MÁSTER UNIVERSITARIO EN BIO-INFORMÁTICA Y ANÁLISIS DE DATOS BIOMÉDICOS

"Análisis no supervisado de perfiles clínico-dietético-metabolómicos para la identificación de patrones asociados a la adherencia a la dieta EAT-Lancet"

Autora: Gloria Álvarez Alegre

Tutor académico: Rodrigo Madurga de Lacalle

Supervisor: José Aguilar

Curso académico *2024/2025*

Universidad Francisco de Vitoria

DECLARACIÓN PERSONAL DE NO PLAGIO

Yo, Dña Gloria Álvarez Alegre con NIF/NIE 20093972E estudiante del **Máster Universitario en Bioinformática y Análisis de Datos Biomédicos** de la Universidad de Francisco de Vitoria, como autor/a de este documento académico, titulado "Análisis no supervisado de perfiles clínico-dietético-metabolómicos para la identifica-ción de patrones asociados a la adherencia a la dieta EAT-Lancet" presentado como Trabajo de Fin de Máster, para la obtención del Título correspondiente, **declaro que, es fruto de mi trabajo personal, que no copio, que no utilizo ideas, formulaciones, citas integrales e ilustraciones diversas, sacadas de cualquier obra, artículo, memoria, etc., (en versión impresa o electrónica), sin mencionar de forma clara y estricta su origen, tanto en el cuerpo del texto como en la bibliografía.**

Así mismo, soy plenamente consciente de que el hecho de no respetar estos extremos es objeto de sanciones universitarias y/o de otro orden.

En Madrid a 7 de Julio de 2025.

Fdo.

Experimental Master's Final Project

Análisis no supervisado de perfiles clínico-dietético-metabolómicos para la identificación de patrones asociados a la adherencia a la dieta EAT-Lancet

Gloria Álvarez Alegre [1], Rodrigo Madurga de Lacalle [1,*], Dr. José Aguilar [2,], Dr. J. Alfredo Martínez[3,**], Edwin Fernández Cruz[3,**]**

[1] Universidad Francisco de Vitoria; gloria.alvale@gmail.com, rodrigo.madurga@ufv.es
[2] IMDEA Networks; jose.aguilar@imdea.org
[3] IMDEA Nutricion; jalfredo.martinez@imdea.org, edwin.fernandez@nutricion.imdea.org
* Academic tutor
** Institution/Company Tutor

Abstract

Antecedentes y objetivos: La dieta EAT-Lancet ha emergido como un modelo global para fomentar patrones alimentarios saludables y sostenibles. Sin embargo, la evaluación objetiva de la adherencia individual se ve limitada por la complejidad de integrar datos clínicos, dietéticos y metabolómicos. Este estudio propone un enfoque innovador basado en aprendizaje automático no supervisado y minería de reglas para identificar perfiles clínico-dietético-metabolómicos asociados a la adherencia a la dieta EAT-Lancet. **Métodos:** Se analizaron 54 variables clínicas, dietéticas y metabolómicas en una cohorte adulta. Se aplicaron múltiples algoritmos de clustering no supervisado — GMM, K-means, clustering aglomerativo, DBSCAN y HDBSCAN— sobre conjuntos de datos preprocesados con y sin ingeniería de características basada en filtrado por correlación de Spearman ($\rho \geq 0.8$). La calidad y estabilidad del clustering se evaluaron mediante índices de calidad. Además, se utilizó minería de reglas a priori para descubrir asociaciones significativas entre metabolitos y variables dietéticas. Los análisis se llevaron a cabo en Python, empleando las bibliotecas scikit-learn, mlxtend y statsmodels. **Resultados:** El modelo GMM con dos clusters (k=2) mostró la mayor estabilidad y precisión en todas las bases. La caracterización demostró diferencias significativas en colesterol HDL, edad, adherencia a la dieta EAT-Lancet y consumo de alimentos clave como frutos secos, legumbres y cereales integrales. Asimismo, se identificaron metabolitos diferenciales —incluyendo aminoácidos ramificados y hipurato— asociados a patrones metabólicos distintos. La minería de reglas complementó estos hallazgos, evidenciando patrones integrados de dieta y metabolismo.

Conclusiones: Este estudio demuestra que la combinación de técnicas no supervisadas con minería de reglas es una estrategia potente para identificar perfiles clínico-dietético-metabolómicos relacionados con la adherencia a la dieta EAT-Lancet. Los resultados resaltan el valor de los metabolitos como biomarcadores objetivos y allanan el camino hacia estrategias avanzadas de nutrición personalizada y salud pública basada en evidencia multivariante.

Keywords: EAT-Lancet, aprendizaje automático, ingeniería de características, clustering, análisis no supervisado, algoritmo a priori, metabolómica, nutrición personalizada.

1. Introducción.

El impacto de la alimentación sobre la salud humana y la sostenibilidad del planeta ha adquirido una relevancia sin precedentes en las últimas décadas. El incremento global de enfermedades crónicas no transmisibles, junto con el deterioro ambiental asociado a los sistemas alimentarios, ha impulsado la necesidad de adoptar patrones dietéticos más saludables y sostenibles. En este contexto, la Comisión EAT-Lancet propuso en 2019 un modelo alimentario de referencia —la denominada “dieta de salud planetaria”— diseñado para mejorar los resultados en salud pública y, al mismo tiempo, mitigar el impacto ambiental de la producción y el consumo de alimentos (Willett et al., 2019). Esta propuesta promueve una dieta rica en alimentos de origen vegetal (frutas, verduras, legumbres, cereales integrales, frutos secos) y restringida en productos animales, especialmente carnes rojas, así como en alimentos ultraprocesados y azúcares añadidos. Su adopción se ha asociado con la prevención de hasta 11 millones de muertes al año a nivel mundial, así como con una notable reducción en el uso de recursos naturales y en las emisiones derivadas de la cadena alimentaria (Stubbendorff et al., 2022).

