

PERFILES DE RIESGO DE SALUD MENTAL EN ESTUDIANTES UNIVERSITARIOS MEDIANTE GHQ-28 Y TÉCNICAS DE *MACHINE LEARNING*

PROFILES OF MENTAL HEALTH RISK IN UNIVERSITY STUDENTS USING GHQ-28 AND MACHINE LEARNING TECHNIQUES



¹ Doris Paola Sanabria García,
² Jesus Antonio Villarraga Palomino

^{1,2} Universidad Nacional de Colombia

Cómo citar: Sanabria García, D. P., & Villarraga, J. A. (2025). Perfiles de riesgo de salud mental en estudiantes universitarios mediante GHQ-28 y técnicas de Machine Learning. *Publicaciones E Investigación*, 19(2). <https://doi.org/10.22490/25394088.10120>

Recibido 12 Agosto 2025, Aprobado 21 Septiembre 2025

RESUMEN

El bienestar psicológico de los estudiantes universitarios es clave para su permanencia y rendimiento académico. Este estudio tuvo como objetivo identificar y caracterizar perfiles de riesgo psicológico en estudiantes de Ingeniería Agronómica y Ciencias Sociales de la Universidad de Cundinamarca, a partir del cuestionario GHQ-28, que evalúa cuatro subescalas: somatización, ansiedad/insomnio, disfunción social y depresión. La investigación, de tipo cuantitativo con enfoque descriptivo, exploratorio y predictivo, incluyó a 227 estudiantes que completaron voluntariamente el instrumento. Las respuestas fueron codificadas con el método Likert (0-1-2-3) y analizadas junto con variables sociodemográficas. Se aplicaron análisis descriptivos, reducción de dimensionalidad mediante análisis de componentes principales (PCA) y agrupamiento no supervisado con K-Means, con lo cual se identificaron tres perfiles de riesgo: bajo, intermedio y alto. Para la clasificación predictiva, se implementaron modelos de regresión logística y random forest, que evidenciaron un mejor desempeño del modelo logístico (AUC = 0.96) frente al random forest (AUC = 0.57). Los hallazgos permiten proponer estrategias de intervención diferenciadas y resaltan la utilidad del GHQ-28 como herramienta de cribado en entornos universitarios.

¹ Magíster en enseñanza de las ciencias exactas / ORCID: <https://orcid.org/0009-0003-5130-6544>
Correo electrónico: dosanabriag@unal.edu.co

² Licenciado en matemáticas. Especialista en estadística aplicada. Magíster en Ciencia de Datos. Master en enseñanza de las ciencias exactas y naturales / ORCID: <https://orcid.org/0000-0002-5379-5156> / Correo electrónico: dosanabriag@unal.edu.co

Palabras clave: salud mental; GHQ-28; somatización; ansiedad; disfunción social; depresión; PCA; K-Means; regresión logística

ABSTRACT

University students' psychological well-being is essential for their academic persistence and performance. This study aimed to identify and characterize psychological risk profiles among students of Agricultural Engineering and Social Sciences at the University of Cundinamarca, using the General Health Questionnaire (GHQ-28), which assesses four subscales: somatization, anxiety/insomnia, social dysfunction, and severe depression. This quantitative research, with a descriptive, exploratory, and predictive approach, included 227 students who voluntarily completed the instrument. Responses were coded using the Likert method (0-1-2-3) and analyzed along with sociodemographic variables. Descriptive statistics, dimensionality reduction through Principal Component Analysis (PCA), and unsupervised clustering with K-Means led to the identification of three distinct risk profiles: low, intermediate, and high. For predictive classification, logistic regression and random forest models were implemented, with logistic regression showing superior performance (AUC = 0.96) compared to random forest (AUC = 0.57). Findings support the design of targeted intervention strategies and highlight the GHQ-28's utility as a screening tool in university settings.

Keywords: mental health; GHQ-28; somatization; anxiety; social dysfunction; depression; PCA; K-Means; logistic regression.



