PREDICCIÓN DE DIABETES MELLITUS EN EL CONTEXTO COLOMBIANO A PARTIR DE PATRONES DE CONSUMO Y COMPUESTOS MOLECULARES MEDIANTE EL USO DE TÉCNICAS SUPERVISADAS DE MACHINE LEARNING

Jeison Suescun Holguín

Código 8934976

Proyecto Aplicado para optar al título de

Magister en Ciencia de Datos

Director(a)

Delia Ortega Lenis

Codirector(a)

Juliana Chaura Cortés

FACULTAD DE INGENIERÍA Y CIENCIAS

MAESTRÍA EN CIENCIA DE DATOS

SANTIAGO DE CALI, JUNIO 9 DE 2025

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

INTRODUCCIÓN

La diabetes mellitus tipo 2 (DM2) es una enfermedad metabólica crónica de alta prevalencia que representa una de las principales causas de morbilidad y mortalidad a nivel mundial. Su progresión silenciosa y la fuerte relación con factores dietarios y de estilo de vida la convierten en un problema prioritario en salud pública [1], [2]. En el caso de Colombia, las cifras recientes evidencian un aumento sostenido en los diagnósticos, afectando de forma desproporcionada a poblaciones con menor acceso a servicios de prevención y diagnóstico temprano. Esta situación demanda enfoques innovadores que permitan anticipar el riesgo de enfermedad a partir de información accesible, como los patrones de consumo alimentario.

La literatura científica ha documentado ampliamente la asociación entre la dieta y el riesgo de desarrollar DM2, reconociendo el papel de componentes como los azúcares simples, las grasas saturadas y la fibra dietaria [3], [4], [5], [6]. Sin embargo, muchos estudios se centran en variables generales o en patrones dietarios amplios, sin desagregar adecuadamente la influencia específica de nutrientes individuales ni considerar la complejidad molecular de los alimentos. Esta limitación es especialmente relevante en contextos como el colombiano, donde la diversidad alimentaria y los hábitos culturales podrían dar lugar a patrones particulares de riesgo que no han sido suficientemente caracterizados.

En este escenario, los avances en ciencia de datos y aprendizaje automático abren nuevas posibilidades para analizar relaciones complejas entre múltiples variables alimentarias y de salud. El uso de modelos predictivos permite no solo anticipar la aparición de condiciones como la diabetes, sino también identificar factores relevantes que podrían guiar intervenciones nutricionales específicas. No obstante, la aplicación de estos modelos requiere superar desafíos metodológicos como el desbalance entre clases, la integración de bases de datos heterogéneas y la necesidad de interpretabilidad para su uso en contextos reales.

El presente trabajo se enmarca en esa necesidad, proponiendo un modelo predictivo de riesgo de DM2 que considera tanto la composición nutricional como la composición molecular de los alimentos consumidos por la población colombiana. A partir de datos de la Encuesta Nacional de la Situación Nutricional (ENSIN), se plantea una estrategia de modelado capaz de identificar variables dietarias de alta relevancia, al tiempo que se exploran enfoques de balanceo de clases y métodos de interpretación de modelos orientados a contextos epidemiológicos. Con ello, se busca contribuir al desarrollo de herramientas que complementen los sistemas tradicionales de monitoreo y prevención de enfermedades crónicas no transmisibles en el país.

Los resultados obtenidos a partir de esta estrategia no solo evidenciaron un desempeño predictivo notable, sino que también permitieron identificar un conjunto de variables clave asociadas al riesgo de diabetes. Entre los compuestos más influyentes se destacaron la vitamina B6, la proteína de origen animal, la fibra cruda y la vitamina A. En términos de alimentos, se observó una asociación inversa entre el consumo de frutas y el riesgo de diabetes, mientras que un mayor consumo de dulces mostró una relación positiva con la enfermedad. Estas asociaciones son coherentes con hallazgos de estudios internacionales [7], [8], [9], [10], [11] y refuerzan la importancia de considerar tanto la calidad como la cantidad de los alimentos dentro de las estrategias de prevención. No obstante, se requiere investigación adicional para validar estos patrones en la población local, idealmente mediante estudios longitudinales o clínicos que complementen los hallazgos del presente modelo.

Este documento está estructurado de la siguiente manera: el Capítulo 1 presenta la problemática; los Capítulos 2 y 3 abordan los objetivos y el contexto del estudio; el Capítulo 4 revisa antecedentes y metodologías previas; y los Capítulos 5 a 8 se centran en el análisis exploratorio de datos y el desarrollo del modelo. Finalmente, el trabajo concluye en el Capítulo 9 con una interpretación de los resultados obtenidos para dar paso a las conclusiones y trabajos futuros.

1. DEFINICIÓN DEL PROBLEMA

1.1. Planteamiento del problema

La diabetes y en particular la diabetes mellitus tipo 2 (DM2) representa un desafío significativo para la salud pública mundial, con aproximadamente 422 millones de casos a nivel global y 1,6 millones de casos reportados en Colombia. Esto subraya la urgencia de abordar esta enfermedad de manera eficaz [12], [13]. Entre los principales factores que agravan la DM2 se encuentra la dieta, que ha evolucionado hacia un menor grado de diversidad en los ingredientes y un mayor consumo de alimentos potencialmente perjudiciales. Numerosos estudios han explorado la relación entre la alimentación y la DM2, identificando la dieta como un pilar fundamental para su control y prevención [14], [15], [16]. Estos estudios han destacado alimentos específicos, patrones alimentarios y compuestos moleculares relacionados con la enfermedad [17], [18], [19], [20], [21]. Sin embargo, persiste un notable vacío de conocimiento sobre cómo los alimentos y sus compuestos moleculares específicos influyen en la aparición y progresión de la DM2.

Los enfoques tradicionales para el manejo y prevención de la diabetes, aunque efectivos hasta cierto punto, no han sido suficientes para detener el aumento global de casos [22]. Además, disciplinas emergentes como la alimentómica (del inglés *Foodomics*) han puesto de manifiesto la limitada comprensión de cómo los alimentos y sus compuestos moleculares afectan nuestra salud en general [23], [24], [25], [26]. Muchas recomendaciones dietéticas se han basado en apenas 150 componentes (entre macronutrientes y micronutrientes), ignorando miles de otros componentes presentes en los alimentos [27]. Este contexto resalta la necesidad de adoptar un enfoque más integral y basado en evidencia para comprender mejor la relación entre la dieta y la diabetes.

En este sentido, la ciencia de datos se presenta como una herramienta clave para abordar este problema. El análisis de grandes volúmenes de datos, incluyendo información sobre ingesta alimentaria, perfiles moleculares de los alimentos y datos clínicos de personas con diabetes, permite identificar patrones y relaciones que podrían pasar desapercibidos con métodos convencionales. Este enfoque ofrece una comprensión más profunda y detallada de cómo la dieta influye en la DM2, y va más allá de considerar únicamente alimentos específicos o frecuencias de consumo [28], [29], [30].

No obstante, este campo enfrenta desafíos significativos. Las bases de datos sobre la composición molecular de los alimentos suelen ser limitadas, desactualizadas o inexactas, lo que dificulta un análisis riguroso [31]. Asimismo, la variabilidad en los hábitos alimentarios entre individuos y regiones complica la generalización de los resultados. Para superar estas limitaciones, es crucial implementar técnicas avanzadas de ciencia de datos, como el aprendizaje automático, para analizar datos no estructurados y desarrollar modelos predictivos. Estas herramientas permiten identificar factores de riesgo y patrones dietéticos asociados con la diabetes, proporcionando nuevas perspectivas para su manejo y prevención.

1.2. Formulación del problema

La investigación se centró en una cuestión clave: determinar el impacto de la ingesta de distintos alimentos, junto con su composición molecular y nutricional, en la incidencia de la diabetes en la población colombiana. Además, se buscó analizar cómo la ciencia de datos puede profundizar en esta relación para predecir posibles riesgos de desarrollar la enfermedad. Para abordar este interrogante, fue imprescindible

desglosar diversas cuestiones técnicas y prácticas fundamentales para el desarrollo del estudio.

Primero, fue crucial abordar la disponibilidad y calidad de los conjuntos de datos necesarios para la investigación. Las preguntas clave fueron:

• ¿Qué bases de datos estuvieron accesibles y qué nivel de fiabilidad poseían en términos de información sobre la composición molecular y nutricional de los alimentos, los hábitos alimenticios y la incidencia de la diabetes?

• ¿Qué tipo de preparación o preprocesamiento se debió realizar sobre esta información con el objetivo de adaptarla para la construcción de modelos?

También fue necesario identificar las estrategias analíticas óptimas que permitieran explorar de manera efectiva la asociación entre la composición molecular y nutricional de los alimentos y la diabetes. Esto implicó responder a las siguientes preguntas:

• ¿Qué tipo de análisis estadísticos exploratorios debieron ser efectuados y sobre qué variables para identificar relaciones y comportamientos importantes para el análisis?

Una vez listos los datos y exploradas las relaciones, se entró en la fase de construcción del modelo. Las preguntas a responder en esta etapa fueron:

• ¿Cuáles fueron los métodos más adecuados, de acuerdo con los datos disponibles, para construir el modelo?

• ¿Cuáles fueron las métricas adecuadas para evaluar la precisión y efectividad del modelo predictivo?

Finalmente, se consideró el alcance y el impacto potencial de los hallazgos en la prevención y manejo de la diabetes. Las preguntas esenciales aquí fueron:

• ¿Cómo pudieron las conclusiones de la investigación influir en la formulación de recomendaciones dietéticas dirigidas a la prevención de la diabetes y la mejora de la salud pública?

Estas interrogantes no solo fueron cruciales para la investigación en curso, sino que también determinaron la relevancia de la aproximación y su potencial aplicación para el estudio de otras enfermedades donde la dieta jugó un papel fundamental.

2. OBJETIVOS

Objetivo General

Desarrollar un modelo predictivo utilizando algoritmos supervisados de Machine Learning que permita relacionar los alimentos y sus compuestos moleculares con la diabetes, con el fin de predecir potenciales riesgos de padecer la enfermedad asociados a la dieta en el contexto colombiano.

Objetivos Específicos

OE1. Recopilar y preparar datos relacionados con la composición molecular y nutricional de los alimentos, la ingesta dietética y el estado de salud de los individuos con respecto a la diabetes, realizando procesos de limpieza y preprocesamiento para garantizar la calidad de los datos.

OE2. Realizar un análisis exploratorio de los datos para comprender la distribución y relaciones entre las variables, así como seleccionar las características más relevantes para la construcción del modelo predictivo.

OE3. Desarrollar un modelo predictivo utilizando algoritmos supervisados de Machine Learning, como clasificadores, entrenándolo con los datos recopilados y ajustando los hiperparámetros para mejorar su rendimiento en la predicción de los riesgos de enfermedad asociados a la dieta.

OE4. Evaluar el rendimiento del modelo predictivo mediante técnicas de validación cruzada y métricas adecuadas, como la precisión, el recall y el F1-score, realizando ajustes adicionales según sea necesario para garantizar su fiabilidad y generalización.

OE5. Analizar e interpretar los resultados del modelo predictivo para identificar los factores más influyentes en la relación entre la composición molecular y nutricional de los alimentos y la diabetes, extrayendo conclusiones relevantes para la formulación de recomendaciones dietéticas y la prevención de la enfermedad.

3. MARCO TEÓRICO

3.1. Diabetes mellitus

La diabetes mellitus, comúnmente conocida como diabetes, es una de las enfermedades crónicas no transmisibles más prevalentes y constituye una de las principales causas de mortalidad a nivel mundial. En 2016, se estimó que 422 millones de adultos vivían con diabetes, y las tendencias indican un aumento constante tanto en la prevalencia como en la mortalidad asociada, con un incremento de hasta el 3% en la tasa de mortalidad durante los últimos 20 años. La diabetes se clasifica en varios tipos, siendo los más comunes la diabetes tipo 1 (DM1), la diabetes tipo 2 (DM2) y la diabetes gestacional [1], [12].

De estos, la DM2 es la más frecuente, afectando aproximadamente al 90-95% de las personas diagnosticadas con algún tipo de diabetes, lo que la convierte en un problema crítico para la salud pública. Esta condición se caracteriza por una resistencia a la insulina y una producción insuficiente de insulina por parte del páncreas. A menudo se desarrolla de manera progresiva y suele estar precedida por un estado conocido como prediabetes. A diferencia de la DM1, la DM2 puede ser manejada de manera efectiva mediante cambios en el estilo de vida, como una dieta equilibrada, actividad física regular, y, en algunos casos, el uso de medicamentos [13], [14].

Por su parte, la diabetes tipo 1, aunque menos prevalente, también es significativa. Esta condición autoinmune se produce cuando el sistema inmunológico del cuerpo destruye las células beta del páncreas, responsables de la producción de insulina. Como resultado, las personas con diabetes tipo 1 dependen de la administración de insulina exógena para regular los niveles de glucosa en sangre [27].

La diabetes gestacional, en cambio, se diagnostica durante el embarazo y comparte características con la diabetes tipo 2, particularmente en lo relacionado con la resistencia a la insulina. Sin embargo, su origen está estrechamente vinculado a los cambios hormonales propios de esta etapa. Aunque en la mayoría de los casos desaparece tras el parto, la diabetes gestacional incrementa significativamente el riesgo de desarrollar diabetes tipo 2 en el futuro, tanto para la madre como para el hijo.

Este estudio se enfoca en identificar factores de riesgo asociados a la dieta, dada la importancia de este conocimiento para la prevención efectiva de la diabetes tipo 2.

3.2. Composición nutricional y molecular de los alimentos

La composición molecular de los alimentos se refiere a la combinación y proporción de diversos elementos, compuestos orgánicos e inorgánicos, y otras sustancias que confieren características particulares a cada alimento. Los alimentos están compuestos principalmente por agua, lípidos, carbohidratos y proteínas, que constituyen la mayor parte del alimento y le otorgan propiedades funcionales y sensoriales, como valor nutricional y sabor. Es importante destacar que algunos alimentos pueden contener alérgenos y moléculas tóxicas, cuyo consumo puede requerir estrategias específicas de recolección o cocción para hacerlos seguros [32].

Durante los procesos de producción, distribución y cocción, los alimentos pueden sufrir cambios significativos. Por ejemplo, los alimentos fritos absorben grasas saturadas y trans del aceite utilizado, y experimentan oxidación y degradación térmica, lo que modifica sus propiedades nutricionales [15]. También existen formas de contaminación que afectan la composición final de los alimentos, como el uso de pesticidas, métodos de conservación y crecimiento microbiano, que pueden ser incluso carcinogénicos

[16].

Controlar la composición molecular de un alimento antes de su consumo es una tarea complicada debido a la interacción de múltiples factores que la afectan. La complejidad se ve exacerbada por la existencia de diferentes metodologías para medir la proporción de cada compuesto o estimarla cuando sea necesario. El agua, uno de los compuestos más esenciales, determina la capacidad de conservación y las propiedades funcionales y sensoriales del alimento. Se mide con diversas técnicas de radiación, como infrarroja y microondas, utilizando analizadores halógenos de humedad.

La cantidad de proteína se determina indirectamente mediante la cuantificación del porcentaje de nitrógeno, utilizando métodos como Kjeldahl y, más recientemente, el método Dumas. Para medir los ácidos grasos, se utilizan métodos como Soxhlet y la hidrólisis ácida/alcalina, que preparan los ácidos grasos para su análisis en cromatógrafos de gases. Para medir azúcares, se emplean métodos enzimáticos, de absorbancia y cromatografía líquida de alta resolución acoplada a detectores de dispersión de luz evaporativa o masas, que presentan ventajas en términos de sensibilidad [17]. Estos métodos son ampliamente utilizados en la industria, proporcionando valores aproximados de macronutrientes y otros elementos relevantes para el consumidor y la salud pública.

Es fundamental diferenciar entre composición molecular y composición nutricional. La composición nutricional se enfoca en grandes grupos de nutrientes, como proteínas, grasas y carbohidratos, e informa sobre su cantidad total en un alimento. Esta aproximación es útil para evaluaciones dietarías, pero no proporciona detalles sobre la identidad química de los compuestos individuales. En contraste, la composición molecular busca identificar y cuantificar todas las moléculas presentes en un alimento, incluyendo compuestos bioactivos, metabolitos secundarios y otros elementos con potencial efecto sobre la salud. Esta caracterización requiere tecnologías analíticas de alta resolución, como la espectrometría de masas acoplada a técnicas de separación avanzadas, cuyo uso aún es limitado en estudios nutricionales debido a su complejidad y costo [33]. Por ello, la mayoría de estudios disponibles ofrecen datos agregados de tipo nutricional, sin alcanzar el nivel de detalle que exige un análisis molecular completo.

3.3. La diabetes y la alimentación

La comprensión de la variedad y composición de los alimentos es fundamental para el control de enfermedades crónicas, como la DM2. Esto se debe a que permite identificar alimentos específicos asociados con el riesgo de enfermedad en contextos determinados. Por ejemplo, en China, los estudios de dieta total han detectado fuentes de toxicidad por pesticidas en ciertos alimentos, lo que ha llevado a medidas para mitigar estos riesgos [18]. Además, ciertos compuestos presentes en los alimentos pueden aumentar el riesgo de desarrollar o agravar enfermedades. Investigaciones han revelado que el consumo frecuente de alimentos fritos está correlacionado con un mayor riesgo de padecer DM2, obesidad y enfermedades cardíacas [19], [20], [21]. De manera similar, los azúcares añadidos y una baja ingesta de proteínas se han asociado con un aumento en las complicaciones de enfermedades crónicas [23].

La identificación de compuestos y factores dietéticos relacionados con la evolución de enfermedades crónicas ha impulsado la creación de nuevas estrategias y políticas públicas. La Organización Mundial de la Salud (OMS) ha informado sobre avances a nivel mundial, incluyendo sistemas de etiquetado más comprensibles, impuestos a las bebidas azucaradas y reducciones en los límites de grasas saturadas permitidas por producto. Para abordar enfermedades crónicas como la diabetes, la reducción del consumo de sodio es una prioridad, con el objetivo de disminuirlo en un 30% para 2025. Decisiones informadas

sobre la composición de los alimentos ya están marcando la diferencia en el control de enfermedades en África y en países como India y Ruanda[24], [28].

La relación entre la composición nutricional y molecular de los alimentos y la diabetes es un área de estudio crucial. Los factores dietéticos comúnmente relacionados con DM2 incluyen azúcares añadidos, carbohidratos refinados, grasas saturadas, sodio, alimentos procesados y carne roja, especialmente cuando se consumen en exceso. Por otro lado, la falta de fibra, frutas, lácteos y verduras en la dieta también se asocia con un mayor riesgo de desarrollar DM2 [29], [30], [31].

Existen compuestos y alimentos cuyo consumo se considera terapéutico y preventivo en el manejo de la diabetes, conocidos como "alimentos funcionales". Estos alimentos, que poseen propiedades bioactivas y protectoras contra los procesos diabéticos, son comúnmente encontrados en productos agrícolas. De hecho, el bajo consumo de frutas, granos y verduras se correlaciona con un mayor riesgo de diabetes [34]. Otro ejemplo muy conocido de protección proveniente de los alimentos es la dieta mediterránea, rica en alimentos con compuestos bioactivos. De forma individual, se ha demostrado que algunos compuestos como los polifenoles tienen propiedades antidiabéticas significativas, destacando los subgrupos de ácidos fenólicos y flavonoides [35].

3.4. Balanceo de Clases en Problemas de Clasificación

En problemas de clasificación aplicados a contextos de salud pública, como el diagnóstico de enfermedades, es frecuente encontrar conjuntos de datos desbalanceados, donde la clase de interés (por ejemplo, personas con diagnóstico positivo) está subrepresentada en comparación con la clase negativa. Este desbalance puede deteriorar el desempeño del modelo predictivo, ya que los algoritmos tienden a favorecer la clase mayoritaria, comprometiendo métricas clave como la sensibilidad o la tasa de verdaderos positivos [36], [37]. Para abordar esta limitación, se han desarrollado diversas estrategias de sobremuestreo que buscan mejorar la representación de la clase minoritaria mediante la generación de ejemplos sintéticos. Entre las más utilizadas se encuentran SMOTE (Synthetic Minority Oversampling Technique) y ADASYN (Adaptive Synthetic Sampling), ambas orientadas a reforzar el aprendizaje del modelo en presencia de desbalance sin recurrir a la simple duplicación de instancias.

