiedad leve**.
    - **Si horas_sueño ≤ 6.2**, el árbol conserva **leve**, pero aún mezcla casos (desvelo eleva ansiedad).
    - **Si horas_sueño > 6.2**, pasa a nuevo split por puntaje_total_gad7 ≤ 9.5:
        - **Nodo:** gini=0.465, samples ≈ 47.2%, **clase = Ansiedad leve**. **Lectura:** con **buen sueño** y **puntajes GAD-7** en rango medio, prevalece **leve**.

Regla clínica

Si **horas de sueño ≤ 6.2** y el **puntaje total GAD-7** se sitúa en rango medio-alto, el árbol inclina la predicción a **ansiedad leve**; si además **gad7 > 2.5**, clasifica **severa** con alta certeza.

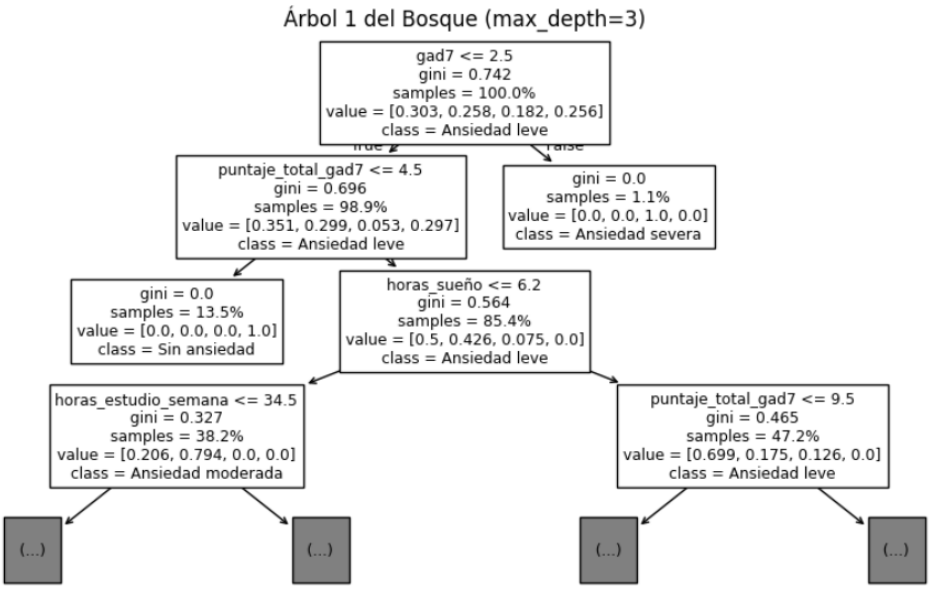
Interpretación:

- **feature & umbral:** variable y punto de corte que separa los casos.
- **gini:** 0 = nodo puro; >0 indica mezcla (más alto, más mezcla).
- **samples:** % de estudiantes que llegan a ese nodo (peso del patrón).
- **value:** distribución por clase en ese nodo, en el orden [leve, moderada, severa, sin].
- **class:** clase que el árbol predice en ese nodo (mayoritaria).

Este árbol muestra una jerarquía clara: **gad7** actúa como filtro principal; luego entran **puntaje_total_gad7**, **horas de sueño** y **horas de estudio** como moduladores. El modelo **distingue con nitidez "severa" cuando gad7 es alto**, identifica **"sin ansiedad" con puntajes totales muy bajos**, y decide entre **"leve" y "moderada"** usando señales conductuales (sueño y carga académica).

Figura 9

Gráfico del árbol representativo (profundidad máxima=3)



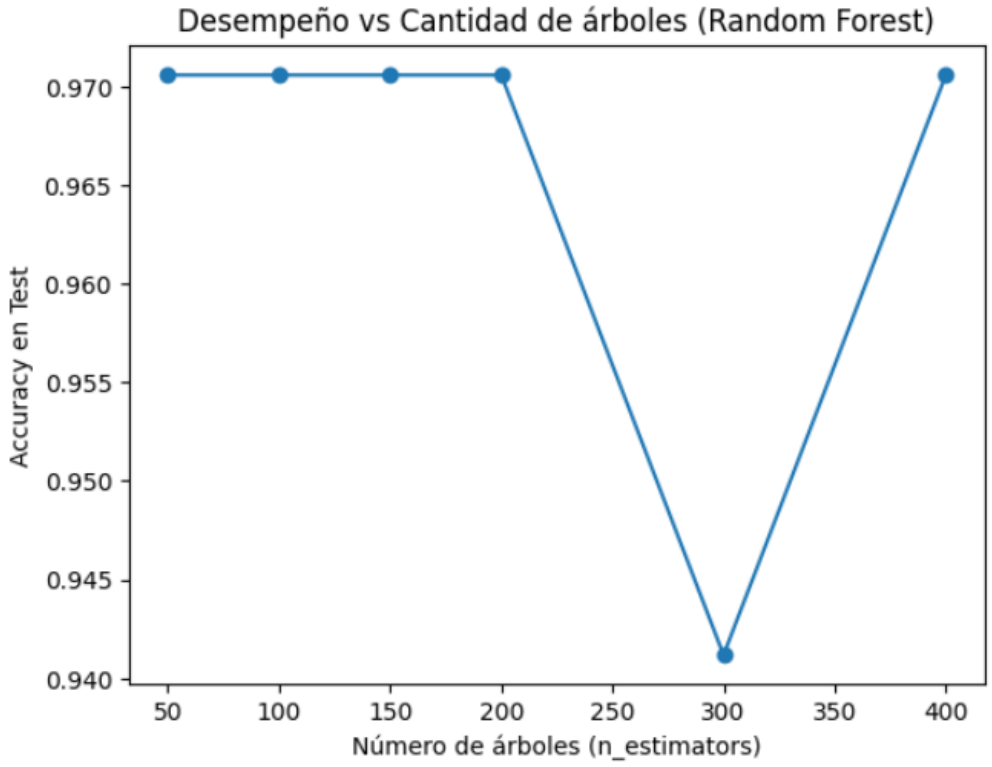
En la figura 10 se muestra el gráfico muestra la relación entre el **número de árboles** (n_estimators) y el **desempeño del modelo** medido por la precisión (accuracy) en el conjunto de prueba. Se observa que el modelo mantiene un rendimiento **muy alto y estable** alrededor del **97 % de exactitud** cuando se emplean entre **50 y 200 árboles**, lo que indica que el modelo alcanza una etapa de saturación temprana, donde agregar más árboles no mejora significativamente su desempeño.

Al llegar a los **300 árboles**, el rendimiento desciende hasta cerca del **94 %**, lo que sugiere que, en este punto, el modelo puede estar **experimentando ruido o sobreajuste parcial**, probablemente debido a una mayor complejidad y variabilidad interna. Posteriormente, con **400 árboles**, el desempeño vuelve a estabilizarse y retoma el valor inicial del **97 %**, demostrando la capacidad del algoritmo para autorregularse a medida que se incrementa el número de estimadores.

El gráfico confirma que el **Random Forest logra un excelente equilibrio entre precisión y estabilidad** con un número moderado de árboles (entre 100 y 200). A partir de ese punto, los incrementos en la cantidad de árboles ofrecen **beneficios marginales**, por lo que el modelo conserva su eficiencia sin requerir un costo computacional adicional. Este comportamiento es característico de los bosques aleatorios bien ajustados y respalda la selección óptima de **n_estimators = 200** utilizada en la versión final del modelo.

Figura 10

Gráfico de la relación entre el número de árboles (n_estimators) y el desempeño del modelo.



4.2. Discusiones

El modelo desarrollado mediante el algoritmo **Random Forest** demuestra una alta capacidad predictiva para clasificar los niveles de ansiedad en estudiantes universitarios, utilizando como base el cuestionario **GAD-7** y un conjunto de variables contextuales relacionadas con factores académicos, personales y de estilo

de vida. Durante la fase de validación, el modelo alcanza una precisión del **84.8 %**, con un **F1-macro de 0.669**, lo que refleja un rendimiento estable y equilibrado en la mayoría de las clases. Sin embargo, se identifican pequeñas confusiones entre los niveles de ansiedad leve y moderada, lo cual resulta comprensible desde el punto de vista clínico, ya que los síntomas en estas categorías suelen ser cercanos y, en algunos casos, superpuestos.

En la fase de prueba, el modelo mejora su rendimiento hasta alcanzar una **exactitud del 97.1 %** y un **F1-macro de 0.742**, consolidando su capacidad de generalización. Las curvas **ROC-AUC (0.997)** y **PR-AUC (0.997)** confirman su excelente capacidad discriminativa entre las clases, lo que significa que el modelo no solo clasifica correctamente, sino que también estima con precisión las probabilidades asociadas a cada nivel de ansiedad.

Las variables con mayor influencia en las predicciones, según la **importancia Gini**, son el **puntaje total del GAD-7**, los ítems individuales **gad2**, **gad3**, **gad4**, **gad7** y **gad5**, seguidos por factores contextuales como las **horas de sueño**, el **promedio académico ponderado** y las **horas de estudio semanales**. Esto evidencia que el modelo reconoce adecuadamente tanto los factores clínicos como los hábitos y condiciones académicas que contribuyen al desarrollo de la ansiedad.

El análisis por permutación también revela el papel destacado de variables personales como el **estado civil (con pareja)** y el **sexo**, lo que sugiere que las condiciones emocionales y sociales influyen directamente en la estabilidad psicológica del estudiante. Este hallazgo coincide con estudios previos que vinculan el apoyo afectivo y la percepción de acompañamiento emocional con menores niveles de ansiedad.

El árbol de decisión representativo muestra que la variable **gad7** actúa como el principal punto de división, seguida por el **puntaje total del GAD-7**, las **horas de sueño** y las **horas de estudio**. Esta secuencia de decisiones refleja un razonamiento clínico consistente: a medida que aumentan los síntomas emocionales y disminuye el descanso, la probabilidad de ansiedad leve o moderada

se eleva; mientras que los puntajes más bajos y un sueño adecuado se asocian con estados sin ansiedad.

En general, los resultados confirman que el modelo **aprende patrones clínicamente coherentes**, identifica correctamente las clases con mayor frecuencia (leve y moderada) y mantiene un comportamiento estable frente a datos no vistos. No obstante, se reconoce una limitación importante: el bajo número de casos de **ansiedad severa**, lo que afecta la capacidad del modelo para generalizar en esa categoría. Para futuras versiones, se recomienda ampliar la muestra y equilibrar las clases, lo que permitiría mejorar la sensibilidad del modelo en los casos clínicos más críticos.

5. CONCLUSIONES

El modelo de **Random Forest** logra una predicción precisa y confiable del nivel de ansiedad generalizada en estudiantes universitarios, con un desempeño global superior al **97 % de exactitud** en la etapa de prueba, demostrando su eficacia como herramienta de apoyo al tamizaje psicológico.

Las variables derivadas del **cuestionario GAD-7** constituyen los predictores más relevantes del modelo, confirmando la validez del instrumento como base diagnóstica. A la vez, los factores contextuales como el **sueño**, la **carga académica** y el **rendimiento estudiantil** complementan el análisis y fortalecen la interpretación integral del fenómeno.

El modelo muestra una **capacidad discriminativa casi perfecta** (ROC-AUC y PR-AUC > 0.99), lo que significa que distingue de manera efectiva entre los distintos niveles de ansiedad, especialmente en las categorías leve, moderada y sin ansiedad.

La fase de interpretación revela que el algoritmo no funciona como una "caja negra", sino como un sistema explicativo que permite entender qué variables influyen en la clasificación, favoreciendo la transparencia y la utilidad clínica del modelo.

Se concluye que la combinación de técnicas de **Machine Learning** con instrumentos clínicos estandarizados ofrece un enfoque innovador y efectivo para la detección temprana de ansiedad, aportando un valor agregado tanto al ámbito académico como al de la salud mental universitaria.

El estudio cumple con éxito su objetivo principal al **desarrollar un modelo predictivo basado en el algoritmo Random Forest** para la predicción del riesgo de ansiedad generalizada en estudiantes universitarios de Ayacucho, 2025. El modelo logra una **exactitud global del 97.1 %**, demostrando su capacidad para clasificar con alta precisión los diferentes niveles de ansiedad —sin ansiedad, leve,

moderada y severa— a partir de los ítems del cuestionario **GAD-7** y variables contextuales relacionadas con el entorno académico y personal. Este resultado evidencia la efectividad del enfoque de *Machine Learning* aplicado al ámbito de la salud mental, constituyéndose en una herramienta potencial de apoyo para la detección temprana y la intervención psicológica en el contexto universitario.

