# APLICACIÓN DE LA ESPECTROSCOPÍA NIR Y HERRAMIENTAS QUIMIOMÉTRICAS PARA LA DETERMINACIÓN DE COMPONENTES QUÍMICOS DEL CAFÉ VERDE ESPECIAL PRODUCIDO EN LA PROVINCIA DE CHIRIQUÍ, PANAMÁ

UNIVERSIDAD AUTÓNOMA DE CHIRIQUÍ

VICERRECTORÍA DE INVESTIGACIÓN Y POSGRADO

PROGRAMA DE MAESTRÍA EN INVESTIGACIÓN

# APLICACIÓN DE LA ESPECTROSCOPÍA NIR Y HERRAMIENTAS QUIMIOMÉTRICAS PARA LA DETERMINACIÓN DE COMPONENTES QUÍMICOS DEL CAFÉ VERDE ESPECIAL PRODUCIDO EN LA PROVINCIA DE CHIRIQUÍ, PANAMÁ

ELABORADO POR: STEPHANY M. REYES E.

ASESORA PRINCIPAL: DRA. ARACELLY VEGA, UNACHI

COASESORES:

DR. JAVIER SÁNCHEZ GALÁN, UTP

DR. PEDRO GONZÁLEZ BEERMANN, UNACHI

TRABAJO PARA OPTAR POR EL TÍTULO DE

MAGISTER EN INVESTIGACIÓN

DAVID, CHIRIQUÍ

REPÚBLICA DE PANAMÁ

2024

## DEDICATORIA

Este trabajo está dedicado especialmente a mis hijos Juan Arturo y David Alejandro, quienes fueron mi motivación y no mi excusa para culminar con éxito esta etapa.

A mi esposo José Augusto, por el apoyo incondicional, por sus palabras de motivación que me impulsaron a iniciar con este reto y por todo su amor.

A mi madre Laura Isabel por apoyarme constantemente y por su cariño.

## AGRADECIMIENTO

Al Centro de Investigación en Recursos Naturales de la Universidad Autónoma de Chiriquí, especialmente a su directora la Dra. Aracelly Vega por aceptar ser mi asesora y por sus contribuciones para que este trabajo se desarrollará de manera exitosa.

A mis coasesores el Dr. Javier Sánchez-Galán porque su asesoría fue clave para el inicio y la culminación de este trabajo y al Dr. Pedro González Beermann por su tiempo y por los aportes significativos para la mejora de la investigación.

A la Asociación de Cafés Especiales de Panamá por facilitar las muestras de café verde para la realización de este estudio.

A la Secretaría Nacional de Ciencia y Tecnología (SENACYT), especialmente a la Dirección de Desarrollo de Capacidades Científicas y Tecnológicas, quienes a través de la Convocatoria de Nuevos Investigadores proporcionaron los fondos para la ejecución de esta investigación.

Al Instituto de Innovación Agropecuaria de Panamá (IDIAP) y al Dr. Audino Melgar quienes permitieron el uso de sus equipos para realizar parte de los análisis químicos de esta investigación.

A mis compañeros Stepfanie Miranda y Javier De León por su apoyo durante la ejecución del proyecto.

Al programa de Maestría en Investigación de la Vicerrectoría de Investigación y Posgrado de la Universidad Autónoma de Chiriquí, y a todos los que de una u otra forma me ayudaron a culminar esta meta.

# ÍNDICE

## ÍNDICE DE TABLAS

## ÍNDICE DE FIGURAS

## RESUMEN

El cultivo de café en Panamá se realiza en diferentes provincias, siendo Chiriquí la provincia de mayor producción y en donde se produce y exporta café de especialidad. El café de especialidad es apreciado a nivel internacional por sus atributos sensoriales, y estos están relacionados con la composición química del café verde. Determinar estos compuestos por métodos tradicionales tomaría mucho tiempo y sería muy costoso por lo que el objetivo de este estudio fue aplicar la espectroscopía NIR y herramientas quimiométricas para la determinación rápida del contenido de cafeína, proteína y grasa en café verde especial, producido en la provincia de Chiriquí.

El contenido de proteína se determinó por el método Kjeldahl, la grasa cruda por el método AOCS AM 5-04 y la cafeína por HLPC-DAD. El espectro NIR se obtuvo en la región 570-1850 nm en modo de reflectancia difusa. Los métodos de calibración multivariada aplicados fueron PLS y PCR, con diferentes pretratamientos espectrales y en diferentes regiones del espectro. Además, para superar el problema de desbalance en las clases se aplicó el algoritmo SMOTE.

El mejor modelo para la estimación de la grasa cruda fue PLS-2da. derivada (15,2,2) con SMOTE aplicado a todo el rango espectral y se obtuvo un $R^2$ de 0.787, 0.631 y 0.601 para calibración, validación cruzada y predicción respectivamente. Los errores asociados a este modelo fueron 0.306, 0.407 y 0.444 para $RMSE_C$, $RMSE_{CV}$ y $RMSE_p$ respectivamente. Además, el RPD fue de 1.62 y el REP de 5.53.

El mejor modelo para la predicción de proteína cruda fue PLS-2da. derivada (5,2,2) con SMOTE y presentó $R^2$ de 0.965, 0.817 y 0.793 para la calibración, la validación cruzada y la predicción respectivamente. En cuanto a los errores asociados a este modelo se obtuvieron valores de 0.108, 0.249 y 0.199 para $RMSE_C$, $RMSE_{CV}$ y $RMSE_P$ respectivamente, además un RPD=1.86 y un REP= 1.45.

El mejor modelo para estimar el contenido de cafeína en café verde especial fue PLS-2da. derivada (5,2,2) con SMOTE y presentó $R^2$ de 0.939, 0.865 y 0.822 para la calibración, la validación cruzada y la predicción respectivamente. En cuanto a los errores asociados a este modelo se obtuvieron valores de 0.038, 0.057 y 0.057 para $RMSE_C$, $RMSE_{CV}$ y $RMSE_P$ respectivamente, además un RPD=2.35 y un REP= 5.42.

Los resultados de este estudio confirman que la espectroscopía NIR acoplada a herramientas quimiométricas proveen un método rápido para determinar componentes químicos en café verde especial

# CAPÍTULO 1.

# ASPECTOS GENERALES

## 1.1 Antecedentes del problema

El café es la bebida más consumida alrededor del mundo, después del agua (Barbin *et al*., 2014; Zhu *et al*., 2021) y es el segundo producto básico, más comercializado a nivel mundial (Núñez *et al*., 2020).

Las dos variedades de café más consumidas en el mundo son la variedad Robusta (*Coffea canephora*) y Arábica (*Coffea arabica*) siendo esta última la más apreciada por sus atributos sensoriales (Núñez *et al*., 2020).

El sabor y el aroma del café que se produce durante el tostado está estrechamente relacionado con la composición química del grano de café verde (Liu *et al*., 2019; Zhu *et al*., 2021) por lo que la determinación de estos componentes parece clave en la predicción de la calidad de la bebida.

En este sentido, los atributos sensoriales del café como fragancia/aroma, cuerpo, acidez, balance entre otros, son evaluados en eventos de catación, nacionales e internacionales, en donde los jueces aplican protocolos establecidos por la Asociación de Cafés Especiales de América (SCAA, por sus siglas en inglés)(Barbosa *et al*., 2019).

Por otra parte, en la última década la espectroscopía de infrarrojo cercano (NIRS, por sus siglas en inglés) ha sido una herramienta analítica muy utilizada para la determinación simultánea de diferentes componentes químicos en el café verde y tostado (Barbin *et al*., 2014).

La mayoría de las investigaciones basadas en esta técnica se han centrado en desarrollar una alternativa a los métodos de referencia para la predicción de los atributos del café (Barbin *et al*., 2014), reportándose el uso de esta técnica junto a las herramientas

quimiométricas, para la determinación del origen geográfico de granos de café verde (Giraudo *et al*., 2019), determinación de lípidos y proteínas en café verde (Zhu *et al*., 2021), determinación de pH y acidez en granos de café verde (Araújo *et al*., 2020), así como muchas otras aplicaciones.

En Panamá, se ha reportado la implementación de un modelo matemático para predecir el puntaje que indica la calidad sensorial del café, utilizando 11 parámetros fisicoquímicos de 34 muestras de café tostado de la variedad Arábica (Geisha y Pacamara) (Vega *et al*., 2021); sin embargo no se han encontrado otros estudios a nivel nacional aplicando la espectroscopía NIR para predecir las propiedades químicas del café producido en Panamá.

### 1.2 Situación actual del problema

A nivel internacional el uso de la espectroscopía acoplada a quimiometría para el análisis de alimentos es de uso rutinario, siendo una de las técnicas más utilizadas en la industria alimentaria. En Panamá, los estudios desarrollados en esta temática involucran investigaciones como la evaluación de la calidad en un sistema de clasificación de sandías (Rangel *et al*., 2021), así como para clasificación de diferentes variedades de arroz producido localmente (Sánchez-Galán *et al*., 2021).

En cuanto al café, aunque existen trabajos previos en donde se utilizan herramientas quimiometricas para predecir la calidad del café especial producido en la provincia de Chiriquí (Vega *et al*., 2021), no se han encontrado estudios o aplicaciones de la espectroscopía NIR acoplada a quimiometría para el monitoreo del café de especialidad.

### 1.2.1 Planteamiento del problema

Los granos de café verde contienen una gran cantidad de compuestos químicos, que reaccionan e interactúan entre sí en todas las etapas del tostado del café, lo que da como resultado productos finales con diferentes características sensoriales (Barbin *et al*., 2014).

En este sentido la industria del café tiene una necesidad creciente de implementar métodos de referencia consistentes, robustos, precisos y rápidos, para determinar diferentes compuestos químicos asociados con la calidad del producto final y así ser más competitivos en el mercado mundial.

Analizar todos estos compuestos utilizando metodologías convencionales tomaría mucho tiempo y sería muy costoso, por lo que la utilización de la espectroscopía NIR junto con las herramientas quimiométricas, representaría una alternativa viable a la demanda de la industria del café.

## 1.3 Hipótesis

La espectroscopía NIR y las herramientas quimiométricas pueden ser utilizadas como una técnica analítica alternativa a metodologías tradicionales para estimar el contenido de proteína, grasa y cafeína en el café verde.

## 1.4 Objetivos

### 1.4.1 General

Aplicar la espectroscopía NIR y herramientas quimiométricas para la determinación rápida del contenido de cafeína, proteína y grasa en café verde especial, producido en la provincia de Chiriquí, Panamá.

### 1.4.2 Específicos

- Caracterizar fisicoquímicamente (grasa, proteína y cafeína) las muestras de café verde especial utilizando métodos de química húmeda, HPLC y la técnica de espectroscopia infrarroja.
- Aplicar métodos quimiométricos para generar modelos que permitan determinar contenido de grasa, proteína y cafeína en el café verde especial.

## 1.5 Delimitación, Alcance o Cobertura

Esta investigación fue realizada en la provincia de Chiriquí, específicamente en el Centro de Investigación en Recursos Naturales (CIRN) de la Universidad Autónoma de Chiriquí (UNACHI) e incluyó un total de 98 muestras de café verde de la variedad *C. arabica*, provenientes de diferentes fincas, ubicadas en los distritos de Boquete, Renacimiento y Tierras Altas.

A través de este estudio se logró generar modelos de predicción basados en espectroscopía NIR y quimiometría, para determinar el contenido de cafeína, proteína y grasa en café verde especial.

## 1.6 Restricciones y/o Limitaciones

Para generar modelos de predicción robustos, el número de muestras es un factor relevante. Estas deben ser analizadas previamente por metodologías analíticas oficiales y así obtener los valores de referencia que serán utilizados para generar los modelos. Algunas de estas metodologías analíticas son costosas, representando así una importante limitación en la investigación.

# CAPÍTULO 2.

# MARCO TEÓRICO

## 2.1 Café

El café pertenece al género *Coffea* de la familia de las *Rubiaceae.* Las especies comerciales más relevantes son *C. arabica* (Arábica) y *C. canephora* (Robusta), representando aproximadamente un 64% y 35% de la producción mundial de café respectivamente (Tsegay *et al*., 2020). *C. arabica* es cultivada a alturas mayores a los 1000msnm, en regiones tropicales y subtropicales donde el clima es más fresco, mientras que la especie robusta se cultiva por debajo de los 800 msnm y es resistente a las altas temperaturas. Brasil es el mayor productor y exportador de café a nivel mundial (Franca *et al*., 2005; Haile y Kang, 2019).

El mejoramiento del café para resistir a enfermedades como la roya del café en combinación con producción, vigor y calidad comenzó a principios de la década de 1920 en la India y tuvo mayor auge en la mitad del siglo 20, en donde países como India, Brazil, Costa Rica, Colombia, Kenia, Tanzania, Etiopía, Angola y Portugal, aplicaron técnicas genéticas, fitopatología, fisiología de cultivos y agronomía, dando paso a nuevas variedades o cultivares (van der Vossen *et al*., 2015).

La Figura 1 muestra algunas de las variedades o cultivares de mayor importancia comercial incluyendo la variedad *Coffea Charrieriana* descubierta en África Central, que es la primera variedad naturalmente libre de cafeína (Torres Castillo *et al*., 2020).

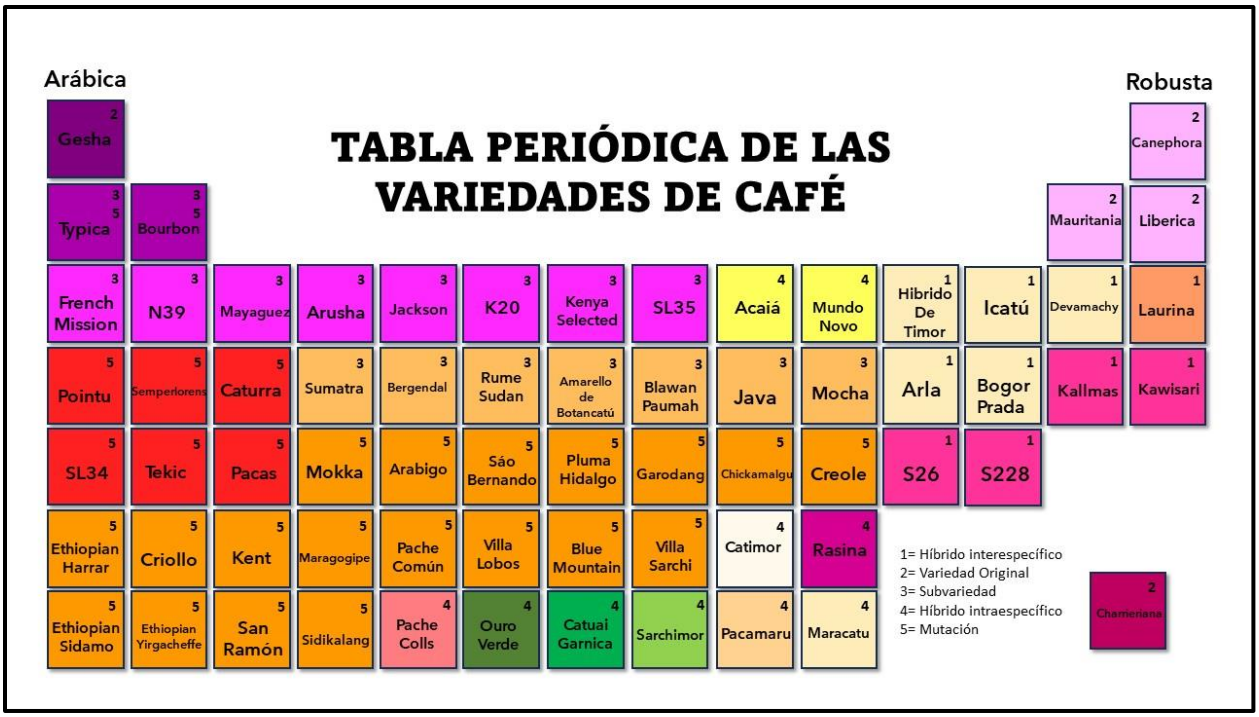



**Figura 1. Tabla periódica de las variedades de café. Fuente: The Coffee CoMission, 2015.**

La variedad *C. arabica* es apreciada por sus grandes atributos sensoriales (Núñez *et al*., 2020), incluso siendo cultivada bajo el concepto de café de especialidad es países como Panamá.

## 2.2 Café de Especialidad

En Panamá el cultivo de café se realiza en las provincias Coclé, Colón, Chiriquí, Panamá, Panamá Oeste, Veraguas y la Comarca Ngäbe Buglé. Para el año agrícola 2022/23, la producción nacional fue de 127600 quintales de café pilados, siendo la provincia de Chiriquí la mayor productora con aproximadamente 89600 quintales pilados, lo que representa el 70.22% de la producción total, según cifras preliminares del INEC (Instituto Nacional de Estadística y Censo, 2023); además, se dedica también a la producción y exportación de café de especialidad.

Un café de especialidad se produce cuando se estandariza toda la cadena de producción, desde su cultivo en el campo hasta el consumidor final, cumpliendo con estándares de calidad para resaltar las características de la bebida (Laukaleja y Kruma, 2018). En este sentido, el instrumento más utilizado para examinar la calidad de un café es el análisis sensorial, este se realiza utilizando los protocolos y catadores certificados por la Asociación de Cafés Especiales de América (SCAA) que califican en una escala de 0 a 10 atributos como fragancia (olor seco), aroma (olor de infusión), sabor, regusto (astringencia), acidez, cuerpo, equilibrio, dulzor, limpieza en taza, uniformidad y defectos y cuando este puntaje sumado es superior a 80 puntos se considera como café especial (Gumecindo-Alejo *et al*., 2021).

La Asociación de Cafés Especiales de Panamá (SCAP, por sus siglas en Inglés) realiza un evento anual de catación nacional para evaluar los atributos del café de especialidad siguiendo los protocolos desarrollados y estandarizados por la industria del café de especialidad a nivel global (Vega *et al*., 2021). Este evento se conoce como el Best of Panama y en los últimos 10 años se ha registrado un incremento en el precio del café de especialidad ofertado a mercados internacionales como se muestra en la Figura 2, llegando a alcanzar precios de hasta 10000.00 dólares por kilogramo de café.

La calidad en taza del café está relacionada con varios factores como variedad de café, ubicación geográfica, altitud de la finca, práctica de cosecha, manejo postcosecha (tipo de proceso), almacenamiento, tostado (Abubakar *et al*., 2019), suelo y variaciones de temperatura (Tsegay *et al*., 2020).

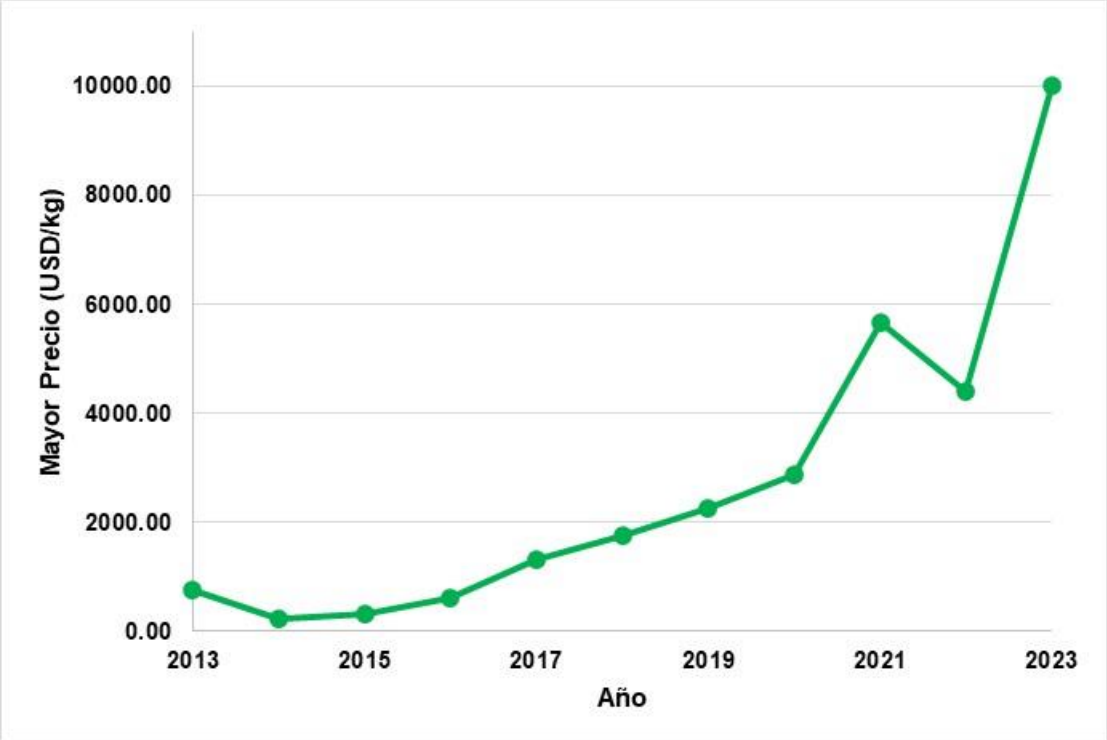



**Figura 2. Precio del café de especialidad de Panamá desde 2013-2023 (SCAP, 2023).**

Durante el tostado, se producen compuestos químicos que le dan las características sensoriales finales a la bebida; la cantidad y tipo de estos compuestos, dependen de la composición química en el café verde (Abubakar *et al*., 2019; Zhu, Long, Chen, *et al*., 2021).