1. INTRODUCCIÓN

La salud mental ha sido reconocida por la Organización Mundial de la Salud como un pilar esencial del bienestar humano y del desarrollo sostenible. Particularmente, la población de estudiantes universitarios constituye un grupo especial que atraviesa un periodo transitorio crítico, en el que pasan de la adolescencia a la adultez, intentan encajar, mantener un promedio, cumplir en el trabajo y adaptarse a estar lejos de casa, en medio de múltiples presiones (Sarokhani et al., 2013). En este escenario, la evaluación del estado mental de los estudiantes se convierte en una tarea prioritaria para garantizar trayectorias formativas integrales.

En este contexto, uno de los instrumentos más utilizados a nivel internacional para la evaluación de la salud mental es el General Health Questionnaire (GHQ), diseñado por David Goldberg en la década de 1970 en el Instituto de Psiquiatría del King's College London, con el objetivo de detectar la severidad

de trastornos psiquiátricos menores. A pesar de que su denominación sugiere la medición de la salud general, su enfoque real consiste en valorar el estado de salud mental, especialmente síntomas como depresión, ansiedad, inadecuación social e hipocondría (García, 1999; Goldberg, 1972). La versión abreviada de 28 ítems (GHQ-28), traducida y validada en múltiples idiomas, agrupa sus preguntas en cuatro subescalas: síntomas somáticos, ansiedad e insomnio, disfunción social y depresión grave (Goldberg y Hillier, 1979). Se trata de un cuestionario autoadministrado, ampliamente reconocido por su eficacia como herramienta de cribado y por su versatilidad de uso tanto en entornos universitarios como en el ámbito de la salud pública, lo que respalda su elección en el presente estudio (Silva-Gutiérrez y De la Cruz-Guzmán, 2017).

En Colombia, esta problemática ha sido reconocida como un asunto de salud pública, dado el incremento

sostenido en los reportes de trastornos psicoemocionales, intentos de suicidio y abandono educativo relacionados con el malestar psicológico.

Sin embargo, los procesos del sistema de salud no solo son complejos desde la parte conceptual, sino que, por sus múltiples fuentes de información, generan grandes volúmenes de datos, y el procesamiento de estos a través de técnicas de aprendizaje automático o *machine learning* resulta determinante para garantizar resultados efectivos, pues estas técnicas permiten implementar modelos con alta precisión. Por ello, se han convertido en una de las vías más utilizadas a la hora de apoyar el sistema de salud en la predicción, detección y prevención de enfermedades mentales (Tran y Kavuluru, 2017, citados por Castrillón, 2022).

En este contexto, las técnicas de machine learning emergen como herramientas poderosas para detectar patrones en grandes volúmenes de información, lo que permite segmentar a la población estudiantil según indicadores clínicos y sociodemográficos relevantes. Investigaciones recientes han demostrado el potencial de modelos predictivos en la comprensión de conductas suicidas, trastornos depresivos y alteraciones emocionales severas.

En este sentido, resulta fundamental contar con herramientas y metodologías que permitan no solo medir el estado de salud mental de los estudiantes, sino también segmentar y predecir perfiles de riesgo de forma precisa. El presente trabajo integra el uso del GHQ-28 con técnicas de machine learning como el análisis de componentes principales (PCA), el agrupamiento no supervisado (K-Means) y modelos supervisados de clasificación, con el propósito de caracterizar perfiles de riesgo de salud mental en estudiantes de la Universidad de Cundinamarca.

2. MATERIALES Y MÉTODOS

Este estudio corresponde a una investigación de tipo cuantitativa, con enfoque descriptivo, exploratorio y predictivo. El objetivo fue identificar

y caracterizar perfiles de riesgo psicológico en estudiantes universitarios a partir de sus respuestas al cuestionario GHQ-28 y variables sociodemográficas. Se utilizó un diseño no experimental, transversal y de análisis secundario de datos.

El universo de estudio estuvo conformado por estudiantes de los programas de Ingeniería Agronómica y Ciencias Sociales de la Universidad de Cundinamarca, sede Fusagasugá. La muestra fue de tipo censal, compuesta por 227 estudiantes que completaron voluntariamente el cuestionario GHQ-28 a través de un formulario de Google Forms, aplicado durante el primer semestre académico de 2025. Como criterio de inclusión, se consideraron únicamente las respuestas completas y con autorización de uso de datos. No se aplicaron criterios de exclusión adicionales.

