Artículo de Investigación Original

Análisis de Big Data para la identificación de factores asociados a diagnósticos de salud mental en Puno.

Big Data analysis for the identification of factors associated with mental health diagnoses in Puno.

Yanina Maritza Chambi Arucutipa [1][0000-0003-0406-2914], Edwin Cruz Cruz[2][0000-0003-1353-6421], Daisy Mestas Yucra [3][0000-0002-0969-5529], Alicia Zuaña Quispe [4][0009-0004-6142-4074], Jaime Huanca Aracayo [5][0009-0002-2706-7443]

[1-5] Universidad Nacional del Altiplano – Perú.
yaninachambi@unap.edu.pe, edwin.cruz@unap.edu.pe, daisy.mestas@unap.edu.pe, aliciaz@unap.edu.pe, jaime.huanca@unap.edu.pe

CITA EN APA:
Chambi Arucutipa , Y. M., Cruz Cruz, E., Mestas Yucra, D., Zuaña Quispe , A., & Huanca Aracayo , J. (2026). Análisis de Big Data para la identificación de factores asociados a diagnósticos de salud mental en Puno. *Technology Rain Journal*, *5*(2). https://doi.org/10.55204/trj.v5i2.e155

Recibido: 06 de mayo-2026
Aceptado: 15 de agosto-2026
Publicado: 19 de agosto-2026

Technology Rain Journal
ISSN: 2953-464X

Resumen. El estudio identificó variables asociadas a diagnósticos de salud mental en Puno, Perú, una región con volúmenes significativos de datos clínicos y demográficos. Se empleó un diseño no experimental, observacional, retrospectivo, con un corpus de 583,390 registros de atención de salud mental (2021-2023), tras eliminar 2,874 duplicados exactos. La unidad de análisis fue el registro de atención con diagnóstico (no el paciente, dado que la fuente no incluye un identificador único de paciente). Se construyó un modelo Random Forest para predecir la categoría diagnóstica CIE-10 agrupada (10 clases), evaluado con partición estratificada 70/30 y validación cruzada de 5 particiones, obteniendo una exactitud de 52.2%, F1-macro de 0.44 y F1-ponderado de 0.54. La variable con mayor importancia predictiva relativa fue la edad del paciente (60%), seguida del mes (11%) y el año (6%) de atención. La segmentación con K-Means (K=5, seleccionado por coeficiente de silueta) alcanzó una silueta de 0.547. Las categorías de ansiedad/estrés (24.4%), psicóticos/esquizofrenia (20.2%) y trastornos del ánimo (19.9%) fueron las más frecuentes. Los trastornos psicóticos representaron el 21.4% de los diagnósticos en establecimientos de primer nivel de atención frente al 10.3% en hospitales; los trastornos del ánimo no mostraron diferencia relevante entre tipos de establecimiento. Las consultas de noviembre superaron el promedio mensual en 22.9%. Las conclusiones destacan que la edad y las variables temporales (mes y año) son las más asociadas a la categoría diagnóstica en este corpus, sin que ello implique relaciones causales, dado el diseño observacional del estudio.

Palabras Clave: Big Data, aprendizaje automático, salud mental, modelado predictivo, análisis de conglomerados, minería de datos, sistemas de salud.

Abstract: The study identified variables associated with mental health diagnoses in Puno, Peru, a region with significant volumes of clinical and demographic data. A non-experimental, observational, retrospective design was used, with a corpus of 583,390 mental health care records (2021-2023), after removing 2,874 exact duplicates. The unit of analysis was the care record with diagnosis (not the patient, since the source lacks a unique patient identifier). A Random Forest model was built to predict the grouped ICD-10 diagnostic category (10 classes), evaluated with a stratified 70/30 split and 5-fold cross-validation, obtaining an accuracy of 52.2%, macro-F1 of 0.44, and weighted-F1 of 0.54. The variable with the highest relative predictive importance was patient age (60%), followed by month (11%) and year (6%) of visit. K-Means segmentation (K=5, selected via silhouette coefficient) reached a silhouette score of 0.547. Anxiety/stress disorders (24.4%), psychotic disorders/schizophrenia (20.2%), and mood disorders (19.9%) were the most frequent categories. Psychotic disorders represented 21.4% of diagnoses in first-level facilities versus 10.3% in hospitals; mood disorders showed no meaningful difference between facility types. November visits exceeded the monthly average by 22.9%. The conclusions highlight that age and temporal variables (month and year) are the most associated with diagnostic category in this corpus, without implying causal relationships, given the study's observational design.

