

## Diseño de un dataset multidimensional para el análisis de factores asociados a la ansiedad en estudiantes universitarios mexicanos

Edgar Jahir Hernández-Andrade<sup>1</sup>, Mónica Ruiz-Martínez<sup>1</sup>,  
Maritza Bustos-López<sup>2</sup>, Alfonso Flores-Leal<sup>1</sup>, Ignacio López-Martínez<sup>1</sup>

<sup>1</sup> Instituto Tecnológico de Orizaba/Tecnológico Nacional de México, Orizaba,  
México

<sup>2</sup> Instituto Tecnológico de Orizaba,  
Secretaría de Ciencia, Humanidades, Tecnología e Innovación, Orizaba,  
México

{m24010203, monica.rm, maritza.bl, alfonso.fl,  
ignacio.lm} @orizaba.tecnm.com.mx

**Resumen.** Los trastornos de ansiedad constituyen uno de los problemas de salud mental más prevalentes en estudiantes universitarios, con una incidencia del 19.3% en México. A pesar de ello, no existe en la literatura un dataset que integre simultáneamente variables demográficas, académicas, socioeconómicas junto con una medición psicométrica validada, orientado al contexto universitario mexicano. Este trabajo presenta el diseño de un esquema de dataset multidimensional que combina la subescala de ansiedad del instrumento DASS-21 con cuatro categorías de variables de contexto, materializado a través de AnxiTech: un sistema web desarrollado con Vue.js, PHP y Python-FastAPI. Para validar técnicamente la viabilidad del esquema, se empleó un dataset híbrido de 500 registros y se entrenó un modelo Random Forest con análisis de *feature importance*, confirmando la viabilidad técnica del esquema. Adicionalmente, se realizó un estudio piloto con 42 estudiantes reales de una universidad pública mexicana, recolectando respuestas del DASS-21 junto con el conjunto completo de variables contextuales. Los resultados del piloto revelaron una redistribución significativa de la importancia de variables: si bien el promedio académico se mantuvo como factor principal (25.2%), variables como la edad (18.2%) y la situación laboral (13.8%) adquirieron mayor relevancia que en los datos sintéticos, lo que sugiere que los datos empíricos capturan patrones más complejos. La principal contribución es el esquema multidimensional y el sistema de recolección como base para la generación de un dataset empírico institucional en México.

**Palabras clave:** Análisis predictivo de datos, ciencia de datos, dataset, estudiantes universitarios, factores que identifican niveles de ansiedad, *feature importance*, machine learning, random forest, trastorno de ansiedad.

## Design of a Multidimensional Dataset for the Analysis of Factors Associated with Anxiety in Mexican University Students

**Abstract.** Anxiety disorders are among the most prevalent mental health issues among university students, with an incidence rate of 19.3% in Mexico. Despite this, the literature lacks a dataset that simultaneously integrates demographic, academic, and socioeconomic variables with a validated psychometric measure tailored to the Mexican university context. This paper presents the design of a multidimensional dataset schema combining the anxiety subscale of the DASS-21 instrument with four categories of contextual variables, implemented via AnxiTech, a web-based system developed using Vue.js, PHP, and Python-FastAPI. To technically validate the schema's feasibility, a hybrid dataset of 500 records was used to train a Random Forest model with feature importance analysis, confirming the schema's technical viability. Additionally, a pilot study was conducted with 42 actual students from a Mexican public university, collecting DASS-21 responses alongside the full set of contextual variables. Pilot results revealed a significant redistribution of variable importance: while academic average remained the primary factor (25.2%), variables such as age (18.2%) and employment status (13.8%) gained greater relevance compared to the synthetic data, suggesting that empirical data capture more complex patterns. The primary contribution is the multidimensional schema and data collection system, serving as a foundation for generating an institutional empirical dataset in Mexico.

**Keywords:** Predictive data analysis, data science, dataset, university students, factors identifying anxiety levels, feature importance, machine learning, random forest, anxiety disorder.

### 1. Introducción

El trastorno de ansiedad en estudiantes universitarios es un problema de salud mental prevalente que impacta negativamente el rendimiento académico, incrementa las tasas de deserción y compromete significativamente el bienestar integral del alumnado. La ansiedad se manifiesta mediante un conjunto de síntomas físicos como taquicardia, sudoración y mareos, acompañados de síntomas psíquicos que incluyen nerviosismo, inquietud, preocupación excesiva y alteraciones del sueño [1, 2]. Estos síntomas tienden a producir distorsiones en la percepción de la realidad y el entorno, afectando directamente la funcionalidad del individuo en el contexto académico y social.

Las estadísticas globales indican que aproximadamente el 4% de la población mundial experimenta trastornos de ansiedad, mientras que en México según datos de la Secretaría de Salud la situación es particularmente alarmante, con un 19.3% de la población padeciendo ansiedad severa y un 30% presentando ansiedad moderada [3]. En el contexto universitario mexicano, donde solo el 22% de los jóvenes accede a educación superior, resulta fundamental ofrecer herramientas efectivas para la identificación temprana de factores de riesgo asociados a este trastorno.