## 2.3 Composición química del café verde

La composición química del grano de café verde es compleja y los carbohidratos representan la mayor parte de los componentes, como se observa en la Tabla 1 (Saud y Salamatullah, 2021).

Los carbohidratos, las proteínas, los aminoácidos y azúcares son precursores importantes involucrados en las reacciones de Maillard durante el proceso del tostado, influenciando el aroma, color (caramelización) y sabor del café (Mahmud *et al*., 2020; Zhu, Long, Ma, *et al*., 2021).

**Tabla 1. Composición química del café verde.**

| Componentes | Concentración | Referencia |
|---|---|---|
| Carbohidratos | 50-60% | (Saud y Salamatullah, 2021; Yeager *et al*., 2023) |
| Lípidos | 8-20% | (Dong *et al*., 2021; Pimpley *et al*., 2020; Yeager *et al*., 2023) |
| Proteínas | 9-15% | (Pimpley *et al*., 2020; Saud y Salamatullah, 2021; Yeager *et al*., 2023) |
| Minerales | 3-5% | (Pimpley *et al*., 2020; Yeager *et al*., 2023) |
| Cafeína | 1-4% | (Mahmud *et al*., 2020; Saud y Salamatullah, 2021; Yeager *et al*., 2023) |
| Ácidos Clorogénicos | 3.5-12% | (Mahmud *et al*., 2020; Pimpley *et al*., 2020) |
| Sacarosa | 5-9% | (Saud y Salamatullah, 2021; Yeager *et al*., 2023) |

Los compuestos fenólicos más abundantes en el café verde son los ácidos clorogénicos, que son ésteres de ácido *trans*-cinámico y ácido quínico (Pimpley *et al*., 2020), se producen naturalmente y se acumulan en el grano a medida que el fruto del café madura; estos se dividen en ácidos cafeoilquínicos (CQA), ácidos dicafeoilquínicos (diCQA) y ácidos feruloilquínicos (FQA) (Yeager *et al*., 2023).

La cafeína es la sustancia más estudiada en el café (dos Santos y Boffo, 2021) y junto a la trigonelina y los ácidos clorogénicos son responsables del amargor (Mahmud *et al*., 2020) y precursores del aroma en el café (Sualeh *et al*., 2020).

Los lípidos en los granos de café verde están compuestos de triacilgliceroles, tocoferoles, esteroles y diterpenos, estos últimos constituyen hasta el 20% del total de

lípidos (Pimpley *et al*., 2020) y durante el proceso de tostado participan en reacciones de descomposición y oxidación contribuyendo al sabor y aroma del café (Zhu, *et al*., 2021).

Para evaluar la calidad y características químicas del café, se utilizan tradicionalmente los métodos de química húmeda, pero estos toman mucho tiempo, son costosos y destructivos. Por lo tanto, investigaciones recientes han demostrado que la espectroscopía infrarroja cercana (NIR) e infrarroja media (MIR), son técnicas útiles en el análisis de café (Wójcicki, 2022).

## 2.4 Espectroscopía NIR

La espectroscopía de infrarrojo cercano es una técnica utilizada para caracterizar la composición química de materiales complejos como alimentos y su rango espectral va de 12000-4000 $cm^{-1}$ .Como técnica analítica, ha experimentado una evolución constante, impulsada principalmente por mejoras en los instrumentos y herramientas matemáticas empleadas para extraer y procesar información analítica de los datos espectrales, como el método estadístico de mínimos cuadrados parciales (PLS) (Zhu *et al*., 2021).

Según Pasquini (2018), la espectroscopía NIR moderna se basa en tres pilares: fundamentos de la espectroscopía (teoría), instrumentación (medición) y quimiometría (análisis).

En cuanto al fundamento de la técnica, la interacción de la radiación NIR con la muestra se puede dar en diferentes modos como transmitancia, reflectancia difusa, transmitancia difusa, interactancia y transflectancia; siendo la reflectancia difusa el modo que más se emplea (Okazaki, 2012).

En el modo de reflectancia difusa el espectro resultante muestra la absorción y dispersión de la radiación al interaccionar con los componentes de la muestra, siendo ampliamente utilizado para el análisis de muestras sólidas de manera no destructiva como frutas, carnes, cereales, vino, cervezas, huevos, productos farmacéuticos, tejidos, suelos, entre otros (Munyendo *et al*., 2022; Pasquini, 2018).

En contraste con la espectroscopía de infrarrojo medio y lejano, la interacción de la radiación electromagnética NIR se produce excitando sobretonos, combinaciones y resonancias de modos vibratorios fundamentales; estos modos están asociados a modos altamente anarmónicos de grupos funcionales que contienen un átomo relativamente pesado (C, O, N y S) unido a un hidrógeno (Pasquini, 2003; Zareef *et al*., 2020).

Los coeficientes de absorción en el rango espectral NIR son 10-1000 veces menores que los de la región del infrarrojo medio, donde las transiciones fundamentales son excitadas. Esto reduce la sensibilidad y selectividad de esta técnica analítica, que produce una respuesta instrumental universal en muestras de composición orgánica. Sin embargo, la sensibilidad se puede compensar con fuentes de radiación más brillantes y mejores detectores, mientras que, la baja selectividad puede superarse mediante una calibración adecuada y el uso de análisis de datos multivariados y regresión (Pasquini, 2018).

En cuanto a la instrumentación, esta evolucionó de simples equipos dispersivos basados en filtros, hasta el espectrofotómetro basado en interferometría, mejorando su robustez, durabilidad y reduciendo su tamaño (Pasquini, 2018). Las partes principales

son un divisor de haz, un detector óptico, un detector de muestras, una fuente de luz y un sistema para el procesamiento de datos (Zareef *et al*., 2020).

Por otra parte, la información obtenida de los espectros NIR es muy compleja, debido a que la mayoría de los componentes de la matriz contribuyen al espectro, sumado a los efectos producidos por propiedades físicas de la muestra como el tamaño de partícula, que imparten efectos significativos y no deseados; por lo que es necesario que el análisis de datos considere los cambios en la totalidad o en una porción significativa del espectro (Pasquini, 2018; Zareef *et al*., 2020).

La espectroscopía NIR es una fuente significativa de información analítica, que se puede aplicar a una gran variedad de muestras, mientras que los algoritmos quimiométricos se desarrollan para superar la falta de selectividad de la técnica espectral, cuando se aplica a fines cuantitativos (Pasquini, 2018).

En el análisis de café, la espectroscopía NIR acoplada a quimiometría, se ha utilizado para determinar adulteración (Manuel *et al*., 2022), clasificación (Okubo y Kurata, 2019) y discriminación entre variedades (Buratti *et al*., 2015), composición química (Araújo *et al*., 2020; Zhu *et al*., 2021), predicción sensorial (Baqueta *et al*., 2019) y grado de tostado (Bertone *et al*., 2016).

La Tabla 2 muestra algunas de las investigaciones recientes en donde se aplica la espectroscopía NIR para el análisis de café verde.

**Tabla 2. Uso de la espectroscopía NIR en diferentes aplicaciones para el análisis de café verde.**

| Aplicación | Técnica espectroscópica | Análisis Multivariado | Referencias |
|---|---|---|---|
| Determinación de lípidos y proteína en granos de café verde de diferentes orígenes | NIR | PLS | (Zhu, *et al*., 2021) |
| Determinación de pH y acidez en granos de café verde | FT-NIR | PLS | (Araújo *et al*., 2020) |
| Determinación de la autenticidad y origen geográfico de café verde | NIR | PLS-DA | (Nguyen Minh *et al*., 2022) |
| | FT-MIR | PLS-DA | (Mendes *et al*., 2022) |
| | HSI-NIR | PLS-DA, SVM, RBF-SVM, Random Forest | (Sim *et al*., 2024) |
| Discriminación entre especies de café verde | UV-Vis/NIR | LDA | (Sim *et al*., 2024) |
| Determinación de cafeína en granos de café verde Arábica | NIR | PLS | (Ayu *et al*., 2020) |
| Clasificación de granos de café verde Arábica | FT-NIR | PCA | (Siregar *et al*., 2020) |
| Predicción del contenido de agua en granos de café verde Lintong | FT-NIR | PLS | (Sahfitri *et al*., 2020) |

NIR: infrarrojo cercano; MIR: Infrarrojo medio; HSI-NIR: imágenes hiperespectrales en infrarrojo cercano; PLS: mínimos cuadrados parciales; PLS-DA: mínimos cuadrados parciales con análisis discriminante; SVM: máquina de soporte de vectores; RBF-SVM: máquina de soporte de vectores con función de base radial de kernel; LDA: análisis discriminate lineal; PCA: análisis de componentes principales.

## 2.5 Quimiometría aplicada a espectroscopía NIR

El término “quimiometría” fue introducido en 1970, con el objetivo de referirse a la extracción, uso e interpretación de datos químicos presentes en grandes conjuntos de datos (Pasquini, 2018) y se ha convertido en un área destacada de la química analítica, que permite desarrollar estrategias multivariadas para evaluar e interpretar los datos químicos (Cardoso Santos *et al*., 2019).

En este sentido, la quimiometría acoplada con NIR permite construir modelos multivariados para predecir propiedades cualitativas o cuantitativas en diversas matrices.

El primer paso en la construcción de cualquier modelo multivariado debe ser, visualizar las señales espectrales de las muestras para detectar diferencias claras y esto se puede lograr utilizando el análisis de componentes principales (PCA, por sus siglas en inglés), que es una técnica para reducir la dimensionalidad de un conjunto de datos, aumentando la interpretabilidad y minimizando la pérdida de información. Esto se logra creando nuevas variables no correlacionadas llamadas componentes principales (Jolliffe y Cadima, 2016).

### 2.5.1 Eliminación de datos atípicos

El PCA es aplicado a los datos previamente preprocesados con centrado, escalado, etc (Rodionova *et al*., 2021). Aplicar centrado a los datos coloca el origen del nuevo sistema de coordenadas en el centro de la nube de datos, esto simplifica la interpretación de las puntuaciones y permite hacer inferencias sobre los intervalos de confianza para las puntuaciones del componente principal, por lo que se recomienda centrar los datos antes de aplicar el PCA. Por otro lado, si las variables se han medido en diferentes unidades (variables con mayor varianza tendrán mayor impacto en el PCA), es importante aplicar el escalado automático (estandarización de variables), ya que el PCA es sensible a la escala de las variables originales. Sin embargo, una desventaja es que las variables con una relación señal-ruido baja contribuirán en la misma cantidad a la varianza total que las variables con una relación señal-ruido alta, por lo que no se recomienda el ajuste de escala automático para variables espectrales que hayan sido medidas con el mismo instrumento y se expresen en las mismas unidades (Saeys *et al*., 2019).

El PCA también es utilizado para detectar valores atípicos en la data espectral. Los valores atípicos pueden dar lugar a resultados incorrectos en relación con el análisis multivariado y en consecuencia disminuir la calidad del modelo multivariado (Zhu *et al*., 2021).

### 2.5.2 Selección de datos de calibración y validación

Para obtener un modelo de regresión multivariado robusto, es necesario que sea entrenado con un conjunto de datos de calibración que sea representativo de la población de la muestra; a este conjunto de datos se le conoce como “datos de calibración”. El muestreo utilizado para obtener las muestras de calibración se debe diseñar con el objetivo de tener una representación adecuada de las fuentes de variación conocidas. Por otra parte, como se desconoce la verdadera dimensionalidad de la relación entre X y Y en un modelo de regresión multivariado, este debe ser evaluado utilizando un conjunto de datos que no fue utilizado para construir el modelo. A este conjunto de datos se les conoce como “datos de validación”. Idealmente los datos de calibración deben ser adquiridos de manera independiente de los datos de calibración y deben contener también los fuentes de variación que enfrentará el modelo durante su uso (Saeys *et al*., 2019).

En la práctica, una de las limitantes es la cantidad de muestras que pueden ser recolectadas y analizadas por métodos de referencia. Cuando se dispone de un número grande de muestras (>150 muestras), se recomienda utilizar 2/3 de los datos para la calibración y 1/3 para la validación. La selección puede realizarse de forma secuencial o aleatoria, lo cual puede representar un riesgo, ya que se puede seleccionar un conjunto de validación con el que se obtengan resultados sobreajustados, o, por el contrario, un

conjunto de datos que genere resultados pobres. Por esta razón se pueden utilizar algoritmos como el de Kennard-Stone o Duplex cuyo objetivo es dividir los datos en dos conjuntos que abarcan el espacio multidimensional igualmente bien (Saeys *et al*., 2019).

### 2.5.3 Desbalance de clases en el conjunto de datos

Cuando la distribución de clases en un conjunto de datos está sesgada, es decir, la proporción de muestras que pertenecen a cada clase en el conjunto de datos es inconsistente, tenemos un problema de desequilibrio de clases. Estos conjuntos de datos desequilibrados son representativos de los problemas de clasificación en el mundo real, como la detección de fraude el diagnóstico médico y el reconocimiento de anomalías, entre otros (Chakravarthy *et al*., 2019).

Con el fin de evitar estos problemas de desequilibrio se utilizan algoritmos que permiten equilibrar las clases, siendo los más comunes el sobremuestreo aleatorio (ROSE, por sus siglas en inglés) y técnica de sobremuestreo de minorías sintéticas (SMOTE, por sus siglas en inglés) (Selamat *et al*., 2022).

La técnica ROSE funciona submuestreando primero la clase mayoritaria y luego sobremuestreando la clase minoritaria y finalmente generando muestras sintéticas en las proximidades de estos datos recién muestreados. Estos tres pasos se repiten hasta lograr el equilibrio en el conjunto de datos (Chakravarthy *et al*., 2019).

Por otro lado, la técnica SMOTE sintetiza nuevas muestras entre dos muestras minoritarias mediante interpolación lineal, evitando el problema de sobreajuste causado por el sobremuestreo aleatorio. SMOTE se considera un estándar “de facto” en el marco del aprendizaje a partir de datos desequilibrados y ha contribuido significativamente a

nuevos paradigmas de aprendizaje supervisado, incluida la clasificación multietiqueta (Sun *et al*., 2021).

### 2.5.4 Preprocesamiento de la data espectral

El rendimiento de los modelos puede variar en función del preprocesamiento espectral utilizado ya que, los diferentes tratamientos eliminan fenómenos físicos que no están relacionados con la información química. Es una buena práctica probar varios métodos de preprocesamiento para comprender cuál ofrece el mejor rendimiento del modelo de regresión multivariante (Caporaso *et al*., 2021).

Las técnicas de preprocesamiento más comunes pueden ser divididas en dos categorías: métodos de corrección de la dispersión y derivadas espectrales. En el primer grupo se incluyen corrección del efecto multiplicativo de la dispersión (MSC), variable normal estándar (SNV) entre otros. Mientras que, para el grupo de derivación uno de los métodos más utilizados es filtros derivados polinomiales de Savitzky-Golay (SG), que aplica el suavizado del espectro antes de aplicar la derivada para disminuir el efecto perjudicial sobre la relación señal ruido que tendrían las derivadas (Rinnan *et al*., 2009).

**El filtro de Savitzky-Golay (SG),** tiene como objetivo suavizar una señal sin corromper su contenido informativo. El método se basa en un enfoque de ventana deslizante en el que se ajusta localmente una curva polinomial a los datos. El método se puede implementar eficientemente como un método de convolución. Además de elegir diferentes grados polinomiales, se pueden aproximar las derivadas de la señal lo cual es útil para muchos tipos de datos espectrales ruidosos, especialmente los de FTIR (Helin *et al*., 2022).

**Centrado medio y Variable normal estándar (SNV),** el centrado medio es el enfoque más simple para eliminar el ruido aditivo de todos los espectros NIR calculando la media y restándola. El SNV aplica el centrado medio eliminando el ruido aditivo y multiplicativo de todos los espectros NIR calculando la media y la varianza, seguido de restar la media y dividiendo por la varianza. Tanto el centrado medio como el SNV suponen que el ruido de sesgo aditivo puede aproximare mediante la media de los datos y el ruido de sesgo multiplicativo puede aproximarse mediante la varianza de los datos (Zhang *et al*., 2022).

**El efecto multiplicativo de la dispersión (MSC),** intenta eliminar los efectos multiplicativos y el desplazamiento constante. Los coeficientes de corrección de cada espectro se calculan haciendo una regresión sobre un espectro de muestra ideal (un espectro representativo del grupo de espectros considerado en un entorno completamente libre de ruido). En otras palabras, cada espectro se ajusta al espectro de muestra ideal (generalmente el espectro promedio) lo más cerca posible utilizando mínimos cuadrados (Gautam *et al*., 2015). SNV elimina eficazmente las interferencias por efecto multiplicativo de dispersión y tamaño de partículas, y los resultados son similares a los obtenidos por MSC (Caporaso *et al*., 2021)

**Primera y segunda derivada,** se utiliza comúnmente la primera derivada para minimizar las desviaciones de la línea base dentro de un conjunto de espectros. Sin embargo, el término de la pendiente se convierte en un término constante en la primera derivada. Por lo tanto, es una práctica común calcular la segunda derivada para eliminar tanto el desplazamiento como la pendiente.

**Corrección de la señal Ortogonal (OSC),** es un método desarrollado para reducir los efectos de dispersión de la luz y las interferencias más generales, eliminando solo los efectos que tienen correlación cero con el valor de referencia Y. La idea de es que toda la información relaciona con Y permanezca en lugar de eliminarse (Pizarro *et al*., 2004).

### 2.5.5 Métodos de regresión multivariado

De los métodos de regresión multivariado más utilizados se encuentra mínimos cuadrados parciales (PLS) y regresión de componentes principales (PCR).

**Mínimos cuadrados parciales (PLS),** es una técnica que combina características básicas del análisis de componentes principales (PCA) y la regresión lineal múltiple (MLR). Es muy útil cuando el número de observaciones está muy por debajo del número de características (Çataltas y Tütüncü, 2021).

**Regresión de componentes principales (PCR),** es un método muy utilizado cuando se tienen problemas de colinealidad. Este método está formado por el análisis de componentes principales (PCA) y por regresión de mínimos cuadrados (LSR). En primer lugar, se extraen los componentes principales de los datos espectrales con PCA y los que muestran la mayor variación en los datos se seleccionan y el modelo se crea utilizando estos componentes principales (Çataltas y Tütüncü, 2021).

### 2.5.6 Métricas de evaluación del modelo

La capacidad predictiva de un modelo de regresión multivariado se evalúa en función de parámetros de mérito como coeficiente de determinación, $RMSE_{CV}$, $RMSE_{p}$, RPD, entre otros que se describen en la Tabla 3 (Araújo *et al*., 2020).

**Tabla 3. Parámetros de mérito evaluados en los modelos de regresión multivariado**

| Métrica | Símbolo | Fórmula | Descripción | Referencia |
|---|---|---|---|---|
| Raíz del error cuadrático medio de validación cruzada | $RMSE_{CV}$ | $\sqrt{\frac{1}{N_{cal}}\sum_{i=1}^{N_{cal}}(y_{cal,i}-\hat{y}_{cal,i})^2}$ | $RMSE_{CV}$ es igual a la raíz del cuadrado medio de la diferencia entre los valores reales y los obtenidos con el conjunto de validación cruzada. Cuanto más se acerque el valor a 0, es mejor. | (Çataltas y Tütüncü, 2021) |
| Raíz del error cuadrático medio de predicción | $RMSE_P$ | $\sqrt{\frac{1}{N_{pred}}\sum_{i=1}^{N_{pred}}(y_{pred,i}-\hat{y}_{pred,i})^2}$ | $RMSE_P$ es igual a la raíz del cuadrado medio de la diferencia entre los valores reales y los obtenidos con el conjunto de predicción. Cuanto más se acerque el valor a 0, es mejor. | |
| La relación entre el rendimiento y la desviación | RPD | $\frac{\sigma}{SEP}$ | El RPD se obtiene dividiendo la desviación estándar del set de predicción entre el error de predicción (SEP o RMSEP). | |
| Coeficiente de determinación | $R^2$ | $\frac{Explained\ variation}{Total\ variation}$ | $R^2$ corresponde a cómo las variables independientes explican gran parte de la variación de una variable dependiente. Puede tomar un valor entre 0 y 1, donde 1 significa que toda la variable de salida se explica completamente por la variable de entrada. | |
| Error relativo de predicción | REP | $\mathrm{REP}=\frac{\mathrm{SEP}}{\bar{y}}\times 100$ | REP está relacionado con la precisión del modelo y se calcula por la relación porcentual entre $RMSE_P$ y la media de los valores reales del conjunto de predicción. | (Dupuy *et al*., 2010) |

Por otro lado, para evaluar qué regiones del espectro infrarrojo sustentan el mejor modelo se utiliza la importancia de la variable dependiente en la proyección (VIP), donde puntuaciones mayores a uno indican que esa región tiene un aporte significativo al modelo (Hashimoto et al., 2018).Todos estos parámetros en conjunto permiten decidir qué modelo tiene una mejor capacidad predictiva.