Sin embargo, evaluar de forma precisa la adherencia individual a este patrón dietético continúa siendo un desafío metodológico importante. Las herramientas más comúnmente utilizadas, como los cuestionarios de frecuencia de consumo (FFQ), los recordatorios de 24 horas o los diarios alimentarios, presentan limitaciones bien conocidas: subregistro, sesgos de memoria, errores en la estimación de porciones y variabilidad en la composición de los alimentos registrado

(De La O et al., 2025). Estas debilidades dificultan la obtención de una medida precisa y reproducible del cumplimiento dietético, especialmente cuando se pretende establecer relaciones con variables clínicas o moleculares.

Ante estas limitaciones, la integración de información dietética con datos clínicos y metabolómicos se presenta como una vía prometedora en el campo de la nutrición de precisión. La metabolómica, al permitir el análisis de metabolitos en biofluidos como sangre u orina, ofrece una representación cuantificable de los procesos metabólicos influenciados por la dieta. Así, biomarcadores como el ácido hipúrico (asociado al consumo de frutas y polifenoles), la trigonelina (relacionada con la ingesta de café) o ciertos ácidos grasos de cadena impar (indicadores del consumo de productos lácteos) han demostrado su utilidad para validar la exposición dietética más allá del autorreporte (De La O et al., 2025; Fernández-Cruz et al., 2025).

El tratamiento de este tipo de datos multidimensionales complejos—que combinan variables clínicas, dietéticas y metabolómicas— requiere enfoques analíticos avanzados. En este sentido, las técnicas de aprendizaje automático han emergido como herramientas clave para el análisis exploratorio en nutrición. En particular, los métodos no supervisados permiten identificar estructuras latentes o agrupamientos naturales dentro de poblaciones heterogéneas, sin necesidad de depender de variables objetivo (Azmi et al., 2025; DeGregory et al., 2018). Estos enfoques resultan especialmente útiles para detectar patrones complejos que podrían estar vinculados a distintos niveles de adherencia dietética.

Un paso clave en este tipo de análisis es la ingeniería de características, entendida como el proceso de selección, transformación y construcción de variables relevantes a partir de los datos originales. Esta etapa permite reducir la redundancia, mejorar la calidad de la información utilizada y favorecer la interpretación de los resultados obtenidos, especialmente en contextos donde se integran dominios heterogéneos como el clínico, dietético y metabolómico. Según Kirk et al. (2022), aplicar estrategias de ingeniería de características es fundamental para mejorar tanto la estabilidad como la interpretabilidad de los modelos en estudios de nutrición de precisión.

A su vez, más allá de la identificación de perfiles mediante agrupamiento, resulta fundamental comprender cómo se relacionan entre sí las variables implicadas. En este sentido, la minería de reglas de asociación permite explorar combinaciones frecuentes de atributos, facilitando la detección de relaciones específicas entre patrones dietéticos y biomarcadores metabolómicos. Este enfoque, al generar reglas del tipo "si… entonces…", complementa la caracterización de grupos y aporta un marco interpretativo sobre posibles mecanismos fisiológicos vinculados a la dieta. El algoritmo a priori, una de las herramientas más utilizadas para este fin, ha demostrado su utilidad tanto en bioinformática como en estudios de elección alimentaria al identificar conjuntos de ítems relacionados dentro de grandes bases de datos (Guan et al., 2018).

En este trabajo se propone un enfoque de análisis no supervisado sobre una base de datos clínico-dietético-metabolómica, proporcionada por el Instituto IMDEA Nutrición (Madrid, España), con el objetivo de identificar perfiles poblacionales diferenciados y descubrir asociaciones relevantes entre variables dietéticas y metabolitos. El análisis combina técnicas de ingeniería de características, agrupamiento no supervisado y minería de reglas de asociación, con el fin de aportar evidencia empírica que contribuya al desarrollo de estrategias de nutrición personalizada y sostenible, basadas en relaciones objetivas entre el consumo alimentario y la expresión metabólica.

2. Objetivos.

Objetivo general

Explorar perfiles clínico-dietético-metabolómicos mediante técnicas de aprendizaje no supervisado, con el fin de identificar patrones dietéticos y descubrir asociaciones relevantes entre variables dietéticas y metabolitos vinculadas a la adherencia a la dieta EAT-Lancet.

Objetivos específicos

1. Preprocesar y depurar la base de datos, incluyendo el tratamiento de valores faltantes y atípicos, y la selección de variables relevantes mediante técnicas de ingeniería de características.
2. Aplicar métodos de agrupamiento no supervisado para identificar perfiles diferenciados de individuos según sus características clínicas, dietéticas y metabolómicas.
3. Analizar los grupos obtenidos e interpretar sus diferencias en relación con la adherencia a la dieta EAT-Lancet.
4. Implementar algoritmos de minería de reglas de asociación para detectar relaciones frecuentes entre variables dietéticas y metabolitos, tanto en la base total como dentro de los grupos definidos.

3. Materiales y métodos.

El análisis se realizó a partir de una base de datos original1 que contenía un total de 54 variables, divididas en: (1) datos clínicos y antropométricos básicos; (2) variables dietéticas procedentes de cuestionarios validados, incluyendo indicadores de adherencia al patrón EAT-Lancet; y (3) perfiles metabolómicos obtenidos mediante análisis de laboratorio en muestras biológicas. Para el primer paso de exploración de construyó un diccionario descriptivo donde se resumen las características estadísticas de cada variable (Anexo).

Preprocesamiento de datos e ingeniería de características.

Los pasos seguidos se basaron en la metodología CRISP-DM (Schröer et al., 2021). Se eliminaron exclusivamente aquellas variables que contenían más de 50 valores ausentes (NA), en una muestra total de 150 individuos, quedando N = 138 individuos. Los valores inferiores al límite de detección (<LOD) del aparato en el caso de las variables metabólicas, se trataron como NA. La imputación de NAs se realizó mediante la mediana o regresión, según el grado de correlación con otras variables, si >0.7. Los valores atípicos fueron identificados mediante el criterio de rango intercuartílico ampliado, pero no se trataron pues podríamos estar limitando el estudio. Se realizó un proceso de selección y transformación de variables para reducir la dimensionalidad y facilitar la interpretabilidad: (a) eliminación de variables con baja varianza; (b) eliminación de redundancias entre variables mediante análisis de correlación de Spearman ($\rho > 0.8$); reducción de dimensionalidad mediante PCA, estas transformaciones no se usaron en el análisis final.

