Determinación del origen geográfico de dos variedades de café mediante espectroscopia NIR
Determination of the geographical origin of two coffee varieties by NIR spectroscopy

Jimy Oblitas-Cruz, M.Sc., Yuleyci Cieza-Rimarachin[2], and Wilson Castro-Silupu, Dr.[3]
[1] Universidad Privada del Norte., Perú, jimy.oblitas@upn.edu.pe
[2] Universidad Nacional de Cajamarca, Perú, yciezar16_1@unc.edu.pe
[3] Universidad Nacional de la Frontera, Perú, wcastro@unf.edu.pe

Abstract*– The objective was to implement a non-invasive classification system for green coffee beans by using near-infrared spectroscopy (NIR) and multivariate data analysis. For this, 4 types of coffee were analyzed, according to variety and geographical location. The samples were repeated 5 times. The observed NIR spectrum was absorbance in the range of 1100 and 2500 nm. In order to reduce the data, the analysis of main components was used by testing 24 classification models, from which the one that reached the highest level of precision was the Linear Support Vector Machine (SVM) algorithm, reaching 98.8%, achieving fairly satisfactory discrimination with values of PC1 (97.9%), PC2 (1.9%) and PC3 (0.1%), reaching a total cumulative variation of the contribution of the first 3 PCs of 99.9%. These values demonstrated that NIR spectroscopy is a valid alternative for classification by geographical origin and variety of green coffee beans.

***Keywords*- Green coffee beans, NIR spectroscopy, Geographical origin**

Resumen*–El objetivo fue implementar un sistema de clasificación no invasivo de granos de café verde haciendo uso de la espectroscopia de infrarrojo cercano (NIR) y el análisis de datos multivariados. Para ello se analizó 4 clases de café, de acuerdo a variedad y ubicación geográfica. Las muestras fueron repetidas 5 veces. El espectro NIR observado fue la absorbancia en el rango de 1100 y 2500 nm. Para poder reducir los datos se usó el análisis de componentes principales probando 24 modelos de clasificación, del cual el que alcanzó el mayor nivel de precisión fue el algoritmo tipo Support vector machine (SVM) del tipo lineal, alcanzado un 98.8%, logrando una discriminación bastante satisfactoria con valores de PC1 (97,9%), PC2 (1,9%) y PC3 (0,1%), alcanzando una variación total acumulada de la contribución de los primeros 3 PC del 99,9%. Estos valores demostraron que la espectroscopia NIR es una alternativa válida para clasificación por origen geográfico y variedad de granos de café verde

***Palabras clave*- Granos de café verde, Espectroscopia NIR, Origen geográfico**

I. Introducción

El café es uno de los cultivos básicos tropicales por los que la creciente demanda mundial está motivando a los caficultores a expandir la tierra cultivada [1], actualmente, el café se ha convertido en el principal producto agrícola de exportación en Perú y representa el 6% del área agrícola peruana, las plantaciones de café están instaladas en 17 regiones, 67 provincias y 338 distritos y un tercio del empleo agrícola está relacionado al mercado del café [2]

Perú produce casi exclusivamente café Arábica, del cual más del 70% es de la variedad Typica, seguido de Caturra (20%) y otras (10%) [3], junto con ello se han determinado cambios debido a nuevas zonas geográficas de cultivo y de gestión en la biodiversidad que han mejorado la adaptación y la sostenibilidad de la producción de café en pequeña escala[4], estos cambios son relacionados a la composición química de los granos de café crudo que, a su vez, está muy relacionada con sus regiones geográficas de cultivo [5]

En el mercado del café existe una relación entre el precio de venta y la calidad, siendo este uno de los criterios clave que pueden ayudar a los productores a ganar mayor posición en el mercado mundial del café [6]; la forma de medir la calidad siempre se ha dado de manera química siendo los compuestos más usuales de medición la cafeína, ácidos clorogénicos, sacarosa y la trigonelina [7]

Se han realizado ensayos de clasificación de calidad en granos de café verde y tostado[8], [9], sin embargo tal como menciona Barboza [10] los parámetros agronómicos como las condiciones edafoclimáticas y la genética del café impactan en la composición de los granos de café verde y, en consecuencia, en la calidad del café resultante.

