Generación de sustitutos alimentarios mediante inteligencia artificial: un enfoque combinado de modelado supervisado y algoritmos genéticos

Ing. Daniel Hernández Mota

Instituto Tecnológico y de Estudios Superiores de Occidente, ITESO Guadalajara, Jalisco

Abstract. This research explores the use of Artificial Intelligence (AI) techniques, specifically Machine Learning (ML) and Genetic Algorithms, to drive practical innovation in food engineering. The innovative approach integrates diverse data sources on relevant food characteristics, such as flavor molecules and profiles with their respective functional groups, as well as nutritional values. A Random Forest binary classification model was developed that compares food products in pairs, learning to identify similarities with promising performance (AUC PR of 0.898). This model was integrated into an iterative genetic algorithm that proposed optimized lists of ingredients to replicate target products such as cheese, milk, and butter. The generated candidates achieved similarity scores between 0.5 and 0.7, indicating an 80% probability of preserving nutritional and sensory properties comparable to the original products, but with a completely different ingredient composition. This methodology demonstrates the potential of AI to innovate in food product design and personalization, contributing to diversification, sustainability, and accessibility in the food industry.

Keywords: Artificial Intelligence, Machine Learning, Genetic Algorithms, food engineering

Resumen. Esta investigación explora el uso de técnicas de Inteligencia Artificial (IA), específicamente Aprendizaje Automático (ML) y Algoritmos Genéticos, para impulsar la innovación práctica en la ingeniería de alimentos. El enfoque innovador integra diversas fuentes de datos sobre características relevantes de los alimentos, como moléculas y perfiles de sabor con sus respectivos grupos funcionales de sabor, así como valores nutricionales. Se desarrolló un modelo de clasificación binaria de Bosque Aleatorio que compara productos alimenticios por pares, aprendiendo a identificar similitudes con un rendimiento prometedor (AUC PR de 0.898). Este modelo se integró en un algoritmo genético iterativo que propuso listas de ingredientes optimizadas para replicar productos objetivo como queso, leche y mantequilla. Los candidatos generados lograron puntajes de similitud entre 0.5 y 0.7, indicando una probabilidad del 80% de conservar propiedades nutricionales y sensoriales comparables a los productos originales, pero con una composición de ingredientes completamente diferente. Esta metodología demuestra el potencial de la IA para innovar en el diseño y personalización de productos alimenticios, contribuyendo a la diversificación, sostenibilidad y accesibilidad en la industria alimentaria.

Palabras Clave: Inteligencia artificial, aprendizaje automático, algoritmos genéticos, ingeniería de alimentos.

1. Introducción

La inteligencia artificial (IA) se está integrando cada vez más en diversas esferas de la vida cotidiana, impactando significativamente en múltiples sectores gracias a su capacidad para procesar y aprender de grandes volúmenes de datos. En particular, el campo de la ingeniería de alimentos ha comenzado a explorar las capacidades del Machine Learning (ML) para innovar en el diseño y la optimización de productos alimenticios. Tradicionalmente, este diseño se ha basado en la experiencia de los expertos; sin embargo, la incorporación de técnicas avanzadas de IA promete superar las limitaciones de los enfoques convencionales mediante el uso de algoritmos que descubren patrones complejos y ofrecen soluciones menos sesgadas y más eficientes [1][2][3].

El uso de ML en la ingeniería de alimentos no es completamente nuevo. Estudios previos han aplicado estos métodos para el desarrollo de productos como alimentos estructurados [4], o incluso productos individuales como la mayonesa [5], y galletas [6], demostrando de esta manera la versatilidad y eficacia de la IA para mejorar las propiedades nutricionales y sensoriales de los alimentos, así como para reducir el desperdicio alimentario. Adicionalmente, los algoritmos genéticos han ganado popularidad dentro del ámbito de la creatividad computacional, usándose para generar innovaciones como nuevas recetas de comida y sopas [7][8]. Estos estudios destacan la capacidad de los modelos de ML para ajustarse y predecir características deseables en alimentos basados en datos tanto objetivos como subjetivos proporcionados durante el entrenamiento [9][10].

En el aprendizaje automático, se distinguen dos enfoques principales: supervisado y no-supervisado. Los modelos supervisados, que son los más utilizados, buscan predecir un valor deseado ajustando un modelo sobre un conjunto de datos con variables de respuesta conocidas [9]. Estos modelos son a menudo descritos como "cajas negras" debido a su capacidad para adaptarse a diversos contextos y tipos de datos, desde tabulares hasta sensoriales, dependiendo de la cantidad de datos y la experiencia del usuario con los modelos [10]. Entre los algoritmos más comunes se encuentran la regresión lineal, la regresión logística, y técnicas más complejas como las redes neuronales y algoritmos de ensamble basados en árboles, como bosques aleatorios y potenciación del gradiente. En contra parte, los algoritmos no-supervisados se pueden emplear para realizar agrupaciones de los datos o reducciones en la dimension de los mismos. Estos algoritmos no-supevisados se pueden emplear en conjunto con los supervisados para mejorar el poder predictivo de los algoritmos al aumentar la calidad de información que contienen las variables.

Este proyecto investiga si es posible mejorar el diseño y la personalización de productos alimenticios utilizando un enfoque combinado de aprendizaje supervisado y algoritmos genéticos para simular un conjunto de recetas con distintos ingredientes y de esta manera llegar a generar un producto similar, tanto nutricional como de sabor, pero completamente diferente en composición. Hipotetizamos que la implementación de un sistema de IA que integra técnicas avanzadas de aprendizaje supervisado con algoritmos genéticos puede ofrecer propuestas innovadoras que no solo compitan, sino que potencialmente sustituyan a productos alimenticios convencionales, mejorando así la accesibilidad y calidad nutricional de los alimentos. Este enfoque pretende no solo satisfacer las expectativas nutricionales y sensoriales de los consumidores, sino también contribuir significativamente a la reducción del desperdicio de alimentos, alineándose con los objetivos de sostenibilidad global.