# CAPÍTULO 3.

# METODOLOGÍA

## 3.1 Tipo de Investigación

### 3.1.1 Definición

La presente investigación posee un enfoque cuantitativo de corte transversal, pues la misma busca estimar el contenido de grasa cruda, proteína cruda y cafeína en café verde especial de la variedad *C. arabica*, utilizando espectroscopía NIR y quimiometría, sin tomar en cuenta estacionalidad o comparación entre años de cosecha.

### 3.1.2 Justificación

La espectroscopia NIR es ampliamente utilizada en la industria alimentaria por su rapidez, simplicidad y seguridad, así como por su capacidad para medir diferentes componentes, sin la necesidad de una extensa preparación de la muestra. Esta técnica permite trabajar, con poco o ningún reactivo químico, razón por la cual es una técnica de bajo costo, amigable al medio ambiente, por lo que es adecuada para su implementación rutinaria en el análisis de alimentos como el café.

En café se han desarrollado modelos para estimar la concentración de componentes como cafeína, humedad, contenido de lípidos, ceniza, teobromina, teofilina entre otros; también para clasificar o discriminar el café de acuerdo con la variedad u origen geográfico, para identificar adulteraciones, predecir la parte sensorial y grado de tostado (Barbin *et al*., 2014).

La implementación de la espectroscopía NIR en el análisis de café permite monitorear diferentes componentes químicos involucrados en la calidad del producto final y a su vez el estudio serviría de base para otras investigaciones acerca del café.

### 3.2 Sujetos, entorno o fuente de información

Las muestras de café verde de la variedad *C. arabica* fueron provistas por la Asociación de Café Especial de Panamá (SCAP, por sus siglas en inglés) y correspondían a los años de cosecha de 2018 y 2019. Estas muestras provenían de diferentes fincas ubicadas en los distritos de Boquete, Renacimiento y Tierras Altas, de la provincia de Chiriquí.

### 3.3 Variables o Categorías de la Investigación

Las variables consideradas en esta investigación comprenden variables cuantitativas como grasa cruda (%), proteína cruda (%) y cafeína (%) y espectro infrarrojo (nm), obtenidas de las muestras de café verde especial.

#### 3.3.1 Definición Conceptual

La grasa cruda consiste principalmente en grasas neutras y lípidos no polares que representan más del 90 % de los lípidos que componen los alimentos y piensos de animales (Brown y Mueller-Harvey, 1999).

La proteína cruda es la determinación del contenido de nitrógeno que componen los alimentos o piensos de animales y que provienen no solo de las proteínas sino de otras fuentes. El contenido de proteína cruda es calculado por la multiplicación del porcentaje de nitrógeno obtenido por un factor de 6.25, ya que la mayoría de las proteínas contienen 16% de nitrógeno (AOAC, 2010).

La cafeína es un alcaloide cuyo nombre común es 1,3,7 trimetilxantina y es encontrado en cantidades diferentes en granos, hojas y frutos de más de 60 plantas como el café el cacao, hojas de té entre otros (Heckman *et al*., 2010).

El espectro infrarrojo surge como consecuencia de la absorción de radiación electromagnética en frecuencias que se correlacionan con los modos de vibración de conjuntos específicos de enlaces químicos dentro de una molécula (Coates, 2000).

### 3.3.2 Definición Operacional

En lo que respecta a la definición operacional de las variables, el contenido de grasa cruda se determinó por la extracción con solventes no polares como el éter de petróleo utilizando un analizador de grasa de la marca ANKOM. La proteína cruda se determinó utilizando un digestor de proteína y destilador de la marca FOSS, mientras que el contenido de cafeína se determinó utilizando un cromatógrafo líquido de alta eficiencia. Los espectros de infrarrojo cercano se adquirieron en un espectrofotómetro infrarrojo cercano de la marca FOSS.

## 3.4 Población

El estudio fue realizado en Chiriquí, por ser la provincia de mayor producción de café a nivel nacional. Las muestras provenían de los distritos de Boquete, Renacimiento y Tierras Altas, que se dedican al cultivo de café especial variedad *C. arabica*.

### 3.4.1 Muestra

Un total de 98 muestras de café verde especial de la variedad *C. arabica,* fueron provistas por el SCAP; estas fueron agrupadas en dos clases: 72 pertenecían a la clase Geisha y 26 a la clase Tradicional. Estas muestras corresponden a los lotes de café participantes en el evento de catación internacional Best of Panama de los años 2018 y 2019.

### 3.4.2 Tipo de muestreo

Se realizó un muestreo no probabilístico por conveniencia, ya que se consideraron para el estudio las muestras de café verde especial perteneciente a los lotes participantes en el evento de catación internacional organizado por el SCAP en los años 2018 y 2019.

## 3.5 Métodos e instrumentos de investigación

### 3.5.1 Preparación de la muestra

Las muestras de café verde fueron liofilizadas (Millrock Technologies, Estados Unidos) a -40 °C y 500 mTorr, luego molidas y tamizadas utilizando un mesh #20 (Wiley Mill Thomas Scientific, Estados Unidos). Las muestras molidas fueron guardadas en bolsas plásticas con cierre hermético y almacenadas a -20°C hasta el análisis (ver Anexo I, Figura A1).

### 3.5.2 Determinación de Humedad

La determinación de humedad se realizó por triplicado utilizando el método oficial AOAC (934.01), en donde se utilizó 1 gramo de la muestra de café verde molido y se colocó en horno a 100 ± 5 °C hasta peso constante (AOAC, 2010).

### 3.5.3 Determinación del contenido de grasa cruda

Para la determinación del contenido de grasa cruda, se utilizó el método AOCS Am 5-04. Se pesó 1 gramo de la muestra de café verde en la bolsa filtrante (XT4) previamente tarada y se procedió a sellar con una selladora de pulso. Luego se procedió a secar las bolsas con la muestra en un horno (Quincy Lab Inc, Estados Unidos) a 102 ± 2 °C por 3 horas. Se dejó enfriar hasta temperatura ambiente y se anotó el peso. El proceso de extracción se realizó en un analizador de grasa XT10 (ANKOM Technology,

Estados Unidos), a una temperatura de 90°C por 60 minutos utilizando éter de petróleo. Luego de la extracción las muestras se llevaron al horno por 30 minutos a una temperatura de 102 ± 2 °C, se dejó enfriar hasta temperatura ambiente y se anotó el peso (AOCS, 2005). El análisis fue realizado por triplicado y el contenido de grasa cruda expresado en porcentaje (%) y se calculó utilizando la siguiente fórmula:

$$\% \text{ Grasa Cruda} = \frac{(W2 - W3)}{W1} * 100\ \%$$

Donde:

W1=Peso de la muestra (g).

W2=Peso de la muestra seca más la bolsa filtrante (g).

W3=Peso de la muestra más la bolsa filtrante, después de la extracción (g).

### 3.5.4 Determinación del contenido de proteína cruda

El contenido de proteína cruda se determinó por el método Kjeldahl AOAC (2001.11). Se pesó 1 gramo de muestra de café verde molido y se colocaron en los tubos de digestión. Luego se adicionó una pastilla catalizadora y 12 mL de ácido sulfúrico concentrado. Se procedió a digerir la muestra en un digestor FOSS Tecator 20, a una temperatura de 420 °C por 45 minutos, se dejó enfriar y se adicionó 75 mL de agua destilada. Luego se le adicionó 50 mL de hidróxido de sodio al 40% y se destiló por 5 minutos en un destilador FOSS Kjeltec 2100, recolectando en un erlenmeyer que contenía 25 mL de ácido bórico al 4 % con indicador. Se tituló con HCl al 0.1 N hasta obtener un color morado suave y se anotó el volumen de HCl consumido (AOAC, 2010). El análisis fue realizado por triplicado y el contenido de proteína cruda fue expresado en porcentaje y se calculó utilizando la siguiente fórmula:

$$N\ (\%) = \frac{(T\text{-}B)\ x\ (N)\ x\ 14.01}{muestra\ inicial\ (g) x 10}\ x\ 100\%$$

$$Proteína\ cruda\ (\%) = N\ (\%) * F$$

Donde:

$V_a$= Volumen ácido clorhídrico (mL).

$V_b$ = Volumen de titulación blanco (mL).

N = Normalidad del ácido clorhídrico.

14.01= peso atómico del nitrógeno.

F = Factor de conversión 6.25 para alimentos en general.

### 3.5.5 Determinación del contenido de cafeína por HPLC-DAD

El contenido de cafeína de las muestras de café verde se determinó de acuerdo con lo propuesto por Naegele, (2016) con algunas modificaciones. Se pesaron 500 mg de la muestra en un tubo falcón de 50 mL, se adicionó 10 mL de una solución metanol/agua (1:1) y se agitó en un vortex (Scilogex, Estados Unidos) por 2 minutos. Luego se centrifugó por 5 minutos a 6000 rpm. Se tomó una alícuota de 1 mL, se llevó a un matraz volumétrico de 10 mL y se aforó con agua tipo I hasta la marca. Previo al análisis cromatográfico la muestra fue filtrada con filtros de jeringa de nylon con un tamaño de poro de 0.45 µm.

Se utilizó un equipo Agilent 1260 Infinity (Agilent Technologies Inc, Estados Unidos) con desgasificador (G1322A), bomba cuaternaria (G1311C), automuestreador (G7129A), compartimento de columna termostatizado (G1316A) y detector de arreglo de

diodos (G1315D). En la Tabla 4 se muestran las condiciones cromatográficas para la determinación de cafeína.

**Tabla 4.Condiciones cromatográficas para la determinación de cafeína.**

| Parámetro | Valor |
|---|---|
| Solventes | A: Agua (90 %) |
| | B: Acetonitrilo (10 %) |
| Columna: | Zorbax SB-C18, 4.6 mm x 150 mm, 5 µm |
| Temperatura: | 25 °C |
| Flujo: | 1.0 mL/min |
| Condiciones de elución | Isocrático |
| Volumen Inyección: | 20 µL |
| Longitud de onda: | 275 nm |
| Tiempo de corrida. | 10 min |

Los cromatogramas se procesaron con el software OpenLABCDS ChemStation C.01.05 SPI [61]. La cuantificación de la cafeína se realizó utilizando una recta de calibración preparada con un estándar.

### 3.5.6 Validación del método cromatográfico

Los parámetros de validación que se realizaron fueron : linealidad instrumental, límite de detección (LD), límite de cuantificación (LC), precisión y exactitud (recuperación) (A.E.F.I, 2001).

La linealidad se evaluó preparando una recta de calibración con 5 niveles de concentración y cada nivel fue analizado por triplicado (15 determinaciones).

Para evaluar el LD y LC se utilizó el método basado en la extrapolación de la recta de calibrado a concentración 0, en donde se prepararon 3 niveles de concentración por

debajo del nivel más bajo de la recta de calibración y cada nivel se analizó por triplicado (9 determinaciones). Posteriormente se graficó concentración vs área**,** se utilizó la ecuación de la recta y en ella se sustituyó X=0, el valor obtenido correspondía a señal del blanco (Ybl). También se graficó concentración vs desviación estándar, se utilizó la ecuación de la recta y se sustituyó X=0, el valor obtenido correspondía a la desviación del blanco (Sbl). Luego de obtenidos esos valores se procedió a calcular el límite de detección y límite de cuantificación teóricos, utilizando las siguientes fórmulas:

$$LD=Ybl + 3*Sbl/b*raíz(3)$$

$$LC=Ybl + 10*Sbl/b*raíz(3)$$

Para confirmar el LD y el LC, se prepararon soluciones a las concentraciones teóricas y se inyectaron. Se evaluó visualmente que el pico correspondiente a la cafeína se diferenciara de la línea base. El análisis se realizó por triplicado y se reportó el LD y el LC obtenido experimentalmente.

La precisión y la exactitud del método cromatográfico se evaluaron en función de la desviación estándar relativa (RSD) y la recuperación (Rec) respectivamente, obtenidos a través de la fortificación de una muestra a tres niveles de concentración (n=3), cada nivel se analizó por triplicado y en tres días diferentes (n=9).

### 3.5.7 Obtención de los espectros NIR

Para obtener los espectros se utilizó un espectrofotómetro NIR InfraXact™ Pro (FOSS, Estados Unidos). Los espectros fueron tomados en un rango de 570-1850 nm en intervalos de 2 nm en modo de reflectancia difusa (640 variables). La muestra de café verde molida (5 gramos aproximadamente) se colocó en un recipiente redondo de 10.2 cm de diámetro con una base de borosilicato, al momento de obtener los espectros el

recipiente rota y adquiere mediciones desde 10 posiciones diferentes, realiza un promedio y se obtiene el espectro. Cada muestra fue analizada por triplicado.

### 3.5.8 Calibración multivariada

Los modelos de calibración para estimar el contenido de grasa cruda, proteína cruda y cafeína en café verde especial se construyeron comparando dos métodos: regresión de componentes principales (PCR) y regresión de mínimos cuadrados parciales (PLS), junto a diferentes pretratamientos de la data espectral que incluían métodos comunes como: normalización vectorial (SNV), corrección multiplicativa de la dispersión (MSC), primera y segunda derivada, así como suavizado de Savitzky-Golay.

La identificación de los datos atípicos es importante, ya que disminuyen significativamente la calidad del modelo, por lo que antes de construir los modelos, se utilizó el análisis de componentes principales (PCA) para eliminar los datos atípicos de la data espectral. Luego de remover los datos atípicos la data de café verde restante fue dividida en data de entrenamiento (80%) y data de prueba (20%).

Con la finalidad de evaluar los efectos causados por el desbalance de las clases en la data de entrenamiento, se aplicó la técnica de sobremuestreo de minorías sintéticas (SMOTE) a la data de entrenamiento y se compararon los modelos.

Para la construcción de los modelos la data fue centrada y el número de variables latentes (LVs) o componentes principales (PCs) fueron elegidos basados en el menor valor de la raíz del error cuadrático medio de validación cruzada ($RMSE_{CV}$), mediante la validación cruzada en bloques continuos de 10 muestras, utilizando un máximo de 20 variables latentes o componentes.

La exactitud del modelo fue determinada por la raíz del error cuadrático medio de calibración ($RMSE_C$), por el error cuadrático medio de validación cruzada ($RMSE_{CV}$) y el coeficiente de determinación ($R^2_{CV}$).

En cuanto a la validación externa, se consideró el error cuadrático medio de predicción ($RMSE_P$), la desviación residual predictiva (RPD) y el error relativo de predicción (REP).

De manera general valores altos de $R^2$ y RPD así como valores pequeños de RMSE indican una buena capacidad predictiva del modelo (Xu *et al*., 2018).

### 3.5.9 Análisis estadístico

El análisis descriptivo de las variables químicas, la partición en conjunto de calibración y predicción y la aplicación del algoritmo SMOTE se realizó con “R” y “RStudio” (v. 2022.12.0+353) utilizando las librerías stats v. 4.2.3 (R Core Team, 2023) , caret v. 6.0.94 (Kuhn y Max, 2008) y smotefamily v. 1.3.1 (Siriseriwan, 2019) (ver códigos en el Anexo II).

La identificación de valores atípicos, preprocesamiento espectral y la construcción de los modelos predictivos se llevó a cabo en el software SOLO versión 9.2 (Eigenvector Research, Inc).

# CAPÍTULO 4.

# ANÁLISIS E INTERPRETACIÓN DE RESULTADOS

## 4.1 Determinación de parámetros fisicoquímicos

El contenido de grasa cruda, proteína cruda y cafeína determinados en las 98 muestras de café verde especial se resumen en la Tabla 5.

**Tabla 5. Rango, promedio y desviación estándar del contenido de grasa, proteína cruda y cafeína del café verde especial por variedad.**

| Clase | *Grasa Cruda (%) | | *Proteína Cruda (%) | | *Cafeína (%) | |
|---|---|---|---|---|---|---|
| | Rango | Promedio± Desviación | Rango | Promedio± Desviación | Rango | Promedio± Desviación |
| Geisha (n=72) | 6.22-10.79 | 8.04 ± 0.90 | 12.69-14.83 | 13.85 ± 0.43 | 0.70-1.11 | 0.88 ± 0.09 |
| Tradicional (n=26) | 6.99-9.45 | 8.32 ± 0.66 | 12.16-14.51 | 13.54 ± 0.64 | 0.69-1.53 | 1.16 ± 0.15 |

*Los datos son presentados en base seca

La grasa cruda estuvo en un rango de 6.22-10.79 % y 6.99-9.45 % para Geisha y Tradicional respectivamente. Estos valores de grasa cruda fueron menores a los obtenidos por Zhu *et al*., (2021), quienes encontraron contenidos de grasa de 12.88 a 16.29 g $100^{-1}$ para muestras de café verde de la variedad Arábica procedente de ocho países diferentes. Otros autores mencionan que el contenido de lípidos en el café verde puede variar entre 8-20 % (Dong *et al*., 2021; Pimpley *et al*., 2020; Yeager *et al*., 2023). La variación en la composición química puede ser afectada significativamente por condiciones edafoclimáticas y la genética del café (Barbosa *et al*., 2019). De acuerdo con la Figura 3, para la clase Geisha se observó una mayor dispersión de los datos en comparación con la clase tradicional, sin embargo, se podría decir que en el contenido de grasa es similar entre la clase Geisha y Tradicional.

El contenido de lípidos influye en la calidad del café, ya que durante el proceso de tostado participan en reacciones de descomposición y oxidación contribuyendo al sabor y aroma del café (Zhu, Long, Ma, *et al*., 2021).

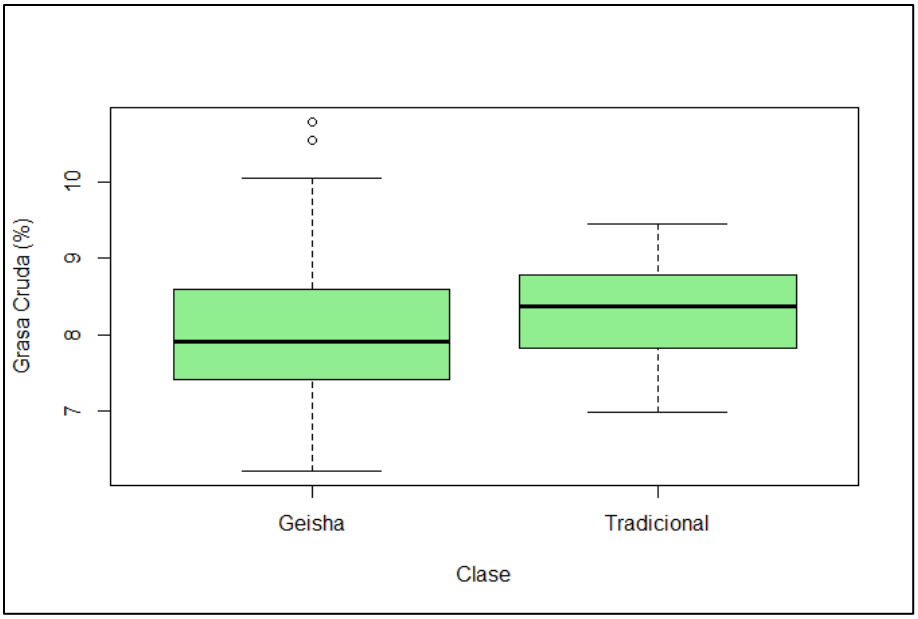



**Figura 3.Gráfico de caja y bigote para el contenido de grasa cruda en café verde especial.**

El contenido de proteína cruda estuvo en un rango de 12.69-14.83 % y 12.16-14.51 % para Geisha y Tradicional respectivamente. Según Saud y Salamatullah, (2021), el contenido promedio de proteína en muestras de café verde es de aproximadamente 13 %, lo que concuerda con lo obtenido en este estudio (Tabla 5). En cuanto a la Figura 4, se observa que los valores de proteína cruda son similares en ambas clases, pero se observa una mayor dispersión de los valores en la clase Tradicional. La proteína junto a otros componentes químicos son precursores de otros compuestos durante el tostado, que se relacionan con el sabor y el aroma del café (Zhu, *et al*., 2021).