Debido a las diferencias en términos de precio y calidad, la disponibilidad de instrumentos efectivos para discriminar entre café Arábica y Robusta es extremadamente importante [11], por ello es imprescindible generar estudios de métodos que puedan diferenciar rápidamente la calidad y origen del producto, los métodos químicos y bioquímicos son costosos, complejos de usar y consumen mucho tiempo [12], por ello observar la viabilidad de usas técnicas espectroscópicas alternativas, como la espectroscopia infrarroja cercana, brindan una solución válida para superar algunos de los inconvenientes antes mencionados, ya que permiten realizar análisis verdes, simples, rápidos y no invasivos [13].

Las herramientas espectroscópicas usadas y aplicadas en productos agrarios es diversa como las imágenes hiperespectrales [14], espectroscopia del Dominio del tiempo

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2021.1.1.111
ISBN: 978-958-52071-8-9 ISSN: 2414-6390

[15], espectroscopia dieléctricas [16], en particular, la espectroscopia del Infrarrojo cercano a sido utilizada para discriminar café [11]. La espectroscopia de infrarrojo cercano (NIR) es una técnica de análisis rápida y no destructiva con buena reproducibilidad, ha sido ampliamente aplicada para la detección rápida de composiciones de alimentos y evaluación de calidad en productos alimenticios [17].

Junto con el análisis espectral es necesario generar análisis multivariado como el análisis de componentes principales (PCA), que es una técnica multivariante que se usa para describir la relación entre varias variables de respuesta y para explicar la variación total en los datos, este tipo de análisis es muy útil cuando las variables en estudio están altamente correlacionadas (positiva o negativamente) o cuando el número de variables independientes es grande [18].

El objetivo buscado en la investigación es evaluar el potencial de la espectroscopia del infrarrojo cercano (NIR) y proponer una metodología para poder diferenciar 2 variedades de café de 2 zonas geográficas distintas en Perú usando un enfoque cualitativo y cuantitativo junto con un análisis multivariado.

II. MATERIALES Y METODOS

A. Muestras de café

En el presente estudio se utilizaron 4 muestras de café, que se clasificaron en 4 clases de acuerdo a la variedad y origen geográfico como se muestra en la Figura 1 y tabla 1.

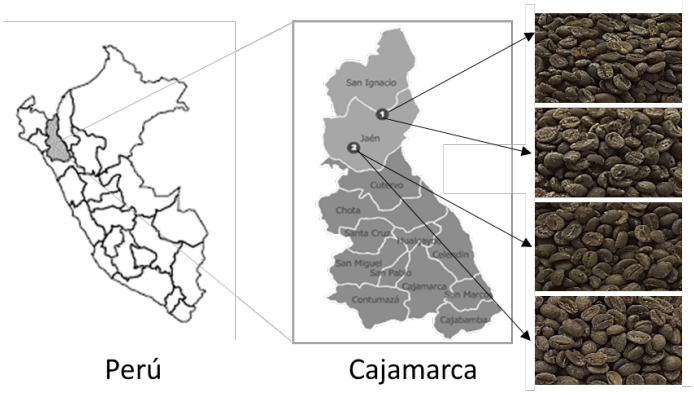

Fig. 1 Ubicación geográfica de muestras

TABLA I
CLASIFICACION DE MUESTRAS

Muestras de café	Variedades	Origen Geográfico
Clase 1.	Caturra	Jaén
Clase 2.	Caturra	San Ignacio
Clase 3.	Typica	Jaén
Clase 4.	Typica	San Ignacio

Luego del lavado de frutos de cereza y secado, se tomaron las muestras por triplicado de tal manera que de cada muestra se divida en 20 submuestras de 70g para luego ser puesta en un espectrómetro infrarrojo cercano.

B. Toma de datos NIR

Cada muestra de 70g se mezcló siempre cuidadosamente para cada una de las tres repeticiones. En la investigación se utilizó un espectrómetro infrarrojo cercano Unity Scientific NIRS (SpectraStar 2500XL, EE.UU.) equipada con lámpara halógena de tungsteno como fuente de luz y detector InGaAs (Indio – Galio – Arsénico) en el rango de 1100 y 2500 nm, con una resolución de de 3 nm y 467 longitudes de onda.