Keywords: Big Data, machine learning, mental health, predictive modeling, cluster analysis, data mining, health systems.

1. INTRODUCCIÓN

La salud mental se ha vuelto un tema cada vez más relevante para la salud pública, sobre todo en áreas con recursos escasos, como en Puno, Perú, donde el acceso a servicios especializados está limitado por múltiples condiciones geográficas y de infraestructura (Chávez Fernández, 2018; Quispe Pari, 2021; Reategui Monge, 2022). A pesar de los avances en la atención a la salud mental, estos sistemas de salud siguen teniendo dificultades, tales como la carencia de una infraestructura adecuada que permita el análisis y procesamiento sistemático de la información clínica registrada (Ahidar Tarhouchi & Ortiz de Urbina Criado, 2024). En este sentido, Big Data es una herramienta que puede contribuir significativamente al análisis de patrones en salud mental (Gandomi & Haider, 2015; Stewart & Davis, 2016; Ristevski & Chen, 2018).

El análisis de Big Data en el campo de la salud mental permite explorar información demográfica y clínica que suele pasar inadvertida en los reportes agregados convencionales (Hudson et al., 2023; Passos et al., 2019). La aplicación de aprendizaje automático ha demostrado utilidad para reconocer patrones asociados a distintos trastornos mentales (Chekroud et al., 2021; Ristevski & Chen, 2018; Palacios-Ariza et al., 2023), ayudando al procesamiento eficiente de datos complejos y heterogéneos (Goodfellow et al., 2016; Shmueli et al., 2018).

En la región de Puno, el análisis de datos de salud mediante técnicas de minería de datos y modelado predictivo puede apoyar la planificación de intervenciones y la asignación de recursos (Gonzalez Cieza & Chavez Monteza, 2023; Rosenfeld et al., 2019; Reddy & Aggarwal, 2015; Shatte et al., 2019).

La implementación de técnicas de procesamiento ETL (Extract, Transform, Load) permite integrar y depurar datos de diversas fuentes administrativas, mejorando la calidad de los análisis (Kimball & Caserta, 2004).

Si bien existen estudios de Big Data y aprendizaje automático aplicados a salud mental en poblaciones urbanas o de países de altos ingresos (Madububambachu et al., 2024), la evidencia sobre regiones andinas rurales y de altitud, como Puno, es escasa. No se han identificado estudios previos que documenten de forma reproducible qué variables administrativas y demográficas disponibles rutinariamente —edad, mes y año de atención, tipo de establecimiento— se asocian con la categoría diagnóstica registrada en los sistemas de información de salud mental de la región.

Este estudio busca cubrir ese vacío, aportando un pipeline reproducible y una evaluación honesta del poder predictivo de las variables administrativas disponibles, sin

recurrir a fuentes clínicas adicionales no disponibles en este corpus, y sin atribuir relaciones causales a un diseño observacional.

El objetivo general de este estudio es identificar las variables demográficas, temporales y administrativas más asociadas con la categoría diagnóstica de salud mental registrada en los establecimientos de salud de Puno entre 2021 y 2023 (583,390 registros analizados), mediante un modelo predictivo y técnicas de segmentación evaluados con métricas de validación completas.

2. METODOLOGÍA O MATERIALES Y MÉTODOS

Ámbito o lugar de estudio

Se realizó el estudio en la región de Puno, en el sureste del Perú, ubicada en la alta meseta del Collao a una altitud media de 3,827 m s. n. m. Esta región presenta una topografía compleja, con mesetas y el lago Titicaca, condición que impacta la disponibilidad y el acceso a servicios de salud mental. El estudio abarcó establecimientos de salud de primer nivel de atención y hospitales de la región.

Descripción de métodos

a) Periodo de estudio .- Se utilizó un diseño no experimental, observacional, de tipo retrospectivo, con datos correspondientes al periodo enero de 2021 a diciembre de 2023. El corpus consolidado contenía inicialmente 586,264 registros (unión de tres archivos anuales); tras eliminar 2,874 registros duplicados exactos, atribuibles a la consolidación de archivos, el corpus analítico quedó en 583,390 registros.

