# UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

**INSTITUTO DE INGENIERÍA**

***MAESTRÍA Y DOCTORADO EN CIENCIAS E INGENIERÍA***

**"ESTUDIO DE LA COMPOSICIÓN NUTRICIONAL DE UN ALIMENTO POR ESPECTROSCOPÍA DE INFRARROJO CERCANO CON TRANSFORMADA DE FOURIER (FT-NIRS) Y QUIMIOMETRÍA"**

**TESIS PARA OBTENER EL GRADO DE:**
**MAESTRO EN CIENCIAS**

**PRESENTA**
**OLIVIA FLORES PEÑALOZA**

**DIRECTOR**
**DRA. MÓNICA CARRILLO BELTRÁN**

**Mexicali, B. C.** **Mayo de 2017**

# CONTENIDO

**ÍNDICE DE FIGURAS:**

**ÍNDICE DE TABLAS:**

Para mi esposo Benjamín, por su amor y apoyo incondicional. A mi hijo Benji, esperando le sirva como testimonio de superación personal y para mi hija Isabella, que me cuida desde el cielo.

# Agradecimientos


## *A mi tutora*

Gracias Dra. Mónica Carrillo Beltrán por su gran apoyo, motivación, guía y tiempo compartido durante el desarrollo del proyecto. Sus conocimientos y profesionalismo han sido esenciales para mi formación como investigador. Además, agradezco su comprensión y consejos que recibí al convertirme en madre y cuando atravesé situaciones difíciles de salud y familiares. Me siento en deuda con usted y espero poder probar que su dedicación hacia mi persona no ha sido en vano.

## *A mis maestros*

Por sus asignaturas impartidas, me han ayudado de manera directa o indirecta en el desarrollo del proyecto y escritura de la tesis. Especialmente al Dr. José Luis Arcos por su orientación, así como al Dr. Benjamín Valdés y a la Dra. Gisela Montero por permitirnos trabajar en su laboratorio.

*"A veces, el replanteamiento de un problema es más decisivo que el hallazgo de la solución, que puede ser un puro asunto de habilidad matemática o experimental. La capacidad de suscitar nuevas cuestiones, nuevas posibilidades de mirar viejos problemas, requiere una imaginación creativa y determina los avances científicos auténticos".*

*Albert Einstein*

# 1

# Introducción

Los análisis de alimentos se realizan de manera rutinaria para garantizar su calidad e inocuidad. Para lograr que cumplan con los estándares establecidos por las regulaciones gubernamentales, además de examinar el producto final se evalúan la materia prima y se monitorean los cambios durante el proceso de producción (Nielsen, 2010).

Dentro de los parámetros de calidad solicitados está el determinar la composición química del alimento cuyos constituyentes principales son los carbohidratos, lípidos, proteínas, vitaminas, minerales y agua, también conocidos como nutrientes. La información resultante se puede utilizar para distintos propósitos como determinar si se alteraron las propiedades físicas y químicas del producto (Cheung y Mehta, 2015) así como el establecer el valor nutricional, dato que les interesa a muchos

consumidores al momento de elegir un alimento sobre otro por motivos de salud, y que se reporta por medio de la etiqueta nutrimental.

A los métodos empleados para cuantificar a los nutrientes se le conoce como análisis proximales. Debido a que el alimento se considera un sistema complejo, para evaluar un componente en particular se requiere de múltiples técnicas analíticas cuyo desempeño dependerá de la matriz alimenticia a la que pertenece (Nielsen, 2010).

Actualmente, se realizan investigaciones sobre metodologías analíticas alternativas que, comparadas con las tradicionales, sean más rápidas, menos costosas, se realicen en cada etapa de la producción, no dependan de la matriz alimenticia y que no destruyan la muestra (Bhunia, Kim y Taitt, 2015).

La espectroscopía de infrarrojo cercano (NIRS, por sus siglas en inglés: Near Infrared Spectroscopy) es una técnica secundaria que permite evaluar distintos componentes de una muestra con niveles de exactitud y precisión comparables con los métodos de referencia primarios. Al prescindir de la separación física de los analitos o de una dilución, no requiere el uso de solventes ni reactivos tóxicos para el análisis, teniendo un beneficio con la reducción de tiempos y costos. Además, el material a investigar no es destruido, lo que permite realizar el estudio en la línea de producción (Moros, Garrigues y De la Guardia, 2010). Esto ha permitido que sea utilizada como técnica analítica en la industria farmacéutica, petroquímica, alimenticia, agrícola, medio ambiental, médica, textil, cosmética y estudio de químicos como polímeros (Manley, 2014).

Los primeros estudios que se realizaron con la espectroscopía en el infrarrojo cercano (NIRS) comenzaron a inicios del siglo XIX con el descubrimiento de la energía relacionada con la radiación infrarroja por Sir William Herschel. Las investigaciones consistían principalmente en caracterizar las moléculas orgánicas sencillas. Transcurrieron aproximadamente 150 años para que esta técnica analítica se utilizara en la industria agrícola y alimentaria. La inclusión se le atribuye a Karl Norris con su trabajo para el Departamento de Agricultura de Estados Unidos (USDA) sobre la determinación de humedad y proteínas en cereales y granos. Además, fue pionero en el uso del análisis multivariante, herramienta indispensable para la interpretación de los espectros en NIRS (Workman y Weyer, 2008). Para la década de 1970, Phil Williams,

investigador de la Comisión de Granos Canadiense (CGC), reemplazó exitosamente la prueba Kjeldahl para la determinación de proteína en granos por NIRS lo que significó un ahorro de costos por análisis de más de 500,000 dólares anuales (Skoog, Holler y Nieman, 2001). Al convertirse en una técnica analítica útil en productos agrícolas, la transición hacia el área de alimentos procesados fue rápida. Gradualmente se ha incrementado su uso durante la cadena de producción de chocolates, panes, lácteos, carnes, aceites y bebidas (Irudayaraj y Reh, 2008). El establecer el lugar de origen de un producto o ingrediente, se ha convertido en uno de los estudios recurrentes debido a que son parte de los parámetros que determinan su valor económico. Tal es el caso del aceite de oliva, miel, quesos, carnes y café. Además, para garantizar que es seguro para el consumidor, se han desarrollado modelos que detectan si han sido adulterados, como el utilizado para identificar la presencia de melamina en la leche en polvo para lactantes. La industria de carnes la emplea para obtener un análisis de la composición nutricional de manera rápida, no destructiva y constante durante el proceso de producción. Con un solo espectro, se puede cuantificar simultáneamente el contenido de grasas, humedad y proteína. También se han desarrollado modelos para monitorear la concentración de azúcar y etanol durante la fermentación en la producción de vinos, así como determinar el contenido de alcohol en bebidas alcohólicas (Picó, 2012)

Existen métodos oficiales internacionales basados en el uso de NIRS para cuantificar nutrientes en alimentos, en su mayoría productos agrícolas (ver tabla 1), además de guías para el desarrollo de los modelos de predicción y manejo del instrumento NIR, como la AOCS AM1a-09, AACC 39-00.01 y AACC 39-01.01. Sin embargo, México carece de normas sobre este campo.

*Tabla 1. Lista de métodos oficiales internacionales basados en NIRS*

| Método | Tipo de Método | Analito | Matriz |
|---|---|---|---|
| AOAC 967.19 | Cuantitativo | Agua | Vegetales deshidratados |
| AOAC 997.06 | Cuantitativo | Proteína cruda | Granos/trigo |
| AOAC 991.01 | Cuantitativo | Humedad | Forraje |
| AOAC 989.03 | Cuantitativo | Fibra y proteína cruda | Forraje |
| AOAC 2007.04 | Cuantitativo | Grasa, humedad y proteína | Carnes (res, puerco y pollo) |
| AOCS Am1-92 | Cuantitativo | Aceite, humedad, materia volátil y proteína | Semillas oleaginosas |
| AOCS Cd 1e-01 | Cuantitativo | Yodo | Aceites y grasas |
| AOCS Cd14f-14 | Cuantitativo | Ácidos grasos saturados, monoinsaturados, poliinsaturados y trans. | Aceites vegetales |

| | | | |
|---|---|---|---|
| AACC 08-21.01 | Cuantitativo | Cenizas | Harina de trigo |
| AACC 39-70-.02 | Cuantitativo | Dureza | Trigo |
| AACC 39-10.01 | Cuantitativo | Proteína | Granos pequeños |
| AACC 39-11.01 | Cuantitativo | Proteína | Harina de trigo |
| AACC 39-20.01 | Cuantitativo | Proteína y aceite | Soya |
| AACC 39-25.01 | Cuantitativo | Proteína | Grano de trigo |
| ISO 21543:2006 (IDF 201:2006) | Cuantitativo | Sólidos totales, grasas y proteínas. | Quesos |
| | | Humedad, grasas, proteínas y lactosa | Leche, suero y mantequilla deshidratados. |
| | | Humedad, grasas, sólidos no grasos y sal. | Mantequilla |
| ISO 12099:2010 | Cuantitativo | Humedad, grasas, proteína, almidón y fibra cruda. | Alimento para animales, cereales y productos de cereales molidos. |

Fuente: AOAC, AOCS, AACC, ISO.

El trabajo de investigación en el desarrollo de modelos predictivos continúa. Tan solo en el año 2015, se publicaron en revistas científicas un aproximado de 3,000 artículos relacionados con el uso de NIRS en el campo de las ciencias de alimentos (ScienceDirect, 2016).

## 1.1 Planteamiento del Problema

El municipio de Mexicali, Baja California, se caracteriza por su actividad agropecuaria e industrial. En el sector agrícola destaca en la producción de trigo, cebada, algodón y hortalizas; en la pecuaria en el criadero de bovinos de engorda y lecheros. En el sector industrial, la producción de alimentos ocupa el primer lugar. Dentro de los más importantes se encuentran la pasteurización de lácteos, empacadoras de carnes, tortillerías, molinos de trigo y embotelladoras. Además, la presencia de empresas como Bimbo, Nestlé, Sabritas, Maseca y Coca-Cola comprueba que esta es una actividad económica relevante (Gobierno del Estado de Baja California-GEBC, 2015).

Aunque en algunos laboratorios privados de la ciudad, se realizan los análisis tradicionales para determinar la composición química de los alimentos, estos son costosos y la mayoría los envían a la matriz que se encuentran en otras ciudades como Guadalajara o Ciudad de México. La Universidad Autónoma de Baja California (UABC)

a través del Instituto de Ciencias Agrícolas, también realiza algunos de estos análisis, enfocándose principalmente a los productos agrícolas y alimento para ganado.

Por tal motivo, el Instituto de Ingeniería de la UABC debe tener la opción de ofrecer este servicio a una comunidad y de participar en investigaciones relacionadas utilizando de preferencia, una metodología que no destruya la muestra, disminuya los desperdicios y que reduzca el uso de sustancias tóxicas. La espectroscopía en el infrarrojo cercano (NIRS) cumple con esos puntos.

De acuerdo a lo anterior, la presente investigación comprende una evaluación del desempeño de los modelos generados por NIRS y quimiometría que permitan estimar los componentes químicos de un alimento, en particular el contenido de carbohidratos, proteínas y lípidos.

## 1.2 Hipótesis

Se determina el contenido de carbohidratos, lípidos y proteínas en un producto alimenticio por medio de modelos de predicción diseñados mediante metodología NIRS con una alta correlación entre los datos estimados contra los de referencia.

## 1.3 Objetivos

Para solucionar el problema planteado y contrastar la hipótesis, se establecieron los siguientes objetivos:

### *1.3.1 General*

Evaluar la capacidad de la espectroscopía de infrarrojo cercano para cuantificar, de manera rápida, no destructiva y sin utilizar sustancias tóxicas, el contenido de proteínas, grasas y carbohidratos presentes en un producto alimenticio.

### *1.3.2 Específicos*

- Realizar lecturas con el equipo de infrarrojo cercano con transformada de Fourier (FT-NIR) para obtener espectros del producto alimenticio.

- Desarrollar y validar modelos de predicción para cuantificar carbohidratos, lípidos y proteínas a través de técnicas quimiométricas.

## 1.4 Importancia del Estudio

Una de las razones principales del estudio es iniciar con el uso del equipo FT-NIR ubicado en el Instituto de Ingeniería de la UABC. La relevancia radica en evaluar su potencial como técnica analítica alternativa para cuantificar nutrientes, al analizar los resultados generados por los modelos de predicción. Además, en México este tipo de investigaciones se realizan con poca frecuencia a pesar de las bondades que presenta, lo que permite incursionar en un área.

## 1.5 Delimitación del Estudio

Los modelos de predicción generados para determinar el contenido de carbohidratos, lípidos y proteínas en un producto alimenticio, solo serán aplicables a los espectros obtenidos por el equipo FT-NIR del Instituto de Ingeniería de la UABC de Mexicali, B. C.

*“Los que se enamoran de la práctica sin la teoría son como pilotos sin timón ni brújula, que nunca podrán saber a dónde van”*

*Leonardo Da Vinci*

# 2

# Marco Teórico

La espectroscopía de infrarrojo cercano (NIR) está basada en la absorción de la radiación y pertenece al conjunto de técnicas vibracionales. Las señales de los compuestos químicos que se observan en el espectro se deben a los sobretonos y modos de combinación de los estados de vibración de las moléculas, lo que la hace única comparada con otras técnicas y dificulta su interpretación.

La finalidad de este capítulo es contextualizar las bases de la espectroscopía NIR con la presente investigación para comprender la naturaleza compleja del espectro resultante y facilitar su análisis. Comienza con la descripción general de los principios teóricos de la espectroscopía vibracional, continuando con las características del espectro NIR. Posteriormente, se incluye una revisión sobre los componentes químicos de los alimentos y las señales que manifiestan en el espectro. Finalmente, en la sección de quimiometría, se explica cómo corregir los efectos de dispersión y obtener la información relevante de los datos espectrales para construir los modelos de calibración para los análisis cuantitativos y cualitativos.

## 2.1 Principios generales

La espectroscopía analiza la materia al estudiar su interacción con la radiación electromagnética. El espectro electromagnético está dividido en distintas regiones (fig. 1) y cada una de ellas representa un tipo de transición atómica o molecular y por tanto, una técnica espectroscópica específica.

