CENTRO UNIVERSITARIO SANTA ANA

ALMENDRALEJO

Del 2 al 5 de mayo de 2023

Joaquín Sorolla Bastida. Comiendo uvas, 1898. Acualera sobre papel. Museo Sorolla, nº inv. 00427

XLV JORNADAS DE VITICULTURA Y ENOLOGÍA TIERRA DE BARROS

V CONGRESO AGROALIMENTARIO DE EXTREMADURA

XLV JORNADAS DE VITICULTURA Y ENOLOGÍA
DE LA TIERRA DE BARROS

V CONGRESO AGROALIMENTARIO DE EXTREMADURA

Edita:
Centro Universitario Santa Ana
C/ IX Marqués de la Encomienda, nº 2
Almendralejo
Tel. 924 661 689
http//www.univsantana.com

Colabora: Cajalmendralejo

Ilustración de portada:
Joaquín Sorolla Bastida. "Comiendo uvas". 1898. Acuarela sobre papel.
Museo Sorolla. n: inv. 00427. © Fundación Museo Sorolla

Diseño original:

Tecnigraf S.A.

Maquetación: María Sabater

ISBN: 84-7930-113-9

D.L.: BA-000169-2024

Imprime: Impresal

Cromatografía de gases ultrarrápida acoplada a técnicas de machine learning para predecir el nivel de adulteración en miel de azahar y girasol

PUNTA-SÁNCHEZ, I.[1]

DYMERSKI, T.[2]

CALLE. J. L.[1]

RUIZ-RODRÍGUEZ, A.[1]

FERREIRO-GONZÁLEZ, M.[1]

PALMA, M.[1]

1 Departamento de Química Analítica, Facultad de Ciencias, Universidad de Cádiz.

2 Department of Analytical Chemistry, Faculty of Chemistry, Gdansk University of Technology, Poland.

RESUMEN

La adulteración de la miel es un problema importante en la industria alimentaria, y la detección de estas prácticas fraudulentas es esencial para garantizar la calidad y autenticidad del producto. La cromatografía de gases ultrarrápida (CG-ultrarrápida) es una técnica analítica rápida y sensible para detectar adulteraciones en miel y, combinada con algoritmos de machine learning, ha demostrado ser una herramienta efectiva para desarrollar modelos precisos y

fiables para detectar la adulteración de forma automatizada y objetiva.

En este estudio, se evaluó la capacidad de diferentes algoritmos de ML en combinación con CG-ultrarrápida para predecir el nivel de adulteración en miel de azahar (OB) y girasol (SF). Las muestras de miel de azahar y girasol se adulteraron utilizando una mezcla de otras mieles de diferentes orígenes florales como adulterante. Se obtuvieron muestras adulteradas con un rango de pureza de miel que osciló entre el 50% y el 95%. Se encontró que la regresión de vectores soporte (SVR) mostró el mejor rendimiento con un R^2 de 0.9086 en el conjunto de prueba para la matriz de datos conjunta de miel de azahar y girasol. Para mejorar la precisión de los modelos de regresión, se propuso clasificar las muestras de miel en función de su origen botánico y luego aplicar los modelos de regresión por separado. Todos los modelos de regresión probados en miel de azahar y de girasol por separado obtuvieron un rendimiento superior. El modelo de operador de selección y contracción mínima absoluta (LASSO) resultó ser el mejor para predecir las propiedades de la miel de azahar y de girasol, con un R^2 de 0.9987.

Palabras clave: Miel, Adulteraciones, Cromatografía de gases ultrarrápida, Machine learning.

SUMMARY

Honey adulteration is a major problem in the food industry, and detection of these fraudulent practices is essential to ensure product quality and authenticity. Ultra-fast gas chromatography (ultra-fast GC) is a fast and sensitive analytical technique for detecting adulteration in honey and, in combination with machine learning algorithms, has proven to be an effective tool for developing accurate and reliable models to detect adulteration.

In this study, several machine learning techniques were compared to predict the level of adulteration in orange blossom (OB) and sunflower (SF) honey using ultrafast GC. The OB and SF honey samples were adulterated using a mixture of other honeys from different floral origins as adulterant. Adulterated samples were obtai-

ned with a range of honey purity that oscillated between 50% and 95%. It was found that the supported vector regression (SVR) showed the best performance with an R^2 of 0.9086 for the data set containing orange blossom and sunflower honey. To improve the accuracy of the regression models, it was proposed to classify the honey samples based on their botanical origin and then apply the regression models separately. All the regression models tested on orange blossom and sunflower honey separately obtained superior performance. The least absolute shrinkage and selection operator (LASSO) turned out to be the best to predict the properties of orange blossom and sunflower honey, with an R^2 of 0.9987.

Keywords: Honey, Adulteration, Ultra-fast gas chromatography, Machine learning.

INTRODUCCIÓN

La miel es un edulcorante natural producido por las abejas *Apis mellifera L.* a partir del néctar de las flores o de las secreciones de las partes vivas de las plantas o de las excreciones de los insectos chupadores de plantas [1–3]. La composición y las propiedades de la miel dependen del origen botánico, del origen geográfico, de las prácticas apícolas y de las condiciones climáticas y ambientales. La miel se puede clasificar en dos categorías según las secreciones de las plantas utilizadas para su síntesis: miel de flores elaborada a partir del néctar de las flores y miel de melaza elaborada a partir de las secreciones de todas las partes vivas de las plantas distintas de las flores o excreciones de insectos [4,5].