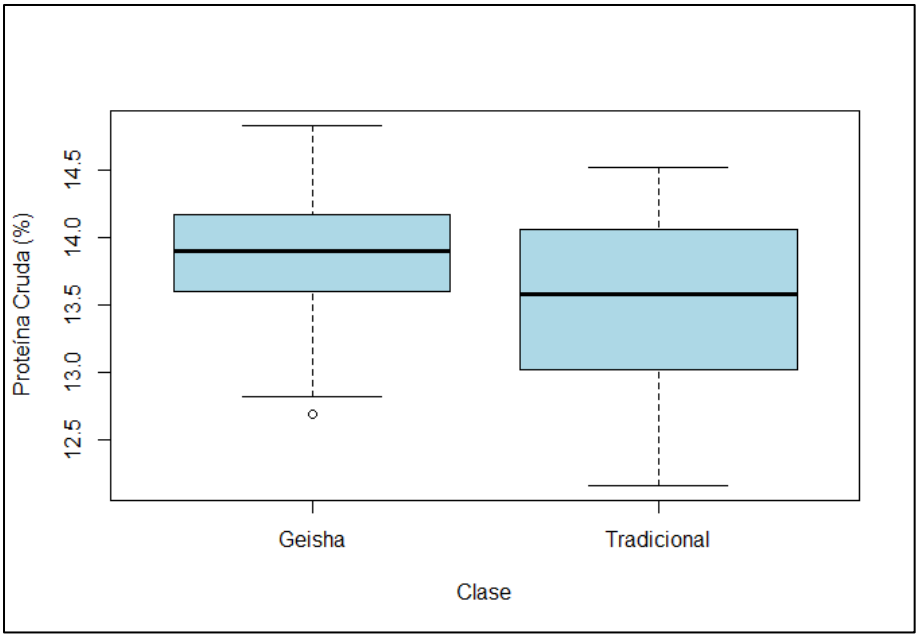



**Figura 4.Gráfico de caja y bigote para el contenido de proteína cruda en café verde especial.**

La determinación del contenido de cafeína se realizó utilizando un método cromatográfico previamente validado y la Tabla 6 muestra los resultados de esta validación.

**Tabla 6. Parámetros de validación analizados para la determinación de cafeína en café verde especial.**

| | | |
|---|---|---|
| **Rango de trabajo en matriz (%)** | | 0.21-0.98 |
| **LD (%)** | | 0.005 |
| **LC (%)** | | 0.008 |
| **Ecuación de la recta** | | y=53717x + 2.6595 |
| **$R^2$** | | 0.9997 |
| **Rec. ± RSD** | ***n1*** | 110.46 ± 14.59 |
| | ***n2*** | 102.61 ± 8.62 |
| | ***n3*** | 102.43 ± 5.47 |

LD=Límite de detección, LC= límite de cuantificación, $R^2$= coeficiente de determinación, Rec.= recuperación, RSD= desviación estándar relativa.

De acuerdo con los resultados de la validación, se encontró una excelente relación lineal ($R^2$=0.9997), en un rango de concentración 0.21-0.98 % de cafeína. El límite de

detección y límite de cuantificación obtenidos de manera experimental fueron de 0.005 % y 0.008 % respectivamente; estos parámetros nos permitieron evaluar la sensibilidad del método cromatográfico. La recuperación estuvo dentro del rango de 102.43-110.46 %, y el RSD estuvo por debajo del 15 % para los tres niveles de concentración ensayados. Estos resultados permiten confirmar la confiabilidad del método cromatográfico para determinar el contenido de cafeína en las muestras de café verde.

El contenido de cafeína estuvo en un rango de 0.70-1.11 % y 0.69-1.53 % para Geisha y Tradicional respectivamente. En la Figura 5 se observa que la concentración de cafeína es mayor en la clase Tradicional, mientras que se observa más dispersión en los valores de la clase Geisha.

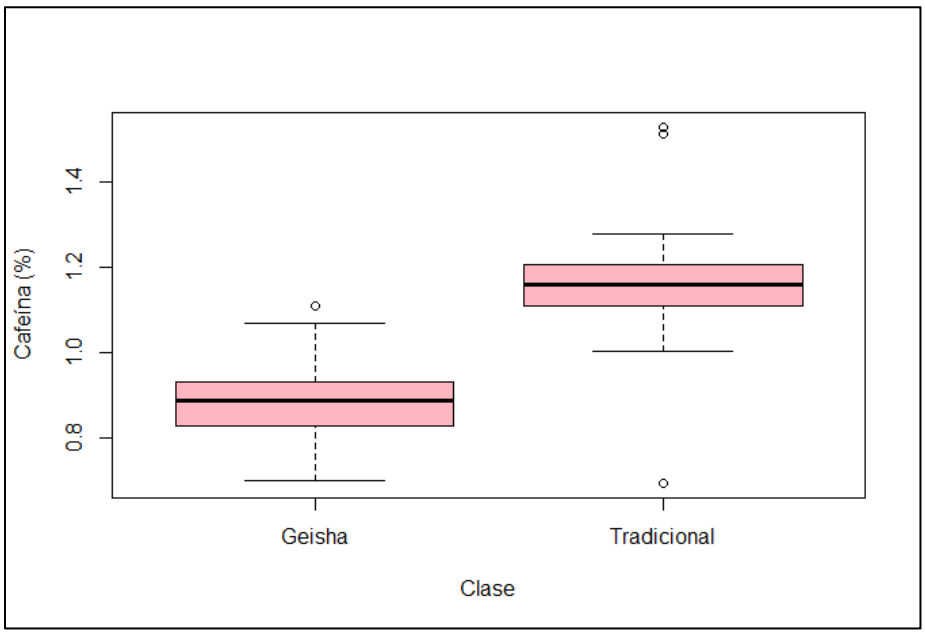



**Figura 5. Gráfico de caja y bigote para el contenido de cafeína en café verde especial.**

El contenido de cafeína promedio obtenido en este estudio (0.88 %) para Geisha, coincidió con el reportado por Santamaría *et al*., (2023), quienes reportaron contenidos de cafeína de 0.84 % y 0.87 %, para café Geisha natural y lavado respectivamente. La cafeína junto a otras sustancias, se relacionan con el amargor (Mahmud *et al*., 2020) y el aroma en la bebida de café. (Sualeh *et al*., 2020).

## 4.2 Calibración Multivariada

### 4.2.1 Espectro infrarrojo cercano

El espectro original, promedio y con segunda derivada, obtenido para las muestras de café verde especial en un rango de 570-1850 nm se presenta en la Figura 6.

La espectroscopía de infrarrojo cercano abarca la región del espectro electromagnético que va de 800-2500 nm y se caracteriza por bandas de absorción relacionadas a combinaciones y bandas de sobretonos correspondientes a enlaces C-H, N-H, OH y SH (Zhu *et al*., 2021). El espectro obtenido en este estudio contempla una parte de la región NIR, que va de 800-1850 nm debido a la configuración del equipo utilizado para obtener los espectros e incorpora una parte de la región del espectro visible que va de 570-800nm (figura 6). A pesar de no contar con todo el rango espectral del infrarrojo cercano, la región de 570-1850 nm contempla bandas de absorción para enlaces $CH_3$, $CH_2$ y CH, que están presentes en la cafeína, proteína y grasa (Munyendo *et al*., 2022).

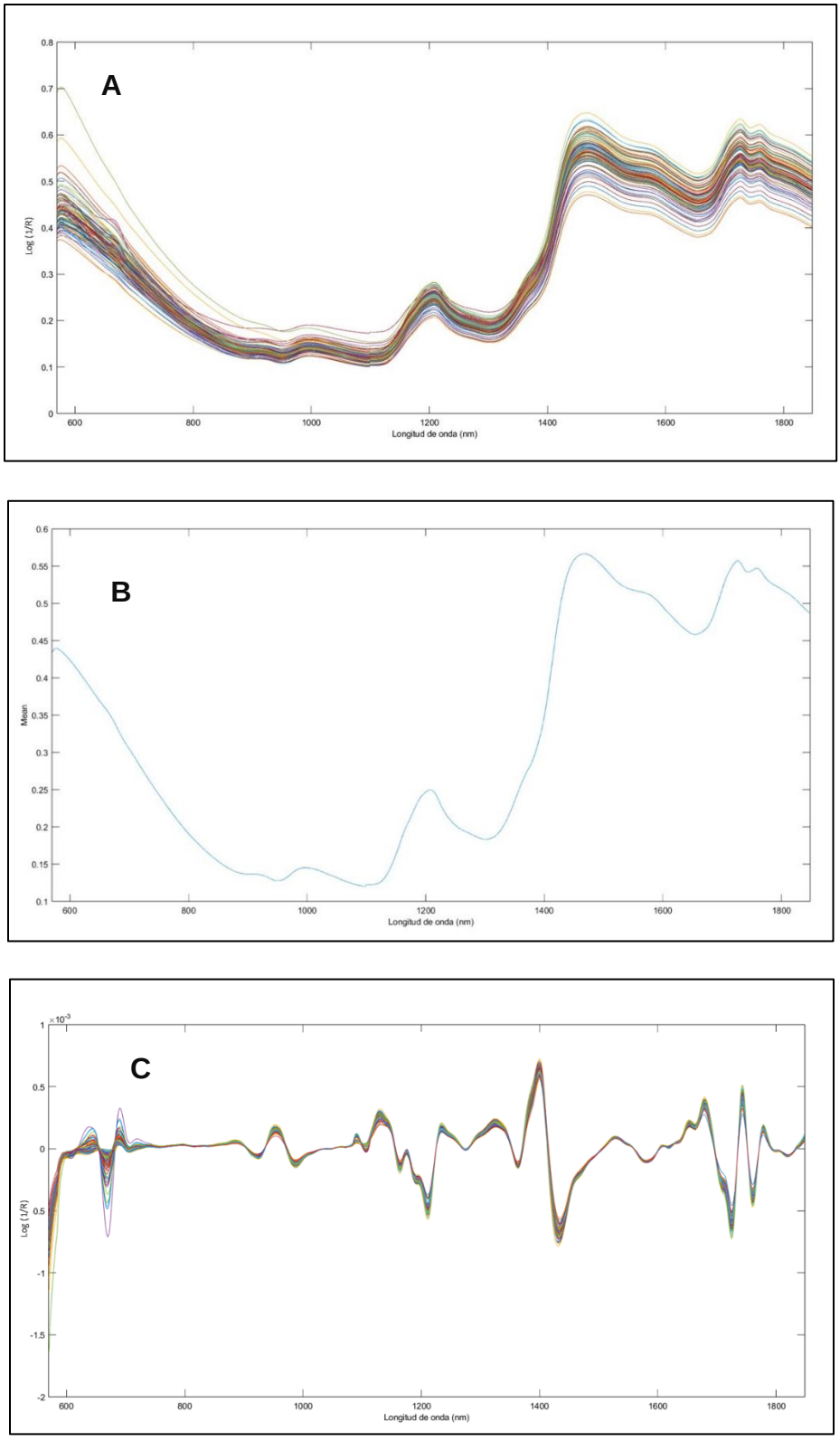



**Figura 6. Espectro de café verde en el rango de 570-1850nm: (A) espectro original; B) espectro promedio y (C) espectro con segunda derivada.**

La sensibilidad de las mediciones espectroscópicas depende de las intensidades de las bandas. La información espectral se repite a lo largo de los sucesivos armónicos y regiones de combinación; sin embargo, las señales de las bandas implicadas se debilitan en un orden de magnitud. Los terceros sobretonos (400-800 nm), tienen intensidades de banda considerablemente más débiles en comparación con el segundo y primer sobretono en la región NIR (800-1800 nm) (Barbin *et al*., 2014).

### 4.2.2 Eliminación de datos atípicos

De acuerdo con el gráfico de puntajes del PCA de la data espectral que se muestra en la Figura 7A, se consideraron datos atípicos a todas las muestras que estaban fuera de la elipse al 95b% de confianza, por lo que se identificaron un total de 13 datos atípicos, de los cuales 2 pertenecían a la clase Geisha y 11 a la clase Tradicional.

Los datos atípicos son muestras que tienen un comportamiento diferente del resto de los datos y pueden ser removidos hasta un 22.2b% del total de muestras en el conjunto de datos, según la Sociedad Estadounidense para Pruebas y Materiales (ASTM, por sus siglas en inglés) (Baqueta *et al*., 2020), lo cual fue considerado en este estudio.

Después de eliminar los datos atípicos la data fue dividida en dos subconjuntos de datos: la data para entrenamiento (80b%) y la data para predicción (20b%) y se presentan en la Tabla 7.

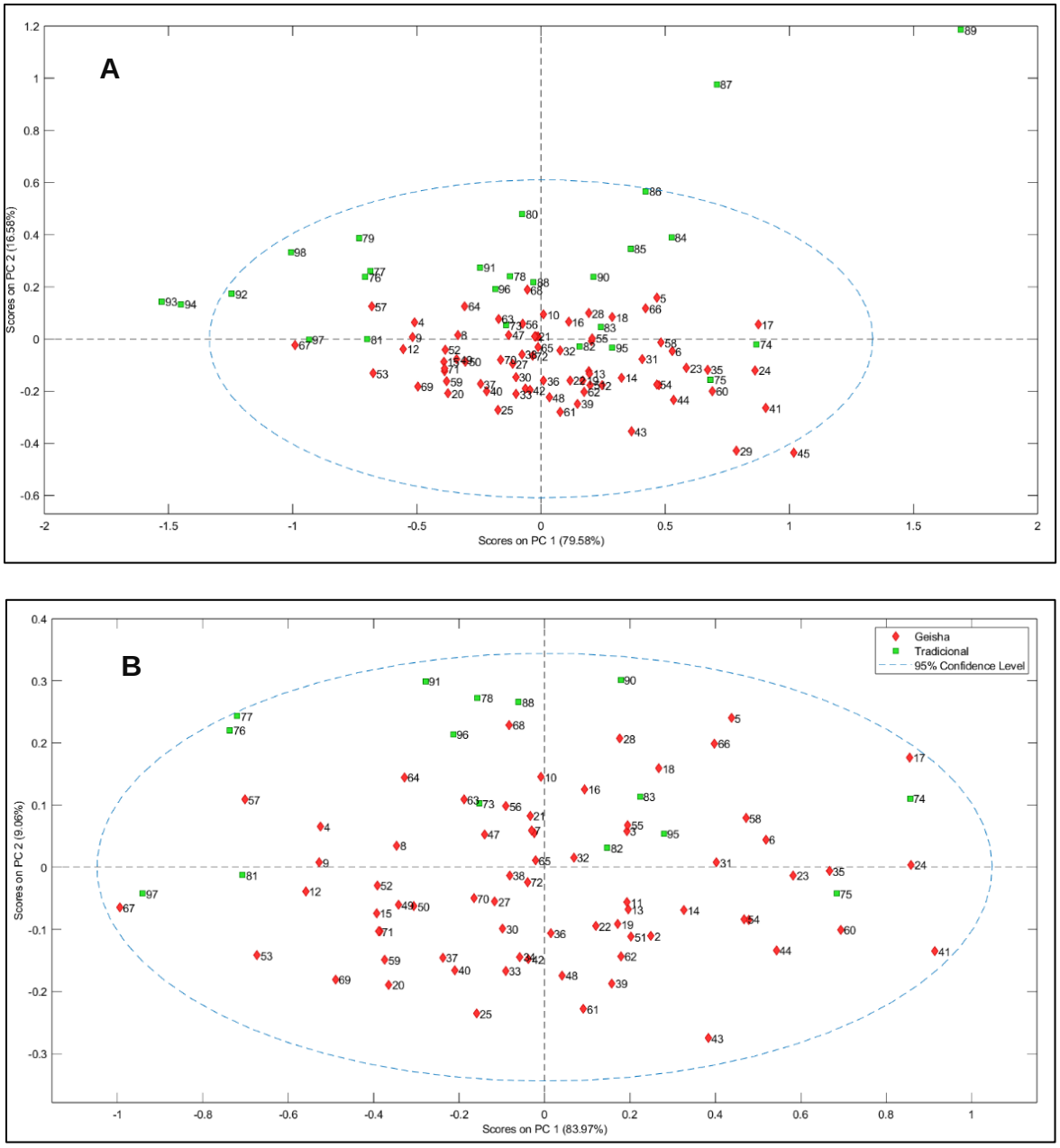



**Figura 7. Gráfico de puntaje de PCA de la data espectral de las muestras de café verde especial: (A) Gráfico de puntaje original; (B) Gráfico de puntaje después de eliminar datos atípicos.**

**Tabla 7. Resumen de la estadística descriptiva para el contenido de grasa cruda, proteína cruda y cafeína de las muestras de café verde especial en los conjuntos de datos de entrenamiento y predicción.**

| Contenido | Conjunto de datos | Número | Rango | Promedio | Desviación |
|---|---|---|---|---|---|
| **Grasa Cruda (%)** | **Entrenamiento** | 69 | 6.22-10.79 | 8.07 | 0.87 |
| | **Predicción** | 16 | 6.33-9.22 | 7.94 | 0.81 |
| **Proteína Cruda (%)** | **Entrenamiento** | 69 | 12.16-14.83 | 13.77 | 0.55 |
| | **Predicción** | 16 | 13.26-14.33 | 13.77 | 0.37 |
| **Cafeína (%)** | **Entrenamiento** | 69 | 0.69-1.28 | 0.92 | 0.13 |
| | **Predicción** | 16 | 0.77-1.16 | 0.94 | 0.11 |

### 4.2.3 Generación de muestras sintéticas

El modo de sobremuestreo de SMOTE fue utilizado para generar nuevas muestras para la clase minoritaria en el conjunto de datos de entrenamiento. La clase minoritaria fue la Tradicional (12 muestras) en comparación con la clase Geisha (57 muestras) como se observa en la Figura 8A. Luego de aplicado el método SMOTE la clase minoritaria contenía un total de 48 muestras como se muestra en la Figura 8B.

Según lo mencionado por Demir y Şahin, (2022), el conjunto de datos utilizado para validar el método no debe incluir muestras generadas de manera sintética, por lo que en este estudio el método de sobremuestreo se aplicó al conjunto de datos de entrenamiento, esto debido a que la capacidad predictiva del modelo puede verse afectada al contener muestras que no son reales.

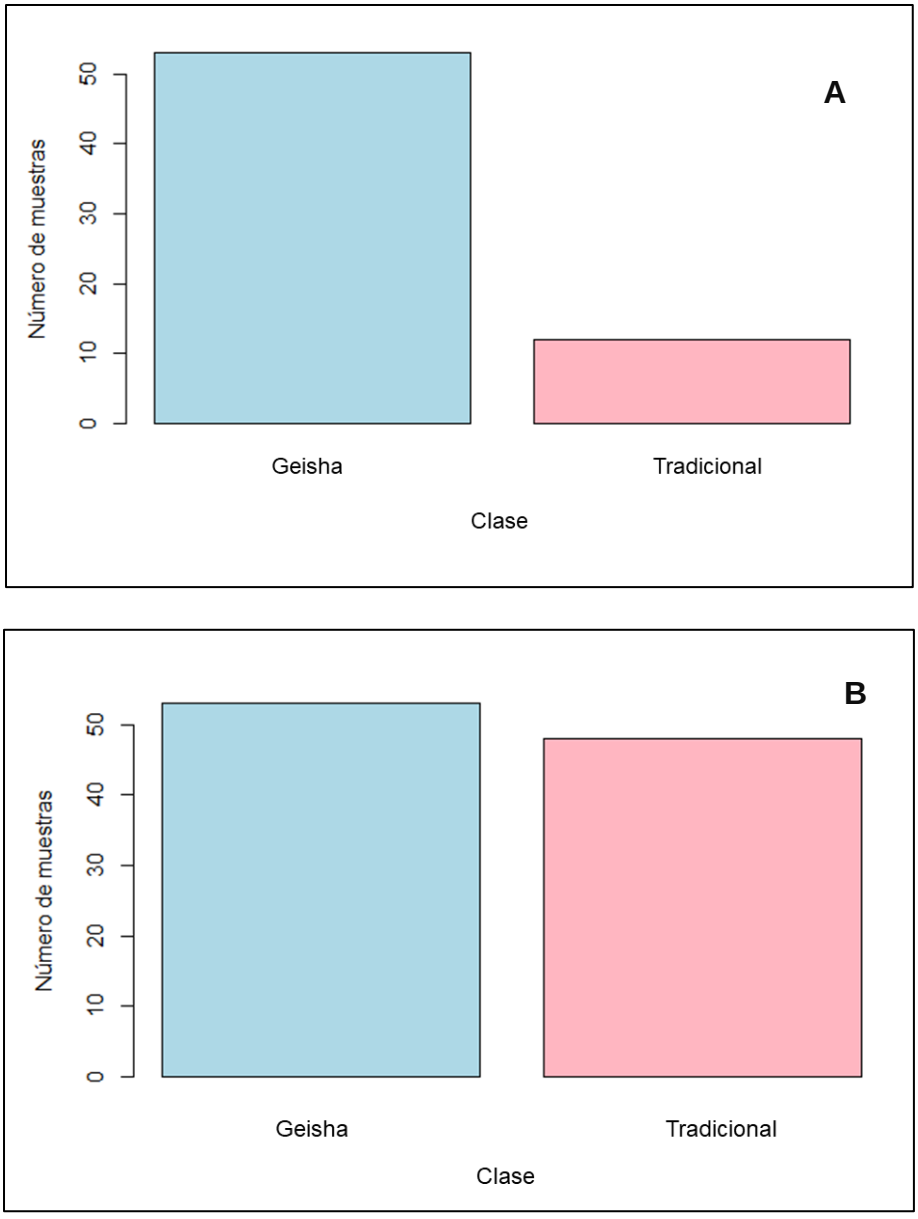



**Figura 8. Número de muestras por clase Geisha y Tradicional: (A): conjunto de datos de entrenamiento desbalanceado; (B) conjunto de datos de entrenamiento con SMOTE.**

El desequilibrio en las clases introduce sesgos inherentes y puede conducir a un rendimiento bajo para las clases minoritarias, ya que los algoritmos de aprendizaje tienden a estar sesgados hacia la clase mayoritaria (Nemade *et al*., 2023).