2. Métodología

Las principales herramientas y software utilizados en este estudio incluyen:

- **Base de Datos de FlavorDB y USDA Branded Foods, Edamam API:** Utilizados para obtener datos de ingredientes e información nutricional [11] [12] [13].
- **Python3 :** Lenguaje de programación principal [14].
- **Librerias pandas (2.0.3), numpy (1.24.4), nltk (3.8.1), matplotlib (1.3.1), seaborn (), Scikit-Learn (3.7.3), Shap (0.43.0):** Librerías desarrolladas en python, utilizadas para realizar el análisis de datos, la manipulación de la información numérica, así como la limpieza de la información y la visualización de los resultados [15] [16] [17] [18] [19]. También utilizada para implementar modelos de aprendizaje automático, tanto supervisados (Bosque Aleatorio) como no supervisados (Analisis de componentes Principales) y finalmente la explicabilidad del modelo [20] [21].
- **Algoritmos Genéticos:** Se desarrolló un algoritmo genético para optimizar combinaciones de ingredientes en la generación de recetas.

2.1 Procedimiento

El procedimiento de investigación se llevó a cabo en varios pasos secuenciales:

Adquisición de Datos: Se recopilaron datos de FlavorDB, de la base de datos USDA Branded Foods y de la API Edamam. FlavorDB proporcionó información respecto a nombres de moléculas de sabor, perfiles de sabor detallados y grupos funcionales de las moléculas asociadas con cada ingrediente alimenticio. La USDA proporcionó información respecto a los ingredientes de cada producto y también las categorías a las que pertenecen dichos productos. Y la API de Edamam proporcionó información nutricional, estandarizada a medidas de 100g para cada entidad.

Selección y Limpieza de Datos: Se excluyeron manualmente las entidades de FlavorDB consideradas demasiado generales (por ejemplo, "jugo de frutas", "otros quesos", etc.) o que fueran específica a productos ya generados (por ejemplo, "tacos", "hotcakes", etc.) para centrarse solamente en los ingredientes fundamentales, obteniendo 677 distintos entidades de los 936 originles. Respecto a la USDA, la limpieza de texto implicó transformar todas las descripciones de los ingredientes, utilizando técnicas convencionales de procesamiento de lenguaje natural: cambiar el texto a minúsculas, eliminar caracteres especiales y aplicar técnicas de derivación (stemming) a cada una de las palabras, esto mismo se realizó para cada ingrediente de FlavorDB.

Integración de Datos: Los datos limpios de la USDA se cruzaron con los de FlavorDB para obtener una selección tanto de registros de la USDA que contenieran información que estuviera de igual manera en FlavorDB como para ayudar a generar una lista simplificada de ingredientes.

Ingeniería de Características: De manera individual, se sometieron tanto los perfiles de sabor, grupos funcionales y datos de moléculas a un Análisis de Componentes Principales (PCA) para obtener una reducción de dimensionalidad, creando un conjunto condensado de características para cada entidad. El perfil de sabor se redujo de 592 elementos a 50 componentes, los grupos funcionaes de 84 elementos se redujeron a 20 componentes, y las distintas moléculas se redujeron de 1702 a 100 componentes. Las propiedades nutricionales se mantuvieron como los 34 elementos que originalmente fueron proporcionados por la API de Edamam. Esto generó que cada entidad (e_i) tuviera en total 204 descriptores.

$$e_i = [c_1, c_2, c_3, c_4, \ldots, c_{204}]$$

Donde $c_{[1\text{-}100]} \in$ *Moléculas de sabor,* $c_{[101\text{-}150]} \in$ *Perfil de sabor,* $c_{[151\text{-}170]} \in$ *Grupos funcionales de sabor,* $c_{[171\text{-}204]} \in$ *Propiedades nutricionales.*

Representación de producto: Un producto (p_i) se define de manera simplificada como la lista de ingredientes, es decir un vector de entidades.

$$p_i = [e_1, e_2, e_3, e_4, \ldots, e_n]$$

Entonces también se puede interpretar a un producto como una matriz donde cada elemento fila representa la información de las características de cada entidad, en otras palabras, un producto consiste en elementos que cuentan con características multidimensionales, es decir que cada caracteristica tiene su propia dimensionalidad interna, esto se aborda de manera similar a [22].

$$p_i = [e_1, e_2, \ldots, e_n] = [[c_1, c_2, c_3, c_4, \ldots, c_{204}]_1, [c_1, c_2, c_3, c_4, \ldots, c_{204}]_2, \ldots, [c_1, c_2, c_3, c_4, \ldots, c_{204}]_n]$$

Por simplicidad se redefine un producto obteniendo el valor promedio de cada característica (o columna):

$$p'_i = [c'_1, c'_2, c'_3, c'_4, \ldots, c'_{204}] \text{ donde } c'_k = \frac{1}{n}\sum_{j=1}^{n} c_{kj}$$