Este fenómeno ha sido tradicionalmente abordado desde disciplinas como la psicología clínica y las ciencias de la educación mediante instrumentos diagnósticos estandarizados como el Inventario de Ansiedad de Beck, la Escala de Ansiedad de Hamilton y la escala DASS-21 [4, 5].

Sin embargo, estos enfoques presentan limitaciones significativas al evaluar de manera aislada las diversas dimensiones del problema, sin capturar la naturaleza multifactorial y compleja de la ansiedad en el entorno universitario.

La literatura científica actual evidencia una brecha importante en el desarrollo de herramientas integrales para la identificación efectiva de factores de riesgo asociados a los niveles de ansiedad en estudiantes universitarios en México. Los métodos actuales, basados principalmente en cuestionarios y encuestas, abordan aspectos aislados del fenómeno y no permiten establecer correlaciones significativas. Los estudios existentes abordan el fenómeno de manera parcial: algunos analizan variables académicas de forma aislada [6], otros se centran en factores demográficos o psicosociales [7], y los enfoques basados en aprendizaje automático suelen trabajar con señales fisiológicas o texto [8, 9]. Ninguno integra todas estas dimensiones en un esquema de dataset estructurado para el contexto universitario mexicano.

Este trabajo presenta el diseño de un esquema de dataset multidimensional que combina la escala DASS-21 con variables demográficas, académicas y socioeconómicas, así como su implementación a través de AnxiTech, un sistema web que actúa como instrumento digital de recolección. La validación técnica del esquema se realizó en dos fases: primero, mediante un dataset híbrido de 500 registros y un modelo Random Forest con análisis de Feature Importance; y segundo, mediante un estudio piloto con 42 estudiantes reales del Instituto Tecnológico de Orizaba, cuyos resultados evidenciaron una redistribución significativa en la importancia de variables.

El presente artículo se estructura de la siguiente manera: en la sección 2 se presentan los trabajos relacionados; la sección 3 describe el diseño del esquema multidimensional; la sección 4 presenta el sistema AnxiTech; la sección 5 expone la validación técnica y el estudio piloto con datos reales; y la sección 6 presenta las conclusiones.

## **2. Trabajos relacionados**

La investigación sobre ansiedad universitaria y el uso de técnicas computacionales en salud mental estudiantil se agrupa en tres líneas temáticas, cada una con aportaciones relevantes, pero con brechas específicas que motivan el presente trabajo.

### **2.1. Factores de ansiedad en estudiantes universitarios**

Esta línea aborda la identificación de variables asociadas a la ansiedad mediante instrumentos psicométricos y análisis estadístico.

**Tabla 1.** Comparación de trabajos previos por cobertura de variables.

| <b>Característica</b>                  | <b>Chavez<br/>Marquez</b> | <b>Morales<br/>et. al</b> | <b>Farrer et. al</b> | <b>Mental Health<br/>Dass-21</b> | <b>Anxitech</b> |
|----------------------------------------|---------------------------|---------------------------|----------------------|----------------------------------|-----------------|
| Instrumento<br>psicométrico            | IDARE                     | DASS-21                   | Encuesta<br>propia   | DASS-21                          | DASS-21         |
| Variables<br>demográficas              | Parcial                   | Parcial                   | Si                   | Parcial                          | Si              |
| Variables<br>académicas.               | Si                        | Si                        | No                   | Parcial                          | Si              |
| Variables<br>socioeconómicas           | No                        | No                        | Parcial              | No                               | Si              |
| Contexto<br>universitario<br>mexicano. | Si                        | Si                        | No                   | No                               | Si              |
| Dataset /<br>Sistema<br>reutilizable   | No                        | No                        | No                   | Si                               | Si              |

Chávez-Márquez y Del Río [10] aplicaron el instrumento IDARE (Inventario de Ansiedad Rasgo-Estado) en 1,006 estudiantes mediante muestreo aleatorio estratificado, identificando el área de estudio como el único factor significativamente relacionado; su análisis, sin embargo, se enfocó en variables académicas sin dimensiones socioeconómicas ni demográficas.

Morales et al. [6] aplicó el instrumento DASS-21 a 186 estudiantes y encontraron asociación estadísticamente significativa entre ansiedad y rendimiento académico, aunque tampoco incorporaron variables de contexto adicionales. Farrer et al. [7] ampliaron el espectro al identificar predictores demográficos y psicosociales de ansiedad generalizada en universitarios australianos, incluyendo género, movilidad y situación financiera, mediante regresión logística, siendo el trabajo con mayor cobertura de variables entre los revisados. En común, estos estudios abordan dimensiones aisladas del fenómeno y ninguno está orientado al contexto universitario mexicano.