Objetivo específico a:

Se logra **identificar las variables que más influyen en la predicción del riesgo de ansiedad**, siendo el **puntaje total del GAD-7** y los ítems individuales **gad2, gad3, gad4, gad7 y gad5** los de mayor relevancia según el índice de importancia Gini. Asimismo, las **horas de sueño**, el **promedio académico**, las **horas de estudio por semana**, el **estrés académico** y el **apoyo familiar** destacan como variables complementarias que aportan contexto y mejoran la capacidad explicativa del modelo. Estos hallazgos confirman que los factores clínicos y académicos interactúan significativamente en la manifestación de la ansiedad estudiantil.

Objetivo específico b:

Se **diseña un modelo predictivo robusto** utilizando el algoritmo **Random Forest**, optimizado mediante la técnica **GridSearchCV** para determinar los mejores hiperparámetros. El modelo incorpora un *pipeline* que automatiza el preprocesamiento de los datos —imputación, codificación y partición estratificada—, asegurando un flujo reproducible y libre de fuga de información. La estructura del modelo y su jerarquía de decisiones, representada en los árboles del bosque, permiten visualizar de forma transparente cómo las variables clínicas y de estilo de vida influyen en el nivel de ansiedad. Este diseño cumple con los criterios metodológicos establecidos y garantiza interpretabilidad, precisión y estabilidad.

Objetivo específico c:

El modelo alcanza **excelentes métricas de rendimiento**, superando ampliamente el umbral mínimo esperado. En la fase de prueba obtiene una **precisión (accuracy) del 97.1 %**, un **F1-macro de 0.742**, una **sensibilidad promedio (recall) de 0.750** y un **ROC-AUC macro de 0.997**, lo que demuestra una discriminación casi perfecta entre los distintos niveles de ansiedad. Estas métricas confirman que el modelo no solo clasifica correctamente, sino que también calibra adecuadamente las probabilidades de cada predicción. Además, los resultados muestran una sólida capacidad para generalizar frente a nuevos datos, manteniendo consistencia entre las fases de validación y prueba.

La investigación demuestra que el modelo predictivo basado en **Random Forest** constituye una herramienta eficaz, transparente y confiable para estimar el riesgo de ansiedad generalizada en estudiantes universitarios. Su aplicación práctica permitiría **fortalecer los procesos de evaluación psicológica y prevención** en las instituciones de educación superior, ofreciendo una alternativa tecnológica que combina la rigurosidad científica con el compromiso ético de promover el bienestar emocional de la comunidad estudiantil.

6. RECOMENDACIONES

Ampliar el tamaño y equilibrio del dataset, especialmente en la clase de ansiedad severa, para fortalecer la capacidad de aprendizaje del modelo y reducir posibles sesgos asociados al desbalance de datos.

Integrar el modelo en una aplicación web o sistema de tamizaje automatizado, que permita a los servicios de bienestar universitario aplicar el GAD-7 y obtener en tiempo real una predicción del nivel de ansiedad de cada estudiante.

Incluir nuevas variables contextuales, como la calidad del sueño, el apoyo social percibido o el uso de dispositivos digitales, para enriquecer la interpretación y aumentar la sensibilidad del modelo frente a factores emergentes en la vida estudiantil.

Comparar el desempeño del Random Forest con otros algoritmos, como Gradient Boosting, XGBoost o Redes Neuronales, con el fin de identificar posibles mejoras en la precisión, robustez o interpretabilidad del sistema.

Validar el modelo en diferentes contextos universitarios, replicando el estudio en otras instituciones o regiones, para garantizar su generalización y utilidad en diversos entornos culturales y académicos.

Fomentar la colaboración interdisciplinaria entre ingenieros, psicólogos y profesionales en ciencias de la salud mental, de modo que la inteligencia artificial se utilice de forma ética, transparente y responsable en la prevención de problemas psicológicos.

Referencias bibliográficas

Dwyer, D. B., Harrison, B. J., Yücel, M., & Whittle, S. (2018). A machine learning approach to predicting anxiety and depression in adolescents using neuroimaging data. *Journal of Neural Transmission*, 125(4), 581-593. https://doi.org/10.1007/s00702-018-1870-6

Eisenberg, D., Golberstein, E., & Hunt, J. B. (2009). Mental health and academic success in college students: Evidence from a national survey. *Journal of Epidemiology & Community Health*, 63(7), 594-600. https://doi.org/10.1136/jech.2008.078873

Mantzoukas, S., & George, S. (2021). Predicting anxiety levels in college students using machine learning techniques. *Computers in Human Behavior*, 115, 106637. https://doi.org/10.1016/j.chb.2020.106637

Tran, U. S., & Glatz, C. (2019). Development of a predictive model for generalized anxiety disorder using machine learning and self-reported data. *Psychological Medicine*, 49(10), 1632-1640. https://doi.org/10.1017/S0033291718002274

Wang, S., & Ayer, T. (2020). Prediction of mental health conditions using machine learning techniques: An application to college students. *IEEE Access*, 8, 36788-36800. https://doi.org/10.1109/ACCESS.2020.2975197

Cohen, A. J., Chang, C., & Lee, J. (2023). Machine learning approaches to predicting anxiety in college students: A systematic review. Journal *of Anxiety Disorders, 97*, 1-15. https://doi.org/10.1016/j.janxdis.2023.102634

Duncan, S., Iqbal, N., & Garcia, C. (2022). Factors influencing mental health among engineering students: A qualitative study. *Engineering Education*, 17(1), 39-52. https://doi.org/10.1080/03043797.2021.1953424

Eisenberg, D., Gollust, S. E., Golberstein, E., & Hefner, J. L. (2016). The importance of mental health in college students. *International* Journal *of Environmental Research and Public Health, 13*(5), 446. https://doi.org/10.3390/ijerph13050446

González, A., Cordero, J. M., & Jiménez, A. (2020). Emotional well-being and academic performance: The role of anxiety in engineering students. *Journal of College Student Development, 61*(6), 728-733. https://doi.org/10.1353/csd.2020.0071

Kessler, R. C., Berglund, P., Demler, O., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry, 62*(6), 593-602. https://doi.org/10.1001/archpsyc.62.6.593

Liu, Z., Li, X., & Hu, Y. (2021). Application of machine learning algorithms for predicting anxiety among college students: A systematic review. *Psychological Bulletin, 147*(5), 510-532. https://doi.org/10.1037/bul0000284

Nath, R. K., Thapliyal, H., & Caban-Holt, A. (2020). Validating physiological stress detection model using cortisol as a stress biomarker. *IEEE International Conference on Consumer Electronics*. https://doi.org/10.1109/ICCE46568.2020.9042972

Petrescu, L., Petrescu, C., Mitruț, O., Moise, G., Moldoveanu, A., Moldoveanu, F., & Leordeanu, M. (2020). Integrating biosignals measurement in virtual reality environments for anxiety detection. *Sensors, 20*(1), 1-32. https://doi.org/10.3390/s20010088

Roslan, N. S., & Ahmad, A. (2020). The prevalence risk of anxiety and its associated factors among university students in Malaysia: A national cross-sectional study. *BMC Public Health, 20*(1), 1-12. https://doi.org/10.1186/s12889-020-08907-9

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine, 166*(10), 1092-1097. https://doi.org/10.1001/archinte.166.10.1092

ANEXO 1: Encuesta

Encuesta Estudiantil

Estimado(a) estudiante de la **Universidad Nacional de San Cristóbal de Huamanga:**

El presente cuestionario tiene como propósito **evaluar y comprender el nivel de ansiedad** que experimentan los estudiantes universitarios, considerando no solo los **síntomas emocionales** asociados a este estado (de acuerdo con el instrumento estandarizado **GAD-7**), sino también los **factores académicos, personales y de hábitos** que pueden influir en su bienestar psicológico.

Tu participación es **voluntaria, anónima y confidencial**. Los datos obtenidos serán utilizados **exclusivamente con fines académicos y de investigación**, contribuyendo al diseño de estrategias de apoyo psicológico y al fortalecimiento del bienestar estudiantil en nuestra comunidad universitaria. No existen respuestas correctas o incorrectas; lo importante es que respondas **de manera sincera y reflexiva**.

a. Datos generales del estudiante

Pregunta	Tipo de respuesta
(Asignado automáticamente)	Numérico
¿Cuál es tu edad?	Numérico (años)
¿Con qué sexo te identificas?	Hombre / Mujer / Otro
¿Cuál es tu estado civil actual?	Soltero(a) / Con pareja / Casado(a) / Otro
¿A qué carrera universitaria perteneces?	Ingeniería / Salud / Educación / Sociales / Económicas
¿En qué ciclo o semestre académico te encuentras actualmente?	Numérico (1–10)

b. Factores académicos

Pregunta	Tipo de respuesta
¿Cuál es tu promedio ponderado general en la universidad?	Numérico (0–20)
¿Cuántas horas dedicas al estudio durante la semana?	Numérico (horas)
En una escala del 0 al 10, ¿cuánto estrés académico sientes actualmente?	Escala (0–10)

c. Factores laborales

Pregunta	Tipo de respuesta
¿Actualmente trabajas mientras estudias?	Sí / No
En caso afirmativo, ¿cuántas horas trabajas a la semana?	Numérico (0–30)
En una escala del 0 al 10, ¿cuánto estrés financiero experimentas actualmente?	Escala (0–10)

d. Factores de salud y hábitos de vida

Pregunta	Tipo de respuesta
¿Cuántas horas duermes por noche en promedio?	Numérico (4–9 horas)
¿Cuántos días a la semana realizas actividad física?	Numérico (0–7)
¿Cuántas tazas de café o bebidas con cafeína consumes al día?	Numérico (0–5)
¿Cuántas horas al día pasas en redes sociales?	Numérico (1–10)
En una escala del 0 al 10, ¿cuánto apoyo sientes de tu familia?	Escala (0–10)
¿Has recibido atención psicológica o psiquiátrica en los últimos 12 meses?	Sí / No

e. Cuestionario GAD-7 (Escala de Ansiedad Generalizada)

Estas son las **7 preguntas oficiales del GAD-7**, cada una con **respuestas en escala Likert de 0 a 3**:

Pregunta (últimas 2 semanas)	Opciones de respuesta
¿Con qué frecuencia te has sentido nervioso, ansioso o al borde?	0 = Nada / 1 = Varios días / 2 = Más de la mitad de los días / 3 = Casi todos los días
¿Con qué frecuencia no has podido dejar de preocuparte o controlar la preocupación?	0–3
¿Con qué frecuencia te has preocupado demasiado por diferentes cosas?	0–3
¿Con qué frecuencia te ha resultado difícil relajarte?	0–3
¿Con qué frecuencia te has sentido tan inquieto que no puedes quedarte quieto?	0–3
¿Con qué frecuencia te has sentido fácilmente molesto o irritable?	0–3
¿Con qué frecuencia has sentido miedo como si algo terrible fuera a pasar?	0–3