El instrumento principal fue el cuestionario General Health Questionnaire (GHQ-28), que mide la salud mental percibida en cuatro subescalas: somatización, ansiedad/insomnio, disfunción social y depresión. La codificación se realizó mediante el método Likert (0-1-2-3), que permite una mayor sensibilidad ante variaciones en la percepción del estado mental. Adicionalmente, se recopilaban variables sociodemográficas como edad, género, programa académico, semestre, lugar de residencia, estado civil y ocupación laboral.

El procesamiento de los datos se efectuó en el entorno de Google Colab, utilizando Python 3.11 y librerías especializadas como Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn y K-modes. No se empleó equipamiento físico adicional, dado que el análisis fue totalmente computacional. Para la reducción de dimensionalidad y visualización de perfiles, se aplicó el PCA. Para el análisis no supervisado, se empleó el agrupamiento con K-Means, y para la clasificación predictiva del riesgo psicológico (definido como $\text{GHQ-28} \geq 24$), se utilizaron modelos de regresión logística y *random forest*. La variable de riesgo fue binarizada con base en el umbral propuesto por Goldberg (1979).

La validación del desempeño de los modelos predictivos se realizó mediante la división del conjunto de datos en entrenamiento y prueba (70/30) y el cálculo de métricas como exactitud, precisión, *recall*, *F1-score* y curva Receiver Operating Characteristic, ROC con Area Under the Curve, AUC. Las variables predictoras estuvieron conformadas por las subescalas del GHQ-28 y las variables sociodemográficas.

Los participantes otorgaron consentimiento informado digital para el uso de sus respuestas en fines académicos y de investigación, asegurando el anonimato de los datos.

3. DESARROLLO

Base de datos

La base de datos se consolidó a partir de la aplicación del cuestionario GHQ-28, promovido entre estudiantes de los programas de Ingeniería Agronómica y Licenciatura en Ciencias Sociales de la Universidad de Cundinamarca, Colombia. La recolección se realizó mediante muestreo por conveniencia, con invitación dirigida a estudiantes de todos los semestres de ambos programas.

El instrumento incluyó una sección de caracterización sociodemográfica (edad, sexo, semestre, lugar de residencia, estado civil, condición laboral, entre otros) y las 28 preguntas del cuestionario GHQ-28, codificadas bajo la modalidad Likert para aumentar la sensibilidad en los niveles de malestar psicológico.

Procesamiento

Recolección y validación de registros

La aplicación del cuestionario GHQ-28 se llevó a cabo mediante un formulario en línea (Google Forms) durante el segundo semestre de 2024 y el primer semestre de 2025. Previamente, los participantes aceptaron un consentimiento informado que especificaba el objetivo del estudio, su rol en la

investigación y las garantías de anonimato y confidencialidad de los datos.

Limpieza y preparación de datos

La limpieza incluyó la eliminación de columnas irrelevantes, como la hora de inicio o el correo electrónico, así como la corrección de valores faltantes (NA) y el ajuste de formatos de texto. Las respuestas al GHQ-28 fueron codificadas utilizando el método Likert scoring (0-1-2-3), lo cual permitió una mayor sensibilidad a la variabilidad en las respuestas. Las variables sociodemográficas fueron estandarizadas y transformadas en factores para facilitar su análisis posterior, tanto en los modelos de agrupamiento como en los supervisados.

Análisis descriptivo y exploratorio

Se calcularon medidas de tendencia central y dispersión para cada subescala y se exploró la distribución de puntajes totales del GHQ-28 por grupo de interés (sexo, programa, semestre, lugar de residencia). También se estimó la proporción de estudiantes en riesgo psicológico, definida a partir de un puntaje total igual o superior a 24, siguiendo el punto de corte propuesto por Goldberg (1979).

Para reducir la dimensionalidad del cuestionario y facilitar la interpretación de patrones, se aplicó un PCA. Esta técnica permitió identificar las dimensiones latentes que agrupan las respuestas de los estudiantes, con más del 70 % de la varianza explicada en los dos primeros componentes. A partir del *biplot* resultante, se visualizaron relaciones entre subescalas y estudiantes, lo cual orientó el análisis posterior de agrupamientos.