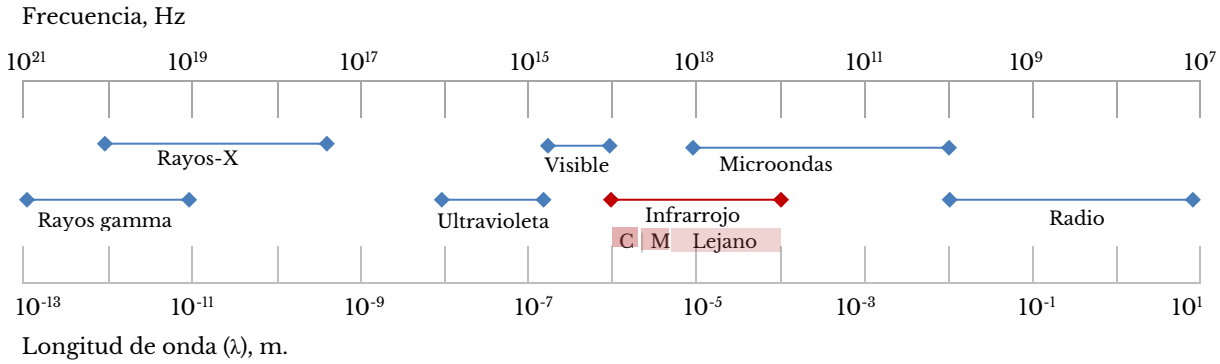



*Figura 1 Regiones del espectro electromagnético*

C: Infrarrojo Cercano, M: Infrarrojo Medio
Fuente: Skoog (2001)

La espectroscopía de infrarrojo cercano (NIR) es la técnica cuyos estudios se limitan a la región de los 800 y 2,500 nm del espectro electromagnético, posicionada entre la visible y la del infrarrojo medio (MIR). Las moléculas que son activas en esta región presentan grandes momentos dipolos y distintos estados vibracionales, específicamente aquellas que contengan enlaces entre carbono – hidrógeno (C-H), oxígeno – hidrógeno (O-H) y nitrógeno – hidrógeno (N-H) son observadas en el espectro NIR. Por esta razón, las muestras de naturaleza orgánica, como los alimentos, pueden ser analizados con este método. Sin embargo, el espectro resultante tiende a manifestar bandas anchas y sobrepuestas debido al mecanismo de interacción de la materia con la radiación NIR.

Cuando una muestra se encuentra en la trayectoria de un rayo NIR de cierta frecuencia y éste coincide con la frecuencia de una vibración específica (tensión o flexión), la molécula absorbe la radiación y presenta transiciones de energía. Una transición del nivel cero al uno se define como frecuencias fundamentales en la molécula, mientras que uno del nivel cero al dos o al tres, como primer y segundo

sobretono respectivamente (fig. 2). Las frecuencias fundamentales se manifiestan en el espectro MIR, mientras que los sobretonos y las combinaciones de los modos vibracionales en el NIR.

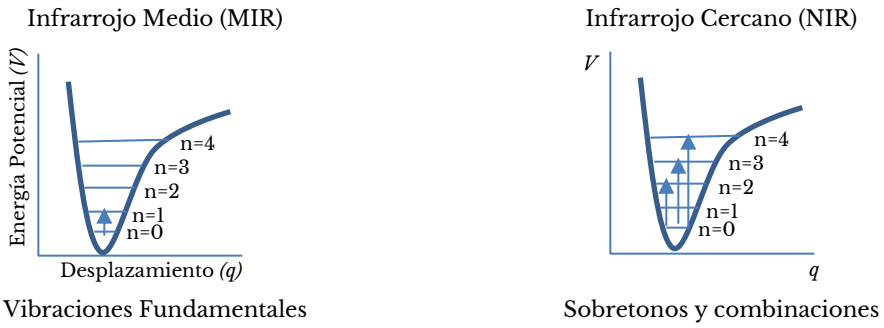



*Figura 2 Principios de la espectroscopía MIR y NIR*

Fuente: Burns (2008)

Las bandas surgidas por los sobretonos y las combinaciones vibracionales son lo que hacen única a esta técnica espectroscópica. Éstas se consideran transiciones prohibidas según el modelo del oscilador armónico simple, que sólo permite aquellos cambios entre niveles de energía que sean vecinos y equidistantes. De esta manera, la anarmonicidad permite que exista más transmisión de la radiación y menor absorción en esta región sin que los enlaces sean destruidos, generando un espectro con bandas débiles y sobrepuestas. Para fines de interpretación espectral esto puede ser una desventaja, por eso la necesidad de emplear quimiometría; sin embargo, la baja intensidad de las bandas facilita el manejo de la muestra en términos de tipo, espesor y concentración. Por esta razón, se pueden analizar sustancias en cualquier estado de agregación que tengan centímetros de espesor y no necesitan ser diluidas, lo que permite que esta técnica se utilice directamente en distintas etapas de producción de un alimento (Burns, 2008).

### *2.1.1 Características del espectro NIR*

La espectroscopía NIR al ser una técnica basada en la absorción de la radiación está fundamentada en la ley Beer – Lambert que relaciona la absorbancia de una sustancia con la concentración de un componente específico de esta manera:

$$A = \varepsilon bc$$

Donde:

A= Absorbancia

ε = coeficiente de absortividad molar: $L*mol^{-1}*cm^{-1}$

*b* = longitud de la trayectoria de la radiación: cm

*c* = concentración del analito: $mol* L^{-1}$

Según lo anterior, la intensidad de las bandas espectrales generadas está relacionada con la absortividad molar del enlace químico que ha sido excitado. Las frecuencias del sobretono tienen un coeficiente bajo de absortividad comparadas con las fundamentales; entre mayor sea el sobretono, menor la intensidad de la banda (Swarbrick, 2016). De esta forma, la región NIR se puede subdividir en tres, aunque las fronteras no son estrictas (ver figura 3):

- Región I: También conocida como la región Herschel, abarca de los 800 a 1200 nm. Se observan bandas débiles por ser los de tercer y cuarto sobretonos. Es ideal para mediciones de transmisión.
- Región II: De los 1200 a 1800 nm. Se presentan bandas por los segundos sobretonos de vibraciones así como algunos modos de combinación.
- Región III: De los 1800 a 2500 nm. Se manifiestan los primeros sobretonos y las bandas de combinación.

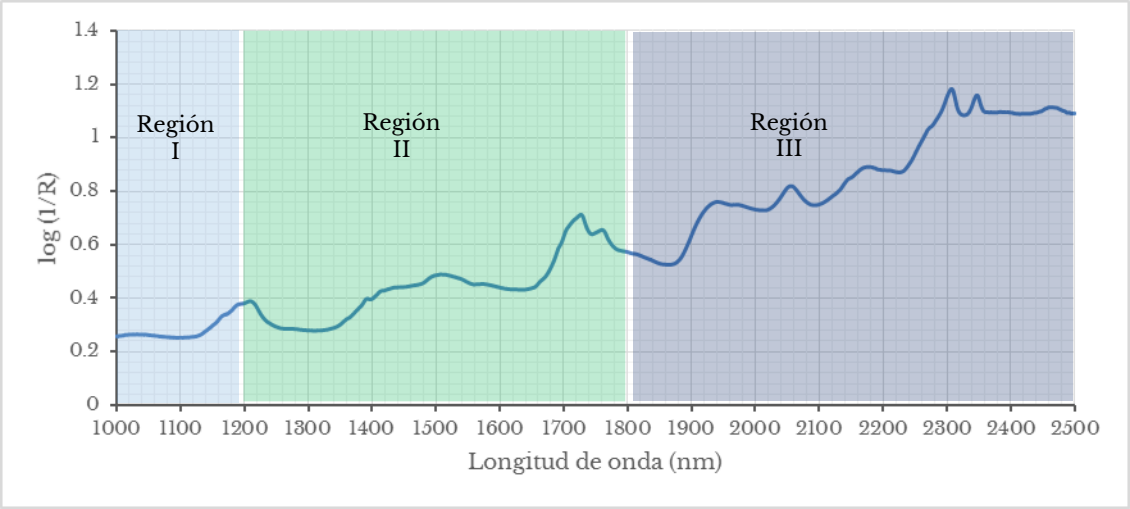



*Figura 3 Regiones del espectro NIR*

Fuente: Yukihiro (2012)

Las regiones II y III tienden a ser las más empleadas en los distintos campos de aplicación, como en el área de alimentos (Yukihiro, 2012). La información de absorción está presentada en forma de un espectro donde la longitud de onda, en unidades de nanómetros (nm), se encuentra sobre en el eje X y la intensidad de absorbancia, usualmente en log (1/R), sobre el eje Y.

Una de las características principales del espectro NIR es que presenta una línea base cuadrática y esto se relaciona directamente con la longitud de la trayectoria de la radiación (*b*), basados en la ley Beer – Lambert. A partir de los 1500 nm, la longitud de onda de la radiación es similar al tamaño de las partículas que se analizan lo que lleva al fenómeno de reflectancia difusa. Por tanto, al ser desviado el rayo NIR el valor de *b* no es constante y la relación de la absorbancia con la concentración deja de ser lineal. Se necesitan aplicar las distintas técnicas quimiométricas para corregir los efectos de dispersión de la luz y para restablecer la relación lineal entre absorbancia y concentración (Swarbrick, 2016).

## 2.2 Componentes principales de los alimentos y su análisis de espectros NIR

Los alimentos son sistemas complejos conformados principalmente por agua, proteínas, carbohidratos, lípidos y en menor grado, por minerales. Sin embargo, la exacta composición química de un producto alimenticio raramente se conoce. Factores como el origen de los ingredientes naturales y la temporada de su obtención, pueden alterar la proporción de dichos constituyentes.

La espectroscopía de infrarrojo cercano (NIR) ha demostrado ser una herramienta eficiente para el análisis cuantitativo y cualitativo de productos alimenticios. Al no alterar la muestra inspeccionada y al obtener información de múltiples analitos con un solo espectro, resulta una técnica valiosa (Zou y Zhao, 2015).

### *2.2.1 Agua*

El agua es un componente esencial de muchos alimentos y su contenido varía según el tipo de comestible que se trate. Las frutas contienen entre un 80 y 95% de agua, las carnes de 50 a 82%, leche de 84 a 86% y el pan de 30 a 35% por mencionar algunos (Cheung, 2015), incluso los que aparentan estar deshidratados presentan entre un 10 a 12% de ella (Badui, 2006). Su presencia puede ser tanto intracelular como extracelular, lo que le permite actuar como solvente o como medio de dispersión. Influye en la apariencia, sabor y estructura, así como en el proceso de deterioro del alimento debido al efecto que tiene en la velocidad de muchas reacciones químicas y de crecimiento microbiano (Cheung, 2015).

La molécula del agua está conformada por dos átomos de hidrógeno unidos en forma covalente con uno de oxígeno y es altamente polar. Tiene la capacidad de crear puentes de hidrógeno estables con otras moléculas de agua y también con aquellas que sean polares (fig. 4), como con los grupos hidrófilos de las proteínas, carbohidratos y ácidos grasos (Badui, 2006).

En el espectro de infrarrojo cercano se observan las bandas de absorción del agua en la región de 1450 (1er sobretono) y 1940 nm (banda de combinación) por las vibraciones OH. Sin embargo, pueden presentarse cambios en su ubicación y en la anchura de la banda debido a la formación o rompimiento de puentes de hidrógeno. Factores como la temperatura y las interacciones químicas con otras moléculas de la muestra a analizar favorecen dichas alteraciones (Burns, 2008).

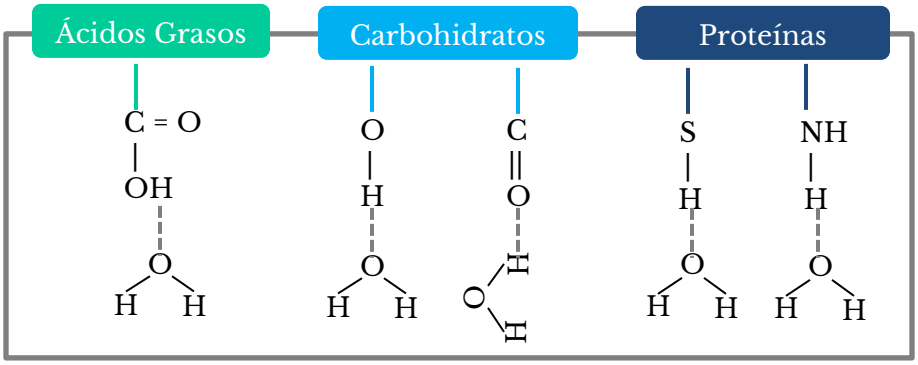



*Figura 4 - Puentes de hidrógeno y su interacción con distintos grupos funcionales*
Fuente: Badui (2006)

### *2.2.2 Proteínas*

Las proteínas son las macromoléculas más abundantes y diversas presentes en todos los seres vivos, además de tener un papel central en muchos procesos biológicos. Se encuentran principalmente en la leche, huevos, carnes, cereales, legumbres y semillas oleaginosas. Aunque aportan 4 kcal/g de energía, su valor nutritivo dependerá del contenido de aminoácidos esenciales y su digestibilidad. Las de origen animal tienden a ser mejor digeridas que las de origen vegetal y se utilizan más en la producción de alimentos (Cheung, 2015).

Las proteínas son polímeros de aminoácidos unidos por un tipo específico de enlace peptídico (fig. 5). Se pueden encontrar veinte distintos aminoácidos en las proteínas y cada uno de ellos están formados por un grupo amino y uno carboxílico unido al mismo átomo de carbono. La diferencia entre ellos es el grupo R, una cadena lateral que varía en tamaño, estructura y carga eléctrica (Nelson y Cox, 2005). Se ha estimado en porcentaje peso que en total una proteína contiene del 50 a 55% de carbono, de 20 a 23% de oxígeno, 12 a 19% de nitrógeno, 6 a 7% de hidrógeno y entre 0.2 a 3 % de azufre (Cheung, 2015).

