MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA Y ANÁLISIS DE DATOS BIOMÉDICOS

"Modelos de Predicción y Clasificación a la adherencia a la dieta EAT Lancet usando técnicas de aprendizaje automático aplicados a datos clínicos, dietéticos y metabolómicos en población adulta"

Guillermo Encinas Sabater

Curso académico *2024/2025*

DECLARACIÓN PERSONAL DE NO PLAGIO

Yo, Don Guillermo Encinas Sabater (nombre y apellidos) con NIF/NIE 02573960F , 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 "Modelos de Predicción y Clasificación a la adherencia a la dieta EAT Lancet usando técnicas de aprendizaje automático aplicados a datos clínicos, dietéticos y metabolómicos en población adulta", 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 (Localidad), a 03 (fecha) de Julio (mes) de 2025 (año)

Fdo.

Trabajo de Fin de Máster experimental

Modelos de Predicción y Clasificación a la adherencia a la dieta EAT Lancet usando técnicas de aprendizaje automático aplicados a datos clínicos, dietéticos y metabolómicos en población adulta

Guillermo Encinas Sabater[1], Dr. Jose Aguilar[2], Dr. J. Alfredo Martínez[3], Edwin Fernández Cruz[3]

[1] Universidad Francisco de Vitoria 1 ; guillermoesabater@gmail.com

[2] IMDEA Networks ; aguilarjos@gmail.com

[3] IMDEA Nutrición ; jalfredo.martinez@imdea.org, edwin.fernandez@nutricion.imdea.org

***** Academic tutor – Ramiro Perezzan Rodríguez

** Institution/Company Tutor – Dr. Jose Aguilar

Abstract

Antecedentes y Objetivos:
El modelo alimentario propuesto por la Comisión EAT-Lancet ha despertado gran interés por su potencial para promover dietas saludables y sostenibles. Sin embargo, su implementación individual y la evaluación precisa de la adherencia continúan siendo retos importantes, especialmente debido a la subjetividad de los métodos de medición convencionales y la introducción de sesgos que limitan la precisión. Este trabajo propone un enfoque basado en aprendizaje automático para predecir la adherencia al patrón EAT-Lancet, integrando variables clínicas, dietéticas y metabolómicas mediante un proceso sistemático de selección de características.

Materiales y Métodos:
Se analizaron 54 variables organizadas en bloques funcionales. Se aplicaron algoritmos de clasificación y regresión (Random Forest, KNN, regresión logística y árboles de decisión) combinados con filtrado previo por correlación de Spearman ($\rho \geq 0.20, 0.25, 0.30$). La evaluación de los modelos incluyó métricas como R^2, RMSE, MAPE, Accuracy y F1-score, con validación cruzada.

Resultados:
El modelo Random Forest con un umbral de $\rho \geq 0.30$ fue el que obtuvo el mejor rendimiento global. Entre las variables más relevantes se encontraron tanto alimentos clave del patrón EAT-Lancet, como frutas, legumbres, cerdo y frutos secos.

Además de varios metabolitos urinarios, especialmente alanina, leucina, 3-methyl-2-oxovalerate y 3-hydroxyisobutyrate. En cambio, los biomarcadores clínicos tradicionales, como glucosa o colesterol, mostraron una contribución mucho más limitada al poder predictivo del modelo.

Conclusiones:

La combinación de aprendizaje automático con técnicas de ingeniería de características mejora significativamente la predicción de la adherencia dietética en contextos multivariantes. Estos hallazgos refuerzan el valor de los metabolitos como biomarcadores objetivos y aportan una base sólida para avanzar hacia estrategias de nutrición personalizada y salud pública más precisas y basadas en datos.

Keywords: EAT-Lancet; machine learning; ingeniería de características; metabolómica; adherencia dietética; Random Forest; nutrición personalizada

1. Introducción

En las últimas décadas, la dieta ha cobrado una importancia creciente, no solo desde la perspectiva de la salud individual, sino también desde el punto de vista medioambiental. En este contexto, la Comisión EAT-Lancet propuso un modelo alimentario sostenible y saludable, basado principalmente en alimentos de origen vegetal, con una reducción considerable del consumo de productos animales, especialmente carnes rojas (Stubbendorff et al., 2022).

Este patrón dietético ha sido reconocido por su potencial para disminuir la incidencia de enfermedades crónicas, como la diabetes tipo 2, enfermedades cardiovasculares y determinados tipos de cáncer, además de contribuir positivamente a la mitigación del impacto ambiental asociado a la producción alimentaria (Stubbendorff et al., 2022).