b) Descripción detallada de los materiales .- La fuente de datos no incluye un identificador único de paciente: la columna ID_CITA identifica una cita/atención (567,869 valores únicos sobre 583,390 filas), no una persona, y un mismo paciente puede aportar varios registros a lo largo del periodo. Por lo tanto, la unidad de análisis de este estudio es el registro de atención con diagnóstico, no el paciente (ver Limitaciones). La columna Tipo_Diagnostico distingue diagnóstico nuevo (“D”, n=71,048), reingreso o diagnóstico repetido (“R”, n=497,987) y diagnóstico presuntivo o provisional (“P”, n=14,355); esta distinción explica la relación entre el total del corpus (583,390) y el total de la Tabla 1, que reporta específicamente los diagnósticos nuevos.

c) Variables analizadas .- Se utilizaron las variables demográficas y administrativas disponibles: edad del paciente, tipo de edad, provincia, distrito, red de salud, categoría del establecimiento, y variables temporales (año, mes). La variable objetivo es la categoría diagnóstica CIE-10 agrupada por capítulo (10 clases), obtenida del prefijo del código CIE-10 (CODIGO_ITEM), no del texto libre de la descripción, lo que la hace

reproducible y evita que la variable objetivo comparta información directamente con las variables predictoras. El conjunto de datos no contiene una variable de zona urbana/rural; los resultados se reportan según la categorización de establecimientos del MINSA (Categoría I = establecimientos de primer nivel de atención; Categoría II = hospitales). La lectura de estas categorías como aproximación de zona rural/urbana se emplea únicamente como proxy indirecto y con cautela interpretativa, dado que reflejan nivel de complejidad del establecimiento y no una clasificación geográfica censal (ver Limitaciones).

d) Prueba estadística .- Se entrenó un modelo Random Forest (class_weight="balanced_subsample", para atenuar el desequilibrio entre categorías diagnósticas, que va de 24.4% a 0.5% del corpus) con 100 árboles y profundidad máxima 18, semilla aleatoria fija (42). Los datos se dividieron en 70% entrenamiento (408,373 registros) y 30% prueba (175,017 registros), con partición estratificada por categoría diagnóstica, y se realizó adicionalmente validación cruzada de 5 particiones (StratifiedKFold) sobre el conjunto de entrenamiento. Ninguna de las variables predictoras se deriva del texto o código del diagnóstico que define la variable objetivo, lo que descarta la fuga de información (data leakage).

La evaluación del modelo se realizó con Accuracy, Precision, Recall y F1-score (macro y ponderado, dado el desequilibrio de clases) y matriz de confusión. La importancia de variables se calculó mediante el atributo feature_importances_ de scikit-learn (Mean Decrease in Impurity, MDI) y se reporta en un gráfico independiente del desempeño del modelo; estos valores expresan importancia predictiva relativa dentro del modelo, no magnitudes de asociación, efectos causales ni proporción de varianza explicada. Para la segmentación se utilizó K-Means sobre variables numéricas estandarizadas (StandardScaler): edad, mes, año y categoría de establecimiento. Se evaluaron valores de K entre 3 y 7, seleccionando el que maximiza el coeficiente de silueta; se documentan también los parámetros de DBSCAN (eps=0.8, min_samples=15).