ANEXO 2: Código Fuente

Random Forest (multiclase) con GridSearchCV para "nivel_ansiedad"

```
# ============================================================
# Random Forest (multiclase) con GridSearchCV para "nivel_ansiedad"
# Train/Valid/Test + métricas, gráficos y explicación de nodos
# Ejecutar en Jupyter (Anaconda)
# ============================================================

# -------------------- 0) Librerías --------------------
import os, re, warnings
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split, StratifiedKFold, GridSearchCV
from sklearn.preprocessing import OneHotEncoder, label_binarize
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.impute import SimpleImputer
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import (
    accuracy_score, f1_score, precision_score, recall_score,
    confusion_matrix, classification_report, brier_score_loss,
    roc_auc_score, roc_curve, precision_recall_curve, average_precision_score,
    make_scorer, log_loss
)
from sklearn.inspection import permutation_importance
from sklearn import tree

warnings.filterwarnings("ignore")
np.Random.seed(42)

# -------------------- 1) Rutas --------------------
EXCEL_PATH = "/mnt/data/Data_Procesada_GAD_13_10_2025_V4.xlsx"  # <-- AJUSTAR SI ES NECESARIO
OUT_DIR    = "./salidas_rf"
os.makedirs(OUT_DIR, exist_ok=True)
assert os.path.exists(EXCEL_PATH), f"No se encontró el archivo: {EXCEL_PATH}"

# -------------------- 2) Carga de datos --------------------
df = pd.read_excel(EXCEL_PATH).copy()

print("Columnas detectadas:")
for i, c in enumerate(df.columns):
    print(f"{i:2d} -> {c}")

TARGET_COL = "nivel_ansiedad"
```

```
assert TARGET_COL in df.columns, f"No se encontró la columna target '{TARGET_COL}'."

# -------------------- 3) (Opcional) Soporte a ítems GAD-7 en texto "(n)" --------------------
score_pat = re.compile(r"\((\d)\)")
def to_score(x):
    if pd.isna(x): return np.nan
    m = score_pat.search(str(x))
    if m: return float(m.group(1))
    txt = str(x).strip().lower()
    manual = {"nada":0,"nunca":0,"varios días":1,"varios dias":1,
              "más de la mitad de los días":2,"mas de la mitad de los dias":2,
              "casi todos los días":3,"casi todos los dias":3}
    return manual.get(txt, np.nan)

gad_cols = []
if df.shape[1] >= 15:
    gad_cols = [df.columns[i] for i in range(8, 15)]
if len(gad_cols) != 7:
    pref = "¿con qué frecuencia le han molestado los siguientes problemas"
    gad_cols = [c for c in df.columns if str(c).strip().lower().startswith(pref)]
    gad_cols = gad_cols[:7]
for c in gad_cols:
    if c in df.columns and df[c].dtype == "object":
        df[c] = df[c].apply(to_score)

# -------------------- 4) Features/Target --------------------
exclude_cols = {TARGET_COL, "subject_id", "id", "dni", "codigo"}
feature_candidates = [c for c in df.columns if c not in exclude_cols]
categorical_feats = [c for c in feature_candidates if df[c].dtype == "object"]
numeric_feats     = [c for c in feature_candidates if c not in categorical_feats]

print("\nVariables numéricas (muestra):", numeric_feats[:10], "..." if len(numeric_feats)>10
else "")
print("Variables categóricas (muestra):", categorical_feats[:10], "..." if
len(categorical_feats)>10 else "")

data = df[feature_candidates + [TARGET_COL]].copy()
data = data[data[TARGET_COL].notna()]

X_full = data[feature_candidates]
y_full = data[TARGET_COL].astype(str)
classes = sorted(y_full.unique().tolist())
print("\nClases detectadas:", classes)

# -------------------- 5) Split Train/Valid/Test (60/20/20) --------------------
X_temp, X_test, y_temp, y_test = train_test_split(
    X_full, y_full, test_size=0.20, stratify=y_full, Random_state=42
)
X_train, X_valid, y_train, y_valid = train_test_split(
```

```
    X_temp, y_temp, test_size=0.25, stratify=y_temp, Random_state=42
)
print(f"\nTamaños -> Train: {X_train.shape[0]} | Valid: {X_valid.shape[0]} | Test:
{X_test.shape[0]}")

# -------------------- 6) Pipeline base (Imputación + OneHot + RF) --------------------
numeric_transformer = Pipeline(steps=[("imputer", SimpleImputer(strategy="median"))])
categorical_transformer = Pipeline(steps=[
    ("imputer", SimpleImputer(strategy="most_frequent")),
    ("onehot", OneHotEncoder(handle_unknown="ignore"))
])
preprocess = ColumnTransformer(
    transformers=[
        ("num", numeric_transformer, numeric_feats),
        ("cat", categorical_transformer, categorical_feats)
    ],
    remainder="drop"
)

rf = RandomForestClassifier(
    Random_state=42, n_jobs=-1, class_weight="balanced"
)

pipe = Pipeline(steps=[("prep", preprocess), ("rf", rf)])

# -------------------- 7) GridSearchCV (5-fold Estratificado) --------------------
# Scorers múltiples y refit por f1_macro
scorers = {
    "f1_macro": make_scorer(f1_score, average="macro"),
    "accuracy": make_scorer(accuracy_score),
    "roc_auc_ovr": make_scorer(roc_auc_score, needs_proba=True, average="macro",
multi_class="ovr"),
    "neg_log_loss": make_scorer(log_loss, greater_is_better=False, needs_proba=True,
labels=classes)
}

param_grid = {
    "rf__n_estimators":    [100, 200, 350],
    "rf__max_depth":       [None, 8, 12, 16],
    "rf__min_samples_split":[2, 4, 8],
    "rf__min_samples_leaf":[1, 2, 4],
    "rf__max_features":    ["sqrt", "log2", 0.5]
}

cv = StratifiedKFold(n_splits=5, shuffle=True, Random_state=42)

grid = GridSearchCV(
    estimator=pipe,
    param_grid=param_grid,
```

```
    scoring=scorers,
    refit="f1_macro",
    cv=cv,
    n_jobs=-1,
    verbose=1,
    return_train_score=True
)

print("\n⏳ Ejecutando GridSearchCV (esto puede tardar según el tamaño/PC)...")
grid.fit(X_train, y_train)

print("\nMejores parámetros (según f1_macro):")
print(grid.best_params_)
print(f"Mejor f1_macro CV: {grid.best_score_:.4f}")

# Guardar resultados completos del grid
cv_results = pd.DataFrame(grid.cv_results_)
cv_results.to_csv(os.path.join(OUT_DIR, "gridsearch_cv_results.csv"), index=False)

best_model = grid.best_estimator_

# -------------------- 8) Funciones de evaluación --------------------
def eval_multiclass(y_true, y_pred, y_proba, class_labels, split_name, out_dir):
    print(f"\n=== Resultados en {split_name} ===")
    acc        = accuracy_score(y_true, y_pred)
    f1_macro   = f1_score(y_true, y_pred, average="macro")
    f1_micro   = f1_score(y_true, y_pred, average="micro")
    prec_macro = precision_score(y_true, y_pred, average="macro", zero_division=0)
    rec_macro  = recall_score(y_true, y_pred, average="macro")
    print(f"Accuracy: {acc:.3f}")
    print(f"F1 (macro): {f1_macro:.3f}")
    print(f"F1 (micro): {f1_micro:.3f}")
    print(f"Precision (macro): {prec_macro:.3f}")
    print(f"Recall (macro): {rec_macro:.3f}")

    cm = confusion_matrix(y_true, y_pred, labels=class_labels)
    print("\nMatriz de confusión (filas=real, columnas=predicho):\n", cm)
    print("\nReporte por clase:")
    print(classification_report(y_true, y_pred, digits=3))

    Y_bin = label_binarize(y_true, classes=class_labels)
    briers = [brier_score_loss(Y_bin[:, i], y_proba[:, i]) for i in range(len(class_labels))]
    brier_mean = float(np.mean(briers))
    print(f"Brier score (promedio de clases): {brier_mean:.4f}")

    try:
        roc_auc_macro_ovr = roc_auc_score(Y_bin, y_proba, average="macro", multi_class="ovr")
    except Exception:
        roc_auc_macro_ovr = np.nan
```

```
    print(f"ROC-AUC (macro, OVR): {roc_auc_macro_ovr:.3f}" if not np.isnan(roc_auc_macro_ovr)
else "ROC-AUC (macro, OVR) no disponible.")
    pr_aucs = [average_precision_score(Y_bin[:, i], y_proba[:, i]) for i in
range(len(class_labels))]
    pr_auc_macro = float(np.mean(pr_aucs))
    print(f"PR-AUC (macro): {pr_auc_macro:.3f}")

    # Graficar
    plt.figure()
    plt.imshow(cm, interpolation="nearest")
    plt.title(f"Matriz de Confusión - {split_name}")
    plt.colorbar()
    ticks = np.arange(len(class_labels))
    plt.xticks(ticks, class_labels, rotation=45, ha="right")
    plt.yticks(ticks, class_labels)
    for i in range(cm.shape[0]):
        for j in range(cm.shape[1]):
            plt.text(j, i, str(cm[i, j]), ha="center", va="center")
    plt.ylabel("Real"); plt.xlabel("Predicción")
    plt.tight_layout()
    plt.savefig(os.path.join(out_dir, f"confusion_matrix_{split_name.lower()}.png"),
bbox_inches="tight")
    plt.show()

    for i, cls in enumerate(class_labels):
        fpr, tpr, _ = roc_curve(Y_bin[:, i], y_proba[:, i])
        plt.figure()
        plt.plot(fpr, tpr)
        plt.xlabel("1 - Especificidad (FPR)"); plt.ylabel("Sensibilidad (TPR)")
        plt.title(f"Curva ROC - {split_name} - Clase: {cls}")
        plt.savefig(os.path.join(out_dir, f"roc_{split_name.lower()}_{i}_{cls}.png"),
bbox_inches="tight")
        plt.show()

    for i, cls in enumerate(class_labels):
        precisions, recalls, _ = precision_recall_curve(Y_bin[:, i], y_proba[:, i])
        plt.figure()
        plt.plot(recalls, precisions)
        plt.xlabel("Recall (Sensibilidad)"); plt.ylabel("Precisión (PPV)")
        plt.title(f"Curva Precision-Recall - {split_name} - Clase: {cls}")
        plt.savefig(os.path.join(out_dir, f"pr_{split_name.lower()}_{i}_{cls}.png"),
bbox_inches="tight")
        plt.show()

    return {
        "split": split_name, "accuracy": acc,
        "f1_macro": f1_macro, "f1_micro": f1_micro,
        "precision_macro": prec_macro, "recall_macro": rec_macro,
        "brier_mean": brier_mean, "roc_auc_macro_ovr": roc_auc_macro_ovr,
```

```
        "pr_auc_macro": pr_auc_macro
    }

# -------------------- 9) Evaluación del mejor modelo en Validación --------------------
y_valid_pred  = best_model.predict(X_valid)
y_valid_proba = best_model.predict_proba(X_valid)
metrics_valid = eval_multiclass(y_valid, y_valid_pred, y_valid_proba, classes, "Validación",
OUT_DIR)

# -------------------- 10) Reentrenar con Train+Valid y evaluar en Test --------------------
X_trval = pd.concat([X_train, X_valid], axis=0)
y_trval = pd.concat([y_train, y_valid], axis=0)

final_model = grid.best_estimator_  # misma estructura/params
final_model.fit(X_trval, y_trval)

y_test_pred  = final_model.predict(X_test)
y_test_proba = final_model.predict_proba(X_test)
metrics_test = eval_multiclass(y_test, y_test_pred, y_test_proba, classes, "Prueba", OUT_DIR)

pd.DataFrame([metrics_valid, metrics_test]).to_csv(os.path.join(OUT_DIR,
"metrics_valid_test.csv"), index=False)

# -------------------- 11) Importancias --------------------
rf_fitted = final_model.named_steps["rf"]
ohe = final_model.named_steps["prep"].named_transformers_["cat"].named_steps.get("onehot",
None)
cat_names = list(ohe.get_feature_names_out(categorical_feats)) if (ohe is not None and
len(categorical_feats)>0) else []
feature_names = list(numeric_feats) + cat_names

gini_importances = pd.DataFrame({
    "feature": feature_names,
    "gini_importance": rf_fitted.feature_importances_
}).sort_values("gini_importance", ascending=False)

print("\nTop 20 importancias (Gini):")
display(gini_importances.head(20))
gini_importances.to_csv(os.path.join(OUT_DIR, "importances_gini_rf.csv"), index=False)

plt.figure(figsize=(8, max(4, 0.3*min(20, len(gini_importances)))))
top_gini = gini_importances.head(20)
plt.barh(range(top_gini.shape[0]), top_gini["gini_importance"].iloc[::-1].values)
plt.yticks(range(top_gini.shape[0]), top_gini["feature"].iloc[::-1].values)
plt.xlabel("Importancia (Gini)"); plt.title("Top Importancias (Gini) - Random Forest")
plt.tight_layout(); plt.savefig(os.path.join(OUT_DIR, "importances_gini_rf.png"),
bbox_inches="tight")
plt.show()
```

```
# Importancia por permutación (en Test para evitar sesgo)
perm = permutation_importance(final_model, X_test, y_test, n_repeats=5, Random_state=42,
n_jobs=-1)
perm_importances = pd.DataFrame({
    "feature": feature_names[:len(perm.importances_mean)],
    "mean_importance": perm.importances_mean,
    "std_importance": perm.importances_std
}).sort_values("mean_importance", ascending=False)
print("\nTop 20 importancias (Permutación) - Test:")
display(perm_importances.head(20))
perm_importances.to_csv(os.path.join(OUT_DIR, "importances_permutation_rf_test.csv"),
index=False)

plt.figure(figsize=(8, max(4, 0.3*min(20, len(perm_importances)))))
top_perm = perm_importances.head(20)
plt.barh(range(top_perm.shape[0]), top_perm["mean_importance"].iloc[::-1].values)
plt.yticks(range(top_perm.shape[0]), top_perm["feature"].iloc[::-1].values)
plt.xlabel("Importancia por permutación (disminución media)")
plt.title("Top Importancias (Permutación) - Test")
plt.tight_layout(); plt.savefig(os.path.join(OUT_DIR, "importances_permutation_rf_test.png"),
bbox_inches="tight")
plt.show()

# -------------------- 12) “Gráfico de bosque” --------------------
# (A) Visualizar un árbol del bosque (profundidad limitada)
est0 = rf_fitted.estimators_[0]
plt.figure(figsize=(12, 6))
tree.plot_tree(
    est0, max_depth=3, filled=False,
    feature_names=feature_names, class_names=classes, proportion=True
)
plt.title("Árbol 1 del Bosque (max_depth=3)")
plt.savefig(os.path.join(OUT_DIR, "Forest_tree_example_depth3.png"), bbox_inches="tight")
plt.show()

# (B) Desempeño vs n_estimators (con mejores parámetros excepto n_estimators)
sizes = [50, 100, 150, 200, 300, 400]
accs  = []
base_params = grid.best_params_.copy()
base_params.pop("rf__n_estimators", None)

for n_est in sizes:
    rf_tmp = RandomForestClassifier(
        n_estimators=n_est,
        max_depth=base_params.get("rf__max_depth"),
        min_samples_split=base_params.get("rf__min_samples_split", 2),
        min_samples_leaf=base_params.get("rf__min_samples_leaf", 1),
        max_features=base_params.get("rf__max_features", "sqrt"),
        class_weight="balanced",
```

```
        Random_state=42, n_jobs=-1
    )
    pipe_tmp = Pipeline(steps=[("prep", preprocess), ("rf", rf_tmp)])
    pipe_tmp.fit(X_trval, y_trval)
    y_test_tmp = pipe_tmp.predict(X_test)
    accs.append(accuracy_score(y_test, y_test_tmp))

plt.figure()
plt.plot(sizes, accs, marker="o")
plt.xlabel("Número de árboles (n_estimators)")
plt.ylabel("Accuracy en Test")
plt.title("Desempeño vs Cantidad de árboles (Random Forest)")
plt.savefig(os.path.join(OUT_DIR, "Forest_size_vs_accuracy_test.png"), bbox_inches="tight")
plt.show()

# -------------------- 13) Explicación detallada de nodos --------------------
def explain_tree_nodes(decision_tree, class_names, feature_names_transformed, out_csv_path):
    t = decision_tree.tree_
    rows = []
    for node_id in range(t.node_count):
        is_leaf = (t.children_left[node_id] == t.children_right[node_id])
        v = t.value[node_id]
        if v.ndim == 2 and v.shape[0] == 1:
            v = v[0]
        v = v.astype(int).tolist()
        pred_idx = int(np.argmax(v)) if sum(v) > 0 else 0
        pred_class = class_names[pred_idx]
        feat_idx = t.feature[node_id]
        feat_name = feature_names_transformed[feat_idx] if (not is_leaf and 0 <= feat_idx <
len(feature_names_transformed)) else ""
        thr = float(t.threshold[node_id]) if not is_leaf else np.nan
        rows.append({
            "node_id": int(node_id),
            "is_leaf": bool(is_leaf),
            "feature_index": int(feat_idx),
            "feature_name_transformed": feat_name,
            "threshold": thr,
            "impurity_gini": float(t.impurity[node_id]),
            "n_node_samples": int(t.n_node_samples[node_id]),
            "class_counts": v,
            "predicted_class": pred_class
        })
    df_nodes = pd.DataFrame(rows)
    df_nodes.to_csv(out_csv_path, index=False)
    return df_nodes

nodes_csv = os.path.join(OUT_DIR, "tree_1_nodes_detail.csv")
nodes_df  = explain_tree_nodes(est0, classes, feature_names, nodes_csv)
print("\n=== Nodos del Árbol 1 (primeras 15 filas) ===")
```

```
display(nodes_df.head(15))
print(f"\nCSV con todos los nodos guardado en: {nodes_csv}")

# -------------------- 14) Resumen y archivos --------------------
print("\n=== Mejores hiperparámetros (GridSearchCV) ===")
print(grid.best_params_)
print("\nResultados de CV guardados en: gridsearch_cv_results.csv")

print("\n=== Archivos generados en ./salidas_rf ===")
for fn in [
    "gridsearch_cv_results.csv",
    "metrics_valid_test.csv",
    "confusion_matrix_validación.png",
    "confusion_matrix_prueba.png",
    "importances_gini_rf.csv",
    "importances_permutation_rf_test.csv",
    "Forest_tree_example_depth3.png",
    "Forest_size_vs_accuracy_test.png",
    "tree_1_nodes_detail.csv",
]:
    print(" -", fn)
```