Pese a las ventajas documentadas de este modelo dietético, la evaluación precisa de la adherencia individual continúa siendo un desafío metodológico significativo. Habitualmente, las herramientas utilizadas para medir la dieta, como los cuestionarios de frecuencia de consumo, acaban introduciendo sesgos que limitan la precisión de los resultados. Debido a estas limitaciones, resulta necesario explorar métodos más objetivos y fiables que permitan evaluar el grado real de cumplimiento con las recomendaciones dietéticas del modelo EAT-Lancet.

1.1 Aprendizaje automático

En los últimos años, las técnicas de aprendizaje automático (machine learning) se han consolidado como herramientas fundamentales en la investigación nutricional debido a su capacidad para identificar patrones complejos en grandes conjuntos de datos. Estos algoritmos permiten integrar simultáneamente múltiples variables clínicas, dietéticas y metabolómicas, aprendiendo de forma automática a partir de datos previos

Usando el aprendizaje automático, es posible definir modelos capaces de detectar de forma precisa y objetiva perfiles dietéticos específicos y evaluar sus implicaciones en la salud de los individuos, ofreciendo resultados difícilmente alcanzables mediante métodos analíticos tradicionales. Diversos estudios han aplicado estas técnicas en el ámbito de la nutrición y la obesidad, mostrando su utilidad para predecir riesgos y clasificar perfiles clínicos complejos a partir de grandes bases de datos poblacionales como NHANES (DeGregory et al., 2018). Más recientemente, se ha demostrado su aplicabilidad también en la predicción de la retención de nutrientes tras el procesamiento de alimentos vegetales, comparando modelos como Random Forest y SVR, y aplicando técnicas avanzadas de selección de variables para mejorar la precisión de los modelos (Muthukumar et al., 2024)

No obstante, el uso de datos multivariantes y con alta dimensionalidad plantea importantes retos metodológicos, especialmente a la hora de distinguir entre variables relevantes para la predicción y aquellas que aportan ruido o redundancia. Para abordar esta complejidad y mejorar la capacidad explicativa de los modelos, es fundamental aplicar técnicas específicas de selección, transformación y reducción de variables, integradas dentro del marco de la *ingeniería de características*. En este contexto, algoritmos como Random Forest son especialmente útiles, ya que además de ofrecer alta precisión en clasificación y predicción, permiten estimar la importancia relativa de cada variable, mejorando así la interpretabilidad del modelo en contextos clínico-nutricionales.

1.2 Ingeniería de características

La ingeniería de características es fundamental para optimizar modelos predictivos en estudios con datos complejos y multivariantes, como los que combinan variables clínicas, dietéticas y metabolómicas. Este proceso implica seleccionar, transformar o reducir variables con el objetivo de mejorar tanto el rendimiento predictivo como la interpretabilidad clínica y nutricional de los resultados.

En este trabajo, la técnica utilizada fue la selección mediante correlación de Spearman, particularmente eficaz para datos heterogéneos y relaciones no lineales. Se aplicaron distintos umbrales específicos ($\rho \geq 0.30$, 0.25 y 0.20) con el fin de comparar la influencia de estos en el rendimiento de los modelos. Se identificaron y priorizaron variables relevantes, logrando reducir significativamente la dimensionalidad de la base de datos original. Este enfoque mejoró notablemente el desempeño predictivo del modelo y facilitó la identificación de factores clave relacionados con la adherencia dietética.

1.3 Controversias, retos y propósito

Aunque el modelo alimentario propuesto por la Comisión EAT-Lancet cuenta con amplio respaldo científico, existen controversias sobre su aplicabilidad individual y la precisión de los métodos empleados para medir su adherencia. La variabilidad interindividual y la complejidad de integrar datos clínicos, dietéticos y metabolómicos suponen retos aún no resueltos. En este contexto, el objetivo de este trabajo es desarrollar y evaluar modelos de predicción y clasificación de la adherencia a la dieta EAT-Lancet mediante técnicas de aprendizaje automático, identificando las variables más relevantes a través de ingeniería de características. Los resultados obtenidos refuerzan el valor de este enfoque para avanzar en la nutrición personalizada y apoyar estrategias de salud pública.

2. Objetivos

2.1 Objetivo general

Evaluar la adherencia a la dieta EAT-Lancet mediante el desarrollo y comparación de modelos predictivos y clasificatorios basados en aprendizaje automático, integrando variables clínicas, dietéticas y metabolómicas, usando técnicas de ingeniería de características para optimizar la selección de las variables predictoras.

2.2 Objetivos específicos

1. Analizar y aplicar las técnicas de ingeniería de características para optimizar el conjunto de variables predictoras a ser empleadas en los modelos.
2. Desarrollar modelos de regresión y clasificación capaces de estimar el nivel de adherencia a la dieta EAT-Lancet, tanto en su forma numérica (EAT score) como categórica (alta/baja adherencia) usando las variables optimizadas por la ingeniería de características.
3. Comparar el rendimiento de los modelos predictivos y clasificatorios bajo diferentes criterios considerados durante el proceso de ingeniería de características.
4. Identificar las variables clínicas, dietéticas y metabolómicas más relevantes asociadas a la adherencia, y analizar su importancia relativa en los modelos generados.