C. Pretratamiento de perfiles espectrales

Se aplicó mejoras espectrales como filtrado espectral, suavizado, normalización, centrado medio y auto escalado, esto de acuerdo a ElMasry [19] es necesario para mejorar los perfiles espectrales extraídos ya que contienen ruido y variabilidad.

D. Analisis multivariante

Se usará el método de Análisis de Componentes Principales (PCA) como un método de reconocimiento común sin supervisión, se aplicó en primer lugar para la exploración inicial para visualizar el marco de datos e identificar observaciones confusas o valores atípicos. PCA se ha convertido en una de las herramientas más amplias para explorar similitudes y patrones ocultos entre muestras donde la relación en los datos y la agrupación son hasta poco claras[20]. En la presente investigación se usó para reducir las dimensiones de la matriz de datos de las muestras y se extrajo la información principal registrados para obtener una visión general utilizando PCA.

III. RESULTADOS

A. Determinación de características espectrales

los espectros NIR se muestran en la figura 2, donde se graficó con valores de la Reflectancia, los picos mostrados en la figura son similares a los reportados en trabajos de café verde [21]. El gráfico ha sido pretratado ya que debido a que la longitud de onda en la radiación electromagnética NIR es comparable al tamaño de partícula en muestras biológicas, el espectro NIR registrado puede verse afectado por efectos de dispersión no deseados[22], eso paso es necesario para el modelado posterior.

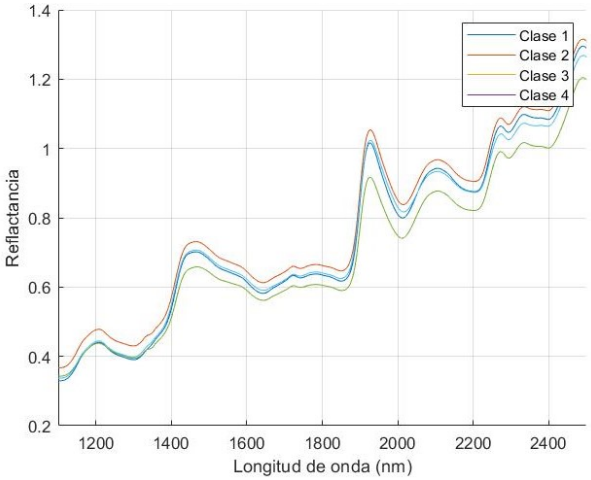

Fig. 2 Espectro NIR promedio de muestras de café

B. Análisis exploratorio de datos

El análisis de componentes principales solo se puede utilizar como un método de reconocimiento de patrones no supervisado, este procedimiento puede indicar la tendencia de los datos de forma gráfica, por ello se aplicó a todo el conjunto de datos NIR para evaluar si se podía clasificar adecuadamente a las muestras de café de acuerdo a su origen geográfico y variedad.

La precisión del entrenamiento de los modelos utilizados se evaluó mediante el indicador de precisión (%). Todos los modelos utilizaron una validación cruzada (15 pliegues). Se probó 24 modelos para el análisis, se muestra en la tabla I los tres modelos con mejor precisión de clasificación. Los tipos de modelos que se probaron incluyen modelos del tipo: decision trees, discriminant analysis, support vector machines, logistic regression, nearest neighbors, naive Bayes, and ensemble classification. Además, se tuvo en cuenta el sobreajuste del modelo, para lo cual todos los modelos usaron una validación cruzada.

TABLA II
MODELOS CON MEJOR PRECISIÓN

Modelo	Precisión
Linear SVM	98.8 %
Quadratic SVM	97.5 %
Ensemble of subspace discriminant	96.3 %

Tal como se observa el mejor modelo lo plantea el algoritmo tipo Support vector machine (SVM) del tipo lineal, el cual se usará en los análisis posteriores.