## 2.2. Sistemas digitales y Machine Learning para la salud mental estudiantil

Una segunda línea integra tecnología digital con instrumentos psicométricos para la detección o predicción de ansiedad. Lara Flores [11] desarrolló una plataforma para detectar ansiedad mediante señales fisiológicas complementadas con el IDARE en 21 estudiantes, demostrando la viabilidad de integrar biometría con autorreporte. Flores Moscoso [12] propuso un sistema predictivo de estrés académico con aprendizaje automático bajo la metodología CRISP-DM, logrando precisión significativa en la predicción de niveles de estrés a partir de carga académica y calidad del sueño.

Drira et al. [13] realizaron una revisión sistemática de ML en salud mental estudiantil, evidenciando que el 65% de los estudios analizados ignora principios de privacidad al manejar datos sensibles, lo que subraya la necesidad de esquemas de recolección éticamente diseñados. Estos trabajos demuestran la pertinencia técnica de aplicar ML en este dominio, pero ninguno diseña un esquema de variables multidimensional explícito como contribución principal.

**Tabla 2.** Esquema de variables por dimensión.

| <b>Dimensión</b> | <b>Variables</b>                                                    | <b>Justificación</b>                                  |
|------------------|---------------------------------------------------------------------|-------------------------------------------------------|
| Demográfica      | Edad, género, estado civil                                          | Perfil base del estudiante                            |
| Académica        | Promedio, semestre, materias inscritas, carrera, beca               | Carga y rendimiento académico como predictores claves |
| Socioeconómica   | Situación laboral, tipo de vivienda, Medio de transporte            | Factores de estrés extracurricular                    |
| Psicométrica     | DASS-21 -subescala de ansiedad (nivel de ansiedad: Bajo/Medio/Alto) | Instrumento validado internacionalmente               |

### 2.3. Dataset existentes y brecha del contexto mexicano

La tercera línea, y la más directamente relacionada con la contribución de este trabajo, corresponde a los Datasets disponibles para el análisis de salud mental universitaria. El Dataset llamado “Mental Health Dataset based on DASS-21” [14] es el referente más cercano: integra variables de ansiedad, depresión y estrés junto con datos demográficos y académicos. No obstante, su construcción no contempla variables de contexto local como medio de transporte, si el estudiante tiene beca o no, ni situación de vivienda, además de que no está enfocado al contexto universitario.

En la literatura en español no se identificó ningún dataset publicado que combine un instrumento psicométrico validado con variables contextuales multidimensionales para universitarios en México. Esta brecha constituye la motivación central del presente trabajo: diseñar un esquema que capture la naturaleza multifactorial de la ansiedad en su contexto local específico, y un sistema que lo materialice como instrumento de recolección estandarizado.

## 3. Diseño del esquema multidimensional

Esta sección presenta la contribución central de este trabajo: el esquema del dataset enfocado en el contexto universitario mexicano que integra cuatro dimensiones de análisis sobre el mismo individuo, superando el enfoque poco dimensional predominante en la literatura.

Además, se describen los criterios que guiaron el diseño, así como las dimensiones y variables que conforman el esquema.

### 3.1. Criterios de diseño

El esquema se diseñó bajo tres criterios rectores:

- Cobertura multidimensional: que garantice la representación de factores de distintos dominios que la literatura identifica como asociados a la ansiedad.

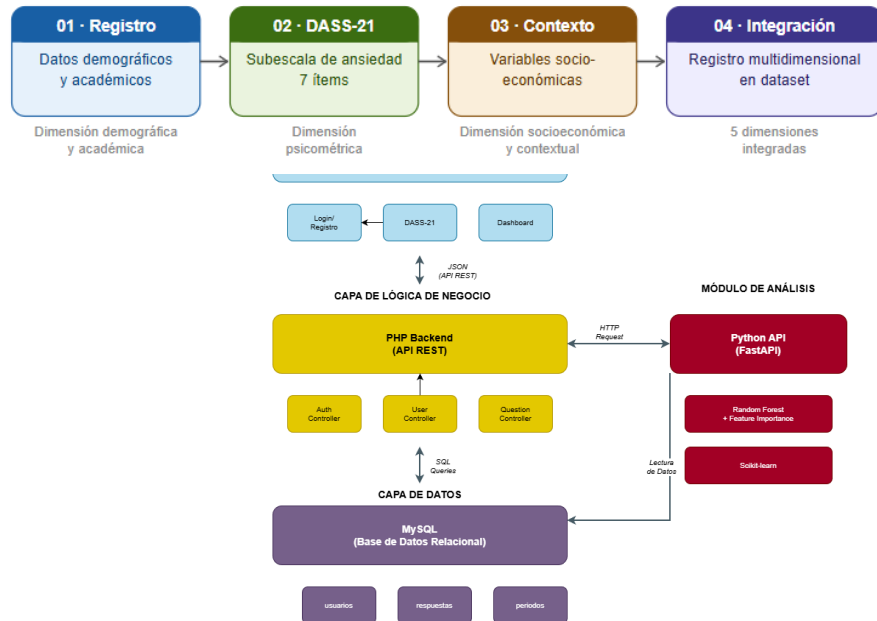


Fig. 1. Arquitectura del sistema: Anxitech.