De acuerdo con la Directiva del Consejo de la Unión Europea 2001/110/EC (European Commission, 2001) y el Codex Alimentarius (Codex Alimentarius, 1987), la miel se define como un producto dulce natural, y tanto la adicción de cualquier sustancia, como la declaración falsa de los orígenes botánicos o geográficos se consideran prácticas fraudulentas. Sin embargo, la miel se encuentra entre los alimentos más adulterados en el mercado. La adición de edulcorantes baratos a la miel es una de las formas más comunes de adulteración de la miel, cuyo objetivo es aumentar el volumen y la dulzura del producto y reducir los costos de producción [6–8].

Además, esta práctica también puede comprometer la calidad y seguridad del producto. La miel adulterada puede tener una composición, sabor y aroma diferente a la miel pura [9].

Con el objetivo de garantizar la calidad y seguridad de la miel, varias organizaciones nacionales e internacionales han establecido estándares y pautas para la producción, el procesamiento y el etiquetado de la miel. La Norma del Codex para la Miel es una norma reconocida a nivel mundial que establece los criterios mínimos de calidad y pureza para la miel, incluido su contenido de humedad, composición de azúcar y ausencia de aditivos y contaminantes [10,11].

En los últimos años, se han utilizado diversas técnicas analíticas para detectar la adulteración de la miel, incluido el análisis de la relación de isótopos de carbono estables (SCIRA) [12], cromatografía de gases (GC) [13] UV-visible (UV-Vis) [16], infrarrojo cercano (NIR) [17], espectroscopia Raman [18], imágenes termográficas [19], y tecnología de biosensores [20].

La cromatografía de gases ultrarrápida (CG-ultrarrápida) es una técnica analítica que permite el análisis rápido y sensible de compuestos volátiles en mezclas complejas como la miel. El uso de estas técnicas en conjunto con algoritmos de machine learning (ML) permite el desarrollo de modelos precisos y fiables para la detección de la adulteración de la miel. De este modo, los resultados de este estudio pueden ser utilizados por la industria de la miel y las agencias reguladoras para garantizar la autenticidad y la calidad de la miel.

El objetivo de este estudio es evaluar la capacidad de diferentes algoritmos de ML en combinación con CG-ultrarrápida para predecir el nivel de adulteración en miel de azahar y girasol. Se evaluaron los siguientes algoritmos: el operador de selección y contracción mínima absoluta (LASSO), regresión de Ridge (RIDGE), red elástica (ENET), mínimos cuadrados parciales (PLS), Random Forest (RF) y vectores de soporte de regresión (SVR). Este estudio es el primero en examinar y comparar el rendimiento de algunas técnicas ML para el análisis de miel adulterada empleando CG-ultrarrápida.

MATERIALES Y MÉTODOS

Preparación de muestras adulteradas

Las mieles puras de azahar y girasol utilizadas en este estudio fueron suministradas por la Subdirección General de Control e Inspección Agroalimentaria de la Consejería de Agricultura, Ganadería, Pesca y Desarrollo Sostenible.

La miel de azahar (OB) se preparó mezclando 13 mieles de azahar puras diferentes. La miel de girasol (SF) se preparó mezclando 7 mieles de girasol pura diferentes. Las muestras adulteradas se hicieron mezclando cada miel con distinta proporción de adulterante para tener una pureza final entre el 50-95%. En la Tabla I se presentan todas las muestras preparadas para el estudio. El adulterante que se utilizó para las adulteraciones fue una mezcla en igual proporción de miel de eucalipto, de romero y de mil flores. Cada tipo de muestra se preparó por duplicado.

Análisis de las muestras

Los análisis se realizaron en el sistema de GC- ultrarrápida Heracles II con muestreador automático HS100 (Alpha M.O.S., Toulouse, Francia). El sistema estaba equipado con dos columnas paralelas de diferente polaridad (MXT-5 y MXT-1701) (Restek, Bellefonte, PA, EE. UU.) cada una acoplada a un detector de ionización de llama (μFID). El gas portador fue hidrógeno de pureza 6N proporcionado por el generador Precision Hydrogen Trace 250 (Peak Scientific Instruments, Inchinnan, Reino Unido). Las muestras se incubaron a 40 °C durante 20 minutos con una agitación de 500 rpm. La temperatura del inyector se estableció en 200°C. El volumen de muestreo del espacio de cabeza inyectado fue de 2500 μL a 250 μL/s. Los analitos se retuvieron en un material absorbente compuesto por *Tenax*® a 40 °C y luego se desorbieron térmicamente a 240 °C durante 20 s. La temperatura del horno se programó para comenzar a 40 °C durante 2 s y luego aumentar de 40 a 270 °C a una velocidad de 3 °C/s, manteniendo 270 °C durante 18 s. La temperatura de los detectores FID se fijó en 270 °C. El tiempo de adquisición de datos fue de 97 s.

Procesamiento de datos

El análisis de datos se realizó con *RStudio* (R versión 4.0.5, Boston, MA, EE. UU.). El conjunto de datos se obtuvo concatenando la información de ambos detectores FID. Dando como resultado una matriz de datos de bidimensional D_{nxp} de $D_{44}x_{20002}$ donde *p* es el número de variables (tiempos de retención) y *n* es el número de muestras (miel de azahar y girasol adulterada).

