



Universidad de
SanAndrés

**Patrones alimenticios y salud mental:
Un Análisis Predictivo Basado en Healthy Survey IPUMS en la
población estadounidense"**

Autoras:

Milagros Arjona

Regina Zasiakin Drittanti

Mila Palacios Vieira

Profesores:

Ignacio Spiousas

María Noelia Romero

Asignatura:

Ciencias de Datos

Tutorial:

Martes 19:00hs.

Semestre y año de presentación:

2do semestre, 2024

Introducción

Los hábitos alimenticios no saludables en Estados Unidos han emergido como uno de los principales problemas de salud pública en las últimas décadas, con consecuencias devastadoras para la salud física y mental de la población. Los patrones dietéticos en el país han experimentado una notable transformación, caracterizándose por un aumento en el consumo de alimentos procesados, ricos en azúcares añadidos, grasas saturadas y sal, mientras que la ingesta de alimentos frescos y nutritivos ha disminuido considerablemente. Por ejemplo, un informe del USDA explora cómo la tendencia de comer fuera de casa y el aumento de la comida rápida han influido en los patrones dietéticos poco saludables de los estadounidenses (Saksena *et al.*, 2018).

Esta tendencia se ve impulsada por diversos factores, como el acceso limitado a alimentos saludables en muchas comunidades, especialmente en zonas de bajos ingresos, la proliferación de la comida rápida, y las agresivas estrategias de marketing de la industria alimentaria. Como resultado, enfermedades como la obesidad, la diabetes tipo 2 y las afecciones cardiovasculares se han convertido en epidemias, afectando a una parte significativa de la población estadounidense. Además, la mala alimentación no solo tiene repercusiones físicas, sino que también está estrechamente relacionada con problemas de salud mental, como la depresión y la ansiedad, lo que agrava aún más la carga de los trastornos de salud en el país.

Teniendo en cuenta el contexto estadounidense, la presente propuesta se plantea como pregunta de investigación: ¿Cómo los patrones alimenticios pueden servir como indicadores predictivos de trastornos de salud mental en diversos grupos poblacionales en función de sus características demográficas?

Para abordar esta pregunta, se propondrá utilizar métodos de visualización de datos como gráficos de dispersión e histogramas, con la finalidad de identificar tendencias y patrones iniciales. Además, se aplicará el método de clúster K-means agrupando a los individuos en tres clústeres predefinidos: alimentación no saludable, alimentación variada y alimentación muy saludable. Esto permitirá identificar patrones comunes en los comportamientos alimenticios y asociarlos con variables de salud mental.

Utilizando enfoques complementarios como la regresión logística y el modelo Random Forest, se buscará identificar las relaciones entre estas variables y la probabilidad de desarrollar trastornos mentales. Mientras que la regresión logística proporciona interpretabilidad y permite analizar la magnitud de las asociaciones lineales, el Random Forest captura relaciones más complejas y no lineales entre las variables. Este enfoque combinado permitirá equilibrar precisión, interpretabilidad y robustez en el análisis, con el objetivo de generar resultados accionables que puedan contribuir a la prevención y gestión de los trastornos de salud mental desde una perspectiva de salud pública.

Literatura previa

A partir de la revisión bibliográfica, se observó que las relaciones entre los hábitos alimentarios, el bienestar psicológico y los trastornos mentales han sido ampliamente estudiadas en la literatura de psicología y salud pública. En este contexto, se utilizaron diferentes enfoques metodológicos. A continuación, se presentan algunos estudios relevantes que sirven como base para la propuesta de investigar las asociaciones entre la calidad de la dieta y la salud mental en jóvenes.

El estudio de Battista *et al.* (2022) emplea árboles de decisión para explorar las asociaciones entre diversos factores de riesgo y tres dimensiones clave del bienestar psicológico en jóvenes: ansiedad, depresión y florecimiento. A través de esta técnica, los autores pudieron

identificar no solo los factores más relevantes que influyen en cada uno de estos resultados, sino también subgrupos específicos dentro de la población que presentan patrones diferenciados de ansiedad, depresión y florecimiento. Esta metodología, que permite una segmentación más precisa de la población, facilita la comprensión de cómo diferentes variables, como el entorno social, las características personales y los hábitos de vida, interactúan de manera compleja para afectar la salud mental de los jóvenes.