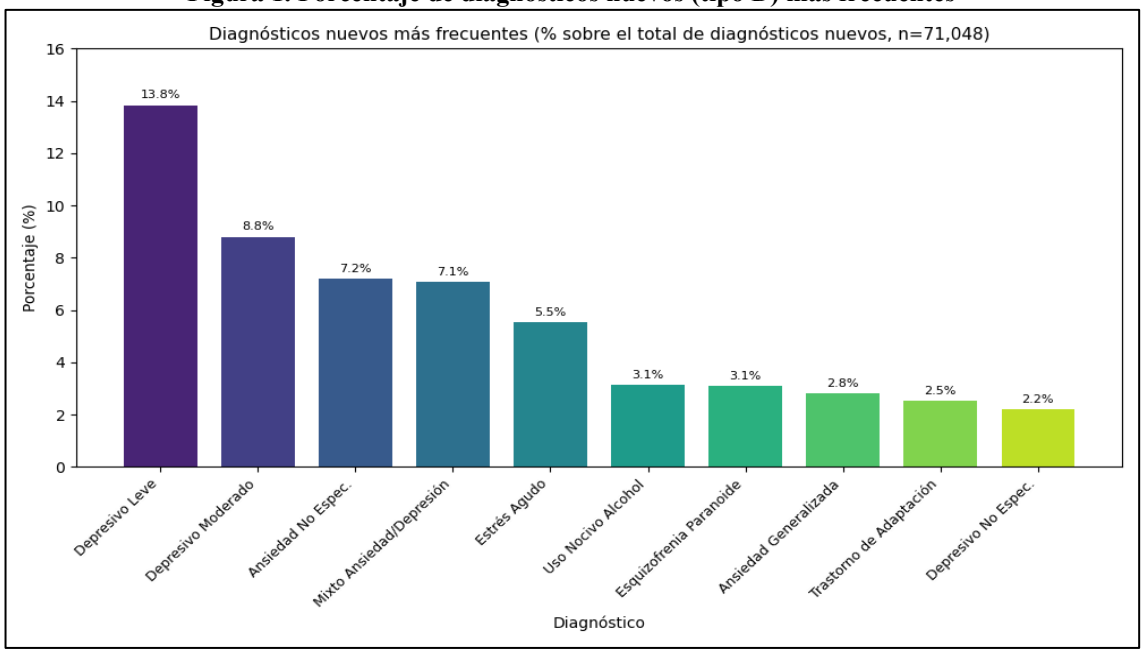
3. RESULTADOS

La variable objetivo del modelo predictivo es la categoría diagnóstica CIE-10 agrupada (10 clases), obtenida del prefijo del código CIE-10. Esta variable es reproducible, clínicamente interpretable y no comparte información con las variables predictoras, lo que evita la fuga de datos (data leakage).

El modelo Random Forest, evaluado sobre el conjunto de prueba (175,017 registros), obtuvo una Accuracy de 52.2%, Precision macro de 0.43, Recall macro de 0.57, F1-macro de 0.44 y F1-ponderado de 0.54. La validación cruzada de 5 particiones confirmó un F1-

macro de 0.441 ± 0.003, consistente con el resultado de prueba. En el patrón temporal, las atenciones de noviembre superaron el promedio mensual en 22.9%, el valor más alto del periodo analizado, en línea con la importancia predictiva del mes de atención. El modelo ofrece una estimación honesta del poder predictivo real de las variables administrativas disponibles.

Figura 1. Porcentaje de diagnósticos nuevos (tipo D) más frecuentes



Nota. El gráfico de barras muestra el porcentaje de los diez diagnósticos nuevos (Tipo_Diagnostico = "D") más frecuentes, calculado sobre el total de diagnósticos nuevos (n=71,048). Estos valores corresponden exactamente a la Tabla 1.

Tabla 1. Diagnósticos nuevos (tipo D) más frecuentes en el corpus (n=71,048)

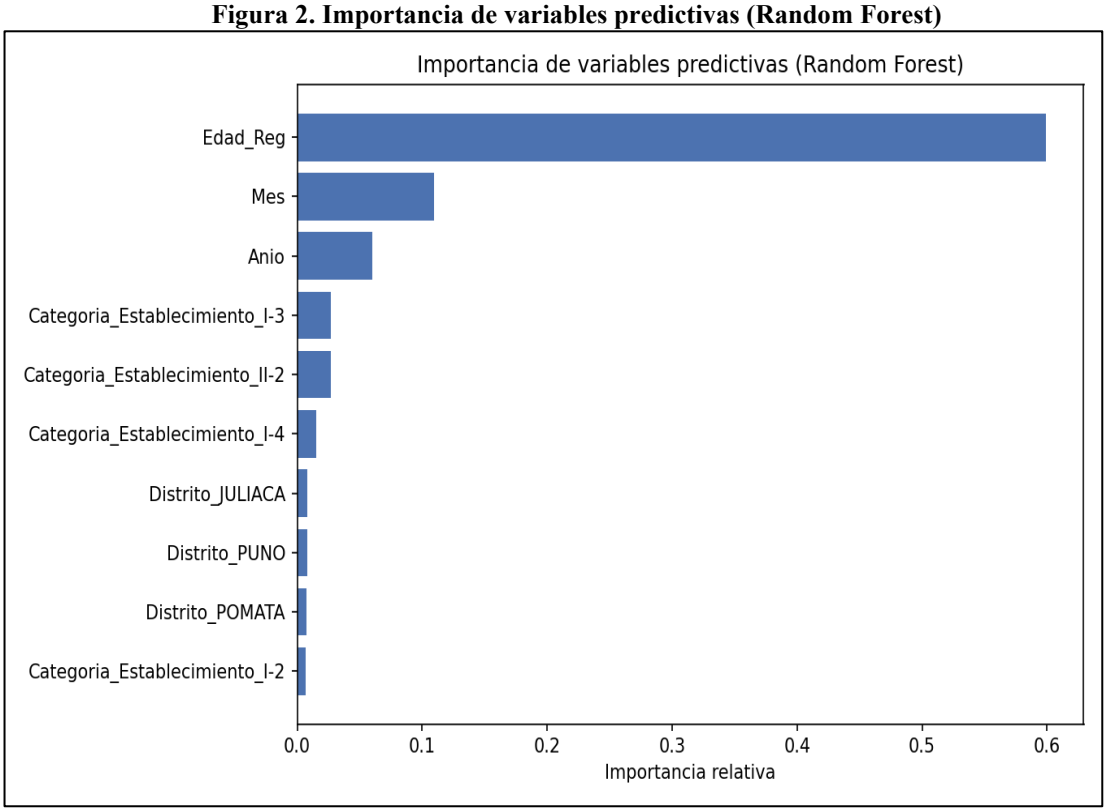
Diagnóstico	Frecuencia Absoluta	Frecuencia Relativa (%)
Depresivo leve	9,830	13.84
Depresivo moderado	6,255	8.80
Ansiedad no especificada	5,120	7.21
Mixto ansiedad/depresión	5,020	7.07
Estrés agudo	3,942	5.55
Uso nocivo de alcohol	2,217	3.12
Esquizofrenia paranoide	2,199	3.10
Ansiedad generalizada	2,006	2.82
Trastornos de adaptación	1,791	2.52
Depresivo no especificado	1,561	2.20
Otros diagnósticos nuevos	31,107	43.78
Total	71,048	100.00