Se utilizaron diversos paquetes de RStudio durante el análisis: *Boruta* para la selección de variables, *factoextra* para el análisis de cluster jerárquico, *ggplot2* para las visualizaciones gráficas, *caret* para la división de los datos en conjunto de entrenamiento y de test y la creación de modelos de machine learning, tanto de clasificación como de regresión. Los modelos de regresión generados con estos algoritmos se evaluaron utilizando métricas como el R-cuadrado (R^2) y el error cuadrático medio (RSME). El R^2 representa la bondad de ajuste del modelo, donde un valor más alto indica un mejor ajuste. El RSME mide la diferencia promedio entre los valores reales y predichos, donde un valor más bajo indica un mejor ajuste. La métrica utilizada para evaluar el rendimiento de los modelos de clasificación fue la precisión, que se calcula como el número de instancias correctamente clasificadas dividido por el número total de instancias.

DISCUSIÓN Y RESULTADOS

Análisis exploratorio de datos

En primer lugar, se realizó un Análisis Jerárquico de Conglomerado (HCA) de la matriz de datos completa ($D_{44x20002}$), sin aplicar pretratamiento de datos. El HCA es un tipo de análisis de conglomerados que crea una jerarquía de grupos en función de la similitud de las muestras. Se empleó la distancia euclídea para determinar la similitud entre las muestras. El método de unión, utilizado para fusionar o dividir grupos, se seleccionó mediante comparación de diferentes métodos (único, completo, promedio, de ward y centroide). Dado que el HCA se realizó sin pretratamiento y concatenando la información obtenida por ambas columnas para comprender las tendencias brutas, los valores en el método de vinculación oscilaron entre 0,6216 y 0,7474. El método promedio obtuvo el mejor resultado (0,7474). El HCA se representa gráficamente en el dendrograma de la Figura 1.

En términos generales, el dendrograma (Figura 1) agrupó las muestras en cuatro clústeres principales identificados como A, B, C y D. Las muestras se agruparon en primer lugar en función de su origen botánico, ya que todas las muestras de azahar (OB) (clúster A y B) se encuentran agrupadas a una mayor distancia de las de girasol (SF) (clúster C y D). En segundo lugar, se observa como dentro de cada origen botánico, las muestras tienden a clasificarse en función del nivel de adulteración, estando las muestras sin adulterar agrupadas en un único grupo y separadas del resto de muestras adulteradas tanto en el caso de las muestras OB como las SF. Las muestras de SF adulteradas mostraron una ligera tendencia a agruparse por el nivel de adulteración en subgrupos. Sin embargo, esta tendencia fue menos evidente en el caso de miel de azahar.

Este análisis sugiere que la CG-ultrarrápida permite distinguir de manera efectiva entre diferentes orígenes botánicos, así como también diferenciar entre muestras de miel adulteradas y no adulteradas. Sin embargo, mediante este análisis exploratorio no se ha obtenido una clasificación perfecta en el caso del nivel de adulteración en las muestras de OB.

Modelos supervisados para la predicción del nivel de adulterante en miel

La aplicación de modelos de regresión supervisada, como el operador de selección y contracción mínima absoluta (LASSO), regresión de Ridge (RIDGE), red elástica (ENET), mínimos cuadrados parciales (PLS), regresión de vectores de soporte (SVR) y Random Forest (RF), permite predecir el porcentaje de adulterante en muestras de miel a partir de los datos obtenidos a través de CG-ultrarrápida.

Para desarrollar modelos de regresión robustos y evitar sobreajustes, se realizó un pretratamiento de los datos aplicando el algoritmo de selección de variables *Boruta*, para identificar las variables más relevantes generadas al combinar datos de ambas columnas de CG-ultrarrápida. Este algoritmo utiliza una versión modificada de Random Forest para clasificar la importancia de las características.

La matriz de datos $D_{44x20002}$ que contenía la información de ambos sensores se dividió aleatoriamente considerando únicamente el nivel de adulteración en un conjunto de entrenamiento que contenía el 75 % (n=33) y en un conjunto de test con el 25% restante (n=11). El conjunto de test incluyó muestras

independientes que no se utilizaron en la construcción del modelo, sino que se reservaron para la validación externa con el objetivo de obtener una estimación imparcial del error para todos los modelos entrenados.

Para evaluar los diferentes modelos supervisados, previamente se procesó el conjunto de entrenamiento ($D_{33x20002}$) utilizando el algoritmo *Boruta*, que seleccionó las 30 características más importantes (D_{33x30}) para construir los modelos. La optimización de los parámetros del modelo se evaluó mediante validación cruzada (VC) con 5 pliegues en el conjunto de entrenamiento.

Operador de selección y contracción mínima absoluta (LASSO)

LASSO es un modelo de regresión lineal que realiza tanto la regularización como la selección de variables mediante la aplicación de una penalización a los coeficientes de regresión, lo que reduce algunos coeficientes a cero y establece que las variables correspondientes se excluyan del modelo. El grado de penalización está controlado por el hiperparámetro lambda (λ), que se puede ajustar para equilibrar la compensación entre la complejidad del modelo y la bondad del ajuste.