Posteriormente, se realizó un análisis de agrupamiento no supervisado mediante el algoritmo K-Means, estableciendo tres clústeres como solución óptima. Esta clasificación permitió identificar perfiles diferenciados de estudiantes con bajo, moderado y alto riesgo psicológico, según sus puntajes en las subescalas del GHQ-28. Los perfiles se visualizaron en el plano bidimensional generado por el PCA.

Entrenamiento

Finalmente, se implementaron modelos predictivos supervisados para clasificar a los estudiantes según su pertenencia al grupo de riesgo. Se construyeron dos modelos principales: una regresión logística entrenada con subescalas GHQ-28 y variables sociodemográficas, y un modelo random forest entrenado exclusivamente con variables sociodemográficas. Los modelos fueron evaluados con métricas como exactitud, sensibilidad, especificidad y AUC, utilizando un conjunto de prueba (30 %) separado del total. El desempeño del modelo logístico resultó sobresaliente ($AUC = 0.96$), mientras que el modelo random forest mostró limitaciones ($AUC = 0.57$), lo que reforzó la importancia del GHQ-28 como herramienta fundamental para la predicción.

4. DISCUSIÓN

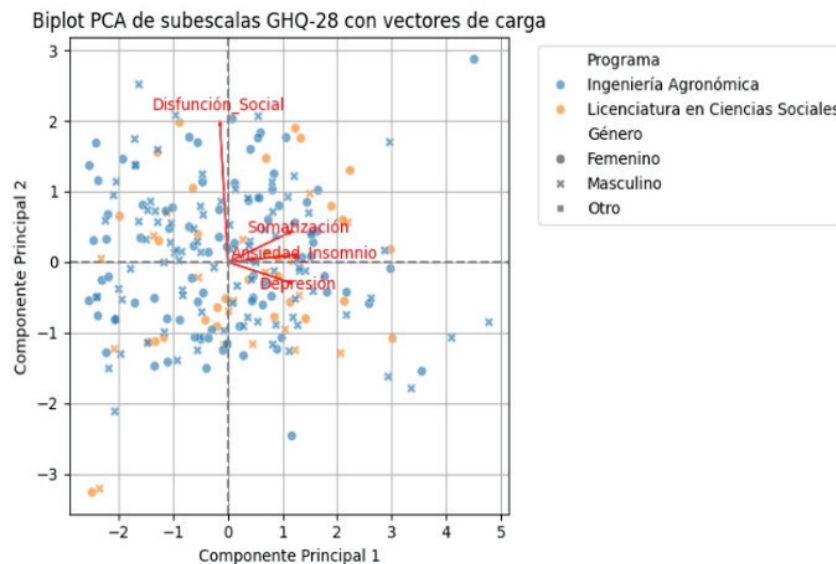
El análisis de las respuestas al cuestionario GHQ-28 permitió calcular los puntajes promedio por subescala para cada estudiante. Posteriormente, se aplicaron técnicas de aprendizaje no supervisado (K-Means) y supervisado (*random forest*) con el objetivo de identificar perfiles de riesgo psicológico y explorar predictores asociados.

En la descripción por subescala, los puntajes promedio revelaron que la disfunción social es la dimensión con mayor afectación ($M = 1.54$), seguida de somatización ($M = 1.17$), ansiedad e insomnio ($M = 1.18$) y, en menor medida, depresión severa ($M = 0.49$).

El PCA permitió identificar patrones de covariación entre las subescalas del GHQ-28. Los dos primeros componentes principales explicaron más del 70 % de la varianza total, lo que indica una estructura latente bien definida.

La figura 1 muestra la distribución bidimensional de los estudiantes a partir de sus puntajes en las cuatro subescalas del GHQ-28. Los vectores en color rojo representan las cargas de cada subescala sobre los dos primeros componentes principales, mientras que los puntos corresponden a cada participante, diferenciados por programa académico (Ingeniería Agronómica y Licenciatura en Ciencias Sociales) y género (femenino, masculino, otro). El Componente Principal 1 captura la mayor proporción de la varianza explicada, seguido del Componente Principal 2.

Figura 1 . Biplot del análisis de componentes principales (PCA) de las subescalas del GHQ-28 con vectores de carga, según el programa académico y género



Fuente: elaboración propia.