El estudio de Jacka *et al.* (2009) utiliza un modelo de regresión logística que incluye variables como los hábitos alimentarios, el índice de masa corporal, la actividad física y el tabaquismo, entre otras. Se seleccionó este modelo debido a que la variable dependiente, la depresión, es dicotómica (sí/no), lo que la hace apropiada para un análisis de regresión logística. El propósito principal de la investigación es estimar las probabilidades de depresión sintomática en función de la calidad de la dieta, ajustando por las covariables mencionadas. Los odds ratios obtenidos mediante la regresión logística permitieron evaluar la magnitud del efecto de las variables independientes sobre la probabilidad de desarrollar depresión.

El estudio de De Melo et al. (2024) emplea un Análisis Factorial de Componentes Principales (PCA) para identificar y analizar patrones dietéticos, con el objetivo de explorar la relación entre hábitos alimentarios no saludables y los trastornos mentales comunes (CMD), como la depresión y la ansiedad. Utilizando esta técnica estadística, que permite reducir la dimensionalidad de los datos y destacar los patrones subyacentes, los investigadores lograron agrupar los alimentos en tres categorías distintas. Al segmentar los alimentos en patrones específicos, los autores pudieron determinar cómo cada uno de estos patrones dietéticos está vinculado a un mayor o menor riesgo de desarrollar trastornos mentales, contribuyendo así a una comprensión más profunda de cómo la dieta puede influir en la salud mental.

El estudio de Vermeulen-Smit *et al.* (2014) utiliza un análisis de clases latentes para identificar clusters de comportamientos de riesgo para la salud, como tabaquismo, consumo excesivo de alcohol, inactividad física y dieta poco saludable. Este método es adecuado para clasificar patrones de comportamiento en grupos no observables directamente, y analizar su relación con trastornos mentales según criterios del DSM-IV. El propósito principal de la investigación es explorar cómo los clusters de comportamientos están asociados con trastornos mentales, como depresión, dependencia de drogas y abuso de alcohol, utilizando la Encuesta de Salud Mental de los Países Bajos (NEMESIS-2). Los resultados muestran que las personas en los tres clusters no saludables presentan un doble riesgo de depresión y están fuertemente asociadas con trastornos como la dependencia de drogas y el abuso de alcohol.

En resumen, se identificaron puntos en común entre los estudios analizados, destacando el enfoque en la segmentación de grupos poblacionales mediante diferentes metodologías que sirven para comprender las relaciones entre los hábitos alimentarios, los comportamientos de riesgo y la salud mental. Todos los estudios comparten el objetivo de identificar patrones o *clusters* que permiten analizar cómo las conductas específicas se asocian con trastornos mentales, con distintos métodos como PCA, árboles de decisión y regresión logística. En conjunto, estos estudios refuerzan la relevancia de aplicar técnicas analíticas que segmenten y expliquen patrones complejos en el comportamiento humano para diseñar intervenciones dirigidas y efectivas en salud pública.

Base de datos

Para abordar la pregunta de investigación, se extrajeron las variables de interés del IPUMS Healthy Survey, que proporciona datos históricos y contemporáneos de encuestas de salud en EE.UU., abarcando el periodo de 2015 a 2023. Las variables seleccionadas incluyen datos sociodemográficos, así como datos sobre hábitos alimenticios y comportamientos de salud.

En total, la base de datos contiene 35 variables. Algunas de las más relevantes son: edad, sexo, frecuencia de consumo de frutas y verduras, percepción general de la salud, nivel de actividad física, horas de sueño, situación socioeconómica (si la persona se encuentra por debajo del umbral de pobreza), frecuencia de episodios de pánico, niveles de depresión y ansiedad, y búsqueda de información sobre salud, entre otras.

Metodología

Para abordar la pregunta de investigación, se implementarán distintos métodos estadísticos y de *machine learning*, con el objetivo de explorar y comprender las variables para lograr predecir cómo los patrones alimenticios servirían como indicadores predictivos de trastornos de salud mental.

En primer lugar, se realizará un análisis exploratorio de los datos. Esta etapa inicial implica examinar detalladamente los datos demográficos y de comportamiento obtenidos sobre la salud y la alimentación. Durante esta fase, se utilizarán diversas técnicas de visualización de datos, como histogramas y gráficos de dispersión, para identificar tendencias y patrones iniciales. Este análisis preliminar es crucial, ya que permitirá detectar cualquier valor atípico en los datos, comprender las relaciones básicas entre las variables y obtener una visión general de la distribución de los consumidores, sus patrones alimenticios y cómo se relaciona con su salud mental.