Se desarrolló un modelo predictivo usando regularización con un valor de λ optimizado de 0.1321. Lambda se optimizó mediante un método de búsqueda de rejilla utilizando secuencias exponenciales de 10^{-5} a 10 cada 100. El modelo logró un buen rendimiento con un RMSE de 4,0412 y un R^2 de 0,9373 en un conjunto de datos independiente. En los conjuntos de entrenamiento y test, el RMSE obtenido fue de 3,7446 y 6,3353 con valores de R^2 de 0,9433 y 0,8739, respectivamente. El modelo seleccionó 2 variables de las 30 variables previamente seleccionadas con *Boruta.*

Regresión de Ridge (RIDGE)

En RIDGE, el valor del hiperparámetro lambda (λ) controla la cantidad de contracción aplicada a los coeficientes. Los valores más grandes de lambda dan como resultado una mayor contracción y coeficientes más pequeños, pero los coeficientes de las variables menos importantes nunca se reducen a 0.

El valor óptimo de lambda obtenido por un método de búsqueda en cuadrícula utilizando secuencias exponenciales de 10^{-5} a 10 cada 100 fue en este caso 10, lo que resultó en un RMSE de 6.3528 y un R^2 de 0.9007.

Los valores RMSE y R^2 para el conjunto de entrenamiento fueron 5,4803 y 0,8774, respectivamente, mientras que, para el conjunto de test, RMSE y R^2 fueron 6,7255 y 0,8363, respectivamente.

Red elástica (ENET)

ENET es un modelo de regresión que combina la regularización LASSO y RIDGE, con dos hiperparámetros para optimizar: lambda (λ), que controla la fuerza general de la penalización, y alfa (α), que determina el equilibrio entre los dos tipos de sanciones. Cuando alfa se acerca a 1, la ENET es similar a la regularización LASSO, mientras que cuando alfa se acerca a 0, se acerca más a la regularización de RIDGE. La ENET puede generar modelos reducidos al generar un coeficiente de valor cero, similar a la regularización de Lasso.

La combinación óptima que se obtuvo empleando el modelo de red elástica fue un valor de λ de 0,0294 y un α de 1, lo que indica que el modelo es similar a LASSO. El RMSE y R^2 para el modelo fue de 4.0412 y 0.9373, respectivamente. En el conjunto de entrenamiento se obtuvo un RMSE de 2,8975 y un R^2 de 0,9656, mientras que en el conjunto de test se obtuvo un R^2 de 0,8822 y un RMSE de 6,9413.

Regresión de mínimos cuadrados parciales (PLS)

PLS es un modelo de regresión que utiliza el análisis de componentes principales para optimizar el poder explicado de las variables de respuesta. Estima los coeficientes de regresión para cada variable latente y determina el número óptimo de variables latentes minimizando el RMSE entre las variables de respuesta previstas y observadas.

El número óptimo de componentes utilizadas para el modelo PLS y determinado por VC fue 3, con un RMSE de 2,9778 y un R^2 de 0,9721. En el conjunto de entrenamiento, el RMSE fue 2,8826 y el R^2 fue 0,9659. En el conjunto de test, el RMSE fue 7,3079 y el R^2 fue 0,8750.

Random forest (RF)

RF es un modelo de machine leraning que combina múltiples árboles de decisión para mejorar la precisión y la solidez del modelo. Cada árbol de decisión en el modelo de RF se entrena en una muestra *bootstrap* del

conjunto de datos original, lo que significa que algunos puntos de datos quedan fuera del proceso de entrenamiento y se usan como muestras *out-of-bag* (OOB). RF selecciona aleatoriamente un subconjunto de características antes de evaluar cada división en un árbol individual, lo que reduce la correlación entre los árboles y evita el sobreajuste. El hiperparámetro *mtry* determina el número de características muestreadas aleatoriamente en cada división y se eligió probando diferentes valores y seleccionando el que dio como resultado el mejor rendimiento, utilizando VC con 5 pliegues. El mejor mtry empleando un método de búsqueda de rejilla de 1 a 30, fue 1, y el número de árboles se estableció en 500. El RSME y R^2 alcanzado por el modelo fue de 6.0952 y 0.8812, respectivamente. En el conjunto de entrenamiento se obtuvo un RMSE de 3.4298 y un R^2 de 0.9528, mientras que en el conjunto de test se obtuvo un R^2 de 0.8683 y un RMSE de 6.2846.

Regresión de vectores soporte (SVR)

SVR es un modelo de machine learning que utiliza un hiperplano para aproximar una función de mapeo entre las variables de entrada y las variables de salida. Encontrando el hiperplano que maximiza el margen entre los puntos más cercanos del conjunto de entrenamiento y el hiperplano. SVR utiliza dos parámetros importantes que deben ajustarse: la función de costo de pérdida (C) y el parámetro de regularización (γ). Además de seleccionar la función kernel, que determina el ancho del kernel. En este estudio, SVR se utilizó con el kernel de función de base radial (FBR).