Se presenta el modelo para PCA no supervisado para las variedades y orígenes del café. El modelo de PCA se presentan como modelos de múltiples clases, es decir, cada muestra de café se presenta por separado como una clase. Como PCA es un algoritmo no supervisado, la información de la clase no afectará al modelo, ya que crea una separación de las muestras basada solo en los datos espectrales[23], con el objetivo de encontrar el número adecuado de Componente principales (PC), se analizó el que alcanzara un nivel de explicación d el varianza de más de 99%, este se observa en la figura 3, donde con 3 PC se logra este objetivo.

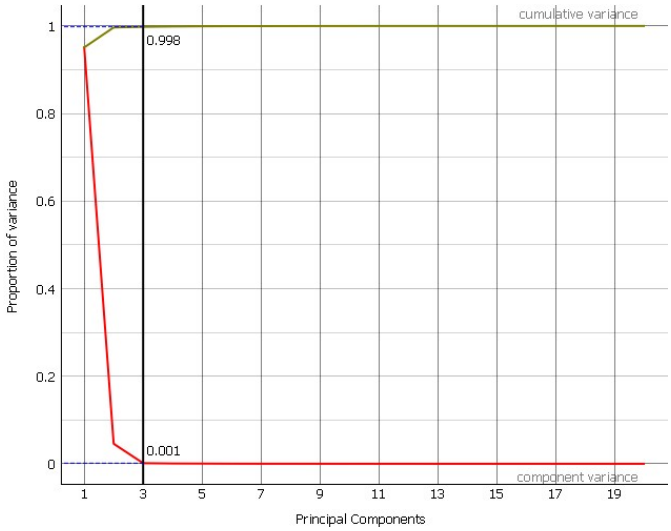

Fig. 3. Varianza explicada por PC

El modelo de PCA aplicado llegó a una discriminación bastante satisfactoria, PC1 dio una explicación del 97,9% de la varianza, PC2 dio una explicación del 1,9% de la varianza y PC3 dio una explicación del 0,1% de la varianza. La tasa de variación total acumulada de la contribución de los primeros 3 PC fue del 99,9%. Como el modelo de PCA sin supervisión indicó una buena separación entre las clases, esto podría mejorarse aún más con el uso de un algoritmo supervisado, con la adición de información de clase conocida[24]

Para visualizar la tendencia de agrupamiento de estas muestras, se trazó un gráfico de dispersión con los primeros 3 componentes principales (PC) emitidos desde PCA (es decir, PC1, PC2, PC3), que se muestran en la figura 4.

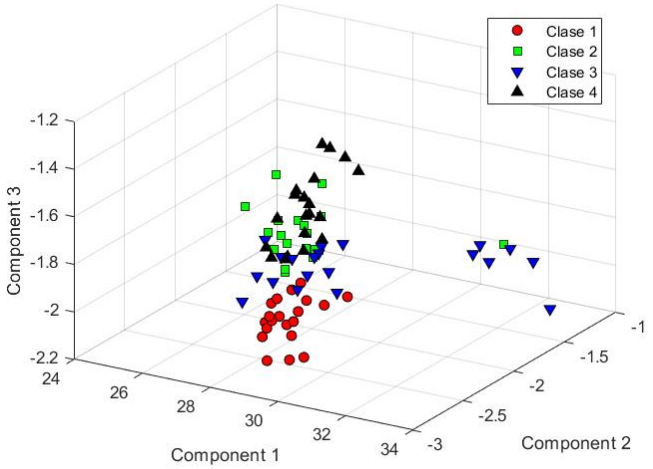

Fig. 4. Resultados de PCA

El gráfico de puntuación de PCA demuestra la posibilidad de separar las muestras de acuerdo a su origen geográfico esto ya que los granos de café dan como resultado cambios en sus espectros. Dado que es posible determinar el origen geográfico de los granos de acfé mediante NIRS, cabe señalar que todas las muestras se han recolectado en la misma temporada. Sin embargo, las diferencias estacionales podrían tener más influencia en los espectros que las diferencias regionales [25].