3. Materiales y métodos

3.1 Fuente y estructura del conjunto de datos

El conjunto de datos utilizado en este estudio está compuesto por 54 variables, organizadas en cuatro bloques según su función analítica. Incluye información clínica básica, indicadores bioquímicos y dietéticos, perfiles metabolómicos obtenidos por análisis de laboratorio, y dos variables que evalúan el grado de adherencia al EAT Lancet. Esta estructura permite abordar el análisis desde una perspectiva integral que combina datos sociodemográficos, nutricionales y moleculares. Como primer paso en la exploración del conjunto de datos, se elaboró un diccionario descriptivo que permitió resumir las características estadísticas clave de cada variable (ver sección 4). Se calcularon medidas de tendencia central y dispersión, se identificaron valores atípicos mediante criterios robustos, y se evaluaron los valores faltantes y aquellos por debajo del límite de detección (LOD), especialmente en variables metabolómicas. Esta síntesis inicial resultó esencial para orientar con solidez las etapas posteriores de imputación, transformación y selección de variables.

3.2 Flujo metodológico general

El procedimiento seguido se basó en la metodología CRISP-DM. Primero, se realizó la exploración y depuración del conjunto de datos. Los valores faltantes se imputaron por regresión (cuando existía alta correlación con otras variables) o mediana. Los valores atípicos se detectaron, pero no fueron modificados. Posteriormente, se aplicaron técnicas de ingeniería de características usando correlación de Spearman con diferentes umbrales ($\rho \geq 0.30, 0.25, 0.20$) para reducir la dimensionalidad y optimizar la selección de predictores. A continuación, se desarrollaron modelos de regresión y clasificación utilizando algoritmos de aprendizaje automático, principalmente Random Forest, y se evaluó su rendimiento mediante validación cruzada con métricas como R^2, RMSE, MAPE, exactitud (accuracy) y F1-score según sea el tipo de modelo a desarrollar (predictivo o de clasificación). Finalmente, se compararon los modelos generados para determinar el umbral de selección más eficiente (proceso descrito en Figura 1).

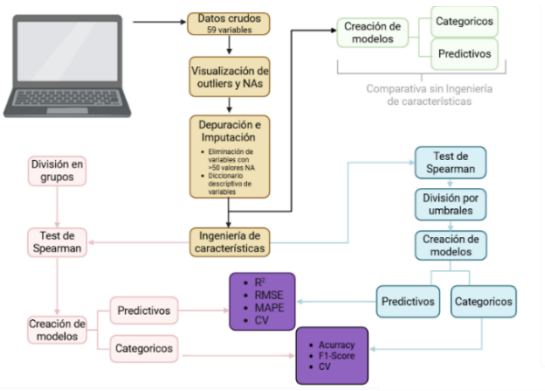

Figura 1. *Esquema del flujo metodológico.* El estudio incluyó depuración e imputación de datos, ingeniería de características por correlación de Spearman, y desarrollo de modelos predictivos y clasificatorios.

4. Resultados y discusión

4.1 Clasificación de variables y depuración

Como se comentó en la sección anterior, para facilitar su análisis, el conjunto de datos fue organizado en cuatro bloques analíticos principales: variables clínicas, colesterol y dieta EAT-Lancet, metabolómicas y variables objetivo. El grupo clínico incluyó información sociodemográfica y antropométrica básica, como el identificador del paciente (id), sexo, edad, minutos de actividad física semanal (minutes_pa_tot_wk) e índice de masa corporal (BMI). El bloque denominado colesterol y EAT-Lancet agrupó tanto biomarcadores bioquímicos clásicos (glucosa, colesterol total, HDL, LDL calculado y colesterol no-HDL calculado) como indicadores dietéticos derivados del patrón alimentario EAT-Lancet. Entre ellos se encuentran variables como EAT_fruits, EAT_legumes, EAT_dairy, EAT_pork o EAT_fish, las cuales indican si se ingiere frutas, legumbres, y otros grupos alimentarios, sumando un total de 19 predictores. Por su parte, el grupo metabolómico estuvo compuesto por 28 metabolitos urinarios, incluyendo aminoácidos, compuestos nitrogenados, ácidos orgánicos intermedios y otros derivados volátiles. Entre los metabolitos identificados se encuentran alanina, un aminoácido no esencial y de pequeño tamaño con propiedades hidrófobas, implicado en la síntesis de proteínas y el metabolismo energético; leucina, un aminoácido esencial de cadena ramificada, fundamental para la síntesis proteica y el mantenimiento de la masa muscular; 3-metil-2-oxovalerato, un compuesto que actúa como intermediario en la degradación de la isoleucina, reflejando el metabolismo de aminoácidos de cadena ramificada; 3-hidroxiisobutirato, derivado del catabolismo de la valina y vinculado a procesos mitocondriales y producción energética; y citrato, un metabolito central en el ciclo del ácido cítrico, esencial para la generación de energía en la célula. Finalmente, se incluyeron dos variables objetivo: EATlancet, que representa la puntuación continua de adherencia al patrón dietético, y EAT_lancet_adherencia, su versión binaria (alta vs. no alta adherencia) utilizada en los modelos de clasificación.