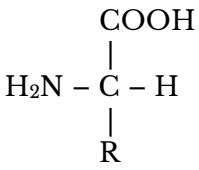



*Figura 5 Estructura general de un aminoácido*
Fuente: Nelson (2005)

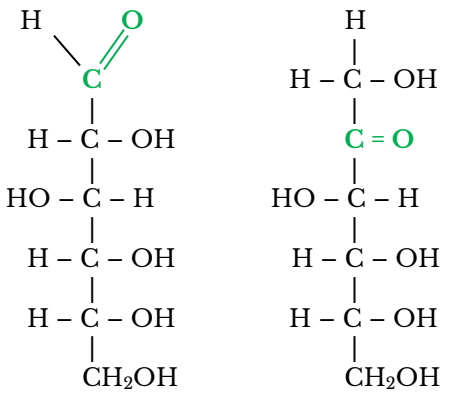
Debido a la estructura compleja de las proteínas se presentan diversas bandas de absorción en los espectros NIR. Entre las longitudes de onda más importantes para su medición directa se encuentra la región de los 2050 a 2060 nm en donde se manifiesta un estiramiento del grupo carbonil de la amida primaria; la región entre los 2168 a 2180 nm, 1640 a 1680 nm por los grupos aromáticos y finalmente, de los 1500 a 1530 nm (Burns, 2008).

### *2.2.3 Carbohidratos*

Los carbohidratos, también conocidos como hidratos de carbono, son las biomoléculas más abundantes de la naturaleza. La mayoría de ellos presentan la fórmula general $(CH_2O)n$ y dependiendo de la ubicación del grupo funcional carbonilo, pueden ser polihidroxialdehídos o polihidroxicetonas. Según el número de

monómeros que los conforman o de unidades de azúcar, se clasifican en tres categorías: monosacáridos, oligosacáridos y polisacáridos (Nelson, 2005).

A los monosacáridos también se les conoce como azúcares simples y están conformados por una sola unidad del grupo funcional aldehído o cetona (fig. 6). Dentro de los más relevantes en los alimentos se encuentran la glucosa, también conocida como dextrosa, y la fructosa. Ésta última se encuentra principalmente en las frutas y en las mieles.




D-Glucosa, polihidroxialdehído — D-Fructosa, polihidroxicetona

*Figura 6 Monosacáridos representativos*

Fuente: Nelson (2005)

Los oligosacáridos son cadenas cortas que contienen hasta diez unidades de azúcar y los más abundantes son los disacáridos, formados por dos monosacáridos. De éstos últimos destacan la sacarosa, mejor conocida como azúcar de mesa compuesta por una glucosa y una fructosa; el azúcar de la leche, lactosa, por una glucosa y una galactosa, así como la maltosa formada por dos moléculas de glucosa. Finalmente, los polisacáridos son polímeros que contienen más de 20 monosacáridos, entre los que destacan los almidones, celulosa, glucógeno, pectinas y gomas.

Los carbohidratos son los más consumidos por el hombre, representan entre el 50 y 80% de la dieta y proporcionan 4 kcal/g de energía. Para la elaboración de distintos productos alimenticios en la industria se utilizan tradicionalmente la glucosa, sacarosa y lactosa aunque han ganado popularidad los llamados polioles o azúcares–alcoholes, como el xilitol y sorbitol. El uso de estos carbohidratos tiene diversos fines como el de endulzar, inhibir el crecimiento microbiano o el de darle al alimento propiedades sensoriales deseadas, como el brillo en los chocolates (Badui, 2006).

En los espectros NIR los carbohidratos presentan una banda de absorción en la longitud de onda de los 2100 nm por las vibraciones de los enlaces O-H/C-O y C=O-O. Sin embargo, para su estudio también se utilizan otras regiones dependiendo del tipo de molécula que se requiera analizar. Por ejemplo, la celulosa presenta bandas en los 1490, 1780, 1820, 2335, 2347, 2352 y los 2488 nm (Burns, 2008).

### *2.2.4 Lípidos*

Los lípidos son biomoléculas cuya característica principal es la insolubilidad en el agua. Su función en los sistemas biológicos depende de su estructura química. Los derivados de ácidos grasos almacenan energía de muchos organismos en forma de grasas o aceites; los fosfolípidos y esteroles son componentes estructurales de la membrana celular. Otros compuestos lipídicos participan como cofactores enzimáticos, pigmentos, hormonas y mensajeros intracelulares, por mencionar algunos (Nelson, 2005).

Podemos encontrar a los lípidos en los alimentos en forma de grasas y aceites principalmente en tejidos animales y en semillas oleaginosas aunque también en algunos frutos como aguacate y aceitunas. Las proporciones varían según el consumible, por ejemplo, la mantequilla contiene un 80% de esta macromolécula, la leche de un 4 a 6%, el queso de 25 a 30% y el pollo un 7% (Cheung, 2015). En la dieta, son la principal fuente de energía al proveer 9 kcal por gramo. En la industria alimenticia los utilizan para proporcionar textura, lubricación, color, sabor, entre otras propiedades sensoriales y como aditivos para emplearlos como emulsionantes y antiaglomerantes (Badui, 2006).

Las grasas y aceites son triacilglicéridos, también llamados triglicéridos. Éstos últimos están formados por tres ácidos grasos unidos a un glicerol por medio de enlaces ésteres (fig. 7). A su vez, los ácidos grasos son ácidos carboxílicos que contienen cadenas de hidrocarburos de hasta 36 átomos de carbonos. Dichas cadenas pueden ser lineales o ramificadas y si presentan enlaces dobles se les conoce como insaturadas (Nelson, 2005).

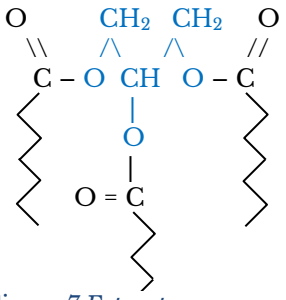



*Figura 7 Estructura general de un triacilglicérido*
Fuente: Nelson, 2005

En los espectros NIR las longitudes de onda relacionadas con los lípidos son varias y dependen del tipo de muestra a analizar y de la molécula objeto de estudio. Se les asocia las vibraciones de enlaces manifestados en los 1410, 2070, 2140, 2310 y 2380 nm (Burns, 2008).

### *2.2.5 Minerales*

Se les conoce como minerales a los elementos presentes en los alimentos diferentes al carbono, hidrógeno, oxígeno y nitrógeno. Se les puede encontrar en forma de sales orgánicas e inorgánicas y como componente de algunas biomoléculas (Cheung, 2015).

Su clasificación está relacionada con la concentración presente en el organismo. De esta forma, los minerales esenciales son el calcio, sodio, potasio, magnesio, cloro y fósforo al tener una presencia mayor al 0.1%. Los elementos traza constituyen el resto, como el hierro y el flúor. Al combinar adecuadamente los alimentos de origen animal y vegetal, el ser humano obtiene los minerales necesarios para mantener buena salud. Aun así, a muchos consumibles se les añade alguno de ellos para fortificarlos y a otros como aditivos (Badui, 2006).

Las formas iónicas y las sales minerales no manifiestan bandas de absorción en la región del infrarrojo cercano. En algunas ocasiones se les logra detectar si están unidos a compuestos orgánicos. Sin embargo, de manera indirecta se puede determinar su presencia en muestras con alto contenido de humedad si se observan desplazamientos en las bandas de absorción del agua (Burns, 2008). Por estas razones, se han realizado investigaciones para determinar el contenido de minerales utilizando espectroscopía NIR en distintas muestras como en los cacahuates, leguminosas, queso, forrajes, pastizales, lomo de puerco Ibérico y vino (Phan-Thien, 2011).

## 2.3 Quimiometría

La quimiometría consiste en la aplicación de métodos matemáticos y estadísticos para obtener información relevante de muestras a partir de sus datos químicos (Burns, 2008). La interpretación de los espectros NIR podría ser una tarea difícil debido a la naturaleza de las bandas, éstas tienden a ser débiles y anchas, además de que se sobreponen unas sobre otras y tienden a desplazarse por la influencia de los puentes de hidrógeno (Jha, 2010). Asimismo, dicho espectro es un conjunto de datos multivariante porque cada longitud de onda que lo conforma aporta una medición

(Wang, 2007). Por ejemplo, un espectro que abarque el rango de los 1000 a 2500 nm puede proporcionar 1501 variables.

La aplicación de los métodos de análisis multivariante permite relacionar una propiedad en particular con el espectro NIR al extraer la información útil e ignorar el ruido. De esta manera, acciones como clasificar las muestras alimenticias según su origen geográfico o determinar su contenido de proteínas, son posibles al desarrollar modelos de clasificación o de regresión respectivamente (Zou, 2015).

Los métodos quimiométricos pueden ser categorizados según el propósito de su uso en cuatro grupos:

- Pretratamientos matemáticos.
- Análisis exploratorio de datos.
- Modelos para análisis cualitativos: Discriminación y clasificación.
- Modelos para análisis cuantitativos: Regresión y predicción.

### *2.3.1 Pretratamientos matemáticos*

Los pretratamientos o preprocesamientos matemáticos preparan los datos espectrales para el desarrollo de modelos cualitativos y cuantitativos. Remueven señales indeseadas como el ruido y los desplazamientos de la línea base, corrigen los efectos de la dispersión de la luz y los datos son linealizados (Wang, 2007).

Existe una gran variedad de estas técnicas que puede resultar complicado elegir de entre todas la más apropiada. Su selección dependerá del tipo de muestra a analizar y la propiedad a evaluar. Por tanto, para la presente investigación se aplicaron los pretratamientos espectrales más populares y otros representativos, la mayoría disponibles en el software del espectrofotómetro.

**Centrado a la media**

Consiste en restar el valor promedio de un conjunto de datos a cada variable (Unscrambler). Esto permite que el modelo de regresión sea más sencillo de interpretar al remover la necesidad de que tenga un intercepto, y por tanto, las concentraciones estimadas de los analitos sean más precisos (Burns, 2008).

La figura 8a, consiste en varios espectros NIR de suplementos alimenticios sin pretratamientos, mientras que la figura 8b es el resultado de la transformación después de haber sido centrado a la media.

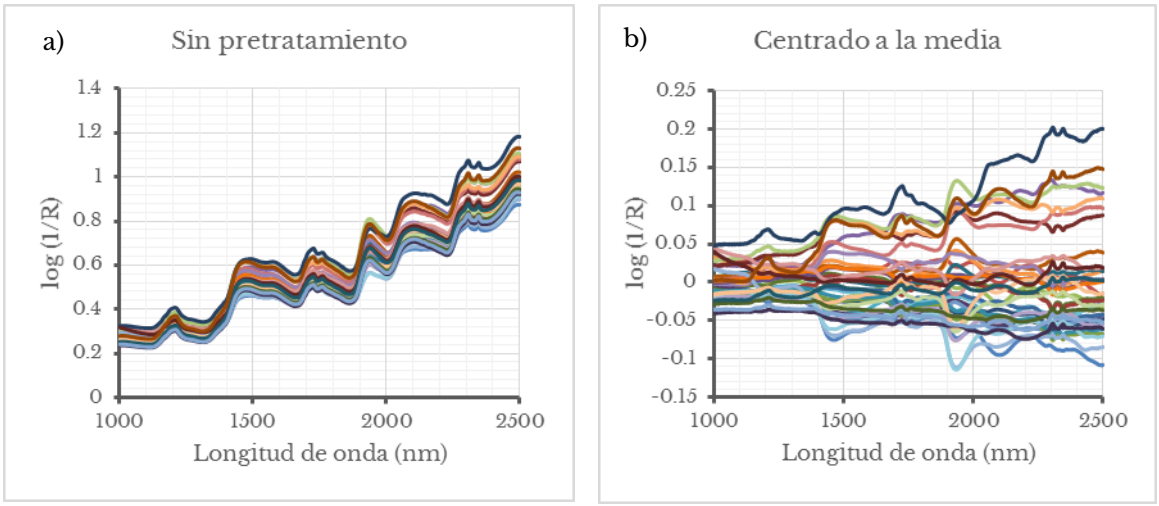



*Figura 8 Comparación de espectros FT-NIR: a) antes y b) después del Centrado a la media*

Fuente: autor

**Corrección de línea base**

Se utiliza para corregir el desplazamiento de la línea base del espectro (fig. 9). Puede ser a través de una compensación (offset) que consiste en restar el valor mínimo de todas las variables o en aplicar una transformación que convierte la línea base inclinada en una horizontal al elegir dos variables entre las cuales se definirá la nueva base (CAMO Software AS, 2016).

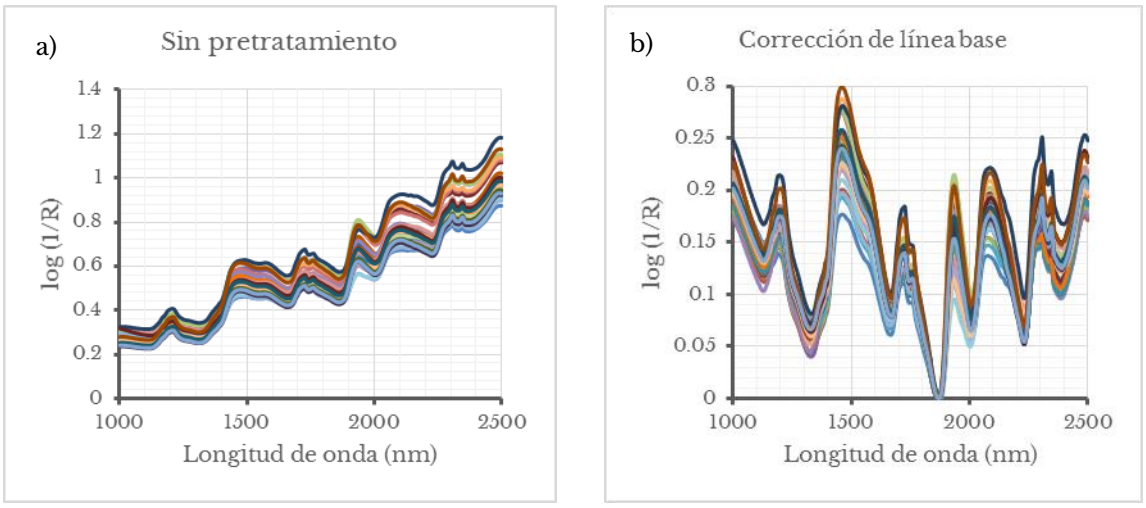



*Figura 9 Comparación de espectros FT-NIR: a) antes y b) después de la corrección de la línea base*

Fuente: autor

**Suavizado**

Permite reducir el ruido en los datos espectrales sin reducir el número de variables (CAMO Software AS, 2016). Existen varios métodos de suavizado, el empleado en este estudio es el Savitzky-Golay (fig. 10) que es el que emplea el software Spectrum.