Nota. La Tabla 1 reporta los diez diagnósticos nuevos (Tipo_Diagnostico = "D") más frecuentes, con porcentajes calculados sobre el total de diagnósticos nuevos (n=71,048), no sobre la totalidad del corpus (583,390 registros, que incluye además reingresos y diagnósticos presuntivos). La categoría "Otros diagnósticos nuevos" agrupa los diagnósticos fuera de los diez más frecuentes. Los porcentajes pueden no sumar exactamente 100 por redondeo.

Segmentación de la población atendida en grupos con perfiles diagnósticos y demográficos diferenciados

Se utilizó K-Means sobre variables numéricas estandarizadas (edad, mes, año, categoría de establecimiento). Se evaluaron valores de K entre 3 y 7 sobre una muestra de 30,000 registros (por costo computacional), seleccionando K=5 por maximizar el coeficiente de silueta (0.547), seguido de K=6 (0.528) y K=4 (0.489). DBSCAN (eps=0.8, min_samples=15) se usó como método complementario para identificar registros atípicos.

Figura 2. Importancia de variables predictivas (Random Forest)



Nota. El gráfico muestra únicamente la importancia relativa de las variables predictoras (Random Forest), sin mezclarla con métricas de desempeño del modelo. La variable más importante fue la edad del paciente (60% de la importancia relativa), seguida del mes de atención (11%) y el año (6%); las variables geográfico-administrativas tuvieron, individualmente, una contribución menor al 3% cada una. La importancia corresponde al método Mean Decrease in Impurity (feature_importances_ de scikit-learn) y expresa importancia predictiva relativa dentro del modelo; no debe interpretarse como magnitud de asociación, efecto causal ni proporción de varianza explicada.

Los clústeres 1 y 2 concentran la mayoría de los registros y están dominados por diagnósticos de ansiedad/estrés, diferenciándose principalmente por tipo de establecimiento (primer nivel vs. hospitales). El clúster 3, de menor edad promedio (23 años), concentra proporcionalmente más trastornos del desarrollo psicológico, consistente con diagnósticos típicamente detectados en la niñez y adolescencia. Los clústeres 4 y 5, ambos concentrados en establecimientos de primer nivel, concentran la mayor proporción relativa de trastornos del ánimo (40.0% y 26.9% respectivamente). Estos resultados deben interpretarse como patrones asociados, no como categorías clínicas discretas.