Durante la fase de depuración, se eliminaron seis metabolitos —Glycolate, Glucose, Galactose, Taurine, Betaine y Methylsuccinate— por presentar más de 50 valores ausentes, lo que superaba el umbral definido para su inclusión en el análisis. Asimismo, se excluyeron 12 sujetos cuya información estaba completamente vacía (todos los valores eran NA), imposibilitando su participación en cualquier fase del modelado.

4.2 Desempeño de los modelos según selección de variables y grupos predictivos

Se evaluó el rendimiento de distintos modelos predictivos y clasificatorios aplicados a la estimación de la adherencia a la dieta EAT-Lancet, comparando configuraciones con y sin

ingeniería de características. Para ello, se entrenaron modelos tanto con todas las variables combinadas como de forma independiente cada bloque (clínicas, colesterol y dieta EAT-Lancet, y metabolitos), utilizando diferentes umbrales de selección de variables según la correlación de Spearman de la variable a la variable objetivo. Esta estrategia permitió identificar qué enfoque ofrecía mejores resultados y qué tipo de variables contribuía más eficazmente a la predicción dentro del contexto del modelado.

La calidad de los modelos se evaluó empleando métricas estándar para regresión y clasificación. En el caso de predicción, se utilizaron el coeficiente de determinación (R^2), el error cuadrático medio (RMSE) y el error porcentual absoluto medio (MAPE), que permiten valorar conjuntamente la capacidad explicativa, la precisión en unidades reales y el error relativo. Para los modelos de clasificación se utilizaron el accuracy y el F1-score, siendo este último especialmente relevante en situaciones de posible desbalance entre clases. Todas las métricas se calcularon tanto sobre los datos de entrenamiento como mediante validación cruzada (CV), con el objetivo de estimar la estabilidad y generalizabilidad de los modelos. Además, en una fase inicial se incluyeron tres modelos adicionales —regresión logística, K-Nearest Neighbors (KNN) y árbol de decisión— con el fin de comprobar que los resultados obtenidos con Random Forest no fueran producto de un sobreajuste. El rendimiento inferior y menos estable de estos modelos confirmó que el algoritmo principal era robusto y no mostraba indicios de overfitting.

Modelo	R^2	CV R^2	RMSE	CV RMSE	MAPE	CV MAPE	Accuracy	CV Accuracy	F1-Score	CV F1 Score
Umbral 0.20 - RF Predicción	0,475	0,593	2,312	2,143	0,08	0,071				
Umbral 0.20 - RF Clasificación							1	1	1	1
Umbral 0.20 - Logistic Regression							1	1	1	1
Umbral 0.20 - KNN (k=3)							0,964	0,935	0,974	0,924
Umbral 0.20 - Decision Tree							1	1	1	1
Umbral 0.25 - RF Predicción	0,459	0,607	2,346	2,114	0,079	0,07				
Umbral 0.25 - RF Clasificación							1	1	1	1
Umbral 0.25 - Logistic Regression							1	1	1	1
Umbral 0.25 - KNN (k=3)							1	0,957	1	0,951
Umbral 0.25 - Decision Tree							1	1	1	1
Umbral 0.30 - RF Predicción	0,618	0,633	1,972	2,057	0,071	0,068				
Umbral 0.30 - RF Clasificación							1	1	1	1
Umbral 0.30 - Logistic Regression							1	1	1	1
Umbral 0.30 - KNN (k=3)							0,964	0,971	0,974	0,967
Umbral 0.30 - Decision Tree							1	1	1	1

Tabla 1. Evaluación del rendimiento de los modelos con ingeniería de características

Sin Ing. Caracteristicas				
Grupo	R2	RMSE	MAPE	Accuracy
Colesterol_EAT	-0,17171	3,997431765	0,152	0,535714286
Clínicos	-0,19557	4,037938831	0,149	0,571428571
Metabolitos	-0,12137	3,910624666	0,148	0,535714286
Total	-0,0666	3,813917184	0,139	0,535714286

Tabla 2. Evaluación del rendimiento de los modelos sin ingeniería de características