SMOTE, propuesta por [38], genera nuevos ejemplos sintéticos al interpolar entre un punto de la clase minoritaria y uno de sus vecinos más cercanos. Este enfoque crea ejemplos adicionales distribuidos de forma más continua en el espacio de características, lo que permite suavizar las fronteras de decisión del clasificador y reduce el riesgo de sobreajuste que podría derivarse del uso de copias exactas. No obstante, SMOTE distribuye los nuevos ejemplos de manera uniforme, sin tener en cuenta la dificultad particular de clasificar ciertos casos, lo que puede limitar su eficacia en regiones donde las clases están solapadas o donde la frontera de decisión es especialmente compleja.

Para superar esta limitación, [39] desarrollaron ADASYN, una técnica que incorpora un componente adaptativo al proceso de sobremuestreo. A diferencia de SMOTE, ADASYN asigna mayor peso a la generación de ejemplos sintéticos en regiones del espacio donde la clase minoritaria es más difícil de aprender, es decir, donde los ejemplos están más cercanos a la clase mayoritaria. De este modo, se refuerza el entrenamiento del modelo en zonas donde la probabilidad de clasificación errónea es mayor, lo que puede traducirse en una mejora de la capacidad de generalización. Sin embargo, esta estrategia también implica riesgos, ya que, al generar ejemplos en regiones de alta complejidad, puede introducir ruido o ambigüedad si no se controla adecuadamente el solapamiento entre clases.

Ambas técnicas han demostrado ser efectivas para mejorar el rendimiento de clasificadores en contextos con datos desbalanceados, particularmente en aplicaciones biomédicas y epidemiológicas donde los casos positivos son escasos. La elección entre SMOTE y ADASYN depende en gran medida de la estructura del conjunto de datos y de los objetivos del análisis. Mientras SMOTE ofrece una solución más controlada y estructurada, adecuada para conjuntos de datos con fronteras de clase bien definidas, ADASYN es preferido en escenarios donde se requiere mayor adaptabilidad y atención a los patrones locales de dificultad en la clasificación.

3.5. Modelos de clasificación, votación y marcos de interpretación

Este estudio emplea algoritmos supervisados para predecir condiciones de salud a partir de variables dietarias y de composición nutricional. Los modelos seleccionados fueron XGBoost, LightGBM, CatBoost, Random Forest y K-Nearest Neighbors (KNN), los cuales son ampliamente reconocidos por su alto rendimiento en datos tabulares y por su capacidad para capturar relaciones no lineales. XGBoost combina árboles débiles mediante gradient boosting, corrigiendo iterativamente los errores de predicciones previas y ofreciendo gran eficiencia computacional [40]; LightGBM acelera dicho proceso con un crecimiento del árbol leaf-wise, reduciendo tiempo de entrenamiento y uso de memoria sin sacrificar precisión [41]; CatBoost, por su parte, maneja variables categóricas de forma nativa y emplea esquemas de ordenamiento por permutación para mitigar el overfitting y mejorar la estabilidad del modelo [42]; Random Forest entrena múltiples árboles sobre subconjuntos aleatorios de datos y características, promediando sus resultados para reducir la varianza y proporcionar estimaciones internas de importancia de variables [43]; finalmente, KNN clasifica cada observación según la clase predominante entre sus k vecinos más cercanos, siendo su simplicidad inversamente proporcional a la sensibilidad frente a la dimensionalidad y la escala de los datos [44].

Con el fin de capitalizar las fortalezas particulares de cada clasificador y amortiguar sus debilidades, las salidas de estos modelos se integraron mediante un ensemble de votación. En la votación dura (hard voting) la clase final se decide por mayoría simple de etiquetas, mientras que en la votación blanda (soft voting) se promedian las probabilidades estimadas, produciendo predicciones más calibradas. De forma general, los ensembles mejoran la precisión al reducir simultáneamente sesgo y varianza, ofrecen robustez frente a datos ruidosos y favorecen la generalización cuando los modelos base aportan errores decorrelacionados [45], [46]. No obstante, incrementan la complejidad computacional y diluyen la interpretabilidad al superponer decisiones individuales [47].

La transparencia del ensemble depende, además, de la compatibilidad entre el método interpretativo y los algoritmos subyacentes. SHAP (SHapley Additive exPlanations) dispone de una versión exacta y eficiente, TreeSHAP, para modelos basados en árboles, lo que permite calcular valores Shapley de XGBoost, LightGBM, CatBoost y de los árboles de Random Forest en tiempo lineal respecto al número de nodos. Para métodos no arbóreos, como KNN, solo es viable la variante agnóstica Kernel SHAP, cuyo coste computacional es mayor y la varianza de la estimación, más alta; por ello se aplica sobre subconjuntos de datos y su contribución se pondera en la síntesis global. Esta diferencia metodológica se tiene en cuenta al fusionar las explicaciones de todos los miembros del ensemble.

Finalmente, SHAP se empleó para descomponer cada predicción en la contribución aditiva de cada variable, asignando un valor Shapley justo, consistente y local [48]. En conjunto, esta estrategia permitió (i) identificar a nivel global los nutrientes o alimentos que el ensemble considera más influyentes y (ii)

explicar caso por caso por qué un individuo concreto se clasifica como sano o diabético, ofreciendo una trazabilidad esencial para la generación de hipótesis etiológicas y la detección de posibles sesgos del modelo.

4. ANTECEDENTES

De acuerdo a una revisión sistemática de literatura en Scopus realizada para la construcción de estos antecedentes, no se encontraron trabajos reportados a la fecha, en donde se aborde la relación entre la diabetes mellitus y la composición molecular y nutrcional de los alimentos usando ciencia de datos con el objetivo de entender relaciones profundas entre la dieta y la enfermedad (**Tabla 1**). Sin embargo, varios estudios son relevantes para este trabajo ya sea por las técnicas empleadas o por su acercamiento desde otros frentes a la pregunta de investigación planteada en este trabajo.

Tabla 1. Lista de términos empleada en la sentencia de búsqueda de Scopus

Diabetes	Composición de alimentos	Ciencia de datos
• diabetes • type 1 diabetes • type 2 diabetes • diabetes mellitus • gestational diabetes • juvenile diabetes • adult-onset diabetes • insulin-dependent diabetes • non-insulin-dependent diabetes	• food composition • nutrient composition • food content • dietary composition • nutritional content • food components • nutritional value • nutritional composition • dietary content • nutrient content • food makeup	• data science • machine learning • artificial intelligence • AI • big data • data mining • computational analysis • predictive modeling • algorithm development • statistical learning • deep learning • neural networks • predictive analytics • data-driven modeling

Un estudio relevante fue publicado por Anjun et al. (2024) [4], quienes abordan la mejora de la salud en pacientes con diabetes tipo 2 (DM2) mediante un sistema digital de monitoreo continuo de glucosa (CGM) asistido por IA. Utiliza técnicas de machine learning como XGBoost, SARIMA y Prophet para predecir niveles de glucosa en sangre y ajustar la dieta diaria de los pacientes. Los resultados permiten mantener los niveles de glucosa dentro de un rango normal y alertar sobre fluctuaciones inminentes, con el objetivo de reducir los niveles de HbA1c a ≤ 5.7% en tres meses. Este trabajo es relevante porque aplica técnicas de machine learning para gestionar la diabetes a través de la dieta. Sin embargo, se aleja bastante de cualquier posible metodología que pueda ser aplicada en esta propuesta, pero refleja el interés persistente de encontrar relaciones entre la dieta y la DM2 como una forma de control y tratamiento incluso en el presente año.

El trabajo de Liu et al. (2022) [49] evaluó exhaustivamente las características asociadas con los registros de accidentes cerebrovasculares utilizando datos de nutrientes dietéticos, biomarcadores sanguíneos e información clínica del National Health and Nutrition Examination Survey (NHANES) 2015-16. Se calcularon las importancias de las características con BoostARoota, se construyeron modelos de clasificación para datos desbalanceados como isolation forest, H2O Driverless y cost-sensitive Neura network. Los resultados mostraron que las características clínicas tienen el mayor poder predictivo en comparación con los

nutrientes dietéticos y biomarcadores sanguíneos, con un aumento del 22.8% en el área bajo la curva ROC (AUROC). Este trabajo es relevante porque demuestra cómo el análisis de datos clínicos, dietéticos y biomarcadores puede predecir enfermedades mediante modelos de machine learning. Aunque se centra en el accidente cerebrovascular (derrame cerebral), sus metodologías son aplicables a el proyecto descrito en este documento, que pretende utilizar datos de salud, alimentación y técnicas de machine learning para identificar relaciones entre la dieta y la diabetes mellitus. Sin embargo, nuestro enfoque se distingue al profundizar en la composición molecular y nutricional de los alimentos y su impacto específico en la diabetes.

El capítulo 20 del libro "Nutrición de Precisión", titulado "Ciencia de Redes y Aprendizaje de Máquina para la Nutrición de Precisión" [50], es un recurso relevante para este proyecto de investigación. En este capítulo, se revisan los métodos y esfuerzos para avanzar en la comprensión de los alimentos, pasando de una visión limitada basada en compuestos individuales a un enfoque más holístico. Esto permite profundizar en el impacto de los diferentes compuestos y sus configuraciones en la salud. El uso de un marco de medicina de redes muestra que es posible predecir posibles asociaciones de salud de los bioquímicos alimenticios. Además, se discute la importancia potencial del uso de técnicas de machine learning e inteligencia artificial para identificar compuestos dentro de los alimentos y sus implicaciones para la salud. Este trabajo resalta la importancia de ir más allá de los nutrientes esenciales para comprender cómo los alimentos afectan nuestra salud, lo cual se alinea con nuestro objetivo de analizar la composición molecular de los alimentos en relación con la diabetes, lo cual no se ha hecho hasta ahora. Aunque el enfoque de este estudio se centra en ampliar el conocimiento sobre la composición bioquímica de los alimentos y sus efectos generales en la salud, nuestro proyecto se enfoca específicamente en cómo estos compuestos afectan la DM2.

Los estudios de monitorización de la dieta basada en datos como el publicado por Das et al. (2022) [51] reflejan el potencial de aplicación que pueden tener trabajos como el planteando en este anteproyecto. Este estudio aborda la importancia de la monitorización dietética en el manejo de enfermedades como la DM2 y las enfermedades cardiovasculares. Los métodos actuales de monitorización de la dieta son engorrosos y a menudo imprecisos. Se ha demostrado previamente que los monitores continuos de glucosa (CGMs) pueden predecir los macronutrientes de las comidas mediante el análisis de la respuesta glucémica postprandial. En este estudio, se investigan biomarcadores adicionales en sangre para mejorar la predicción de macronutrientes en comparación con el uso exclusivo de los CGMs. Se llevó a cabo un estudio nutricional con 10 participantes que consumieron nueve comidas mixtas con cantidades conocidas de macronutrientes, analizando la concentración de 33 biomarcadores dietéticos en varios momentos postprandiales. Luego, se desarrollaron modelos de machine learning para predecir las cantidades de macronutrientes utilizando estos biomarcadores. La conclusión principal es que la integración de estos biomarcadores dietéticos con los CGMs mejora la precisión de la predicción de macronutrientes, lo que podría llevar al desarrollo de métodos automatizados para monitorizar la ingesta nutricional, especialmente en el contexto del control de enfermedades. Este estudio es relevante para nuestro proyecto porque muestra cómo la identificación de marcadores dietéticos relacionados con enfermedades, como los compuestos moleculares de los alimentos, puede integrarse en sistemas de tratamiento asistido.

En el estudio publicado en 2018 por Panaretos et al. [52], se examinó la precisión predictiva de métodos estadísticos y de aprendizaje automático (ML) respecto a la asociación de patrones dietéticos con el riesgo de enfermedad cardiovascular (CVD). Utilizando datos del estudio ATTICA, se inscribieron 3042 participantes entre 2001 y 2002, con un seguimiento de 10 años de CVD en 2020 de ellos. Se aplicaron

técnicas de Teoría de Respuesta al Ítem para crear un puntaje de riesgo cardio metabólico combinado de 10 años, incluyendo incidencias de CVD, diabetes, hipertensión e hipercolesterolemia. Se realizó análisis factorial para identificar patrones dietéticos, y regresión lineal para evaluar su asociación con el puntaje cardio metabólico. Además, se utilizaron dos técnicas de ML (algoritmo de vecinos más cercanos y árbol de decisión de bosques aleatorios) para evaluar la salud de los participantes basada en información dietética. Los resultados destacaron la superioridad de ML sobre la regresión lineal en la clasificación precisa de individuos según su salud, sugiriendo su utilidad para evaluar riesgos de enfermedades en nutrición epidemiológica. Estas técnicas pueden ser aplicables en este proyecto, con un enfoque específico en la diabetes, para clasificar, limpiar y construir relaciones entre variables nutricionales y clínicas.

5. BÚSQUEDA Y EXPLORACIÓN DE DATOS.

En este capítulo se describen las fuentes de datos consultadas para responder a la pregunta de investigación, evaluando su alineación con la temática y el cumplimiento de diversas características. Además, se reportan las pautas mediante las cuales se realizó el proceso de exploración y preparación de los datos.

5.1. Recopilación de datos

Los datos de las tres versiones de la ENSIN y la ENDS 2010 se solicitaron a través del sistema de solicitud online del Ministerio de Salud de Colombia, un proceso formal que permitió obtener acceso a las bases de datos completas. Una vez recibidos los datos, se procedió a un análisis preliminar de las variables, utilizando los reportes nacionales y los diccionarios de variables proporcionados por el Ministerio. Este análisis permitió identificar y seleccionar las variables de interés, especialmente aquellas relacionadas con alimentación, diabetes y la composición molecular y nutricional de los alimentos.

5.2. Exploración de fuentes de datos: ENSIN 2005, 2010 y 2015

Para la recopilación de datos, se utilizaron las versiones 2005, 2010 y 2015 de la Encuesta Nacional de la Situación Nutricional en Colombia (ENSIN), una encuesta realizada periódicamente por el Ministerio de Salud y Protección Social, que recoge información sobre los hábitos alimenticios, el estado nutricional y de salud de la población colombiana. Estas encuestas son una referencia fundamental para el análisis de la relación entre la alimentación y algunas enfermedades, entre ellas la diabetes.

- **ENSIN 2005**: Esta versión incorporó el auto-reporte de diabetes por parte de los participantes, quienes respondieron a una pregunta específica sobre si alguna vez habían sido diagnosticados con esta enfermedad, lo que facilitó la identificación de individuos con dicha condición. Sin embargo, esta pregunta sobre diagnóstico previo no se realizó a la totalidad de la población de la ENSIN 2005, ya que dependía de una pregunta anterior que indagaba si la persona alguna vez se había realizado un examen de glicemia. En caso de una respuesta negativa, no se aplicaba la pregunta sobre el auto-reporte de diabetes. Como resultado, un gran porcentaje de la muestra de la ENSIN 2005 carece de esta etiqueta. Además, el conjunto de datos presenta cierto grado de desactualización debido al tiempo transcurrido desde su recolección.

  Esta versión de la encuesta empleó el método del Recordatorio de 24 horas (R24) para recolectar información alimentaria, lo que presentó importantes desafíos en la estandarización de los nombres de los alimentos debido a la diversidad de denominaciones específicas y regionales reportadas por los participantes. La sistematización de estos datos fue compleja, ya que la encuesta incluyó miles de registros a nivel nacional. Además, el método R24 exige que los participantes detallen todo lo consumido en las últimas 24 horas, lo que puede variar considerablemente entre individuos y no siempre refleja su dieta habitual. Por otro lado, los datos de composición alimentaria, como calorías y nutrientes, ya estaban parcialmente calculados en

esta versión, aunque su formato original requirió una reestructuración significativa para poder ser integrados adecuadamente en el análisis. [53].

- **ENSIN 2010:** En esta versión, los datos alimentarios no se recopilaron mediante el método de Recordatorio de 24 horas (R24), sino a través de frecuencias de consumo, un enfoque diseñado para reflejar la dieta habitual de las personas. Este método consiste en preguntar con qué frecuencia se consumen ciertos grupos de alimentos predeterminados, lo que facilita la estandarización de las respuestas al basarse en categorías claramente definidas y reduce la variabilidad asociada a las denominaciones específicas de los alimentos reportados. Además, permite obtener métricas clave, como la frecuencia de consumo de alimentos específicos, proporcionando una visión más precisa de los patrones dietéticos a nivel poblacional.
  Un aspecto destacado de la ENSIN 2010 es que se realizó en conjunto con la Encuesta Nacional de Demografía y Salud (ENDS) 2010, la cual aborda temas relacionados con la salud reproductiva, el estado nutricional y otros factores de salud pública en Colombia. Esta integración permitió obtener una visión más amplia del estado de salud de la población, al combinar información nutricional con datos demográficos y de salud general.
  Sin embargo, una limitación importante de esta versión fue el uso de los datos de la ENDS para abordar el tema de la diabetes. La ENDS 2010 incluyó un auto-reporte de diabetes, pero este solo estuvo disponible para la población mayor de 60 años. Dado que este grupo no representa a toda la población evaluada en la ENSIN, no fue posible realizar un análisis integral de la prevalencia de diabetes en esta versión. [54].
- **ENSIN 2015:** La versión más reciente de la ENSIN no se realizó conjuntamente con la ENDS y no cuenta con un auto-reporte o una variable específica que permita vincular claramente los datos de alimentación con la presencia de diabetes. Después de una exploración exhaustiva de la base de datos, los reportes y los diccionarios de variables, se concluyó que esta falta de vinculación impide realizar un análisis directo entre la dieta y la diabetes en esta versión [55].

Tras una exploración preliminar, se determinó que la única versión que cumplía con los requerimientos en cuanto a variables y estructura de datos era la ENSIN 2005. Por esta razón, se decidió enfocar el trabajo exclusivamente en esta versión, a pesar de la antigüedad de los datos. Las versiones de la ENSIN correspondientes a 2010 y 2015 fueron descartadas por completo y no se utilizaron en ninguna etapa posterior del análisis.

A continuación, se detallan las tareas específicas realizadas sobre la base de datos seleccionada.

5.3. Limpieza y preprocesamiento

Se revisaron los cuestionarios y las tablas para extraer las variables relevantes con base a los códigos asignadas a las mismas. Las variables relacionadas con diabetes, alimentación y composición de alimentos fueron rescatadas y reestructuradas en un nuevo formato para facilitar su análisis. En aquellos casos donde los datos estaban incompletos o inconsistentes, se aplicaron técnicas de imputación utilizando el método MICE (Multivariate Imputation by Chained Equations), garantizando que los conjuntos de datos fueran lo más completos posible para el análisis posterior. Los valores atípicos reportados en algunas casillas fueron

analizados de forma individual con base a la codificación asignada a cada variable y se verificó si este tenía algún significado interno definido por el instrumento de captura de información (Por ejemplo: índices de masa corporal cercanos de 998 es equivalente a no medido) en caso contrario fue simplemente imputado/calculado si era posible teniendo en cuenta otras variables o eliminado.

Los participantes que contaban con una etiqueta de diabetes fueron filtrados para generar una submuestra de la población de la ENSIN, con un tamaño de 5744 personas. Adicionalmente, se generaron dos grupos grandes de variables que podían agruparse como "alimentos" por su información centrada en consumo y "compuestos" por su información relacionada con composición nutricional y derivados.

Alimentos

En el caso de la ENSIN 2005, la reestructuración de los datos alimenticios requirió un trabajo de sistematización más complejo debido al uso del cuestionario R24, donde cada individuo reportaba alimentos de manera diferente (**Figura 1**). Fue necesario crear un proceso de sistematización para unificar los nombres de los alimentos y garantizar que la información fuera utilizable en el análisis, especialmente porque esta encuesta contenía más de 900 mil registros.