Variable de respuesta: Se desarrolló un conjunto de etiquetas (y_i) con granularidad producto. Para esto se consideró de manera manual un refinamiento de las categorías proporcionadas de la USDA. Estas etiquetas en

total fueron las siguientes: *bread*, *butter*, *cheese*, *egg*, *fruit*, *honey*, *meat*, *milk*, *oil*, *seafood*, *vegetable*, y *yogurt*. Para ilustrar el refinamiento, la etiqueta y_{milk} estaba compuesta por las siguientes categorías de la USDA: "*Milk*", "*Plant Based Milk*" y "*Milk/Milk Substitutes*". Sin embargo, una vez teniendo estas etiquetas, en lugar de abordar el problema como un problema supervisado de multiclasificación, se decidió transformar la notación a un problema de clasificaciòn binaria. Al hacer esta transformación, ya no se requieren las etiquetas (y_i) propuestas sino que se realiza el análisis utilizando pares de datos (p'_i, p'_j) con insumo del modelo y la similitud por pares S_{ij} para la variable de respuesta, donde se define que ambos dos elementos pertenecen a la misma etiqueta, entonces son similares ($S_{ij}=1$). Y si pertenecen de manera individual a distintas etiquetas entonces no son similares ($S_{ij} = 0$) [23]. Es decir

$$\forall\ p'_i \in y_a \wedge \forall\ p'_j \in y_b \text{ donde } a = b, \text{ entonces } S_{ij} = 1$$
$$\forall\ p'_i \in y_a \wedge \forall\ p'_j \in y_b \text{ donde } a \neq b, \text{ entonces } S_{ij} = 0$$

2.2 Análisis de Datos

Entrenamiento del Modelo: Se entrenó un modelo de Bosque Aleatorio[24], de la librería de Scikit Learn manteniendo los parámetros iniciales, utilizando un sistema de clasificación binaria basado en la similitud de productos alimenticios dentro de categorías específicas. El conjunto de datos de 5,851 productos se expandió a más de 34 millones al realizar las comparaciones para determinar su similitud. emparejando cada producto con otro. De este conjunto, debido a capacidades de memoria y procesamiento, se obtuvo una muestra aleatoria del 10% de los registros reduciendo el volumen de información a solo 3.4 millones de registros.

Validación del Modelo: El conjunto de datos para entrenar y validar el modelo se dividió en dos subconjuntos: el conjunto de entrenamiento (considerando sólamente el 70% de los productos únicos disponibles) y prueba (contepmlando el 30% de los productos únicos disponibles sobrantes), asegurando que no hubiera superposición de productos entre los conjuntos para prevenir sesgo en la evaluación del modelo. En otras palabras el no existían ningún producto en el entrenamiento que estuviera en el conjunto de prueba y viceversa. Una vez entrenado el modelo, se procedió a evaluar su desempeño con estos productos nunca antes vistos. Además de esto, se pudo verificar el desempeño del modelo con otro subconjunto de datos, utilizando algunos de los valores sobrantes no utilizados para el proceso es decir del conjunto original de los 34 millones, se obtuvieron aquellos donde no había productos que hayan estado en el entrenamiento para mejorar la validación del modelo.

Implementación del Algoritmo Genético: Una vez validado el modelo y cuantificado su desempeño, se utilizó como función de optimización para un algoritmo genético destinado a generar productos de alimentos [25], buscando la similitud de un producto específico a la vez. El algoritmo iteró a través de varias propuestas de listas de ingredientes para proponer una combinación óptimas basadas en las características, tales como el contenido nutricional y el sabor. Primeramente, para cada etiqueta, se define un producto a desarrollar a través de la lista de sus entidades correspondientes.

Después se inicializa una población contemplando distintas listas de ingredientes (llamados también candidatos) generadas de manera aleatoria y se hace la evaluación del modelo para seleccionar los mejores candidatos los cuales pasarían sus genes a la siguiente iteración. Esta selección aleatoria puede estar restringida tal que no se utilicen ciertas entidades no deseadas. Una vez se tiene un conjunto de candidatos selectos, se aplican técnicas de reproducción por pares de manera aleatoria, donde se busca realizar distintas combinaciones de las listas de ingredientes: concatenación simple de la lista de ingredientes, mantener la primer mitad y concatenar la segunda mitad, mantener la segunda mitad y concatenar la primer mitad, etc. Luego se genera un proceso de mutaciones donde de manera aleatoria se agregan o quitan ingredientes. Posteriormente se vuelve a realizar una evaluación con el modelo comparando los productos con el modelo. Este proceso se sigue ejecutando a través de varias generaciones hasta llegar a algun criterio de detención. En este caso, la cantidad específica de 100 de generaciones fue utilizado como criterio de detención debido a restricciones de tiempo. Finalmente, se guarda la información de los candidatos que tengan el valor más cercano a 1 (el cual es un valor

que nos indica una alta similitud de productos), y estos serán contemplados para el desarrollo del producto deseado.

Este enfoque metódico aprovecha tanto el poder predictivo del aprendizaje automático como el potencial creativo de los algoritmos genéticos para proponer productos alimenticios innovadores que cuenten no solo con propiedades nutricionales similares sino que también con un perfil de sabor similar, cambiando por completo los ingredientes base.

3 Resultados

3.1 Modelo

Para evaluar el desempeño del modelo, se utilizaron dos conjuntos de datos distintos. El primero es el conjunto de prueba, que incluye el 30% de los productos disponibles. El segundo conjunto, denominado conjunto remanente, consiste en una muestra del 90% restante de los datos originales (34 millones de registros) que no se consideraron para el entrenamiento principal ni para la evaluación. Para este conjunto remanente también se contempló que no hubiera superposición con el conjunto de entrenamiento, resultando en un tamaño similar al del conjunto de prueba.

Las métricas seleccionadas para medir el rendimiento del modelo no requirieron la definición de un umbral específico. Se utilizó el área bajo la curva ROC (AUC ROC), el AUC de la Precisión-Sensibilidad (AUC PR), y la Precisión Promedio como indicadores de desempeño. Los resultados obtenidos demuestran un rendimiento prometedor del modelo en la identificación de similitudes entre productos, y estos se detallan en la **Tabla 1**.