C. Modelo de clasificación

La clasificación es un tipo de aprendizaje automático supervisado en el que un algoritmo "aprende" a clasificar nuevas observaciones a partir de ejemplos de datos etiquetados, ebido a que la variabilidad siguió un perfil lineal, el modelo de clasificación denominado Support Vector Machines (SVM), fuel que arrojo la mejor precisión, la literatura muestra que este modelo de regresión obtenidos con espectros NIR dan muy buenos resultados[26]. SVM tiene un rendimiento relativamente bueno y fácil accesibilidad, siendo su mayor ventaja la capacidad para modelar relaciones no lineales [27]

Support Vector Machines dio un nivel de precisión de 75.0% para la discriminación por origen geográfico y variedad en cafés verdes, para ello se calculó la matriz de confusión para cada clase, como se muestra en la Figura 5.

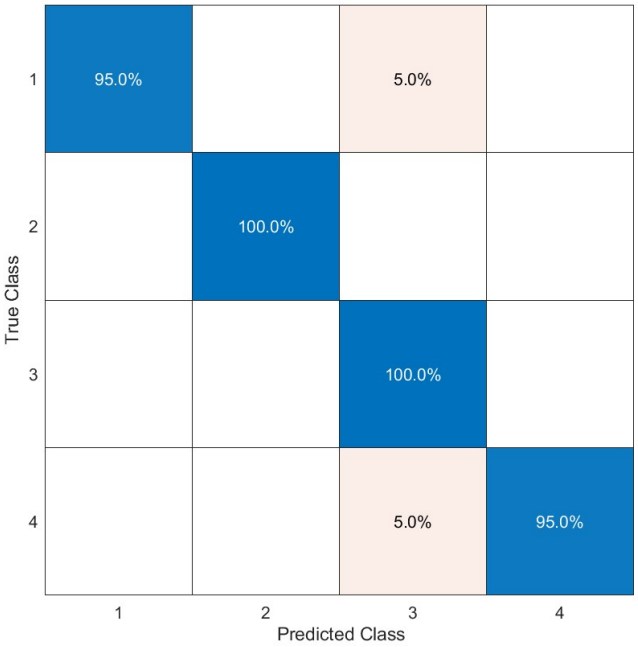

Fig. 5. Matriz de Confusión para modelo SVM

A partir de la matriz de confusión, se puede observar que aún hay clases que no se pueden predecir adecuadamente, lo cual se podría mejorar utilizando tratamientos de pre-procesamiento de datos para obtener un mayor nivel de discriminación entre clase del café verde.

Con el objetivo de mejorar el nivel de precisión se optimizó el algoritmo de clasificación SVM usando el algoritmo de error de clasificación mínimo [28] obteniéndose un nivel de precisión de 98.8 %, en la figura 6 se muestra el gráfico de error de clasificación mínimo.

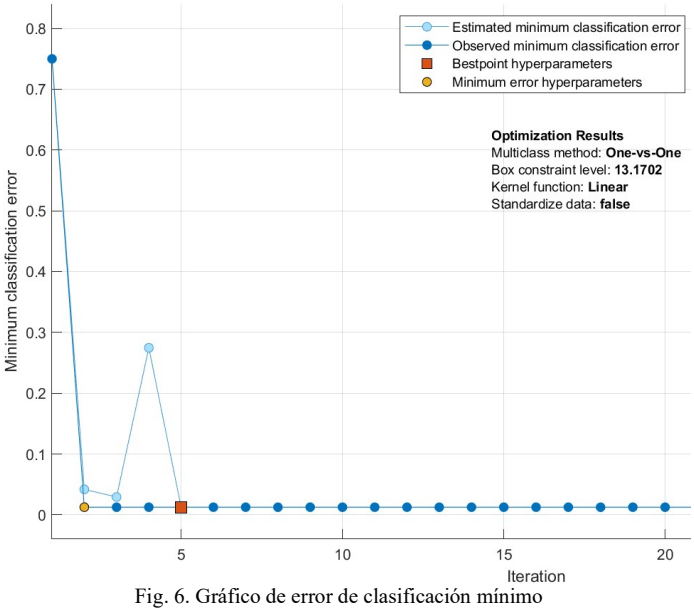

Fig. 6. Gráfico de error de clasificación mínimo

Este tipo de optimización bayesiana prueba una combinación diferente de valores de hiperparámetros y actualiza el gráfico con el error mínimo de clasificación de validación observado hasta esa iteración, indicado en azul oscuro. Cuando la aplicación completa el proceso de optimización, selecciona el conjunto de hiperparámetros optimizados, indicado por un cuadrado rojo.