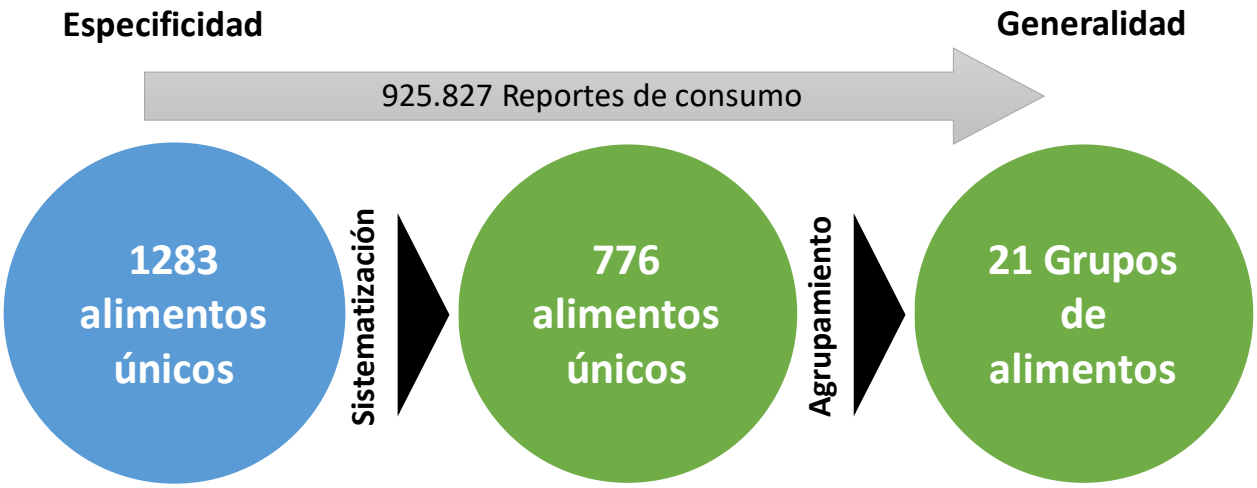

Figura 1. Manejo de los datos alimentarios en la ENSIN 2005.

El proceso de sistematización inicial consistió en la generación de una lista de nombres únicos de los alimentos disponibles en toda la lista de alimentos. Posteriormente se eliminaron espacios o caracteres invisibles que estuvieran diferenciando los registros, se eliminó la puntuación y se homogenizo el texto a minúsculas. Potenciales sinónimos y diferencias gramaticales sutiles fueron encontradas y corregidas. Para el proceso de agrupación de los alimentos se usó como base los grupos de alimentos estandarizados en versiones más actualizadas de la ENSIN, los cuales también fueron ajustados de acuerdo a la disponibilidad de alimentos observada en los registros. Los grupos de alimentos definidos se detallan en la Tabla 2.

Tabla 2. Grupos de alimentos generados mediante el proceso de sistematización.

Grupo de Alimentos	Descripción

Lácteos	Productos derivados de la leche, como queso y yogur.
Grano y Cereales	Cereales y sus derivados, como arroz, pan y pasta.
Aceites o Grasas	Grasas y aceites utilizados en la alimentación.
Verduras	Hortalizas y vegetales consumidos en la dieta.
Proteínas	Fuentes de proteínas de origen animal y vegetal.
Otro	Alimentos no clasificados en otras categorías.
Bebidas Calientes	Bebidas como café, té y chocolate caliente.
Dulce	Productos azucarados y postres.
Frutas	Frutas frescas y naturales.
Carne Blanca	Carnes como pollo, pavo y pescado.
Tubérculos	Raíces y tubérculos como papa, yuca y ñame.
Alimento Procesado	Alimentos industrializados con procesamiento.
Bebida Azucarada	Jugos artificiales, gaseosas y otras bebidas dulces.
Carne Roja	Carnes como res, cerdo y cordero.
Snacks	Bocadillos y aperitivos de consumo rápido.
Frutas Jugo	Jugos naturales y procesados de frutas.
Bebidas Alcohólicas	Bebidas con contenido de alcohol.
Suplemento	Suplementos nutricionales y vitamínicos.
Salsas	Condimentos y aderezos utilizados en la cocina.
Legumbres	Frijoles, lentejas, garbanzos y otras legumbres.
Caldos y Sopas	Sopas, caldos y cremas preparados.

Composición e ingesta nutricional

Aunque la ENSIN 2005 está desactualizada, proporcionó datos clave sobre la composición de los alimentos y la ingesta de nutrientes de cada persona. A partir de la información sobre la composición de los alimentos y los consumos reportados en el R24, se calculó la ingesta estimada de nutrientes para cada individuo. De estos datos, se seleccionaron 32 variables relacionadas con los macronutrientes y 8 variables con índices derivados (como la ingesta recomendada). Estas incluyen calorías, proteínas, grasas totales, colesterol, carbohidratos, así como diversos minerales y vitaminas.

Las variables relacionadas con los nutrientes y la composición molecular en la ENSIN 2005 abarcan un amplio espectro de macronutrientes y micronutrientes esenciales para el análisis nutricional. Dentro de los macronutrientes, se evaluaron las kilocalorías, las proteínas (incluyendo las de origen animal), los carbohidratos totales y concentrados, así como las grasas, desglosadas en saturadas, monoinsaturadas y poliinsaturadas. En el caso de los micronutrientes, se analizaron minerales como calcio, fósforo, hierro, sodio, potasio, magnesio, zinc, cobre y manganeso, esenciales para funciones metabólicas y estructurales. También se incluyeron vitaminas liposolubles, como la vitamina A en sus formas retinoide (AU) y carotenoide (AE), y vitaminas hidrosolubles, como tiamina, riboflavina, niacina, vitamina B6, ácido fólico, vitamina B12 y ácido ascórbico, todas cruciales para el metabolismo energético y la salud celular. Además, se evaluaron compuestos relacionados con la fibra, incluyendo fibra cruda y dietética, que son indicadores de la calidad alimentaria, y las cenizas, que reflejan el contenido de minerales en los alimentos. Finalmente, también se incluyó el colesterol como marcador de lípidos dietarios y su impacto en la salud cardiovascular [53].

En la ENSIN 2005, además de los valores absolutos de los nutrientes, se calcularon razones (R) para algunos nutrientes clave. Estas razones corresponden a la relación entre el valor del nutriente consumido y la recomendación diaria sugerida (ecuación 1), lo que permite evaluar la adecuación del consumo en relación con las necesidades nutricionales establecidas.

$$R = \frac{Consumo\ total\ del\ nutriente}{Consumo\ recomendado} \qquad (1)$$

Las variables disponibles en este sentido incluyen la R del calcio, R del ácido fólico, R de la vitamina A, R de la vitamina C, R de la vitamina B12, R de la fibra y R del zinc (**Tabla 3**). Este análisis es particularmente útil para identificar deficiencias o excesos en el consumo de nutrientes en diferentes grupos poblacionales, proporcionando una perspectiva más detallada sobre la calidad de la dieta y el cumplimiento de los requerimientos nutricionales establecidos.

Tabla 3. Tabla de compuestos disponibles para el análisis.

Nombre de la Variable	Descripción
Kilocalorías (kcal)	Energía total proporcionada por los alimentos consumidos.
Proteína (g)	Contenido total de proteínas en la dieta.
Grasa (g)	Cantidad total de grasas consumidas.
Grasa Polisaturada (g)	Porción de grasas poliinsaturadas en la dieta.
Grasa Monosaturada (g)	Porción de grasas monoinsaturadas en la dieta.
Grasa Saturada (g)	Porción de grasas saturadas en la dieta.
Carbohidratos (g)	Cantidad total de carbohidratos consumidos.
Calcio (mg)	Contenido de calcio en la dieta.

Colesterol (mg)	Cantidad de colesterol presente en los alimentos.
Fibra Cruda (g)	Porción de fibra no soluble presente en los alimentos.
Fibra Dietética (g)	Contenido total de fibra dietética en la dieta.
Cenizas (g)	Minerales residuales presentes en los alimentos.
Fósforo (mg)	Cantidad de fósforo consumido.
Hierro (mg)	Contenido de hierro en la dieta.
Sodio (mg)	Cantidad total de sodio consumido.
Potasio (mg)	Contenido de potasio en los alimentos.
Magnesio (mg)	Cantidad de magnesio presente en la dieta.
Zinc (mg)	Contenido de zinc en la dieta.
Cobre (mg)	Cantidad de cobre consumido.
Manganeso (mg)	Contenido de manganeso en los alimentos.
Vitamina A (µg RE) ***(RE = equivalentes de retinol)***	Vitamina A en su forma de retinol equivalente.
Vitamina E (mg α-TE) ***(α-TE = equivalentes de alfa-tocoferol)***	Vitamina A expresada en equivalentes de actividad retinol.
Tiamina (mg)	Cantidad de vitamina B1 en la dieta.
Vitamina B6 (mg)	Contenido de vitamina B6 consumido.
Ácido Fólico (µg)	Cantidad de ácido fólico en la dieta.
Vitamina B12 (µg)	Contenido de vitamina B12 en los alimentos.
Ácido Ascórbico (mg)	Vitamina C presente en la dieta.
Riboflavina (mg)	Cantidad de vitamina B2 en los alimentos.
Niacina (mg)	Contenido de vitamina B3 en la dieta.
Proteína Animal (g)	Porción de proteínas de origen animal en la dieta.
Carbohidratos Concentrados (g)	Contenido de carbohidratos de alta densidad energética.
Proteína ***(repetida)*** **(g)**	Cantidad total de proteína consumidas.
R Del Calcio	Relación entre el calcio consumido y la recomendación.
R Del Ácido Fólico	Relación entre el ácido fólico consumido y la recomendación.
R Vitamina A	Relación entre la vitamina A consumida y la recomendación.

R Vitamina C	Relación entre la vitamina C consumida y la recomendación.
R Vitamina B12	Relación entre la vitamina B12 consumida y la recomendación.
R Fibra	Relación entre la fibra consumida y la recomendación.
R Zinc	Relación entre el zinc consumida y la recomendación.
R Proteína	Relación entre la proteína consumida y la recomendación.

6. ANÁLISIS EXPLORATORIO DE DATOS.

6.1. La población

La población analizada en este estudio corresponde a una submuestra de los participantes de la Encuesta Nacional de la Situación Nutricional (ENSIN) 2005. Esta submuestra se conformó a partir de las personas a quienes se les aplicaron preguntas relacionadas con el diagnóstico de diabetes. Cabe señalar que algunas de estas preguntas estaban condicionadas por respuestas previas, por lo que no todos los encuestados respondieron al mismo conjunto de preguntas. La más relevante indagaba si el participante había sido diagnosticado alguna vez con diabetes por un profesional de salud. De los 35.297 participantes de la encuesta, únicamente 5.744 respondieron esta pregunta de forma afirmativa o negativa. Estas 5.744 personas constituyen la submuestra incluida en el presente análisis: 5.455 reportaron no haber recibido nunca un diagnóstico de diabetes, mientras que 289 indicaron haber sido diagnosticadas con la enfermedad, lo que representa una prevalencia del 5 % dentro de este grupo.

La proporción en la variable sexo (**Figura 2**) demuestra una predominancia de mujeres en la submuestra seleccionada y esta proporción se mantiene al considerar solamente los casos de enfermedad reportados, por lo cual no parece existir una afectación más drástica relacionadas con el sexo de la persona.

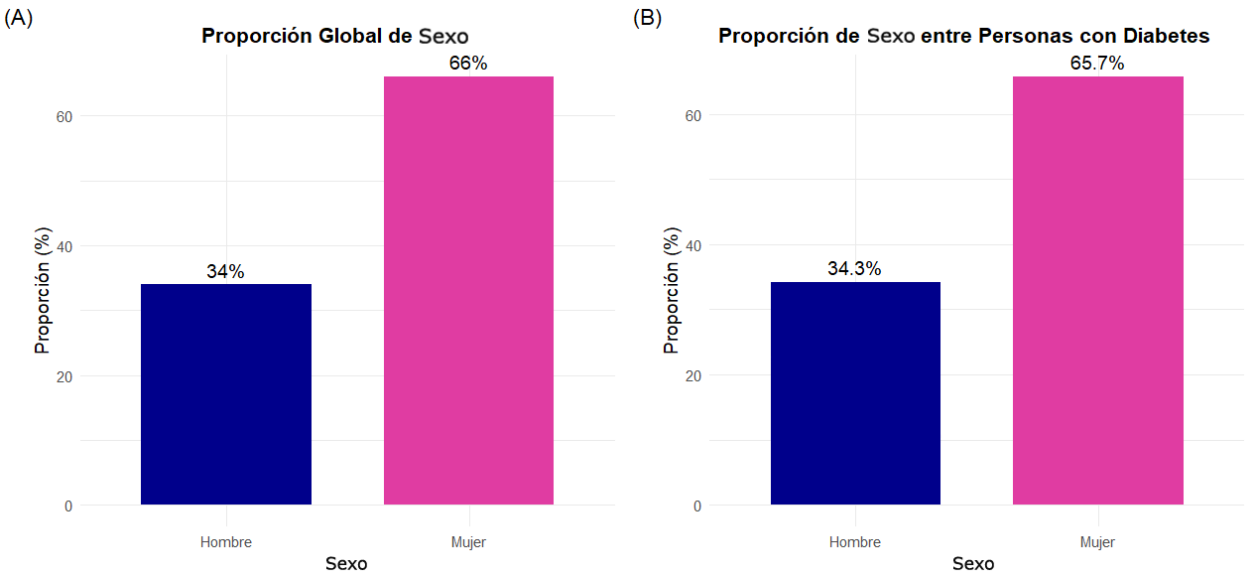

Figura 2. (A) Proporción de sexo observada para la submuestra de personas que cuentan con un auto-reporte de diabetes. (B) Proporción dentro del grupo confirmado de diagnosticados con la enfermedad

El rango de edad de las personas incluidas en el estudio abarca desde los 17 hasta los 64 años con una media muestral de 37 años. Las mediciones antropométricas reportadas en la encuesta permitieron calcular indicadores como el índice de masa corporal (IMC) para el 81% de las personas, una métrica ampliamente utilizada para evaluar el sobrepeso y la obesidad, condiciones reconocidas como factores de riesgo importantes para diversas enfermedades, incluida la diabetes.

La distribución del IMC en la muestra analizada presenta un comportamiento cercano a una distribución

normal, aunque con una cola más prolongada hacia valores altos, correspondientes a niveles de obesidad. Sin embargo, al examinar la relación entre el diagnóstico de diabetes y los valores de IMC, no se identificó un patrón claro que permita asociar el IMC con el diagnóstico de esta enfermedad (**Figura 3**).

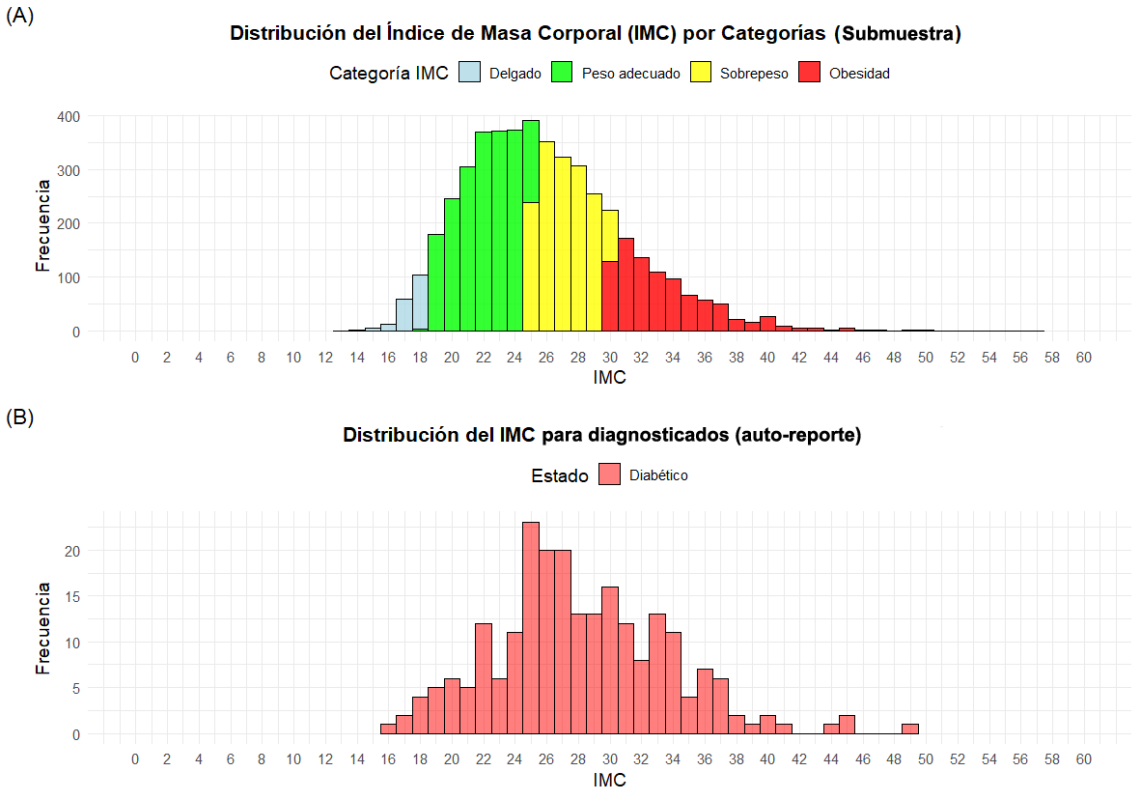

Figura 3. Comparativa de la distribución del índice de masa corporal en la submuestra de personas. (A) Distribución global del IMC en las submuestra (B) Distribución del IMC observada exclusivamente en las personas diabéticas.

El índice de masa corporal promedio para la submuestra es de 25.94, indicando una ligera tendencia hacia el sobrepeso en el momento en que se realizó la encuesta.

En la **Figura 4** se observa que los participantes de la submuestra no se distribuyen uniformemente en todo el país. La mayor concentración se encuentra en departamentos como Valle del Cauca y algunos del centro-occidente, lo que sugiere un posible sesgo hacia áreas con mayor población urbana. Sin embargo, la presencia de individuos en múltiples regiones muestra que el esfuerzo por representar diversas zonas geográficas de la ENSIN se mantiene hasta cierto punto en la submuestra generada. No obstante, la menor densidad en algunas áreas podría significar que ciertos contextos locales, especialmente en regiones más apartadas, estén menos representados en el análisis.

