“Modelos de Machine Learning para el análisis de factores socioeconómicos en la detección de riesgos de salud mental”

Autores: Jiménez A. ,Cerdán A., Pesantes A, Gonzales R.

Resumen- La detección anticipada de riesgos en salud mental es fundamental, especialmente al considerar la compleja influencia de factores socioeconómicos. Este estudio tiene como objetivo principal aplicar modelos de machine learning para analizar y predecir cómo estos factores contribuyen a la vulnerabilidad en salud mental. Se desarrollaron y compararon diversas técnicas predictivas utilizando una base de datos que integraba información socioeconómica y de bienestar psicológico. Los modelos lograron identificar con considerable acierto a individuos en situaciones de riesgo, revelando que variables como el nivel de ingresos, la estabilidad laboral y el acceso a educación son indicadores significativos. Se concluye que el aprendizaje automático representa un avance prometedor para diseñar intervenciones preventivas más efectivas y personalizadas en el ámbito de la salud mental.

1. Introducción

La salud mental es uno de los factores más importantes para el bienestar humano. Múltiples estudios han demostrado la relación entre condiciones socioeconómicas y su aportación en el surgimiento y desarrollo de trastornos mentales como ansiedad, depresión y estrés crónico.

En este contexto, el análisis de datos recolectados en encuestas por distrito permite identificar patrones en la relación antes mencionada. Emplear esta información mediante modelos de machine learning puede ofrecer una herramienta predictiva capaz de identificar grupos de riesgos de manera temprana, lo cual nos permitirá realizar acciones preventivas en salud mental.

La hipótesis que orienta esta investigación es que las condiciones socioeconómicas del hogar, medidas a través de variables como el acceso a servicios y el entorno, permiten predecir con un grado significativo de precisión el riesgo de que algún miembro enfrente problemas de salud mental en los diferentes distritos del Perú. Además, se busca determinar qué tipos de modelos de regresión se ajustan mejor a la relación entre dichas variables. Finalmente, se pretende identificar cuáles son las variables socioeconómicas que presentan una mayor relación con los problemas de salud mental.

Con todo lo anterior, el objetivo general del estudio es elaborar un modelo de Machine Learning que nos permita identificar los distritos que presentan un mayor riesgo de afectación en la salud mental de su población. Para ello, se busca realizar un análisis para la relación entre los factores socioeconómicos y la salud mental, procesar los datos de manera adecuada (con una limpieza previa a los datos) y evaluar distintos modelos de ML.

2. Trabajos relacionados

A continuación se presenta un resumen de diversas publicaciones científicas que aplican modelos y análisis de datos a distintos problemas en el ámbito de la salud:

"Accounting for racial bias and social determinants of health in a model of hypertension control" [1]: Este estudio busca que los modelos predictivos para controlar la hipertensión arterial sean más justos y eficaces para todas las personas. En lugar de solo mirar datos clínicos, los investigadores exploran cómo factores cruciales como el origen étnico-racial de una persona y sus condiciones socioeconómicas (dónde vive, su acceso a recursos, etc.) pueden y deben ser considerados al diseñar estas herramientas tecnológicas. El gran objetivo es que los avances en informática médica realmente ayuden a reducir las desigualdades en salud y ofrezcan un manejo más equitativo de esta condición tan común.

"Development and validation of a predictive model for depression risk in the U.S. adult population: Evidence from the 2007–2014 NHANES" [2]: ¿Sería posible anticipar quién podría estar en mayor riesgo de sufrir depresión? Este trabajo se enfoca precisamente en esa pregunta. Los investigadores desarrollaron y probaron un modelo que intenta predecir qué adultos en Estados Unidos tienen una mayor probabilidad de desarrollar depresión. Para lograrlo, analizaron una gran cantidad de datos sobre salud y nutrición de miles de personas (provenientes de la encuesta NHANES). La esperanza es que este tipo de herramientas puedan ayudar a identificar señales tempranas, permitiendo ofrecer apoyo y estrategias de prevención de manera más oportuna.

"A multifactorial study on duration of temporary disabilities in Spain" [3]: Cuando una persona sufre una enfermedad o lesión que la incapacita temporalmente para trabajar, ¿qué factores influyen en cuánto tiempo tardará en recuperarse y volver a sus actividades? Este estudio realizado en España se sumerge en esta cuestión, analizando una variedad de elementos —que pueden ir desde el tipo de dolencia o el sector laboral, hasta características personales o del sistema de salud— que podrían alargar o acortar estos periodos de incapacidad. Comprender mejor estas dinámicas es fundamental para diseñar políticas de salud pública más efectivas y mejorar la gestión de las prestaciones y el apoyo a los trabajadores.