Ambos hiperparámetros (C, γ) se optimizaron mediante un método de búsqueda de rejilla utilizando secuencias exponenciales de $\log_{2\gamma}$, $\log_{2C}$ en un rango de [–10, 10] cada 0,5. Los mejores resultados se obtuvieron para un γ de 1.381068×10^{-3} y un C de 1024 logrando un RMSE de 2.9527 y un R^2 de 0.9727. En el conjunto de entrenamiento, el RMSE fue de 2,7004 y el R^2 de 0,9701, mientras que, en el conjunto de test, el RMSE fue de 6,3364 y el R^2 de 0,9086. En resumen, la regresión SVR mostró el mejor rendimiento en la predicción del nivel de adulterante en muestras de miel utilizando un conjunto de datos que contenía miel de azahar y girasol.

Con el objetivo de mejorar la precisión de los modelos de regresión, se propuso clasificar las muestras de miel en función de su origen botánico y luego aplicar los modelos de regresión por separado. La discriminación de las muestras de miel según su origen botánico puede ayudar a garantizar que

los modelos de regresión se entrenen en datos que sean más homogéneos y que puedan capturar mejor las relaciones entre los diferentes tipos de miel, y esto puede conducir a predicciones más precisas del nivel de adulteración en muestras de miel.

CLASIFICACIÓN DE LAS MIELES POR SU ORIGEN BOTÁNICO

Análisis de componentes principales (ACP)

Se realizó un análisis de componentes principales (ACP) para identificar las diferencias entre la miel de azahar y de girasol, empleando todas las muestras ($D_{44x20002}$). La Figura 2 muestra las puntuaciones obtenidas por las muestras para las dos primeras componentes principales (CP). La primera componente principal (CP1) y la segunda componente principal (CP2) representaron el 86,7 % y el 7,5 % de la varianza acumulada, respectivamente, cubriendo un 94,2 % de la varianza total del conjunto de datos. En la gráfica de puntuaciones (Figura 2) las muestras de diferente origen botánico se distribuyeron en dos zonas claramente diferenciadas en base a sus puntuaciones con respecto a la CP1 y CP2.

La primera componente principal (CP1) permite distinguir entre los dos tipos de miel, ya que representaba un gran porcentaje de la varianza total en el conjunto de datos. Las muestras con puntuaciones positivas en la CP1 se asociaron con miel de azahar, mientras que los puntajes negativos correspondieron a miel de girasol. Además, la proximidad del valor de CP1 de las muestras a 0 indicó su nivel de adulteración, siendo las muestras con niveles más altos de adulterantes las más cercanas a 0 en ambos tipos de miel.

La segunda componente principal (CP2) no proporcionó información adicional significativa más allá de CP1, ya que representó un porcentaje relativamente pequeño de la varianza total en el conjunto de datos.

Modelos supervisados de machine learning para la clasificación según el origen botánico

El análisis de componentes principales puede ser un primer paso útil para identificar patrones en los datos. En este caso, parece que los dos grupos son claramente separables en función de los datos obtenidos por

GC-ultrarrápida. Sin embargo, para clasificar con precisión las nuevas muestras en uno de los dos grupos, es necesario aplicar algoritmos de ML supervisados. Algunos de los algoritmos empleados para construir los modelos predictivos fueron las máquinas de vectores soporte (MVS) y Random Forest (RF).

Para la clasificación, el conjunto de datos ($D_{44x20002}$) se dividió aleatoriamente en un conjunto de entrenamiento del 75% (n=33) y un conjunto de test del 25% (n=11). Luego, el conjunto de entrenamiento fue preprocesado de la misma manera que los modelos de regresión, utilizando el algoritmo de selección de variables *Boruta*. En este caso, se seleccionaron 330 características como relevantes en la determinación del origen botánico de la miel, lo que resultó en un conjunto de datos de entrenamiento reducido de D_{33x330}. El rendimiento del modelo también se evaluó mediante VC con 5 pliegues, y la métrica utilizada para evaluar el rendimiento de los modelos máquinas de vectores soporte (SVM) y RF generados fue la precisión.

En la Tabla II se muestra un resumen de la precisión alcanzada por estos modelos y los parámetros optimizados. Como puede verse, ambos modelos lograron una precisión del 100 % en la VC con 5 pliegues, el conjunto de entrenamiento y el conjunto de test.

Los resultados obtenidos a través de los modelos RF y SVR confirmaron la aplicabilidad de estas técnicas para la discriminación de mieles según su origen botánico.

Predicción del nivel de adulterante en miel de azahar y girasol

Una vez identificado el origen floral, el siguiente paso fue establecer el nivel de adulteración en cada tipo de miel. Para ello, se aplicó el algoritmo de selección de variables *Boruta* al conjunto de datos perteneciente a cada tipo de miel. En el caso de la miel de azahar se seleccionaron 50 variables (D_{22x50}), y 59 variables para la miel de girasol (D_{22x59}). Por último, se evaluaron los mismos modelos de regresión empleados previamente para el conjunto de datos que contenía todas las muestras de miel. En la Tabla III y la Tabla IV se muestra un resumen de los parámetros optimizados, el rendimiento de los modelos de regresión y el rendimiento en el conjunto de test y entrenamiento aplicado a todas las muestras de miel y solo a la miel de azahar y de girasol.