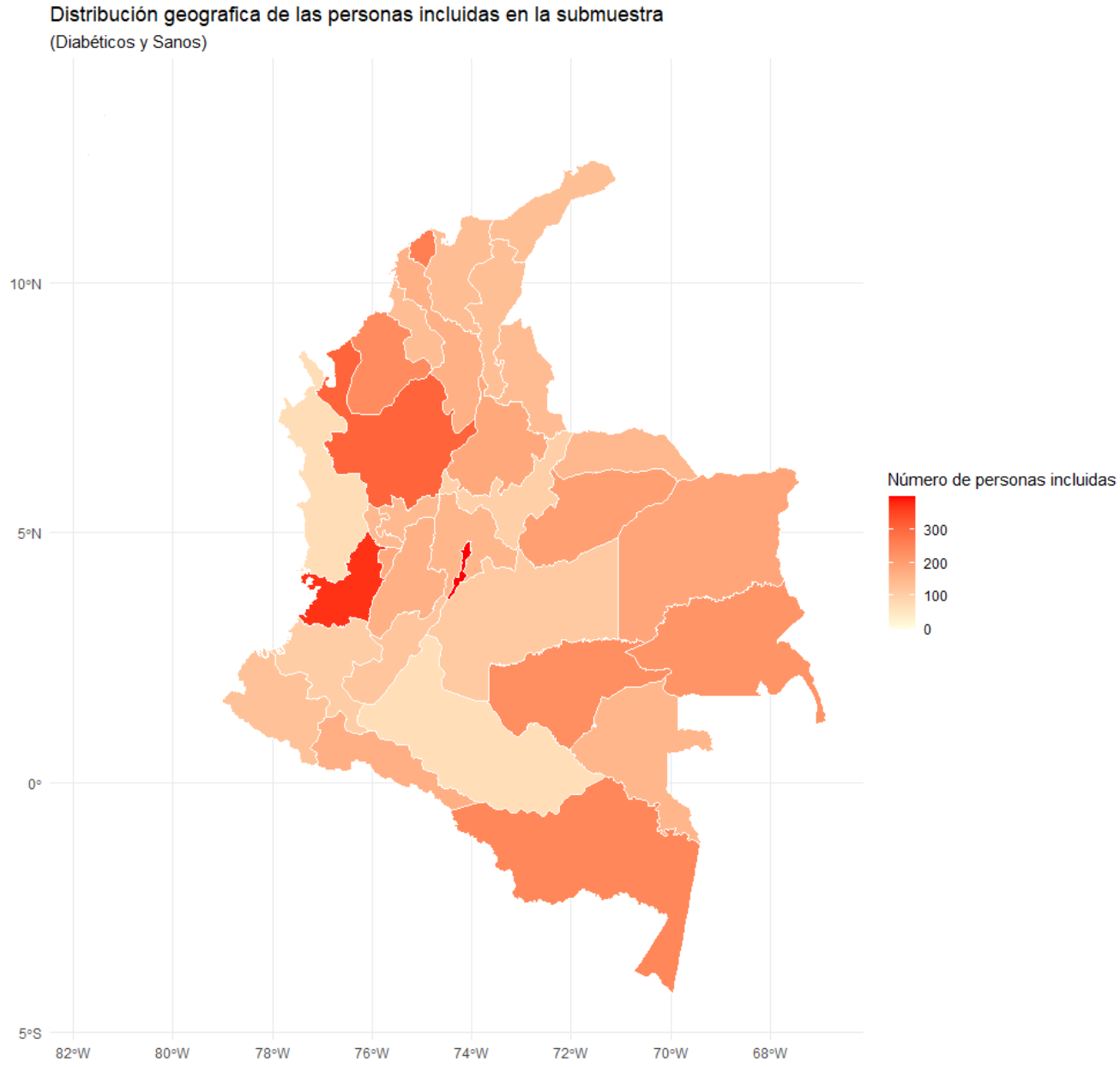

Figura 4. Número de personas incluidas en la submuestra por departamento.

6.2. Los alimentos

Se revela una marcada variabilidad en el consumo de los diferentes grupos de alimentos que se resumen en la **Tabla 4**. Todas las variables relacionadas con alimentos se expresan en unidades de veces consumidas por día. Se identifican categorías con valores medios y medianas relativamente altas, como los granos y cereales (media = 3,31; mediana = 3), aceites y grasas (media = 3,50; mediana = 3) y frutas (media = 2,89; mediana = 3), lo que sugiere una mayor presencia de estos alimentos en la dieta de la población evaluada.

En contraste, grupos como las legumbres (media = 0,11; mediana = 0), caldos y sopas (media = 0,08; mediana = 0) y salsas (media = 0,01; mediana = 0) presentan valores considerablemente más bajos, reflejando un consumo menos frecuente.

Tabla 4. Resumen estadístico de las variables alimenticias. La unidad de medida es veces consumido por día (reportada en el formato de 24 horas/R24).

Variable (Veces consumido)	Media	Mediana	Desviación Estándar	Mínimo	Máximo	Coeficiente de variación
Lácteos	1,48	1	1,44	0	11	0,97
Grano y cereales	3,31	3	2,01	0	18	0,61
Aceites / grasas	3,5	3	2,35	0	19	0,67
Verduras	2,27	2	2,36	0	20	1,04
Proteínas	0,52	0	0,72	0	9	1,38
Otro	0,17	0	0,5	0	7	2,94
Bebidas calientes	1,04	1	1,2	0	10	1,15
Dulce	2,86	3	2,06	0	18	0,72
Frutas	2,89	3	2,27	0	16	0,79
Carne blanca	0,65	0	0,9	0	9	1,38
Tubérculos	1,39	1	1,37	0	10	0,99
Alimento procesado	0,17	0	0,44	0	3	2,59
Bebidas azucaradas	0,53	0	0,85	0	7	1,60
Carne roja	0,93	1	0,97	0	6	1,04
Snacks	0,12	0	0,37	0	4	3,08
Frutas en jugo	0,08	0	0,31	0	4	3,88
Bebidas alcohólicas	0,09	0	0,36	0	5	4,00
Suplemento	0,14	0	0,46	0	8	3,29
Salsas	0,15	0	0,52	0	7	3,47
Legumbres	0,11	0	0,36	0	3	3,27
salsas	0,01	0	0,11	0	3	11,00
Caldos y sopas	0,08	0	0,33	0	4	4,13

La dispersión de los datos, evaluada a través de la desviación estándar y el coeficiente de variación (CV), varía entre las diferentes categorías de alimentos. Se observa una muy alta variabilidad relativa en el consumo de frutas en jugo (CV = 3,88), bebidas alcohólicas (CV = 4,00), caldos y sopas (CV = 4,13), snacks (CV = 3,08) y suplementos (CV = 3,29), lo cual indica diferencias marcadas en los patrones de consumo en relación con la media de consumo de estos productos, a pesar de tener promedios bajos. Esto puede sugerir que solo una parte de la población consume estos alimentos con frecuencia, mientras que la mayoría no los incluye regularmente en su dieta.

Por otro lado, categorías como granos y cereales (CV = 0,61), dulces (CV = 0,72) y aceites y grasas (CV = 0,67) presentan una variabilidad relativa menor, lo que sugiere un consumo más homogéneo en la muestra comparado con las categorías ya mencionadas, aunque igual es alto. Alimentos como los lácteos (CV = 0,97), verduras (CV = 1,04) y frutas (CV = 0,79) muestran una dispersión intermedia, reflejando diferencias moderadas entre los individuos.

Los valores mínimo y máximo permiten identificar la amplitud del consumo dentro de cada grupo. Mientras que algunas categorías presentan un rango amplio, como los aceites y grasas (mínimo = 0; máximo = 19), otras muestran un rango más reducido, como los snacks (mínimo = 0; máximo = 4). La presencia de valores máximos elevados en ciertos grupos indica la existencia de individuos con consumos notablemente altos, lo que podría estar asociado a hábitos alimentarios específicos dentro de la población estudiada.

Relación entre las variables - Alimentos

También se evaluó la relación entre las variables predictoras y la variable de respuesta, que en este caso corresponden a los alimentos y a la diabetes respectivamente, además entre las variables predictoras para analizar posibles colinealidades. En la **Figura 5** se observa que no existen correlaciones altas (>0.8) entre las variables predictoras para el conjunto de alimentos al usar el coeficiente de correlación de Pearson.

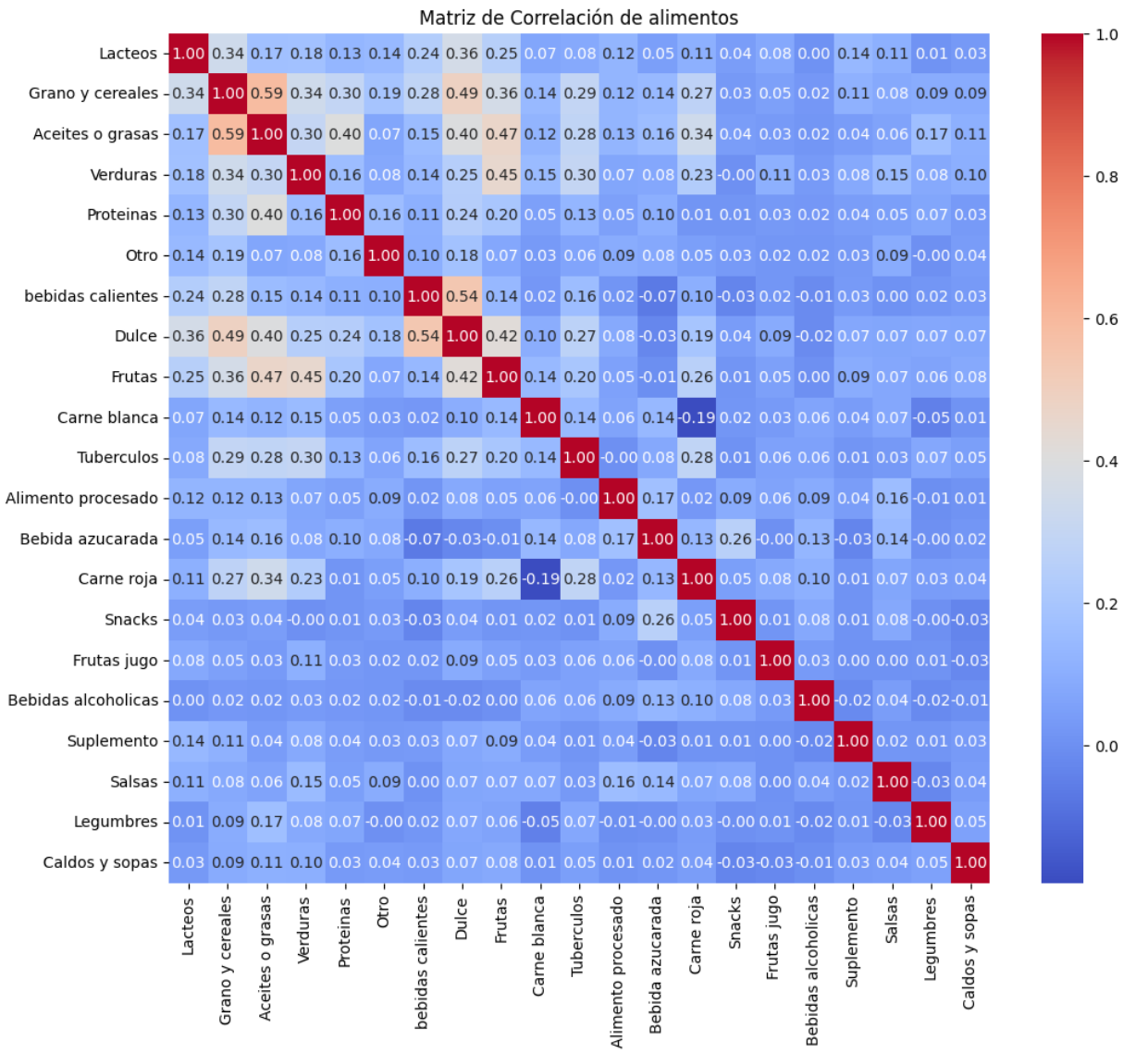

Figura 5. Matriz de Correlación de Pearson entre las 21 variables alimenticias.

Adicionalmente, se calculó la correlación punto biserial entre las 21 variables alimenticias y la variable objetivo, diabetes. Los resultados mostraron una relación débil, cercana a cero, entre todas las variables y la presencia de diabetes (**Figura *6***). Esto indica que no se observan diferencias significativas en la distribución de estas variables entre los dos grupos definidos por la variable diabetes.

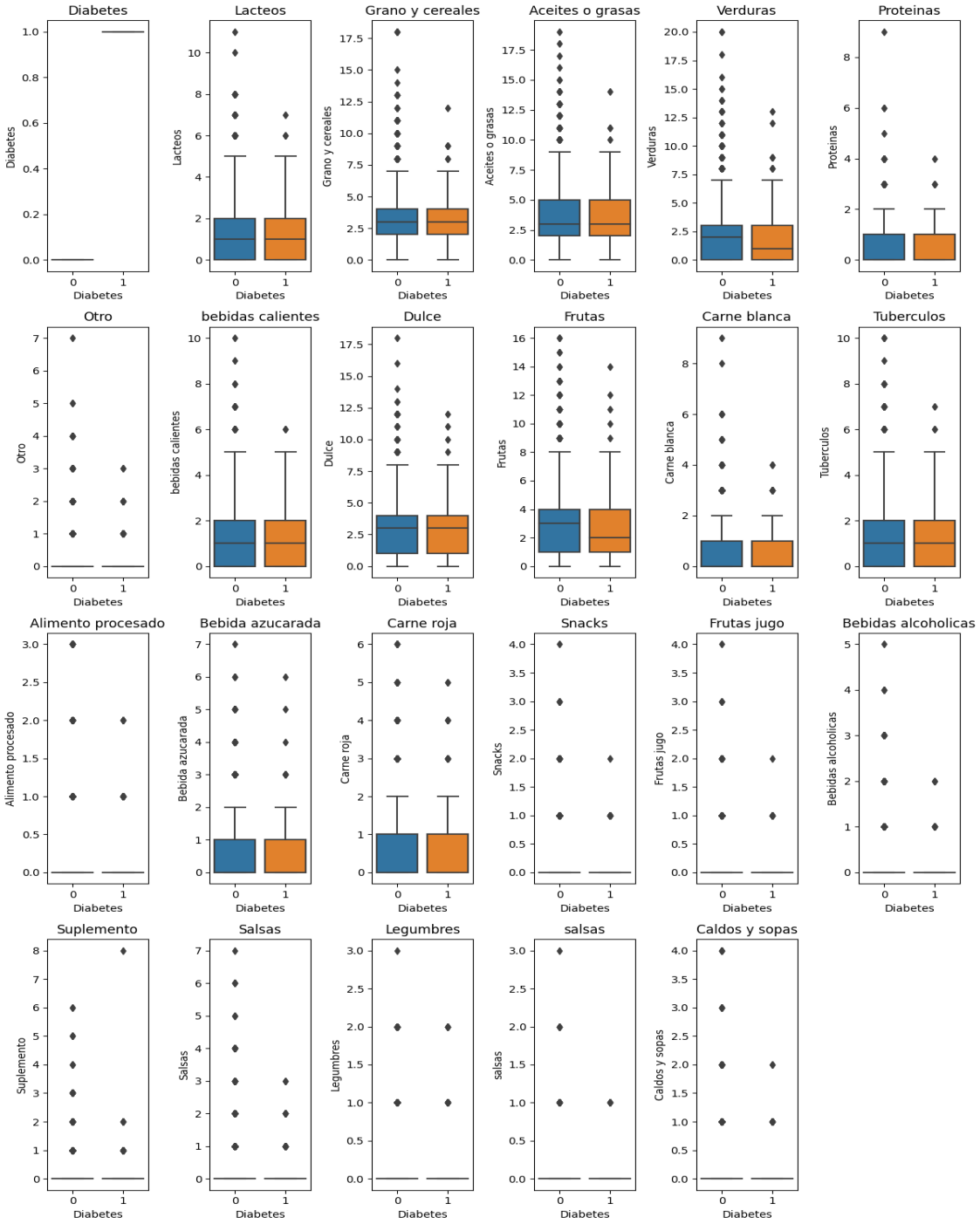

Figura 6. Relación de la variable Diabetes con las variables alimenticias.

6.3. Nutrientes de los alimentos

La **Tabla 5** muestra un resumen del comportamiento de las variables asociadas a los nutrientes ingeridos por las personas. Las kilocalorías tienen un promedio de 1629,64 y en términos de los macronutrientes principales, los carbohidratos lideran con una media de 251,42 g, seguidos por las proteínas con 54,49 g y las grasas con 44,86 g. Dentro de las grasas, las saturadas, monoinsaturadas y poliinsaturadas presentan valores similares en torno a 10-15 g.

Tabla 5. Resumen estadístico de las variables nutricionales/compuestos

Variable	Media	Mediana	Desviación Estándar	Mínimo	Máximo	Coeficiente de variación
Kilocalorías (kcal)	1629,64	1474,21	840,81	27,92	9330,7	0,52
Proteína (g)	54,5	49,91	27,93	0,88	252,49	0,51
Grasa (g)	44,86	38,08	30,4	0,07	310,14	0,68
Grasa Polisaturada (g)	10,5	8,06	8,81	0	87,15	0,84
Grasa Monosaturada (g)	15,68	12,95	11,55	0	138,15	0,74
Grasa Saturada (g)	15,81	12,94	12,53	0	147,57	0,79
Carbohidratos (g)	251,42	229,89	126,2	2,33	1194,65	0,50
Calcio (mg)	437,6	356,35	335,85	1,65	3011,11	0,77
Colesterol (mg)	237,61	179,86	205,44	0	1513,2	0,86
Fibra Cruda (g)	0,9	0,53	1,3	0	32,13	1,44
Fibra Dietética (g)	16,95	14,31	11,31	0	98,48	0,67
Cenizas (g)	9,1	8,19	5,15	0,11	104,85	0,57
Fósforo (mg)	818,21	756,3	415,37	14,24	3559,13	0,51
Hierro (mg)	10,99	8,62	10,37	0,09	141,39	0,94
Sodio (mg)	750,72	559,72	699,89	0	10737,57	0,93
Potasio (mg)	2377,03	2126,4	1338,68	0	11374,61	0,56
Magnesio (mg)	217,96	195,87	119,34	0	1234,94	0,55
Zinc (mg)	7,56	6,67	4,59	0	50,77	0,61
Cobre (mg)	2,94	0,91	37,62	0	1656,69	12,80
Manganeso (mg)	3,31	2,21	5,65	0	153,1	1,71
Vitamina A (µg)	6481,48	3388,17	9113,1	0	143065,4	1,41
Vitamina E (mg)	938,246	466,97	1704,96	0	18260,02	1,82
Tiamina (mg)	2,27	0,86	20,34	0,01	900,74	8,96
Vitamina B6 (mg)	0,61	0,03	3,11	0	80	5,10
Ácido Fólico (µg)	268,3	210,72	211,94	0	1807,37	0,79
Vitamina B12 (µg)	5,21	2,29	12,38	0	184,74	2,38
Ácido Ascórbico (mg)	100,96	63,56	123,49	0	1472,46	1,22
Riboflavina (mg)	1,33	1,01	1,32	0,01	20,46	0,99
Niacina (mg)	15,22	11,74	39,18	0,09	2860,51	2,57
Proteína Animal (g)	30,93	27,22	21,09	0	182,95	0,68
Carbohidratos Concentrados (g)	44,77	35,47	42,72	0	575,34	0,95
R DEL CALCIO	0,5	0,4	0,39	0	3,58	0,78
R DEL FOLICO	0,84	0,66	0,66	0	5,65	0,79
R VITAMINA A	1,74	0,86	3,17	0	34,74	1,82
R VITAMINA C	1,55	0,99	1,89	0	21,38	1,22

R VITAMINA B12	2,61	1,15	6,19	0	92,37	2,37
R FIBRA	0,6	0,52	0,39	0	3,94	0,65
R ZINC	0,78	0,69	0,48	0	6,35	0,62
R PROTEINA	1,02	0,95	0,5	0,02	4,46	0,49

Entre los minerales, destacan el calcio (437,59 mg; CV = 0,77), el fósforo (818,21 mg; CV = 0,51), el sodio (750,72 mg; CV = 0,93) y el potasio (2377,03 mg; CV = 0,56). Aunque el sodio tiene un valor medio elevado, su coeficiente de variación indica una alta dispersión relativa, lo que sugiere grandes diferencias en su consumo entre individuos. De forma similar, el hierro (CV = 0,94) y el zinc (CV = 0,61) presentan una variabilidad moderada, indicando diferencias relevantes en su ingesta dentro de la población.

Las vitaminas presentan aún mayor dispersión relativa. La vitamina A (6481,48 µg; CV = 1,41) y la vitamina E (938,25 mg; CV = 1,82) exhiben una alta variabilidad, lo que indica que mientras algunos individuos presentan consumos muy elevados, otros apenas las incorporan. La vitamina C (100,96 mg; CV = 1,22), el ácido fólico (268,3 µg; CV = 0,79) y la vitamina B12 (5,21 µg; CV = 2,38) también reflejan altos niveles de dispersión. Casos extremos como el del cobre (CV = 12,80) o la tiamina (CV = 8,96) sugieren que, aunque el promedio pueda parecer aceptable, el comportamiento del consumo entre individuos es muy desigual. En cuanto a los lípidos, la grasa total (44,86 g; CV = 0,68) y sus diferentes tipos, poliinsaturada (CV = 0,84), monoinsaturada (CV = 0,74) y saturada (CV = 0,79), también presentan una variabilidad considerable. Esto podría reflejar diferencias en las fuentes alimentarias o hábitos dietarios dentro del grupo estudiado.