La *tabla 1* muestra una comparación del rendimiento de los modelos utilizados para predecir y clasificar la adherencia a la dieta EAT-Lancet, según los tres umbrales de selección de variables por correlación de Spearman (≥ 0.20, 0.25 y 0.30), junto con los resultados obtenidos sin aplicar ingeniería de características. En los modelos de predicción, el mejor desempeño general se alcanzó con Random Forest y un umbral ≥ 0.30, logrando un R^2 de 0.618 y un R^2 con validación cruzada de 0.633, además de los valores más bajos de error (RMSE = 1.972; CV RMSE = 2.057). Esto refleja tanto una alta capacidad explicativa como una buena estabilidad del modelo. En clasificación, todos los modelos con ingeniería de características mostraron métricas elevadas de F1-score y precisión, destacando nuevamente el umbral 0.30 como el más eficaz, especialmente Random Forest, arboles de decisión y regresión logística. En cambio, los modelos que no incluyeron una fase previa de selección de variables (ingeniería de características) incluidos en la *tabla 2*, ofrecieron un rendimiento muy inferior, con valores de R^2 negativos y una precisión generalizada en torno a 0.53. Estos resultados confirman que la aplicación de ingeniería de características fue determinante para mejorar tanto la capacidad predictiva como la clasificatoria de los modelos utilizados.

Grupo	***Umbral***	***R^2***	***R^2 CV***	***RMSE***	***RMSE CV***	***MAPE***	***MAPE CV***	***Accuracy***	***Accuracy CV***	***F1***	***F1 CV***
CLINICAS	0,2	-0,064	-0,282	3,289	3,723	0,115	0,124	0,643	0,631	0,773	0,73
CLINICAS	0,25	-0,143	-0,368	3,409	3,831	0,12	0,13	0,429	0,572	0,579	0,685
CLINICAS	0,3	-0,143	-0,368	3,409	3,831	0,12	0,13	0,429	0,572	0,579	0,685
COLESTEROL_EAT	0,2	0,125	0,437	2,983	2,529	0,097	0,084	0,821	0,826	0,872	0,882
COLESTEROL_EAT	0,25	0,247	0,461	2,767	2,485	0,096	0,085	0,786	0,797	0,85	0,861
COLESTEROL_EAT	0,3	0,32	0,383	2,63	2,64	0,093	0,092	0,75	0,819	0,821	0,872
METABOLITOS	0,2	-0,067	-0,052	3,295	3,42	0,115	0,119	0,679	0,63	0,791	0,74
METABOLITOS	0,25	-0,05	-0,097	3,268	3,484	0,12	0,124	0,571	0,638	0,7	0,751
METABOLITOS	0,3	0,024	-0,055	3,15	3,422	0,115	0,121	0,643	0,667	0,75	0,772

Tabla 3. Rendimiento por grupo de variables y efecto del umbral de selección tras aplicación de ingeniería de características

La *tabla 3* muestra el rendimiento de los modelos predictivos y clasificatorios aplicados por separado a los tres bloques de variables: clínicas, colesterol y dieta EAT-Lancet (Colesterol_EAT), y metabolitos. El análisis indica que el grupo Colesterol_EAT fue el que ofreció mejores resultados globales en ambos tipos de modelos. En el caso de predicción, este grupo alcanzó un R^2 máximo de 0.32 con un umbral de Spearman ≥ 0.30, acompañado de un RMSE tan bajo como 2.63. En clasificación, también presentó los valores más altos de F1-score (hasta 0.882) y precisiones superiores al 80%. Por el contrario, los modelos

construidos únicamente con variables clínicas o metabolómicas mostraron un rendimiento claramente inferior, especialmente en predicción, donde los valores de R^2 fueron negativos o cercanos a cero. Tal como se observó en la *tabla 2*, los modelos que no aplicaron ingeniería de características también mostraron un poder predictivo muy limitado, lo que vuelve a subrayar la importancia de seleccionar adecuadamente las variables. Además, estos resultados sugieren que la combinación de información clínica con predictores dietéticos bien definidos resultan útiles para obtener los patrones asociados a la adherencia del patrón EAT-Lancet de forma más efectiva.

4.3 Importancia relativa de las variables en los modelos optimizados

Una vez identificados los modelos con mejor rendimiento, se analizó qué variables aportaban más al proceso predictivo a través de la medida de importancia relativa obtenida con Random Forest. Este análisis permitió identificar cuáles fueron los predictores más relevantes para estimar la adherencia a la dieta EAT-Lancet, tanto en los modelos de predicción como en los de clasificación. A continuación, se detallan los resultados obtenidos en función del umbral de selección aplicado en cada caso.