Análisis no supervisado y minería de reglas.

Se aplicaron diversos algoritmos de clustering no supervisado: k-means, Gaussian Mixture Models (GMM), clustering aglomerativo, DBSCAN y HDBSCAN. Para k-means y GMM se realizaron 10 ejecuciones independientes con distintas semillas para evaluar la estabilidad. Se evaluaron diferentes k, seleccionando el más robusto según índices de calidad (silueta, DBCV). El algoritmo a priori se empleó para minería de reglas de asociación entre metabolitos y variables dietéticas, usando umbrales mínimos de soporte y confianza de 0.1 y 0.6, filtrando reglas frecuentes y confiables para su posterior interpretación nutricional.

Entorno de desarrollo y herramientas.

Los análisis se realizaron en Python 3.8, utilizando el entorno de desarrollo Spyder. Se emplearon bibliotecas especializadas como pandas para manipulación de datos, scikit-learn para clustering y mlxtend para minería de reglas de asociación.

4. Resultados y discusión.

4.1. Preprocesamiento y selección de variables.

4.1.1. Preprocesamiento de datos.

El conjunto original constaba de 59 variables cuantitativas continuas que incluían datos clínicos, dietéticos y metabolómicos, y una muestra inicial de 150 individuos. Se homogeneizó el formato decimal, seguido de la conversión de los valores inferiores al límite de detección (<LOD) en valores faltantes (NA), con el fin de evitar sesgos en los análisis subsiguientes. Esta decisión se fundamenta en que los valores por debajo del LOD pueden no reflejar la concentración real del metabolito y su imputación o tratamiento incorrecto podría introducir errores en los modelos.

Se eliminaron 6 variables que contenían más de 50 valores ausentes, aproximadamente el 33,3 % de la muestra, umbral considerado adecuado para garantizar calidad en los datos sin perder información relevante. Las variables eliminadas fueron: Glycolate, Glucose, Galactose, Taurine, Betaine y Methylsuccinate. Además, se excluyeron 12 individuos con ausencia total o casi total de datos, quedando una muestra final de 138 participantes. Los valores atípicos fueron identificados mediante el criterio de rango intercuartílico ampliado y se conservaron por su posible valor biológico.

La presencia de datos faltantes representa un obstáculo para la mayoría de los algoritmos de agrupamiento (clustering). Por ello, se aplicó un procedimiento automatizado de imputación, utilizando la mediana para variables sin correlaciones fuertes, y regresiones lineales para aquellas con correlaciones altas (> 0.7) con otras variables. Este método minimiza la distorsión de la distribución original y contribuye a la estabilidad y robustez del análisis posterior.

4.1.2. Ingeniería de características.

La ingeniería de características es una etapa fundamental en el análisis de datos multivariantes. Su principal objetivo es reducir la redundancia entre variables y facilitar una interpretación biológica y nutricional precisa, sin comprometer la integridad de la información.

Se implementó un filtrado inicial basado en la varianza para eliminar variables con baja variabilidad, que aportan escaso valor discriminatorio y pueden introducir ruido en el análisis, seguido de un filtrado por correlación de Spearman ($\rho > 0.8$) para identificar y eliminar variables altamente correlacionadas, minimizando la multicolinealidad (Fig. 1A).

El impacto de este proceso se refleja en la reducción desigual del número total de variables entre los bloques clínicos, dietéticos y metabolómicos (Fig. 1B). La disminución es notablemente mayor en los bloques clínico y metabolómico, mientras que el bloque dietético se mantiene igual. Este proceso mejora la calidad y robustez de los análisis no supervisados posteriores, facilitando la interpretabilidad clínica y nutricional de los patrones identificados. Se priorizó la conservación de variables con relevancia biológica conocida, especialmente dentro del bloque dietético, para garantizar la integridad biológica del estudio.

Aunque se exploraron técnicas de reducción dimensional mediante análisis de componentes principales (PCA), estas no fueron incorporadas en el análisis final debido a la dificultad para interpretar biológicamente los componentes resultantes y a que los análisis basados en las variables originales filtradas mostraron mayor calidad e interpretación clínica.

Por último, se procedió a normalizar todas las variables mediante estandarización Z-score. Esta transformación fue esencial para garantizar que todas las variables contribuyeran de forma equitativa en los análisis posteriores, evitando que diferencias en las escalas numéricas distorsionaran las métricas de similitud o la relevancia de las asociaciones detectadas.

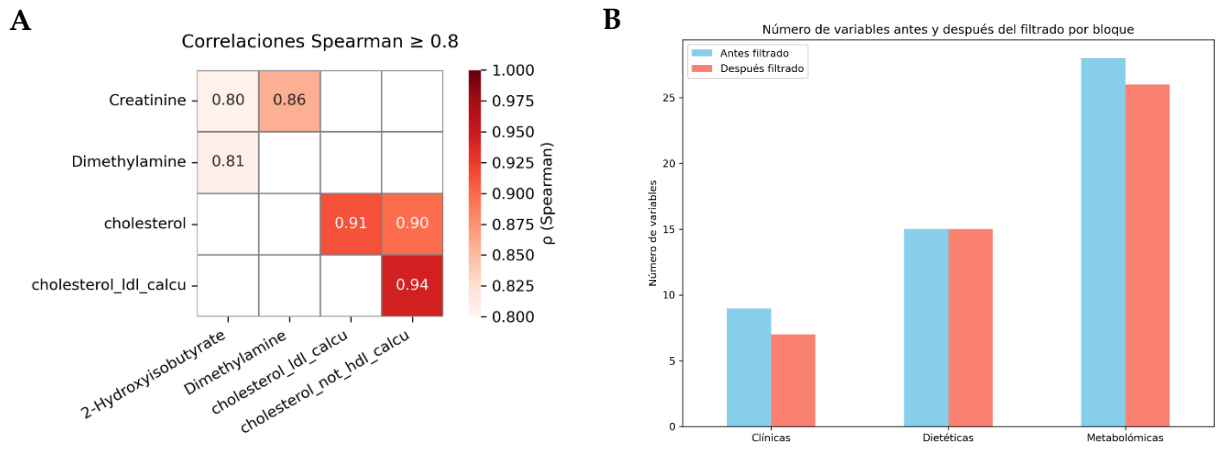

Figura 1. Ingeniería de características. A) Mapa de calor entre las variables seleccionadas tras el filtrado por ($p \geq 0.8$). Los colores y los valores numéricos dentro de cada celda reflejan la intensidad de las asociaciones, evidenciando fuertes correlaciones que motivaron la eliminación de variables redundantes. B) Número de variables antes y después del filtrado desglosadas en tres bloques. Se observa una disminución en los bloques clínico y metabolómico. Esta reducción contribuye a un conjunto de datos más manejable y balanceado.