Tabla 2. Perfil de los clústeres identificados (K=5, muestra de 30,000 registros)

Cluster	Descripción	Características principales	Diagnósticos más frecuentes
Cluster 1	Grupo mayoritario en establecimientos de primer nivel	n=23,136 (muestra); edad promedio 30 años, sin banda etaria diferenciada	Ansiedad, estrés y somatomorfos (25.3% del grupo)
Cluster 2	Grupo en hospitales, perfil similar al Cluster 1	n=3,186 (muestra); edad promedio 28 años	Ansiedad, estrés y somatomorfos (22.6% del grupo)
Cluster 3	Grupo en hospitales con mayor proporción de trastornos del desarrollo	n=1,982 (muestra); edad promedio 23 años, el más joven de los 5 grupos	Trastornos del desarrollo psicológico (29.3% del grupo)
Cluster 4	Grupo de primer nivel con mayor concentración relativa de trastornos del ánimo	n=1,354 (muestra); edad promedio 30 años	Trastornos del ánimo/depresivos (40.0% del grupo)
Cluster 5	Grupo de primer nivel, pequeño, también con predominio de trastornos del ánimo	n=342 (muestra); edad promedio 30 años	Trastornos del ánimo/depresivos (26.9% del grupo)

Nota. La Tabla 2 presenta el perfil de los cinco clústeres identificados mediante K-Means sobre una muestra de 30,000 registros. Los clústeres no muestran una separación etaria limpia (todas las edades se distribuyen en rangos amplios y superpuestos); la variable más diferenciadora entre grupos es la categoría de establecimiento y la categoría diagnóstica dominante.

El desempeño moderado del modelo (F1-macro=0.44) es consistente con la naturaleza de los datos disponibles: variables puramente administrativas y demográficas, sin información clínica de severidad, comorbilidad o tratamiento. Esto no invalida el hallazgo de que la edad y las variables temporales presentan la mayor importancia predictiva relativa para la clasificación de la categoría diagnóstica dentro del modelo.

De forma exploratoria y complementaria, se procesó el texto de las descripciones diagnósticas para identificar los términos más frecuentes (nube de palabras). Esta técnica de minería de texto descriptiva tiene menor poder inferencial que el modelo predictivo y la segmentación reportados arriba, y se presenta únicamente con fines exploratorios.

Los términos más prominentes incluyen “trastorno”, “depresivo” y “episodio”, consistentes con el predominio de trastornos depresivos y de ansiedad ya reportado en la Tabla 1 y la Tabla 2.

En síntesis, la segmentación con K-Means (K=5, silueta=0.547) permite una caracterización reproducible de subgrupos de registros de atención, mientras que la nube de palabras se mantiene únicamente como exploración descriptiva, sin sustento inferencial formal.

Discusión

Este estudio aporta a la literatura sobre la aplicación de Big Data en salud mental en regiones con recursos limitados como Puno, Perú (Ristevski & Chen, 2018; Palacios-Ariza et al., 2023; Madububambachu et al., 2024). En el modelo final, la edad del paciente concentra la mayor parte de la importancia predictiva relativa (60%, según MDI), un resultado clínicamente coherente con la distribución etaria conocida de los trastornos mentales. La diferencia relativa en la proporción de registros diagnósticos de trastornos psicóticos entre establecimientos de primer nivel y hospitales (21.4% vs. 10.3%) es consistente en dirección con estudios previos sobre mayor carga de trastornos psicóticos graves en zonas con menor acceso a servicios especializados (Gonzalez Cieza & Chavez Monteza, 2023; Rosenfeld et al., 2019), aunque la magnitud no es directamente comparable por tratarse de definiciones distintas de "zona".

El uso de registros administrativos recolectados rutinariamente comparte, además, los retos de calidad de datos y de desempeño predictivo descritos en otros dominios clínicos (Shillan et al., 2019). Asimismo, la identificación de patrones en salud mental mediante Big Data exige salvaguardas éticas explícitas para evitar la estigmatización de posibles pacientes (Palomino & Berdugo, 2023).

Como línea futura, la integración de variables clínicas y de instrumentos de tamizaje validados para población peruana (Vega Dienstmaier, 2024) permitiría mejorar el poder predictivo y la utilidad sanitaria de estos modelos.

Limitaciones:

(1) no fue posible identificar pacientes únicos, por lo que los resultados corresponden a registros de atención, no a personas, y un mismo paciente puede estar representado por más de un registro;

(2) la variable de zona urbano/rural es un proxy de nivel de complejidad del establecimiento (Categoría I/II), no una clasificación geográfica censal, y puede reflejar patrones de referencia de pacientes hacia hospitales más que diferencias epidemiológicas geográficas puras;

(3) el corpus no incluye variables clínicas de severidad, comorbilidad ni tratamiento, lo que limita el desempeño predictivo del modelo (F1-macro=0.44) a lo explicable por variables administrativas y demográficas;

(4) la ausencia de un identificador de paciente impide analizar trayectorias diagnósticas longitudinales individuales;

(5) las técnicas de UMAP, modelado de tópicos (LDA), detección de comunidades (Louvain) y modelos tipo Transformers no se aplicaron de forma reproducible en este estudio y se identifican como líneas de trabajo futuro; y

(6) la importancia de variables reportada corresponde al método MDI, que puede favorecer a variables de mayor cardinalidad, y no se contrastó con la importancia por permutación, lo que se plantea como verificación futura.