La figura 7 muestra las 4 clases en un gráfico de caja y bigote, donde los valores en color azul muestran los datos reales y los de color amarillo los datos generados por el mejor modelo de ajuste, en este caso SVM.

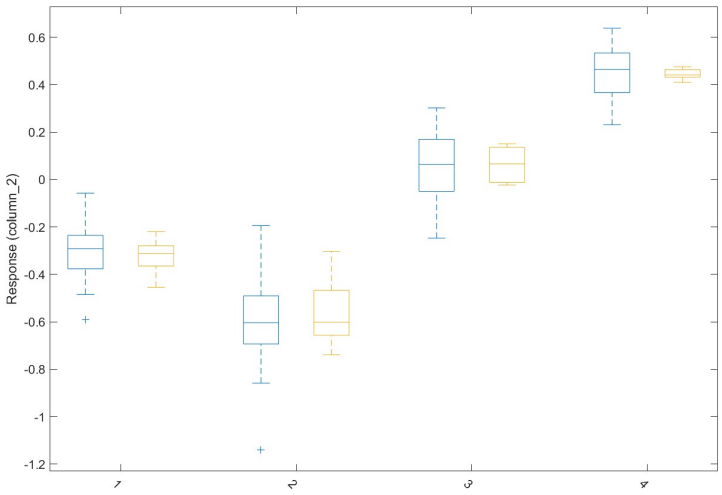

Fig. 7. Comparación de datos reales versus Modelo SVM

IV. CONCLUSIONES

Se ha demostrado la viabilidad de diferenciación por clase debido a origen geográfico y a variedad del tipo de café de la zona de Cajamarca en Perú usando espectros del Infrarrojo cercano (NIR), sumado a métodos de análisis multivariado donde se analizó los diversos algoritmos para el mejor reconocimiento de los patrones a identificar. De acuerdo al estudio se llegó a la conclusión que los métodos usados en el análisis de los datos son factibles de ser automatizados, con lo cual los resultados pasan a ser un punto inicial para poder ayudar a la clasificación rápida, no destructiva y fiable en los centros de acopio y exportación de este tipo de producto, convirtiéndose en una herramienta poderosa para detectar posibles adulteraciones por mezclas de café.

Dado que el método desarrollado tiene una buena capacidad para identificar el origen geográfico y variedad de muestras de café pelado, sería muy interesarte observar si existe la misma relación cuando el café aun no llega esta etapa, esto puede considerarse como una perspectiva para estudios futuros, además de incluir otras variedades o lugares de producción de una zona mucho más amplia en Perú.

AGRADECIMIENTOS

Los autores agradecen el apoyo financiero del Proyecto Concytec - Banco Mundial "Desarrollo de Modelos Predictivos de Calidad de Alimentos Basados en Tecnología de Imágenes THz", a través de su unidad ejecutora Fondecyt. [contrato número 006-2018-FONDECYT / BM-Mejoramiento de la infraestructura para la investigación (equipamiento)]

REFERENCES

[1] S. Jha, C. M. Bacon, S. M. Philpott, V. Ernesto Méndez, P. Läderach, y R. A. Rice, «Shade Coffee: Update on a Disappearing Refuge for Biodiversity», *BioScience*, vol. 64, n.º 5, pp. 416-428, may 2014, doi: 10.1093/biosci/biu038.

[2] «El Café Peruano». http://minagri.gob.pe/portal/485-feria-scaa/10775-el-cafe-peruano (accedido ene. 14, 2021).

[3] «El café de Perú», *Revista Fórum Café*. https://www.revistaforumcafe.com/el-cafe-de-peru (accedido ene. 14, 2021).

[4] R. E. Jezeer, P. A. Verweij, R. G. A. Boot, M. Junginger, y M. J. Santos, «Influence of livelihood assets, experienced shocks and perceived risks on smallholder coffee farming practices in Peru», *Journal of Environmental Management*, vol. 242, pp. 496-506, jul. 2019, doi: 10.1016/j.jenvman.2019.04.101.