4.2. Resultados del agrupamiento.

4.2.1. Selección y comparación de métodos.

El análisis del agrupamiento se llevó a cabo con el objetivo de identificar perfiles clínico-dietético-metabolómicos asociados a la adherencia a la dieta EAT-Lancet. Para ello, se aplicaron múltiples métodos no supervisados (ver Tabla 1), con el fin de evaluar su rendimiento y estabilidad en diferentes versiones preprocesadas de la base de datos.

Tabla 1. Métodos evaluados durante el estudio.

Método	Descripción
K-means	Minimiza la suma de distancias cuadráticas dentro de grupos (clusters) esféricos.
Gaussian Mixture Models (GMM)	Modelo probabilístico que permite identificar clusters con formas y tamaños variados.
Clustering aglomerativo	Fusiona clusters jerárquicamente según proximidad.
DBSCAN / HDBSCAN	Algoritmos basados en densidad que detectan regiones densas separadas por zonas menos densas.

Estos algoritmos se aplicaron a dos bases de datos, original (A) y con ingeniería de características (B), para la calidad de los agrupamientos generados. En todos los casos, se probó un rango de clusters (k = 2 a 6) y se realizaron múltiples ejecuciones con diferentes semillas para garantizar la estabilidad de los resultados.

La calidad de los resultados de los algoritmos de agrupamiento se determinó mediante índices de calidad, incluyendo el índice de silueta, que evalúa la cohesión y separación de los clusters, y el índice DBCV (Density-Based Clustering Validation), específico para métodos basados en densidad. El índice de silueta varía entre -1 y 1, donde valores más altos indican agrupamientos más definidos y separados, mientras que el DBCV evalúa la densidad relativa y estructura de los clusters para validar agrupamientos no esféricos.

Los algoritmos basados en densidad (DBSCAN y HDBSCAN) presentaron limitaciones debidas a la alta dimensionalidad de los datos, lo que generó resultados con agrupamientos de poca calidad y abundancia de puntos clasificados como ruido. En contraste, GMM destacó por ofrecer consistentemente los mejores índices de calidad (silueta) en todas las bases de datos analizadas, siendo k = 2 el número óptimo de clusters en todas las versiones analizadas, lo que sugiere la existencia de dos perfiles diferenciados en la muestra (ver Tabla 2, Anexo).

La calidad de GMM frente a K-means puede atribuirse a la capacidad del modelo para representar clusters con formas elípticas y covarianzas heterogéneas, adecuándose mejor a la estructura biológica y nutricional subyacente que el modelo rígido y esférico de K-means (Wang et al., 2019).

Tabla 2. Resultados de clustering para las bases A y B con el número óptimo de clusters (k = 2). Se presentan la media y desviación estándar del índice de silueta (variabilidad de este índice a lo largo de múltiples ejecuciones con diferentes semillas), y el promedio de outliers para métodos basados en densidad. A: base de datos original normalizada; B: incluye ingeniería de características.

Dataset	Método	k	Media Silueta	Desviación	Media Outliers
A	GMM	2	0.2094	0.0959	-
	K-means	2	0.1462	0.0014	-

	Agglomerative	2	0.1334	-	-
	DBSCAN	0	-	-	138.0
	HDBSCAN	2	0.0248	-	111.5
B	GMM	2	0.2165	0.0830	-
	K-means	2	0.1502	0.0016	-
	Agglomerative	2	0.1659	-	-
	DBSCAN	0	-	-	138.0
	HDBSCAN	2	0.0383	-	94.0

4.2.2. Caracterización de los clusters.

Los análisis posteriores que siguen se realizan con la base de datos B (con ingeniería de características) pues dio mejores resultados. La caracterización demográfica de los clusters (ver Tabla 3) reveló una distribución medianamente equilibrada de participantes, con diferencias moderadas en edad y sexo entre grupos. El Cluster 1 agrupó a individuos de mayor edad y proporción de hombres, lo que sugiere que las variables demográficas podrían contribuir a la variabilidad metabolómica observada.

Tabla 3. Caracterización demográfica básica de los clusters identificados. Se muestra el número total de individuos, edad media, mediana y desviación estándar de la edad, así como el porcentaje de hombres en cada cluster.

Cluster	N	Edad media (años)	Mediana edad (años)	SD edad	% hombres
0	51	39.7	41	12.2	60.8
1	87	44.9	47	15.0	74.7

En el ámbito clínico (ver Tabla 4), el Cluster 1 presentó niveles significativamente más altos de colesterol HDL. Además, la edad media también fue superior en este grupo. En todos los

casos, las diferencias entre clusters fueron estadísticamente significativas según el test no paramétrico de Kruskal-Wallis ($p < 0.05$), lo que indica que los grupos difieren de manera consistente en estas variables clínicas.

Tabla 4. Variables clínicas con diferencias significativas entre clusters. Se muestran medias, desviaciones estándar, el test estadístico empleado y su p-valor.