El colesterol (237,61 mg; CV = 0,86) muestra una variabilidad alta, lo que sugiere que ciertos individuos tienen ingestas considerablemente superiores a la media. Por su parte, la fibra dietética (16,95 g; CV = 0,67) mantiene una dispersión más moderada, aunque con algunos casos de consumo nulo. Al analizar los datos en relación con los requerimientos diarios (R), se observa que la proteína (CV = 0,49) y el ácido fólico (CV = 0,79) presentan valores más cercanos a lo recomendado, y con menor variabilidad comparativa. Sin embargo, nutrientes como la vitamina A (CV = 1,82), la vitamina C (CV = 1,22) y la vitamina B12 (CV = 2,37) muestran una elevada variabilidad relativa respecto a los requerimientos, lo cual sugiere una ingesta excesiva en ciertos casos y deficiencias en otros.

Relación entre las variables - Compuestos

A diferencia de lo observado en la matriz de correlación de alimentos, para los compuestos existen múltiples relaciones fuertes que podrían traducirse en redundancia y problemas de multicolinealidad en un modelo (**Figura 7**). Este resultado era esperable debido a la forma en que se conformaron las variables y como algunas son derivados directos de otras. Por ejemplo, los R, son métricas calculas a partir de variables presentes, de manera que por ejemplo el R del calcio no es más que la división de la variable Calcio entre un numero predefinido. También hay otras variables que se podrían considerar como acumulativas o generales, por ejemplo, la variable “Grasa” implícitamente está definida por la unión de otras variables como Grasa Polisaturada, Grasa Monosaturada y Grasa Saturada y se observa que alcanzan valores de correlación >0.8 con la Grasa. La corrección de los problemas observados se realizó mediante la eliminación de la variable redundante con menos importancia.

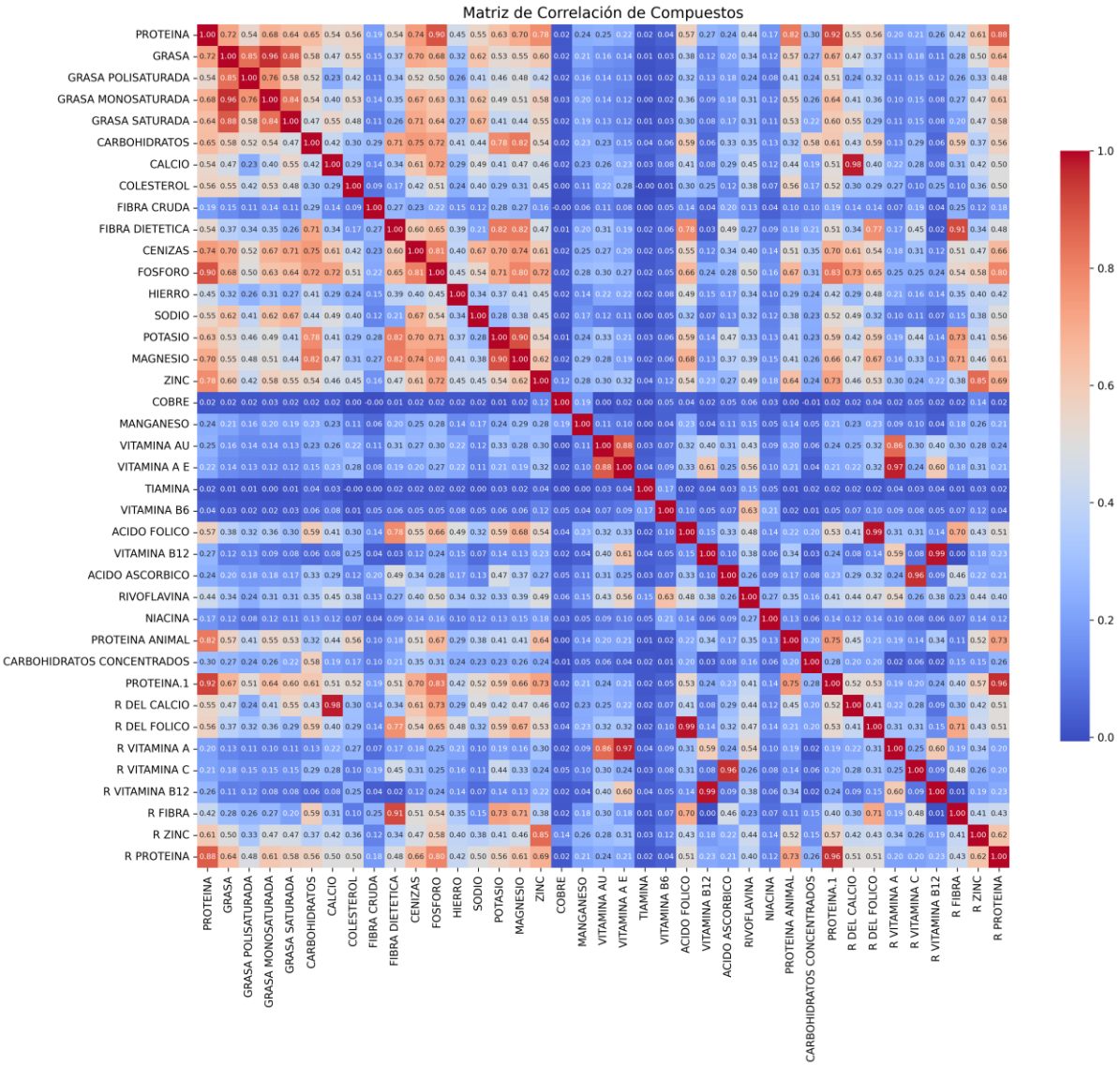

Figura 7. Matriz de Correlación de Pearson entre las 40 variables alimenticias.

La relación de las variables predictoras con diabetes en este caso, muestra un comportamiento muy similar a lo observado para alimentos, relaciones con correlaciones cercanas a cero (**Figura *8***). Por lo anterior no se puede asegurar que alguno de los dos conjuntos de datos preparados (Alimentos/Compuestos) tenga más potencial predictivo que el otro, al menos basándose solo en las relaciones de este nivel.

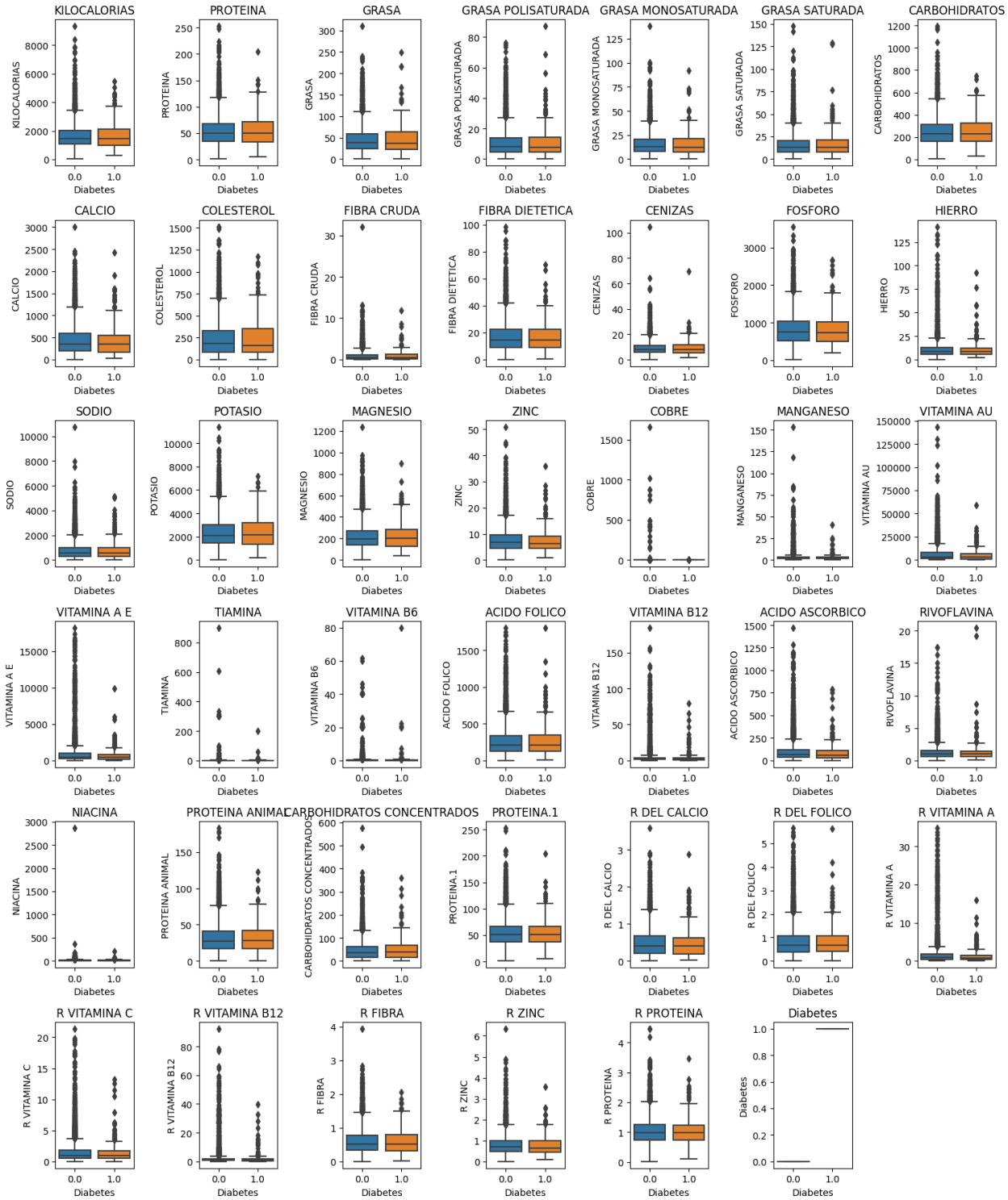

Figura 8. Relación de la variable Diabetes con las variables de composición/nutrientes.

7. DESARROLLO DEL MODELO

7.1. Definición de los conjuntos de entrenamiento

El análisis exploratorio de los datos determinó la viabilidad de los alimentos y su consumo como un conjunto de variables adecuado para generar un modelo. De manera similar, el conjunto de variables que agrupan datos de composición y nutrientes también demostró ser viable, aunque con algunas recomendaciones para el manejo de posibles redundancias. Por lo tanto, ambos conjuntos de datos fueron preparados junto con la etiqueta de diabetes y se usaron en paralelo para entrenar los modelos y evaluar cuál de los dos tiene un mayor potencial predictivo.

No se optó por integrar ambos conjuntos de datos en un único modelo ni por seleccionar un subconjunto común de variables debido a diversas razones. Primero, es importante destacar que el conjunto de datos de nutrientes ya incluye los datos de consumo de alimentos, ya que la ingesta reportada de cada nutriente se calcula a partir del consumo de alimentos por persona y la composición nutricional de cada alimento. Este proceso de cálculo de la ingesta nutricional produce un valor final acumulado que refleja la cantidad de cada nutriente consumido. Unificar estos conjuntos implicaría una redundancia, pues los datos de consumo ya están representados dentro de los valores de ingesta nutricional.

Además, los datos de consumo de alimentos se recogen mediante diversos instrumentos o cuestionarios, que pueden presentar variaciones inherentes debido a las diferencias en los patrones alimentarios de cada país o región geográfica. Este contexto variable hace que los datos del conjunto de consumo de alimentos sean difíciles de comparar entre diferentes contextos geográficos, a menos que se implementen diseños experimentales estrictos. Por otro lado, el conjunto de datos de compuestos y nutrientes está conformado por indicadores ampliamente utilizados en contextos internacionales, ya que se basan en un estándar para medir el estado nutricional de diversas poblaciones. Esto garantiza que los datos sean altamente comparables entre diferentes contextos internacionales, lo que los hace adecuados para análisis en diversas poblaciones.

Para asegurar la relevancia y aplicabilidad a nivel nacional, se desarrolló un modelo centrado en el consumo de alimentos, mientras que, con miras a realizar análisis más profundos en poblaciones internacionales, se creó también un modelo centrado en la ingesta nutricional. Esta estrategia permite abordar las necesidades de predicción tanto dentro del contexto colombiano como en el ámbito internacional, respetando las particularidades de cada enfoque.

7.2. Balanceo de datos

Para mejorar el rendimiento de los clasificadores, se implementaron procesos adicionales de balanceo de clases, una práctica recomendada en contextos de diagnóstico de enfermedades donde la cantidad de ejemplos positivos es limitada [36]. Con este objetivo, se evaluaron dos estrategias principales para ajustar el conjunto de datos al problema: SMOTE y ADASYN.

1. **SMOTE (Synthetic Minority Oversampling Technique):** Esta técnica genera ejemplos sintéticos de la clase minoritaria al interpolar entre instancias cercanas en el espacio de características, equilibrando los datos sin duplicar ejemplos idénticos y reduciendo el riesgo de sobreajuste.
2. **ADASYN (Adaptive Synthetic Sampling):** Crea ejemplos sintéticos de manera adaptativa, priorizando las instancias de la clase minoritaria más difíciles de clasificar. Esto mejora la capacidad

predictiva del clasificador frente a datos desbalanceados.

Ambas estrategias aumentan la representación de la clase minoritaria, fortalecen la robustez del modelo y optimizan su desempeño, especialmente en la detección de enfermedades raras.

Se evaluó el impacto de cada técnica y el porcentaje de creación de instancias sintéticas sobre el rendimiento de diferentes clasificadores (**Figura 9**). Los resultados mostraron un impacto positivo en el desempeño de los modelos, especialmente en el XGBoost. La relación entre el porcentaje de instancias creadas y el desempeño siempre fue positiva. Se estableció un límite de creación de instancias en 400% y se usó el F1 score para evaluar la correcta clasificación de la clase minoritaria. Las gráficas (**Figura 9**, A & B) mostraron un desempeño ligeramente superior al usar SMOTE, con un crecimiento exponencial y un F1 score final más alto en todos los modelos.

También se evaluó la reducción de la clase mayoritaria, que por sí sola no mejoró notablemente el desempeño de los modelos. Sin embargo, combinada con SMOTE y ADASYN, permitió cerrar la brecha entre clases más rápidamente. En las gráficas (**Figura 9**, C & D) se observa que combinar incrementos de la clase minoritaria con reducciones de la clase mayoritaria en valores crecientes (dos, tres, hasta cinco veces) logra un alto desempeño sin alcanzar el límite del 400% de síntesis de datos.

De forma independiente (caso no mostrado en las gráficas) se determinó que la configuración de estrategias de balanceo que mejor resultados obtiene en términos del f1 score es SMOTE al 400% sobre la clase minoritaria y reducción de la clase mayoritaria a un tercio.

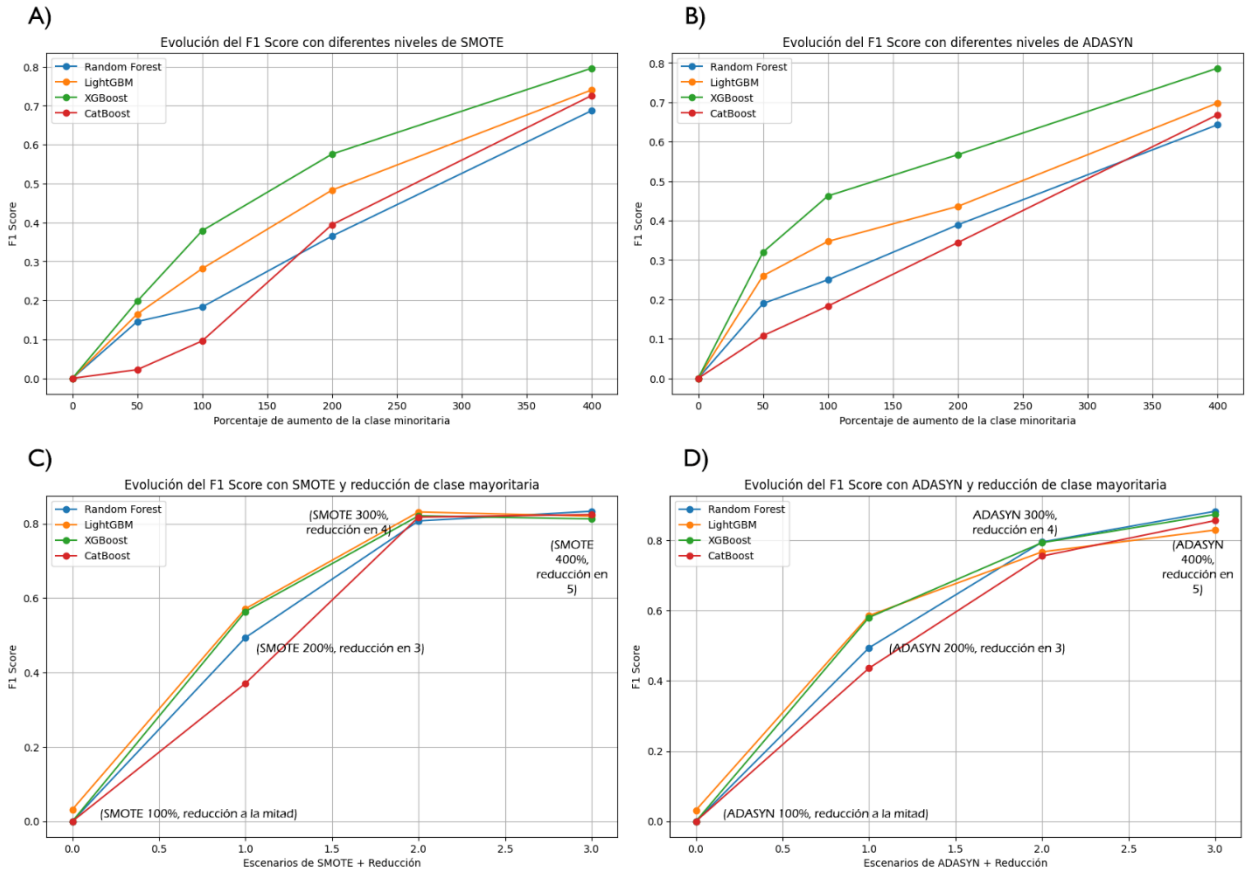

Figura 9. Demostración grafica del impacto de las estrategias de balanceo sobre el desempeño del modelo medido como F1 Score.

7.3. Pruebas preliminares de clasificadores

Los conjuntos de datos generados fueron preparados separando los alimentos de la etiqueta de diabetes en el caso del primer conjunto y los compuestos de la etiqueta de diabetes en el segundo. Posteriormente fueron sometidos a una evaluación preliminar utilizando diferentes clasificadores.

- Bosques aleatorios (Random forest)
- Máquinas de soporte vectorial (SVM)
- Regresión logística (Logistic Regression)
- K vecinos cercanos (KNN)
- CatBoost
- XGBoost
- Potenciación de gradiente (Gradient Boosting)
- Bayes (Naive Bayes)
- LightGBM

- AdaBoost
- Arboles de decisión (Decision Tree)

Durante este proceso, se observó una alta variabilidad en el poder predictivo de los modelos y un desempeño pobre en el recall en gran parte de los clasificadores evaluados, los mejores modelos se muestran en la **Tabla 6**. El desempeño general de los clasificadores fue mejor sobre el conjunto de datos con las variables de compuestos en comparación con el conjunto de alimentos.

Tabla 6. Métricas de los 7 mejores clasificadores evaluados con los dos conjuntos de datos.

Compuestos					
Classifier	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**Confusion Matrix**
XGBoost	0.870968	0.863014	0.790795	0.825328	[351, 30, 50 ,189]
CatBoost	0.866129	0.893939	0.740586	0.810069	[360, 21, 62 ,177]
RandomForest	0.867742	0.919786	0.719665	0.807512	[366, 15, 67 ,172]
LightGBM	0.854839	0.846512	0.761506	0.801762	[348, 33, 57 ,182]
KNN	0.753226	0.633540	0.853556	0.727273	[263, 118, 35 ,204]
DecisionTree	0.719355	0.630522	0.656904	0.643443	[289, 92, 82 ,157]
GradientBoosting	0.769355	0.811688	0.523013	0.636132	[352, 29, 114 ,125]

Alimentos					
Classifier	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**Confusion Matrix**
RandomForest	0.835294	0.855615	0.692641	0.765550	[337, 27, 71 ,160]
KNN	0.742857	0.614035	0.909091	0.732984	[232, 132, 21 ,210]
CatBoost	0.805042	0.804233	0.658009	0.723810	[327, 37, 79 ,152]
LightGBM	0.798319	0.776119	0.675325	0.722222	[319, 45, 75 ,156]
XGBoost	0.788235	0.746479	0.688312	0.716216	[310, 54, 72 ,159]
DecisionTree	0.734454	0.646586	0.696970	0.670833	[276, 88, 70 ,161]
GradientBoosting	0.726050	0.675258	0.567100	0.616471	[301, 63, 100 ,131]

Con el objetivo de generar un modelo robusto con una mayor capacidad de generalización se optó por modelar un método de aprendizaje por conjuntos (ENSEMBLE) que integrara los 5 mejores modelos obtenidos [56]. La decisión de incorporar el método ensemble se tomó tras observar el bajo rendimiento de los modelo en la métrica de recall, de suma importancia en este trabajo, pues identifica a las personas enfermas de diabetes. El modelo KNN presentó el recall más alto, pero también una de las precisiones más bajas. El ensemble puede aprovechar el potencial individual de cada modelo y crear una sinergia que les permita adaptarse a una mayor variedad de contextos.