ANEXO 3: Interfaces de la encuesta estudiantil - cuestionario

a. Datos generales del estudiante

¿ Cual es tu edad? *

Tu respuesta

¿Con que sexo te identificas? *

- Hombre
- Mujer
- Otro

¿Cual es tu estado civil actual? *

- Soltero
- Con pareja
- Casado
- otro

¿A que carrera universitaria perteneces? *

- Ingeniería
- Salud
- Educación
- Sociales
- Económicas

¿En que ciclo o semestre académico te encuentras actualmente? *

Tu respuesta

b. Factores académicos

¿Cuál es tu promedio ponderado general en la universidad? *

Tu respuesta

¿Cuántas horas dedicas al estudio durante la semana? *

Tu respuesta

En una escala de 0 a 10 ¿ cuanto estrés académico sientes actualmente? *

0	1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○	○

c. Factores laborales

¿Actualmente trabajas mientras estudias? *

- ○ Si
- ○ No

En caso afirmativo ¿Cuántas horas trabajas a la semana? *

Tu respuesta

En una escala del 0 al 10 ¿Cuánto estrés financiero experimentas actualmente? *

1 2 3 4 5 6 7 8 9 10

d. Factores de salud y hábitos de vida

¿Cuántas horas duermes por las noches en promedio? *

Tu respuesta

¿Cuántos días a la semana realizas actividad física ? *

Tu respuesta

¿Cuántas tazas de café o bebidas con cafeína consumes al día? *

Tu respuesta

¿Cuántas horas al día pasas en redes sociales? *

Tu respuesta

En una escala del 0 al 10 ¿Cuánto apoyo sientes de tu familia? *

1 2 3 4 5 6 7 8 9 10

¿Has recibido atención psicológica o psiquiátrica en los últimos 12 meses? *

- ◯ SI
- ◯ No

e. Cuestionario GAD-7(Escala de Ansiedad Generalizada)

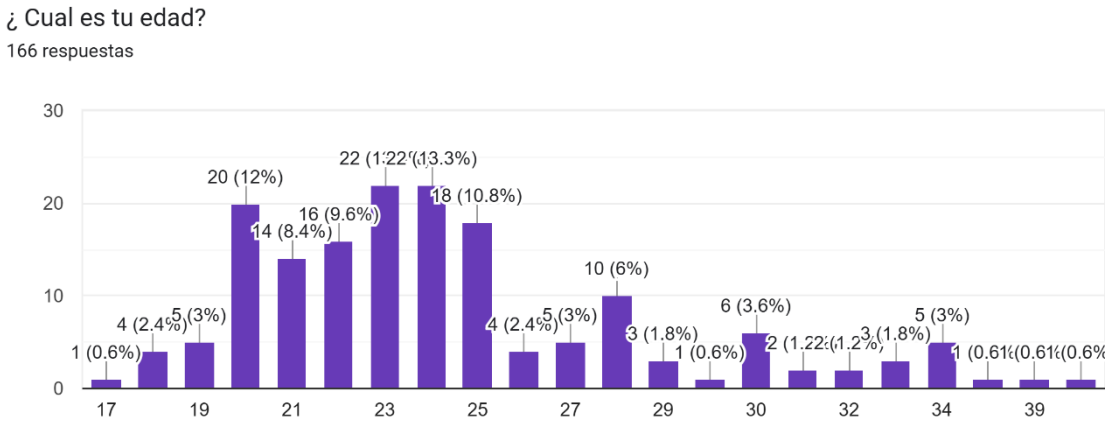
¿Con qué frecuencia te has sentido nervioso, ansioso o al borde? *

	Nada	Varios días	Más de la mitad de los días	Casi todos los días
¿Con qué frecuencia te has sentido nervioso, ansioso o al borde?	◯	◯	◯	◯
¿Con qué frecuencia no has podido dejar de preocuparte o controlar la preocupación?	◯	◯	◯	◯
¿Con qué frecuencia te has preocupado demasiado por diferentes cosas?	◯	◯	◯	◯
¿Con qué frecuencia te ha resultado difícil relajarte?	◯	◯	◯	◯
¿Con qué frecuencia te has sentido tan inquieto que no puedes quedarte quieto?	◯	◯	◯	◯
¿Con qué frecuencia te has sentido fácilmente molesto o irritable?	◯	◯	◯	◯
¿Con qué frecuencia has sentido miedo como si algo terrible fuera a pasar?	◯	◯	◯	◯

ANEXO 4: Estadísticas de la encuesta

Figura 11

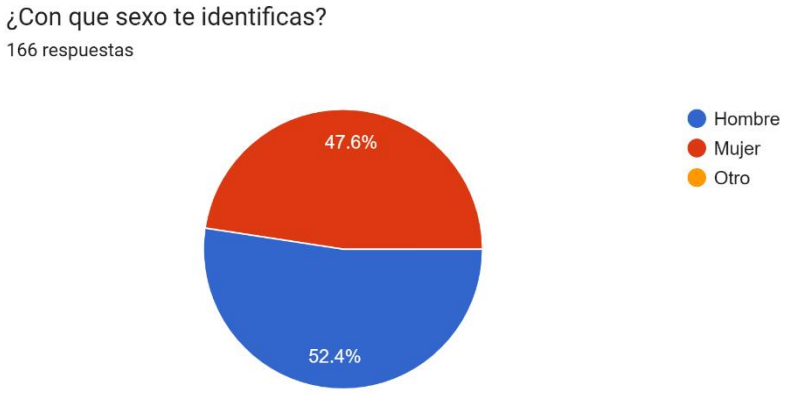
Gráfico en barras de la distribución de los estudiantes su según edad.



Fuente: Elaboración propia a partir de los resultados de la encuesta "Evaluación del nivel de ansiedad (Escala GAD-7) y factores relacionados" (2025).

Figura 12

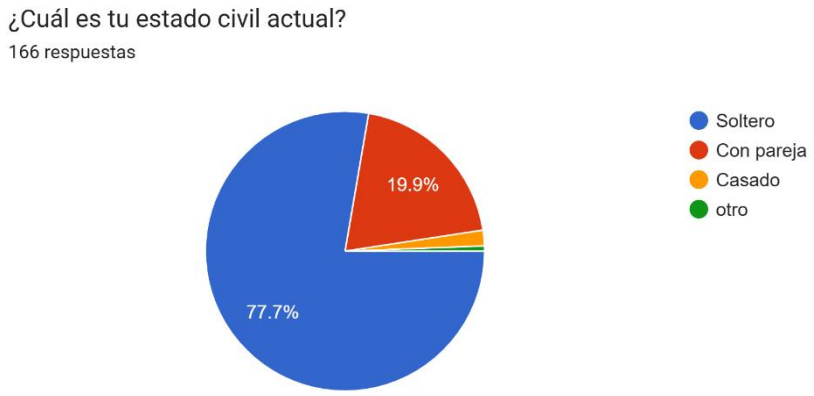
Gráfico circular de la distribución de los estudiantes según sexo.