Variable	Media C0	Media C1	SD C0	SD C1	Test	p-valor
cholesterol_hdl (mg/dL)	57.67	64.94	13.61	18.19	Kruskal-Wallis	0.0251
Age (años)	39.71	44.94	12.18	15.01	Kruskal-Wallis	0.0435

Respecto al perfil dietético (ver Tabla 5), el Cluster 1 mostró una mayor adherencia a la dieta EAT-Lancet y un mayor consumo de alimentos saludables como frutos secos, legumbres y cereales integrales. Por el contrario, el Cluster 0 presentó un consumo relativamente superior de carne de ave. Estas diferencias evidencian la existencia de distintos hábitos alimentarios entre grupos, que podrían estar vinculados a las variaciones metabólicas observadas (ver tabla 6). El test de Kruskal-Wallis confirmó que estas diferencias son estadísticamente significativas ($p < 0.05$).

Tabla 5. Variables dietéticas con diferencias significativas entre clusters. Se presentan medias, desviaciones estándar, el test estadístico utilizado y el p-valor correspondiente.

Variable (g/día)	Media C0	Media C1	SD C0	SD C1	Test	p-valor
EATlancet	22.88	24.62	2.83	3.74	Kruskal-Wallis	0.0014
EAT_nuts	20.18	28.19	22.87	25.55	Kruskal-Wallis	0.0218
EAT_poultry	65.46	59.37	30.23	47.99	Kruskal-Wallis	0.0232
EAT_legumes	20.44	24.59	14.21	16.44	Kruskal-Wallis	0.0391
EAT_whole_grains	33.25	55.35	51.83	68.46	Kruskal-Wallis	0.0490

En análisis metabolómico (ver Tabla 6) se identificaron 26 metabolitos con niveles significativamente distintos entre clusters (Anexo). Entre ellos, creatinina, valina, leucina y hipurato fueron más elevados en el Cluster 0, lo que podría reflejar alteraciones en el metabolismo proteico, función renal y actividad de la microbiota intestinal. Estas diferencias subrayan la estrecha relación entre dieta y metabolismo, configurando fenotipos metabólicos diferenciados.

Tabla 6. Top 10 variables metabolómicas en orina con diferencias significativas entre clusters. Se incluyen medias, desviaciones estándar, tipo de test estadístico y p-valor.

Variable (mmol/L)	Media C0	Media C1	SD C0	SD C1	Test	p-valor
Xanthosine	0.33	0.19	0.08	0.07	ANOVA	1.21E-18
Creatinine	27.94	13.92	8.63	6.10	Kruskal-Wallis	4.61E-17
Valine	0.11	0.05	0.04	0.02	Kruskal-Wallis	6.65E-17
Leucine	0.08	0.04	0.03	0.01	Kruskal-Wallis	9.66E-16
3-Methyl-2-oxovalerate	0.34	0.18	0.14	0.06	Kruskal-Wallis	1.68E-15
3-Hydroxyisobutyrate	0.28	0.13	0.11	0.06	Kruskal-Wallis	3.69E-15
Dimethylamine	1.08	0.46	1.41	0.22	Kruskal-Wallis	3.17E-14
Cis-Aconitate	0.45	0.24	0.17	0.11	Kruskal-Wallis	6.55E-14
3-Hydroxyisovalerate	0.97	0.45	0.46	0.28	Kruskal-Wallis	3.43E-13
3-Indoxylsulfate	0.72	0.29	0.54	0.18	Kruskal-Wallis	5.36E-12

4.3. Minería de reglas de asociación.

Con el objetivo de identificar relaciones interpretables entre patrones dietéticos y perfiles metabolómicos, se aplicó un análisis de reglas de asociación mediante el algoritmo a priori, exclusivamente sobre la base de datos que mostró el mejor rendimiento en la etapa de clustering.

El análisis de reglas se orientó a descubrir asociaciones frecuentes y robustas entre variables discretizadas (por terciles) en ambos clusters, limitando cada regla a un máximo de tres elementos por lado para facilitar la interpretación. Se aplicaron umbrales mínimos de soporte (≥ 0,10), confianza (≥ 0,60) y lift (≥ 2,0), priorizando reglas con lift ≥ 2,2 por su mayor relevancia estadística y biológica.

Posteriormente, se filtraron específicamente aquellas reglas en las que los metabolitos y las variables dietéticas estuvieran en lados opuestos de la regla, excluyendo combinaciones redundantes o difícilmente interpretables (ver Tabla 7).

Tabla 7. Reglas de asociación más relevantes entre metabolitos y variables dietéticas, por cluster.

Cluster	Regla	Lift	Interpretación biológica
0	EAT_whole_grains_0 + EAT_legumes_0 → Xanthosine_2	2,10	Posible acumulación de xantosina asociada a menor ingesta de fuentes vegetales ricas en purinas; sugiere alteraciones en el reciclaje de nucleótidos.
0	Acetate_2 + Trigonelline_2 → EAT_poultry_2	≈2,08	Perfil mixto de dieta vegetal y proteína magra; refleja una microbiota activa (acetato) y consumo de alimentos vegetales (trigonelina) en combinación con carne blanca.
1	EAT_fruits_2 + EATlancet_2 → Hippurate_2	≈2,08	Hipurato: biomarcador robusto de ingesta de frutas, polifenoles y adherencia dietética saludable; recientemente propuesto también como marcador de consumo de frutos secos (cita)
1	Isobutyrate_0 + Hippurate_2 → EAT_fruits_2 / EAT_nuts_2	2,25 / 2,04	Perfil fermentativo de la microbiota intestinal asociado a dietas ricas en fibra y compuestos fenólicos; coherente con la ingesta de frutas y nueces.
1	Urea_0 + TMAO_2 + Acetate_0 → EAT_fish_2	≈2,30	TMAO es un marcador validado de ingesta de pescado. Su combinación con urea y acetato bajos puede reflejar un patrón animal con menor fermentación colónica.
1	Valine_2 + Creatinine_2 → EAT_vegetables_2	≈2,08	Indicadores proteicos (valina, creatinina) asociados a dietas equilibradas con alto consumo de vegetales y buen estado muscular.