El análisis de componentes principales permitió identificar dos dimensiones latentes que agrupan las respuestas del GHQ-28. El primer componente principal (CP1) se asocia principalmente con las subescalas de ansiedad/insomnio y depresión, lo que refleja un eje de afectación emocional. El segundo componente principal (CP2) está fuertemente influenciado por la disfunción social y, en menor medida, por la somatización, representando un eje de funcionamiento social.

La distribución de los participantes sugiere que los estudiantes con puntajes más altos en disfunción social se ubican hacia valores positivos del CP2, mientras que quienes presentan mayores niveles de ansiedad/insomnio y depresión se sitúan hacia valores positivos del CP1. No se observan agrupamientos totalmente definidos por programa académico o género, lo que indica que las diferencias en

la salud mental se explican más por la variabilidad individual en las subescalas que por estas variables sociodemográficas.

Agrupamiento no supervisado

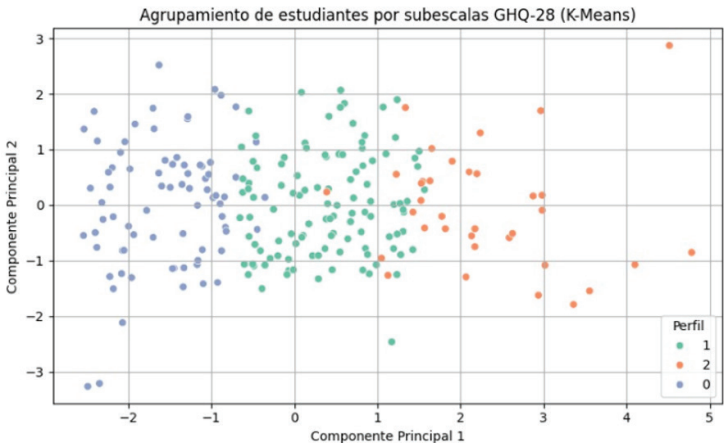
El agrupamiento con K-Means (K = 3) permitió clasificar a los estudiantes en tres perfiles psicológicos distintos. El grupo 0 incluyó a estudiantes con puntuaciones bajas en todas las subescalas (perfil de bajo riesgo); el grupo 1 representó un perfil intermedio, con afectaciones puntuales, y el grupo 2 concentró a los estudiantes con puntuaciones altas en todas las dimensiones (perfil de alto riesgo). Esta clasificación empírica ofrece una base sólida para el diseño de intervenciones diferenciales, ya que permite priorizar esfuerzos hacia el grupo con mayor nivel de afectación.

Tabla 1. Promedios por subescala del GHQ-28 y descripción de perfiles identificados mediante K-Means

Clúster	Somatización	Ansiedad / insomnio	Disfunción social	Depresión	n	Descripción del perfil
0	0.70	0.30	1.30	0.05	76	Perfil funcional
1	1.42	1.63	1.85	0.23	66	Perfil con sobrecarga emocional
2	1.63	1.82	1.59	1.77	38	Perfil de riesgo crítico

Fuente: elaboración propia.

Figura 2. Proyección bidimensional de los perfiles de salud mental obtenidos mediante el análisis de componentes principales (PCA)



Nota. La figura muestra la distribución de tres perfiles identificados a partir de los puntajes promedio por subescala del GHQ-28: perfil funcional (verde), perfil con sobrecarga emocional (naranja) y perfil de riesgo crítico (rojo).

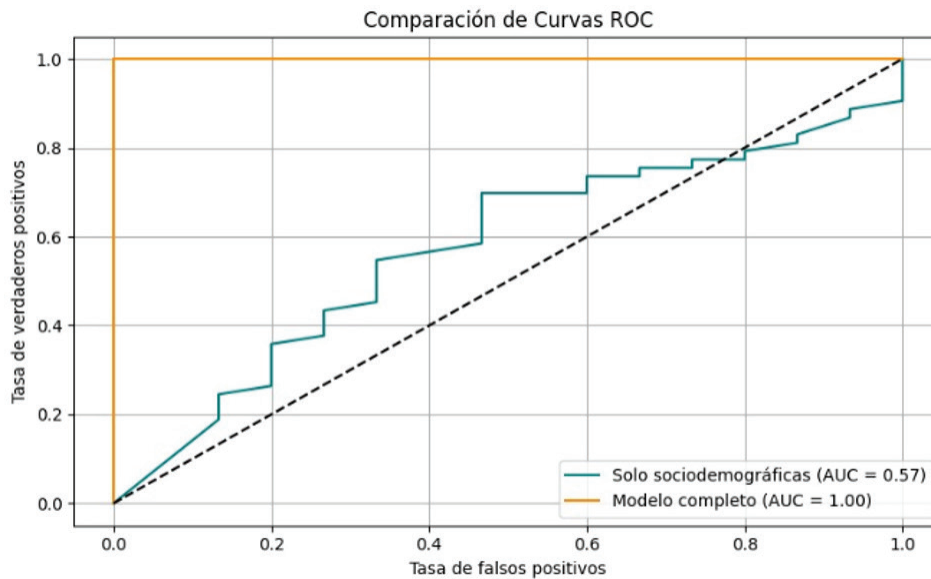
Fuente: elaboración propia.