3. Metodología

Se busca comparar un conjunto de modelos de machine learning que permiten evaluar la cantidad de personas que desarrollaron problemas de salud mental en función a su entorno socioeconómico.
Se trata de datos sociales y características demográficas, como información sobre la vivienda, la población, la cantidad de personas en ese sector y entre otros.
La metodología empleada en el desarrollo de los algoritmos se organiza en las siguientes etapas:

A. Adquisición de datos

Inicialmente se tomaron 2 datasets, uno recuperado de GeoPerú y abordaba todo lo relacionado al entorno económico general de diferentes distritos del país, el otro dataset muestra información sobre problemas de salud mental en personas también dividido por distritos y provincias. Con dicha información se procedió a unir ambos datasets, obteniendo más de 1000 registros y más de 100 columnas, valor que es bastante considerable, es por eso que se usarán métodos para discriminar los atributos menos significativos a través de algoritmos de selección diversos, en este caso se considerará los métodos de Filtros y Wrappers, posteriormente, seleccionaremos los atributos en común obtenidos de aplicar los criterios. Además, luego de obtener el modelo elegido, se optará por el uso del método embebido.

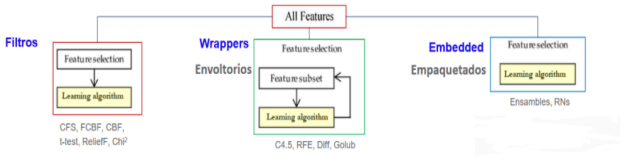

Figura 1: Diagrama de Wrappers, Embedded y filtros

B. Preprocesamiento de datos

Primero se realizó una exploración visual de los datos mediante el uso de boxplots con la finalidad de identificar los atributos más relevantes y así generar un dataset más completo. Se puede apreciar la existencia de outliers en prácticamente todas las variables representadas, siendo especialmente notables en los indicadores sociales relacionados con salud, vivienda y acceso al agua (phs_Inter, phs_sh, phs_agua_r y hbi1_porc), donde los valores atípicos se extienden por encima del rango intercuartílico. Esto refleja una alta variabilidad en dichas características sociales. También se observan valores extremos en variables relacionadas con la salud mental, lo cual podría deberse a la baja frecuencia de casos positivos en comparación con el total de la población. Será fundamental implementar técnicas para el tratamiento de valores faltantes y outliers, la limpieza de datos atípicos y la transformación de los atributos, con el fin de mejorar la calidad del análisis posterior.

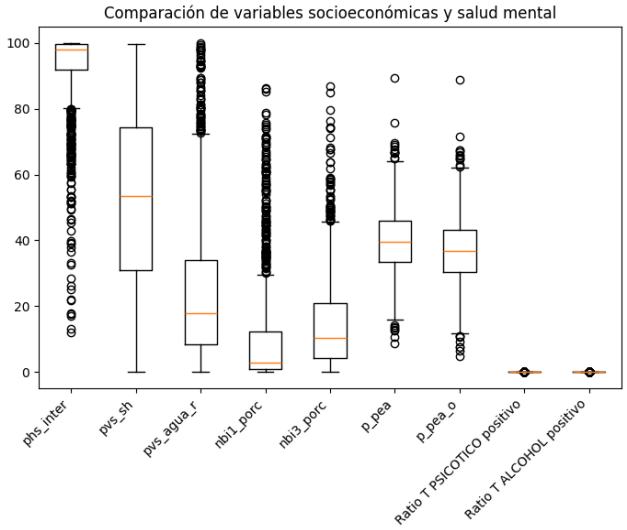

Figura 2: Gráfica de los boxplots de la data
Fuente: Elaboración propia

C. Entrenamiento de los modelos

Con el dataset ya obtenido, para el entrenamiento de los datos, se decidió separar la data en 80% para el entrenamiento y 20% para realizar el test.