Fuente: Elaboración propia a partir de los resultados de la encuesta "Evaluación del nivel de ansiedad (escala GAD-7) y factores relacionados" (2025).

Figura 13

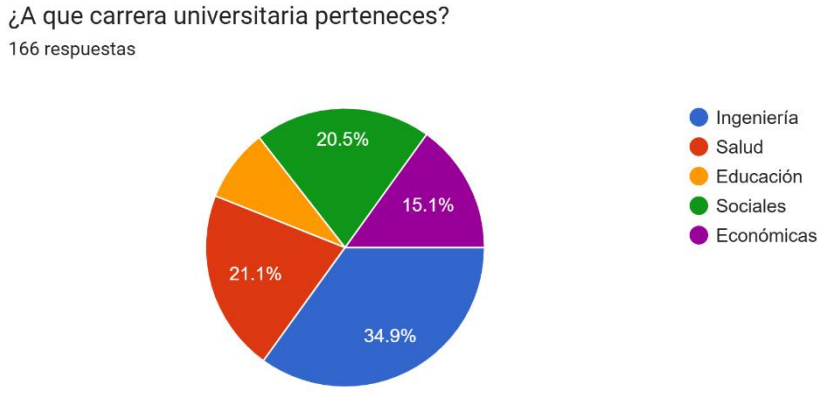
Gráfico circular de la distribución de los estudiantes según su estado civil.



Fuente: Elaboración propia a partir de los resultados de la encuesta "Evaluación del nivel de ansiedad (escala GAD-7) y factores relacionados" (2025).

Figura 14

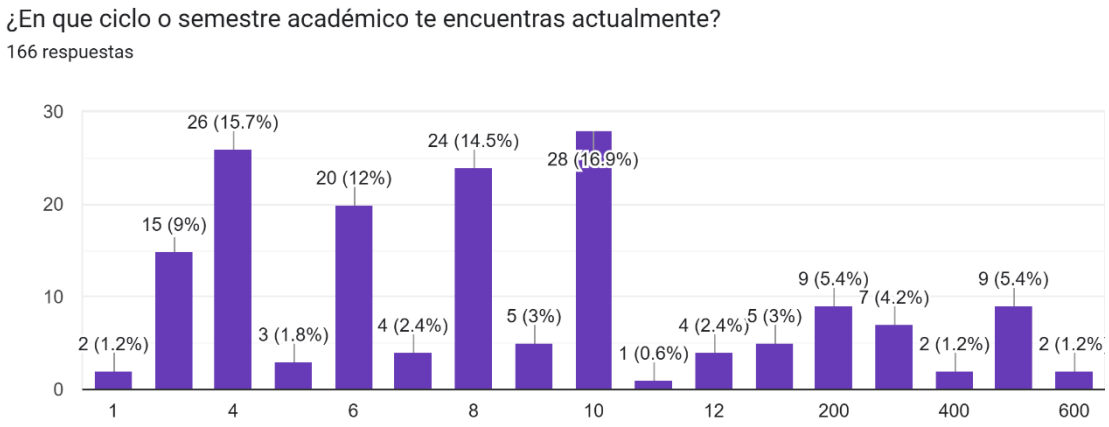
Gráfico circular de la distribución de los estudiantes según su carrera universitaria que pertenece.



Fuente: Elaboración propia a partir de los resultados de la encuesta "Evaluación del Nivel de Ansiedad (Escala GAD-7) y Factores Relacionados" (2025).

Figura 15

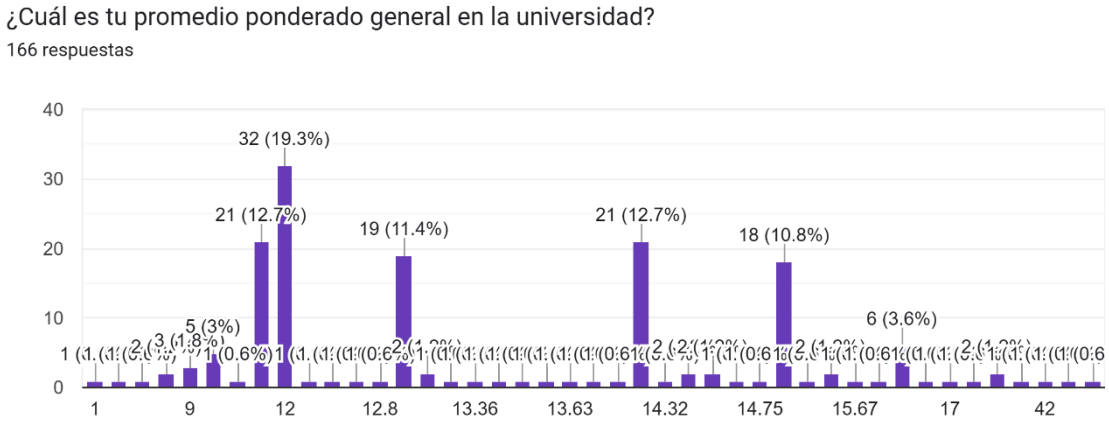
Gráfico en barras de la distribución de los estudiantes de acuerdo al ciclo o semestre académico.



Fuente: Elaboración propia a partir de los resultados de la encuesta "Evaluación del nivel de ansiedad (escala GAD-7) y factores relacionados" (2025).

Figura 16

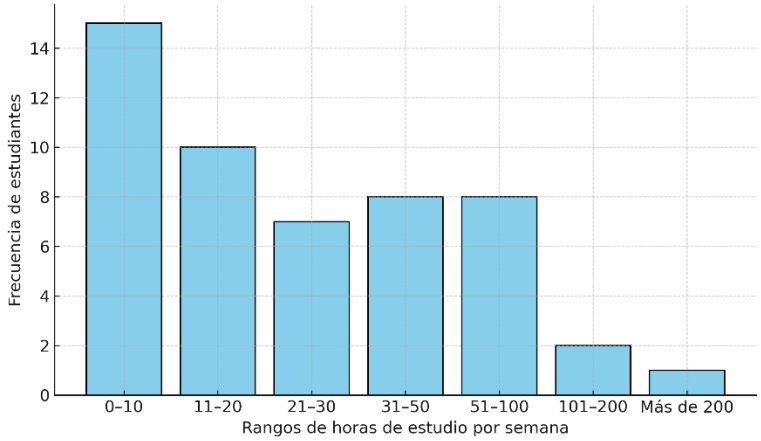
Gráfico en barras de la distribución de los estudiantes de acuerdo al promedio ponderado general.



Fuente: Elaboración propia a partir de los resultados de la encuesta "Evaluación del nivel de ansiedad (escala GAD-7) y factores relacionados" (2025).

Figura 17

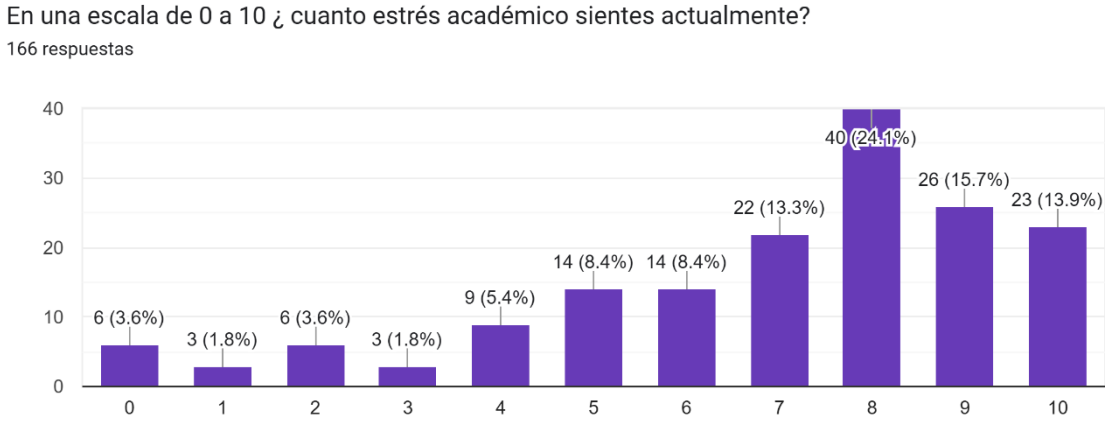
Gráfico en barras de la distribución de horas de estudio semanales en los estudiantes semanales.



Fuente: Elaboración propia a partir de los resultados de la encuesta "Evaluación del nivel de ansiedad (escala GAD-7) y factores relacionados" (2025).

Figura 18

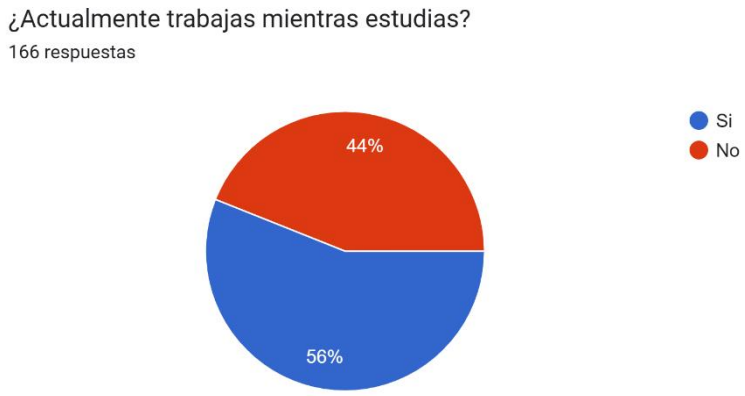
Gráfico en barras de la distribución de los estudiantes en cuanto al estrés académico.



Fuente: Elaboración propia a partir de los resultados de la encuesta "Evaluación del nivel de ansiedad (escala GAD-7) y factores relacionados" (2025).

Figura 19

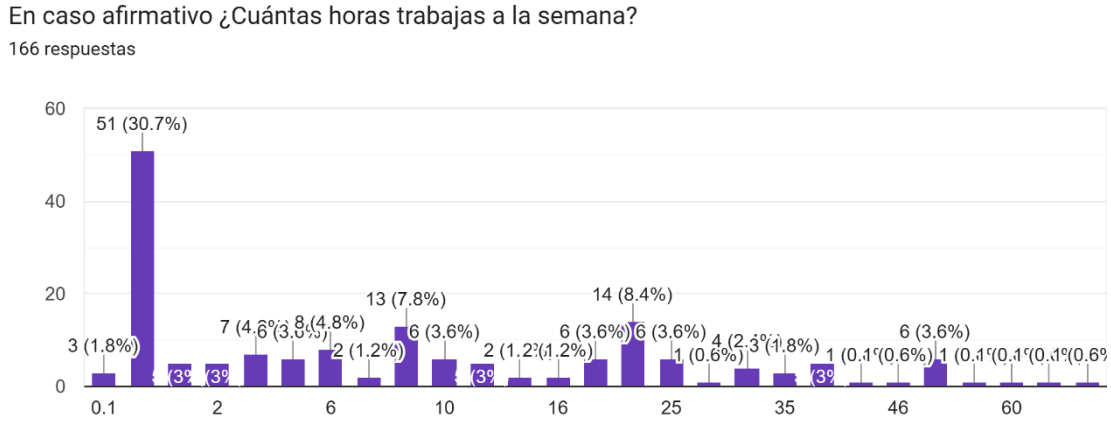
Gráfico circular de la distribución de los estudiantes que trabaja mientras estudia.



Fuente: Elaboración propia a partir de los resultados de la encuesta "Evaluación del nivel de ansiedad (escala GAD-7) y factores relacionados" (2025).