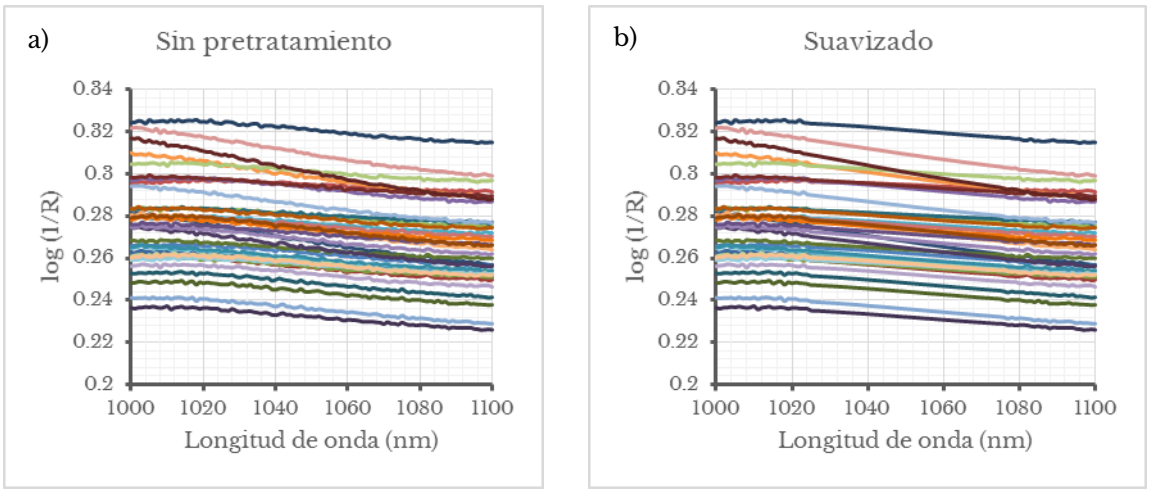



*Figura 10 Comparación de espectros FT-NIR: a) antes y b) después del suavizado*

Fuente: autor

**Derivadas**

La aplicación de esta técnica corrige la línea base y permite resolver el problema del traslape de las bandas en los espectros NIR al obtener información que aparenta estar oculta. Sin embargo, puede incrementar el ruido y remover datos que podrían ser útiles (Li Vigni, Durante y Cocchi, 2013)

Las derivadas pueden ser de primer, segundo, tercer y cuarto orden, siendo las primeras dos las más populares. En esta investigación se utilizaron las de Savitzky-Golay por ser las que el software Spectrum tiene por default.

En los espectros resultantes de la aplicación de la primera derivada (fig. 11b), los picos presentes en los datos originales se convierten en puntos que cruzan cero y en los de la segunda (fig. 11c), cambian de signo y se vuelven picos negativos (CAMO Software AS, 2016).

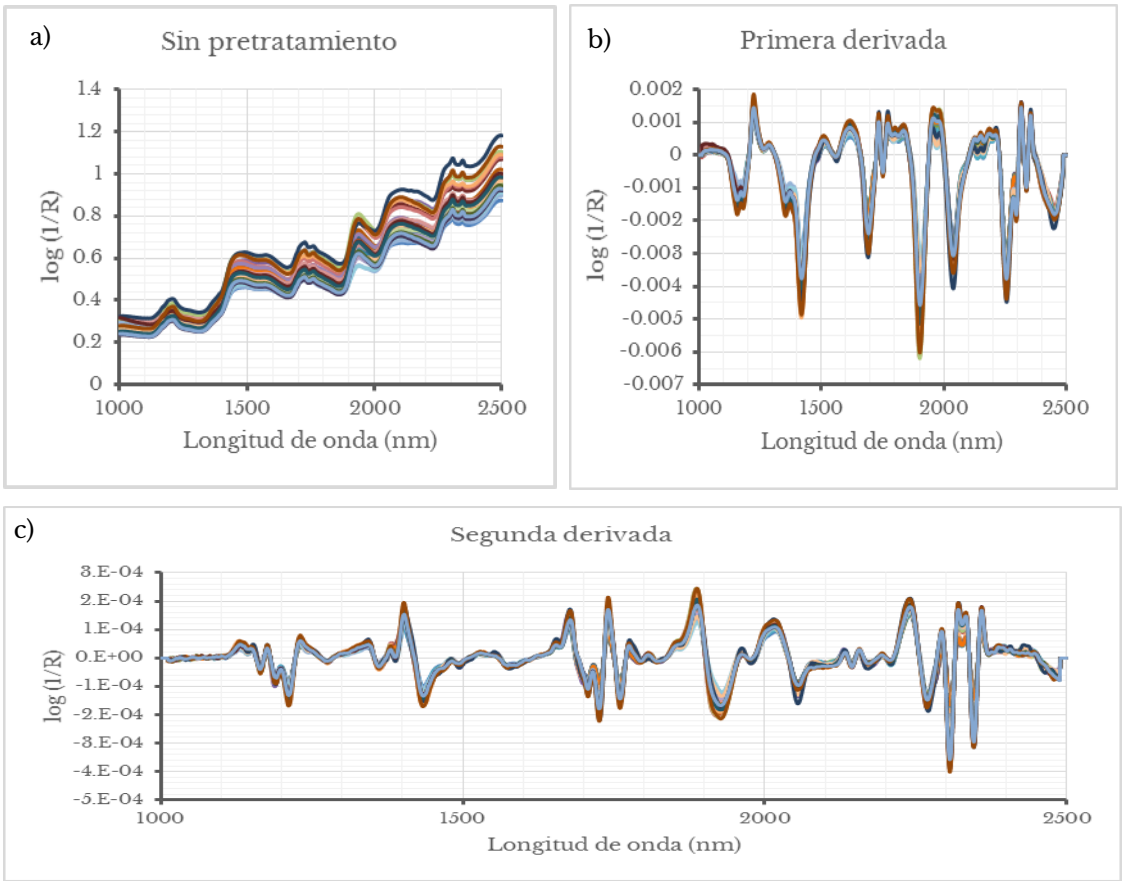



*Figura 11 Comparación de espectros FT-NIR: a) sin pretratamiento y b) con la derivada*

Fuente: autor

**Detrending**

Es un pretratamiento que consiste en remover las tendencias no lineales del espectro (Li Vigni, 2013). Corrige el desplazamiento de la línea base, así como su curvatura (fig. 12b). Usualmente, se aplica junto con el preprocesamiento Variación Normal Estándar (SNV por sus siglas en inglés).

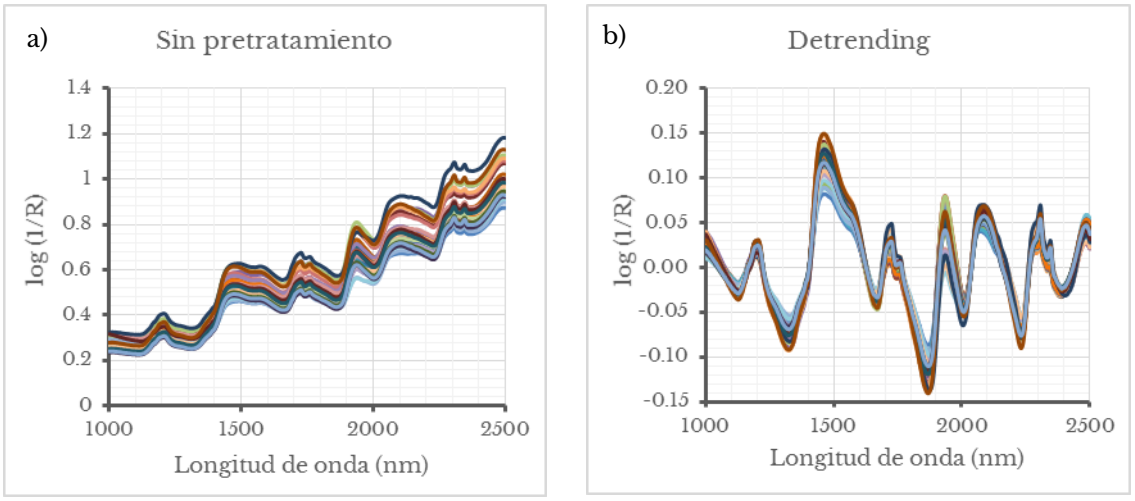



*Figura 12 Comparación de espectros FT-NIR: a) sin pretratamiento y b) con Detrending*

Fuente: autor

**Corrección de los efectos de dispersión de la luz**

Existen dos pretratamientos que permiten remover las interferencias producidas por la dispersión de la luz y los efectos del tamaño de la partícula en los espectros obtenidos por reflectancia, como el desplazamiento de la línea base entre muestras de un mismo conjunto de datos.

El método de corrección de dispersión multiplicativa (MSC por sus siglas en inglés) calcula un coeficiente que permite remover los efectos aditivos y multiplicativos de la dispersión de la luz (fig. 13a). Para esto, utiliza el espectro promedio de un conjunto de datos como referencia para futuras correcciones de espectros obtenidos bajo las mismas condiciones experimentales y asume que dichos efectos no están relacionados con las características químicas de las muestras.

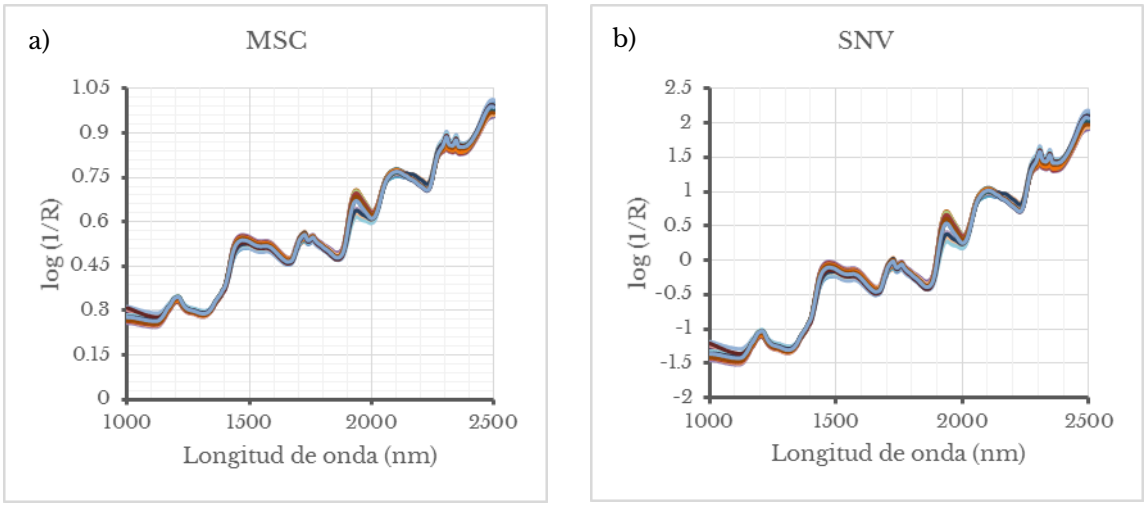



*Figura 13 Comparación de espectros FT-NIR: a) con MSC y b) con SNV*

Fuente: autor

El preprocesamiento Variación Normal Estándar (SNV, por sus siglas en inglés) remueve los efectos de dispersión de la luz al centrar a cero y escalar de manera individual a cada espectro integrante de un conjunto de datos (fig. 13b). El resultado es similar al obtenido por MSC y la principal diferencia entre ellos, haciendo a un lado la escala, consiste en que SNV no utiliza el espectro promedio de un set de datos para realizar correcciones futuras. Es cuestión de gusto emplear cualquiera de los dos (Azzouz, Puigdoménech, Aragay y Tauler, 2003).

**Corrección ortogonal de la señal**

El método de corrección ortogonal de la señal (OSC por sus siglas en inglés) fue diseñado originalmente para corregir datos obtenidos por espectroscopía de reflectancia del infrarrojo cercano. Remueve las variaciones de los datos espectrales que no estén relacionados a la propiedad de interés, lo que permite mejorar el desempeño de los modelos de calibración (Burns, 2008).

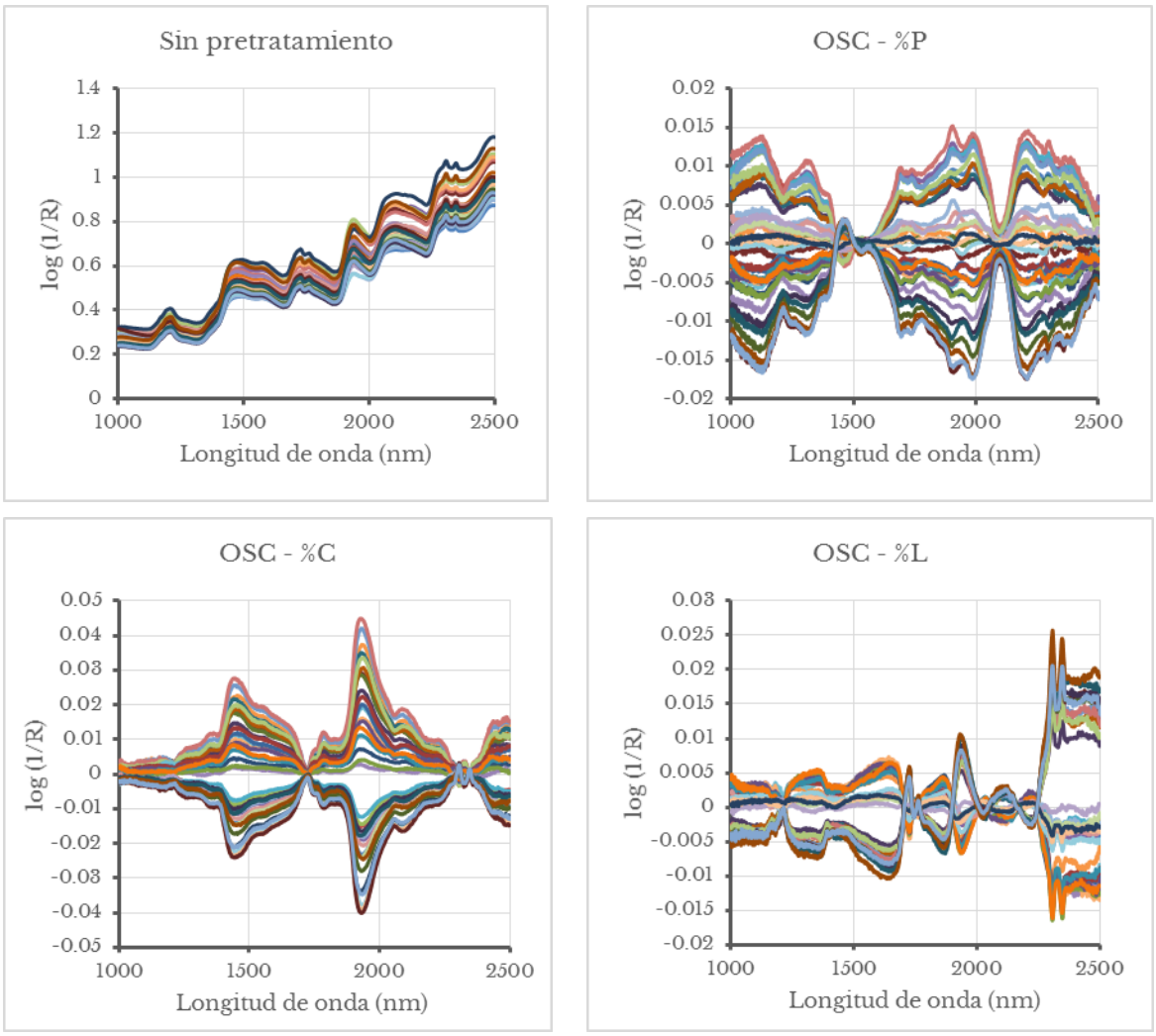



*Figura 14 Espectros FT-NIR con Corrección Ortogonal de la Señal (OSC)*

a) Sin pretratamiento, b) OSC basado en la concentración de proteínas, c) OSC basado en la concentración de carbohidratos, d) OSC basado en la concentración de lípidos.
Fuente: autor

### *2.3.2 Análisis exploratorio de datos*

Se aplican antes de desarrollar el modelo de regresión para determinar posibles relaciones entre las muestras y detectar aquellas que no pertenezcan al grupo (Manley, 2014). Generalmente, los resultados se muestran en gráficos y la búsqueda de patrones se hace de manera visual. La técnica más utilizada es el Análisis del Componente Principal, mejor conocida como PCA por sus siglas en inglés.