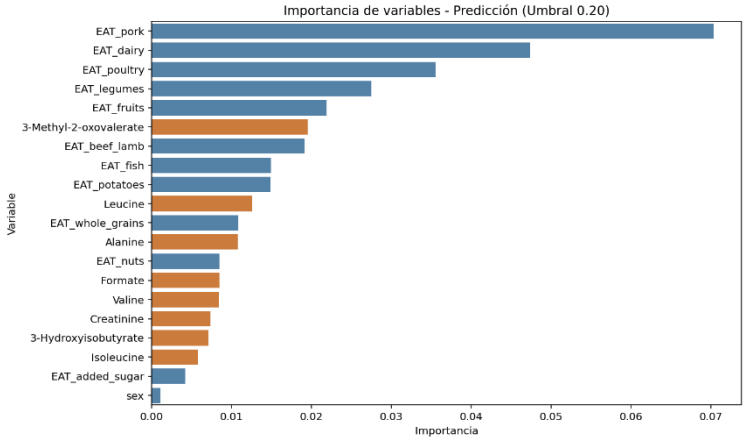

Figura 2. Importancia de variables en el modelo para la predicción del score EAT-Lancet utilizando todas las variables

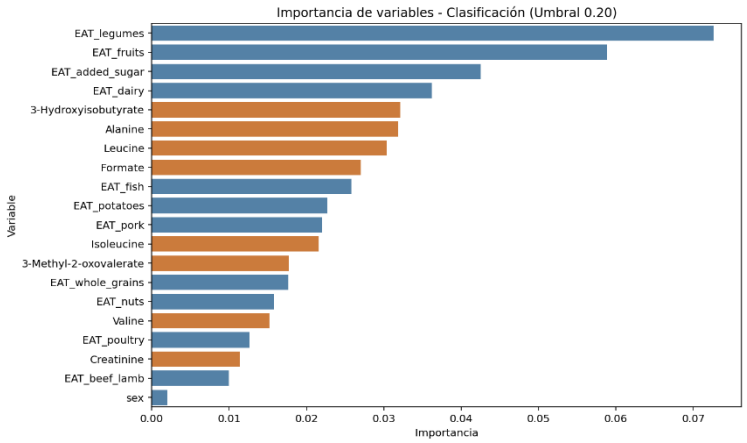

Figura 3. Importancia de variables en el modelo de clasificación de la adherencia al patrón EAT-Lancet (alta vs. No alta),

Las *Figuras 2 y 3* presentan la importancia relativa de las variables predictoras en los modelos de predicción y clasificación construidos con un umbral de selección de Spearman ≥ 0.20. Este umbral se eligió de forma intencionada para su representación visual, ya que permite incluir un mayor número de variables, lo que facilita la observación de cómo se distribuye su peso relativo en comparación con modelos más restrictivos, cuando se aplica un umbral ≥ 0.30. En el modelo de predicción (*Figura 2*), destacan principalmente variables dietéticas como EAT_pork, EAT_dairy y EAT_poultry, seguidas por EAT_legumes y EAT_fruits. También aparecen con cierto peso algunos metabolitos como 3-Methyl-2-oxovalerate, Leucina y Alanina. En el modelo de clasificación (*Figura 3*), el patrón de importancia cambia ligeramente: EAT_legumes, EAT_fruits y EAT_added_sugar ocupan los primeros puestos, seguidos por EAT_dairy y metabolitos como 3-Hydroxyisobutyrate y Alanina. Este contraste entre ambos modelos pone de manifiesto cómo ciertas variables cambian su relevancia según el tipo de tarea (predicción continua o clasificación), y cómo un umbral de selección más bajo permite visualizar un conjunto más amplio de predictores activos, aunque con un peso más repartido entre ellos.

Las Figuras 4 y 5 muestran las variables más importantes según los modelos de Random Forest construidos con un umbral de Spearman ≥ 0.30, que fue el modelo con mejor rendimiento global. Al hacerse una selección más rigurosa, estas figuras permiten identificar con mayor claridad cuáles son los predictores más relevantes en ambos modelos.