Tabla 1. Resultados del modelo en distintos conjuntos de evaluación: entrenamiento, prueba y remanente. Los resultados muestran que el desempeño del modelo en diversas métricas es notablemente alto. En particular, un valor superior a 0.8 en el AUC de la Precisión-Sensibilidad (AUCPR) es un indicador de que el modelo está identificando las clases de manera efectiva y precisa.

Métrica	**Conjunto de entrenamiento**	**Conjunto de prueba**	**Conjunto *remanente***
Cantidad de registros	1674846	307924	308354
AUC ROC	0.999828	0.975239	0.975707
AUC PR	0.998695	0.898611	0.898407
Precisión Promedio	0.998694	0.895873	0.895532

La distribución del score del modelo muestra una marcada localización en los extremos, dependiendo del valor de la variable de respuesta. Cuando el valor de esta variable es 0 ($S_{ij}=0$), lo cual indica que dos productos no son similares, se observa que el score del modelo se concentra en rangos bajos. Por el contrario, un valor de 1 en la variable de respuesta ($S_{ij}=1$) señala que dos productos son similares, y en estos casos, el score del modelo tiende a ser alto, evidenciando una clara separación en las distribuciones. Esta tendencia se ilustra claramente en las **figuras 1 y 2.**

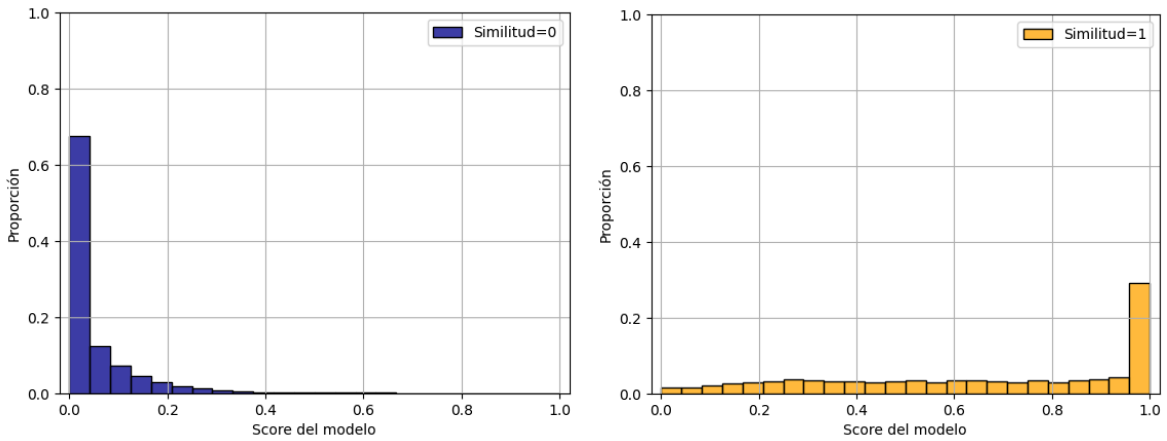

Figura 1. Histograma que muestra la distribución proporcional del score del modelo para cada etiqueta en el conjunto de prueba; en este caso, la suma de todas las barras en cada gráfico debe resultar en 1. En el lado izquierdo, se presenta la distribución para la clase donde la similitud entre productos es 0 ($S_{ij}=0$). Aquí se observa una concentración predominante de scores bajos, prácticamente el 70% de valores de no-similitud se encuentran en el intervalo donde el score es lo mas bajo; lo que indica que el modelo identifica de manera efectiva los casos en que dos productos no son similares. En el lado derecho, se muestra la distribución para los casos donde la similitud entre productos es 1 ($S_{ij}=1$). Esta gráfica refleja una concentración en el extremo de valores altos, el 30% de valores de similitud se encuentran en el intervalo donde el score del modelo es el mas alto; sin embargo, también se aprecia que una menor proporción de casos se distribuye en scores más bajos de manera uniforme. En otras palabras, el 70% de casos de similitud están distribuidos similarmente a lo largo de distintos scores del modelo. Esto indica que existen varios casos en las que al modelo le resulta más difícil identificar similitudes entre productos, asignándoles scores relativamente bajos, lo que conduce a un aumento en la incidencia de falsos negativos.

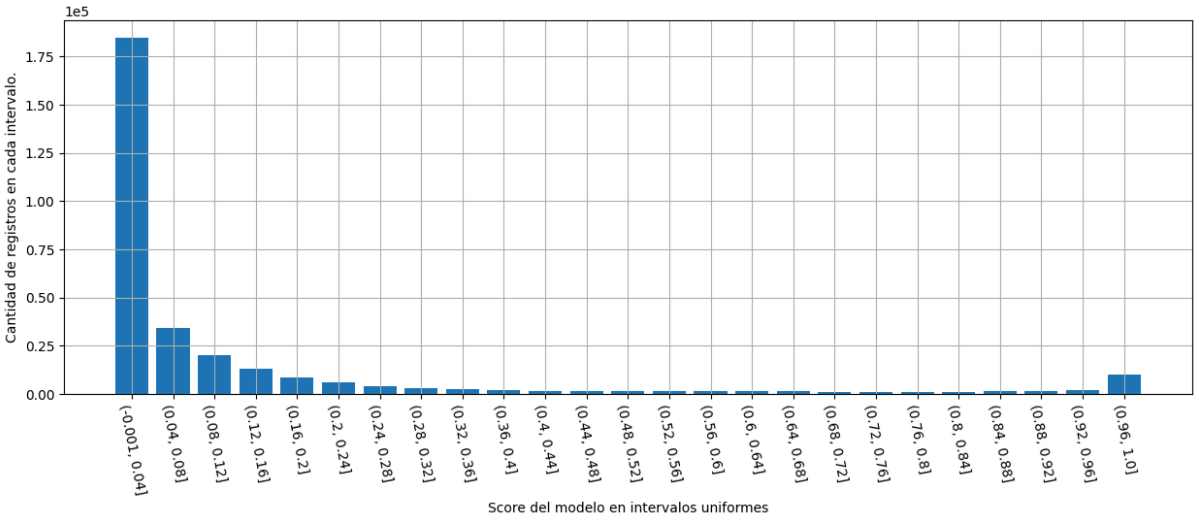

Figura 2. Histograma que muestra la distribución del score del modelo para todos los registros, utilizando intervalos uniformes. Se observa que la mayoría de los registros se localizan en el lado izquierdo del histograma, mientras que solo unos pocos alcanzan los valores más altos en el lado derecho. A partir de estos datos, se puede calcular la proporción de valores que caen en cada intervalo para evaluar el desempeño del modelo según el rango de score **(véase Figura 3).**