7.4. Aprendizaje por conjuntos (ENSEMBLE)

Los 5 modelos seleccionados para el ENSEMBLE (XGBoost, CatBoost, Random Forest, LightGBM, KNN) fueron sometidos a un proceso de búsqueda de parámetros usando el método aleatorio. El número total de pruebas por modelo fue de 200 modelos diferentes (triales) con 5 pruebas de validación cruzada cada

uno, para un total de 1000 modelos entrenados para cada uno de los clasificadores del ENSEMBLE. Este proceso se realizó de forma independiente para el conjunto de Alimentos y para el conjunto de Compuestos. Modelos con menor número de parámetros como KNN fueron automáticamente entrenados con el número máximo de combinaciones.

Los mejores hiperparámetros encontrados para cada modelo se describen en la **Tabla 7**, junto con los desempeños de los cinco modelos bajo dichos parámetros. Se encontró que los mejores hiperparámetros encontrados en ambos conjuntos de datos (alimentos & compuestos) demostraban métricas similares, indicando una convergencia en el rendimiento, por lo cual se optó por usar el conjunto de hiperparámetros de los alimentos, que tenían métricas ligeramente superiores en el recall y F1-Score cuando se probaron en ambos conjuntos.

Tabla 7. Métricas de los 5 modelos incluidos en el ENSEMBLE cuando se usan los mejores hiperparámetros encontrados en el conjunto de datos de compuestos y alimentos.

Modelo	Conjunto	Accuracy	Precision	Recall	F1 Score	TN	FP	FN	TP	Hiperparámetros
XGBoost	Compuestos	0.9065	0.9132	0.8368	0.8734	362	19	39	200	'subsample': 0.8751, 'n_estimators': 300, 'min_child_weight': 1, 'max_depth': 12, 'learning_rate': 0.266, 'gamma': 0.0, 'colsample_bytree': 0.75
	Alimentos	0.8303	0.8125	0.7316	0.7699	325	39	62	169	
LightGBM	Compuestos	0.9081	0.8991	0.8577	0.8779	358	23	34	205	'subsample': 0.775, 'n_estimators': 475, 'min_child_weight': 1, 'max_depth': 7, 'learning_rate': 0.1, 'colsample_bytree': 0.85
	Alimentos	0.8134	0.7970	0.6970	0.7436	323	41	70	161	
CatBoost	Compuestos	0.9161	0.9083	0.8703	0.8889	360	21	31	208	'subsample': 0.775, 'learning_rate': 0.09, 'iterations': 450, 'depth': 9, 'colsample_bylevel': 0.925
	Alimentos	0.8336	0.8300	0.7186	0.7703	330	34	65	166	
Random Forest	Compuestos	0.8952	0.9350	0.7824	0.8519	368	13	52	187	'n_estimators': 350, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_depth': 50, 'bootstrap': False
	Alimentos	0.8336	0.8367	0.7100	0.7681	332	32	67	164	
KNN	Compuestos	0.7445	0.6158	0.9091	0.7343	233	131	21	210	'weights': 'uniform', 'n_neighbors': 1, 'metric': 'manhattan'
	Alimentos	0.7933	0.6667	0.9351	0.7784	256	108	15	216	

Los resultados muestran que el ajuste de hiperparámetros mediante búsqueda aleatoria produjo mejoras significativas en las métricas de desempeño de los modelos. En particular, los clasificadores basados en boosting (XGBoost, CatBoost y LightGBM) experimentaron aumentos notables en recall y F1-score, lo que indica una mejora en su capacidad para identificar correctamente las instancias positivas sin comprometer la precisión. Por ejemplo, XGBoost incrementó su recall de 0.7908 a 0.8368 para los compuestos y de 0.6883 a 0.7316 en el caso de los alimentos. LightGBM pasó de un recall de 0.7615 a 0.8577 en compuestos

y de 0.6753 a 0.7436 en alimentos, evidenciando una optimización en su capacidad de generalización al mejorar el rendimiento en ambos conjuntos.

De manera similar, Random Forest mostró una mejora en recall de 0.7197 a 0.7824 en Compuestos y de 0.6926 a 0.7100 en Alimentos, el F1-score aumento de 0.8075 a 0.8519 en Compuestos el incremento fue muy pequeño, de 0.7655 A 0.7681, lo que sugiere una reducción en la cantidad de falsos negativos em ambos casos. Por otro lado, KNN, si bien continúa presentando un desempeño inferior en comparación con los modelos de boosting y Random Forest, mostró un ligero incremento en F1-score en Compuestos pasando de 0.7273 a 0.7343 y un aumento sustancial en Alimentos pasando de 0.7329 a 0.7784.

8. EVALUACIÓN DEL RENDIMIENTO DEL MODELO

8.1. Métricas de desempeño

El modelo ENSEMBLE, construido a partir de cinco clasificadores (XGBoost, CatBoost, Random Forest, LightGBM y KNN) con hiperparámetros optimizados, mostró un desempeño superior en todas las métricas evaluadas. A lo largo de los 10 folds de validación cruzada, el modelo alcanzó un F1-score promedio de 0.9236, con una accuracy de 0.9419, precision de 0.9321 y un recall de 0.9159 (**Tabla 8**). Estos valores indican que el modelo logra un buen equilibrio entre la cantidad de predicciones correctas y su capacidad para identificar correctamente las clases positivas.

Tabla 8. Resumen del rendimiento del EMSEMBLE en los diferentes ensayos de validación cruzada.

Compuestos					Alimentos				
Fold	Accuracy	Precision	Recall	F1	Fold	Accuracy	Precision	Recall	F1
1	0,913	0,912	0,858	0,884	1	0,849	0,814	0,793	0,804
2	0,932	0,930	0,892	0,911	2	0,869	0,808	0,871	0,838
3	0,923	0,886	0,916	0,901	3	0,896	0,921	0,802	0,857
4	0,942	0,932	0,916	0,924	4	0,822	0,794	0,733	0,762
5	0,932	0,915	0,908	0,911	5	0,842	0,822	0,759	0,789
6	0,945	0,955	0,899	0,926	6	0,852	0,827	0,785	0,805
7	0,939	0,939	0,899	0,919	7	0,838	0,813	0,757	0,784
8	0,906	0,909	0,840	0,873	8	0,805	0,752	0,739	0,746
9	0,955	0,965	0,916	0,940	9	0,852	0,809	0,809	0,809
10	0,916	0,904	0,874	0,889	10	0,872	0,829	0,844	0,836
Promedio	0,930	0,925	0,892	0,908	Promedio	0,850	0,819	0,789	0,803

La capacidad del modelo para diferenciar entre clases es notablemente alta, como se evidencia en la curva ROC donde se observa un área bajo la curva (AUC-ROC) de 0.96 para los compuestos y 0.92 para los alimentos (**Figura 10**). Esto significa que el modelo tiene una alta capacidad de discriminación entre las clases (Diagnosticados con diabetes y no diagnosticados).

Asimismo, la curva Precision-Recall refuerza la efectividad del modelo en contextos donde el balance entre precisión y sensibilidad es crucial, con un AUC-PR de 0.96 para Compuestos y 0.90 para Alimentos (**Figura 10**). Esto indica que el modelo mantiene una alta precisión incluso en casos donde la clase positiva es menos frecuente, lo que sugiere un manejo efectivo del desbalanceo de clases sobre todo al usar el conjunto de datos de compuestos.

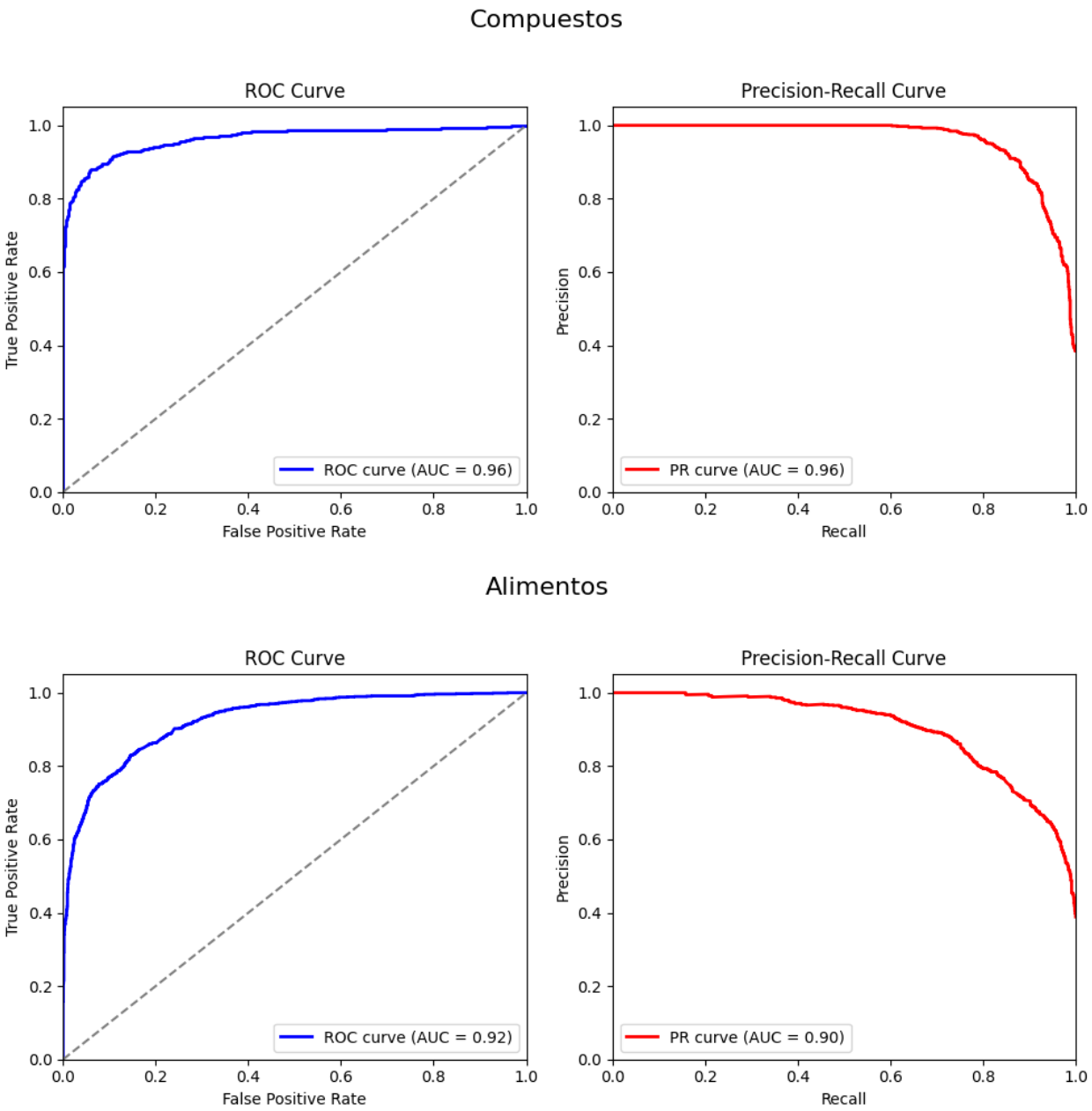

Figura 10. Curvas de desempeño del modelo EMSEMBLE. A la izquierda la curva ROC y a la derecha la curva Precision-recall (PR).

8.2. Desempeño del ENSEMBLE

Para corroborar si el modelo ENSEMBLE es la mejor opción, se realizaron pruebas estadísticas (t-test) sobre los valores de las métricas obtenidas por validación cruzada para el modelo ensemble y los cinco modelos

individuales (**Tabla 9**), utilizando el mismo número de folds de validación cruzada. Este proceso se llevó a cabo para extraer valores de las cuatro métricas principales, las cuales demostraron tener un comportamiento normal mediante la prueba de Shapiro.

Los resultados de las pruebas t mostraron que el modelo ensemble tiene un desempeño significativamente superior al de los cinco modelos individuales en casi todas las métricas. La única métrica en la que no se encontraron diferencias significativas fue el recall del modelo KNN (0.90 en la **Tabla 7**). Por lo tanto, se concluye que el modelo ensemble es significativamente superior a los mejores modelos individuales, lo que demuestra que su capacidad de predicción se ajusta mejor a este problema de clasificación particular.

Tabla 9. Pruebas estadísticas comparativas entre los cinco mejores modelos y el ENSEMBLE usando el conjunto de datos de Compuestos.

Métrica	Modelo Comparado	t (Prueba t)	p (Prueba t)	W (Wilcoxon)	p (Wilcoxon)	Interpretación
Accuracy	XGB	8,21	0,0000	0,00	0,00	Diferencia significativa en ambas pruebas (Ensemble muestra un desempeño superior)
Precision	XGB	3,99	0,0032	1,00	0,00	Diferencia significativa en ambas pruebas (Ensemble muestra un desempeño superior)
Recall	XGB	7,61	0,0000	0,00	0,00	Diferencia significativa en ambas pruebas (Ensemble muestra un desempeño superior)
F1	XGB	8,24	0,0000	0,00	0,00	Diferencia significativa en ambas pruebas (Ensemble muestra un desempeño superior)
Accuracy	LightGBM	11,07	0,0000	0,00	0,00	Diferencia significativa en ambas pruebas (Ensemble muestra un desempeño superior)
Precision	LightGBM	6,59	0,0001	0,00	0,00	Diferencia significativa en ambas pruebas (Ensemble muestra un desempeño superior)
Recall	LightGBM	6,09	0,0002	0,00	0,01	Diferencia significativa en ambas pruebas (Ensemble muestra un desempeño superior)
F1	LightGBM	11,27	0,0000	0,00	0,00	Diferencia significativa en ambas pruebas (Ensemble muestra un desempeño superior)
Accuracy	CatBoost	4,41	0,0017	0,00	0,00	Diferencia significativa en ambas pruebas (Ensemble muestra un desempeño superior)
Precision	CatBoost	2,62	0,0277	6,00	0,03	Diferencia significativa en ambas pruebas (Ensemble muestra un desempeño superior)
Recall	CatBoost	3,95	0,0033	1,00	0,01	Diferencia significativa en ambas pruebas (Ensemble muestra un desempeño superior)
F1	CatBoost	4,49	0,0015	0,00	0,00	Diferencia significativa en ambas pruebas (Ensemble muestra un desempeño superior)
Accuracy	RandomForest	6,33	0,0001	0,00	0,00	Diferencia significativa en ambas pruebas (Ensemble muestra un desempeño superior)

Precision	RandomForest	-2,92	0,0170	5,00	0,02	Diferencia significativa en ambas pruebas (Ensemble muestra un desempeño superior)
Recall	RandomForest	10,73	0,0000	0,00	0,00	Diferencia significativa en ambas pruebas (Ensemble muestra un desempeño superior)
F1	RandomForest	6,84	0,0001	0,00	0,00	Diferencia significativa en ambas pruebas (Ensemble muestra un desempeño superior)
Accuracy	KNN	24,00	0,0000	0,00	0,00	Diferencia significativa en ambas pruebas (Ensemble muestra un desempeño superior)
Precision	KNN	24,94	0,0000	0,00	0,00	Diferencia significativa en ambas pruebas (Ensemble muestra un desempeño superior)
Recall	KNN	0,85	0,4158	21,00	0,56	No hay evidencia suficiente en ninguna prueba
F1	KNN	24,96	0,0000	0,00	0,00	Diferencia significativa en ambas pruebas (Ensemble muestra un desempeño superior)

9. INTERPRETACIÓN DE DATOS Y MODELOS

Se hizo uso de gráficos SHAP, los cuales muestran el impacto de cada variable en la predicción del modelo. En estos gráficos, cada punto representa una observación, y el eje horizontal indica cómo la variable contribuye a la predicción. Los valores en el gráfico se centran en el impacto de cada variable, con valores positivos que indican una contribución hacia la predicción de la clase de interés (en este caso, diabetes), mientras que los valores negativos indican una contribución hacia la clase opuesta (personas no diagnosticadas).

Los colores en el gráfico (como los mostrados en la figura 11-14) reflejan el valor de la característica para cada observación. El color rojo corresponde a valores altos de la variable, y el color azul indica valores bajos. De esta forma, se puede observar cómo una variable influye en la predicción dependiendo de su valor específico. Por ejemplo, si una variable como frutas muestra una mayor concentración de puntos azules en la parte negativa del gráfico, podría interpretarse que el bajo consumo de frutas está asociado con una menor probabilidad de ser clasificado como diabético. Por otro lado, si una variable como bebidas azucaradas muestra puntos rojos en el lado positivo, sugiere que un mayor consumo de estas bebidas contribuye a una mayor probabilidad de ser clasificado como diabético.

9.1. Alimentos

9.1.1. Modelos individuales

Se realizó un análisis individual de cada uno de los modelos que conforman el ensemble y que son compatibles con el enfoque interpretativo utilizado: CatBoost, XGBoost, LightGBM y Random Forest. El modelo K-Nearest Neighbors (KNN) fue excluido del análisis debido a su incompatibilidad con SHAP, dado que no se basa en árboles de decisión ni posee una función de predicción diferenciable que permita atribuir valores de forma robusta.

Los resultados obtenidos para los cuatro modelos muestran una convergencia en la importancia de variables como Frutas, Bebidas Calientes y Dulces, en las cuales se observa una amplia dispersión de los valores a lo largo del eje x/SHAP (**Figura *11***). Esto indica que estas variables tienen un impacto significativo tanto en la predicción de individuos no diagnosticadas como de individuos diagnosticados con diabetes. En particular, Frutas resulta ser la variable más relevante en todos los modelos, mostrando un mayor consumo entre las personas no diagnosticadas y un consumo más bajo (menor frecuencia) en los diagnosticados con diabetes. Bebidas Calientes está dentro del top 5 en todos los modelos, con un patrón similar al de las frutas. En contraste, los Dulces presentan un comportamiento opuesto, donde un mayor consumo se asocia con la predicción de individuos diagnosticados con diabetes, mientras que un menor consumo está vinculado a las personas no diagnosticadas. Este patrón refleja una correlación, pero no debe interpretarse como una relación causal directa.

Otras variables que se destacaron fueron las Proteínas, que aparecieron en el top 5 de importancia en Random Forest y LightGBM (**Figura *11*** C, D), con un patrón que sugiere un mayor consumo en personas no diagnosticadas y un menor consumo en personas diagnosticadas con diabetes. Los Aceites o Grasas también fueron relevantes, siendo una de las variables más importantes en LightGBM y XGBoost, aunque su relación con los dos grupos de personas muestra una dispersión más amplia y los valores que toma la

variable no forman una tendencia, indicando una asociación más compleja.

El comportamiento observado en las variables es coherente con el conocimiento actual sobre la relación entre la dieta y la diabetes [57], [58], [59], [60]. Sin embargo, es importante recalcar que, aunque estas asociaciones son consistentes con lo que se sabe en la literatura, no se puede asumir causalidad debido a que los valores SHAP reflejan asociaciones estadísticas, no relaciones causales. Para establecer causalidad, se necesitarían estudios adicionales con una temporalidad prospectiva y más controlados que validen estas observaciones.