Figura 20

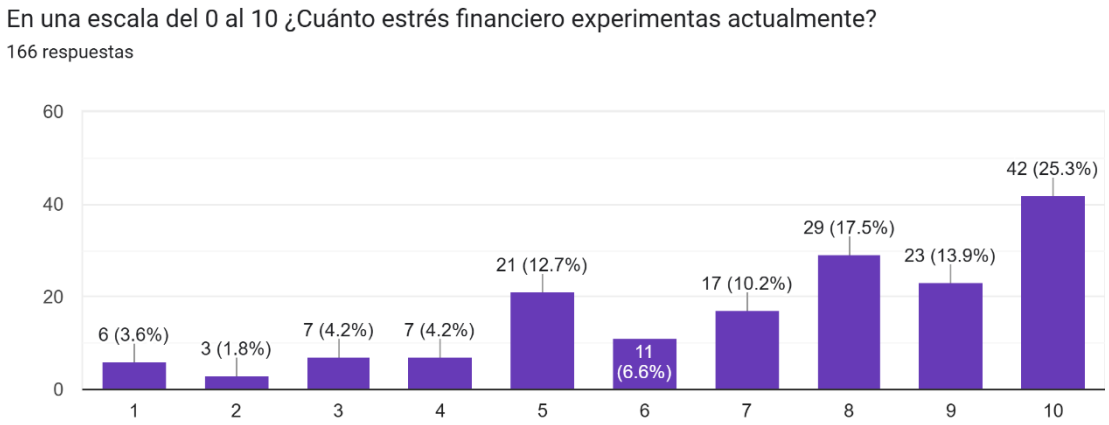
Gráfico en barras de la distribución de los estudiantes en caso de ser afirmativo las horas que trabaja a la semana.



Fuente: Elaboración propia a partir de los resultados de la encuesta "Evaluación del nivel de ansiedad (escala GAD-7) y factores relacionados" (2025).

Figura 21

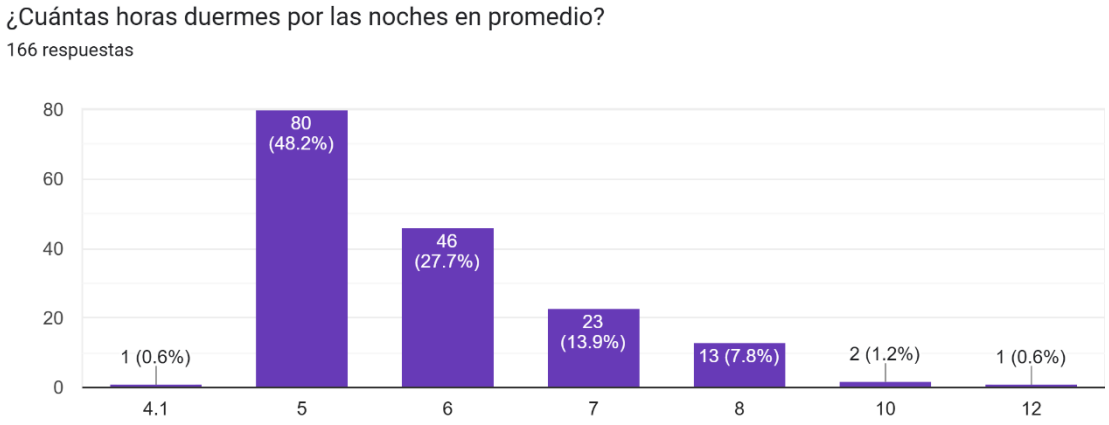
Gráfico en barras de la distribución de los estudiantes en cuanto al estrés financiero.



Fuente: Elaboración propia a partir de los resultados de la encuesta "Evaluación del nivel de ansiedad (escala GAD-7) y factores relacionados" (2025).

Figura 22

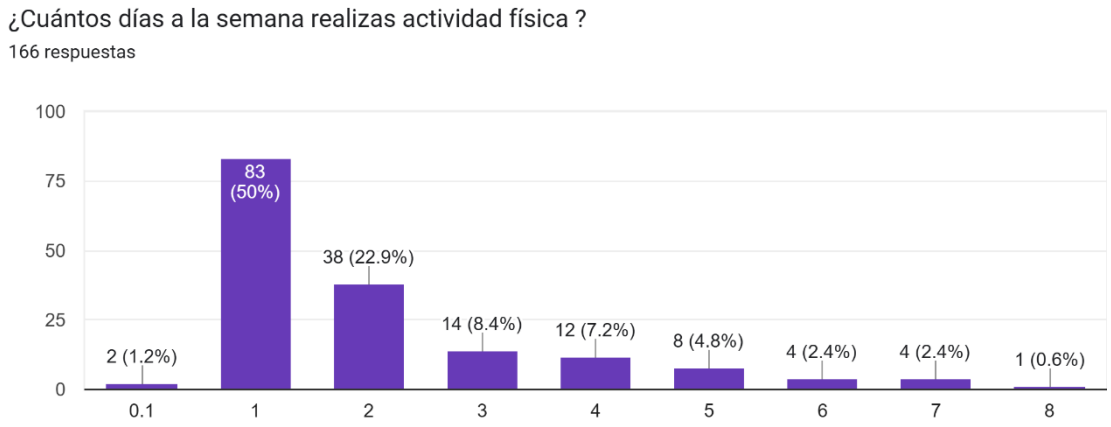
Gráfico en barras de la distribución de los estudiantes las horas promedio que duerme por las noches.



Fuente: Elaboración propia a partir de los resultados de la encuesta "Evaluación del nivel de ansiedad (escala GAD-7) y factores relacionados" (2025).

Figura 23

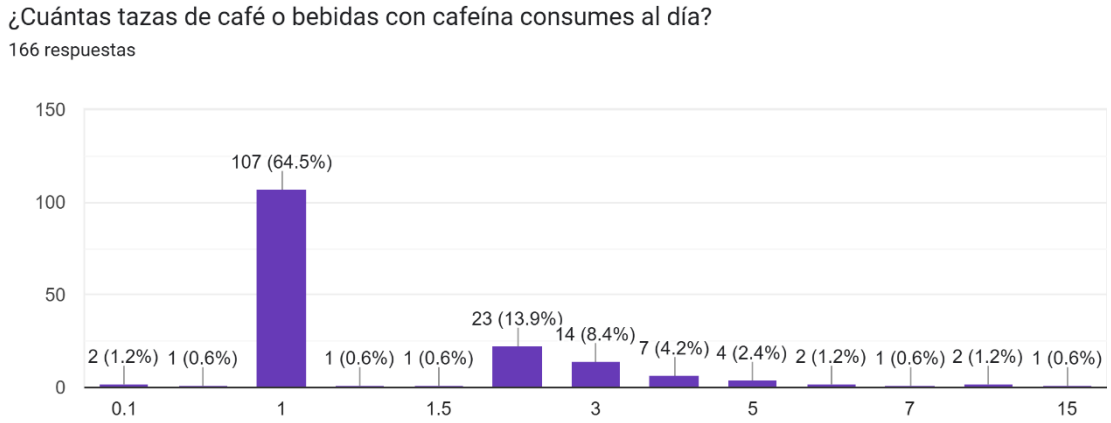
Gráfico en barras de la distribución de los estudiantes las horas de actividad física realiza.



Fuente: Elaboración propia a partir de los resultados de la encuesta "Evaluación del nivel de ansiedad (escala GAD-7) y factores relacionados" (2025).

Figura 24

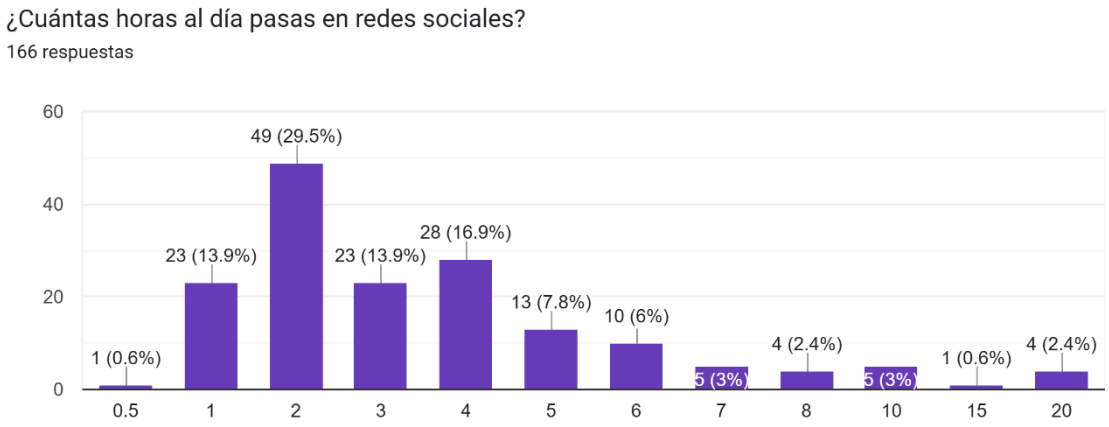
Gráfico en barras de la distribución de los estudiantes de cuantas tazas de café o cafeína consume por día.



Fuente: Elaboración propia a partir de los resultados de la encuesta "Evaluación del nivel de ansiedad (escala GAD-7) y factores elacionados" (2025).

Figura 25

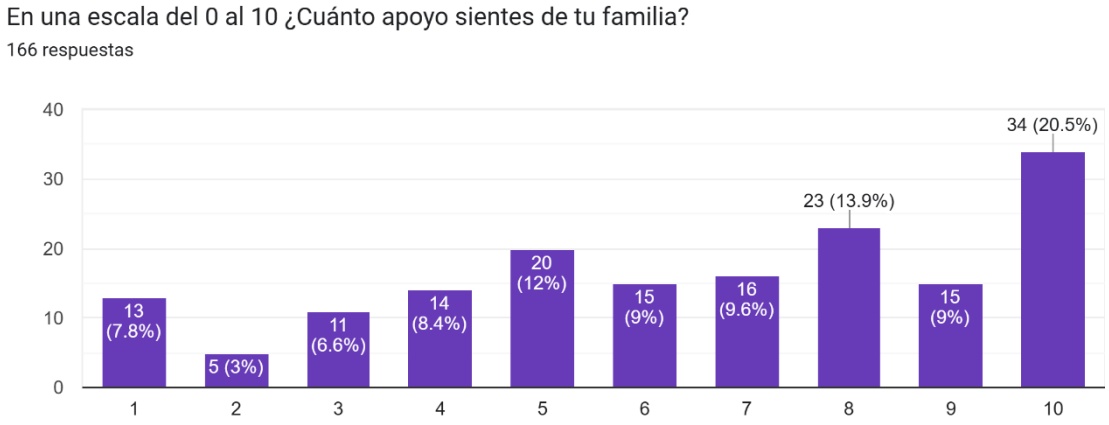
Gráfico en barras de la distribución de los estudiantes las horas que pasa por día en las redes sociales.



Fuente: Elaboración propia a partir de los resultados de la encuesta "Evaluación del nivel de ansiedad (escala GAD-7) y factores relacionados" (2025).

Figura 26

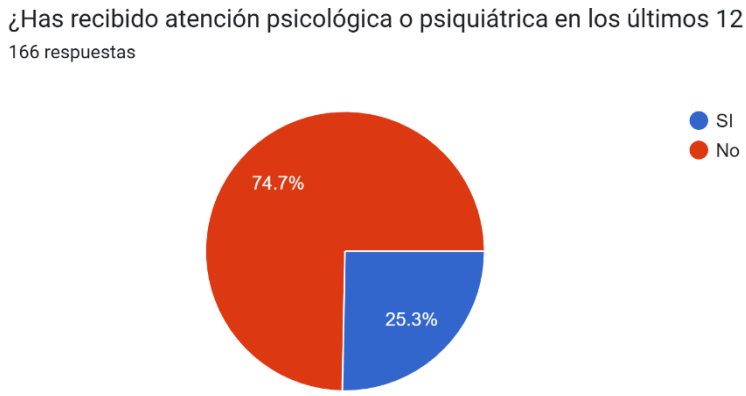
Gráfico de barras de la distribución de los estudiantes que recibe el apoyo de su familia.



Fuente: Elaboración propia a partir de los resultados de la encuesta "Evaluación del nivel de ansiedad (escala GAD-7) y factores relacionados" (2025).

Figura 27

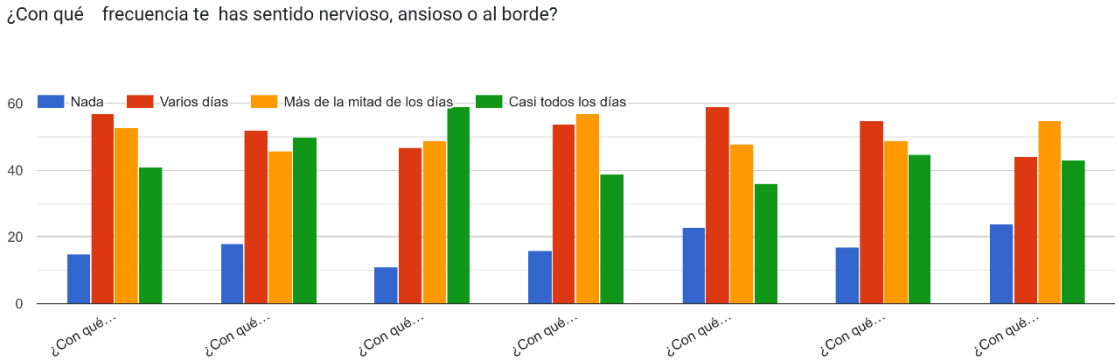
Gráfico circular de la distribución de los estudiantes que ha recibido atención psicológica o psiquiátrica en los últimos 12 meses.