Al analizar las **Figuras 1 y 2,** se deduce que la variable de respuesta está notablemente desbalanceada. La proporción de valores positivos ($S_{ij}=1$) drespecto al total es solo del 11.5%, lo que es relevante al evaluar los resultados del modelo, ya que este desequilibrio puede afectar la precisión del modelo. Para determinar una

métrica de precisión por intervalo, se analiza la proporción de similitud calculando el promedio de la variable de respuesta en cada intervalo de score, como se muestra en la **Figura 3.**

Cada intervalo presenta un coeficiente de similitud distinto, indicando que en los scores bajos la proporción de similitud es baja. Por ejemplo, en el conjunto de intervalos que van desde [0, 0.24] se puede apreciar que la proporción de similitud es menor al 20%, indicando que a scores bajos dos productos no seran similares. Sin embargo, esta proporción aumenta conforme el score se incrementa. Esto sugiere que para detectar similitudes con alta probabilidad, los scores más altos son mejores indicativos, ya que en estos intervalos, la proporción de valores similares es significativamente más alta. Por ejemplo, en el intervalo de (0.96 a 1], aproximadamente el 99% de los valores se consideran similares. Por ende, si la comparación de dos productos genera un score del modelo de 0.97, tenemos una alta certeza que van a ser productos similares.

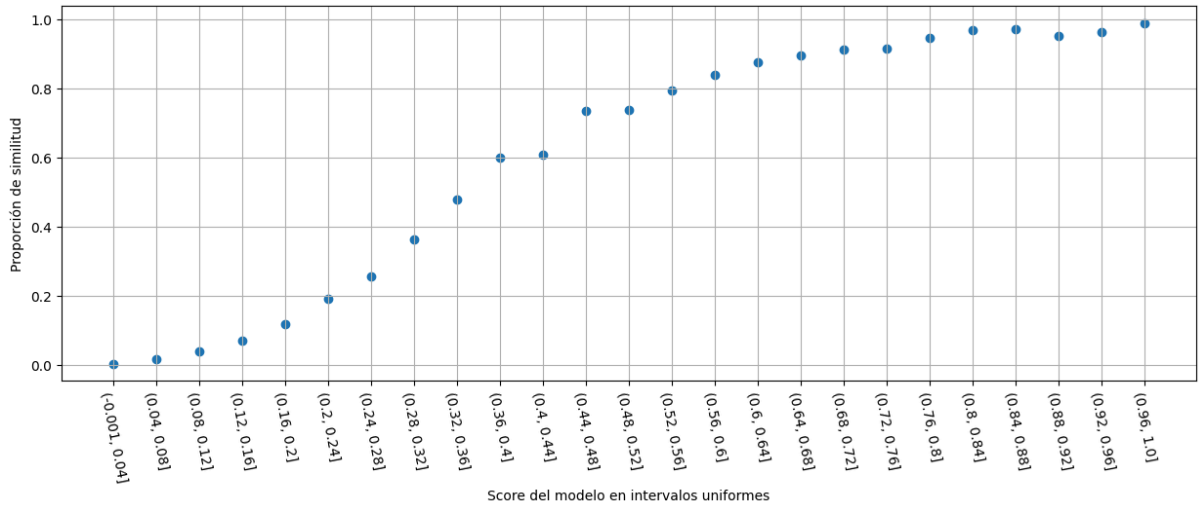

Figura 3. Proporción de similitud (valor promedio de la variable de respuesta) en cada intervalo de score utilizando el conjunto de prueba. Se observa que a medida que el score aumenta, también lo hace el coeficiente de similitud. Por ejemplo, en los intervalos con scores del modelo que van de 0.72 a 1, la proporción de similitud se aproxima al 90%. Esto indica que, dentro de este rango, es mucho más probable encontrar una alta similitud entre dos productos; es decir, si la comparación de dos productos resulta en un score dentro de estos valores, es altamente probable que sean similares, y esta probabilidad aumenta a medida que el score es más alto. Este patrón se puede utilizar como un mecanismo de calibración para determinar la probabilidad de que un candidato sea similar dado un valor de score específico.

3.2 Algoritmo Genético

El algoritmo genético se empleó para generar nuevos candidatos de productos utilizando el resultado del modelo de clasificación binaria como elemento fundamental para optimizar la similitud. Se implementaron tres escenarios específicos, y en cada uno, el algoritmo iteró a lo largo de 100 generaciones, obteniendo así un candidato final que tuviera el score del modelo más alto.