[5] N. Q. Bitter, D. P. Fernandez, A. W. Driscoll, J. D. Howa, y J. R. Ehleringer, «Distinguishing the region-of-origin of roasted coffee beans with trace element ratios», *Food Chemistry*, vol. 320, p. 126602, ago. 2020, doi: 10.1016/j.foodchem.2020.126602.

[6] R. E. Jezeer, M. J. Santos, R. G. A. Boot, M. Junginger, y P. A. Verweij, «Effects of shade and input management on economic performance of small-scale Peruvian coffee systems», *Agricultural Systems*, vol. 162, pp. 179-190, may 2018, doi: 10.1016/j.agsy.2018.01.014.

[7] C.-L. Ky, J. Louarn, S. Dussert, B. Guyot, S. Hamon, y M. Noirot, «Caffeine, trigonelline, chlorogenic acids and sucrose diversity in wild Coffea arabica L. and C. canephora P. accessions», *Food Chemistry*, vol. 75, n.º 2, pp. 223-230, nov. 2001, doi: 10.1016/S0308-8146(01)00204-7.

[8] A. Sualeh, K. Tolessa, y A. Mohammed, «Biochemical composition of green and roasted coffee beans and their association with coffee quality from different districts of southwest Ethiopia», *Heliyon*, vol. 6, n.º 12, p. e05812, dic. 2020, doi: 10.1016/j.heliyon.2020.e05812.

[9] J. O. Cruz y W. C. Silupu, «Computer vision system for the optimization of the color generated by the coffee roasting process according to time, temperature and mesh size», *Ingenieria y Universidad*, vol. 18, n.º 2, pp. 355-368, 2014, doi: 10.11144/Javeriana.IYU18-2.cvso.

[10] M. de S. G. Barbosa, M. B. dos S. Scholz, C. S. G. Kitzberger, y M. de T. Benassi, «Correlation between the composition of green Arabica coffee beans and the sensory quality of coffee brews», *Food Chemistry*, vol. 292, pp. 275-280, sep. 2019, doi: 10.1016/j.foodchem.2019.04.072.

[11] R. Calvini, J. M. Amigo, y A. Ulrici, «Transferring results from NIR-hyperspectral to NIR-multispectral imaging systems: A filter-based simulation applied to the classification of Arabica and Robusta green coffee», *Analytica Chimica Acta*, vol. 967, pp. 33-41, may 2017, doi: 10.1016/j.aca.2017.03.011.

[12] I. Marquetti, J. V. Link, A. L. G. Lemes, M. B. dos S. Scholz, P. Valderrama, y E. Bona, «Partial least square with discriminant analysis and near infrared spectroscopy for evaluation of geographic and genotypic origin of arabica coffee», *Computers and Electronics in Agriculture*, vol. 121, pp. 313-319, feb. 2016, doi: 10.1016/j.compag.2015.12.018.

[13] A. Giraudo, S. Grassi, F. Savorani, G. Gavoci, E. Casiraghi, y F. Geobaldo, «Determination of the geographical origin of green coffee beans using NIR spectroscopy and multivariate data analysis», *Food*

Control, vol. 99, pp. 137-145, may 2019, doi: 10.1016/j.foodcont.2018.12.033.

[14] W. Castro, J. Oblitas, J. Maicelo, y H. Avila-George, «Evaluation of Expert Systems Techniques for Classifying Different Stages of Coffee Rust Infection in Hyperspectral Images», *International Journal of Computational Intelligence Systems*, vol. 11, n.º 1, pp. 86-100, ene. 2018, doi: 10.2991/ijcis.11.1.8.

[15] J. O. Cruz, «Terahertz Time-domain Spectroscopy (THz-TDS) for classification of blueberries according to their maturity», presentado en Proceedings of the 2020 IEEE Engineering International Research Conference, EIRCON 2020, 2020, doi: 10.1109/EIRCON51178.2020.9254046.