4. CONCLUSIONES

En el corpus analizado (583,390 registros de atención de salud mental en Puno, 2021-2023), la edad del paciente fue la variable con mayor importancia predictiva relativa respecto de la categoría diagnóstica registrada (60%, según MDI), seguida de las variables temporales (mes y año de atención).

El modelo predictivo alcanzó una exactitud de 52.2% y un F1-macro de 0.44 para la clasificación en diez categorías diagnósticas CIE-10, un desempeño moderado pero honesto, con salvaguardas explícitas contra la fuga de información (data leakage).

La segmentación en cinco grupos mediante K-Means, con K justificado por coeficiente de silueta (0.547), permite una caracterización reproducible de subgrupos de pacientes atendidos.

Los trastornos psicóticos y de esquizofrenia mostraron una proporción de registros diagnósticos relativa notablemente mayor en establecimientos de primer nivel que en hospitales, un patrón que amerita mayor investigación con datos geográficos censales antes de derivar recomendaciones de política sanitaria específicas. Estos resultados deben interpretarse como variables asociadas y de valor predictivo, no como relaciones causales, dado el diseño observacional del estudio.

AGRADECIMIENTOS

Se agradece a la Dirección Regional de Salud (DIRESA) Puno por autorizar institucionalmente el acceso a los datos administrativos anonimizados utilizados en este estudio, y a la Universidad Nacional del Altiplano por el marco institucional para su realización.

DECLARACIÓN ÉTICA

El estudio utilizó registros administrativos de atenciones de salud mental anonimizados, sin variables identificatorias directas. El acceso a los datos fue autorizado institucionalmente por DIRESA Puno para fines de investigación académica. El estudio no contó con evaluación de un comité de ética de investigación formalmente constituido, al tratarse de un análisis secundario de datos administrativos ya anonimizados por la entidad de origen.

CONFLICTO DE INTERESES

Los Autores declaran que no existe conflicto de interés en el desarrollo de este estudio. No se ha recibido financiamiento de ninguna entidad con intereses comerciales relacionados con el tema de investigación. El rol de "adquisición de fondos" consignado en la matriz de contribución de autoría corresponde a la gestión de apoyo institucional no comercial (acceso a datos y recursos de la universidad), no a financiamiento externo.

CONTRIBUCIÓN DE AUTORÍA

	Chambi M.	Cruz E.	Mestas D.	Zuaña A.	Huanca J.
Participar activamente en:					
Conceptualización		X		X	
Análisis formal	X	X			X
Adquisición de fondos	X		X		
Investigación	X	X			X
Metodología			X		X
Administración del proyecto	X		X		
Recursos	X	X	X	X	X
Redacción –borrador original	X				
Redacción –revisión y edición				X	X
La discusión de los resultados	X	X	X	X	X
Revisión y aprobación de la versión final del trabajo.	X	X	X	X	X

REFERENCIAS

Ahidar Tarhouchi, B., & Ortiz de Urbina Criado, M. (2024). Temas de investigación sobre Big Data en el sector salud. *ESIC Market Economic and Business Journal*, *54*(2), e316. https://doi.org/10.7200/esicm.54.316

Chávez Fernández, F. A. (2018). *Factores de Riesgo del Síndrome Burnout en los Alumnos de las Clínicas Odontológicas de las Universidades de la Región Puno 2017*. Universidad Nacional del Altiplano, Escuela de Posgrado.