Fuente: Elaboración propia a partir de los resultados de la encuesta "Evaluación del nivel de ansiedad (escala GAD-7) y factores relacionados" (2025).

Figura 28

Gráfico de barras de la Distribución de los estudiantes con la frecuencia que se ha sentido nervioso, ansioso.



Fuente: Elaboración propia a partir de los resultados de la encuesta "Evaluación del nivel de ansiedad (escala GAD-7) y factores relacionados" (2025).

ANEXO 5: Dataset

estudian	o_academ	edio_pond	estudio_s	a_actualm	trabajo_s	oras_sueñ	idad_fisica	sumo_cafe	_redes_h	oyo_famil	es_financ	es_acader	ion_psicol	gad1	gad2	gad3	gad4	gad5	gad6	gad7	vel_ansiedad
1	2	13,5	14	1	29	4,8	3	4	4,1	3	9	10	0	1	3	2	2	2	2	2	Ansiedad moderada
2	8	16,2	37	1	9	6,5	5	5	2,7	0	3	4	0	1	1	1	1	1	1	1	Ansiedad leve
3	1	14,7	20	0	12	8	1	4	8,4	6	0	6	0	1	1	2	1	2	2	1	Ansiedad moderada
4	6	15,2	39	1	24	6,2	2	0	9,2	7	0	10	0	1	2	1	2	1	1	2	Ansiedad moderada
5	6	14,9	7	0	16	6,5	6	2	5,8	4	7	6	0	2	2	1	1	1	1	1	Ansiedad leve
6	4	15,5	9	1	6	8,2	7	5	7,5	1	6	4	0	1	0	0	0	1	1	1	Sin ansiedad
7	10	12,5	30	0	23	7,5	7	0	2,1	7	7	1	0	2	1	2	1	2	2	2	Ansiedad moderada
8	3	11,6	15	0	2	6,5	3	5	6,5	0	3	4	0	1	2	1	1	1	2	1	Ansiedad leve
9	3	17,4	11	0	29	5,3	6	2	7,3	6	5	6	1	3	2	2	2	2	1	2	Ansiedad moderada
10	1	15	14	1	19	8,4	3	5	5,1	0	2	6	0	0	1	0	1	0	0	0	Sin ansiedad
11	3	16	38	1	12	6,9	0	0	3,2	5	10	3	1	1	2	1	1	1	1	1	Ansiedad leve
12	4	11	7	1	18	5,1	6	3	8,9	5	7	0	1	3	1	2	2	1	2	2	Ansiedad moderada
13	5	16,6	12	1	6	5,3	0	2	7,7	4	3	5	1	2	2	1	2	1	2	2	Ansiedad moderada
14	3	13,5	35	0	8	8	7	1	9,1	1	2	2	0	0	1	1	1	1	1	1	Ansiedad leve
15	6	12,9	15	1	18	4,2	5	2	8,4	1	3	10	1	2	1	2	2	2	2	2	Ansiedad moderada
16	5	14,4	11	0	12	9	0	2	9,9	8	5	4	0	2	0	0	1	1	1	0	Ansiedad leve
17	3	12,5	6	1	21	4,7	6	2	2,7	4	10	3	0	2	2	2	2	2	2	1	Ansiedad moderada
18	6	14,5	10	0	7	8	4	3	1,8	4	6	1	0	1	1	0	1	1	1	1	Ansiedad leve
19	6	14,2	36	1	21	6	6	0	4	6	6	6	0	1	1	2	1	1	1	1	Ansiedad leve
20	8	16,6	24	1	17	6,5	3	1	2,2	5	3	4	0	1	1	1	3	2	2	1	Ansiedad moderada
21	2	17,5	38	1	28	8,4	0	0	5,7	1	7	6	0	1	1	2	1	1	0	1	Ansiedad leve
22	1	17,5	20	0	22	6,1	1	5	6,7	10	6	8	0	2	2	2	2	1	1	1	Ansiedad moderada
23	4	10	29	0	1	6,7	4	5	6,2	0	1	3	0	0	2	1	2	1	2	1	Ansiedad leve
24	7	12,6	16	0	17	4,2	5	5	3,3	4	7	4	0	1	1	0	1	1	2	1	Ansiedad leve
25	10	10	11	1	17	4	2	0	9,2	3	8	8	0	3	1	2	2	2	2	2	Ansiedad moderada
26	3	12,1	21	0	10	5,2	3	4	1,2	2	4	9	1	2	1	3	2	2	1	1	Ansiedad moderada
27	3	13,2	15	1	30	7,4	3	3	4,1	1	3	10	0	2	2	1	1	1	2	1	Ansiedad moderada
28	5	15,3	10	0	0	4,6	1	4	7,5	9	3	3	0	1	1	2	2	2	2	2	Ansiedad moderada
29	5	13,5	27	0	13	4,6	2	5	7,9	1	3	4	1	1	3	2	2	2	3	1	Ansiedad moderada
30	10	13,5	31	1	15	6,1	7	3	4,7	3	5	4	0	2	2	2	2	2	1	1	Ansiedad moderada
31	3	14,9	10	0	10	7,2	2	0	1,2	1	10	7	0	2	2	2	3	3	3	2	Ansiedad severa
32	9	11,1	6	1	0	7,6	5	5	6,6	3	3	3	1	2	2	1	2	1	1	1	Ansiedad moderada
33	4	12,8	26	1	29	4,9	5	0	1,6	2	6	3	1	1	2	1	1	1	2	2	Ansiedad moderada
34	3	13,5	24	0	6	4,4	4	3	2	8	6	6	0	2	2	1	3	1	1	2	Ansiedad moderada
35	2	14,1	29	0	30	8,7	7	0	6,9	6	4	10	0	2	2	3	1	2	1	2	Ansiedad moderada
36	10	14,7	19	0	28	5,3	3	0	3,8	10	4	4	1	2	1	2	2	2	2	2	Ansiedad moderada
37	9	17,5	33	0	15	8,3	0	1	4,1	2	9	5	0	0	1	2	0	0	1	0	Sin ansiedad
38	4	16,3	12	1	29	8,3	5	4	5,4	7	7	5	0	0	2	1	1	1	1	1	Ansiedad leve
39	3	13	37	0	13	7,9	1	4	9,8	8	6	5	0	1	0	1	1	0	1	0	Sin ansiedad
40	7	14	26	0	14	4,8	1	4	9,7	6	6	9	0	2	1	1	1	2	1	1	Ansiedad leve
41	8	13,8	9	0	13	8,1	1	2	4,7	5	5	4	0	1	1	1	0	1	1	2	Ansiedad leve
42	2	13,5	22	0	27	7,8	6	1	3,4	3	6	4	1	0	1	0	0	1	1	0	Sin ansiedad
43	7	15	34	0	20	7	7	5	3,7	4	6	8	0	2	2	1	1	2	1	2	Ansiedad moderada
44	5	10,8	35	1	10	5,7	7	5	3,4	7	9	7	0	1	2	1	2	2	1	1	Ansiedad moderada
45	5	16,3	14	1	15	5,3	5	5	5,8	2	1	3	0	2	1	2	1	1	1	1	Ansiedad leve
46	3	17,9	19	1	18	4,6	5	4	6,5	1	9	6	0	1	1	2	1	1	1	2	Ansiedad leve
47	5	14	22	1	16	6,5	1	4	5,4	2	1	7	0	1	1	2	0	2	2	1	Ansiedad leve
48	2	13,9	27	0	26	6,9	7	1	4,7	6	9	8	1	2	1	1	1	1	1	1	Ansiedad leve
49	4	14,8	37	1	25	5,8	7	0	2,9	7	6	4	1	1	1	1	2	1	1	1	Ansiedad leve
50	6	16,5	20	0	5	6,4	5	4	5,3	0	9	2	0	1	0	0	0	1	0	0	Sin ansiedad
51	1	13,9	39	0	10	5,8	6	5	4,1	6	6	3	0	1	1	1	1	2	2	1	Ansiedad leve
52	10	16,4	38	0	16	7,2	0	0	5,9	8	6	9	0	1	1	1	1	1	2	1	Ansiedad leve
53	4	15,5	32	0	8	8,5	0	4	7,1	2	4	5	1	1	2	2	0	1	1	1	Ansiedad leve
54	6	19,2	6	1	17	4,8	4	3	6,7	2	0	9	0	2	2	2	2	2	2	2	Ansiedad moderada
55	9	13,8	34	0	16	5,9	7	2	5,2	7	1	10	1	2	3	3	2	1	2	1	Ansiedad moderada
56	2	10,5	17	1	30	4,7	3	4	6	3	8	4	0	1	2	2	2	2	2	2	Ansiedad moderada
57	7	11,6	20	0	11	7,1	0	3	3,6	7	7	8	0	2	2	2	2	1	1	2	Ansiedad moderada
58	6	14,5	9	0	22	7,6	4	1	1,2	4	7	10	0	2	2	2	2	2	2	2	Ansiedad moderada
59	2	16,6	32	1	21	8,7	2	3	6,6	2	9	10	0	0	1	1	1	1	2	1	Ansiedad leve
60	8	13,2	24	0	21	5,4	3	1	3,3	3	0	2	1	1	1	1	1	1	2	1	Ansiedad leve
61	7	14	28	0	29	8,8	7	4	7	3	5	1	0	2	1	1	2	1	1	1	Ansiedad leve
62	5	15,9	31	0	4	4,3	0	1	7,7	7	8	5	0	2	1	1	2	3	2	2	Ansiedad moderada
63	5	12,6	34	1	10	7	6	3	9,2	5	7	6	0	1	1	1	1	2	1	2	Ansiedad leve
64	7	13,3	31	1	3	4,3	2	2	9,6	6	2	4	0	1	2	2	2	2	2	2	Ansiedad moderada
65	9	12,2	30	0	26	8,5	4	0	1,1	8	3	0	0	0	2	0	0	1	1	1	Ansiedad leve
66	9	16,9	11	1	3	6,6	0	0	9,9	5	2	2	0	1	1	1	2	1	1	1	Ansiedad leve
67	9	17,6	9	1	2	7,3	7	2	7,1	2	5	2	0	2	2	2	1	1	1	1	Ansiedad moderada
68	6	11,3	27	1	25	8,2	0	5	6,8	2	8	3	0	1	0	1	1	0	0	1	Sin ansiedad
69	10	14,2	32	1	23	4,8	5	0	1,1	10	7	9	0	2	3	3	2	2	3	2	Ansiedad severa
70	6	17,3	38	0	22	4,2	2	4	8,8	2	9	5	0	2	2	2	2	3	2	2	Ansiedad severa
71	3	15,3	19	1	24	8,4	0	5	2,3	9	3	4	0	1	0	0	0	1	1	1	Sin ansiedad
72	1	13,5	36	0	16	5,6	1	4	8,5	1	1	0	0	2	2	2	3	2	1	1	Ansiedad moderada
73	7	11,4	30	0	6	4,6	6	3	1,1	0	7	5	0	2	2	1	2	2	1	2	Ansiedad moderada
74	4	16,8	35	1	25	6,3	5	3	6,8	10	7	8	0	2	1	1	1	1	1	1	Ansiedad leve
75	1	15,2	23	1	3	7	5	5	4,9	9	10	10	0	1	1	1	2	1	2	2	Ansiedad moderada
76	8	11,6	13	0	22	4,2	0	5	2,8	0	5	0	0	2	2	2	3	2	2	3	Ansiedad severa
77	6	16,9	33	0	12	5,4	6	5	2	3	7	7	0	1	2	2	2	1	1	1	Ansiedad moderada
78	6	11,3	8	1	11	4,2	5	5	5	2	8	7	0	1	2	2	2	2	1	1	Ansiedad moderada
79	2	14,1	7	0	0	6,1	3	1	2,1	4	0	2	0	2	3	2	1	1	1	2	Ansiedad moderada
80	1	14	33	0	20	8,5	2	5	5,4	8	8	8	1	1	1	1	1	2	0	1	Ansiedad leve
81	2	17,7	15	0	18	8,9	3	5	6,6	7	1	10	0	2	2	2	1	1	2	0	Ansiedad moderada
82	4	14	37	0	20	7,2	4	3	8,4	8	0	10	0	1	0	1	1	2	1	1	Ansiedad leve
83	2	14	15	1	19	4,8	3	4	7,1	4	7	7	0	3	3	2	3	2	3	2	Ansiedad severa
84	1	11,8	31	0	17	8,5	6	3	1,6	7	7	2	0	0	1	1	0	1	1	0	Sin ansiedad
85	7	13,2	37	1	10	6,6	1	0	3,8	3	10	9	0	0	1	2	2	1	2	2	Ansiedad moderada
86	8	12,9	25	0	19	4,8	5	1	3,6	4	3	4	1	2	2	2	1	2	1	2	Ansiedad moderada
87	6	13	35	1	2	8,6	6	1	9,2	2	7	2	0	1	1	0	1	1	1	1	Ansiedad leve
88	4	14,5	6	1	28	7,8	5	4	8,8	4	6	4	0	1	1	0	0	1	1	0	Sin ansiedad
89	6	9,6	31	1	16	7,7	7	5	4,3	2	1	8	0	0	1	1	2	1	1	1	Ansiedad leve
90	3	11,5	40	1	11	7,5	1	3	1,6	4	4	3	0	2	1	1	1	1	1	1	Ansiedad leve
91	8	14,3	15	1	9	5	5	0	9	10	9	5	0	2	2	2	2	1	2	2	Ansiedad moderada
92	10	14,7	22	0	7	5,1	3	3	9	4	3	2	1	2	3	2	2	3	3	3	Ansiedad severa
93	2	14	5	1	14	7,4	7	5	8,7	5	2	2	1	2	1	1	1	1	1	1	Ansiedad leve
94	5	14,4	39	1	4	5,1	7	5	8,6	8	0	4	0	2	1	2	2	1	1	1	Ansiedad moderada
95	1	13,7	39	0	3	6,2	4	2	5,3	0	4	7	1	2	1	1	1	1	1	2	Ansiedad leve