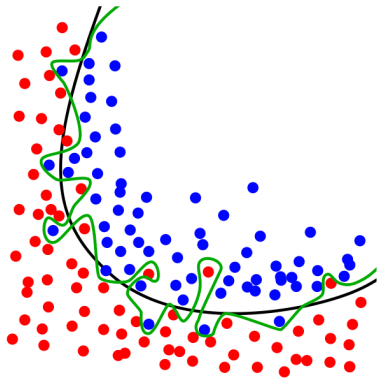

Figura 3: Imagen representativa de ajuste de datos
Fuente: Recuero, Paloma (2022)

D. Selección del modelo óptimo

Se usará la estrategia de validación cruzada para evaluar la capacidad de generalización de los modelos considerados, de modo que se pueda seleccionar el que tenga mejor desempeño en cuanto a predicción de los datos.

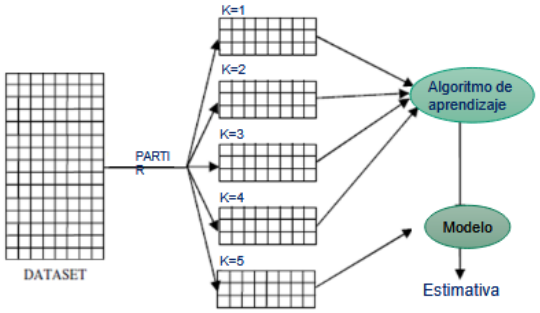

Figura 4: Tabla de tratamiento k-folds para dataset

4. Experimentación y Resultados

Aquí se detallan los procedimientos experimentales que hemos realizado, las métricas empleadas para evaluar el desempeño de los modelos, y el análisis de los resultados obtenidos. Esto con la finalidad de poder validar la hipótesis propuesta en la introducción.

- Setup experimental:

El proceso experimental se centró en el análisis de factores socioeconómicos para la detección de cantidad de casos de problemas sobre salud mental, utilizando un conjunto de datos que combina información de atenciones en salud mental (SALUD MENTAL - ATENDIDOS) con datos geográficos y socioeconómicos a nivel distrital del Perú (GeoPeru-peru_distritos). Se cargaron ambos datasets y, antes del entrenamiento de los modelos, se realizó un preprocesamiento. Inicialmente se eliminaron columnas innecesarias, luego se realizaron métodos de selección de atributos como el método filtro y el método Wrapper. Se realizó todo este procedimiento con el objetivo de filtrar la mayor cantidad de columnas del dataset inicial, que eran más de cien, para quedarnos con las más importantes para el modelo.

Una vez seleccionadas las características más relevantes obtenidas de los algoritmos y a criterio propio, se evaluó la pertinencia de las mismas a través de una matriz de correlación con el objetivo de seleccionar exclusivamente las variables más significativas, evitando redundancias innecesarias:

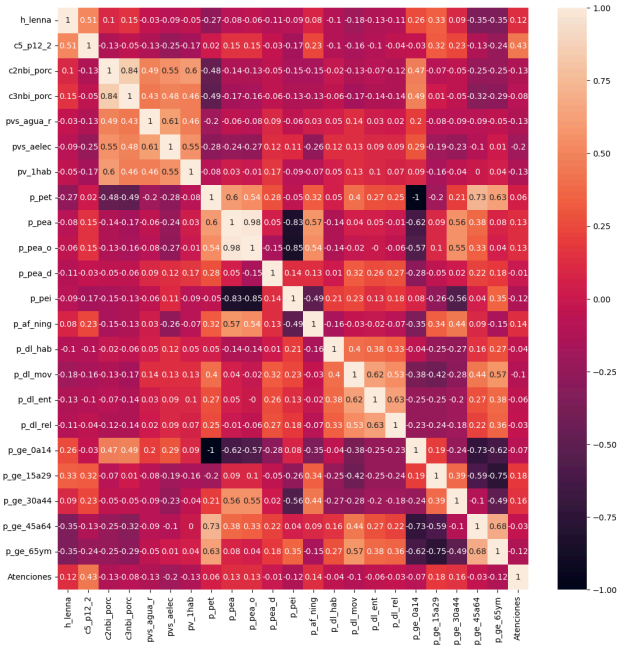

Figura 5: Matriz de correlación de atributos

La matriz de correlación ayuda a identificar relaciones cercanas o distantes entre las variables del conjunto de datos Por ejemplo, se nota una relación fuerte entre c2nbi_porc y c3nbi_porc (0.84). También hay una correlación casi perfecta entre p_pea_o y p_pea_c (0.98), lo que sugiere que se comportan de manera muy similar. En cambio, otras variables como p_pet y p_ge_0a14 tienen una correlación negativa fuerte (−1), lo que indica que cuando una aumenta, la otra disminuye.