Por ejemplo, se identificará si ciertos grupos de edad tienen una mayor tendencia a consumir alimentos más saludables o si existen patrones en la ingesta de frutas y verduras en función de factores como el nivel socioeconómico o la ubicación geográfica. Además, utilizando gráficos de dispersión, se analizará la relación entre el consumo de frutas y verduras y la

presencia de problemas de salud mental, con el fin de identificar posibles correlaciones y patrones subyacentes en los datos.

El análisis más importante para la identificación de perfiles de salud será el análisis de *clústers*, utilizando específicamente el método de K-means. Este método agrupa a los individuos en segmentos homogéneos basados en características compartidas, como patrones alimenticios, hábitos de vida y factores sociodemográficos. En el análisis de *clústers K-means*, se seleccionará un número predefinido de tres grupos (K), y los datos se asignan a estos grupos de manera que se minimice la variabilidad dentro de cada grupo y se maximice la variabilidad entre ellos.

Para esta investigación, se establecerán tres *clústers* predefinidos:

- Alimentación no saludable, caracterizada por un alto consumo de alimentos ultraprocesados, bebidas azucaradas y bajo consumo de frutas y verduras.
- Alimentación variada, con patrones de consumo equilibrados que incluyen alimentos procesados, pero también una cantidad moderada de alimentos frescos.
- Alimentación muy saludable, definida por un alto consumo de frutas, verduras, pescado y otros alimentos asociados con una dieta rica en nutrientes esenciales.

Este enfoque permitirá identificar patrones comunes en los comportamientos alimenticios y asociarlos con variables de salud mental, como la frecuencia de ansiedad, depresión y pánico, así como con características demográficas, como edad y nivel socioeconómico. Al aplicar K-means, no solo será posible entender mejor las interacciones entre alimentación y salud mental, sino también diseñar estrategias de intervención específicas para cada grupo, optimizando así las recomendaciones en salud pública y nutrición.

En este estudio, se utilizarán regresión logística y Random Forest para predecir cómo los patrones alimenticios pueden servir como indicadores de trastornos de salud mental. La regresión logística se seleccionará como un modelo base por su capacidad para modelar relaciones entre múltiples variables explicativas y una variable dependiente binaria (presencia o ausencia de trastornos de salud mental). Este método permitirá interpretar la magnitud y la dirección de las asociaciones entre los hábitos alimenticios y otros factores como el nivel socioeconómico, el consumo de alimentos ultraprocesados y la actividad física con la probabilidad de presentar trastornos de salud mental. Paralelamente, se implementará un Random Forest, que es un modelo no paramétrico que logra capturar relaciones complejas y no lineales entre las variables predictoras. Random Forest entrena múltiples árboles de decisión promediando sus resultados, lo que proporcionará robustez frente al ruido en los datos y permite identificar las variables más importantes para la predicción.

Se evaluará la performance de cada modelo utilizando la métrica de Error Cuadrático Medio (MSE) a través de validación cruzada, lo que garantiza un mejor control del sobreajuste y una evaluación más confiable del desempeño de los modelos en datos no vistos. Para mejorar el desempeño y manejar posibles problemas de multicolinealidad, se aplicarán técnicas de regularización como Ridge y Lasso tanto a la regresión logística como al Random Forest. Estas técnicas restringen los coeficientes del modelo, reduciendo la sensibilidad a variables irrelevantes y estabilizando las predicciones.

Cada método tiene ventajas y limitaciones específicas en este contexto. La regresión logística permite una interpretación clara y sencilla de los coeficientes, lo que facilita la identificación de los factores que más contribuyen al riesgo de trastornos de salud mental. Sin embargo, asume una relación lineal entre las variables predictoras y la *log-odds* de la variable dependiente, lo que podría limitar su capacidad para capturar relaciones complejas o no

lineales. Por otro lado, el Random Forest no requiere suposiciones sobre la forma funcional de los datos y puede manejar relaciones complejas y no lineales entre las variables predictoras. Sin embargo, su principal desventaja es la falta de interpretabilidad y el mayor costo computacional, especialmente con conjuntos de datos grandes.

Finalmente, se realizará una comparación entre los modelos en función del Error Cuadrático Medio (MSE) obtenido en la validación cruzada y seleccionaremos el modelo con el mejor desempeño. Esta combinación de modelos y técnicas analíticas permitirá balancear interpretabilidad y precisión, garantizando resultados robustos y accionables en el contexto de la salud pública.