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., & Trivedi, M. H. (2021). Cross-trial prediction of treatment outcome in depression: a machine learning approach. *The Lancet Psychiatry*, *8*(6), 449–456. https://doi.org/10.1016/S2215-0366(21)00066-1

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Gonzalez Cieza, A. E., & Chavez Monteza, G. (2023). *Sistema para el mapeo, control y monitoreo de síntomas depresivos en escolares* [Universidad Peruana de Ciencias Aplicadas]. https://repositorio.upc.edu.pe/handle/10757/659890

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. https://mitpress.mit.edu/books/deep-learning

Hudson, C., Branjerdporn, G., Hughes, I., Todd, J., Bowman, C., Randall, M., & Stapelberg, N. J. C. (2023). Using machine learning to mine mental health diagnostic groups from emergency department presentations before and during the COVID-19 pandemic. *Discover Mental Health*, 3(1), 22. https://doi.org/10.1007/s44192-023-00047-0

Kimball, R., & Caserta, J. (2004). *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. Wiley.

Madububambachu, U., Ukpebor, A., & Ihezue, U. (2024). Machine learning techniques to predict mental health diagnoses: A systematic literature review. *Clinical Practice and Epidemiology in Mental Health*, *20*, e17450179315688. https://doi.org/10.2174/0117450179315688240607052117

Palacios-Ariza, M. A., Morales-Mendoza, E., Murcia, J., Arias-Duarte, R., Lara-Castellanos, G., Cely-Jiménez, A., Rincón-Acuña, J. C., Araúzo-Bravo, M. J., & McDouall, J. (2023). Prediction of patient admission and readmission in adults from a Colombian cohort with bipolar disorder using artificial intelligence. *Frontiers in Psychiatry*, *14*, 1266548. https://doi.org/10.3389/fpsyt.2023.1266548

Palomino, K., & Berdugo, C. (2023). Big Data Analytics and Mental Health: Would Ethics Be the Only Safeguard Against the Risks of Identifying "Potential Patients"? *IEEE Intelligent Systems*, *38*(5), 37–44. https://doi.org/10.1109/MIS.2023.3287409

Passos, I., Lemos Ballester, P., Vinícius Pinto, J., & Mwangi, B. (2019). Big Data and Machine Learning Meet the Health Sciences: Big Data Analytics in Mental Health. In *Personalized Psychiatry* (pp. 1–13). Springer. https://doi.org/10.1007/978-3-030-03553-2_1

Quispe Pari, M. D. (2021). *Factores de riesgo en la salud mental de los estudiantes de la Facultad de Ingeniería Estadística e Informática de la UNA Puno ante la pandemia de COVID-19* [Universidad Nacional del Altiplano]. https://repositorio.unap.edu.pe/handle/20.500.14082/3848

Reategui Monge, S. K. (2022). *Información digital frente al COVID y riesgos en salud mental en adultos mayores en un centro de salud – Chiclayo 2021* [Universidad Señor de Sipán]. https://repositorio.uss.edu.pe/handle/20.500.12826/4478

Reddy, C. K., & Aggarwal, C. C. (2015). *Healthcare Data Analytics* (C. K. Reddy & C. C. Aggarwal, Eds.). CRC Press. https://doi.org/10.1201/b18588

Ristevski, B., & Chen, M. (2018). Big data analytics in medicine and healthcare. *Journal of Integrative Bioinformatics*, *15*(3), 20170030. https://doi.org/10.1515/jib-2017-0030

Rosenfeld, A., Benrimoh, D., Armstrong, C., Mirchi, N., Langlois-Therrien, T., Rollins, C., Tanguay-Sela, M., Mehltretter, J., Fratila, R., Israel, S., Snook, E., Perlman, K., Kleinerman, A., Saab, B., Thoburn, M., Gabbay, C., & Yaniv-Rosenfeld, A. (2019). Big Data Analytics and AI in Mental Healthcare. arXiv. https://doi.org/10.48550/arXiv.1903.12071

Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine*, *49*(9), 1426–1448. https://doi.org/10.1017/S0033291719000151

Shillan, D., Sterne, J. A. C., Champneys, A., & Gibbison, B. (2019). Use of machine learning to analyze routinely collected intensive care unit data: A systematic review. *Critical Care*, *23*(1), 284. https://doi.org/10.1186/s13054-019-2564-9

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl, K. C., Jr. (2018). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. Wiley.

Stewart, R., & Davis, K. (2016). "Big Data" in Mental Health Research: Current Status and Emerging Possibilities. *Social Psychiatry and Psychiatric Epidemiology*, *51*, 1055–1072. https://doi.org/10.1007/s00127-016-1266-8

Vega Dienstmaier, J. M. (2024). *Construcción y validación de un instrumento para detección de trastornos de ansiedad y depresión* [Universidad Peruana Cayetano Heredia]. https://repositorio.upch.edu.pe/handle/20.500.12866/10340