En el caso de datos espectroscópicos, PCA identifica las longitudes de onda que presentan mayor variación en un set de datos al considerar que éstas son quienes aportan información y que no contribuyen al ruido. Proyecta los resultados en una menor cantidad de variables conocidas como variables latentes o como componentes principales (PC, por sus siglas en inglés). El primer PC interpreta la mayor cantidad de información, el segundo exhibe lo que no se tomó en cuenta en el anterior y en cada subsecuente PC aumenta la porción de ruido.

El PCA es una técnica básica en el análisis multivariante debido a que en la construcción de los modelos de regresión se utilizan las variables latentes o los componentes principales calculados con este método. Tal es el caso de la Regresión del Componente Principal (PCR, por sus siglas en inglés), resolución de curva multivariante (MCR, por sus siglas en inglés) y la regresión parcial de mínimos cuadrados, mejor conocida como PLS. Incluso, para determinar si el modelo es sencillo de interpretar, éste debe contener la menor cantidad de PCs posibles (Miller y Miller, 2010).

### *2.3.3 Modelos para análisis cualitativos: Discriminación y clasificación.*

Consisten en técnicas de reconocimiento de patrones que se emplean para comparar el espectro NIR de una muestra conocida con el de una desconocida y establecer similitudes o diferencias entre ellas. En el área de alimentos, principalmente se utilizan para clasificarlos según su origen o ingredientes, detectar una adulteración o determinar su grado de pureza, como en el caso de las mieles y aceites de oliva.

Para desarrollar un modelo de clasificación se tiene que considerar los requerimientos que una muestra debe cumplir para pertenecer a una clase, además de colectar espectros de las muestras conocidas que contengan todas las variaciones posibles para hacer una biblioteca de espectros representativos.

Las técnicas que se podrían emplear para el desarrollo de modelos de clasificación son diversas como el análisis de grupos, análisis discriminante factorial (FDA, por sus siglas en inglés), análisis discriminante PLS (PLS-DA), modelado independiente suave de la analogía de clase (SIMCA, por sus siglas en inglés), análisis de variación canónica (CVA, por sus siglas en inglés), redes neuronales artificiales (ANN,

por sus siglas en inglés), clasificación con apoyo de la máquina vectorial (SVM, por sus siglas en inglés), análisis discriminante lineal (LDA, por sus siglas en inglés) y análisis del vecino k-cercano o k-NN por sus siglas en inglés (Manley, 2014).

### *2.3.4 Modelos para análisis cuantitativos: Regresión y Predicción.*

La espectroscopía NIR es una técnica analítica secundaria que, para poder medir una propiedad química de una muestra, requiere establecer una relación entre el espectro y el valor de dicha propiedad. A este último se le conoce como valor de referencia y debe ser determinado previamente por un método analítico independiente (Manley, 2014). De esta manera, se puede predecir en un alimento su contenido de proteínas, agua, lípidos, carbohidratos y otros constituyentes, utilizando modelos de regresión multivariantes.

Los métodos comúnmente empleados son la regresión de componentes principales (PCR, por sus siglas en inglés), la de mínimos cuadrados parciales (PLS, por sus siglas en inglés) y la regresión múltiple lineal (MLR, por sus siglas en inglés). Los tres métodos tienen en común que utilizan técnicas de mínimos cuadrados para construir modelos lineales relacionando la matriz de los datos espectrales (X) con la matriz de la propiedad a analizar (Y), estimando así el coeficiente de regresión (Zou, 2015).

Para PCR y PLS, el desarrollo de los modelos de calibración se basa en el principio de que sólo unos cuántos componentes lineales del conjunto de datos pueden emplearse en la ecuación de regresión. En PCR se denominan componentes principales (PCs) y en PLS factores (Manley, 2014). La habilidad de predicción de ambas técnicas es similar, incluso cuando para el cálculo de la ecuación de regresión con PLS se utilicen menos factores (Zou, 2015).

MLR se emplea cuando el espectro contiene pocas variables. Comienza con la longitud de onda que presente mayor correlación lineal y posteriormente, elige las siguiente que aumente el coeficiente de determinación ($R^2$) y reduzca el error de predicción. El proceso se detiene cuando al agregar otra longitud de onda el valor de $R^2$ disminuye y el error aumenta (Zou, 2015).

*“Es esencial realizar experimentos para comprobar lo que se ha escrito en lugar de aceptarlo a ciegas como verdadero”.*

*Ibn al-Haytham*

# 3

# Metodología

El presente capítulo contiene la descripción del proceso seguido para desarrollar modelos de regresión de mínimos cuadrados parciales (PLS, por sus siglas en inglés) que permitieron clasificar a las muestras según una categoría, como la marca comercial de los suplementos alimenticios, y que cuantificaran una propiedad en particular, como la concentración de proteínas.

Comienza con la descripción de la selección y preparación de las muestras a analizar, seguido de las especificaciones técnicas para la obtención de los espectros FT-NIR, continuando con una guía para la interpretación visual de los espectros recabados, así como una exploración de datos con el análisis del componente principal (PCA, por sus siglas en inglés).

Posteriormente, se detallan los parámetros empleados para el cómputo de los modelos, tanto para la clasificación como para la cuantificación, estableciendo los indicadores que determinarán si el modelo es eficiente y el procedimiento a seguir si hay indicios de que no lo sea, finalizando con unos diagramas de flujo que resumen la metodología empleada.

## 3.1 Preparación y valores de referencia de las muestras

Con el propósito de desarrollar una metodología utilizando espectroscopía de infrarrojo cercano que permita un análisis cualitativo y cuantitativo en cualquier tipo de alimento, se dividió en dos grupos las muestras a analizar en el presente estudio.

El primero consistió en mezclas elaboradas con cuatro tipos de harinas libres de gluten, representando a los ingredientes empleados para la elaboración de los comestibles. El segundo set lo conforma un producto elaborado. Se eligieron suplementos alimenticios para garantizar que se tuvieran la mayor cantidad de nutrientes, así como conservadores, saborizantes y colorantes que pudieran influir en el desarrollo de los modelos NIR.

En la figura 15 se aprecia cómo los macronutrientes: carbohidratos, lípidos y proteínas, se encuentran presentes desde los ingredientes que integran a un comestible hasta el producto final.

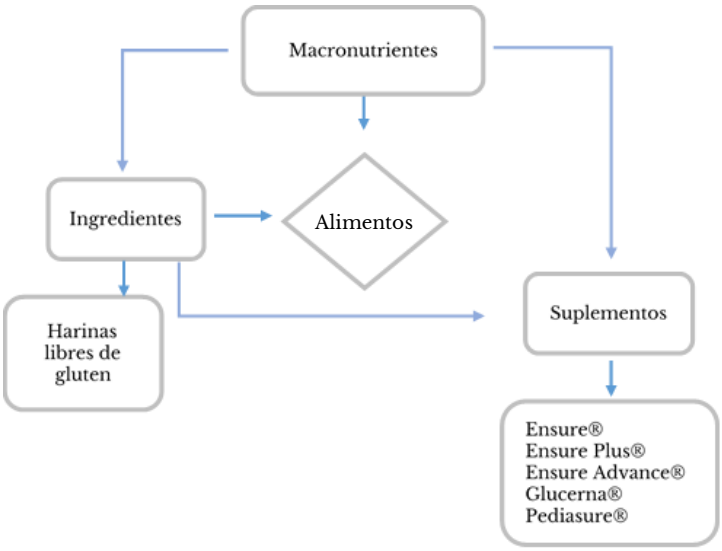



*Figura 15 Relación entre las muestras empleadas en el estudio*
Fuente: autor

### *3.1.1 Harinas libres de gluten*

Se prepararon 40 muestras con distintas proporciones conocidas de harina de almendra, arroz, avena y coco de la marca Bob's Red Mill® que fueron proporcionadas por el laboratorio de Biopelículas del Instituto de Ingeniería de la UABC. Cada mezcla tenía un peso final de 5 gramos.

Se pesaron en la balanza analítica Ohaus® Pioneer™ las cantidades especificadas en la tabla 2, calculadas con la fórmula de porcentaje en peso (3.1) para lograr la composición deseada de cada harina en las muestras. Posteriormente, fueron guardadas en bolsas de polietileno hasta su análisis.

$$\% \text{ P/P} = \frac{\text{g soluto}}{\text{g disolución}} \times 100 \qquad (3.1)$$

*Tabla 2 Valores para la preparación de las Harinas libres de gluten*

| | | Cantidades | | | | Concentración | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Nombre de la Muestra | Almendra (g) | Arroz (g) | Avena (g) | Coco (g) | Almendra (% p) | Arroz (%p) | Avena (%p) | Coco (%p) |
| 1 | ALM-40% | 2.00 | 1.00 | 1.00 | 1.00 | 40 | 20 | 20 | 20 |
| 2 | ALM-55% | 2.75 | 0.75 | 0.75 | 0.75 | 55 | 15 | 15 | 15 |
| 3 | ALM-64% | 3.20 | 0.60 | 0.60 | 0.60 | 64 | 12 | 12 | 12 |
| 4 | ALM-70% A | 3.50 | 0.50 | 0.50 | 0.50 | 70 | 10 | 10 | 10 |
| 5 | ALM-70% B | 3.50 | 0.50 | 0.50 | 0.50 | 70 | 10 | 10 | 10 |
| 6 | ALM-80% | 4.00 | 0.34 | 0.34 | 0.34 | 80 | 6.7 | 6.7 | 6.7 |
| 7 | ARR-40% | 1.00 | 2.00 | 1.00 | 1.00 | 20 | 40 | 20 | 20 |
| 8 | ARR-55% | 0.75 | 2.75 | 0.75 | 0.75 | 15 | 55 | 15 | 15 |
| 9 | ARR-64% | 0.60 | 3.20 | 0.60 | 0.60 | 12 | 64 | 12 | 12 |
| 10 | ARR-70% | 0.50 | 3.50 | 0.50 | 0.50 | 10 | 70 | 10 | 10 |
| 11 | ARR-80% | 0.34 | 4.00 | 0.34 | 0.34 | 6.7 | 80 | 6.7 | 6.7 |
| 12 | AVE-40% | 1.00 | 1.00 | 2.00 | 1.00 | 20 | 20 | 40 | 20 |
| 13 | AVE-55% | 0.75 | 0.75 | 2.75 | 0.75 | 15 | 15 | 55 | 15 |
| 14 | AVE-64% | 0.60 | 0.60 | 3.20 | 0.60 | 12 | 12 | 64 | 12 |
| 15 | AVE-70% | 0.50 | 0.50 | 3.50 | 0.50 | 10 | 10 | 70 | 10 |
| 16 | AVE-80% | 0.34 | 0.34 | 4.00 | 0.34 | 6.7 | 6.7 | 80 | 6.7 |
| 17 | COAR-05153050A | 1.50 | 2.50 | 0.75 | 0.25 | 30 | 50 | 15 | 5 |
| 18 | COAR-05153050B | 1.50 | 2.50 | 0.75 | 0.25 | 30 | 50 | 15 | 5 |
| 19 | COAR-15305005A | 2.50 | 0.25 | 1.50 | 0.75 | 50 | 5 | 30 | 15 |
| 20 | COAR-15305005B | 2.50 | 0.25 | 1.50 | 0.75 | 50 | 5 | 30 | 15 |
| 21 | COAR-25% A | 1.25 | 1.25 | 1.25 | 1.25 | 25 | 25 | 25 | 25 |
| 22 | COAR-25% B | 1.25 | 1.25 | 1.25 | 1.25 | 25 | 25 | 25 | 25 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 23 | COAR-30500515A | 0.25 | 0.75 | 2.50 | 1.50 | 5 | 15 | 50 | 30 |
| 24 | COAR-30500515B | 0.25 | 0.75 | 2.50 | 1.50 | 5 | 15 | 50 | 30 |
| 25 | COAR-50051530A | 0.75 | 1.50 | 0.25 | 2.50 | 15 | 30 | 5 | 50 |
| 26 | COAR-50051530B | 0.75 | 1.50 | 0.25 | 2.50 | 15 | 30 | 5 | 50 |
| 27 | COCO-40% | 1.00 | 1.00 | 1.00 | 2.00 | 20 | 20 | 20 | 40 |
| 28 | COCO-55% | 0.75 | 0.75 | 0.75 | 2.75 | 15 | 15 | 15 | 55 |
| 29 | COCO-64% | 0.60 | 0.60 | 0.60 | 3.20 | 12 | 12 | 12 | 64 |
| 30 | COCO-70% | 0.50 | 0.50 | 0.50 | 3.50 | 10 | 10 | 10 | 70 |
| 31 | COCO-80% | 0.34 | 0.34 | 0.34 | 4.00 | 6.7 | 6.7 | 6.7 | 80 |
| 32 | AROC-07136020 | 0.35 | 0.65 | 3.00 | 1.00 | 7 | 13 | 60 | 20 |
| 33 | AROC-20071360 | 1.00 | 0.35 | 0.65 | 3.00 | 20 | 7 | 13 | 60 |
| 34 | AROC-60200713 | 3.00 | 1.00 | 0.35 | 0.65 | 60 | 20 | 7 | 13 |
| 35 | AROC-13602007A | 0.65 | 3.00 | 1.00 | 0.35 | 13 | 60 | 20 | 7 |
| 36 | AROC-13602007B | 0.65 | 3.00 | 1.00 | 0.35 | 13 | 60 | 20 | 7 |
| 37 | AROC-35451111 | 1.65 | 2.25 | 0.55 | 0.55 | 35 | 45 | 11 | 11 |
| 38 | AROC-11354511 | 0.55 | 1.65 | 2.25 | 0.55 | 11 | 35 | 45 | 11 |
| 39 | AROC-11113545 | 0.55 | 0.55 | 1.65 | 2.25 | 11 | 11 | 35 | 45 |
| 40 | AROC-45111135 | 2.25 | 0.55 | 0.55 | 1.65 | 45 | 11 | 11 | 35 |
| | **Rango** | - | - | - | - | 5-80 | 5-80 | 5-80 | 5-80 |
| | **Media** | - | - | - | - | 25.82 | 25.50 | 24.50 | 24.18 |