- Validez psicométrica: preservando el instrumento DASS-21 sin modificaciones para mantener su integridad científica.
- Pertinencia contextual: incluyendo variables relevantes para el entorno universitario mexicano no contempladas en datasets internacionales, como la situación laboral del estudiante, el medio de transporte utilizado, la situación de vivienda y el acceso a becas académicas, factores que reflejan las condiciones socioeconómicas específicas del estudiante universitario en México.

### 3.2. Variables y dimensiones del esquema

La tabla 2 presenta el esquema completo de variables por dimensión, con su justificación basada en la literatura

La dimensión académica concentra el mayor número de variables dado que la literatura consistentemente la señala como el dominio con mayor asociación a la ansiedad universitaria [10]. La dimensión socioeconómica incorpora variables específicas del entorno mexicano que no aparecen en datasets internacionales como el “Mental Health Dataset based on DASS-21” [14], lo que constituye una aportación diferenciadora del esquema.

La variable objetivo se deriva del puntaje bruto de la subescala de ansiedad del instrumento DASS-21 mediante la clasificación oficial del instrumento para la subescala de ansiedad, garantizando la comparabilidad con estudios previos que utilizan el mismo instrumento.

## **4. AnxiTech (Sistema de recolección)**

AnxiTech es el sistema web que materializa el esquema multidimensional como instrumento digital de recolección estandarizado. Su desarrollo se fundamentó en la metodología Extreme Programming (XP) con cuatro fases iterativas: Planificación, Diseño, Codificación y Pruebas.

### **4.1. Arquitectura del sistema**

El sistema implementa una arquitectura de cuatro capas:

- Capa de Presentación: Desarrollada con Vue.js 3 y Vuetify, que incluye módulos para el registro de estudiantes, aplicación del DASS-21 y captura de variables contextuales, así como un panel administrativo con visualización de la información obtenida desde el modelo.
- Capa de Aplicación: Implementada en PHP con patrón MVC y endpoints RESTful para la lógica de negocio.
- Capa de Machine Learning: Un microservicio Python-FastAPI que ejecuta el entrenamiento del modelo Random Forest y calcula la importancia de características utilizando la tecnología de Feature Importance.
- Capa de Datos: Una base de datos MySQL normalizada que almacena respuestas al DASS-21, variables contextuales y resultados del análisis.

La Figura 1 ilustra la arquitectura del sistema y la comunicación entre capas. El backend PHP actúa como intermediario entre la interfaz de usuario, el microservicio Python-FastAPI y la base de datos, enviando los datos recolectados para su análisis, almacenándolos y recibiendo los resultados de Feature Importance en formato JSON.

### **4.2. Flujo de recolección**

Como se ilustra en la Figura 2, el flujo de recolección sigue cuatro pasos secuenciales: primero, el estudiante se registra proporcionando sus datos personales y demográficos; posteriormente responde la subescala de ansiedad del DASS-21 en su formato original sin modificaciones; a continuación, completa la información complementaria con sus datos académicos y socioeconómicos; y finalmente el sistema integra las cuatro dimensiones en un único registro del dataset. Este diseño garantiza la integridad del instrumento psicométrico al tiempo que amplía el contexto de análisis.

### **4.3. Diseño de bases de datos**

Como se observa en la Figura 3, el esquema multidimensional se materializa en la base de datos mediante nueve tablas especializadas organizadas por dimensión. La gestión de usuarios se estructura en tres tablas: usuarios almacena credenciales y preferencias del sistema, estudiantes concentra los datos demográficos del alumno — nombre, fecha de nacimiento, sexo, estado civil y número de control— y administradores gestiona los permisos de acceso institucional.