Creación de producto similar al Queso:
En este escenario, el objetivo era generar un producto que representase el producto generado por la siguiente lista:

$$p_i = \text{['Cheese']}$$

Para evitar la convergencia del algoritmo hacia productos conocidos y prevenir el estancamiento, se excluyeron las siguientes entidades del proceso de generación: ['Blue Cheese', 'Camembert Cheese', 'Cheddar

Cheese', 'Cheese', 'Comte Cheese', 'Cottage Cheese', 'Cream Cheese', 'Emmental Cheese', 'Feta Cheese', 'Goat Cheese', 'Gruyere Cheese', 'Limburger Cheese', 'Mozzarella Cheese', 'Munster Cheese', 'Parmesan Cheese', 'Provolone Cheese', 'Ricotta Cheese', 'Romano Cheese', 'Roquefort Cheese', 'Sheep Cheese', 'Swiss Cheese', 'Tilsit Cheese']. Tras 100 generaciones, el algoritmo generó la el siguiente producto como el candidato más cercano:

p_j = ['Guinea hen', 'Pacific rockfish', 'Mutton', 'Cream', 'Roe', 'Swordfish', 'Cardamom', 'Milk Powder', 'Sage', 'Sapodilla', 'Margarine like spread', 'Coriander', 'Northern bluefin tuna', 'Rum', 'Raisin', 'Smoked Fish', 'Buttermilk', 'Cognac Brandy', 'Tomato', 'Tamarind', 'Lamb', 'Bonito'].

Comparando ambos productos con el modelo, el score obtenido fue de 0.58, lo que, según los resultados mostrados en la **Figura 3**, equivale a una similitud ligeramente superior al 80%.

Posteriormente, se llevó a cabo un análisis de las variables más influyentes utilizando SHAP (SHapley Additive exPlanations), para cuantificar la contribución de cada variable en función de la categoría de datos a la que pertenecían. En este análisis, se destacó que los factores nutricionales y las moléculas de sabor fueron los elementos que más influyeron en las decisiones del modelo, como se muestra en la **Figura 4.**

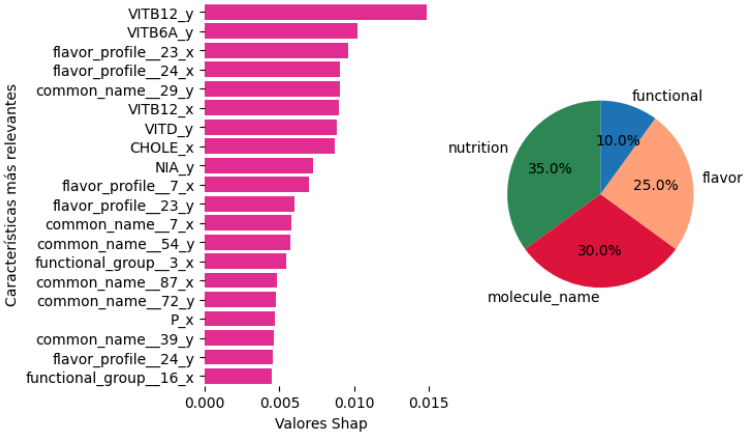

Figura 4: Las 20 variables más importantes a través de SHAP y porcentaje de representación del grupo al cual pertenecen de los resultados obteindos del algoritmo genético al realizar una propuesta de queso. En general el sabor se compone de las tres categorías que excluyen nutrición, por lo que en este caso el sabor se tomó como 65% de indicador, mientras que la nutrición solo un 35%. Pero de ese 65%, lo más relevante era la presencia de las moléculas de sabor.

Creación de producto similar a la Leche:
En este caso específico, el objetivo era generar un producto representado por la lista ['Milk'].

p_i = ['Milk']

Se restringió al algoritmo el acceso a las siguientes entidades: ['Buttermilk', 'Condensed Milk', 'Evaporated Milk', 'Goat Milk', 'Milk', 'Milk Fat', 'Milk Human', 'Milk Powder', 'Milkfish', 'Milkshake', 'Sheep Milk', 'Skimmed Milk', 'Soy Milk']. Tras 100 generaciones, el algoritmo produjo una el siguiente producto:

p_j = ['Salt', 'Rose hip', 'Cocoa']

Cuantifiando la similtiud, se alcanzó un score de 0.63, lo que indica una similitud ligeramente superior al 80%. Es probable que el ingrediente "Cocoa" aparezca recurrentemente en el entrenamiento debido a que probablemente hay mucha leche donde un ingrediente adicional es cocoa, lo que sugiere una alta frecuencia de ocurrencia, para siguientes iteraciones quizá se limite este valor adicional y detectar qué tanto cambian los

resultados. En este caso, las moléculas de sabor fueron el factor más influyente en la decisión del modelo, seguido de cerca por la nutrición y el perfil de sabor, como se puede ver en la **Figura 5**.

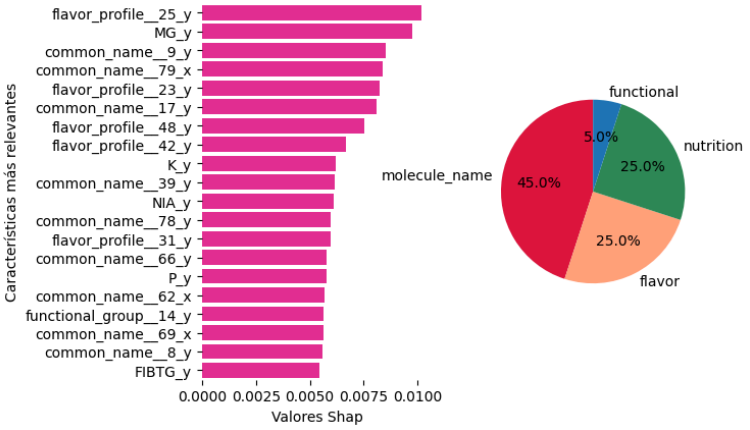

Figura 5: Las 20 variables más importantes a través de SHAP y porcentaje de representación del grupo al cual pertenecen de los resultados obteindos del algoritmo genético al realizar una propuesta de leche.