Fuente: autor

Además, se prepararon 6 mezclas por duplicado con un peso final de 3 gramos, basadas en una formulación para hacer panecillos libres de gluten donde la harina de almendra se encontraba en una proporción de 7:1 con respecto a las de arroz, avena y coco. La concentración de dichas harinas, todas ellas de la marca Bob's Red Mill® fueron modificadas como se especifica en la tabla 3. El resto de los ingredientes se mantuvo constante y éstos fueron: polvo para hornear, goma xantana, sal, mantequilla en polvo, azúcar con stevia y huevo en polvo. Las muestras fueron guardadas en bolsas de polietileno.

*Tabla 3. Valores para la preparación de mezclas basadas en la Formulación del producto libre de gluten*

| | | Cantidades | | | | Concentración | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Nombre de la Muestra | Almendra (g) | Arroz (g) | Avena (g) | Coco (g) | Almendra (%p) | Arroz (%p) | Avena (%p) | Coco (%p) |
| 1 | **MIX-65A** | 1.115 | 0.2 | 0.2 | 0.2 | 37.17 | 6.67 | 6.67 | 6.67 |
| 2 | **MIX-65B** | 1.115 | 0.2 | 0.2 | 0.2 | 37.17 | 6.67 | 6.67 | 6.67 |
| 3 | **MIX-68A** | 1.1662 | 0.183 | 0.183 | 0.183 | 38.87 | 6.10 | 6.10 | 6.10 |
| 4 | **MIX-68B** | 1.1662 | 0.183 | 0.183 | 0.183 | 38.87 | 6.10 | 6.10 | 6.10 |
| 5 | **MIX-70A** | 1.2 | 0.1715 | 0.1715 | 0.1715 | 40.00 | 5.72 | 5.72 | 5.72 |
| 6 | **MIX-70B** | 1.2 | 0.1715 | 0.1715 | 0.1715 | 40.00 | 5.72 | 5.72 | 5.72 |
| 7 | **MIX-72A** | 1.235 | 0.16 | 0.16 | 0.16 | 41.17 | 5.33 | 5.33 | 5.33 |
| 8 | **MIX-72B** | 1.235 | 0.16 | 0.16 | 0.16 | 41.17 | 5.33 | 5.33 | 5.33 |
| 9 | **MIX-75A** | 1.2865 | 0.143 | 0.143 | 0.143 | 42.88 | 4.77 | 4.77 | 4.77 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 10 | **MIX-75B** | 1.2865 | 0.143 | 0.143 | 0.143 | 42.88 | 4.77 | 4.77 | 4.77 |
| 11 | **MIX-100A** | 1.715 | 0 | 0 | 0 | 57.17 | 0.00 | 0.00 | 0.00 |
| 12 | **MIX-100B** | 1.715 | 0 | 0 | 0 | 57.17 | 0.00 | 0.00 | 0.00 |
| | **Rango** | - | - | - | - | 31.17-57.17 | 0-6.67 | 0-6.67 | 0-6.67 |
| | **Media** | - | - | - | - | 42.88 | 4.77 | 4.77 | 4.77 |

Fuente: autor

En la tabla 4 se especifica la composición nutrimental, es decir, el contenido de proteínas, lípidos y carbohidratos de las mezclas de harinas libres de gluten. Se tomó como referencia la base de datos de Productos Alimenticios de Marca del Departamento de Agricultura de los Estados Unidos (USDA) versión 3.7, siendo ésta reconocida internacionalmente y por tanto, aceptada para calcular los valores de composición bromatológica según la norma NOM-051-SCFI/SSA-2010 en su apartado 4.2.8.3.8.

*Tabla 4 Composición nutrimental de las mezclas de harinas libres de gluten*

| | Nombre de la Muestra | Proteínas (g/100 g) | Lípidos (g/100 g) | Carbohidratos (g/100 g) |
|---|---|---|---|---|
| 1 | ALM-40% | 15.93 | 24.61 | 49.00 |
| 2 | ALM-55% | 17.31 | 30.96 | 42.11 |
| 3 | ALM-64% | 18.13 | 34.76 | 37.97 |
| 4 | ALM-70% A | 18.68 | 37.30 | 35.22 |
| 5 | ALM-70% B | 18.68 | 37.30 | 35.22 |
| 6 | ALM-80% | 19.65 | 41.57 | 30.89 |
| 7 | ARR-40% | 12.64 | 14.86 | 60.71 |
| 8 | ARR-55% | 10.73 | 11.46 | 65.54 |
| 9 | ARR-64% | 9.59 | 9.41 | 68.43 |
| 10 | ARR-70% | 8.82 | 8.05 | 70.36 |
| 11 | ARR-80% | 7.62 | 5.88 | 73.76 |
| 12 | AVE-40% | 15.14 | 16.11 | 57.71 |
| 13 | AVE-55% | 15.73 | 13.96 | 59.54 |
| 14 | AVE-64% | 16.09 | 12.66 | 60.63 |
| 15 | AVE-70% | 16.32 | 11.80 | 61.36 |
| 16 | AVE-80% | 16.77 | 10.46 | 62.78 |
| 17 | COAR-05153050A | 12.27 | 17.46 | 59.04 |
| 18 | COAR-05153050B | 12.27 | 17.46 | 59.04 |
| 19 | COAR-15305005A | 18.36 | 29.46 | 42.79 |
| 20 | COAR-15305005B | 18.36 | 29.46 | 42.79 |
| 21 | COAR-25% A | 14.56 | 18.26 | 55.89 |
| 22 | COAR-25% B | 14.56 | 18.26 | 55.89 |
| 23 | COAR-30500515A | 14.86 | 10.72 | 62.71 |
| 24 | COAR-30500515B | 14.86 | 10.72 | 62.71 |
| 25 | COAR-50051530A | 12.73 | 15.40 | 59.03 |
| 26 | COAR-50051530B | 12.73 | 15.40 | 59.03 |
| 27 | COCO-40% | 14.50 | 17.47 | 56.14 |
| 28 | COCO-55% | 14.45 | 16.67 | 56.39 |

| | | | | |
|---|---|---|---|---|
| 29 | COCO-64% | 14.42 | 16.20 | 56.54 |
| 30 | COCO-70% | 14.40 | 15.88 | 56.64 |
| 31 | COCO-80% | 14.42 | 15.43 | 57.03 |
| 32 | AROC-07136020 | 15.51 | 11.02 | 62.33 |
| 33 | AROC-20071360 | 15.49 | 19.64 | 52.62 |
| 34 | AROC-60200713 | 16.94 | 32.63 | 40.84 |
| 35 | AROC-13602007A | 10.29 | 9.75 | 67.79 |
| 36 | AROC-13602007B | 10.29 | 9.75 | 67.79 |
| 37 | AROC-35451111 | 12.82 | 19.46 | 56.51 |
| 38 | AROC-11354511 | 13.45 | 10.86 | 64.29 |
| 39 | AROC-11113545 | 15.11 | 14.54 | 58.32 |
| 40 | AROC-45111135 | 16.83 | 28.18 | 44.45 |
| | **Rango** | 7.62-19.65 | 5.88-41.57 | 30.89-73.76 |
| | **Media** | 14.56 | 18.53 | 55.70 |

Fuente: USDA (2016), autor.

### *3.1.2 Suplementos alimenticios*

Se adquirieron en distintos supermercados y farmacias de la ciudad de Mexicali, B. C., un total de 37 suplementos alimenticios de presentación líquida en botella de 237 mL y en tres sabores: vainilla, fresa y chocolate. Fueron de distinto lote como se especifica en la tabla 5 y de cinco marcas diferentes: Ensure®, Ensure Advance®, Ensure Plus®, Glucerna® y Pediasure®, todos ellos producidos por el laboratorio Abbott.

*Tabla 5 Descripción de los Suplementos alimenticios*

| | ID de la Muestra | Marca | Sabor | Lote |
|---|---|---|---|---|
| 1 | **EAV001** | Ensure Advance® | Vainilla | 48420RR3550003 |
| 2 | **EAV002** | Ensure Advance® | Vainilla | 51142RR0661940 |
| 3 | **EAV003** | Ensure Advance® | Vainilla | 54809RR1742152 |
| 4 | **EAV004** | Ensure Advance® | Vainilla | 54809RR1750113 |
| 5 | **EAV005** | Ensure Advance® | Vainilla | 54809RR1750336 |
| 6 | **EC001** | Ensure® | Chocolate | 50835RR0470335 |
| 7 | **EC002** | Ensure® | Chocolate | 54713RR11641339 |
| 8 | **EF001** | Ensure® | Fresa | 56201RR2342033 |
| 9 | **EF002** | Ensure® | Fresa | 57316RR2480052 |
| 10 | **EPC001** | Ensure Plus® | Chocolate | 55067RR2071009 |
| 11 | **EPF001** | Ensure Plus® | Fresa | 54731RR11790041 |
| 12 | **EPV001** | Ensure Plus® | Vainilla | 48416RR3522336 |
| 13 | **EPV002** | Ensure Plus® | Vainilla | 51114RR0790807 |

| | | | | |
|---|---|---|---|---|
| 14 | **EPV003** | Ensure Plus® | Vainilla | 54803RR1731703 |
| 15 | **EPV004** | Ensure Plus® | Vainilla | 54803RR1732021 |
| 16 | **EV001** | Ensure® | Vainilla | 49555RR0090906 |
| 17 | **EV002** | Ensure® | Vainilla | 51054RR0760202 |
| 18 | **EV003** | Ensure® | Vainilla | 51054RR0760209 |
| 19 | **EV004** | Ensure® | Vainilla | 51054RR0760927 |
| 20 | **EV005** | Ensure® | Vainilla | 51054RR0760932 |
| 21 | **EV006** | Ensure® | Vainilla | 51177RR0890257 |
| 22 | **EV007** | Ensure® | Vainilla | 51177RR0891829 |
| 23 | **EV008** | Ensure® | Vainilla | 56252RR2440347 |
| 24 | **EV009** | Ensure® | Vainilla | 57530RR2760436 |
| 25 | **GC001** | Glucerna® | Chocolate | 52317RR0940515 |
| 26 | **GC002** | Glucerna® | Chocolate | 54738RR1771455 |
| 27 | **GF001** | Glucerna® | Fresa | 52261RR1030630 |
| 28 | **GV001** | Glucerna® | Vainilla | 51124RR0801123 |
| 29 | **GV002** | Glucerna® | Vainilla | 52252RR0990302 |
| 30 | **GV003** | Glucerna® | Vainilla | 57341RR2560158 |
| 31 | **GV004** | Glucerna® | Vainilla | 57444RR2521919 |
| 32 | **PC001** | Pediasure® | Chocolate | 52322RR0951855 |
| 33 | **PF001** | Pediasure® | Fresa | 52260RR1031626 |
| 34 | **PV001** | Pediasure® | Vainilla | 51184RR0902322 |
| 35 | **PV002** | Pediasure® | Vainilla | 54729RR1691823 |
| 36 | **PV003** | Pediasure® | Vainilla | 54729RR1692229 |
| 37 | **PV004** | Pediasure® | Vainilla | 54825RR1801403 |
| | **Totales** | 5 Ensure Advance®<br>6 Ensure Plus ®<br>13 Ensure®<br>7 Glucerna®<br>6 Pediasure® | 26 Vainilla<br>6 Chocolate<br>5 Fresa | |

Fuente: autor

Además, en la tabla 6 se especifica el contenido de carbohidratos, lípidos y proteínas de los suplementos alimenticios según su nombre comercial. Se tomó como referencia la base de datos de Nutrientes del Departamento de Agricultura de los Estados Unidos (USDA) para Referencia Estándar 28 versión del software 3.7, siendo ésta reconocida internacionalmente y por tanto, aceptada para obtener los valores de composición bromatológica según la norma NOM-051-SCFI/SSA-2010 en su apartado 4.2.8.3.8.

*Tabla 6 Composición nutricional de los Suplementos Alimenticios*

| Nombre Comercial | No. de piezas | Proteínas (g/100mL) | Carbohidratos (g/100 mL) | Lípidos (g/100 mL) |
|---|---|---|---|---|
| **Ensure Advance®** | 5 | 5.48 | 13.5 | 3.38 |
| **Ensure Plus®** | 6 | 5.54 | 21.34 | 4.85 |

| | | | | |
|---|---|---|---|---|
| **Ensure®** | 13 | 4.07 | 18.12 | 2.71 |
| **Glucerna®** | 7 | 4.6 | 12.2 | 3.4 |
| **Pediasure®** | 6 | 3.07 | 11.97 | 5.12 |
| **Rango** | - | 3.07-5.54 | 11.97-21.34 | 2.71-5.12 |
| **Media** | - | 4.43 | 15.90 | 3.67 |

Fuente: USDA (2016), autor

Las muestras con un alto porcentaje de humedad, como estos suplementos alimenticios, no son útiles para desarrollar modelos de calibración en NIR (Burns, 2008). Para demostrar lo anterior, se obtuvieron espectros FT-NIR de la muestra EAV001 antes y después de deshidratarse que pueden apreciarse en la figura 16.