Conclusiones y limitaciones

En conclusión, los modelos predictivos desarrollados en este estudio permitirán identificar de manera más precisa los factores alimentarios y demográficos asociados con los trastornos de salud mental en la población estadounidense. Esta información es crucial para diseñar estrategias de intervención dirigidas, que promuevan hábitos alimenticios más saludables y reduzcan la incidencia de problemas de salud mental. Además, los resultados pueden ser utilizados por las autoridades de salud pública de Estados Unidos para orientar políticas nutricionales, priorizar recursos en comunidades vulnerables y crear campañas educativas que aborden tanto la dieta como su impacto en el bienestar mental. Este enfoque basado en datos ofrece una base sólida para implementar acciones efectivas y sostenibles que mejoren la calidad de vida de los ciudadanos y reduzcan las cargas sociales y económicas asociadas con los trastornos de salud mental.

Este estudio presenta algunas limitaciones importantes que deben tenerse en cuenta. En primer lugar, no se contó con datos específicos sobre las dietas detalladas de las personas y su relación directa con indicadores de salud mental, lo que restringe la capacidad para capturar

información clave. Además, los datos provienen de encuestas basadas en la autopercepción y memoria de los participantes, lo que puede introducir sesgos de respuesta o imprecisiones en las variables.

Por otro lado, aunque la regresión logística es útil para interpretar relaciones entre variables, su capacidad para capturar interacciones complejas o no lineales es limitada, lo que podría llevar a una simplificación excesiva de los datos. En cambio, el modelo de Random Forest, al ser más robusto frente a relaciones no lineales, ofrece una mayor flexibilidad en la captura de patrones complejos. Sin embargo, su principal desventaja radica en la falta de interpretabilidad, lo que dificulta su aplicación en el diseño de políticas públicas claras y comprensibles, ya que no proporciona explicaciones intuitivas sobre cómo las variables afectan los resultados. Por otro lado, una limitación del uso de los métodos de regularización es el riesgo de que penalice variables relevantes y afecte la precisión del modelo.

Finalmente, aunque el uso de la regularización es crucial para manejar la multicolinealidad y simplificar los modelos, existe el riesgo de que penalice variables relevantes, lo que podría afectar la precisión del modelo. Sin embargo, estas limitaciones no restan valor al enfoque del presente estudio, el cual ha considerado cuidadosamente diversas variables y metodologías. De todos modos, se recomienda realizar investigaciones futuras que profundicen aún más en estos aspectos, para reforzar la precisión y aplicabilidad de las conclusiones obtenidas.

Bibliografia

- Battista, K., Patte, K. A., Diao, L., Dubin, J. A., & Leatherdale, S. T. (2022). Using decision trees to examine environmental and behavioural factors associated with youth anxiety, depression, and flourishing. *International Journal of Environmental Research and Public Health*, 19(17), 10873.
- de Melo Simplício, A. P., de Almeida Fonseca Viola, P. C., de Carvalho Lavôr, L. C., de Lima Sousa, P. V., de Carvalho, C. A., Lopes Rodrigues, L. A., ... & de Macedo Gonçalves Frota, K. (2024). Unhealthy Dietary Pattern Associated with Common Mental Disorders in Adults and Older Adults: A Population-based Study. *Current Nutrition & Food Science*, 20(9), 1155-1164.
- Jacka, F. N., Kremer, P. J., Leslie, E. R., Berk, M., Patton, G. C., Toumbourou, J. W., & Williams, J. W. (2009). Asociaciones entre la calidad de la dieta y el estado de ánimo depresivo en adolescentes: resultados del estudio australiano Healthy Neighbourhoods. *Medical Journal of Australia*, 190(8), 426-430.
<https://doi.org/10.3109/00048670903571598>
- Vermeulen-Smit, E., Ten Have, M., Van Laar, M., & De Graaf, R. (2015). Clustering of health risk behaviours and the relationship with mental disorders. *Journal of affective disorders*, 171, 111-119.
- Saksena, M. J., Okrent, A. M., Anekwe, T. D., Cho, C., Dicken, C., Effland, A., Elitzak, H., Guthrie, J., Hamrick, K. S., Hyman, J., Jo, Y., Lin, B. H., Mancino, L., McLaughlin, P. W., Rahkovsky, I., Ralston, K., Smith, T. A., Stewart, H., Todd, J., & Tuttle, C. (2018). *America's eating habits: Food away from home* (Economic Information Bulletin No. 196). USDA Economic Research Service. <https://doi.org/10.22004/ag.econ.282348>