### 4.2.4 Modelos de regresión multivariada

Un total de 84 modelos de calibración fueron construidos para cada variable química. Estos modelos fueron construidos tomando en cuenta dos modelos de regresión multivariada (PLS y PCR), tres regiones espectrales diferentes, 7 preprocesamientos espectrales que incluían centrado medio, MSC, SNV, SG, OSC, primera y segunda derivada (ver Anexo I Figura A2-A5), además de considerar la construcción de los modelos con la data original y con muestras sintéticas (SMOTE). Los mejores modelos (resaltados en negrita) fueron elegidos en función del menor $RMSE_{CV}$, mayor $R^2_{CV}$ y mejor capacidad predictiva.

#### *4.2.4.1 Modelos predictivos para el contenido de grasa cruda*

Los resultados de los modelos para predecir grasa cruda (%) en café verde especial se muestran en la Tabla 8. La capacidad de predicción en función de los modelos de regresión PLS y PCR fueron similares (Figura 10). Los métodos de PCR y PLS se adaptan a datos espectrales correlacionados linealmente y se utilizan para ajustar modelos cuando hay más datos espectrales que muestras, como ocurrió en este estudio (Xu *et al*., 2018). Hubo una ligera mejora en los coeficientes de determinación y en consecuencia en los errores de los modelos, por efecto del SMOTE. La exactitud de los modelos para la estimación de la grasa cruda se evaluó en función de la raíz del error cuadrático medio, donde la tendencia general observada para estos errores fue de $RMSE_C < RMSE_{CV} > RMSE_P$, tanto para PLS y PCR.

**Tabla 8. Resultados de los diferentes modelos para la estimación del contenido de grasa cruda (%) en café verde especial.**

| PLS | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Calibración | | Validación Cruzada | | | Predicción | | | |
| **Región Espectral (nm)** | **Preprocesamiento** | **n** | **SMOTE** | **LVs** | **$R^2$** | **$RMSE_C$** | **$R^2_{CV}$** | **$RMSE_{CV}$** | **n** | **$R^2_P$** | **$RMSE_P$** | **RPD** | **REP** |
| 570-1850 | MSC | 69 | - | 4 | 0.458 | 0.633 | 0.336 | 0.710 | 16 | 0.623 | 0.500 | 1.63 | 6.30 |
| 570-1850 | 2da. derivada (15,2,2) | 105 | + | 9 | 0.661 | 0.437 | 0.462 | 0.564 | 16 | 0.654 | 0.469 | 1.74 | 5.90 |
| 800-1850 | Centrado medio | 69 | - | 5 | 0.524 | 0.594 | 0.404 | 0.672 | 16 | 0.623 | 0.510 | 1.60 | 6.41 |
| 800-1850 | SG (15) | 105 | + | 5 | 0.549 | 0.503 | 0.501 | 0.531 | 16 | 0.623 | 0.516 | 1.58 | 6.50 |
| 1480-1850 | SG (15) | 69 | - | 7 | 0.564 | 0.569 | 0.462 | 0.636 | 16 | 0.421 | 0.600 | 1.36 | 7.55 |
| 1480-1850 | Centrado medio | 105 | + | 8 | 0.611 | 0.467 | 0.519 | 0.525 | 16 | 0.517 | 0.555 | 1.46 | 6.99 |
| **570-1850** | **2da. derivada (15,2,2)** | **102** | **+** | **11** | **0.787** | **0.306** | **0.631** | **0.407** | **13** | **0.601** | **0.444** | **1.62** | **5.53** |
| **PCR** | | | | | | | | | | | | | |
| **Región Espectral (nm)** | **Preprocesamiento** | **n** | **SMOTE** | **PCs** | **$R^2$** | **$RMSE_C$** | **$R^2_{CV}$** | **$RMSE_{CV}$** | **n** | **$R^2_P$** | **$RMSE_P$** | **RPD** | **REP** |
| 570-1850 | SG (15) | 69 | - | 8 | 0.510 | 0.602 | 0.359 | 0.708 | 16 | 0.657 | 0.488 | 1.67 | 6.14 |
| 570-1850 | SG (15) | 105 | + | 8 | 0.546 | 0.505 | 0.475 | 0.545 | 16 | 0.642 | 0.500 | 1.63 | 6.30 |
| 800-1850 | SG (15) | 69 | - | 6 | 0.517 | 0.598 | 0.405 | 0.671 | 16 | 0.635 | 0.505 | 1.61 | 6.36 |
| 800-1850 | SG (15) | 105 | + | 6 | 0.569 | 0.505 | 0.498 | 0.533 | 16 | 0.638 | 0.502 | 1.62 | 6.32 |
| 1480-1850 | 2da. derivada (15,2,2) | 69 | - | 11 | 0.585 | 0.555 | 0.434 | 0.653 | 16 | 0.537 | 0.544 | 1.50 | 6.85 |
| 1480-1850 | 2da. derivada (15,2,2) | 105 | + | 11 | 0.620 | 0.462 | 0.536 | 0.512 | 16 | 0.592 | 0.520 | 1.50 | 6.54 |

PLS=mínimos cuadrados parciales, PCR=regresión de componentes principales, n= número de muestras, += con SMOTE, -= sin SMOTE, LVs= variables latentes, PCs= componentes principales, $R^2$= coeficiente de determinación calibración, $R^2_{CV}$= coeficiente de determinación validación cruzada, $R^2_p$= coeficiente de determinación predicción, $RMSE_C$= raíz del error cuadrático medio de calibración, $RMSE_{CV}$= raíz del error cuadrático medio de validación cruzada, $RMSE_P$= raíz del error cuadrático medio de predicción, RPD= desviación predictiva residual, REP= error relativo de predicción, MSC= corrección del efecto multiplicativo de la dispersión, SG: suavizado de Savitzky-Golay.

**Tabla 9. Resultados de los diferentes modelos para la estimación del contenido de proteína cruda (%) en café verde especial.**

| PLS | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Calibración | | Validación Cruzada | | | | Predicción | | |
| **Región Espectral (nm)** | **Preprocesamiento** | **n** | **SMOTE** | **LVs** | **$R^2_C$** | **$RMSE_C$** | **$R^2_{CV}$** | **$RMSE_{CV}$** | **n** | **$R^2_P$** | **$RMSE_P$** | **RPD** | **REP** |
| 570-1850 | SG (15) | 69 | - | 19 | 0.876 | 0.191 | 0.621 | 0.344 | 16 | 0.449 | 0.265 | 1.38 | 1.92 |
| 570-1850 | OSC | 105 | + | 14 | 0.923 | 0.164 | 0.845 | 0.234 | 16 | 0.379 | 0.288 | 1.27 | 2.09 |
| 800-1850 | 2da. derivada (15,2,2) | 69 | - | 9 | 0.810 | 0.236 | 0.617 | 0.339 | 16 | 0.309 | 0.351 | 1.04 | 2.55 |
| 800-1850 | 2da. derivada (15,2,2) | 105 | + | 16 | 0.975 | 0.094 | 0.823 | 0.258 | 16 | 0.325 | 0.357 | 1.03 | 2.59 |
| 1480-1850 | MSC | 69 | - | 6 | 0.722 | 0.286 | 0.580 | 0.353 | 16 | 0.448 | 0.284 | 1.29 | 2.06 |
| 1480-1850 | 2da. derivada (15,2,2) | 105 | + | 8 | 0.830 | 0.243 | 0.718 | 0.316 | 16 | 0.407 | 0.368 | 1.00 | 2.67 |
| **800-1850** | **2da. derivada (5,2,2)** | **99** | **+** | **9** | **0.965** | **0.108** | **0.817** | **0.249** | **13** | **0.793** | **0.199** | **1.86** | **1.45** |
| **PCR** | | | | | | | | | | | | | |
| **Región Espectral (nm)** | **Preprocesamiento** | **n** | **SMOTE** | **PCs** | **$R^2_C$** | **$RMSE_C$** | **$R^2_{CV}$** | **$RMSE_{CV}$** | **n** | **$R^2_P$** | **$RMSE_P$** | **RPD** | **REP** |
| 570-1850 | OSC | 69 | - | 20 | 0.858 | 0.204 | 0.601 | 0.347 | 16 | 0.419 | 0.282 | 1.30 | 2.04 |
| 570-1850 | MSC | 105 | + | 18 | 0.908 | 0.179 | 0.849 | 0.230 | 16 | 0.329 | 0.231 | 1.59 | 1.68 |
| 800-1850 | 2da. derivada (15,2,2) | 69 | - | 17 | 0.787 | 0.250 | 0.584 | 0.352 | 16 | 0.363 | 0.327 | 1.12 | 2.37 |
| 800-1850 | SG (15) | 105 | + | 18 | 0.856 | 0.224 | 0.761 | 0.290 | 16 | 0.340 | 0.328 | 1.12 | 2.38 |
| 1480-1850 | SNV | 69 | - | 8 | 0.720 | 0.287 | 0.574 | 0.357 | 16 | 0.472 | 0.274 | 1.33 | 1.99 |
| 1480-1850 | 2da. derivada (15,2,2) | 105 | + | 13 | 0.761 | 0.288 | 0.691 | 0.328 | 16 | 0.420 | 0.345 | 1.06 | 2.51 |

PLS=mínimos cuadrados parciales, PCR=regresión de componentes principales, n= número de muestras, += con SMOTE, -= sin SMOTE, LVs= variables latentes, PCs= componentes principales, $R^2$= coeficiente de determinación calibración, $R^2_{CV}$= coeficiente de determinación validación cruzada, $R^2_p$= coeficiente de determinación predicción, $RMSE_C$= raíz del error cuadrático medio de calibración, $RMSE_{CV}$= raíz del error cuadrático medio de validación cruzada, $RMSE_P$= raíz del error cuadrático medio de predicción, RPD= desviación predictiva residual, REP= error relativo de predicción, MSC= corrección del efecto multiplicativo de la dispersión, SNV= variable normal estándar, SG: suavizado de Savitzky-Golay, OSC= corrección de señal ortogonal.

**Tabla 10. Resultados de los diferentes modelos para la estimación del contenido de cafeína (%) en café verde especial.**

| PLS | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Calibración | | Validación Cruzada | | | Predicción | | | |
| **Región Espectral (nm)** | **Preprocesamiento** | **n** | **SMOTE** | **LVs** | **$R^2_C$** | **$RMSE_C$** | **$R^2_{CV}$** | **$RMSE_{CV}$** | **n** | **$R^2_P$** | **$RMSE_P$** | **RPD** | **REP** |
| 570-1850 | 2da. derivada (15,2,2) | 69 | - | 10 | 0.787 | 0.061 | 0.500 | 0.099 | 16 | 0.569 | 0.076 | 1.50 | 8.11 |
| 570-1850 | 2da. derivada (15,2,2) | 105 | + | 10 | 0.912 | 0.046 | 0.774 | 0.075 | 16 | 0.572 | 0.073 | 1.55 | 7.84 |
| 800-1850 | 2da. derivada (15,2,2) | 69 | - | 8 | 0.782 | 0.062 | 0.561 | 0.090 | 16 | 0.646 | 0.066 | 1.71 | 7.09 |
| 800-1850 | MSC | 105 | + | 14 | 0.894 | 0.050 | 0.759 | 0.077 | 16 | 0.674 | 0.065 | 1.74 | 7.09 |
| 1480-1850 | 2da. derivada (15,2,2) | 69 | - | 5 | 0.726 | 0.069 | 0.586 | 0.086 | 16 | 0.542 | 0.075 | 1.46 | 8.34 |
| 1480-1850 | MSC | 105 | + | 11 | 0.906 | 0.047 | 0.810 | 0.067 | 16 | 0.474 | 0.084 | 1.35 | 8.97 |
| **570-1850** | **2da. derivada (5,2,2)** | **99** | **+** | **8** | **0.939** | **0.038** | **0.865** | **0.057** | **14** | **0.822** | **0.057** | **2.35** | **5.42** |
| **PCR** | | | | | | | | | | | | | |
| **Región Espectral (nm)** | **Preprocesamiento** | **n** | **SMOTE** | **PCs** | **$R^2_C$** | **$RMSE_C$** | **$R^2_{CV}$** | **$RMSE_{CV}$** | **n** | **$R^2_P$** | **$RMSE_P$** | **RPD** | **REP** |
| 570-1850 | 2da. derivada (15,2,2) | 69 | - | 8 | 0.544 | 0.089 | 0.428 | 0.101 | 16 | **0.474** | 0.081 | 1.40 | 8.7 |
| 570-1850 | 2da. derivada (15,2,1) | 105 | + | 14 | 0.756 | 0.076 | 0.683 | 0.087 | 16 | 0.594 | 0.072 | 1.59 | 7.66 |
| 800-1850 | 2da. derivada (15,2,2) | 69 | - | 13 | 0.725 | 0.069 | 0.542 | 0.090 | 16 | 0.656 | 0.066 | 1.72 | 7.05 |
| 800-1850 | 2da. derivada (15,2,2) | 105 | + | 17 | 0.874 | 0.055 | 0.803 | 0.069 | 16 | 0.620 | 0.069 | 1.65 | 7.36 |
| 1480-1850 | MSC | 69 | - | 14 | 0.760 | 0.065 | 0.582 | 0.086 | 16 | 0.522 | 0.078 | 1.45 | 8.36 |
| 1480-1850 | 2da. derivada (15,2,2) | 105 | + | 10 | 0.840 | 0.062 | 0.792 | 0.070 | 16 | 0.576 | 0.073 | 1.56 | 7.79 |

PLS=mínimos cuadrados parciales, PCR=regresión de componentes principales, n= número de muestras, += con SMOTE, -= sin SMOTE, LVs= variables latentes, PCs= componentes principales, $R^2$= coeficiente de determinación calibración, $R^2_{CV}$= coeficiente de determinación validación cruzada, $R^2_p$= coeficiente de determinación predicción, $RMSE_C$= raíz del error cuadrático medio de calibración, $RMSE_{CV}$= raíz del error cuadrático medio de validación cruzada, $RMSE_P$= raíz del error cuadrático medio de predicción, RPD= desviación predictiva residual, REP= error relativo de predicción, MSC= corrección del efecto multiplicativo de la dispersión.

Los valores para $RMSE_{CV}$ fueron similares a los de $RMSE_C$, lo que indica que las muestras excluidas durante el paso de validación cruzada fueron bien predichas. Por otro lado, los valores de $RMSE_P$ fueron menores a los de $RMSE_{CV}$, esto indica que la dimensión del modelo (variables latentes o componentes principales) se eligieron adecuadamente y no existe un sobreajuste (Baqueta *et al*., 2020).

El mejor modelo para la estimación de grasa cruda fue PLS-2da. derivada (15,2,2) con SMOTE aplicado a todo el rango espectral y se obtuvo un $R^2$ de 0.787, 0.631 y 0.601 para los conjuntos de datos de calibración, validación cruzada y predicción respectivamente (Figura 9). Los errores asociados a este modelo fueron 0.306, 0.407 y 0.444 para $RMSE_C$, $RMSE_{CV}$ y $RMSE_P$ respectivamente. Además, el RPD fue de 1.62 y el REP de 5.53.

Valores de $R^2$ >0.90 indican modelos con excelente capacidad predictiva, $R^2$ entre 0.82-0.90 indican una buena capacidad predictiva, $R^2$ entre 0.65 y 0.82 indican predicciones aproximadas y modelos con $R^2$ menores de 0.65 no producen predicciones exitosas (Xu *et al*., 2018).

En contraste con la calibración univariada, es inusual obtener coeficientes de determinación ($R^2$) mayores de 0.95 en modelos de calibración multivariada para analitos de matrices complejas (Hashimoto *et al*., 2018); además el coeficiente de determinación, está relacionado directamente con los errores incorporados por el método de referencia, por lo que si este presenta altos niveles de errores analíticos, podría disminuir en consecuencia el coeficiente de determinación (Baqueta *et al*., 2019). Por otro lado, el RPD es la relación entre la desviación estándar de los valores de referencia y el $RMSE_P$

en el set de predicción. Cuando los valores de RPD son mayores de 2.5 y el $RMSE_P$ es menor que el $RMSE_{CV}$ el modelo se considera confiable y valores de RPD superiores a 1.5 indican que el modelo puede ser utilizado, mientras que valores menores a 1.5 indican que el modelo no debe ser utilizado (Baqueta *et al*., 2020; Zhu, Long, Chen, *et al*., 2021). Además, valores de REP menores de 2 % se consideran excelentes, entre 2-5 % buenos, entre 5 y 10 se considera razonable y malos cuando son más de 10 % (Olivieri y Escandar, 2014).

Tomando en consideración las métricas anteriormente mencionadas, el modelo basado en PLS-2da. derivada (15,2,2) con SMOTE, utilizando todo el rango espectral indica un modelo con una pobre calidad para la cuantificación del contenido de grasa cruda en las muestras de café verde especial, sin embargo, puede ser utilizado con fines prácticos (screening) ya que el error de predicción ($RMSE_P$) fue menor a la desviación estándar del conjunto de datos utilizados en la validación externa.

Estos resultados fueron similares a los obtenidos por Caporaso *et al*., (2021), quienes reportaron un modelo de $R^2$= 0.62 y 0.52, errores de 3.58 y 4.06 para el conjunto de datos de calibración y predicción respectivamente, además de un RPD= 1.41 utilizando PLS-2da. derivada, para la estimación de lípidos totales en granos de cacao sin cáscara y concluyeron que podía ser utilizado con fines de detección general.

Por el contrario, los resultados de este estudio fueron diferentes a los obtenidos por Zhu *et al*., (2021), quienes reportaron un modelo para la estimación de lípidos en granos de café verde utilizando PLS-OSC con $R^2$= 0.964 y 0.975, errores de 0.173 y 0.123 para la calibración cruzada y la predicción respectivamente, además de un

RPD=6.51, concluyendo que OSC mejoró el rendimiento de los modelos en comparación con los pretratamientos espectrales.

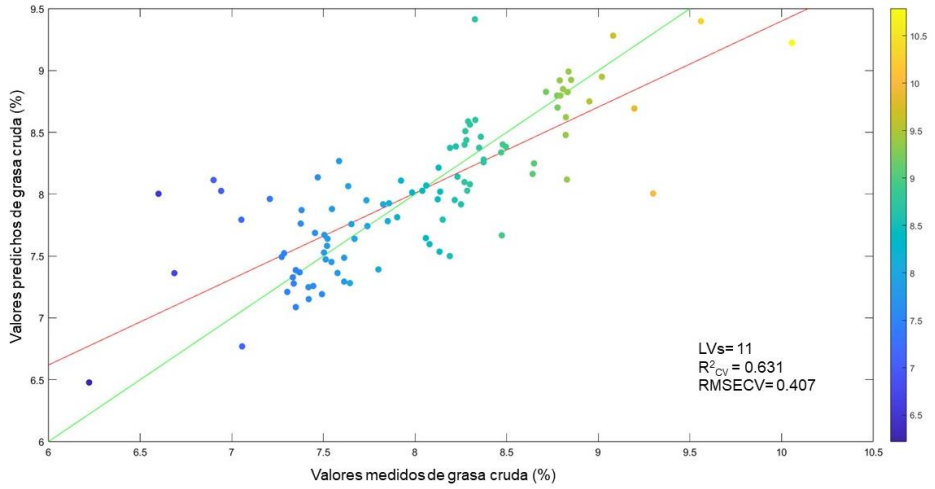
Para obtener modelos con una buena capacidad predictiva es necesario incluir muestras con información química y espectral diversa y que sea representativa de las muestras naturales (Hashimoto *et al*., 2018).