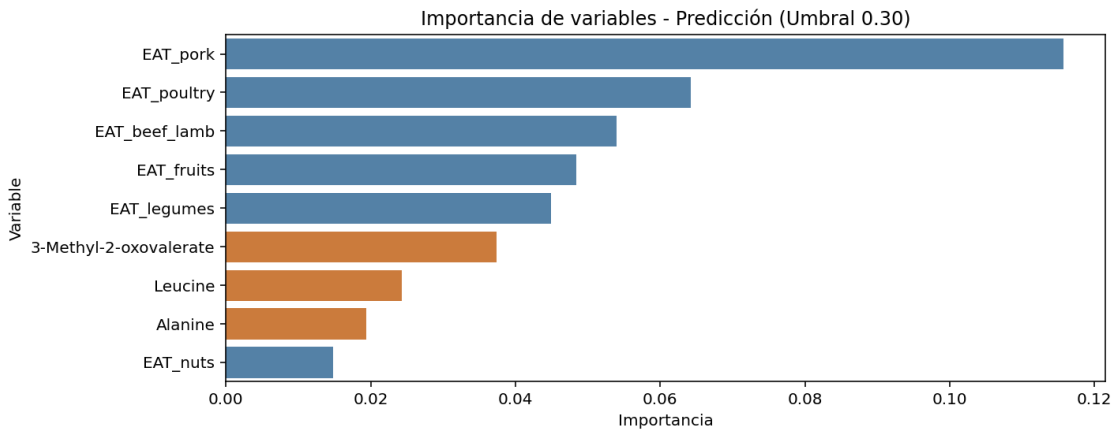

Figura 4. Importancia de variables en el modelo para la predicción del score EAT-Lancet utilizando todas las variables

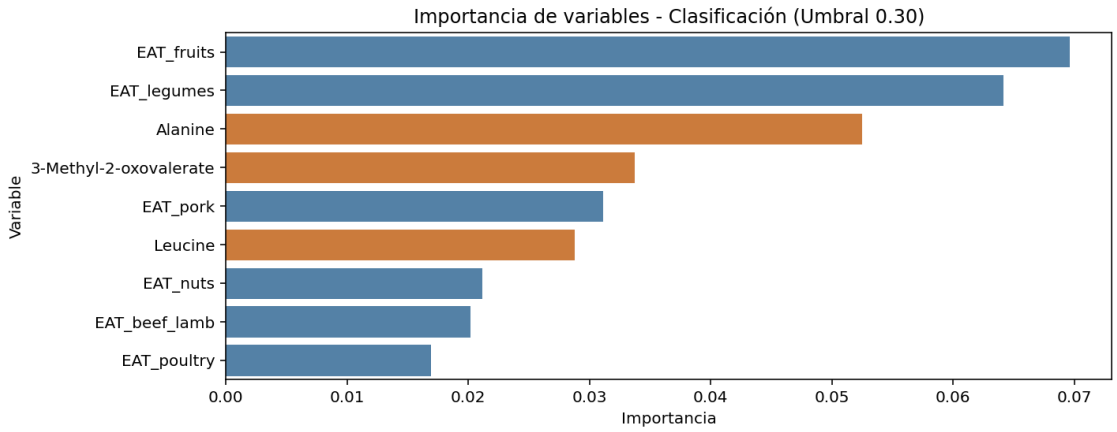

Figura 5. Importancia relativa de las variables predictoras en el modelo de clasificación de la adherencia al patrón dietético

En ambos casos se observa una clara predominancia de variables dietéticas, especialmente aquellas relacionadas con el consumo de cerdo, aves, frutas y legumbres. En el modelo de predicción (*Figura 4*), EAT_pork aparece como el predictor más relevante, mientras que en el modelo de clasificación (*Figura 5*) lideran EAT_fruits y EAT_legumes. Esta diferencia en el orden de importancia, de nuevo, pone de manifiesto cómo el tipo de modelo influye en qué variables resultan más útiles. Por otro lado, los metabolitos Alanina, 3-Methyl-2-oxovalerate y Leucina mantienen una presencia destacada en ambos casos, lo que refuerza su papel como posibles biomarcadores asociados al patrón alimentario EAT-Lancet.

Los resultados globales de este estudio confirman que la aplicación de técnicas avanzadas de ingeniería de características, como la selección mediante correlación de Spearman, mejora de forma significativa tanto la precisión como la fiabilidad de los modelos predictivos y clasificatorios orientados a evaluar la adherencia a la dieta EAT-Lancet. Los modelos optimizados mediante selección individual de variables (umbral ≥ 0.20, 0.25 y 0.30) mostraron un rendimiento muy superior frente a los modelos sin ingeniería de características o aquellos construidos exclusivamente con bloques de variables agrupadas por tipo (clínicas, dietéticas o metabolómicas). En particular, el modelo con umbral ≥ 0.30 alcanzó un rendimiento destacado (R^2 = 0.618; RMSE = 1.972), lo que evidencia que integrar variables individuales seleccionadas con criterios estadísticos claros ofrece mayor capacidad explicativa que usar grupos predefinidos sin filtrado adicional.

Es importante señalar que algunos modelos arrojaron valores negativos de R^2. Esto ocurre cuando el modelo tiene un poder predictivo inferior al que se obtendría simplemente usando la media de la variable objetivo como predicción. Esta situación suele darse cuando las variables incluidas no aportan información útil para explicar la variabilidad observada, especialmente en modelos sin filtrado previo o con grupos de variables de baja relevancia estadística.