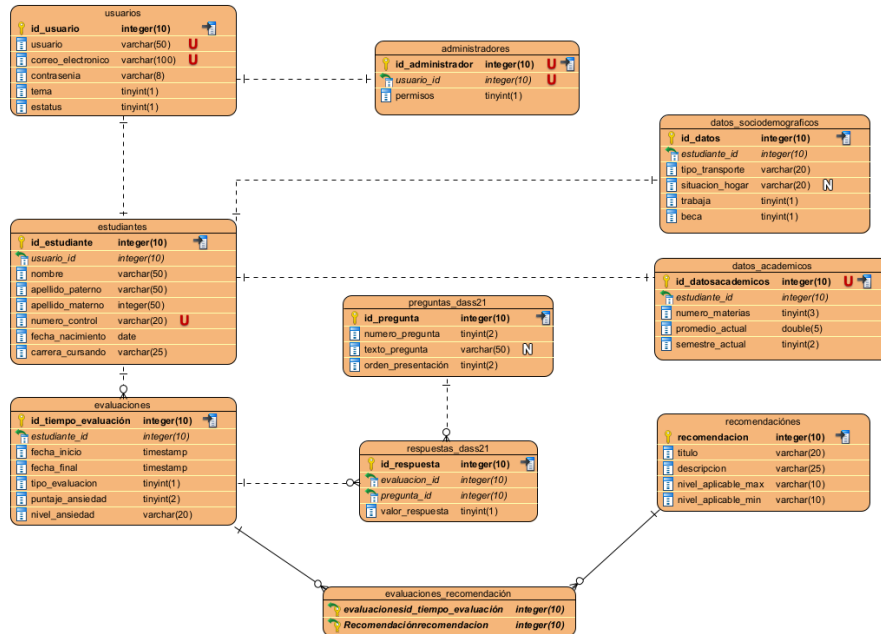


Fig. 3. Diseño de base de datos: Anxitech.

Las dimensiones del esquema se distribuyen en tablas independientes: datos\_academicos registra promedio, semestre actual y número de materias inscritas; datos\_sociodemograficos captura situación laboral, tipo de vivienda, medio de transporte y beca. El instrumento psicométrico se implementa mediante preguntas\_dass21, que almacena los 7 ítems de la subescala de ansiedad, y respuestas\_dass21, que registra el valor de respuesta individual por ítem.

La tabla evaluaciones concentra el resultado integrado por aplicación: puntaje, nivel de ansiedad derivado y tipo de evaluación. Finalmente, recomendaciones y evaluaciones\_recomendacion implementan el módulo de retroalimentación personalizada al estudiante. Esta separación por dimensión facilita tanto la integridad referencial como la extracción selectiva de variables para el análisis de Feature Importance, permitiendo consultas eficientes por categoría sin afectar el rendimiento del sistema.

## 5. Validación técnica

### 5.1. Validación con datos híbridos

Para demostrar la viabilidad técnica del esquema multidimensional antes de la recolección empírica, se construyó un Dataset híbrido de 500 registros en dos componentes. El primero corresponde a los niveles de ansiedad derivados de respuestas reales de la subescala de ansiedad del DASS-21, obtenidas del Dataset "Mental Health Dataset based on DASS-21" [14], con umbrales ajustados empíricamente para

**Tabla 3.** Resultados de análisis: feature importance.

| Numero | Variable            | Importancia   | Dimensión      |
|--------|---------------------|---------------|----------------|
| 1      | Promedio académico  | 0.468 (46.8%) | Académica      |
| 2      | Semestre cursado    | 0.185 (18.5%) | Académica      |
| 3      | Número de materias  | 0.184 (18.4%) | Académica      |
| 4      | Situación laboral   | 0.038 (3.8%)  | Socioeconómica |
| 5      | Edad                | 0.034 (3.4%)  | Demográfica    |
| 6      | Medio de transporte | 0.032 (3.2%)  | Socioeconómica |
| 7      | Tipo de vivienda    | 0.022 (2.2%)  | Socioeconómica |
| 8      | Carrera             | 0.013 (1.3%)  | Académica      |
| 9      | Sexo                | 0.011(1.1%)   | Demográfica    |
| 10     | Beca                | 0.008 (0.8%)  | Académica      |
| 11     | Estado civil        | 0.005 (0.5%)  | Demográfica    |

**Tabla 4.** Métricas de clasificación individuales para cada clase.

| Nivel de ansiedad | Precisión | Recall | F1-score |
|-------------------|-----------|--------|----------|
| Bajo              | 0.94      | 0.97   | 0.96     |
| Medio             | 0.90      | 0.82   | 0.86     |
| Alto              | 1.00      | 1.00   | 1.00     |
| Macro AVG         | 0.95      | 0.93   | 0.94     |

garantizar representatividad en los tres niveles durante esta fase de validación técnica: Bajo (suma  $\leq 4$ ), Medio (suma 5–7) y Alto (suma  $\geq 8$ ).

El segundo componente consiste en variables contextuales generadas con relaciones lógicas establecidas con apoyo de una psicóloga especialista, basadas en patrones clínicamente documentados: estudiantes con mayor carga académica, menor

rendimiento y situación laboral activa presentan mayor probabilidad de síntomas ansiosos. Este enfoque permitió verificar que el sistema identifica y procesa correctamente los patrones esperados.

Se implementó un clasificador Random Forest con los siguientes hiperparámetros: `n_estimators=50`, `max_depth=5`, `min_samples_split=10`, `min_samples_leaf=5`, `max_features='sqrt'`, `class_weight='balanced'`, `random_state=42`. La selección de este algoritmo se justifica por su capacidad para manejar simultáneamente variables categóricas y numéricas, su resistencia al sobreajuste mediante el ensamblado de árboles, y su mecanismo de Feature Importance que permite cuantificar la contribución de cada variable a la clasificación [15].