Estas reglas reflejan patrones coherentes con la literatura actual en metabolómica nutricional. El hipurato, en particular, destaca como un metabolito clave en este análisis. Su asociación con frutas, frutos secos y adherencia al patrón EAT-Lancet ha sido recientemente validada por Fernández-Cruz et al. (2025), quienes identificaron una relación positiva entre hipurato urinario y consumo de nueces, especialmente en hombres. Este hallazgo respalda su uso como biomarcador sexoespecífico de alimentos ricos en polifenoles.

La presencia de TMAO en una regla predictora de consumo de pescado aporta robustez metodológica, dada su amplia validación como marcador específico de alimentos marinos (DeGregory et al., 2018; Lombardo et al., 2021). Asimismo, la combinación de metabolitos como

valina y creatinina, típicamente asociados a metabolismo proteico y función renal, sugiere la existencia de perfiles metabólicos diferenciados según el patrón dietético dominante.

Aunque algunas reglas en la que aparece la xantosina requieren mayor validación empírica, su plausibilidad bioquímica y coherencia con el perfil de baja ingesta vegetal justifican su inclusión como hallazgos exploratorios.

La minería de reglas permitió identificar asociaciones robustas entre componentes de la dieta y perfiles metabolómicos característicos de los grupos generados por clustering. Estas reglas no solo refuerzan la validez de los patrones identificados, sino que también aportan candidatos plausibles a biomarcadores de adherencia dietética, abriendo vías para futuras investigaciones en nutrición de precisión y metabolómica clínica.

5. Conclusiones.

Este estudio ha permitido explorar la caracterización de perfiles clínico-dietético-metabolómicos mediante métodos de análisis no supervisado, aplicados a un conjunto de datos integrados. La estrategia combinó filtrado de variables, agrupamiento probabilístico y minería de reglas, con el objetivo de identificar patrones relacionados con la adherencia a la dieta EAT-Lancet. Las principales conclusiones derivadas del estudio son las siguientes:

1. La aplicación de filtros por varianza y correlación permitió depurar el conjunto de variables y mejorar la coherencia interna de los datos. La base de datos resultante, con variables normalizadas, ofreció mayor estabilidad en los modelos y fue seleccionada para el análisis final.
2. El modelo Gaussian Mixture Models con k = 2 agrupó a los participantes en dos perfiles diferenciados, con diferencias claras en edad, distribución por sexo, perfil lipídico, patrón dietético y niveles de metabolitos. Este método superó en rendimiento y consistencia a otras estrategias evaluadas.

3. Las reglas de asociación identificadas evidenciaron relaciones específicas entre metabolitos y grupos de alimentos, destacando el hipurato como marcador de dietas vegetales ricas en frutas y frutos secos, así como asociaciones robustas entre TMAO y pescado, o entre creatinina-valina y consumo de vegetales.
4. Aunque algunos metabolitos están estrechamente ligados a la dieta, otros pueden estar condicionados por el estado fisiológico, tratamientos farmacológicos o características individuales del metabolismo. Esto debe tenerse en cuenta al interpretar los resultados y al considerar su posible uso como biomarcadores.
5. Los resultados apoyan el potencial de combinar clustering y minería de reglas para avanzar hacia una caracterización más precisa y personalizada de los patrones dietéticos, con aplicaciones en nutrición de precisión y en el desarrollo de herramientas predictivas basadas en datos metabolómicos.

El estudio presenta limitaciones como el tamaño moderado de la muestra y el diseño transversal, que dificultan establecer causalidad entre dieta, metabolismo y salud. Aunque se consideró la actividad física y variables clínicas clave, no se incluyeron otros factores externos como medicación o hábitos de sueño. Futuras investigaciones deberían ampliar cohortes, incluir datos longitudinales y aplicar enfoques multi-ómicos para capturar mejor las complejas interacciones. Además, el desarrollo de modelos predictivos integrados será clave para optimizar intervenciones nutricionales personalizadas.

6. Bibliografía.

Azmi, S., Kunnathodi, F., Alotaibi, H. F., Alhazzani, W., Mustafa, M., Ahmad, I., Anvarbatcha, R., Lytras, M. D., & Arafat, A. A. (2025). Harnessing Artificial Intelligence in Obesity Research and Management: A Comprehensive Review. *Diagnostics*, *15*(3), 396. https://doi.org/10.3390/diagnostics15030396

De La O, V., Fernández-Cruz, E., Valdés, A., Cifuentes, A., Walton, J., & Martínez, J. A. (2025). Exhaustive Search of Dietary Intake Biomarkers as Objective Tools for Personalized Nutrimetabolomics and Precision Nutrition Implementation. *Nutrition Reviews*, *83*(5), 925-942. https://doi.org/10.1093/nutrit/nuae133

DeGregory, K. W., Kuiper, P., DeSilvio, T., Pleuss, J. D., Miller, R., Roginski, J. W., Fisher, C. B., Harness, D., Viswanath, S., Heymsfield, S. B., Dungan, I., & Thomas, D. M. (2018). A review of machine learning in obesity. *Obesity Reviews*, *19*(5), 668-685. https://doi.org/10.1111/obr.12667

Fernández-Cruz, E., De La O, V., Fernández-Diaz, C. M., Matía-Martín, P., Rubio-Herrera, M. Á., Amigó, N., Calle-Pascual, A. L., & Martínez, J. A. (2025). Urinary Hippuric Acid as a Sex-Dependent Biomarker for Fruit and Nut Intake Raised from the EAT-Lancet Index and Nuclear Magnetic Resonance Analysis. *Metabolites*, *15*(6), 348. https://doi.org/10.3390/metabo15060348

Guan, V. X., Probst, Y. C., Neale, E. P., Batterham, M. J., & Tapsell, L. C. (2018). Identifying usual food choices at meals in overweight and obese study volunteers: Implications for dietary advice. *British Journal of Nutrition*, *120*(4), 472-480. https://doi.org/10.1017/S0007114518001587

Kirk, D., Kok, E., Tufano, M., Tekinerdogan, B., Feskens, E. J. M., & Camps, G. (2022). Machine Learning in Nutrition Research. *Advances in Nutrition*, *13*(6), 2573-2589. https://doi.org/10.1093/advances/nmac103

Lombardo, M., Aulisa, G., Marcon, D., Rizzo, G., Tarsisano, M. G., Di Renzo, L., Federici, M., Caprio, M., & De Lorenzo, A. (2021). Association of Urinary and Plasma Levels of Trimethylamine N-Oxide (TMAO) with Foods. *Nutrients*, *13*(5), Article 5. https://doi.org/10.3390/nu13051426

Muthukumar, K.A., Gupta, S. & Saikia, D. Leveraging machine learning techniques to analyze nutritional content in processed foods. *Discov Food* 4, 182 (2024).