Uno de los hallazgos más interesantes fue la relación consistente entre varios metabolitos urinarios identificados como importantes (Alanina, Leucina, 3-Methyl-2-oxovalerate y 3-Hydroxyisobutyrate) y grupos alimentarios EAT-Lancet (como las carnes, frutas y las legumbres) con las variables objetivos (EAT_lancet_adherencia). Esta asociación se alinea con evidencias previas que sugieren su relación directa con la ingesta de alimentos ricos en proteínas animales, frutas y legumbres (De la O et al., 2024). Esta coherencia entre dieta y metabolitos refuerza el valor de estos compuestos como posibles biomarcadores objetivos de adherencia a patrones dietéticos sostenibles.

Por último, es llamativo que algunas variables clínicas tradicionalmente utilizadas en estudios nutricionales, como la glucosa o el colesterol total, no aparecieran como relevantes en los modelos optimizados. Estos resultados sugieren que, en contextos de evaluación dietética, variables más específicas como los metabolitos urinarios y los indicadores dietéticos detallados pueden ofrecer una capacidad predictiva más precisa y útil. Esto pone de relieve la importancia de avanzar hacia una nutrición más personalizada, basada en biomarcadores más sensibles y específicos (Stubbendorff et al., 2022).

En conjunto, estos hallazgos no solo evidencian el valor añadido de integrar variables metabolómicas y dietéticas en el modelado predictivo, sino que también refuerzan el potencial de aplicar enfoques de aprendizaje automático en el ámbito de la nutrición personalizada, con vistas a mejorar la evaluación objetiva de patrones dietéticos y guiar intervenciones más precisas y sostenibles en salud pública.

5. Conclusiones

El presente estudio ha permitido analizar la adherencia a la dieta EAT-Lancet mediante modelos predictivos y clasificatorios basados en aprendizaje automático, integrando variables clínicas, dietéticas y metabolómicas. El uso de técnicas avanzadas de ingeniería de características ha permitido identificar y optimizar las variables predictoras más relevantes, mejorando la precisión y aplicabilidad de los resultados obtenidos. A continuación, se presentan las principales conclusiones extraídas del análisis realizado:

1. Se analizaron y aplicaron técnicas de ingeniería de características que permitieron optimizar las variables predictoras mediante correlación de Spearman, determinando el umbral óptimo de selección en ≥ 0.30.

2. Se desarrollaron modelos de regresión y clasificación altamente efectivos para estimar la adherencia a la dieta EAT-Lancet en su forma continua (EAT score) y categórica (alta/baja adherencia). El modelo basado en Random Forest con umbral ≥ 0.30 alcanzó los mejores resultados, un $R^2 = 0.618$ y excelentes hallazgos en clasificación (Accuracy y F1-score superiores al 0.80), destacando claramente respecto a modelos sin ingeniería de características (DeGregory et al., 2018).

3. La comparación del rendimiento de los modelos evidenció que aquellos generados con selección de variables presentan considerablemente mejor desempeño predictivo, validando la importancia de la selección óptima de predictores basada en correlaciones y características estadísticas.

4. Se identificaron las variables más relevantes en los modelos finales.
    I. Entre los metabolitos destacaron alanina, leucina, 3-Methyl-2-oxovalerate y 3-Hydroxyisobutyrate, apareciendo consistentemente como relevantes en clasificación y regresión, lo que refuerza el papel de las mismas como posibles biomarcadores metabólicos de tipo de dieta.
    II. Las variables dietéticas con mayor peso fueron el consumo de frutas (EAT_fruits), legumbres (EAT_legumes), carne de cerdo (EAT_pork), carne roja (EAT_beef_lamb), aves (EAT_poultry) y frutos secos (EAT_nuts), en línea con las recomendaciones del patrón EAT-Lancet.
    III. Las variables clínicas, como glucosa o colesterol, no mostraron capacidad predictiva relevante en los modelos con umbrales más bajos, lo que refuerza la especificidad de los predictores seleccionados y sugiere un papel secundario de los marcadores clínicos tradicionales en este tipo de enfoques (de la O et al., 2024).

6.Bibliografía

de la O V, Martínez JA, Fernández-Cruz E, Valdés A, Cifuentes A, Walton J. Exhaustive search of dietary intake biomarkers as objective tools for personalized nutrimetabolomics and precision nutrition implementation: a scoping review. Adv Nutr. 2024;15(2):255–272.

DeGregory KW, Kuiper P, DeSilvio T, Pleuss JD, Miller R, Roginski JW, et al. A review of machine learning in obesity. Obes Rev. 2018 May;19(5):668–685.

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.

Stubbendorff A, Sonestedt E, Ramne S, Drake I, Hallström E, Ericson U. Development of an EAT-Lancet index and its relation to mortality in a Swedish population. Am J Clin Nutr. 2022;115(3):705–16.