Creación de producto similar a la mantequilla:
En este escenario específico, se definió que el producto a generar estuviera representado por ['Butter'].

$$p_i = \text{['Butter']}$$

Se restringió al algoritmo el acceso a las siguientes entidades para evitar convergencias no deseadas: ['American Butterfish', 'Butter', 'Buttermilk', 'Butternut', 'Butternut Squash', 'Canola Oil', 'Citrus Peel Oil', 'Cocoa Butter', 'Cooking Oil', 'Corn Oil', 'Fish Oil', 'Giant Butterbur', 'Grapefruit Peel Oil', 'Lemon Peel Oil', 'Lime Peel Oil', 'Mustard Oil', 'Oil Palm', 'Orange Oil', 'Peanut Butter', 'Peanut Oil', 'Soybean Oil']. Tras 100 generaciones, el algoritmo seleccionó el siguiente producto alcanzando un score de 0.52:

$$p_j = \text{['Cognac Brandy', 'Drumstick Leaf', 'Oregano', 'Ghee', 'Filbert', 'Bearded Seal', 'Corn Chip']}$$

Este valor indica una probabilidad del 80% de que el producto resultante tenga similitud con la mantequilla tradicional. En este caso, las propiedades nutricionales fueron el factor más influyente en la toma de decisiones del modelo, como se muestra en la **Figura 6.**

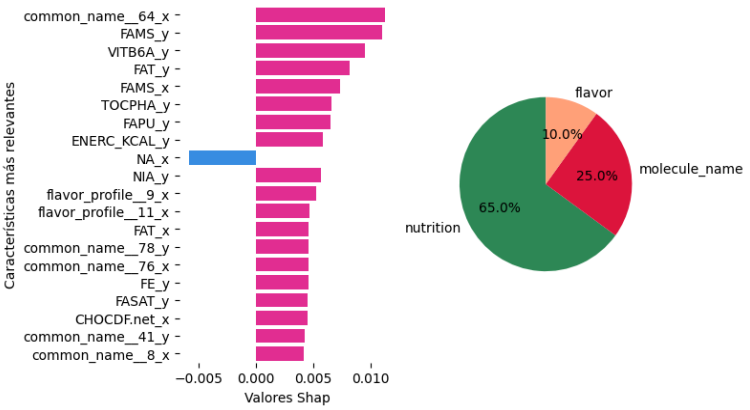

Figura 6: Las 20 variables más importantes a través de SHAP y porcentaje de representación del grupo al cual pertenecen de los resultados obteindos del algoritmo genético al realizar una propuesta de mantequilla.

Discusión

Los resultados obtenidos en este estudio indican que es viable generar sustitutos de productos alimenticios mediante el uso de algoritmos genéticos guiados por un modelo de clasificación supervisada. En los resultados de los sustitutos de productos, los scores obtenidos oscilan entre 0.52 y 0.63 en tan solo 100 generaciones. Esto sugiere que el modelo puede tanto identificar como recomendar ingredientes que potencialmente replican las características de alimentos específicos. Esto responde afirmativamente a nuestra pregunta de investigación sobre si la IA puede mejorar el diseño y personalización de productos alimenticios, al menos en un nivel preliminar.

Los hallazgos tienen amplias implicaciones para la industria alimentaria, particularmente en la innovación de productos. Al poder crear variantes de alimentos existentes que conservan cualidades nutricionales y sensoriales deseables, se podrían diversificar las opciones disponibles para los consumidores y responder mejor a necesidades dietéticas específicas. Además, esta tecnología podría contribuir a la sostenibilidad, permitiendo el desarrollo de productos que maximicen el uso de recursos disponibles o subutilizados.

A pesar de los resultados prometedores, el estudio presenta varias limitaciones. Primero, la necesidad de utilizar más generaciones para obtener resultados más precisos señala una limitación en la capacidad actual del algoritmo para converger rápidamente hacia la solución óptima. Además, la falta de validación práctica de los productos generados con expertos en alimentación y nutrición es un aspecto crítico que podría afectar la aplicabilidad real de los sustitutos desarrollados. Otro aspecto limitante es que el modelo actual no especifica las proporciones de los ingredientes ni sugiere métodos de preparación, lo cual es esencial para la realización práctica de cualquier receta.

Para futuras investigaciones, sería beneficioso extender el modelo para incluir recomendaciones sobre las proporciones de ingredientes y métodos de preparación. Esto haría que los resultados fueran más aplicables en contextos prácticos. Además, incrementar el número de generaciones en los algoritmos genéticos podría mejorar la precisión y relevancia de los productos generados. Sería también esencial implementar estudios que involucren a expertos culinarios y nutricionistas para validar la viabilidad y aceptación de los productos diseñados. Finalmente, explorar la integración de consideraciones sobre aditivos y otros componentes alimentarios en el modelo ampliaría su utilidad.

Conclusión

Este estudio ha desarrollado un modelo de clasificación binaria eficaz que determina la similitud entre dos productos alimenticios, exhibiendo métricas de rendimiento relativamente altas. A través del uso de este modelo, se impulsó un algoritmo genético que identifica iterativamente candidatos potenciales, representados por listas de entidades, seleccionando las mejores propuestas basadas en su similitud con el producto deseado.

Las propuestas generadas demostraron tener scores que indican una probabilidad de aproximadamente el 80% de similitud con el producto objetivo. Además, se ha logrado una interpretación detallada de las caracteristicas más influyentes en cada resultado, proporcionando una cuantificación porcentual de los factores que más contribuyen a la toma de decisiones.

Los hallazgos de esta investigación son significativos para el campo de la ingeniería de alimentos y la tecnología de alimentos, ofreciendo un nuevo enfoque para la creación de productos alimenticios. Este enfoque no solo permite la innovación en términos de desarrollo de productos que puedan satisfacer necesidades

específicas de los consumidores, sino que también contribuye a la optimización de recursos y a la reducción del desperdicio alimentario.