El dataset de 500 registros se dividió en 80% entrenamiento y 20% prueba con estratificación por clase, resultando en 400 registros de entrenamiento y 100 de prueba.

**Tabla 5.** Comparación de importancia de variables: dataset híbrido vs. estudio piloto.

| Numero | Variable            | Híbrido       | Piloto        | Cambio  |
|--------|---------------------|---------------|---------------|---------|
| 1      | Promedio académico  | 0.468 (46.8%) | 0.252 (25.2%) | ↓ 21.6% |
| 2      | Semestre cursado    | 0.185 (18.5%) | 0.121 (12.1%) | ↓ 6.4%  |
| 3      | Numero de materias  | 0.184 (18.4%) | 0.146 (14.6%) | ↓ 3.8%  |
| 4      | Situación laboral   | 0.038 (3.8%)  | 0.138 (13.8%) | ↑ 10.0% |
| 5      | Edad                | 0.034 (3.4%)  | 0.182 (18.2%) | ↑ 14.8% |
| 6      | Medio de transporte | 0.032 (3.2%)  | 0.048 (4.8%)  | ↑ 1.6%  |
| 7      | Tipo de vivienda    | 0.022 (2.2%)  | 0.019 (1.9%)  | ↓ 0.3%  |
| 8      | Carrera             | 0.013 (1.3%)  | 0.063 (6.3%)  | ↑ 5.0%  |
| 9      | Sexo                | 0.011(1.1%)   | 0.030(3%)     | ↑ 1.9%  |
| 10     | Beca                | 0.008 (0.8%)  | 0.000(0.0%)   | ↓ 0.8%  |
| 11     | Estado civil        | 0.005 (0.5%)  | 0.000(0.0%)   | ↓ 0.5%  |

## 5.2. Resultados del Dataset híbrido

El modelo alcanzó una precisión general del 94% en el conjunto de prueba con una F1-score macro de 0.94 y una validación cruzada de 5 pliegues con media del 90% ( $\pm 1.37\%$ ), confirmando que el modelo generaliza correctamente sin sobreajuste significativo (diferencia train/test de 0.75%). La tabla 3 presenta los resultados del análisis de Feature Importance.

Con el fin de evaluar el desempeño del modelo por nivel de ansiedad, la tabla 4 presenta las métricas de clasificación individuales para cada clase.

Cabe destacar que estas métricas validan el funcionamiento técnico del sistema sobre el dataset híbrido de validación y no constituyen una evaluación del modelo sobre población real.

Los resultados muestran que las variables académicas concentran el mayor peso explicativo acumulado, resultado esperado dado que las relaciones lógicas del dataset híbrido fueron construidas con esa dimensión como eje principal, siguiendo la evidencia de Morales et al. [6] y Chávez-Márquez [10].

Es importante destacar que estos valores de Feature Importance reflejan la estructura del dataset de validación técnica y no constituyen conclusiones sobre los factores reales de ansiedad en la población mexicana, las cuales se obtendrán con la recolección empírica planificada.

## 5.3. Estudio piloto con muestra de datos reales

Con el objetivo de evaluar la viabilidad del esquema multidimensional más allá de la validación sintética, se realizó un estudio piloto con estudiantes reales del Instituto Tecnológico de Orizaba (ITO), Veracruz, México. Esta evaluación empírica preliminar

**Tabla 6.** Métricas de clasificación por clase: estudio piloto (n=42).

| Nivel de ansiedad | Precisión | Recall | F1-score |
|-------------------|-----------|--------|----------|
| Bajo              | 0.57      | 1.0    | 0.73     |
| Medio             | 0.50      | 0.50   | 0.50     |
| Alto              | 0.00      | 0.00   | 0.00     |
| Macro AVG         | 0.36      | 0.50   | 0.41     |

buscó verificar dos aspectos: que la plataforma AnxiTech recolecta exitosamente datos multidimensionales de estudiantes reales, y que el análisis de importancia de variables revela patrones diferentes a los observados con datos sintéticos.

La muestra piloto estuvo conformada por 42 estudiantes de licenciatura que completaron tanto la subescala de ansiedad del DASS-21 (7 ítems) como el conjunto completo de variables contextuales a través de la plataforma AnxiTech.

Se entrenó un clasificador Random Forest sobre los datos del piloto utilizando las mismas 11 variables contextuales y una configuración de hiperparámetros adaptada al tamaño de la muestra ( $n\_estimators=200$ ,  $max\_depth=4$ ,  $class\_weight='balanced'$ ). La Tabla 5 presenta el ranking de importancia de variables obtenido del piloto, comparado con los resultados del dataset híbrido.