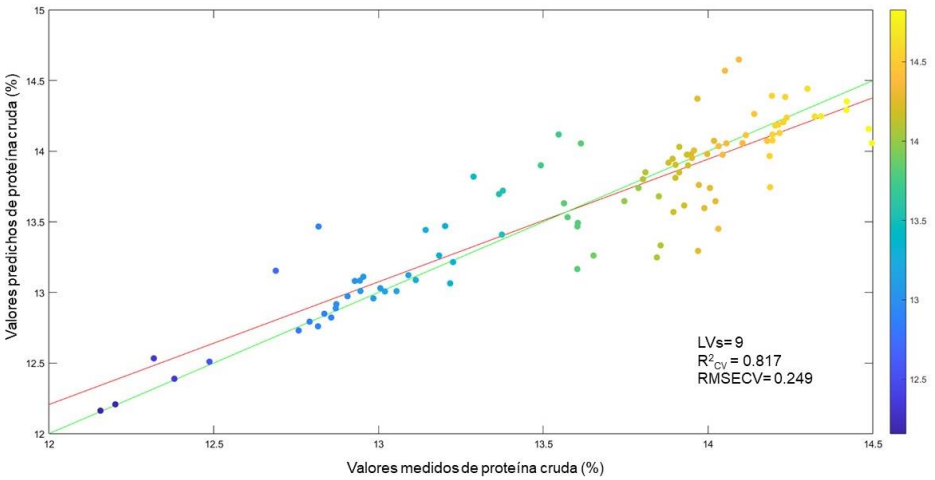
#### *4.2.4.2 Modelos predictivos para el contenido de proteína cruda*

Los resultados de los mejores modelos de predicción para proteína cruda (%) en café verde especial se muestran en la Tabla 9.

De manera general los modelos obtenidos por PLS y PCR fueron similares (Figura 10); en cuanto a la región espectral se observó una disminución en los coeficientes de determinación y un aumento de los errores al cambiar la región espectral (reducir el rango espectral). Al comparar los modelos utilizando SMOTE se pudo observar una mejora significativa de los coeficientes de determinación y los errores (de calibración y validación cruzada) para los modelos de PLS y PCR. El algoritmo SMOTE aumenta efectivamente la clase minoritaria generando muestras sintéticas que capturan las características de las muestras reales. Al cerrar la brecha entre muestras de clases minoritarias mejora el rendimiento de los modelos de aprendizaje automático que inicialmente fueron entrenados con un conjunto de datos desequilibrados (Nemade *et al*., 2023).

El mejor modelo para la predicción de proteína cruda fue el PLS-2da. derivada (5,2,2) con SMOTE y presentó $R^2$ de 0.965, 0.817 y 0.793 para la calibración, la validación cruzada y la predicción respectivamente (Figura 9).



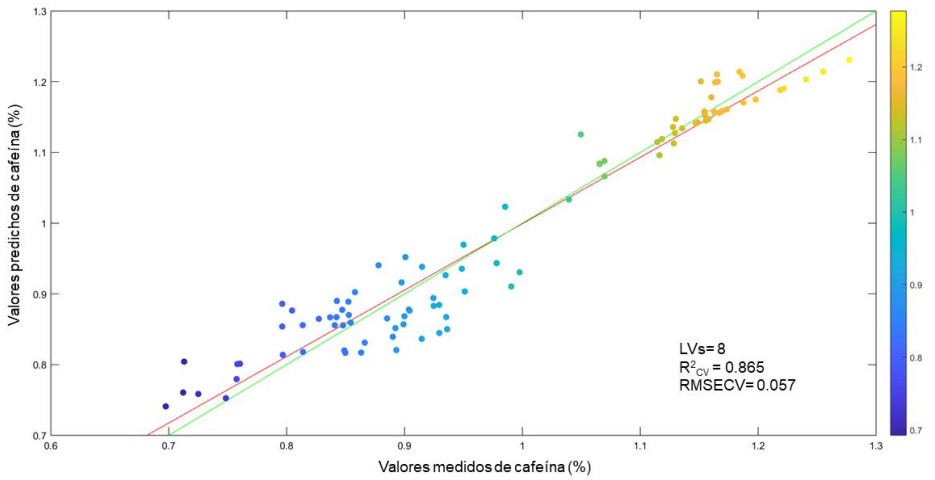




**Figura 9.Correlación entre los valores medidos y predichos para los mejores modelos de PLS construidos con el set de calibración cruzada.**

Según lo mencionado por Xu *et al*., (2018), estos coeficientes de determinación indican modelos con predicciones aproximadas. En cuanto a los errores asociados a este modelo se obtuvieron valores de 0.108, 0.249 y 0.199 para $RMSE_C$, $RMSE_{CV}$ y $RMSE_P$ respectivamente, además un RPD=1.86 y un REP= 1.45. Estos resultados indican que el modelo puede ser utilizado para predicciones aproximadas del contenido de proteína cruda en café verde especial.

Los resultados de este estudio fueron diferentes a los reportados por Zhu *et al*., (2021), quienes inicialmente construyeron el modelo PLS con segunda derivada para predicción de proteína cruda en café verde y obtuvieron un $R^2$ de 0.192 y 0.423 con errores de 0.673 y 0.485 para la validación cruzada y predicción respectivamente, además de un RPD=1.36. Sin embargo, aplicaron luego el algoritmo OSC a la data espectral y mejoraron el rendimiento del modelo obteniendo $R^2$ mayores a 0.965 y errores menores a 0.119 además de un RPD=5.54.

Esta comparación sugiere que, aunque el modelo obtenido en esta investigación permita predicciones aproximadas del contenido de proteína cruda en café verde aún requiere mejoras adicionales.

#### *4.2.4.3 Modelos predictivos para el contenido de cafeína*

Los resultados de los modelos predictivos para la estimación de cafeína (%) en café verde especial se muestran en la Tabla 10.

En función de los métodos de regresión utilizados, PLS mostró una ligera mejora en los coeficientes de determinación y los errores de la mayoría de los modelos (Figura 10). PLS y PCR se basan en la descomposición de la data espectral en componentes principales para PCR y variables latentes para PLS que capturan la mayor parte de la

varianza del espectro sin procesar y crean modelos lineales utilizando las puntuaciones de las características más correlacionadas. Sin embargo, a diferencia de PCR, el algoritmo PLS integra los pasos de compresión y selecciona factores ortogonales sucesivos que maximizan la covarianza entre las variables predictoras y de respuesta (Xu *et al*., 2018).

SMOTE mejoró ligeramente el coeficiente de determinación y los errores en los modelos construidos con PLS y PCR. Por otro lado, el cambio de la región espectral influyó negativamente en el ajuste de los modelos, lo que ocasionó disminución en los $R^2$ y aumentó en los errores utilizando ambos métodos de regresión multivariada; esta misma tendencia fue observada en los modelos para proteína cruda. La selección de variables (longitudes de onda), puede eliminar regiones no informativas del espectro, resolver problemas de colinealidad, superposición de bandas e interacciones; esta selección se puede realizar de manera manual a partir de conocimientos básicos sobre las propiedades espectroscópicas, pero se ha demostrado que utilizar estrategias matemáticas para la selección de variables es más eficiente (Chen *et al*., 2013).

El mejor modelo para estimar el contenido de cafeína en café verde especial PLS-2da. derivada (5,2,2) con SMOTE y presentó $R^2$ de 0.939, 0.865 y 0.822 para la calibración, la validación cruzada y la predicción respectivamente (Figura 9). En cuanto a los errores asociados a este modelo se obtuvieron valores de 0.038, 0.057 y 0.057 para $RMSE_C$, $RMSE_{CV}$ y $RMSE_P$ respectivamente, además un RPD=2.35 y un REP= 5.42. De acuerdo con estos resultados se considera que el modelo tiene buena capacidad predictiva.

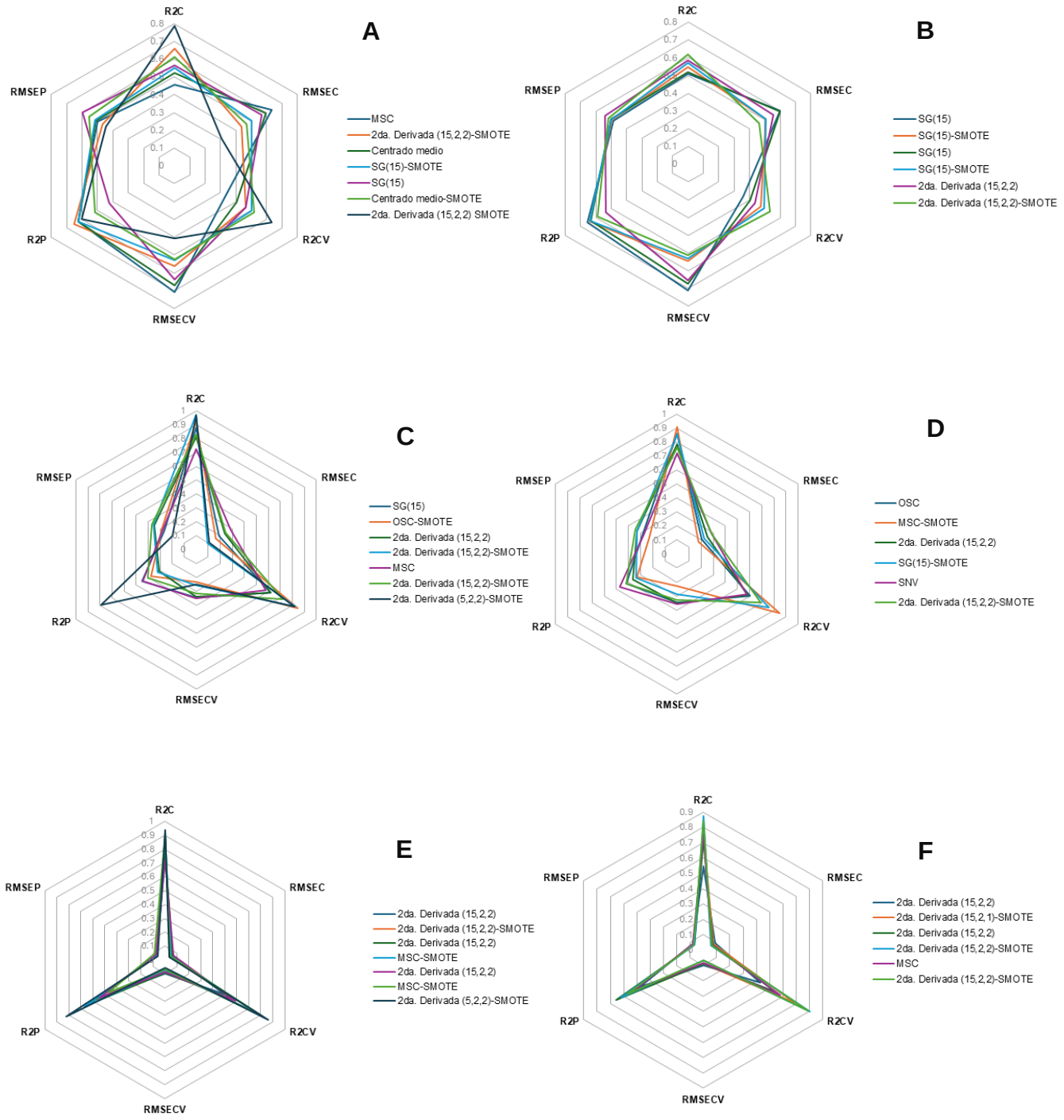



**Figura 10.Comparación de los modelos PCR y PLS: (A) modelo PLS grasa cruda (%); (B) modelo PCR grasa cruda (%); (C) modelo PLS proteína cruda (%); (D) modelo PCR proteína cruda; (E) modelo PLS cafeína (%); (F) modelo PCR cafeína (%).**

Los resultados de este estudio fueron diferentes a los reportados por Yuwita *et al*., (2023), quienes obtuvieron un modelo PLS-MSC donde los coeficientes de determinación fueron de 0.996 y 0.989 además de unos errores de 0.015 y 0.021 para validación cruzada y predicción respectivamente. El RPD reportado fue de 11.869, por lo que el modelo presentado por estos autores, indica una excelente capacidad de predecir el contenido de cafeína en granos de café.

Por otro lado, Ayu *et al*., (2020), construyeron un modelo de predicción para cafeína en café verde, utilizando un total de 240 muestras y adquiriendo el espectro en la región de 1000-2500nm. Reportaron que el mejor modelo de predicción fue obtenido aplicando PLS + MSC con un coeficiente de correlación (r) de 0.936 y un RPD de 2.15. Estos resultados fueron similares a los obtenidos en este estudio.

#### *4.2.4.4 Importancia de la variable dependiente en la proyección (VIP scores)*

Para eliminar efectos aditivos o multiplicativos por la dispersión de la luz, ruido de la línea base, deriva de la línea base en la data espectral se utilizan métodos de pretratamiento espectral, antes de la construcción de los modelos de predicción (Zhu, Long, Chen, *et al*., 2021).

En este estudio se ensayaron modelos utilizando métodos de pretratamiento como MSC, SNV, SG, OSC, primera y segunda derivada, obteniendo los mejores modelos de predicción para grasa cruda, proteína cruda y cafeína, utilizando la segunda derivada de segundo orden polinomial con suavizado de Savitsky-Golay (Figura 6C).

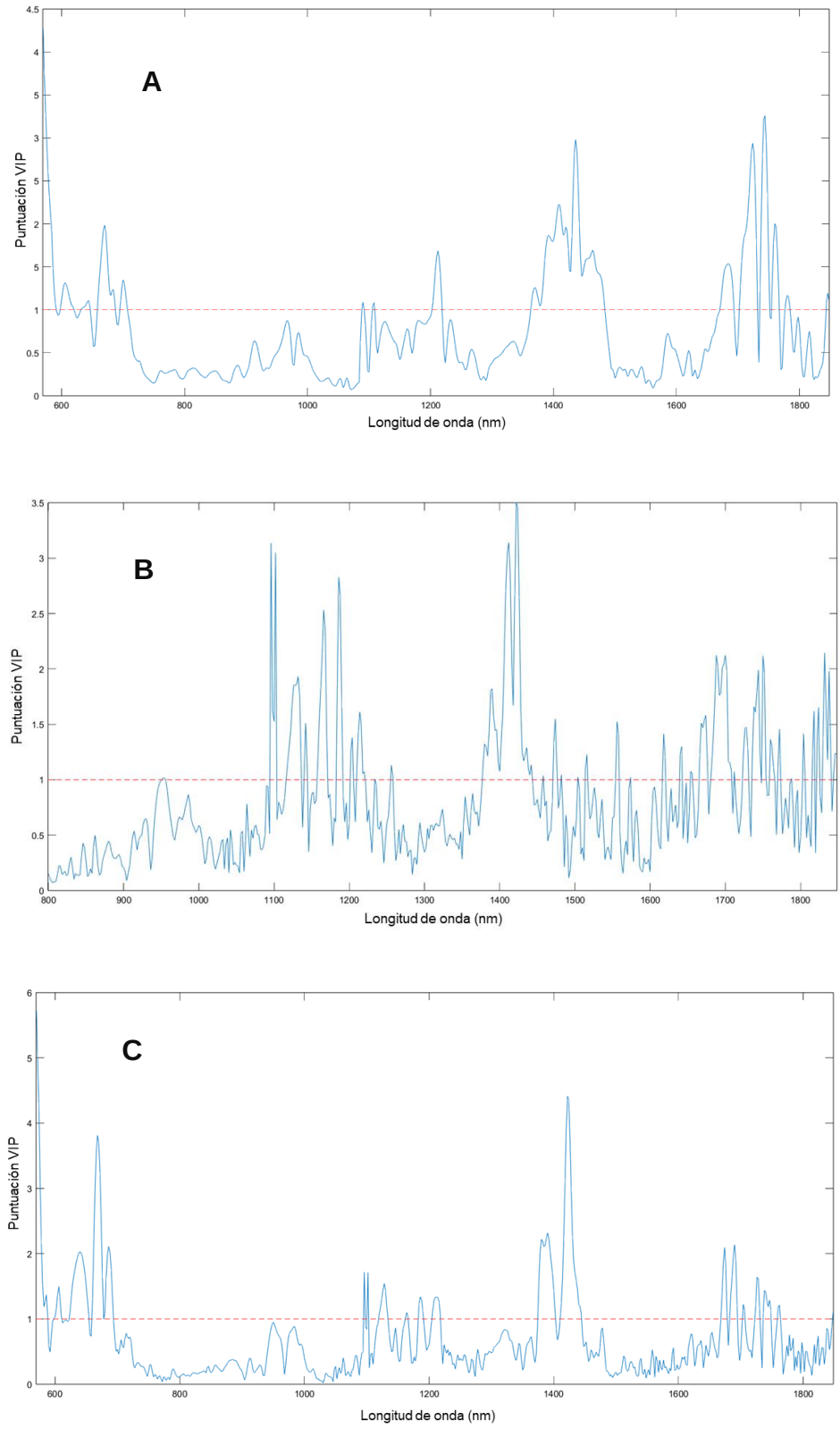



**Figura 11.Puntuación VIP: (A) Modelo de grasa cruda; (B) Modelos de proteína cruda y (C) Modelo de cafeína.**

La primera y la segunda derivada se utilizan para mejorar la resolución de los picos superpuestos y la corrección de la línea base, a expensas de una menor relación señal ruido (Araújo *et al*., 2020; Pizarro *et al*., 2004).

Para determinar qué longitudes de onda fueron las que aportaron a los modelos de PLS para estimar el contenido de grasa cruda, proteína cruda y cafeína en café verde especial se utilizó la importancia de la variable independiente en la proyección (VIP) que se muestran en la Figura 11.

Un gráfico de puntuación VIP permite evaluar la importancia de cada variable (longitud de onda) en la proyección utilizada en un modelo de PLS. Si una variable tiene una puntuación cercana o superior a uno puede considerarse importante en un modelo determinado, mientras que las variables con puntuaciones VIP significativamente menores a uno son menos importantes y podrían ser candidatas para ser excluidas del modelo (Hashimoto *et al*., 2018).

La Figura 11A correspondiente al modelo de predicción de grasa cruda, muestra que una región importante se encuentra entre 1700-1800 nm, que podría atribuirse al estiramiento C-H simétrico y asimétrico del primer sobretono (Daszykowski *et al*., 2008; Munyendo *et al*., 2022), otra región importante es la región entre 1400-1500 nm que está relacionada con bandas del primer sobretono correspondiente a CH, $H_2O$, ROH, RNH que se relacionan con lípidos y otros compuestos como las proteínas y cafeína que también están presentes en el café (Baqueta *et al*., 2020).

Los lípidos contribuyen al aroma y cuerpo del café, ya que durante el tostado se producen reacciones de descomposición y autooxidación (Barbosa *et al*., 2019; Zhu, *et al*., 2021).

Para el modelo de proteína cruda (Figura 11B) una región importante del espectro estuvo entre 1100-1200 nm que puede deberse a enlaces $CH_3$, $CH_2$, CH en la región del segundo sobretono, mientras que para la región entre 1650-1850 nm puede estar asociada a enlaces $CH_3$, $CH_2$, CH en la región del primer sobretono (Barbin *et al*., 2014).

Las proteínas participan durante las reacciones de Maillard en el proceso de tostado para  formando melanoidinas y compuestos volátiles de bajo peso molecular (Barbosa *et al*., 2019). Estas junto a otros compuestos influyen en el aroma, color (caramelización) y sabor del café (Mahmud *et al*., 2020; Zhu, Long, Ma, *et al*., 2021).

Por último, para el modelo de cafeína (figura 11C), la región entre 1390-1450 nm puede estar asociada a el enlace CH del grupo metilo ($CH_3$) del segundo sobretono; la cafeína cuenta con 3 grupos metilos en su estructura cíclica por lo que esa región espectral puede jugar un papel importante en la estimación de la cafeína (Munyendo *et al*., 2022; Sinija y Mishra, 2009). Otra región importante fue la comprendida entre 1650-1750 nm que podría estar asociada a estiramientos CH del primer sobretono (Wójcicki, 2022).

La cafeína es un compuesto termoestable que junto a la trigonelina y los ácidos clorogénicos son responsables del amargor (Barbosa *et al*., 2019; Mahmud *et al*., 2020) y precursores del aroma en el café (Sualeh *et al*., 2020).

Compuestos específicos en la región NIR no están bien determinados, como en la región MIR, debido a que las bandas están combinadas y superpuestas (Munyendo *et al*., 2022), por lo que la asignación de características moleculares se realiza de manera tentativa.

# CAPÍTULO 5.

# CONCLUSIONES

De acuerdo con los resultados de esta investigación se puede concluir lo siguiente:

1. El método de regresión multivariado que permitió obtener los mejores ajustes para estimar el contenido de grasa cruda, proteína cruda y cafeína, en café verde especial fue PLS.
2. La generación de muestras sintéticas permitió mejorar la capacidad predictiva de los modelos en general, mientras que la reducción de la región espectral desmejoró los modelos.
3. El pretratamiento espectral que mejor resultado aportó a los modelos fue la segunda derivada.
4. De acuerdo con las métricas de predicción evaluadas para los tres modelos, el modelo de cafeína se puede utilizar con confianza para estimar el contenido de cafeína en muestras de café verde especial, mientras que el de grasa y proteínas crudas requieren algunas adecuaciones, pero pueden ser utilizados a modo de screening.
5. Los resultados de este estudio confirman que la espectroscopía NIR acoplada a herramientas quimiométricas proveen un método rápido para determinar componentes químicos en café verde especial.
6. Para investigaciones futuras se recomienda incluir más muestras, principalmente muestras de la clase minoritaria, lo que permitirá obtener modelos más robustos.
7. Los modelos podrían mejorar su capacidad predictiva, al ampliar el rango espectral, ya que este estudio estuvo limitado a solo la región del espectro NIR entre 800-1850 nm.