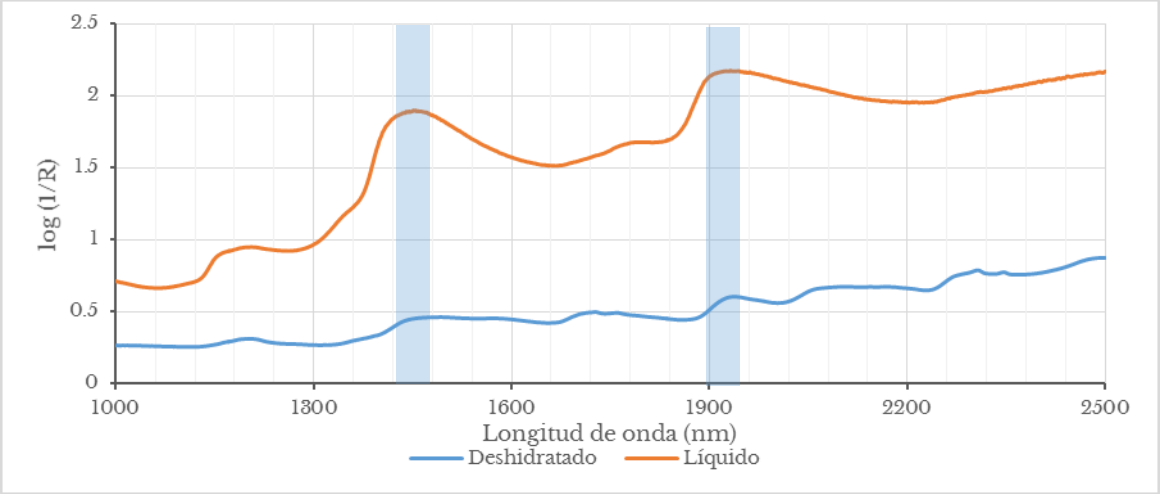



*Figura 16 Espectros FT-NIR de la muestra EAV001 antes y después de someterse al proceso de deshidratación*

Fuente: autor

En la muestra deshidratada se redujo la intensidad de las bandas de absorción del agua que se encuentran en la región de 1450 y 1940 nm, lo que permitió que se manifestaran otras señales que no son visibles en la muestra líquida y que la línea base disminuyera. De esta manera, se confirmó lo dicho en la literatura. Por estas razones, se decidió someter a un proceso de deshidratación a todos los suplementos alimenticios.

Se distribuyeron 30 mL de cada producto en un molde de silicón, luego se colocaron en la incubadora marca Barnstead Lab-Line modelo 120 durante 48 horas a

45°C. La figura 17a muestra el resultado del proceso de deshidratación. Posteriormente, se pulverizaron con un mortero y pistilo, se pasaron por un colador (figura 17b) cuyo poro era de aproximadamente 1 mm para que el tamaño de la partícula fuera similar. Se almacenaron en una bolsa de polietileno y mantuvieron en refrigeración a 4°C hasta su análisis.

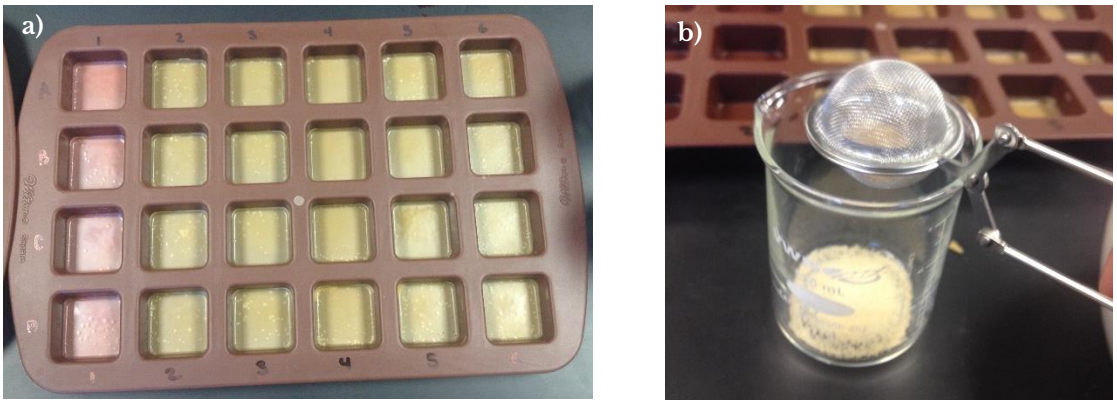



*Figura 17 Proceso de deshidratación de los suplementos alimenticios*

a) Al salir de la incubadora, b) Tamizado.
Fuente: autor

## 3.2 Obtención de espectros FT-NIR

Se utilizó el espectrofotómetro Frontier FT-NIR PerkinElmer® equipado con el accesorio de reflectancia (NIRA) y detector INGAAS, así como el software Spectrum versión 10.4.2 en donde se especificaron las condiciones de obtención de los espectros NIR de las muestras de harinas libres de gluten y de los suplementos alimenticios. Los parámetros fueron los siguientes: supresión de bandas de $CO_2/H_2O$, velocidad de escaneo de 1 cm/s, fondo (background) como corrección de fase y apodización fuerte. Las unidades de la abscisa se establecieron en número de onda y las de ordenadas en log (1/R). La resolución fue de 4 $cm^{-1}$ y las acumulaciones en 32 escaneos en el rango de 10,000 a 4,000 $cm^{-1}$ para dar un total de 6001 puntos de datos.

Se colocó un vial de vidrio de 15 mm de diámetro y 44 mm de altura, conteniendo 1 mL de la muestra a analizar sobre el NIRA del espectrofotómetro como se observa en la figura 18. Triplicados de cada muestra fueron escaneados, rotando el vial un aproximado de 45 grados en cada ocasión. Se promediaron los espectros con el

software Spectrum y el resultante fue empleado para el desarrollo de los modelos de calibración (Kim, Singh y Kays, 2007).

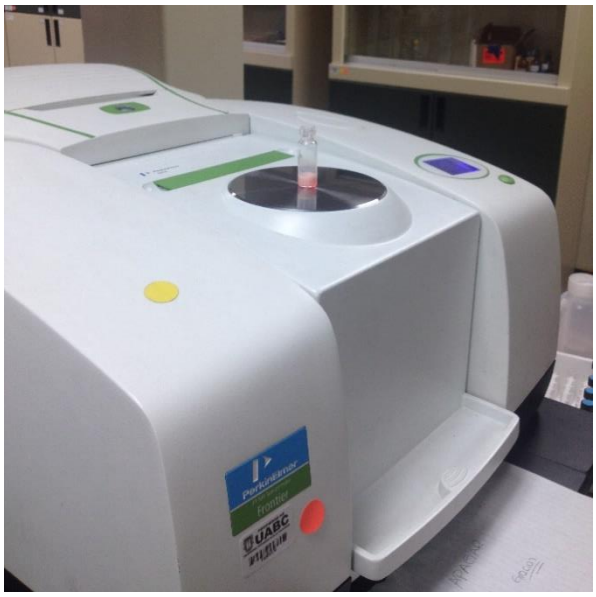


*Figura 18 Espectrofotómetro FT-NIR*

Fuente: autor

## 3.3 Identificación de bandas espectrales

Con el propósito de obtener una mejor interpretación de las señales vibracionales, se realizó la conversión del número de onda ($cm^{-1}$) a longitud de onda (nm) con el software Spectrum. Esto también permitió reducir el número de puntos de datos de 6001 a 1501.

Para determinar las regiones donde aparecen las bandas de absorción características de cada macronutriente, se colectaron espectros FT-NIR de distintos carbohidratos, proteínas y lípidos. Los resultados obtenidos sirvieron como base para la interpretación espectral de las muestras de harinas libres de gluten y de los suplementos alimenticios.

La figura x contiene los espectros FT-NIR sin pretratamientos de la caseína y de la proteína soya, ambas en estado sólido. En ella se resaltan las bandas exclusivas de esta biomolécula, es decir, aquellas que presenten cualquier vibración entre enlaces N-H.

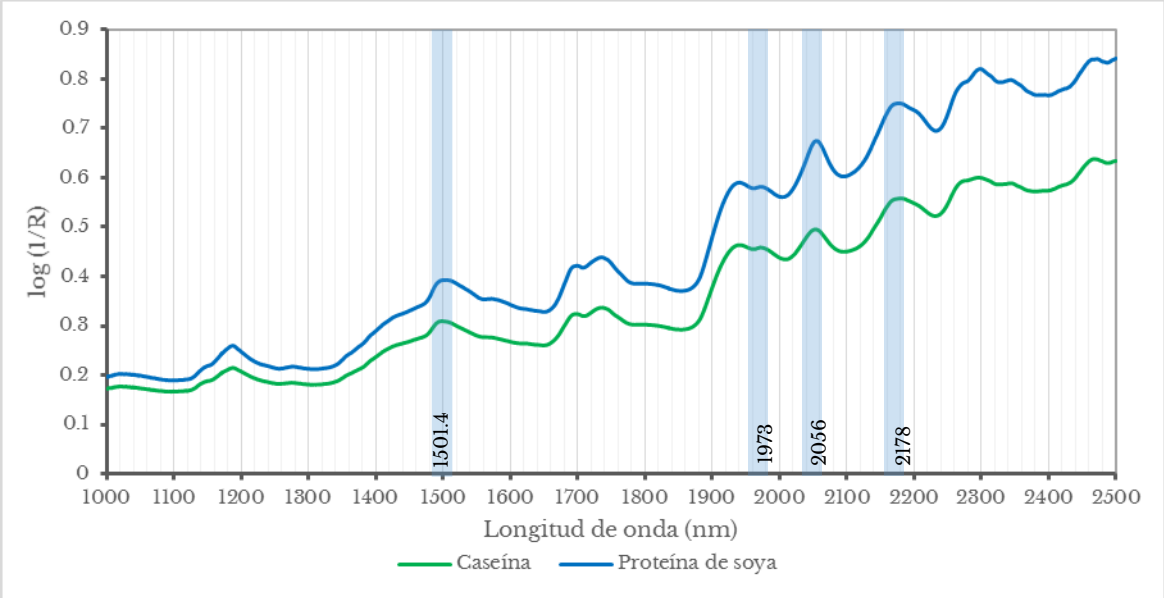



*Figura 19 Espectros FT-NIR con la asignación de las bandas características de las proteínas*

Fuente: autor

Se identificaron en los espectros cuatro señales asociadas a las proteínas. La primera fue la de 1501.4 nm, correspondiente al primer sobretono del enlace N-H. La segunda fue la de 1973 nm, asociada a la combinación de la tensión y flexión de N-H. La siguiente fue la región que comprende entre los 2050 a 2060 nm, asignada a la combinación de la tensión de los enlaces N-H y C=O. Finalmente, la banda de absorción entre los 2168 a 2180 es adjudicada al segundo sobretono del enlace N-H, así como al primer sobretono de la tensión C=O y a la combinación de tensión C-N y flexión de N-H (Workman, 2012).

En cuanto a los carbohidratos, la figura 20 muestra los espectros FT-NIR sin pretratamientos de la glucosa, sacarosa, goma xantana y maltodextrosa en estado sólido, siendo el primero un representante de los monosacáridos, el segundo de los disacáridos y los últimos, de los polisacáridos. Además, se resaltan las bandas de absorción que tienen en común entre ellos.

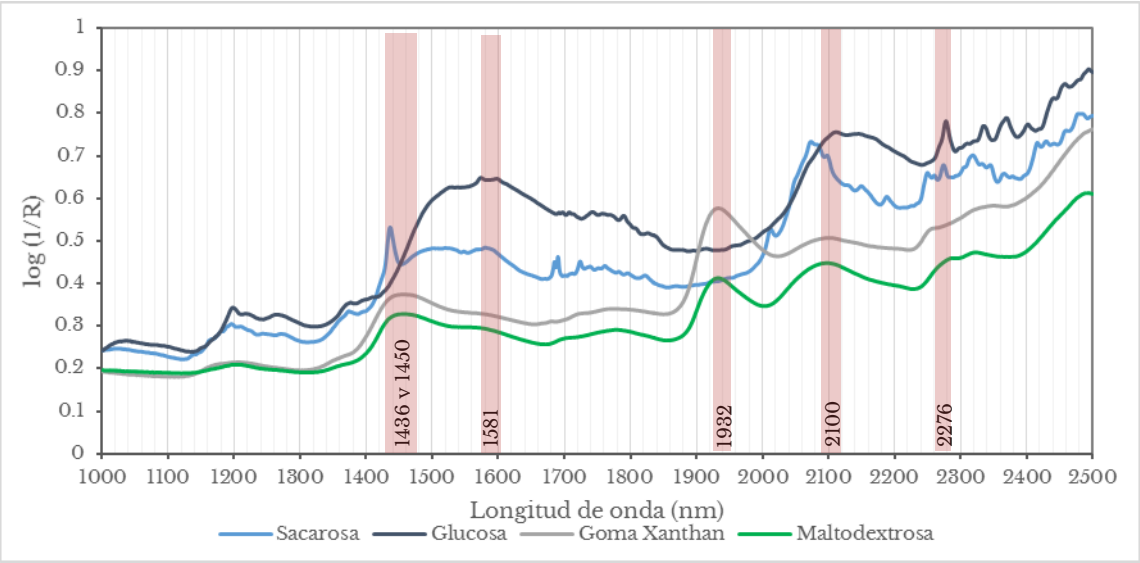



*Figura 20 Espectros FT-NIR con la asignación de las bandas características de los carbohidratos*

Fuente: autor

La banda de absorción de los 2100 nm se manifiesta en las cuatro muestras, así como la de los 1581 nm. Por tanto, es una región que debe buscarse en las harinas y suplementos alimenticios para verificar la presencia de los carbohidratos. Por otro lado, los polisacáridos goma xantana y maltodextrosa exhiben una señal intensa en 1450 y 1930 nm, ausentes en la glucosa y sacarosa. Sin embargo, estas bandas suelen ser asociadas a cualquier tipo de esta biomolécula (Workman, 2012). De esta manera, también se consideraron para identificar a los azúcares. El mismo criterio se aplicó a la región entre 2276 a 2280 nm que sólo fue positivo para la glucosa, sacarosa y maltodextrina.