El hallazgo más significativo de la Tabla 5 es la redistribución de la importancia de variables al transitar de datos híbridos a datos empíricos. En el dataset híbrido, la variable promedio académico concentraba el 46.8% de la importancia total, mientras que en el piloto descendió a 25.2%. Por el contrario, la variable edad ascendió de 3.4% a 18.2%, y situación laboral de 3.8% a 13.8%. Esta redistribución confirma que las correlaciones generadas sintéticamente tienden a amplificar las relaciones predefinidas, mientras que los datos reales revelan un patrón de asociaciones más equilibrado y complejo.

Agrupada por dimensión, la dimensión Académica contribuyó con el 58.3% de la importancia total, seguida de la dimensión Demográfica (21.2%) y la Socioeconómica (20.5%). La mayor relevancia de las variables socioeconómicas en el piloto (20.5% frente a aproximadamente 9.0% en el dataset híbrido) sugiere que factores como la situación laboral y el medio de transporte presentan una asociación más fuerte con la ansiedad en la población estudiantil real de lo que los datos sintéticos anticipaban.

El modelo Random Forest entrenado sobre los 42 registros del piloto alcanzó una precisión en el conjunto de prueba del 55.6% (entrenamiento: 78.8%, 9 muestras de prueba). La validación cruzada de 5 pliegues arrojó una media de 35.3% ( $\pm 20.0\%$ ). La Tabla 6 presenta las métricas por clase.

Estas métricas reflejan las limitaciones esperadas de entrenar un clasificador multiclase sobre un dataset de tamaño reducido con desbalance de clases. La diferencia sustancial entre la precisión de entrenamiento y prueba, junto con la alta varianza en la validación cruzada, indica que 42 registros resultan insuficientes para producir un clasificador confiable.

No obstante, el estudio piloto no fue diseñado para producir un modelo de clasificación desplegable.

Su propósito principal fue validar dos aspectos: (1) que la plataforma AnxiTech recolecta exitosamente datos multidimensionales reales de estudiantes, y (2) que el análisis de importancia de variables revela patrones significativamente diferentes a los obtenidos con datos sintéticos. Ambos objetivos fueron alcanzados. La obtención de un clasificador con rendimiento satisfactorio requerirá una muestra sustancialmente mayor, lo cual constituye la fase central del trabajo futuro.

#### 5.4. Consideraciones éticas

El estudio piloto se realizó bajo la supervisión del programa de posgrado del Instituto Tecnológico de Orizaba. La participación fue voluntaria y todos los estudiantes proporcionaron su consentimiento informado antes de completar la encuesta a través de AnxiTech. El sistema no recopila identificadores directos más allá del número de control institucional, el cual se almacena de forma separada de las respuestas de la encuesta. Todos los datos se transmiten mediante protocolo HTTPS y se almacenan en una base de datos protegida con acceso restringido. El instrumento DASS-21 fue aplicado en su forma original validada sin modificaciones, preservando su integridad psicométrica. El protocolo de investigación se encuentra en proceso de revisión formal por el comité de ética institucional como parte de la fase de recolección empírica a mayor escala.

### 6. Conclusiones y trabajo futuro

A continuación, se presentan las conclusiones obtenidas y las líneas de trabajo futuro planificadas.

#### 6.1. Conclusiones

Este trabajo presentó el diseño e implementación de un esquema de dataset multidimensional para el análisis de factores asociados a la ansiedad en estudiantes universitarios mexicanos, y su materialización a través del sistema AnxiTech. La principal contribución es el esquema en sí mismo: la integración estructurada de cuatro dimensiones demográfica, académica, socioeconómica y psicométrica en un único instrumento de recolección estandarizado, que no tiene precedente en la literatura para el contexto universitario mexicano.

La validación técnica con datos híbridos demostró que el sistema es funcional en sus cuatro capas, que el esquema multidimensional puede ser procesado íntegramente por el modelo Random Forest, y que el módulo de Feature Importance genera salidas interpretables y diferenciadas por dimensión.

Adicionalmente, el estudio piloto con 42 estudiantes reales del ITO confirmó que la plataforma AnxiTech recolecta exitosamente datos multidimensionales y que el análisis de importancia de variables revela patrones que difieren significativamente de los obtenidos con datos sintéticos. El hallazgo más notable fue la redistribución de la importancia de variables: mientras que el promedio académico se mantuvo como factor principal, las variables demográficas y socioeconómicas adquirieron una relevancia sustancialmente mayor en los datos reales (21.2% y 20.5% respectivamente, frente a

aproximadamente 5.0% y 9.0% en el dataset híbrido), lo que sugiere que el enfoque multidimensional captura una complejidad que los análisis unidimensionales no podrían detectar.