Es importante señalar que el modelo de SVR obtuvo el mejor rendimiento para el conjunto de datos que incluía todas las muestras de miel, con un RSME de 6.3364 y un R^2 de 0.9086. Sin embargo, al tratar las muestras de miel de azahar y de girasol por separado, se observó una mejora en el rendimiento de todos los modelos de regresión. Además, se encontró que todos los modelos de regresión probados en miel de azahar y de girasol por separado obtuvieron un rendimiento similar muy alto, con un R^2 en el conjunto de test superior a 0.9900, excepto el modelo de RIDGE en miel de azahar (0.9843). En cuanto al RMSE en el conjunto de test, el modelo LASSO resultó ser el mejor para predecir las propiedades de la miel de azahar y de girasol, obteniendo valores de 1.3064 y 1.3574, respectivamente. Esto subraya la importancia de considerar diferentes modelos de regresión y seleccionar el de mejor rendimiento para un conjunto de datos específico.

CONCLUSIONES

En general, la combinación de CG-ultrarrápida con algoritmos de machine learning es una herramienta eficaz para detectar la adulteración de la miel y garantizar su calidad y autenticidad.

Los modelos desarrollados en este estudio pueden ser utilizados para predecir la adulteración en la miel de azahar y girasol con una precisión y fiabilidad elevada. La regresión SVR mostró el mejor rendimiento en la predicción del nivel de adulterante en muestras de miel utilizando un conjunto de datos que contenía miel de azahar y girasol, obteniendo valores de R^2 superiores 0.90 en el conjunto de test. Además, se encontró que todos los modelos de regresión probados en miel de azahar y de girasol por separado obtuvieron un mejor rendimiento, con valores de R^2 en el conjunto de test superior a 0.99, siendo el modelo LASSO el mejor para predecir las adulteraciones en miel de azahar y de girasol por separado.

Los resultados de este estudio pueden ser muy útiles para la industria de la miel y las agencias reguladoras, ya que proporcionan información valiosa sobre las técnicas de machine learning más efectivas para detectar la adulteración de la miel de forma rápida, objetiva y automatizada.

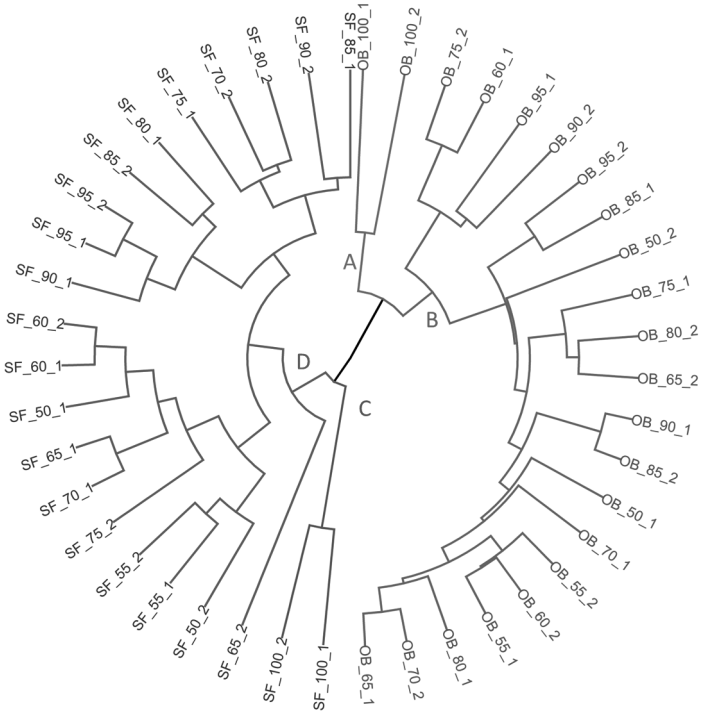

Figura 1. Dendrograma circular resultante del Análisis jerárquico de conglomerados (HCA) del conjunto de datos ($D_{44}x_{20002}$) con CG-ultrarrápida; Las muestras de miel se colorean según su origen botánico: Girasol (SF) en rosa y Azahar (OB) en naranja. Se utilizó el método de promedios con distancias euclidianas.

Figura 2. Gráfico de puntuación de ACP para miel de azahar y girasol. Girasol (SF) en rosa y Azahar (OB) en naranja.

Tabla 1. Descripción de muestras.

Tipo de miel	% Pureza	Nombre	Tipo de miel	% Pureza	Nombre
Girasol (SF)	50	SF_50_1	Azahar (OB)	50	OB_50_1
		SF_50_2			OB_50_2
	55	SF_55_1		55	OB_55_1
		SF_55_2			OB_55_2
	60	SF_60_1		60	OB_60_1
		SF_60_2			OB_60_2
	65	SF_65_1		65	OB_65_1
		SF_65_2			OB_65_2
	70	SF_70_1		70	OB_70_1
		SF_70_2			OB_70_2
	75	SF_75_1		75	OB_75_1
		SF_75_2			OB_75_2
	80	SF_80_1		80	OB_80_1
		SF_80_2			OB_80_2
	85	SF_85_1		85	OB_85_1
		SF_85_2			OB_85_2
	90	SF_90_1		90	OB_90_1
		SF_90_2			OB_90_2
	95	SF_95_1		95	OB_95_1
		SF_95_2			OB_95_2
	100	SF_100_1		100	OB_100_1
		SF_100_2			OB_100_2

Tabla 2. Resumen de los modelos de clasificación de origen botánico en muestras de miel. RF = Random forest; MVS = Máquina de vectores soporte con kernel radial.