Este resultado puso en evidencia la existencia de atributos repetitivos, en ese sentido, se identificó de forma analítica la variable más apropiada a conservar y se eliminó la redundante.

La división del conjunto de datos se estableció en un 80% para el entrenamiento y el 20% restante para la evaluación. Adicionalmente, se optó por un valor de k=10 para el método de validación cruzada K-folds. Esta estrategia hizo posible comparar la solidez de los modelos analizados en este estudio,

evaluando si los resultados obtenidos se mantienen consistentes independientemente de una partición específica del conjunto de datos. Finalmente, se optó por utilizar como métrica de desempeño el neg_mean_squared_error, ya que el objetivo del estudio es predecir la cantidad de personas en riesgo de salud mental en función de variables socioeconómicas. Al tratarse de un problema de regresión enfocado en conteos, esta métrica permite medir la precisión del modelo al comparar sus predicciones con los valores reales observados. Minimizar el error cuadrático medio es clave para garantizar que las estimaciones se acerquen lo más posible a la realidad, lo que resulta fundamental para identificar con mayor certeza los distritos más vulnerables y orientar acciones preventivas eficaces en salud pública.

■ Resultados y Discusión:

En base a la métrica de desempeño elegida, en primera instancia se determinó que el modelo Decision Tree Regressor presentó el mejor rendimiento pero no con resultados positivos.

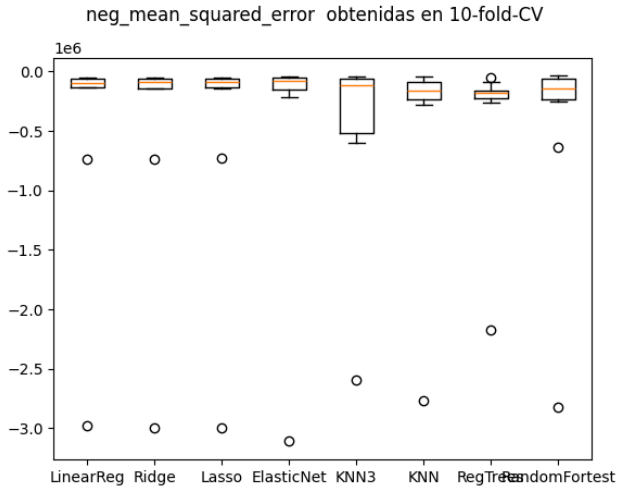

Figura 6: Gráfica de boxplot de métricas para el neg_mean_squared_error.

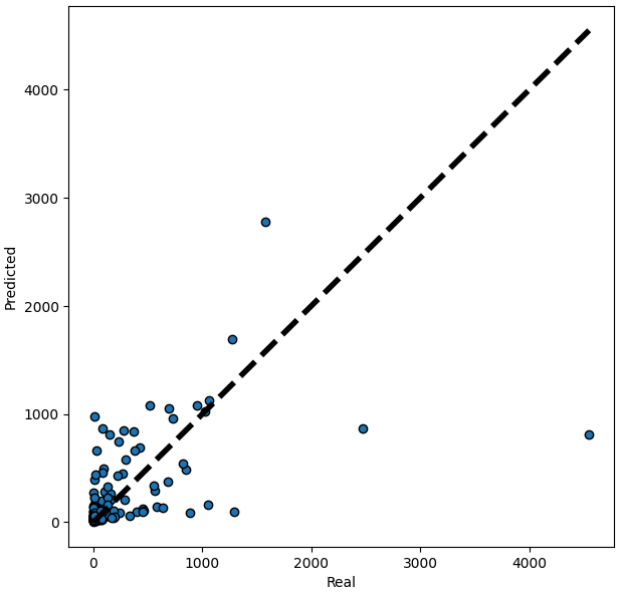

Figura 7: Gráfica de modelo de regresión antes de suavizar las variables.

Nuestra clase respuesta presenta valores muy dispersos y en muchos casos son ceros. Por lo que el modelo no podía deducir patrones notables y el rendimiento es pobre con un R2 score de 0.2715. Por lo tanto, recurrimos al método de suavizar la variable target con la ayuda de la función logaritmo y con esto obtuvimos mejores rendimientos.

Con este cambio realizado, el modelo Random Forest Regressor presentó el mejor rendimiento, y por lo tanto, fue el seleccionado como el mejor respecto a los otros modelos.