Patel M, Patel DA, Gajra B. Validation of Analytical Procedures: Methodology ICH-Q2B. Int J Pharm Innov. 2011;1(2):45–50.

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526-534. https://doi.org/10.1016/j.procs.2021.01.199

Stubbendorff, A., Sonestedt, E., Ramne, S., Drake, I., Hallström, E., & Ericson, U. (2022). Development of an EAT-Lancet index and its relation to mortality in a Swedish population. *The American Journal of Clinical Nutrition*, *115*(3), 705-716. https://doi.org/10.1093/ajcn/nqab369

Wang, Z., Da Cunha, C., Ritou, M., & Furet, B. (2019). Comparison of K-means and GMM methods for contextual clustering in HSM. *Procedia Manufacturing*, *28*, 154-159. https://doi.org/10.1016/j.promfg.2018.12.025

Willett, W., Rockström, J., Loken, B., Springmann, M., Lang, T., Vermeulen, S., Garnett, T., Tilman, D., DeClerck, F., Wood, A., Jonell, M., Clark, M., Gordon, L. J., Fanzo, J., Hawkes, C., Zurayk, R., Rivera, J. A., Vries, W. D., Sibanda, L. M., … Murray, C. J. L. (2019). Food in the Anthropocene: The EAT–Lancet Commission on healthy diets from sustainable food systems. *The Lancet*, *393*(10170), 447-492. https://doi.org/10.1016/S0140-6736(18)31788-4

Anexo

Tabla A1. Estadísticos descriptivos de las variables del estudio. Se presentan el número de valores perdidos (NA), mínimos, máximos, medias, desviaciones estándar (STD), medianas e intervalos intercuartílicos (IQR) de cada variable incluida en el análisis, clasificadas por bloque temático.

Variable	Clase	NA	Mín	Máx	Media	STD	Mediana
id	Otra	0	1	150	73,8	43,8	71,5
sex	Sociodemog	0	0	1	0,7	0,5	1
age	Sociodemog	0	17	79	43	14,2	43
minutes_pa_tot_wk	Actividad f	0	0	2940	662,2	481,3	525
glucose	Clínica	1	76	497	99,1	37,8	93
cholesterol	Clínica	1	1	279	187,8	39,3	185
cholesterol_hdl	Clínica	1	31	156	62,3	17	60
cholesterol_ldl_calcu	Clínica	2	43	186	108,9	30,2	107
choleste-rol_not_hdl_calcu	Clínica	1	69	215	127,4	32,7	125
EAT_whole_grains	Dieta	1	0	200	47,4	63,7	11,4
EAT_potatoes	Dieta	0	0	300	45,6	39	31,4
EAT_vegetables	Dieta	0	0	1414,2	554,9	301,7	473,1
EAT_fruits	Dieta	0	0	2865,2	307	316,4	259,9
EAT_dairy	Dieta	0	0	1062,8	388,9	228,1	348,9
EAT_beef_lamb	Dieta	0	0	281,2	35,6	37,6	21,4
EAT_pork	Dieta	0	0	264	76,8	45,2	69
EAT_poultry	Dieta	0	0	375	61,6	42,3	64,3
EAT_eggs	Dieta	0	0	300	39,2	43,9	25,7
EAT_fish	Dieta	0	0	263	90,6	52,1	81,6
EAT_legumes	Dieta	0	0	102,3	23,1	15,7	18,6
EAT_nuts	Dieta	0	0	125	25,2	24,8	21,4
EAT_unsaturated_oils	Dieta	0	0	675	93,9	82	66,7
EAT_added_sugar	Dieta	0	0	50	7,3	11	1,4
EATlancet	Dieta	0	10,666	35	24	3,5	24
bmi	Sociodemog	1	1,765	45,3	26,6	6,7	25,4
Acetate	Metabolómica	28	0,007	1,7	0,1	0,2	0
Alanine	Metabolómica	4	0,041	2,8	0,5	0,4	0,4
Creatinine	Metabolómica	0	3,653	61,5	19,1	9,8	17,7
Trigonelline	Metabolómica	5	0,056	2,4	0,4	0,4	0,3
Urea	Metabolómica	2	1,042	6,2	3	0,9	2,9
Xanthosine	Metabolómica	23	0,057	0,6	0,2	0,1	0,2
Uracil	Metabolómica	39	0,015	0,8	0,1	0,1	0
Cis-Aconitate	Metabolómica	34	0,038	1,2	0,3	0,2	0,3