Modelos	**Hiperparámetros**	**VC-5 pliegues Precisión (%)**	**Conjunto de entrenamiento Precisión (%)**	**Conjunto de test Precisión (%)**
RF	mtry = 19.467	100	100	100
MVS	γ= 4.882812 x 10^{-4} C = 0.5	100	100	100

Tabla 3. Parámetros optimizados y rendimiento del modelo por validación cruzada con 5 pliegues de los modelos de regresión aplicados a todas las muestras de miel y solo a la miel azahar y girasol. LASSO = Mínima Contracción Absoluta y Operador de Selección; RIDGE = regresión de Ridge; ENET = regresión neta elástica; PLS = regresión de mínimos cuadrados parciales; RF = Random forest; SVR = Soporte de regresión vectorial con núcleo radial; TODOS = Contiene todas las muestras de miel de origen botánico; OB = Miel de azahar; SF = Miel de girasol.

Modelos	**Datos**	**Hiperparámetros**	**Rendimiento VC-5 pliegues**	
			RMSE	**R^2**
LASSO	TODOS	$\lambda = 0.1321$	4.0412	0.9373
	OB	$\lambda = 0.3053$	2.3400	0.9904
	SF	$\lambda = 0.2009$	2.7121	0.9915
RIDGE	TODOS	$\lambda = 10$	6.3528	0.9007
	OB	$\lambda = 10$	3.5080	0.9733
	SF	$\lambda = 10$	3.4041	0.9911
ENET	TODOS	$\alpha = 0.0294$ $\lambda = 1$	4.0412	0.9373
	OB	$\alpha = 0.1$ $\lambda = 0.9234$	2.2973	0.9892
	SF	$\alpha = 0.1$ $\lambda = 0.9995$	2.6309	0.9934
PLS	TODOS	3 CPs	2.9778	0.9721
	OB	3 CPs	2.1123	0.9959
	SF	3 CPs	2.7097	0.9946
RF	TODOS	Mtry = 1	6.0952	0.8812
	OB	Mtry = 46	4.4522	0.9864
	SF	Mtry = 40	3.3990	0.9803
SVR	TODOS	$\gamma = 1.3810 \times 10\text{-}3$ C = 1024	2.9527	0.9727
	OB	$\gamma = 9.7656 \times 10\text{-}4$ C = 32	3.0985	0.9892
	SF	$\gamma = 3.906 \times 10\text{-}3$ C= 45.2548	2.0362	0.9922

Tabla 4. Rendimiento de los modelos de regresión aplicados a todas las muestras de miel y solo a la miel de azahar y de girasol en el conjunto de test y entrenamiento. LASSO = Mínima Contracción Absoluta y Operador de Selección; RIDGE = regresión de Ridge; ENET = regresión neta elástica; PLS = regresión de mínimos cuadrados parciales; RF = Random forest; SVR = Soporte de regresión vectorial con núcleo radial; TODOS = Contiene todas las muestras de miel de origen botánico; OB = Miel de azahar; SF = Miel de girasol.

Modelos	**Datos**	**Rendimiento del conjunto de entrenamiento**		**Rendimiento del conjunto de test**	
		RMSE	**R^2**	**RMSE**	**R^2**
LASSO	TODOS	3.7446	0.9433	6.3353	0.8739
	OB	1.6154	0.9896	1.3064	0.9942
	SF	1.9564	0.9850	1.3574	0.9987
RIDGE	TODOS	5.4803	0.8774	6.7255	0.8363
	OB	3.7368	0.9772	5.0263	0.9843
	SF	3.2928	0.9786	2.3002	0.9983
ENET	TODOS	2.8975	0.9656	6.9413	0.8822
	OB	1.6037	0.9895	1.3645	0.9936
	SF	3.6678	0.9476	1.3985	0.9986
PLS	TODOS	2.8826	0.9659	7.3079	0.8750
	OB	1.7374	0.9874	1.7319	0.9918
	SF	2.1588	0.9816	1.5181	0.9985
RF	TODOS	3.4298	0.9528	6.2846	0.8683
	OB	4.6664	0.9877	5.1115	0.9642
	SF	1.9831	0.9893	2.0996	0.9959
SVR	TODOS	2.7004	0.9701	6.3364	0.9086
	OB	1.6394	0.9898	1.4348	0.9945
	SF	1.4521	0.9917	1.9276	0.9989

BIBLIOGRAFÍA

1. Siddiqui, A.J.; Musharraf, S.G.; Choudhary, M.I.; Rahman, A. Ur "Application of Analytical Methods in Authentication and Adulteration of Honey". *Food Chem.* 2017, *217*, 687–698, doi:10.1016/j.foodchem.2016.09.001.

2. Aliaño-González, M.J.; Ferreiro-González, M.; Espada-Bellido, E.; Palma, M.; Barbero, G.F. "A Screening Method Based on Headspace-Ion Mobility Spectrometry to Identify Adulterated Honey". *Sensors (Switzerland)* **2019**, *19*, doi:10.3390/s19071621.

3. European Commission Council Directive 2001/110/EC of 20 December 2001 Relating to Honey. *Off. J. Eur. Communities, 10(12*. 2002, *110*, 47–50.