ANEXO 6: Matriz de consistencia

Tema: Predicción de la ansiedad generalizada en estudiantes universitarios mediante Random Forest basado en el cuestionario Generalized Anxiety Disorder 7-item (GAD-7), 2025.

Problema de investigación	Objetivo de investigación	Hipótesis	Variables	Indicadores	Diseño metodológico
Problema general: ¿Cómo predecir el riesgo de ansiedad generalizada en estudiantes universitarios utilizando el algoritmo Random Forest, basado en los resultados del cuestionario GAD-7? **Problema específicos** ¿Qué variables influyen más en la predicción del riesgo de ansiedad generalizada en estudiantes universitarios según el cuestionario GAD-7?	**Objetivo general:** Desarrollar un modelo predictivo basado en el algoritmo Random Forest para identificar el riesgo de ansiedad generalizada en estudiantes universitarios mediante el cuestionario GAD-7. **Objetivos específicos** Identificar las variables que influyen en la predicción del riesgo de ansiedad generalizada en estudiantes universitarios utilizando el cuestionario GAD-7.	No se formulará hipótesis ya que el enfoque es predictivo, basado en la construcción y evaluación del modelo Random Forest para predecir ansiedad generalizada.	**Variable dependiente:** Ansiedad Generalizada. **Dimensiones:** Síntomas físicos, síntomas emocionales, impacto en la vida diaria. **Variable independiente:** Modelo Random Forest. **Dimensiones:** Precisión,	**Indicadores VD:** Puntuación del GAD-7 (niveles de síntomas físicos y emocionales, interferencia en la vida diaria). **Indicadores VI:** Matriz de confusión, área bajo la curva ROC, puntuación F1 del modelo Random Forest.	**Población:** Estudiantes de ingeniería de la UNSCH. **Muestra:** 166 estudiantes. **Fórmula de la muestra:** $n = \frac{N \cdot Z^2 \cdot p \cdot (1-p)}{(E^2 \cdot (N-1)) + (Z^2 \cdot p \cdot (1-p))}$ **Instrumentos:** Cuestionario GAD-7, software Python con Scikit-learn para implementar el modelo Random Forest. **Tipo de Investigación:**

		sensibilidad, especificidad.	Aplicada **Diseño de Investigación:** No experimental, transversal y predictivo **Nivel de Investigación:** Explicativo
¿Cómo diseñar un modelo predictivo basado en Random Forest que permita identificar el riesgo de ansiedad generalizada en estudiantes universitarios?	Diseñar un modelo predictivo basado en Random Forest para la predicción del riesgo de ansiedad generalizada en estudiantes universitarios.		
¿Qué nivel de precisión, sensibilidad y especificidad tiene el modelo predictivo Random Forest en la predicción de la ansiedad generalizada en estudiantes universitarios?	Evaluar el rendimiento del modelo predictivo en términos de precisión, sensibilidad y especificidad en la predicción de ansiedad generalizada en estudiantes universitarios.		

UNSCH FACULTAD DE INGENIERÍA DE MINAS, GEOLOGÍA Y CIVIL

ACTA DE SUSTENTACIÓN DE TESIS Nº 07-2026-FIMGC

PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO DE SISTEMAS

En la Universidad Nacional de San Cristóbal de Huamanga, en la ciudad de Ayacucho, en cumplimiento a la **Resolución Decanal No 032-2026-FIMGC-D**, a los **veintisiete días del mes de marzo de 2026**, siendo las **10:00 a.m.**, reunidos en el **Auditorio de la Escuela Profesional de Ingeniería de Minas**, bajo la presidencia del **MSc. Ing. José Ernesto ESTRADA CÁRDENAS**, y los miembros: **Mg. Juan Carlos CARREÑO GAMARRA, Dr. Hubner JANAMPA PATILLA** y **Mg. Eden Jersson TERRAZA HUAMAN**, actuando como secretario docente el **Ing. Saul Walter RETAMOZO FERNANDEZ**, para proceder a la sustentación de tesis para optar el **Título Profesional de Ingeniero de Sistemas**, del Bachiller en Ingeniería de Sistemas:

Irvin Pelayo TAYPE PANIAGUA

Quien presentó la tesis denominada:

Predicción de la ansiedad generalizada en estudiantes universitarios mediante Random Forest basado en el cuestionario Generalized Anxiety Disorder 7-item (GAD-7), 2025.

Los señores miembros del jurado luego de expuesta la tesis y absueltas las preguntas, deliberaron y declararon:

Aprobado con 15 (Quince)

Siendo las **12:04 p.m.** del día **27 de marzo del 2026**, culmina el acto de sustentación de tesis, y en conformidad de lo actuado los miembros del jurado firmamos al pie del presente.

MSc. Ing. José Ernesto ESTRADA CÁRDENAS
Presidente

Mg. Juan Carlos CARREÑO GAMARRA
Miembro

Mg. Eden Jersson TERRAZA HUAMAN
Miembro

Dr. Hubner JANAMPA PATILLA
Miembro - Asesor

Ing. Saul Walter RETAMOZO FERNANDEZ
Secretario docente de la FIMGC

FACULTAD DE INGENIERÍA
DE MINAS Y CIVIL
Av. Independencia S/N
Ciudad Universitaria
Central Tel. 066 312510
Anexo 151

UNSCH FACULTAD DE INGENIERÍA DE MINAS, GEOLOGÍA Y CIVIL

turnitin

CONSTANCIA DE ORIGINALIDAD DE TRABAJO DE INVESTIGACIÓN

CONSTANCIA Nº 036-2026-KPS-FIMGC/UNSCH

El que suscribe; responsable verificador de originalidad de trabajos de tesis de pregrado con el software Turnitin, en segunda instancia para las **Escuelas Profesionales** de la **Facultad de Ingeniería de Minas, Geología y Civil**; en cumplimiento a la **Resolución de Consejo Universitario Nº 039-2021-UNSCH-CU**, Reglamento de Originalidad de Trabajos de Investigación de la Universidad Nacional San Cristóbal de Huamanga y **Resolución Decanal Nº 697-2024-FIMGC-D**, deja constancia de originalidad de trabajo de investigación, que el/la Sr./Srta.

Nombres y Apellidos	:	Irvin Pelayo Taype Paniagua
Escuela Profesional	:	INGENIERÍA DE SISTEMAS
Título de la Tesis	:	Predicción de la Ansiedad Generalizada en Estudiantes Universitarios mediante Random Forest basado en el Cuestionario Generalized Anxiety Disorder 7-item (GAD-7), 2025.
Evaluación de la Originalidad	:	**6%** Índice de Similitud
Identificador de la entrega	:	2945966440

Por tanto, según los **Artículos 12, 13 y 17** del Reglamento de Originalidad de Trabajos de Investigación, es **PROCEDENTE** otorgar la **Constancia de Originalidad** para los fines que crea conveniente.

En señal de conformidad y verificación se firma la presente constancia

Ayacucho, 30 de abril de 2026

FIRMA DIGITAL

Firmado digitalmente por: PERALTA SOTOMAYOR Karel FAU 20143660754 soft
Motivo: Soy el autor del documento
Fecha: 30/04/2026 07:36:49-0500

Con depósito para Sustentación y Tramites
C.c. Archivo

FACULTAD DE INGENIERIA DE MINAS, GEOLOGIA Y CIVIL
Av. Independencia S/N Ciudad Universitaria
Central Tel. 066 312510
Anexo 151

Predicción de la Ansiedad Generalizada en Estudiantes Universitarios mediante Random Forest basado en el Cuestionario Generalized Anxiety Disorder 7-item (GAD-7), 2025.

por Irvin Pelayo Taype Paniagua

Fecha de entrega: 27-abr-2026 05:58p. m. (UTC-0500)

Identificador de la entrega: 2945966440

Nombre del archivo: MEMORANDO_Nº_235-2026-FIMGC-UNSCH-CERTIFICADO_DE_ORIGINALIDAD-IRVIN_PELAYO_TAYPE_PANIAGUA.pdf (8.37M)

Total de palabras: 26031

Total de caracteres: 155368

Predicción de la Ansiedad Generalizada en Estudiantes Universitarios mediante Random Forest basado en el Cuestionario Generalized Anxiety Disorder 7-item (GAD-7), 2025.

INFORME DE ORIGINALIDAD

6%	5%	2%	3%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	Submitted to Universidad Nacional de San Cristóbal de Huamanga Trabajo del estudiante	2%
2	www.researchgate.net Fuente de Internet	1%
3	hdl.handle.net Fuente de Internet	<1%
4	Submitted to Universidad Internacional de la Rioja Trabajo del estudiante	<1%
5	repositorio.unsaac.edu.pe Fuente de Internet	<1%
6	repositorio.ucsg.edu.ec Fuente de Internet	<1%
7	repositorio.uide.edu.ec Fuente de Internet	<1%
8	repositorio.uwiener.edu.pe Fuente de Internet	<1%
9	oai.e-spacio.uned.es Fuente de Internet	<1%
10	Rojas Atachao, Richard Frank. "Aplicación de un modelo de Deep Learning con redes neuronales convolucionales para evaluar la integridad estructural y predecir fallas en puentes de concreto mediante exploración visual de imágenes", Pontificia Universidad Catolica del Peru (Peru), 2025 Publicación	<1%
11	Submitted to Universidad Francisco de Vitoria	

	Trabajo del estudiante	<1 %
12	repositorioinstitucional.buap.mx Fuente de Internet	<1 %
13	riunet.upv.es Fuente de Internet	<1 %
14	www.coursehero.com Fuente de Internet	<1 %
15	riujap.ujap.edu.ve Fuente de Internet	<1 %
16	Submitted to consultoriadeserviciosformativos Trabajo del estudiante	<1 %
17	repositorio.unjbg.edu.pe Fuente de Internet	<1 %
18	Submitted to Aliat Universidades Trabajo del estudiante	<1 %
19	es.slideshare.net Fuente de Internet	<1 %

Excluir citas	Activo	Excluir coincidencias	< 30 words
Excluir bibliografía	Activo		