Las métricas de clasificación del piloto (precisión 55.6%, validación cruzada 35.3%) reflejan las limitaciones inherentes de una muestra pequeña ( $n=42$ ) y no una deficiencia del esquema o del sistema. Estos resultados son consistentes con el comportamiento esperado de Random Forest sobre datasets de tamaño reducido con desbalance de clases, y subrayan la necesidad de una recolección empírica a mayor escala.

## **6.2. Trabajo futuro**

La siguiente fase contempla la ampliación de la recolección empírica con una muestra objetivo de 300 a 500 estudiantes de múltiples carreras del ITO, con un cronograma planificado para mediados de este año 2026. La estrategia de reclutamiento incluirá la coordinación con directores de carrera y la integración con los procesos institucionales de detección de necesidades estudiantiles.

Paralelamente, se validará el esquema de variables con un especialista en psicología clínica para confirmar la pertinencia de las variables contextuales incorporadas. Una vez recolectados los datos empíricos suficientes, se reentrenará el modelo con los umbrales oficiales del DASS-21 y se evaluará mediante protocolos rigurosos de validación cruzada. Se contempla además la comparación con clasificadores alternativos (SVM, XGBoost) para fortalecer la contribución metodológica y evaluar la generalización del esquema propuesto.

**Agradecimientos.** El presente estudio recibió financiamiento de la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) y la Secretaría de Educación Pública (SEP) de México. Se reconoce especialmente la contribución del Tecnológico Nacional de México (TecNM) por proporcionar el marco institucional que hizo posible el desarrollo de esta propuesta.

## **Referencias**

1. Organización Mundial de la Salud: Trastorno de ansiedad.
2. Mayo Clinic: Trastornos de ansiedad - síntomas y causas - Mayo clinic.
3. Secretaría de Salud: 136. Ansiedad puede evolucionar y convertirse en trastorno de salud mental.
4. Beck, A.T., Epstein, N., Brown, G., Steer, R.A.: An Inventory for Measuring Clinical Anxiety: Psychometric Properties. *J Consult Clin Psychol.*, 56, pp. 893–897 (1988). Doi: 10.1037//0022-006x.56.6.893.
5. Hamilton, M.: The Assessment of Anxiety States by Rating. *Br J Med Psychol.*, 32, pp. 50–55 (1959). Doi: 10.1111/j.2044-8341.1959.tb00467.x.
6. Morales, S.T.T., Quinchalef, G. del P.V., Vera, J.A.A., Muñoz, S.I.M., Contreras, K.M.W.: Niveles de depresión, ansiedad, estrés y su relación con el rendimiento académico en estudiantes universitarios. *Investigación en Educación Médica*, 9, pp. 8–16 (2020). Doi: 10.22201/fm.20075057e.2020.36.20229.
7. Farrer, L.M., Gulliver, A., Bennett, K., Fassnacht, D.B., Griffiths, K.M.: Demographic and Psychosocial Predictors of Major Depression and Generalised Anxiety Disorder in

- Australian University Students. *BMC Psychiatry*, 16, pp. 241 (2016). Doi: 10.1186/s12888-016-0961-z.
8. Kyrou, M., Kompatsiaris, I., Petrantonakis, P.C.: Deep Learning Approaches for Stress Detection: A Survey. *IEEE Transactions on Affective Computing*, pp. 1–20 (2024). Doi: 10.1109/TAFFC.2024.3455371.
9. Moran, V.E., Olaz, F.O., Pérez, E.R., Prette, Z.A.P.D.: Desarrollo y validación del test de ansiedad social para estudiantes universitarios (TAS-U). *Liberabit. Revista de Psicología*, 24, pp. 195–212 (2018). Doi: 10.24265/liberabit.2018.v24n2.03.
10. Chávez-Márquez, I.L., Chávez, H.J.D. los R.: Niveles de ansiedad en estudiantes universitarios y factores asociados; estudio cuantitativo. *Dilemas contemporáneos: Educación, Política y Valores* (2023). Doi: 10.46377/dilemas.v10i3.3584.
11. Lara Flores, L.: Integración de una plataforma para la detección de ansiedad en alumnos universitarios (2019)
12. Flores Moscoso, G.N.: Predicción del nivel de estrés académico en estudiantes de la Universidad Nacional José María Arguedas aplicando un modelo de inteligencia artificial. (2024)
13. Drira, M., Ben Hassine, S., Zhang, M., Smith, S.: Machine Learning Methods in Student Mental Health Research: An Ethics-Centered Systematic Literature Review. *Applied Sciences*, 14, pp. 11738 (2024). Doi: 10.3390/app142411738.
14. Rony, G.R., Hossain, M.B., Chowdhury, A.A., Ahmed, F.: Mental Health Dataset Based on DASS-21. 1, (2024). Doi: 10.17632/br82d4xkj7.1.
15. Breiman, L.: Random Forests. *Machine Learning*, 45, pp. 5–32 (2001). Doi: 10.1023/A:1010933404324.