Aunque este es solo el comienzo, la aplicación de inteligencia artificial en este contexto abre múltiples oportunidades para mejorar la calidad, accesibilidad y personalización de los alimentos. La investigación sugiere un camino prometedor hacia la integración de técnicas más sofisticadas, como la inclusión de aditivos que mejoren características como la consistencia, el sabor y la estabilidad de los productos alimenticios. Así, este estudio no solo enriquece la comprensión académica y aplicada de la ingeniería de alimentos, sino que también establece una base sólida para futuras investigaciones y desarrollos en el sector.

Referencias

[1] Arenas, A., Macias, B., Gómez, A. Miramontes, A., Michel, L., Trapero, R., Vela, A., Ramirez, P., Pérez, I., Barrera, M., Ramírez, H., y Valdés, J. (2023) Diseño y Desarrollo de Alimentos con Inteligencia Artificial. Instituto Tecnológico y de Estudios Superiores de Occidente.

[2] Negro, A. (2021) Graph-Powered Machine Learning. Manning Publicactions Co.

[3] Nozaki, N.; Konno, E.; Sato, M.; Sakairi, M.; Shibuya, T.; Kanazawa, Y.; y Georgescu, S. (2017) Application of Artificial Intelligence Technology In product Design. Fujitsu Scientific & Technical Journal 53(4): pp 43-51

[4] Meeuse, F. M., Chapter 6 - Process Synthesis for structured food products. In Computer Aided Chemical Engineering; Ng, K. M., Gani, R., Dam-Johansen, K., Eds.; Elsevier, 2007; Vol. 23, pp 167−179.

[5] Dubbelboer, A., Janssen, J., Krijgsman, A., Zondervan, E., y Meuldijk, J. (2015) Integrated Product and Process Design for the Optimization of Mayonnaise Creaminess,Computer Aided Chemical Engineering, Elsevier, Vol. 37, pp 1133-1138, https://doi.org/10.1016/B978-0-444-63577-8.50034-6

[6] Zhang, X., Zhou, T., Zhang, L., Yip, K. y Ming, K. (2019) Food Product Design: A Hybrid Machine Learning and Mechanistic Modeling Approach. Indistrial and Engineering Chemistry Research 58 (36), 16743-16752 DOI: 10.1021/acs.iecr.9b02462

[7] Varshney, L. R.; Wang, J.; and Varshney, K. R. 2016. Associative algorithms for computational creativity. The Journal of Creative Behavior 50(3):211–223.

[8] Morris, R. G.; Burton, S. H.; Bodily, P. M.; and Ventura, D. 2012. Soup over bean of pure joy: Culinary ruminations of an artificial chef. In International Conference on Computational Creativity, 119–125.

[9] Müller, A. C., Guido, S., & Müller, A. C. (2016). Introduction to machine learning with Python: A guide for data scientists. O'Reilly Media.

[10] Shwartz-Ziv, R., & Armon, A. (2021) Tabular Data: Deep Learning is Not All You Need. Cornell University Recuperado el 15 de abril del 2023 de: https://doi.org/10.48550/arXiv.2106.03253

[11] Garg, N., Sethupathy, A., Tuwani, R., Rakhi, NK, Dokania, S., Iyer, A., Gupta, A., Agrawal, S., Singh, N., Shukla, S., Kathuria, K., Badhwar, R., Kanji, R., Jain, A., Kaur A., Nagpal, R., y Bagler, G. (2017) FlavorDB: A database of flavor molecules, Nucleic Acids Research, Nucleic Acids Research, Volume 46, Issue D1, 4 January 2018, Pages D1210–D1216, https://doi.org/10.1093/nar/gkx957

[12] U.S. Department of Agriculture, Agricultural Research Service, Beltsville Human Nutrition Research Center. FoodData Central. [Internet] [cited (18/04/2024)]. Available from https://fdc.nal.usda.gov/.

[13] EDAMAM (2024). Food database API. [Internet] [cited (18/04/2024)]. Available from: https://developer.edamam.com/food-database-api

[14] Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.

[15] The Pandas development team, pandas-dev/pandas: Pandas, 2020, https://doi.org/10.5281/zenodo.3509134.

[16] Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357–362 (2020). DOI: 10.1038/s41586-020-2649-2.

[17] Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc."

[18] J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.

[19] Waskom, M. L., (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021, https://doi.org/10.21105/joss.03021.

[20] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel,O., Blondel M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., y Duchesnay E., (2011) Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research Vol.12, pp. 2825−2830.

[21] Lundberg, S. & Lee, S.-I (2017). A unified approach to interpreting model predictions. In Adv. Neural Information Processing pp. 4765–4774. Curran Associates.

[22] Olszewski, D. (2014). Fraud detection using self-organizing map visualizing the user profiles. Knowledge-Based Systems. Vol 70, pp 324-334, ISSN: 0950-7051, DOI: https://doi.org/10.1016/j.knosys.2014.07.008

[23] Hsu, Y., Lv, Z., Schlosser, J. , Odom, P., y Kira Z. (2019) Multiclass classification without multiclass labels. Georgia Institute of Technology, Georgia Tech Research Institute. International Conference on Learning Representations.

[24] Breiman, “Random Forests”, Machine Learning, 45(1), 5-32, 2001.

[25] A. Amorim, L. Fabricio, W. Goes, A. Ribeiro, D. Silva, y C. Franc¸A, (2017) “Creative Flavor Pairing: ´ Using RDC Metric to Generate and Assess Ingredients Combination” in Proceedings of the Eighth International Conference on Computational Creativity, pp. 33–40.