Además, realizamos una evaluación de los atributos seleccionados con el Random Forest para obtener los veinte atributos más importantes, con esto evaluamos combinaciones dentro de este top, que nos permita obtener mejores métricas, y así obtuvimos un R2 score de 0.48157, lo cual indica una mejora con respecto al modelo planteado anteriormente.

```
Mean squared error: 1.7684405828538867
Mean absolute error: 1.0663482499478485
Explained variance score: 0.4839909536676399
R2 score: 0.4815743097137821
```

Figura 8: Métricas de desempeño

Obtuviendo esta gráfica del modelo:

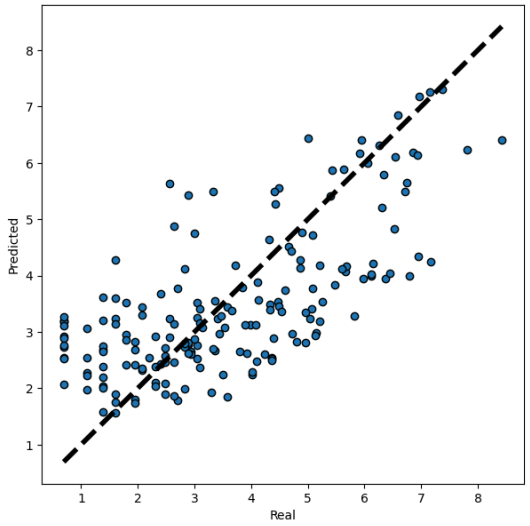

Figura 9: Gráfica de modelo de regresión luego de suavizar la variables y seleccionando 8 atributos.

5. Conclusión

Este estudio exploró la viabilidad de utilizar factores socioeconómicos y demográficos para predecir el riesgo de salud mental en la población peruana. Mediante la integración de datos de atenciones de salud mental con estadísticas geográficas y socioeconómicas a nivel distrital, se estableció una base robusta para el análisis. Se evaluaron diversos modelos de Machine Learning tradicionales, incluyendo Regresión Logística, K-Nearest Neighbors (KNN), Árbol de Decisión y Random Forest. Los resultados muestran avances significativos, aunque también resaltan la complejidad de los datos.

Se inició con una gran cantidad de datos, lo que requirió la selección de atributos para optimizar el análisis. La eliminación de variables redundantes, identificadas mediante la matriz de correlación (alta correlación entre c2nbi_porc y c3nbi_porc de 0.84), fue crucial para la eficiencia del modelo.

El conjunto de datos se dividió para entrenamiento y prueba (80/20), utilizando validación cruzada K-folds con k=10. La métrica neg_mean_squared_error fue clave para medir la precisión en la predicción de casos.

Aunque el modelo inicial Decision Tree Regressor mostró un rendimiento bajo (R2 de 0.2715) debido a la dispersión de los datos, la aplicación de una transformación logarítmica a la variable objetivo mejoró drásticamente los resultados. Con esta mejora, el Random Forest Regressor se destacó como el modelo de mejor desempeño.

Finalmente, al optimizar con los quince atributos más importantes, prescindiendo de alguno de los menos relevantes del Random Forest, se logró un R2 de 0.48157. Este valor, aunque moderado, valida la hipótesis de que las condiciones socioeconómicas pueden predecir el riesgo de problemas de salud mental, sentando las bases para futuras acciones preventivas en salud pública. Asimismo, la selección de las características más importantes permitieron descubrir que la cantidad de personas con primaria incompleta otorga información valiosa al modelo para realizar la predicción del target

6. Sugerencias de trabajos futuros

Para mejorar la capacidad predictiva de los modelos de detección de riesgo de salud mental y abordar las limitaciones identificadas en este estudio, se sugieren las siguientes líneas de trabajo futuro:

Mejora en toma de datos del dataset:

- Incorporar datos adicionales que puedan influir en la salud mental, como acceso a servicios básicos (agua, desagüe, electricidad), educación (tasa de analfabetismo, años de escolaridad), seguridad ciudadana o indicadores de cohesión social a nivel distrital.
- Incluir datos sobre la disponibilidad y acceso a servicios de salud mental a nivel local, número de especialistas o programas de salud mental existentes.