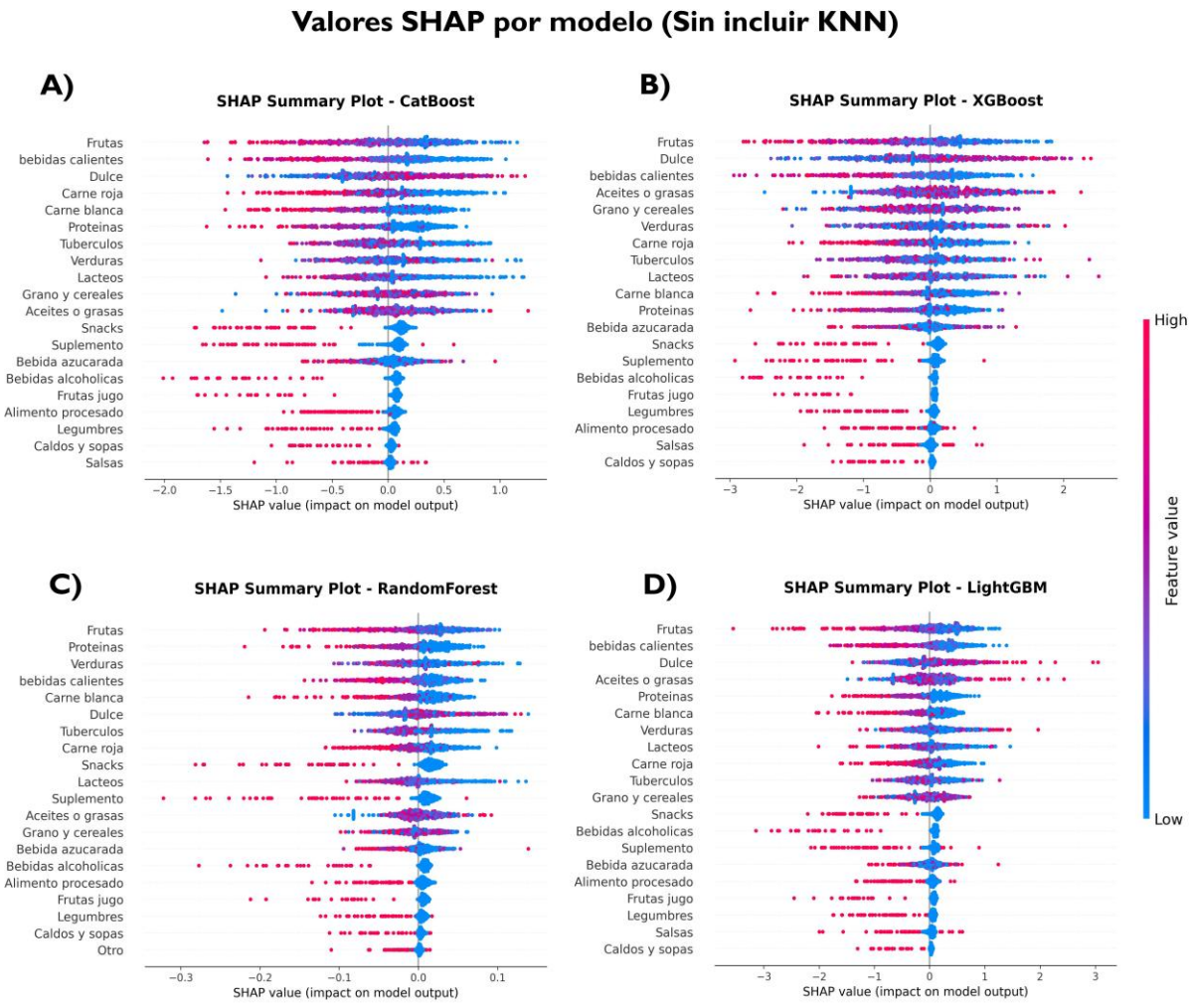

Figura 11. Importancia de variables y valores SHAP asociados a las predicciones de los diferentes modelos que componen el ENSEMBLE evaluados en el Dataset de alimentos. (A) Grafica para el modelo CatBoost. (B) Grafica para el modelo XGBoost. (C) Grafica para el modelo Random Forest. (D) Grafica para el modelo LightGBM.

9.1.2. Modelo ENSEMBLE y Alimentos

El modelo conjunto mantuvo las bebidas calientes como la variable más importante en la predicción,

seguida de dos novedades, los suplementos y los tubérculos. Las bebidas calientes, los suplementos y el consumo de ambos en general se relaciona con personas no diagnosticadas mientras que los tubérculos y su consumo se relaciona con la predicción de diabetes. Los aceites y grasas siguen mostrando un comportamiento difuso sin una tendencia especifica (**Figura *12***, A).

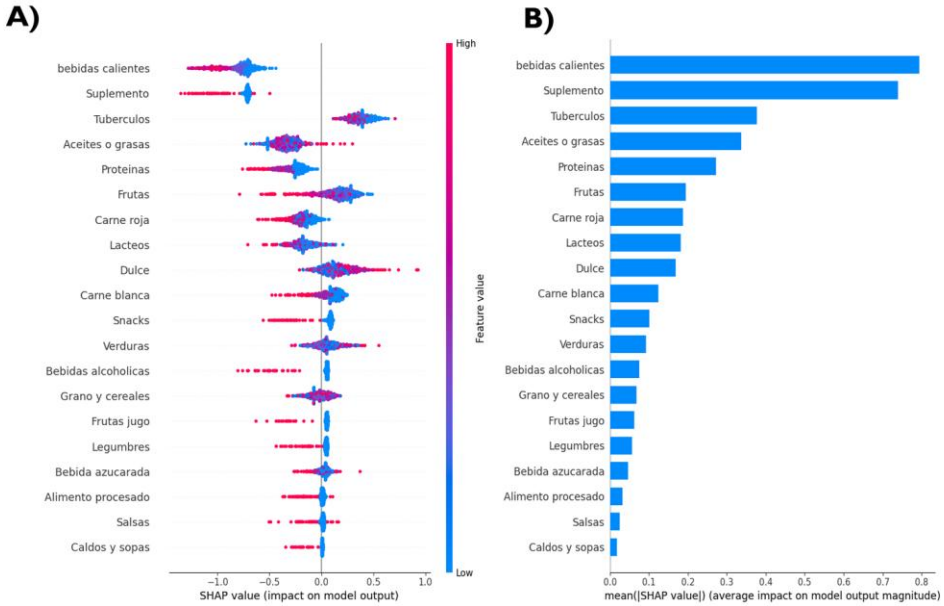

Figura 12. Importancia de variables y valores SHAP asociados a la predicción global del modelo de votación (ENSEMBLE) en el conjunto de Alimentos. (A) grafica de puntos y distribución de las predicciones en cada variable. (B) Las 20 variables más importantes del modelo de acuerdo al método SHAP.

En el ENSEMBLE las proteínas, los dulces y las frutas siguen jugando un papel importante en la predicción y contrario a las tres variables más importantes, los mencionados son relevantes tanto para la predicción de personas no diagnosticadas como de personas diagnosticadas.

9.2. Compuestos

9.2.1. Modelos individuales

Los resultados de la evaluación de los valores shap en los modelos individuales se resumen en la **Figura *13***. En el caso particular del modelo CatBoost, las cinco variables con mayor importancia en la predicción fueron la Vitamina B6, la proteína de origen animal, la fibra cruda, el colesterol y la vitamina A. Estas variables mostraron una amplia dispersión en sus valores SHAP, lo que indica que influyen en la clasificación tanto de casos positivos (personas con diabetes) como negativos (personas no diagnosticadas). La Vitamina B6, la vitamina A y la fibra cruda están en el top 5 de variables más importantes para los cuatro modelos evaluados, indicando que son variables con una importancia global. Por otro lado, el magnesio esta presente en el top 5 de variables mas importantes en tres de los cuatro modelos considerados.

Además de la importancia de las variables, se observaron patrones en los valores que toman las variables clave cuando el modelo clasifica a una persona como diabética. En el modelo CatBoost, por ejemplo, la proteína de origen animal, la fibra cruda, el sodio, el ácido ascórbico y el manganeso tienden a tener valores

más altos en individuos clasificados como diagnosticados con diabetes. Por el contrario, variables como el colesterol y la vitamina A (VitaminaAU) presentaron valores más elevados en personas clasificadas como no diagnosticadas (**Figura *13*A**), lo cual sugiere una asociación estadística entre niveles más altos de estos nutrientes y una menor probabilidad predicha de diabetes en el modelo. Este patrón podría reflejar diferencias de consumo entre los grupos, pero no debe interpretarse como un efecto causal o protector sin un análisis adicional de tipo causal o clínico**.**

En el modelo XGBoost variables como la vitamina A, el Zinc y la Niacina mostraron una correlación entre valores más altos para esas variables y personas no diagnosticadas (**Figura *13*B**). Además, al igual que en el primer modelo, mayores cantidades consumidas de Proteína Animal se relacionan con diabetes en el modelo.

En el caso del RandomForest la vitamina A sigue mostrando el mismo comportamiento que en otros modelos, con valores más altos correlacionados con personas no diagnosticadas. El potasio por otro lado muestra una correlación de valores bajos con diabetes y en este modelo destaca que la dispersión de los valores es menor.

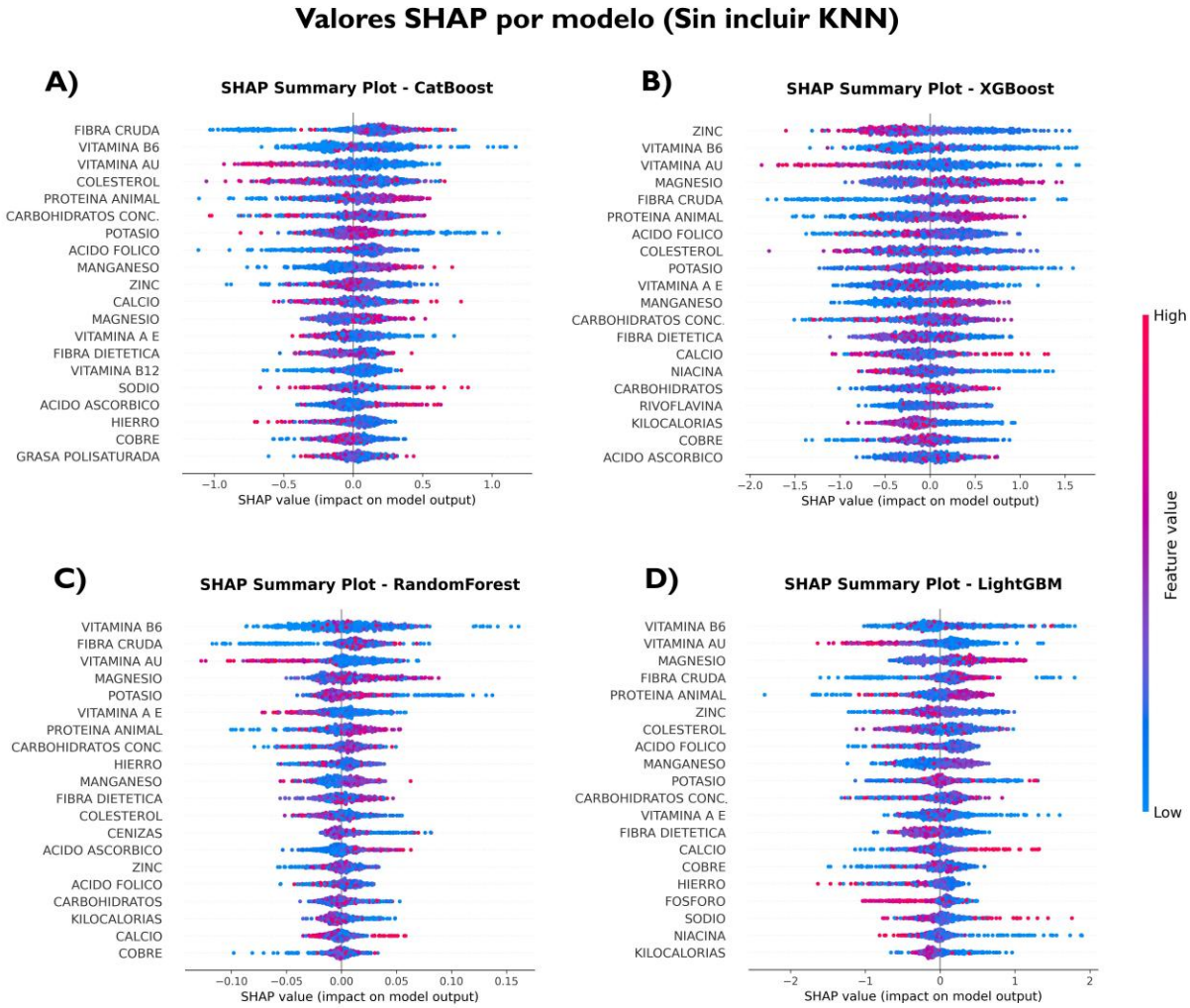

Figura 13. Importancia de variables y valores SHAP asociados a las predicciones de los diferentes modelos que componen el ENSEMBLE evaluados en el Dataset de compuestos/nutrientes. (A) Grafica para el modelo CatBoost. (B) Grafica para el modelo XGBoost. (C) Grafica para el modelo Random Forest. (D) Grafica para el modelo LightGBM.

El análisis individual sugiere que ciertos nutrientes, como la Vitamina B6, la vitamina A y la fibra cruda desempeñan un papel relevante y consistente en la predicción de la diabetes a través de distintos modelos de aprendizaje automático.

9.2.2. Modelo ENSEMBLE y Compuestos/nutrientes

El análisis de importancia global de características del modelo ensemble arrojó resultados sustancialmente diferentes a los obtenidos a partir de los modelos individuales. Nutrientes como la vitamina B6, la fibra cruda y la proteína animal perdieron protagonismo, en su lugar el potasio, la energía y las grasas monosaturadas se posicionan entre las variables más relevantes para diferenciar entre personas con y sin diagnóstico de diabetes (**Figura *14***).

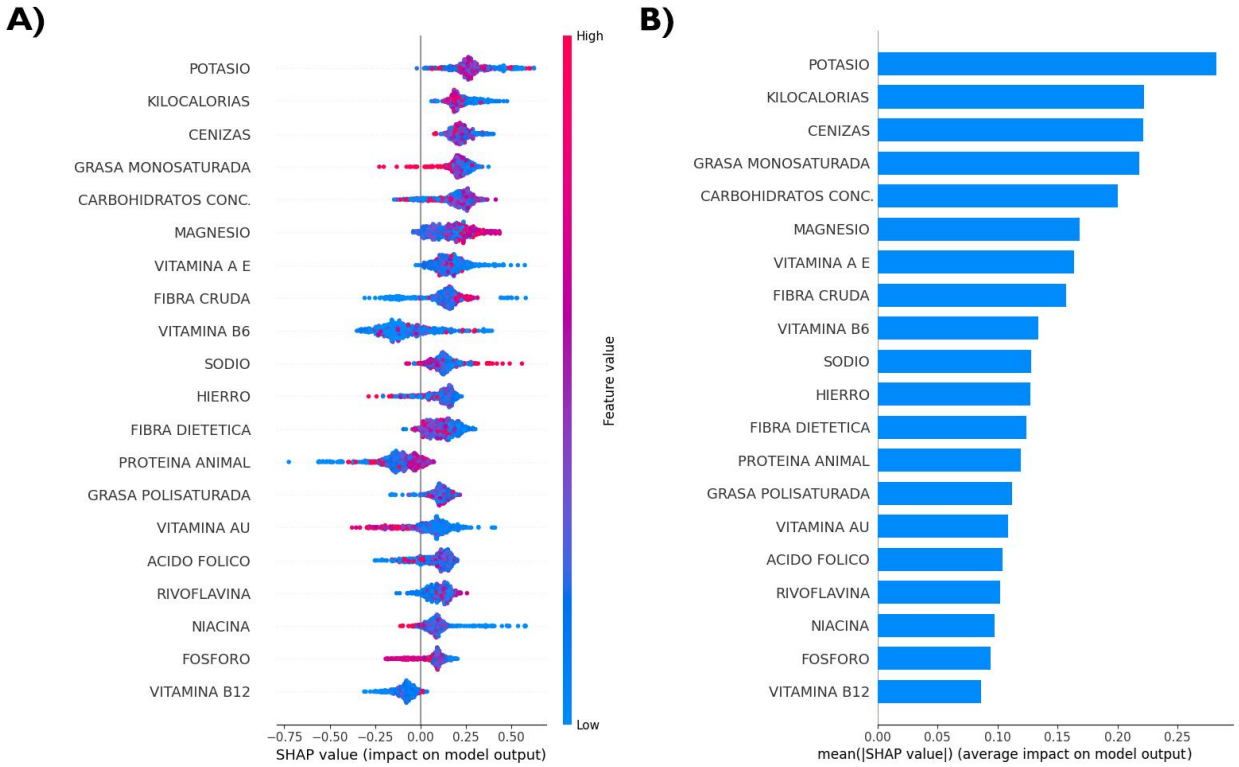

Figura 14. Importancia de variables y valores SHAP asociados a la predicción global del modelo de votación (ENSEMBLE) en el conjunto de compuestos/nutrientes. (A) grafica de puntos y distribución de las predicciones en cada variable. (B) Las 20 variables mas importantes del modelo de acuerdo al método SHAP.

La distribución de los valores SHAP asignados a cada registro revela una diferenciación más clara entre los grupos en comparación con los modelos individuales. Variables como el sodio, el potasio, los carbohidratos y las kilocalorías muestran una fuerte concentración de valores SHAP positivos, lo cual indica que contribuyen mayoritariamente a clasificaciones como diagnosticados con diabetes. En contraste, nutrientes como la vitamina B12, la proteína de origen animal y la vitamina B6 presentan una predominancia de valores SHAP negativos, asociados a clasificaciones como personas no diagnosticadas.

Al examinar los patrones internos dentro de cada variable, se observan comportamientos específicos. En el caso de la vitamina B12, los valores más bajos tienden a estar asociados a predicciones de personas no diagnosticadas, mientras que para el sodio ocurre lo opuesto: niveles más altos de esta variable se asocian con clasificaciones como diagnosticados con diabetes. En el caso de la grasa monosaturada, la vitamina A y el fosforo los valores SHAP más negativos se concentran en niveles altos de consumo, sugiriendo que este patrón es característico en personas clasificadas como no diagnosticadas.

10. CONCLUSIONES Y TRABAJOS FUTUROS

10.1. Conclusiones

Este estudio desarrolló dos modelos predictivos usando un enfoque de modelos de votación o ENSEMBLE, que integró múltiples algoritmos supervisados de machine learning incluyendo XGBoost, CatBoost, Random Forest, LightGBM y K-Nearest Neighbors, para predecir diabetes a partir de información de consumo de alimentos e ingesta nutricional en la población colombiana. Ambos modelos generados alcanzaron desempeños óptimos, pero se evidencio un mayor potencial predictivo cuando se usan datos de ingesta de nutrientes. El ENSEMBLE alcanzo tasas de 78.9% en la predicción adecuada de personas diagnosticadas con diabetes cuando se usó información de consumo de alimentos y un 89.2% cuando se usó información de la ingesta de nutrientes, minerales y proteínas. Las curvas ROC de ambos modelos están por encima del 90%, indicando una capacidad de predicción alta y especifica.

El estudio enfrentó un escenario inicial de desbalance significativo entre los datos de personas diagnosticadas y no diagnosticadas con diabetes. Para abordar este desafío, se implementaron estrategias de balanceo como SMOTE y ADASYN, las cuales permitieron mejorar la representatividad de las clases y optimizar el proceso de modelado, reflejándose en resultados más robustos y generalizables. Durante el desarrollo del proyecto se cumplieron satisfactoriamente los objetivos propuestos, logrando recopilar, procesar y depurar la información proveniente de la Encuesta Nacional de la Situación Nutricional (ENSIN 2005). Este proceso incluyó la sistematización de variables dietéticas y el manejo riguroso de valores atípicos, lo que resultó en una base de datos coherente y adecuada para el análisis estadístico y predictivo.