4. Mădaş, N.M.; Mărghitaş, L.A.; Dezmirean, D.S.; Bonta, V.; Bobiş, O.; Fauconnier, M.L.; Francis, F.; Haubruge, E.; Nguyen, K.B. "Volatile Profile and Physico-Chemical Analysis of Acacia Honey for Geographical Origin and Nutritional Value Determination". *Foods* 2019, *8*, 5–9, doi:10.3390/foods8100445.

5. Manyi-Loh, C.E.; Ndip, R.N.; Clarke, A.M. "Volatile Compounds in Honey: A Review on Their Involvement in Aroma, Botanical Origin Determination and Potential Biomedical Activities". *Int. J. Mol. Sci.* 2011, *12*, 9514–9532.

6. Brar, D.S.; Pant, K.; Krishnan, R.; Kaur, S.; Rasane, P.; Nanda, V.; Saxena, S.; Gautam, S. "A Comprehensive Review on Unethical Honey: Validation by Emerging Techniques". *Food Control* 2023, *145*, 109482, doi:10.1016/J.FOODCONT.2022.109482.

7. Da Silva, P.M.; Gauche, C.; Gonzaga, L.V.; Costa, A.C.O.; Fett, R. "Honey: Chemical Composition, Stability and Authenticity". *Food Chem.* 2016, *196*, 309–323, doi:10.1016/j.foodchem.2015.09.051.

8. Aliaño-González, M.J.; Ferreiro-González, M.; Espada-Bellido, E.; Barbero, G.F.; Palma, M. "Novel Method Based on Ion Mobility Spectroscopy for the Quantification of Adulterants in Honeys". *Food Control* 2020, *114*, 107236, doi:10.1016/j.foodcont.2020.107236.

9. Zhang, G.; Abdulla, W. "On Honey Authentication and Adulterant Detection Techniques". *Food Control* 2022, *138*, 108992, doi:10.1016/J.FOODCONT.2022.108992.

10. Codex Alimentarius Revised Codex Standard for Honey. *Codex stan* 2001, *12*, 1982.

11. Naila, A.; Flint, S.H.; Sulaiman, A.Z.; Ajit, A.; Weeds, Z. "Classical and Novel Approaches to the Analysis of Honey and Detection of Adulterants". *Food Control* 2018, *90*, 152–165.

12. Tosun, M. "Detection of Adulteration in Honey Samples Added Various Sugar Syrups with 13C/12C Isotope Ratio Analysis Method". *Food Chem.* 2013, *138*, 1629–1632, doi:10.1016/j.foodchem.2012.11.068.

13. Ruiz-Matute, A.I.; Soria, A.C.; Martínez-Castro, I.; Sanz, M.L. "A New Methodology Based on GC–MS To Detect Honey Adulteration with Commercial Syrups". 2007, doi:10.1021/jf070559j.

14. Song, X.; She, S.; Xin, M.; Chen, L.; Li, Y.; Heyden, Y. Vander; Rogers, K.M.; Chen, L. "Detection of Adulteration in Chinese Monofloral Honey Using 1H Nuclear Magnetic Resonance and Chemometrics". *J. Food Compos. Anal.* 2020, *86*, 103390, doi:10.1016/J.JFCA.2019.103390.

15. Dumancas, G.G.; Ellis, H. "Comprehensive Examination and Comparison of Machine Learning Techniques for the Quantitative Determination of Adulterants in Honey Using Fourier Infrared Spectroscopy with Attenuated Total Reflectance Accessory". *Spectrochim. Acta–Part A Mol. Biomol. Spectrosc.* 2022, *276*, doi:10.1016/J.SAA.2022.121186.

16. Mitra, P.K.; Karmakar, R.; Nandi, R.; Gupta, S. "Low-Cost Rapid Workflow for Honey Adulteration Detection by UV–Vis Spectroscopy in Combination with Factorial Design, Response Surface Methodology and Supervised Machine Learning Classifiers". *Bioresour. Technol. Reports* **2023**, *21*, 101327, doi:10.1016/J.BITEB.2022.101327.

17. Ferreiro-González, M.; Espada-Bellido, E.; Guillén-Cueto, L.; Palma, M.; Barroso, C.G.; Barbero, G.F. "Rapid Quantification of Honey Adulteration by Visible-near Infrared Spectroscopy Combined with Chemometrics". *Talanta* 2018, *188*, 288–292, doi:10.1016/j.talanta.2018.05.095.

18. Wu, X.; Xu, B.; Ma, R.; Niu, Y.; Gao, S.; Liu, H.; Zhang, Y. "Identification and Quantification of Adulterated Honey by Raman Spectroscopy Combined with Convolutional Neural Network and Chemometrics". *Spectrochim. Acta–Part A Mol. Biomol. Spectrosc.* 2022, *274*, 121133, doi:10.1016/j.saa.2022.121133.

19. Izquierdo, M.; Lastra-Mejías, M.; González-Flores, E.; Cancilla, J.C.; Pérez, M.; Torrecilla, J.S. "Convolutional Decoding of Thermographic Images to Locate and Quantify Honey Adulterations". *Talanta* 2020, *209*, 120500, doi:10.1016/j.talanta.2019.120500.

20. Peris, M.; Escuder-Gilabert, L. "Electronic Noses and Tongues to Assess Food Authenticity and Adulteration". *Trends Food Sci. Technol.* 2016, *58*, 40–54, doi:10.1016/J.TIFS.2016.10.014.