Exploración de nuevos algoritmos:

- Investigar el uso de redes neuronales, especialmente redes neuronales profundas (DNNs), que pueden capturar patrones no lineales y relaciones complejas en los datos, a menudo superando a los modelos tradicionales en datasets grandes y complejos.
- Profundizar en la optimización de los hiperparámetros de Random Forest y explorar otros algoritmos de boosting (LightGBM, CatBoost) o bagging (Bagging Classifier).
- Considerar la combinación de diferentes tipos de modelos o el uso de apilamiento (stacking) o votación (voting) para aprovechar las fortalezas de varios algoritmos.

Consideración del tiempo como variable:

- Si se dispone de datos longitudinales, analizar tendencias y patrones temporales en las atenciones de salud mental y su relación con eventos socioeconómicos.

Personalización del modelo:

- (Este punto no tiene una correspondencia directa en tus sugerencias, pero si quisieras añadirlo, podrías pensar en "Modelos adaptados a subpoblaciones específicas" como, por ejemplo, jóvenes, tercera edad, etc., si los datos lo permiten.)

Integración en sistemas productivos:

- Pensar en cómo el modelo podría ser integrado en un sistema de apoyo a la toma de decisiones para profesionales de la salud o formuladores de políticas.
- Validar los modelos con datos de atenciones de salud mental de periodos posteriores o de otras regiones para evaluar la generalización del modelo.

Optimización de interpretabilidad:

- Aplicar técnicas de explicabilidad como SHAP (SHapley Additive exPlanations) o LIME (Local Interpretable Model-agnostic Explanations) para entender qué características son las más influyentes en las predicciones del modelo. Esto no solo mejora la confianza en el modelo, sino que también puede ofrecer insights valiosos para la formulación de políticas públicas.

7. Implicancias éticas

Posibles sesgos en el modelo:

- **Problema:** El modelo puede contener sesgos estructurales, ya que está basado únicamente en variables censales.
- **Solución:** Incluir una revisión técnica que no se modelaron y complementar el modelo con datos cualitativos o encuestas si se logra escalar.

Sobre privacidad de datos:

- **Problema:** Sin seguridad y anonimato estrictos, la información podría filtrarse o usarse indebidamente, minando la confianza pública y dificultando futuras iniciativas cruciales en salud.
- **Solución:** Se deben implementar protocolos robustos de seguridad y anonimización, como encriptación y acceso restringido. Es vital establecer un marco ético y legal que limite el uso de datos exclusivamente a fines de salud pública. La transparencia en su manejo es clave para mantener la confianza de la población.

Sobre Interpretación:

- **Problema:** Los resultados del experimento podría sesgar que las variables usadas son deterministas para los problemas de salud mental.
- **Solución:** Enfatizar que la salud mental depende de múltiples factores personales, sociales y culturales que van más allá del alcance del modelo.

Uso responsable de predicciones:

- **Problema:** El modelo predictivo podría usarse para negar la atención en zonas que “predicen poca necesidad”.
- **Solución:** Dejar en claro que el modelo no busca reemplazar el juicio clínico ni el trabajo comunitario.

8. Link del repositorio del trabajo

En el siguiente repositorio de github encontrará el código en python de nuestra solución:

https://github.com/jflxx/TA-SALUD_MENTAL

9. Declaración de contribución de cada integrante

Jiménez A: Realización de partes del documento como introducción e implicaciones éticas. Contribución en la búsqueda de datos, preprocesamiento, selección y ajuste del modelo de regresión.
Cerdán A: Realización de partes del informe y ajustes del modelo de regresión junto con selección de atributos.
Pesantes A: Apoyo con la revisión general del informe y código, contribuyendo en el preprocesamiento de las variables y análisis de resultados
Gonzales R: Realización del informe (paper) y apoyo en el proceso y ajuste del modelo de regresión en el código de python. Además de apoyo al realizar el ppt para la presentación.

10. Referencias

[1]. Y. Hu, N. Cordella, R. Mishuris, y I. Paschalidis,«Accounting for racial bias and social determinants of health in a model of hypertension control-Web of Science Core Collection», vol. 25, 2025, doi: 10.1186/s12911-025-02873-4.

[2]. W. Tian *et al.*, «Development and validation of a predictive model for depression risk in the US adult population: Evidence from the 2007-2014 NHANES-Web of Science Core Collection», *BMC Psychol.*, vol. 11, 2023, doi: 10.1186/s40359-023-01278-0.

[3]. C. Gonzales, J. Montanero, y D. Peral, «A multifactorial study on duration of temporary disabilities in Spain-Web of Science Core Collection», vol. 72, pp. 328-335, 2017, doi: 10.1080/19338244.2016.1246410.