Modelo Predictivo

Se construyó un modelo de regresión logística para predecir la pertenencia al grupo de riesgo (GHQ-28). Este modelo mostró un excelente desempeño predictivo, con un AUC de 0.96, exactitud del 95.6 %, sensibilidad del 100 % y especificidad

del 80 %, lo cual indica su utilidad como herramienta de cribado. En contraste, el modelo *random forest*, entrenado únicamente con variables sociodemográficas (edad, sexo, semestre, etc.), tuvo un AUC de 0.57, mostrando una capacidad muy limitada para detectar casos en riesgo.

Figura 3. Comparación de las curvas ROC y AUC



Fuente: elaboración propia.

5. CONCLUSIONES

El estudio permitió identificar que la disfunción social es la dimensión más afectada en la población analizada, seguida de la somatización y de la ansiedad/insomnio, con menor impacto en la depresión severa. El análisis de componentes principales evidenció la existencia de dos dimensiones independientes del malestar psicológico: una funcional y otra emocional. El agrupamiento con K-Means definió tres perfiles diferenciados de riesgo, lo que favorece el diseño de intervenciones focalizadas. El modelo de regresión logística, alimentado con datos del GHQ-28 y variables sociodemográficas,

presentó un desempeño sobresaliente, lo que confirma la relevancia del cuestionario como herramienta de cribado en entornos universitarios. En contraste, el modelo *random forest*, basado solo en variables sociodemográficas, mostró baja capacidad predictiva. Estos resultados resaltan la necesidad de incorporar instrumentos clínicos estandarizados en los procesos de seguimiento institucional de la salud mental y sugieren que el uso combinado de técnicas estadísticas y de *machine learning* puede optimizar la detección temprana y la asignación de recursos para la atención psicológica.

REFERENCIAS

- Cifuentes Madrigal, A., Gómez Méndez, T. S., y Jiménez Zapata, M. (2024). Analítica predictiva como apoyo en la salud pública: Modelos de pronóstico con series de tiempo aplicados a la conducta suicida. *Revista ELA*, 21(42). <https://doi.org/10.24050/reia.v21i42.1763>
- Goldberg, D. P. (1972). *The detection of psychiatric illness by questionnaire*. Oxford University Press.
- Goldberg, D. P., y Hillier, V. F. (1979). A scaled version of the General Health Questionnaire. *Psychological Medicine*, 9(1), 139-145. <https://doi.org/10.1017/S0033291700021644>
- Castrillon, S. O., Marín, L. M. G., Villegas, H. H. J., y Escobar, C. C. P. (2022). Machine learning aplicado en la clasificación y predicción de la depresión: Una revisión sistemática. RISTI - *Revista Iberica de Sistemas e Tecnologias de Informacao*, (Special Issue 47), 363-375.
- Sarokhani, D., Delpisheh, A., Veisani, Y., Sarokhani, M. T., Esmaeli Manesh, R. y Sayehmiri, K. (2013). Prevalence of depression among university students: A systematic review and meta-analysis study. *Depression Research and Treatment*, 2013, 1-7. <https://doi.org/10.1155/2013/373857>
- Silva-Gutiérrez, B. N., y De la Cruz-Guzmán, U. O. (2017). Autopercepción del estado de salud mental en estudiantes universitarios y propiedades psicométricas del Cuestionario de Salud General (GHQ28). *Revista Iberoamericana de producción académica y gestión educativa*, 4(8), 1-26.