El análisis exploratorio de los datos permitió identificar patrones relevantes en la relación entre el consumo alimentario y la presencia de diabetes, así como otras variables asociadas con condiciones de salud y enfermedades crónicas, que podrían ser objeto de futuras investigaciones bajo el mismo enfoque metodológico. La arquitectura del modelo predictivo se optimizó mediante la búsqueda sistemática de hiperparámetros, la aplicación de técnicas de balanceo de clases y la validación cruzada, asegurando una adecuada capacidad de generalización y estabilidad del modelo. La utilización de dos conjuntos de datos, uno basado en patrones alimentarios específicos y otro en la ingesta nutricional cuantificada, permitió tanto el análisis contextualizado a nivel nacional como la posibilidad de comparaciones internacionales gracias a la estandarización de los nutrientes evaluados.

La interpretación de los resultados del modelo permitió identificar variables clave para la predicción de diabetes. Entre los nutrientes, la vitamina B6, la fibra y la vitamina A sobresalieron como factores de importancia, mientras que, en términos de grupos de alimentos, las bebidas calientes, los dulces y ciertas frutas mostraron una influencia significativa en la predicción. Estos hallazgos proporcionan una base sólida para futuras investigaciones clínicas y epidemiológicas orientadas a validar y profundizar en estos vínculos en el contexto nacional. En conjunto, este trabajo demuestra el valor de integrar técnicas avanzadas de inteligencia artificial con datos epidemiológicos y nutricionales para fortalecer la capacidad predictiva y apoyar la toma de decisiones en salud pública.

10.2. Trabajos futuros

Como actividades complementarias a futuro, se ha propuesto refinar el modelo ensemble mediante la evaluación del impacto que tendría la eliminación de uno o varios de los algoritmos que lo componen. Esta estrategia busca reducir el costo computacional del ensemble, optimizando las tareas de clasificación y

predicción sin comprometer de forma significativa su rendimiento. La revisión estructural del modelo permitirá identificar los algoritmos más relevantes y, en consecuencia, justificar una posible simplificación manteniendo una alta capacidad predictiva.

Adicionalmente, es fundamental realizar una validación externa del modelo utilizando datos completamente independientes a los empleados durante el entrenamiento. Este paso es esencial para evaluar su capacidad de generalización y mitigar el riesgo de sobreajuste. No obstante, una limitación importante es la escasa disponibilidad de bases de datos recientes y de alta calidad en el contexto latinoamericano. Para enfrentar esta dificultad, se plantea la exploración de encuestas nutricionales comparables a la Encuesta Nacional de Situación Nutricional (ENSIN) de Colombia en otros países con características socioeconómicas, alimentarias y demográficas similares. Esto permitiría probar el modelo en contextos diversos, incrementando la robustez y validez externa de sus predicciones.

Una limitación adicional identificada en este estudio se relaciona con la definición de la variable dependiente "diabetes", la cual, en la mayoría de encuestas nacionales como la ENSIN, no distingue entre los distintos tipos de esta enfermedad. Esta falta de especificidad dificulta establecer relaciones directas con factores de riesgo que son diferenciales según el tipo de diabetes, particularmente en el caso de la diabetes tipo 2, cuya asociación con el estilo de vida y la dieta está bien documentada. Esta limitación introduce un grado de incertidumbre en la interpretación de los resultados, ya que el modelo podría estar capturando señales provenientes de tipos de diabetes con etiologías distintas. En futuros trabajos, se espera mitigar esta limitación mediante la integración de fuentes de datos que incluyan información clínica más detallada sobre el diagnóstico, clasificación y evolución del tipo de diabetes, permitiendo así un análisis más específico y dirigido.

Finalmente, se considera relevante la incorporación de variables adicionales, como factores genéticos, comportamentales y ambientales, que no fueron contemplados en este estudio, pero que podrían enriquecer el modelo y mejorar su capacidad de capturar la complejidad multifactorial del riesgo de diabetes. La inclusión de estas variables permitiría avanzar hacia una predicción más integral y personalizada.

11. AGRADECIMIENTOS

La realización de esta tesis de maestría fue posible gracias al respaldo financiero del programa **Good Food Fellows**, iniciativa que, junto con **Food EDU**, es facilitada por la **American Heart Association** y la **Alliance of Bioversity & CIAT** en su calidad de co-secretarías de la **Periodic Table of Food Initiative (PTFI),** con el generoso apoyo de la **Fundación Rockefeller**.

Asimismo, expreso mi especial gratitud al Instituto de Investigación en Ciencias Ómicas, centro de excelencia asociado al PTFI, por brindarme un entorno propicio para la discusión científica y el desarrollo de ideas innovadoras. De la misma forma, valoro profundamente la guía de las y los tutores afiliados al programa Good Food Fellows y la colaboración de mis compañeras y compañeros de proyecto, cuyo intercambio de conocimientos y apoyo mutuo resultaron decisivos para el desarrollo de esta investigación.

Agradezco de manera muy especial a las profesoras Delia Ortega (directora) y Juliana Chaura (codirectora), por su guía rigurosa, sus comentarios siempre constructivos y la confianza depositada en mí desde el inicio hasta la culminación de este trabajo.

12. **REFERENCIAS BIBLIOGRÁFICAS**

[1] Organización Mundial de la Salud, “Diabetes,” News. Accessed: Jun. 06, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] Organización panamericana de la salud, “Diabetes - OPS/OMS | Organización Panamericana de la Salud.” Accessed: May 12, 2024. [Online]. Available: https://www.paho.org/es/temas/diabetes

[3] Y. Luo *et al.*, “Diet-Related Lipidomic Signatures and Changed Type 2 Diabetes Risk in a Randomized Controlled Feeding Study With Mediterranean Diet and Traditional Chinese or Transitional Diets,” *Diabetes Care*, vol. 46, no. 9, pp. 1691–1699, Jul. 2023, doi: 10.2337/dc23-0314.

[4] M. Anjum, R. Saher, and M. N. Saeed, “Optimizing type 2 diabetes management: AI-enhanced time series analysis of continuous glucose monitoring data for personalized dietary intervention,” *PeerJ Comput Sci*, vol. 10, p. e1971, Apr. 2024, doi: 10.7717/PEERJ-CS.1971/SUPP-2.

[5] T. T. Fung, M. McCullough, R. M. Van Dam, and F. B. Hu, “A Prospective Study of Overall Diet Quality and Risk of Type 2 Diabetes in Women,” *Diabetes Care*, vol. 30, no. 7, pp. 1753–1757, Jul. 2007, doi: 10.2337/DC06-2581.

[6] S. Jacobs *et al.*, “A priori-defined diet quality indexes and risk of type 2 diabetes: the Multiethnic Cohort,” *Diabetologia*, vol. 58, no. 1, 2015, doi: 10.1007/s00125-014-3404-8.

[7] W. Shen *et al.*, “Associations of a proinflammatory diet, habitual salt intake, and the onset of type 2 diabetes: A prospective cohort study from the UK Biobank,” *Diabetes Obes Metab*, vol. 26, no. 6, pp. 2119–2127, Jun. 2024, doi: 10.1111/DOM.15517,.

[8] X. Wang *et al.*, “Dietary Sodium Intake and Risk of Incident Type 2 Diabetes,” *Mayo Clin Proc*, vol. 98, no. 11, pp. 1641–1652, Nov. 2023, doi: 10.1016/j.mayocp.2023.02.029.

[9] L. Gan *et al.*, “Dietary carbohydrate intake and risk of type 2 diabetes: a 16-year prospective cohort study,” *Sci China Life Sci*, vol. 68, no. 4, Apr. 2025, doi: 10.1007/S11427-024-2804-0,.

[10] J. Zhu, W. Xu, S. Wu, and D. Song, “Vitamin B6 status, type 2 diabetes mellitus, and periodontitis: evidence from the NHANES database 2009–2010,” *BMC Oral Health*, vol. 25, no. 1, Dec. 2025, doi: 10.1186/S12903-025-05597-Z,.

[11] Y. Zhu *et al.*, “Joint B Vitamin Intake and Type 2 Diabetes Risk: The Mediating Role of Inflammation in a Prospective Shanghai Cohort,” *Nutrients*, vol. 16, no. 12, Jun. 2024, doi: 10.3390/NU16121901,.

[12] E. Ekure *et al.*, “A systematic review of diabetes risk assessment tools in sub-Saharan Africa,” *Int J Diabetes Dev Ctries*, vol. 42, no. 3, pp. 380–393, Jul. 2022, doi: 10.1007/S13410-022-01045-8/TABLES/2.

[13] J. Henson, O. Anyiam, and D. Vishnubala, “Type 2 Diabetes,” *Exercise Management for Referred Medical Conditions*, pp. 223–252, Jun. 2023, doi: 10.4324/9781315102399-12.

[14] Organización Panamericana de la Salud, “Diabetes - OPS/OMS | Organización Panamericana de la Salud.” Accessed: May 19, 2024. [Online]. Available: https://www.paho.org/es/temas/diabetes

[15] D. Petras *et al.*, “High-resolution liquid chromatography tandem mass spectrometry enables large scale molecular characterization of dissolved organic matter,” *Front Mar Sci*, vol. 4, no. DEC, p. 301420, Dec. 2017, doi: 10.3389/FMARS.2017.00405/BIBTEX.

[16] D. D. Mondal, U. Chakraborty, M. Bera, S. Ghosh, and D. Kar, “An overview of nutritional profiling in foods: Bioanalytical techniques and useful protocols,” *Front Nutr*, vol. 10, p. 1124409, Mar. 2023, doi: 10.3389/FNUT.2023.1124409/BIBTEX.

[17] B. Zhou, J. F. Xiao, L. Tuli, and H. W. Ressom, “LC-MS-based metabolomics,” *Mol Biosyst*, vol. 8, no. 2, p. 470, 2012, doi: 10.1039/C1MB05350G.

[18] M. Grootveld, B. C. Percival, and K. L. Grootveld, “Chronic non-communicable disease risks presented by lipid oxidation products in fried foods,” *Hepatobiliary Surg Nutr*, vol. 7, no. 4, p. 305, Aug. 2018, doi: 10.21037/HBSN.2018.04.01.

[19] Y. Zhang *et al.*, “Cooking oil/fat consumption and deaths from cardiometabolic diseases and other causes: prospective analysis of 521,120 individuals,” *BMC Med*, vol. 19, no. 1, pp. 1–14, Dec. 2021, doi: 10.1186/S12916-021-01961-2/FIGURES/3.

[20] E. Azzini *et al.*, “Total and Plant Protein Consumption: The Role of Inflammation and Risk of Non-Communicable Disease,” *Int J Mol Sci*, vol. 23, no. 14, Jul. 2022, doi: 10.3390/IJMS23148008/S1.

[21] World Health Organization, “Use of Nutrition Data in Decision Making: A Review Paper,” https://cdn.who.int/media/docs/default-source/nutritionlibrary/team---technical-expert-advisory-group-on-nutrition-monitoring/team-nutrition-data-decisionmaking-reviewpaper.pdf?sfvrsn=43b43e17_2&download=true.

[22] Organización Panamericana de la Salud, “Ante el aumento en el número de casos en todo el mundo, que se han cuadruplicado en los últimos decenios, es necesario tomar medidas urgentes contra la diabetes - OPS/OMS | Organización Panamericana de la Salud.” Accessed: May 27, 2025. [Online]. Available: https://www.paho.org/es/noticias/14-11-2024-ante-aumento-numero-casos-todo-mundo-que-se-han-cuadruplicado-ultimos-decenios

[23] B. Pretorius, J. M. Muka, P. J. M. Hulshof, and H. C. Schönfeldt, “Current practices, challenges and new advances in the collection and use of food composition data for Africa,” *Front Sustain Food Syst*, vol. 7, p. 1240734, Jul. 2023, doi: 10.3389/FSUFS.2023.1240734/BIBTEX.

[24] R. S. Bernstein, M. C. Marshall, and A. L. Carney, “Effects of Dietary Composition on Adipose Tissue Hexokinase-II and Glucose Utilization in Normal and Streptozotocin-diabetic Rats,” *Diabetes*, vol. 26, no. 8, pp. 770–779, Aug. 1977, doi: 10.2337/DIAB.26.8.770.

[25] S. Ahmed *et al.*, “Foodomics: A Data-Driven Approach to Revolutionize Nutrition and Sustainable Diets,” *Front Nutr*, vol. 9, p. 874312, May 2022, doi: 10.3389/FNUT.2022.874312/BIBTEX.

[26] F. Capozzi and A. Bordoni, “Foodomics: a new comprehensive approach to food and nutrition,” *Genes Nutr*, vol. 8, no. 1, p. 1, Jan. 2013, doi: 10.1007/S12263-012-0310-X.

[27] L. A. DiMeglio, C. Evans-Molina, and R. A. Oram, “Type 1 diabetes,” *The Lancet*, vol. 391, no. 10138, pp. 2449–2462, Jun. 2018, doi: 10.1016/S0140-6736(18)31320-5.

[28] A. Alkhatib *et al.*, “Functional Foods and Lifestyle Approaches for Diabetes Prevention and

Management," *Nutrients 2017, Vol. 9, Page 1310*, vol. 9, no. 12, p. 1310, Dec. 2017, doi: 10.3390/NU9121310.

[29] L. T. Crummett and M. H. Aslam, "Diabetes websites lack information on dietary causes, risk factors, and preventions for type 2 diabetes," *Front Public Health*, vol. 11, p. 1159024, Jul. 2023, doi: 10.3389/FPUBH.2023.1159024/BIBTEX.

[30] G. Sikand, P. Kris-Etherton, and N. M. Boulos, "Impact of Functional Foods on Prevention of Cardiovascular Disease and Diabetes," *Curr Cardiol Rep*, vol. 17, no. 6, pp. 1–16, Jun. 2015, doi: 10.1007/S11886-015-0593-9/TABLES/2.

[31] J. B. Xiao and P. Hogger, "Dietary Polyphenols and Type 2 Diabetes: Current Insights and Future Perspectives," *Curr Med Chem*, vol. 22, no. 1, pp. 23–38, Dec. 2014, doi: 10.2174/09298673216661407061308O7.

[32] O. P. Chauhan, *Advances in Food Chemistry: Food Components, Processing and Preservation*. Springer Nature, 2022.

[33] S. Brinkley *et al.*, "The state of food composition databases: data attributes and FAIR data harmonization in the era of digital innovation," *Front Nutr*, vol. 12, p. 1552367, Mar. 2025, doi: 10.3389/FNUT.2025.1552367/BIBTEX.

[34] R. Naz *et al.*, "Food Polyphenols and Type II Diabetes Mellitus: Pharmacology and Mechanisms," *Molecules 2023, Vol. 28, Page 3996*, vol. 28, no. 10, p. 3996, May 2023, doi: 10.3390/MOLECULES28103996.

[35] J. P. P. Vieira *et al.*, "Metabolite Profiling in a Diet-Induced Obesity Mouse Model and Individuals with Diabetes: A Combined Mass Spectrometry and Proton Nuclear Magnetic Resonance Spectroscopy Study," *Metabolites*, vol. 13, no. 7, p. 874, Jul. 2023, doi: 10.3390/METABO13070874/S1.

[36] R. A. Danquah, "Handling Imbalanced Data: A Case Study for Binary Class Problems," Oct. 2020, Accessed: Jan. 16, 2025. [Online]. Available: http://arxiv.org/abs/2010.04326

[37] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst Appl*, vol. 73, pp. 220–239, May 2017, doi: 10.1016/J.ESWA.2016.12.035.

[38] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/JAIR.953.

[39] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," *Proceedings of the International Joint Conference on Neural Networks*, pp. 1322–1328, 2008, doi: 10.1109/IJCNN.2008.4633969.

[40] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785/SUPPL_FILE/KDD2016_CHEN_BOOSTING_SYSTEM_01-ACM.MP4.

[41] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree", Accessed: May 23, 2025. [Online]. Available: https://github.com/Microsoft/LightGBM.

[42] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features", doi: 10.5555/3327757.3327770.

[43] L. Breiman, "Random forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324/METRICS.

[44] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification," *IEEE Trans Inf Theory*, vol. 13, no. 1, pp. 21–27, 1967, doi: 10.1109/TIT.1967.1053964.

[45] T. G. Dietterich, "Ensemble Methods in Machine Learning," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 1857 LNCS, pp. 1–15, 2000, doi: 10.1007/3-540-45014-9_1.

[46] L. Breiman, "Bagging predictors," *Mach Learn*, vol. 24, no. 2, pp. 123–140, 1996, doi: 10.1007/BF00058655/METRICS.

[47] Z. H. Zhou, "Ensemble methods: Foundations and algorithms," *Ensemble Methods: Foundations and Algorithms*, pp. 1–218, Jan. 2012, doi: 10.1201/B12207.

[48] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," *Adv Neural Inf Process Syst*, vol. 2017-December, pp. 4766–4775, May 2017, Accessed: Apr. 10, 2025. [Online]. Available: https://arxiv.org/abs/1705.07874v2

[49] J. Liu, E. L. Chou, K. K. Lau, P. Y. M. Woo, J. Li, and K. H. K. Chan, "Machine learning algorithms identify demographics, dietary features, and blood biomarkers associated with stroke records," *J Neurol Sci*, vol. 440, p. 120335, Sep. 2022, doi: 10.1016/j.jns.2022.120335.

[50] M. Sebek and G. Menichetti, "Network Science and Machine Learning for Precision Nutrition," *Precision Nutrition: the Science and Promise of Personalized Nutrition and Health*, pp. 367–402, Jan. 2024, doi: 10.1016/B978-0-443-15315-0.00012-2.

[51] A. Das *et al.*, "Predicting the Macronutrient Composition of Mixed Meals From Dietary Biomarkers in Blood," *IEEE J Biomed Health Inform*, vol. 26, no. 6, pp. 2726–2736, Jun. 2022, doi: 10.1109/JBHI.2021.3134193.

[52] D. Panaretos *et al.*, "A comparison of statistical and machine-learning techniques in evaluating the association between dietary patterns and 10-year cardiometabolic risk (2002–2012): the ATTICA study," *British Journal of Nutrition*, vol. 120, no. 3, pp. 326–334, Aug. 2018, doi: 10.1017/S0007114518001150.

[53] Ministerio de la Protección Social., *Encuesta Nacional de la Situación Nutricional en Colombia 2005*, ICBF. 2005.

[54] Instituto Colombiano de Bienestar Familiar, *Encuesta Nacional de la Situación Nutricional en Colombia 2010*, ICBF. 2011.

[55] Ministerio de Salud y Protección Social, Instituto Colombiano de Bienestar Familiar, and Departamento Administrativo para la Prosperidad Social, *Encuesta Nacional de la Situación Nutricional en Colombia 2015*, ICBF. 2019.

[56] L. A. Ortega, R. Cabañas, and A. R. Masegosa, “Diversity and Generalization in Neural Network Ensembles,” *Proc Mach Learn Res*, vol. 151, pp. 11720–11743, Oct. 2021, Accessed: Mar. 04, 2025. [Online]. Available: https://arxiv.org/abs/2110.13786v2

[57] G. Pestoni *et al.*, “Association between dietary patterns and prediabetes, undetected diabetes or clinically diagnosed diabetes: results from the KORA FF4 study,” *Eur J Nutr*, vol. 60, no. 3, pp. 2331–2341, 2021, doi: 10.1007/s00394-020-02416-9.

[58] D. de Assumpção, A. M. P. Ruiz, F. S. A. Borim, A. L. Neri, D. C. Malta, and P. M. S. B. Francisco, “Eating Behavior of Older Adults with and Without Diabetes: The Vigitel Survey, Brazil, 2016,” *Arq Bras Cardiol*, vol. 118, no. 2, pp. 388–397, Mar. 2022, doi: 10.36660/ABC.20201204.

[59] F. Jannasch, J. Kröger, and M. B. Schulze, “Dietary Patterns and Type 2 Diabetes: A Systematic Literature Review and Meta-Analysis of Prospective Studies,” *J Nutr*, vol. 147, no. 6, pp. 1174–1182, Jun. 2017, doi: 10.3945/JN.116.242552.

[60] A. Ahmed, A. Lager, P. Fredlund, and L. S. fer Elinder, “Consumption of fruit and vegetables and the risk of type 2 diabetes: a 4-year longitudinal study among Swedish adults,” *J Nutr Sci*, vol. 9, 2020, doi: 10.1017/JNS.2020.7.