8. El rendimiento del modelo se ve afectado por el tipo de preprocesamiento espectral que se utiliza, por lo que se pueden aplicar métodos diferentes a los aplicados en este estudio o incluso utilizar combinaciones de ellos.
9. También aplicar otros métodos multivariados como redes neuronales, máquina de soporte de vectores entre otros.

# CAPÍTULO 6.

# REFERENCIAS BIBLIOGRÁFICAS

A.E.F.I. (2001). *Validación de Métodos Analíticos* (pp. 1–315).

Abubakar, Y., Hasni, D., Muzaifa, M., Sulaiman, Mahdi, & Widayat, H. P. (2019). Effect of varieties and processing practices on the physical and sensory characteristics of Gayo Arabica specialty coffee. *IOP Conference Series: Materials Science and Engineering*, *523*(1), 0–9. https://doi.org/10.1088/1757-899X/523/1/012027

Adnan, A., Naumann, M., Morlein, D., & Pawelzik, E. (2020). 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*, *9*(6), 1–14. https://doi.org/10.3390/foods9060788

AOAC. (2010). *Official Methods of Analysis of AOAC International* (W. Horwitz & G. Latimer (eds.); 18th ed.).

AOCS. (2005). Rapid Determination of Oil/Fat Utilizing High Temperature solvent Extraction. *American Oil Chemists Society*. http://www.ssco.com.tw/Ankom/PDF_file/Crude Fat Method.pdf

Araújo, C. da S., Macedo, L. L., Vimercati, W. C., Ferreira, A., Prezotti, L. C., & Saraiva, S. H. (2020). Determination of pH and acidity in green coffee using near-infrared spectroscopy and multivariate regression. *Journal of the Science of Food and Agriculture*, *100*(6), 2488–2493. https://doi.org/10.1002/jsfa.10270

Ayu, P. C., Budiastra, I. W., & Rindang, A. (2020). NIR spectroscopy application for determination caffeine content of Arabica green bean coffee. *IOP Conference Series: Earth and Environmental Science*, *454*(1), 3–7. https://doi.org/10.1088/1755-

1315/454/1/012049

Baqueta, M. R., Coqueiro, A., Março, P. H., & Valderrama, P. (2020). Quality Control Parameters in the Roasted Coffee Industry: a Proposal by Using MicroNIR Spectroscopy and Multivariate Calibration. *Food Analytical Methods*, *13*(1), 50–60. https://doi.org/10.1007/s12161-019-01503-w

Baqueta, M. R., Coqueiro, A., & Valderrama, P. (2019). Brazilian Coffee Blends: A Simple and Fast Method by Near-Infrared Spectroscopy for the Determination of the Sensory Attributes Elicited in Professional Coffee Cupping. *Journal of Food Science*, *84*(6), 1247–1255. https://doi.org/10.1111/1750-3841.14617

Barbin, D. F., Felicio, A. L. de S. M., Sun, D. W., Nixdorf, S. L., & Hirooka, E. Y. (2014). Application of infrared spectral techniques on quality and compositional attributes of coffee: An overview. *Food Research International*, *61*, 23–32. https://doi.org/10.1016/j.foodres.2014.01.005

Barbosa, M. de S. G., Scholz, M. B. dos S., Kitzberger, C. S. G., & Benassi, M. de T. (2019). Correlation between the composition of green Arabica coffee beans and the sensory quality of coffee brews. *Food Chemistry*, *292*, 275–280. https://doi.org/10.1016/j.foodchem.2019.04.072

Bertone, E., Venturello, A., Giraudo, A., Pellegrino, G., & Geobaldo, F. (2016). Simultaneous determination by NIR spectroscopy of the roasting degree and Arabica/Robusta ratio in roasted and ground coffee. *Food Control*, *59*, 683–689. https://doi.org/10.1016/j.foodcont.2015.06.055

Brown, R. H., & Mueller-Harvey, I. (1999). Evaluation of the novel Soxflo technique for rapid extraction of crude fat in foods and animal feeds. *Journal of AOAC International*, *82*(6), 1369–1374. https://doi.org/10.1093/jaoac/82.6.1369

Buratti, S., Sinelli, N., Bertone, E., Venturello, A., Casiraghi, E., & Geobaldo, F. (2015). Discrimination between washed Arabica, natural Arabica and Robusta coffees by using near infrared spectroscopy, electronic nose and electronic tongue analysis. *Journal of the Science of Food and Agriculture*, *95*(11), 2192–2200. https://doi.org/10.1002/jsfa.6933

Caporaso, N., Whitworth, M. B., & Fisk, I. D. (2021). Total lipid prediction in single intact cocoa beans by hyperspectral chemical imaging. *Food Chemistry*, *344*, 128663. https://doi.org/10.1016/j.foodchem.2020.128663

Cardoso Santos, M., Martins Nascimento, P. A., Nascimento Guedes, W., RodriguesPereira-Filho, E., Filletti, É. R., & Verbi Pereira, F. M. (2019). Chemometrics in analytical chemistry - An overview of applications from 2014 to 2018. *Eclética Quimica*, *44*(2), 11–25. https://doi.org/10.26850/1678-4618eqj.v44.2.11-25

Çataltas, Ö., & Tütüncü, K. (2021). A Review of Data Analysis Techniques Used in Near-Infrared Spectroscopy. *European Journal of Science and Technology*, *25*, 475–484. https://doi.org/10.31590/ejosat.882749

Chakravarthy, A. D., Bonthu, S., Chen, Z., & Zhu, Q. (2019). Predictive models with resampling: A comparative study of machine learning algorithms and their performances on handling imbalanced datasets. *Proceedings - 18th IEEE*

*International Conference on Machine Learning and Applications, ICMLA 2019*, 1492–1495. https://doi.org/10.1109/ICMLA.2019.00245

Chen, J., Ren, X., Zhang, Q., Diao, X., & Shen, Q. (2013). Determination of protein, total carbohydrates and crude fat contents of foxtail millet using effective wavelengths in NIR spectroscopy. *Journal of Cereal Science*, *58*(2), 241–247. https://doi.org/10.1016/j.jcs.2013.07.002

Coates, J. (2000). Infrared spectral interpretation: A systematic approach. In J. W. & S. Ltd. (Ed.), *Encyclopedia of Analytical Chemistry* (pp. 10815–10837). https://doi.org/10.1002/9780470027318

Daszykowski, M., Wrobel, M. S., Czarnik-Matusewicz, H., & Walczak, B. (2008). Near-infrared reflectance spectroscopy and multivariate calibration techniques applied to modelling the crude protein, fibre and fat content in rapeseed meal. *Analyst*, *133*(11), 1523–1531. https://doi.org/10.1039/b803687j

Demir, S., & Şahin, E. K. (2022). Evaluation of Oversampling Methods (OVER, SMOTE, and ROSE) in Classifying Soil Liquefaction Dataset based on SVM, RF, and Naïve Bayes. *European Journal of Science and Technology*, *34*, 142–147. https://doi.org/10.31590/ejosat.1077867

Dong, W., Chen, Q., Wei, C., Hu, R., Long, Y., Zong, Y., & Chu, Z. (2021). Comparison of the effect of extraction methods on the quality of green coffee oil from Arabica coffee beans: Lipid yield, fatty acid composition, bioactive components, and antioxidant activity. *Ultrasonics Sonochemistry*, *74*, 105578. https://doi.org/10.1016/j.ultsonch.2021.105578

dos Santos, H. D., & Boffo, E. F. (2021). Coffee beyond the cup: analytical techniques used in chemical composition research—a review. *European Food Research and Technology*, *247*(4), 749–775. https://doi.org/10.1007/s00217-020-03679-6

Dupuy, N., Galtier, O., Ollivier, D., Vanloot, P., & Artaud, J. (2010). Comparison between NIR, MIR, concatenated NIR and MIR analysis and hierarchical PLS model . Application to virgin olive oil analysis. *Analytica Chimica Acta*, *666*, 23–31. https://doi.org/10.1016/j.aca.2010.03.034

Franca, A. S., Mendonça, J. C. F., & Oliveira, S. D. (2005). Composition of green and roasted coffees of different cup qualities. *LWT - Food Science and Technology*, *38*(7), 709–715. https://doi.org/10.1016/j.lwt.2004.08.014

Gautam, R., Vanga, S., Ariese, F., & Umapathy, S. (2015). Review of multidimensional data processing approaches for Raman and infrared spectroscopy. *EPJ Techniques and Instrumentation*, *2*(1). https://doi.org/10.1140/epjti/s40485-015-0018-6

Giraudo, A., Grassi, S., Savorani, F., Gavoci, G., Casiraghi, E., & Geobaldo, F. (2019). Determination of the geographical origin of green coffee beans using NIR spectroscopy and multivariate data analysis. *Food Control*, *99*, 137–145. https://doi.org/10.1016/j.foodcont.2018.12.033

Gumecindo-Alejo, A. L., Sánchez-Landero, L. A., Ortiz-Ceballos, G. C., Cerdán-Cabrera, C. R., & Alvarado-Castillo, G. (2021). Factors related to coffee quality, based on the “cup of excellence” contest in Mexico. *Coffee Science*, *16*. https://doi.org/10.25186/.v16i.1887

Haile, M., & Kang, W. H. (2019). The Role of Microbes in Coffee Fermentation and Their Impact on Coffee Quality. *Journal of Food Quality*, *2019*. https://doi.org/10.1155/2019/4836709

Hashimoto, J. C., Lima, J. C., Celeghini, R. M. S., Nogueira, A. B., Efraim, P., Poppi, R. J., & Pallone, J. A. L. (2018). Quality Control of Commercial Cocoa Beans (Theobroma cacao L.) by Near-infrared Spectroscopy. *Food Analytical Methods*, *11*(5), 1510–1517. https://doi.org/10.1007/s12161-017-1137-2

Heckman, M. A., Weil, J., & de Mejia, E. G. (2010). Caffeine (1, 3, 7-trimethylxanthine) in foods: A comprehensive review on consumption, functionality, safety, and regulatory matters. *Journal of Food Science*, *75*(3), 77–87. https://doi.org/10.1111/j.1750-3841.2010.01561.x

Helin, R., Indahl, U. G., Tomic, O., & Liland, K. H. (2022). On the possible benefits of deep learning for spectral preprocessing. *Journal of Chemometrics*, *36*(2), 1–19. https://doi.org/10.1002/cem.3374

Instituto Nacional de Estadística y Censo. (2023). *COSECHA DE CAFÉ EN LAS PRINCIPALES PROVINCIAS Y COMARCA INDÍGENA DE LA REPÚBLICA: AÑOS AGRÍCOLAS 2018/19 A 2022/23*. https://www.inec.gob.pa/publicaciones/Default3.aspx?ID_PUBLICACION=1204&ID_CATEGORIA=4&ID_SUBCATEGORIA=12

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis : a review and recent developments. *Philosophical Transactions A*, *374*. https://doi.org/http://dx.doi.org/10.1098/rsta.2015.0202

Kuhn, & Max. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Laukaleja, I., & Kruma, Z. (2018). Quality of Specialty Coffee: Balance between aroma, flavour and biologically active compound composition: Review. *Research for Rural Development*, *1*(December), 240–247. https://doi.org/10.22616/rrd.24.2018.038

Liu, C., Yang, N., Yang, Q., Ayed, C., Linforth, R., & Fisk, I. D. (2019). Enhancing Robusta coffee aroma by modifying flavour precursors in the green coffee bean. *Food Chemistry*, *281*, 8–17. https://doi.org/10.1016/j.foodchem.2018.12.080

Mahmud, M. M. C., Shellie, R. A., & Keast, R. (2020). Unravelling the relationship between aroma compounds and consumer acceptance: Coffee as an example. *Comprehensive Reviews in Food Science and Food Safety*, *19*(5), 2380–2420. https://doi.org/10.1111/1541-4337.12595

Manuel, M. N. B., da Silva, A. C., Lopes, G. S., & Ribeiro, L. P. D. (2022). One-class classification of special agroforestry Brazilian coffee using NIR spectrometry and chemometric tools. *Food Chemistry*, *366*(May 2021), 130480. https://doi.org/10.1016/j.foodchem.2021.130480

Mendes, G. de A., de Oliveira, M. A. L., Rodarte, M. P., de Carvalho dos Anjos, V., & Bell, M. J. V. (2022). Origin geographical classification of green coffee beans (Coffea arabica L.) produced in different regions of the Minas Gerais state by FT-MIR and chemometric. *Current Research in Food Science*, *5*(November 2021), 298–305. https://doi.org/10.1016/j.crfs.2022.01.017

Munyendo, L., Njoroge, D., & Hitzmann, B. (2022). The Potential of Spectroscopic Techniques in Coffee Analysis—A Review. *Processes*, *10*(1), 1–25. https://doi.org/10.3390/pr10010071

Naegele, E. (2016). *Determination of Caffeine in Coffee Products According to DIN 20481* (pp. 4–9). https://www.agilent.com/cs/library/applications/5991-2851EN.pdf

Nemade, B., Bharadi, V., Alegavi, S. S., & Marakarkandy, B. (2023). A Comprehensive Review: SMOTE-Based Oversampling Methods for Imbalanced Classification Techniques, Evaluation, and Result Comparisons. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(9s), 790–803.

Nguyen Minh, Q., Lai, Q. D., Nguy Minh, H., Tran Kieu, M. T., Lam Gia, N., Le, U., Hang, M. P., Nguyen, H. D., Chau, T. D. A., & Doan, N. T. T. (2022). Authenticity green coffee bean species and geographical origin using near-infrared spectroscopy combined with chemometrics. *International Journal of Food Science & Technology*, *57*(7), 4507–4517. https://doi.org/https://doi.org/10.1111/ijfs.15786

Núñez, N., Collado, X., Martínez, C., Saurina, J., & Núñez, O. (2020). Authentication of the origin, variety and roasting degree of coffee samples by non-targeted HPLC-UV fingerprinting and chemometrics. Application to the detection and quantitation of adulterated coffee samples. *Foods*, *9*(3), 1–14. https://doi.org/10.3390/foods9030378

Okazaki, Y. (2012). Near-Infrared Spectroscopy—Its Versatility in Analytical. *Anal. Chem*, *28*(June), 545–562.

Okubo, N., & Kurata, Y. (2019). Nondestructive classification analysis of green coffee beans by using near-infrared spectroscopy. *Foods*, *8*(2). https://doi.org/10.3390/foods8020082

Olivieri, A., & Escandar, G. M. (2014). *Practical three-way calibration* (Elsevier (ed.)). Elsevier.

Pasquini, C. (2003). Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications. *Journal of the Brazilian Chemical Society*, *14*(2), 198–219. https://doi.org/10.1590/S0103-50532003000200006

Pasquini, C. (2018). Near infrared spectroscopy: A mature analytical technique with new perspectives – A review. *Analytica Chimica Acta*, *1026*, 8–36. https://doi.org/10.1016/j.aca.2018.04.004

Pimpley, V., Patil, S., Srinivasan, K., Desai, N., & Murthy, P. S. (2020). The chemistry of chlorogenic acid from green coffee and its role in attenuation of obesity and diabetes. *Preparative Biochemistry and Biotechnology*, *50*(10), 969–978. https://doi.org/10.1080/10826068.2020.1786699

Pizarro, C., Esteban-Díez, I., Nistal, A.-J., & González-Sáiz, J.-M. (2004). Influence of data pre-processing on the quantitative determination of the ash content and lipids in roasted coffee by near infrared spectroscopy. *Analytica Chimica Acta*, *509*, 217–227. https://doi.org/10.1016/j.aca.2003.11.008

R Core Team. (2023). *{R}: A Language and Environment for Statistical Computing*. https://www.r-project.org/

Rangel, F., Sáez, E., Henry, A., Cáceres-Hernández, D., & Galán, J. S. (2021). Using Visible/Near-Infrared Reflectance Spectroscopy and Chemometrics for the Rapid Evaluation of Two Panamanian Watermelon (Citrullus lanatus) Varieties. *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, 1–6. https://doi.org/10.1109/ISIE45552.2021.9576169

Rinnan, Å., Berg, F. van den, & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TrAC - Trends in Analytical Chemistry*, *28*(10), 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007

Rodionova, O., Kucheryavskiy, S., & Pomerantsev, A. (2021). Efficient tools for principal component analysis of complex data— a tutorial. *Chemometrics and Intelligent Laboratory Systems*, *213*, 104304. https://doi.org/10.1016/j.chemolab.2021.104304

Saeys, W., Nguyen Do Trong, N., Van Beers, R., & Nicolaï, B. M. (2019). Multivariate calibration of spectroscopic sensors for postharvest quality evaluation: A review. *Postharvest Biology and Technology*, *158*, 110981. https://doi.org/10.1016/j.postharvbio.2019.110981

Sahfitri, Rindang, A., & Ayu, P. C. (2020). Prediction of water content in Lintong green bean coffee using FT-NIRS and PLS method. *IOP Conference Series: Earth and Environmental Science*, *454*(1). https://doi.org/10.1088/1755-1315/454/1/012047

Sánchez-Galán, J. E., Barranco, F. R., Reyes, J. S., Quirós-McIntire, E. I., Jiménez, J. U., & Fábrega, J. R. (2021). Using Supervised Classification Methods for the Analysis of Multi-spectral Signatures of Rice Varieties in Panama. *Advances in Science, Technology and Engineering Systems Journal*, *6*(2), 552–558.

https://doi.org/10.25046/aj060262

Santamaría, N., Meléndez, F., Arroyo, P., Calvo, P., Sánchez, F., Lozano, J., & Sánchez, R. (2023). Olfactory Evaluation of Geisha Coffee from Panama Using Electronic Nose. *Chemosensors*, *11*(11), 559. https://doi.org/10.3390/chemosensors11110559

Saud, S., & Salamatullah, A. M. (2021). Relationship between the chemical composition and the biological functions of coffee. *Molecules*, *26*(24), 1–14. https://doi.org/10.3390/molecules26247634

SCAP, S. C. A. of P. (2023). *Specialty Coffee Association of Panamá*. http://scap-panama.com/

Selamat, N. A., Abdullah, A., & Mat Diah, N. (2022). Association features of smote and rose for drug addiction relapse risk. *Journal of King Saud University - Computer and Information Sciences*, *34*(9), 7710–7719. https://doi.org/10.1016/j.jksuci.2022.06.012

Sim, J., Dixit, Y., Mcgoverin, C., Oey, I., Frew, R., Reis, M. M., & Kebede, B. (2024). Machine learning-driven hyperspectral imaging for non-destructive origin verification of green coffee beans across continents, countries, and regions. *Food Control*, *156*(October 2023), 110159. https://doi.org/10.1016/j.foodcont.2023.110159

Sinija, V. R., & Mishra, H. N. (2009). FT-NIR spectroscopy for caffeine estimation in instant green tea powder and granules. *Lwt*, *42*(5), 998–1002. https://doi.org/10.1016/j.lwt.2008.12.013

Siregar, S. D., Rindang, A., & Ayu, P. C. (2020). Principle Component Analysis (PCA) -

Classification of Arabica green bean coffee of North Sumatera Using FT-NIRS. *IOP Conference Series: Earth and Environmental Science*, *454*(1). https://doi.org/10.1088/1755-1315/454/1/012046

Siriseriwan, W. (2019). *smotefamily: A Collection of Oversampling Techniques for Class Imbalance Problem Based on SMOTE*. https://cran.r-project.org/package=smotefamily

Sualeh, A., Tolessa, K., & Mohammed, A. (2020). Biochemical composition of green and roasted coffee beans and their association with coffee quality from different districts of southwest Ethiopia. *Heliyon*, *6*(12). https://doi.org/10.1016/j.heliyon.2020.e05812

Sun, Y., Liu, N., Kang, X., Zhao, Y., Cao, R., Ning, J., Ding, H., Sheng, X., & Zhou, D. (2021). Rapid identification of geographical origin of sea cucumbers Apostichopus japonicus using FT-NIR coupled with light gradient boosting machine. *Food Control*, *124*, 107883. https://doi.org/10.1016/j.foodcont.2021.107883

Torres Castillo, N. E., Melchor-Martínez, E. M., Ochoa Sierra, J. S., Ramirez-Mendoza, R. A., Parra-Saldívar, R., & Iqbal, H. M. N. (2020). Impact of climate change and early development of coffee rust – An overview of control strategies to preserve organic cultivars in Mexico. *Science of the Total Environment*, *738*, 140225. https://doi.org/10.1016/j.scitotenv.2020.140225