[16] T. C. Trigoso, J. O. Cruz, H. A. Minano, y W. C. Silupu, «Application of Machine Learning in the Discrimination of Citrus Fruit Juices: Uses of Dielectric Spectroscopy», presentado en Proceedings of the 2020 IEEE Engineering International Research Conference, EIRCON 2020, 2020, doi: 10.1109/EIRCON51178.2020.9253756.

[17] X. Wang, «7 - Near-infrared spectroscopy for food quality evaluation», en *Evaluation Technologies for Food Quality*, J. Zhong y X. Wang, Eds. Woodhead Publishing, 2019, pp. 105-118.

[18] A. F. M. Alkarkhi y W. A. A. Alqaraghuli, «Chapter 8 - Principal Components Analysis», en *Easy Statistics for Food Science with R*, A. F. M. Alkarkhi y W. A. A. Alqaraghuli, Eds. Academic Press, 2019, pp. 125-141.

[19] G. M. ElMasry y S. Nakauchi, «Image analysis operations applied to hyperspectral images for non-invasive sensing of food quality – A comprehensive review», *Biosystems Engineering*, vol. 142, pp. 53-82, feb. 2016, doi: 10.1016/j.biosystemseng.2015.11.009.

[20] J. Oblitas *et al.*, «The Use of Correlation, Association and Regression Techniques for Analyzing Processes and Food Products», *Mathematical and Statistical Applications in Food Engineering*, ene. 30, 2020. https://www.taylorfrancis.com/ (accedido sep. 28, 2020).

[21] A. Adnan, M. Naumann, D. Mörlein, y E. Pawelzik, «Reliable Discrimination of Green Coffee Beans Species: A Comparison of UV-Vis-Based Determination of Caffeine and Chlorogenic Acid with Non-Targeted Near-Infrared Spectroscopy», *Foods*, vol. 9, n.º 6, Art. n.º 6, jun. 2020, doi: 10.3390/foods9060788.

[22] Y. Sun *et al.*, «Rapid identification of geographical origin of sea cucumbers Apostichopus japonicus using FT-NIR coupled with light gradient boosting machine», *Food Control*, p. 107883, ene. 2021, doi: 10.1016/j.foodcont.2021.107883.

[23] T. F. McGrath *et al.*, «What are the scientific challenges in moving from targeted to non-targeted methods for food fraud testing and how can they be addressed? – Spectroscopy case study», *Trends in Food Science & Technology*, vol. 76, pp. 38-55, jun. 2018, doi: 10.1016/j.tifs.2018.04.001.

[24] P. Galvin-King, S. A. Haughey, y C. T. Elliott, «Garlic adulteration detection using NIR and FTIR spectroscopy and chemometrics», *Journal of Food Composition and Analysis*, vol. 96, p. 103757, mar. 2021, doi: 10.1016/j.jfca.2020.103757.

[25] M. Schmutzler y C. W. Huck, «Automatic sample rotation for simultaneous determination of geographical origin and quality characteristics of apples based on near infrared spectroscopy (NIRS)», *Vibrational Spectroscopy*, vol. 72, pp. 97-104, may 2014, doi: 10.1016/j.vibspec.2014.02.010.

[26] C. Malegori, E. J. Nascimento Marques, S. T. de Freitas, M. F. Pimentel, C. Pasquini, y E. Casiraghi, «Comparing the analytical performances of Micro-NIR and FT-NIR spectrometers in the evaluation of acerola fruit quality, using PLS and SVM regression algorithms», *Talanta*, vol. 165, pp. 112-116, abr. 2017, doi: 10.1016/j.talanta.2016.12.035.

[27] U. Thissen, M. Pepers, B. Üstün, W. J. Melssen, y L. M. C. Buydens, «Comparing support vector machines to PLS for spectral regression applications», *Chemometrics and Intelligent Laboratory Systems*, vol. 73, n.º 2, pp. 169-179, oct. 2004, doi: 10.1016/j.chemolab.2004.01.002.

[28] Z. Hu, Y. Peng, y S. Zhao, «A new sparse representation algorithm based on kernel spatial non-minimum residual error for classification», *Optik*, vol. 126, n.º 23, pp. 4665-4670, dic. 2015, doi: 10.1016/j.ijleo.2015.08.088.