Hippurate	Metabolómica	5	0,344	22,8	5,3	4,8	3,6
Citrate	Metabolómica	0	0,161	16,5	4	2,8	3,3
Glycine	Metabolómica	3	0,152	10,7	2,2	1,7	1,6
Glycolate	Metabolómica	53	0,232	9,5	1,5	1,4	1,1
3-Indoxylsulfate	Metabolómica	14	0,04	3,1	0,5	0,4	0,3
3-Aminoisobutyrate	Metabolómica	20	0,026	2,8	0,3	0,5	0,2
Glucose	Metabolómica	58	0,773	481,9	14	72	1,6
Galactose	Metabolómica	75	0,048	0,7	0,2	0,1	0,2
Dimethylamine	Metabolómica	1	0,138	10	0,7	0,9	0,5
Formate	Metabolómica	8	0,029	0,7	0,3	0,1	0,2
Taurine	Metabolómica	66	0,163	10,2	2,5	2,1	1,9
Trimethylami-neNoxide	Metabolómica	35	0,03	32,6	1,9	4,2	0,7
Betaine	Metabolómica	94	0,013	2,2	0,1	0,3	0,1
3-Methyl-2-oxovale-rate	Metabolómica	29	0,019	0,8	0,2	0,1	0,2
Methylsuccinate	Metabolómica	71	0,004	0,2	0	0	0
3-Hydroxyisobutyrate	Metabolómica	6	0,029	0,6	0,2	0,1	0,2
Isobutyrate	Metabolómica	14	0,003	0,1	0	0	0
2-Hydroxyisobutyrate	Metabolómica	7	0,019	0,3	0,1	0	0,1
3-Hydroxyisovalerate	Metabolómica	5	0,059	2,3	0,6	0,4	0,5
Phenylalanine	Metabolómica	10	0,305	10,6	1,8	1,3	1,5
4-Hydroxyphenyllac-tate	Metabolómica	5	0,042	2,1	0,4	0,3	0,3
Tyrosine	Metabolómica	21	0,098	0,7	0,2	0,1	0,2
Valine	Metabolómica	7	0,008	0,3	0,1	0	0,1
Isoleucine	Metabolómica	43	0,005	0,1	0	0	0
Alloisoleucine	Metabolómica	35	0,002	0,3	0	0	0
Leucine	Metabolómica	22	0,005	0,2	0,1	0	0

Tabla A2. Evaluación de métodos de clustering sobre las bases A y B. A: original; B: con ingeniería de variables). Se presenta el número de clusters (k), el valor medio del índice de silueta su desviación estándar y el número medio de observaciones consideradas outliers cuando aplica.

dataset	metodo	k	media_score	desviacion	media_outliers
A	agglo	2	0,133378		
	agglo	3	0,135763		
	agglo	4	0,130345		
	agglo	5	0,127717		
	agglo	6	0,130396		
	dbscan	0			138
	gmm	2	0,209365	0,095867	
	gmm	3	0,094319	0,040515	
	gmm	4	0,046996	0,008894	
	gmm	5	0,042391	0,013166	
	gmm	6	0,035573	0,011972	
	hdbscan	0			138
	hdbscan	2	0,024806		111,5
	kmeans	2	0,146194	0,001427	
	kmeans	3	0,060834	0,031929	
	kmeans	4	0,066661	0,0301	
	kmeans	5	0,043267	0,007635	
	kmeans	6	0,044627	0,013412	
B	agglo	2	0,122636		
	agglo	3	0,124437		
	agglo	4	0,096004		
	agglo	5	0,099986		
	agglo	6	0,102554		
	dbscan	0			138
	gmm	2	0,209228	0,094714	
	gmm	3	0,123839	0,038205	
	gmm	4	0,064497	0,026848	
	gmm	5	0,045093	0,021287	
	gmm	6	0,0347	0,008719	
	hdbscan	0			138
	hdbscan	2	0,042162		88
	kmeans	2	0,151024	0	
	kmeans	3	0,106294	0,050568	
	kmeans	4	0,058254	0,01905	
	kmeans	5	0,041908	0,014095	
	kmeans	6	0,044393	0,015985	

Tabla A3. Comparación de los niveles medios de metabolitos entre los grupos identificados por el algoritmo GMM (k=2). Se muestran la media y desviación estándar en cada cluster, así como el test estadístico aplicado y su p-valor. Solo se incluyen los metabolitos con diferencias significativas ($p < 0.05$).

Variable	Media C0	Media C1	SD C0	SD C1	Test	p-valor
Xanthosine	0.33	0.19	0.08	0.07	ANOVA	1.21E-18
Creatinine	27.94	13.92	8.63	6.10	Kruskal-Wallis	4.61E-17
Valine	0.11	0.05	0.04	0.02	Kruskal-Wallis	6.65E-17
Leucine	0.08	0.04	0.03	0.01	Kruskal-Wallis	9.66E-16
3-Methyl-2-oxovalerate	0.34	0.18	0.14	0.06	Kruskal-Wallis	1.68E-15
3-Hydroxyisobutyrate	0.28	0.13	0.11	0.06	Kruskal-Wallis	3.69E-15
Dimethylamine	1.08	0.46	1.41	0.22	Kruskal-Wallis	3.17E-14
Cis-Aconitate	0.45	0.24	0.17	0.11	Kruskal-Wallis	6.55E-14
3-Hydroxyisovalerate	0.97	0.45	0.46	0.28	Kruskal-Wallis	3.43E-13
3-Indoxylsulfate	0.72	0.29	0.54	0.18	Kruskal-Wallis	5.36E-12
2-Hydroxyisobutyrate	0.11	0.06	0.04	0.03	Kruskal-Wallis	8.46E-12
Alanine	0.75	0.35	0.46	0.18	Kruskal-Wallis	5.17E-11
4-Hydroxyphenyllactate	0.52	0.28	0.31	0.15	Kruskal-Wallis	1.14E-10
Alloisoleucine	0.02	0.01	0.03	0.01	Kruskal-Wallis	1.25E-10
Glycine	3.20	1.52	1.95	1.15	Kruskal-Wallis	2.07E-09
Isoleucine	0.05	0.03	0.03	0.01	Kruskal-Wallis	2.11E-08
Formate	0.36	0.22	0.15	0.11	Kruskal-Wallis	3.05E-08
Citrate	5.72	3.05	3.22	1.84	Kruskal-Wallis	3.56E-08
Isobutyrate	0.02	0.01	0.01	0.01	Kruskal-Wallis	4.33E-08
Phenylalanine	2.51	1.39	1.63	0.69	Kruskal-Wallis	4.71E-08
Tyrosine	0.31	0.20	0.14	0.06	Kruskal-Wallis	1.23E-06
Trigonelline	0.58	0.31	0.46	0.23	Kruskal-Wallis	1.59E-05
Hippurate	7.54	3.95	6.03	3.20	Kruskal-Wallis	4.37E-05
3-Aminoisobutyrate	0.46	0.19	0.65	0.20	Kruskal-Wallis	9.98E-04
TrimethylamineNoxide	2.74	0.96	5.67	1.25	Kruskal-Wallis	2.32E-03
Uracil	0.10	0.05	0.13	0.04	Kruskal-Wallis	1.70E-02