Tsegay, G., Redi-Abshiro, M., Chandravanshi, B. S., Ele, E., Mohammed, A. M., & Mamo, H. (2020). Effect of altitude of coffee plants on the composition of fatty acids of green coffee beans. *BMC Chemistry*, *14*(1), 1–11. https://doi.org/10.1186/s13065-020-00688-0

van der Vossen, H., Bertrand, B., & Charrier, A. (2015). Next generation variety development for sustainable production of arabica coffee (Coffea arabica L.): a review. *Euphytica*, *204*(2), 243–256. https://doi.org/10.1007/s10681-015-1398-z

Vega, A., León, J. A. De, Reyes, S. M., & Gallardo, J. M. (2021). Modelo matemático para determinar la correlación entre parámetros fisicoquímicos y la calidad sensorial de café Geisha y Pacamara de Panamá. *Información Tecnológica*, *32*(1), 89–100. https://doi.org/10.4067/s0718-07642021000100089

Wójcicki, K. (2022). Near-infrared spectroscopy as a green technology to monitor coffee roasting. *Foods and Raw Materials*, *10*(2), 295–303. https://doi.org/10.21603/2308-4057-2022-2-536

Xu, S., Zhao, Y., Wang, M., & Shi, X. (2018). Comparison of multivariate methods for estimating selected soil properties from intact soil cores of paddy fields by Vis–NIR spectroscopy. *Geoderma*, *310*, 29–43. https://doi.org/10.1016/j.geoderma.2017.09.013

Yeager, S. E., Batali, M. E., Guinard, J. X., & Ristenpart, W. D. (2023). Acids in coffee: A review of sensory measurements and meta-analysis of chemical composition. *Critical Reviews in Food Science and Nutrition*, *63*(8), 1010–1036. https://doi.org/10.1080/10408398.2021.1957767

Yuwita, F., Ifmalinda, I., & Makky, M. (2023). Prediction of Caffeine and Protein of Arabica Coffee Beans Using Near Infrared Spectroscopy (NIRS). *Jurnal Teknik Pertanian Lampung (Journal of Agricultural Engineering)*, *12*(4), 852. https://doi.org/10.23960/jtep-l.v12i4.852-862

Zareef, M., Chen, Q., Hassan, M. M., Arslan, M., Hashim, M. M., Ahmad, W., Kutsanedzie, F. Y. H., & Agyekum, A. A. (2020). An Overview on the Applications of Typical Non-linear Algorithms Coupled With NIR Spectroscopy in Food Analysis. *Food Engineering Reviews*, *12*(2), 173–190. https://doi.org/10.1007/s12393-020-09210-7

Zhang, W., Kasun, L. C., Wang, Q. J., Zheng, Y., & Lin, Z. (2022). A Review of Machine Learning for Near-Infrared Spectroscopy. *Sensors*, *22*(24), 1–32. https://doi.org/10.3390/s22249764

Zhu, M., Long, Y., Chen, Y., Huang, Y., Tang, L., Gan, B., Yu, Q., & Xie, J. (2021). Fast determnation of lipid and protein content in green coffee beans from different origins using NIR spectroscopy and chemometrics. *Journal of Food Composition and Analysis*, *102*, 1–8.

Zhu, M., Long, Y., Ma, Y., Chen, Y., Yu, Q., Xie, J., Li, B., & Tian, J. (2021). Comparison of chemical and fatty acid composition of green coffee bean (Coffea arabica L.) from different geographical origins. *Lwt*, *140*. https://doi.org/10.1016/j.lwt.2020.110802

# CAPÍTULO 7.

# ANEXOS

## 7.1 Anexo I


**Figura A1. Muestras de café verde especial molidas.**

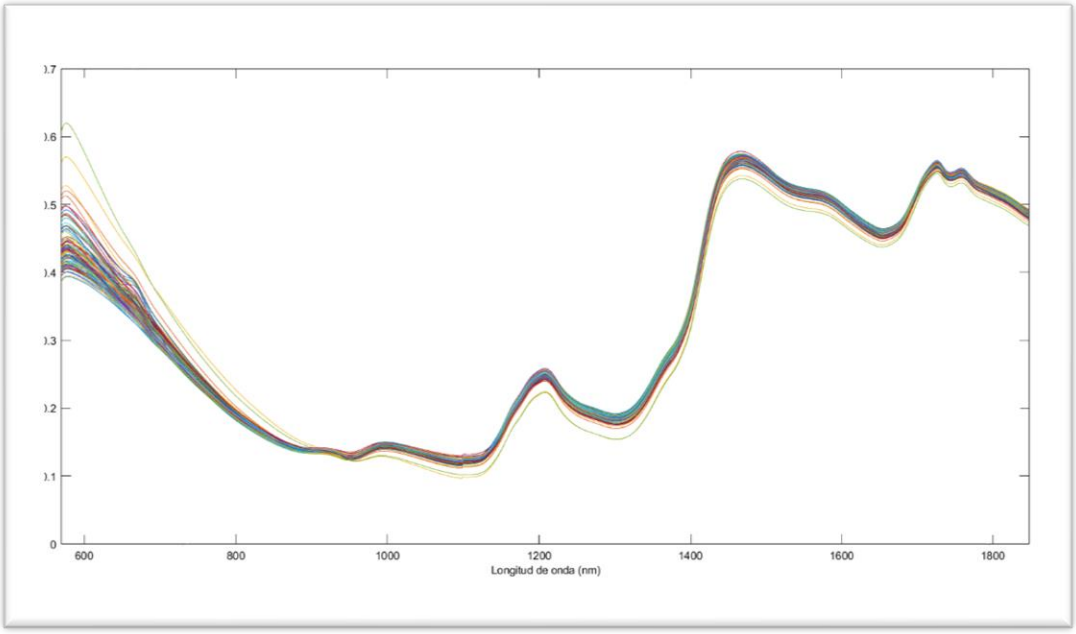



**Figura A2. Espectro café verde con pretratamiento MSC.**

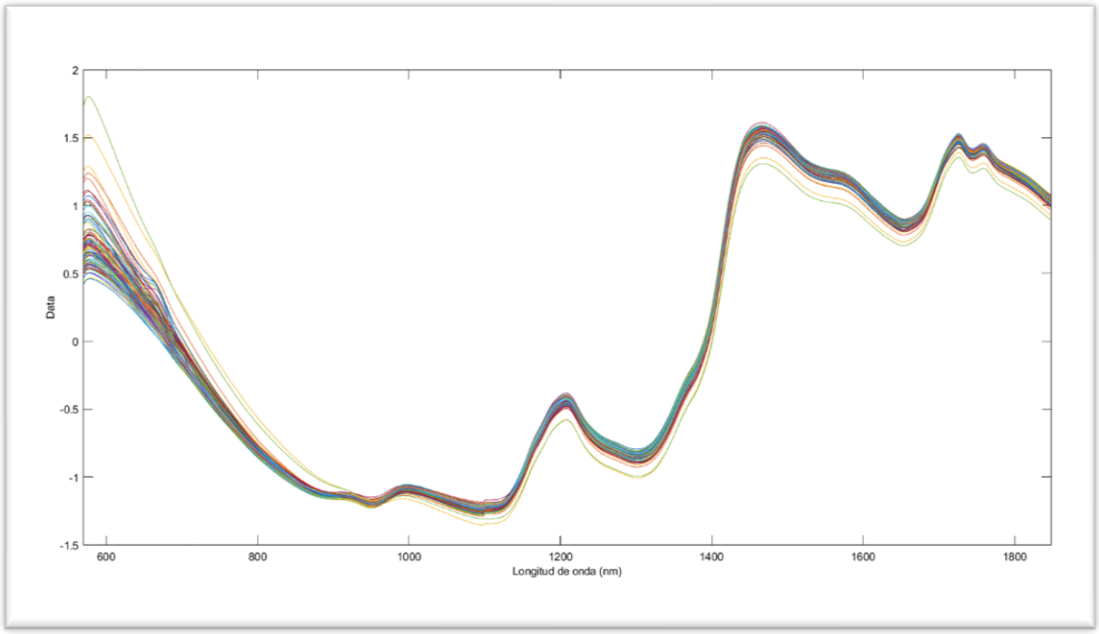



**Figura A3. Espectro café verde con pretratamiento SNV.**

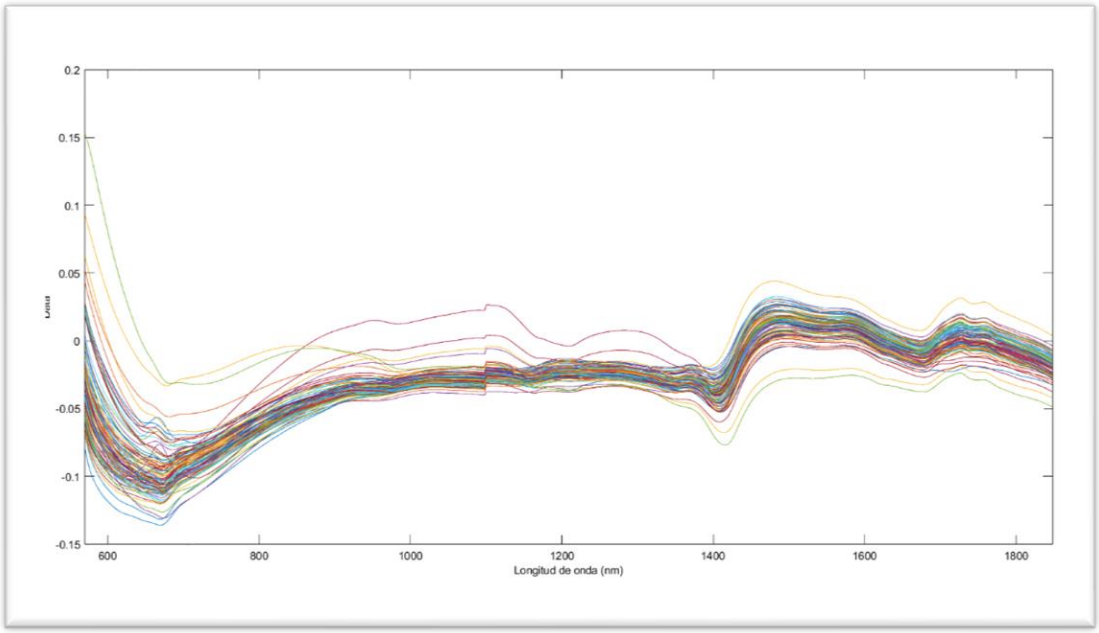



**Figura A4. Espectro café verde con pretratamiento OSC.**

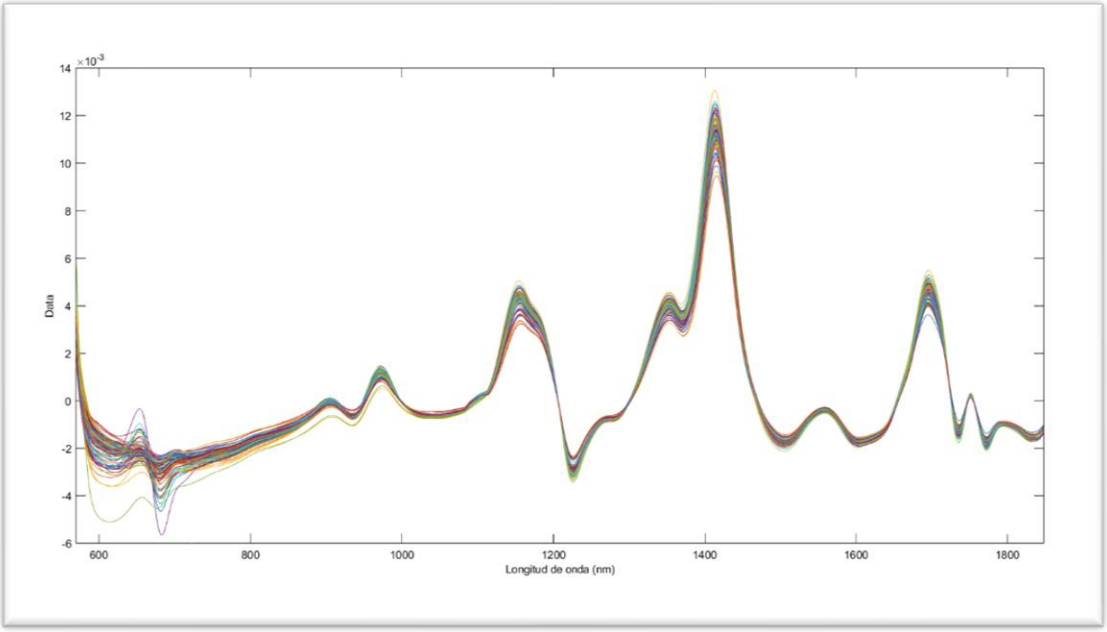



**Figura A5. Espectro café verde pretratamiento primera derivada.**

## 7.2 Anexo II

### Código de R

***#Activando librerías***

*library(openxlsx)*

*library(readr)*

*library(readxl)*

*library(stats)*

*library(corrplot)*

*library(ggplot2)*

*library(lattice)*

*library(caret)*

*library(dplyr)*

*library(RColorBrewer)*

*library(smotefamily)*

***#cambiando el nombre de los datos***

*df=Data_Variables_químicas*

***#Confirmación de estructura de los datos***

*str(df)*

***#Coersionando variable Variedad a factor***

*df$Variedad=factor(df$Variedad)*

***#Confirmación transformación de variable Variedad a factor***

*str(df)*

***#Histograma***

*hist(df$`Proteína Cruda`, main= "Histograma de Proteína Cruda (%)",xlab= "Proteína Cruda (%)", col= "lightblue" )*

*hist(df$`Grasa Cruda`, main= "Histogramas de Grasa Cruda (%)", xlab= "Grasa Cruda (%)",col="lightgreen" )*

*hist(df$`Cafeína`, main= "Histograma de Cafeína(%)", xlab= "Cafeína (%)", col="lightpink")*

***#Gráfico de caja y bigote por Variedad***

*boxplot(df$`Proteína Cruda` ~ df$Variedad, horizontal = FALSE, col= "lightblue", xlab = "Clase", ylab = "Proteína Cruda (%)")*

*boxplot(df$`Grasa Cruda`~ df$Variedad, horizontal = FALSE, col="lightgreen", xlab = "Clase", ylab = "Grasa Cruda (%)")*

*boxplot(df$`Cafeína` ~ df$Variedad, horizontal = FALSE, col="lightpink", xlab = "Clase", ylab = "Cafeína (%)")*

***#Inspeccionando correlación entre variables químicas***

*matcorr <-as.dist( round( cor( df[,3:ncol(df)]),3) )*

*matcorr)*

***#Importando data sin valores atípicos***

*Espectros_y_variables_completo_SO <- read_excel("C:/Users/Stephany Reyes/OneDrive/Escritorio/Tesis Final/Modelos/Espectros y variables completo SO.xlsx")*

*Espectros_y_variables_completo_cafeínaSO <- read_excel("Nuevo Ensayo/Cafeína/Espectros y variables completo-cafeínaSO.xlsx")*

```
df1=Espectros_y_variables_completo_cafeínaSO

#Coersionando variable Variedad

df1$Variedad= as.factor(df1$Variedad)

str(df1)

table(df1$Variedad)

barplot(table(df1$Variedad), xlab="Variedad", ylab ="Número de muestras", col= c("lightblue","lightpink"))

#Renombrando para partición

specud=df1


#Partición Entrenamiento (80%) y Predicción (20%)

set.seed(123)

training.muestra <-specud$`Grasa Cruda` %>%

  createDataPartition(p=.80, list = FALSE)

train.datacaf <-specud[training.muestra, ]

test.datacaf <-specud[-training.muestra, ]

#Guardando en formato CSV

write.csv(t(train.datacaf), file = "Traindatacaf.csv")

write.csv(t(test.datacaf), file = "Testdatacaf.csv")

#Importando data para SMOTE (80%)

Traindata80_para_SMOTE <- read_excel("C:/Users/Stephany Reyes/OneDrive/Escritorio/Tesis Final/Modelos/Muestras sintéticas/Traindata80 para SMOTE.xlsx")

Traindata2=Traindata80_para_SMOTE

str(Traindata2)

Traindata2$Variedad= as.factor(Traindata2$Variedad)

str(Traindata2)

table(Traindata2$Variedad)

barplot(table(Traindata2$Variedad),xlab="Variedad", ylab ="Número de muestras", col= c("lightblue","lightpink"))

#Coersionando la variable Variedad a numérica

Traindata2$Variedad= as.numeric(Traindata2$Variedad)

str(Traindata2)
```

***#SMOTE=Synthetic Minority Oversampling Technique. Create synthetic data by oversampling minority class only.***

*SMOTE1=SMOTE(Traindata2[,2:ncol(Traindata2)], as.numeric(Traindata2$Variedad))*

*smote1=SMOTE1$data*

*table(smote1$class)*

*barplot(table(smote1$class),xlab="Variedad", ylab ="Número de muestras", col= c("lightblue","lightpink"))*

***#Guardando en formato CSV***

*write.csv(t(smote1), file = "Traindatasmote105.csv")*

Universidad Autónoma de Chiriquí
Vicerrectoría Académica
Sistema de Bibliotecas e Información

Fecha: 28/2/24

Facultad: VIP

Escuela: Maestría

Se certifica que, tras llevar a cabo el proceso de análisis de originalidad y detección de similitudes en el trabajo de investigación titulado Aplicación de la espectroscopía y herramientas quimiométricas, presentado por el/la estudiante Stephany M. Reyes E. con número de cédula N° 4-745-14 con la asesoría del profesor DRA. Araelly Vega; el trabajo cumple con el 91% de originalidad, de acuerdo al informe emitido por el profesor asesor.

Es importante señalar que el proceso de análisis de plagio se ha realizado utilizando la herramienta Turnitin y siguiendo procedimientos estandarizados para asegurar la precisión de los resultados.

*Nota: El uso de la herramienta Turnitin fue aprobada por el Consejo Académico #5 - Sesión extraordinaria - 22 de mayo de 2023 y modificada el 6 de octubre de 2023.*

Eibar Amaya
Responsable de departamentco

Ada Chávez
Directora del
SIBIUNACHI

Panamá, 22 de febrero de 2024

Señores,

Universidad Autónoma de Chiriquí

E. S. M.

Estimados señores,

Yo Migdalia de Rodríguez con cédula de identidad personal 6-50-2384, Licenciada idónea de Español, certifico que el trabajo de Stephany M. Reyes C., con cédula de identidad personal 4-745-14, titulado: Aplicación de la espectroscopía Nir y herramientas quimiométricas para la determinación de componentes químicos del café verde especial producido en la Provincia de Chiriquí, Panamá.

Cumple con los requisitos de Ortografía, Redacción y Sintaxis, que debe reunir el mismo.

Código de idoneidad: 120435

Atentamente,

Migdalia de Rodríguez.

Licenciada en Español

**Adjunto diploma y copia de cédula**


REPÚBLICA DE PANAMÁ
TRIBUNAL ELECTORAL
Migdalia
Diaz Santamaria de Rodriguez
NOMBRE USUAL:
FECHA DE NACIMIENTO: 01-JUL-1960
LUGAR DE NACIMIENTO: HERRERA,OCÚ
SEXO: F
TIPO DE SANGRE:
EXPEDIDA: 28-SEP-2022
EXPIRA: 28-SEP-2037
6-50-2384
Migdalia de Rodríguez


TE TRIBUNAL ELECTORAL
6-50-2384
A7186N80199

# UNIVERSIDAD DE PANAMÁ

## LA FACULTAD DE

## Ciencias De La Educación

EN VIRTUD DE LA POTESTAD QUE LE CONFIEREN LA LEY Y EL ESTATUTO UNIVERSITARIO,
HACE CONSTAR QUE

## Migdalia Díaz S.

HA TERMINADO LOS ESTUDIOS Y CUMPLIDO CON LOS REQUISITOS
QUE LE HACEN ACREEDOR AL TÍTULO DE

## Profesora de Educación Media con Especialización en Español

Y EN CONSECUENCIA SE LE CONCEDE TAL GRADO CON TODOS LOS DERECHOS,
HONORES Y PRIVILEGIOS RESPECTIVOS, EN TESTIMONIO DE LO CUAL SE LE EXPIDE
ESTE DIPLOMA EN LA CIUDAD DE PANAMÁ, A LOS diecisiete
DÍAS DEL MES DE septiembre DEL AÑO DOS MIL cuatro.

Secretario General

Diploma 120435

Identificación Personal
6-80-2584

Decano

Rector

REPUBLICA DE PANAMA
MINISTERIO DE EDUCACION
Dirección Regional de Panamá Centro
Panamá, 18 de noviembre de 2004
Nombre del Dueño del Diploma: .....
MIGDALIA DIAZ S.
Folio 2996 Bajo el No. 148811
.....
Oficial de Registro