La sacarosa exhibió una señal intensa en la longitud de onda 1436nm. Sin embargo no se asocia a ningún tipo de vibración. Se considera entonces que existe un desplazamiento de la banda debido a que una señal en los 1441 nm es asignada a la sacarosa cristalina (Workman, 2012).

Para identificar las señales vibracionales características de los lípidos se colectaron espectros FT-NIR, que se exhiben en la figura x, de mono y diglicéridos, así como de dos productos alimenticios ricos en grasas: huevo entero y mantequilla, cuya presentación para ambos casos fue en polvo.

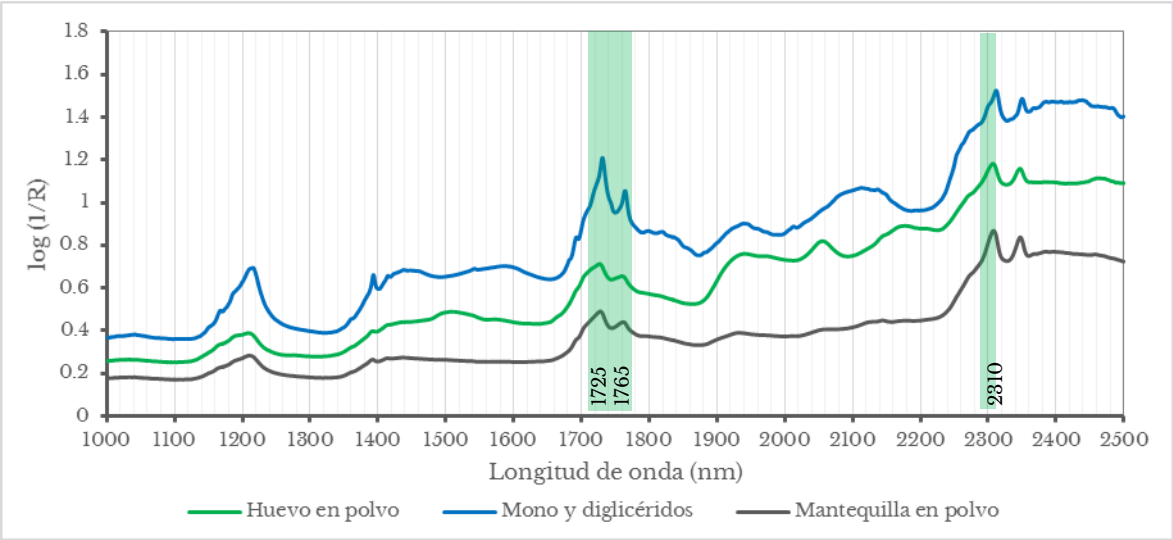



*Figura 21 Espectros FT-NIR con la asignación de las bandas características de los lípidos*

Fuente: autor

En las tres muestras se aprecian bandas intensas en las longitudes de onda 1725 y 1765 nm que corresponden al primer sobretono del estiramiento del enlace C-H de los metilenos ($CH_2$) y otra en la región de 2310 nm asociado al doblamiento del enlace C-H de los aceites (Workman, 2012).

En la tabla x se especifican las regiones características de las proteínas, carbohidratos y lípidos, junto con la descripción del tipo de vibración, y el grupo funcional al que se le asigna esta señal.

*Tabla 7 Bandas de absorción características de las proteínas, carbohidratos y lípidos en el espectro NIR*

| | Longitud de onda (nm) | | Tipo de vibración | Asociado a |
|---|---|---|---|---|
| | Real | Teórico | | |
| Proteínas | 1501.4 | 1502.5 | 1er sobretono de N-H | Amina aromática primaria (*p*-$NH_2$) |
| | 1973 | 1967-1977.5 | Combinación de tensión y flexión de N-H | Amina aromática primaria o-$NO_2$ |
| | 2056 | 2055 | Combinación de tensión N-H y tensión C=O | Amidas: .CONH. y .$CONH_2$ |
| | 2178 | 2180 | 2do sobretono de N-H, 1er sobretono de tensión C=O; combinación de tensión C-N y flexión de N-H | N-H/C-N/C=O de amidas secundarias |
| C | 1436 | 1441 | 1er sobretono O-H | Sacarosa Cristalina, C4-OH. |

| | | | | |
|---|---|---|---|---|
| | 1450 | 1450 | 1er sobretono O-H polimérico (.O-H) | Almidones/Alcoholes poliméricos |
| | 1581 | 1580 | 1er sobretono del puente de hidrógeno O-H | Alcoholes R-C-O-H |
| | 1932 | 1930 | Combinación de la tensión O-H y flexión H-O-H | Polisacáridos |
| | 2100 | 2100 | Combinación de la flexión O-H y tensión C-O, 3er sobretono de C=O-O | Glucosa, Polisacáridos |
| | 2276 | 2273, 2280 | Combinación de tensión O-H y tensión C-O. Combinación de tensión de C-H y deformación de C-H | Glucosa, Polisacáridos |
| **Lípidos** | 1724-1726 | 1725 | 1er sobretono de la tensión C-H | Metilenos (.$CH_2$) |
| | 1764 | 1765 | 1er sobretono de la tensión C-H | Metilenos (.$CH_2$) |
| | 2307-2311 | 2310 | 2do sobretono de la flexión C-H | Aceites, lípidos |

Fuente: Burns (2008), Workman (2012).

## 3.4 Análisis Exploratorio de Datos

Previo al desarrollo de los modelos de calibración, se necesita revisar la información recabada por los espectros con distintos propósitos:

- Determinar el número óptimo de componentes principales o variables latentes. Este es un parámetro que el software Spectrum solicita al momento de desarrollar un modelo con el algoritmo PLS.
- Identificar anormalidades como muestras distintas al resto (outliers) o mala calidad de los espectros.
- Seleccionar los objetos que formarán parte del set de calibración y de validación independiente.

El análisis de componentes principales, PCA por sus siglas en inglés, permite dar respuesta a los puntos anteriores. Para calcularlo, se empleó el software The Unscrambler® X y los valores de los parámetros solicitados para el cómputo fueron los siguientes:

- Número de componentes principales: 10. Se considera que un modelo robusto o "fácil de interpretar" debe utilizar el menor número de factores para su cálculo. Por tanto, se consideró suficiente emplear hasta diez componentes para el cómputo de PCA.

- Rango espectral: 1000 a 2500 nm, con un total de 1501 variables. Aunque sólo interese emplear una sola región del espectro para el desarrollo de un modelo, se recomienda emplear todo el rango NIR en este punto.
- Número de muestras: Todas las pertenecientes al grupo a evaluar.
- Escalado: Centrado a la media y dividido por la desviación estándar del espectro. De esta manera todas las longitudes de onda tienen el mismo peso y ninguna es más influyente que otra.
- Validación: Cruzada "dejar uno fuera".
- Algoritmo: Valor Singular de Descomposición (SVD, por sus siglas en inglés). Recomendado por el software cuando hay pocas muestras y muchas variables. En este estudio se manejan una matriz de 40 x 1501.
- Detección de Anormalidades "Outliers": Residuales-F y distribución T2 de Hotelling. Los límites para los dos parámetros fueron los sugeridos por los establecidos por el software: 5%.

El software devuelve distintas tablas y gráficos que se necesitan interpretar, basadas en el número óptimo de componentes calculado. Los tomados en cuenta para el análisis fueron el de Puntuaciones y el de Influencia.

El gráfico de Puntuaciones, mejor conocido como "Scores", informa sobre los patrones de las muestras al realizar un examen visual. Relaciona los resultados obtenidos del componente principal 1 (PC1) contra el componente principal 2 (PC2). Si los objetos se encuentran distribuidos a lo largo de los ejes, significa que son representativos para el desarrollo del modelo. Si se aprecian agrupaciones, estos se consideran similares y si aparecen otros muy alejados al resto, se califican como diferentes y, por tanto, como posibles anomalías u "outliers". Además, este gráfico especifica el porcentaje de varianza de cada componente, entre mayor sea el porcentaje, menor la cantidad de componentes se tendrán que utilizar para el desarrollo de los modelos.

El gráfico de Influencia relaciona la distribución $T^2$ de Hotelling con los residuales-F permitiendo así identificar a los objetos atípicos. Si una muestra sale de los límites establecidos, significa que es diferente al resto de la población y por tanto, se considera como un "outlier". Ante esos casos se siguió el procedimiento descrito en la figura 22.

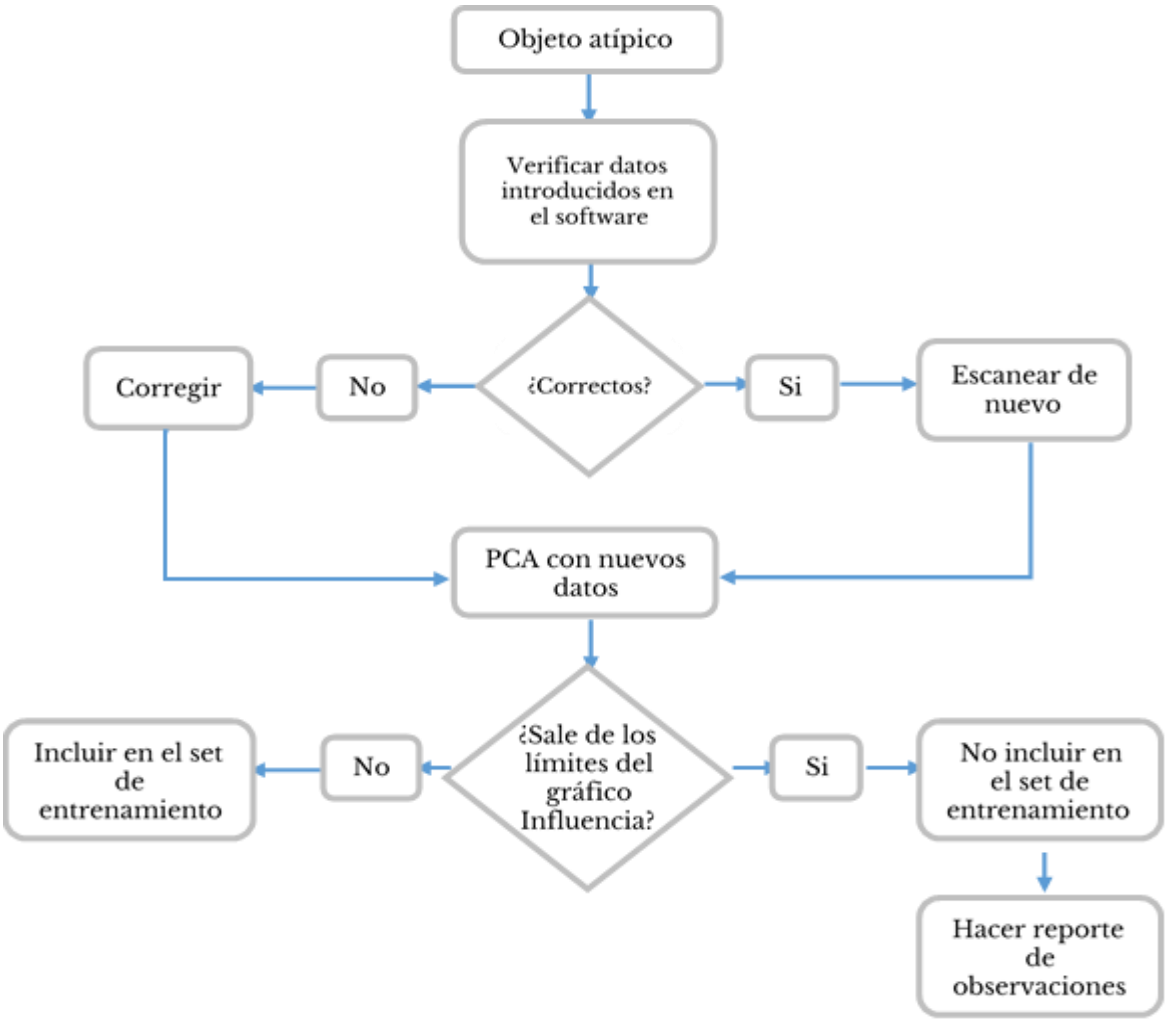



*Figura 22 Proceso a seguir en caso de detectar objetos anómalos "outliers"*

Fuente: autor.

El reporte de observaciones consiste en describir todas las características físicas de la muestra, así como la composición química. Esto permitirá tener un mayor entendimiento del por qué un objeto en particular es distinto al resto de la población.

Finalmente, para seleccionar las muestras que formarían parte del set de calibración y de prueba, se aplicó el algoritmo Kennard – Stone (KS) sobre las puntuaciones obtenidas del componente principal óptimo calculado para elegir a aquellas que sean más representativas. Este algoritmo permite elegir a aquellos objetos que se encuentren distribuidos uniformemente sobre el espacio multivariante para formar parte del grupo que se utilizará para la calibración del modelo (CAMO Software AS, 2016).

Aunque no existe una regla para determinar el número de muestras pertenecientes a cada set, se recomienda que en el de entrenamiento, también llamado de calibración, esté integrado por el 70% del total de la población y el de prueba por el 30% (Mucherino, Papajorgji y Pardalos, 2009).

## 3.5 Desarrollo de modelos de calibración

El software empleado para el desarrollo de los distintos modelos de calibración fue The Unscrambler® X versión 10.4.

El desarrollo de un modelo se podría resumir en tres pasos:

1. Construcción del set de datos: Relacionar los datos espectrales de una muestra con los valores de referencia.
2. Entrenamiento del modelo: Con el set de calibración aplicando técnicas de regresión multivariante y control de anomalías u "outliers".
3. Evaluación: Validación cruzada e independiente. Éste último con el set de prueba (Burns, 2008)

Sin embargo, la metodología no es tan sencilla como se resume en el párrafo anterior. A continuación, se detallan los parámetros, algoritmos e indicadores de desempeño que fueron empleados en el desarrollo de los modelos.

### *3.5.1 Modelos cualitativos*

Se utilizó el método de análisis de discriminación con la regresión de mínimos cuadrados parciales (PLS-DA, por sus siglas en inglés) para crear modelos que permitieran clasificar a las muestras según una categoría. Se empleó un código para discriminar entre una clase y otra: +1 para las que pertenecieran a ella y -1 a las que no. De esta manera, la línea de decisión fue cero facilitando la visualización en el gráfico de regresión (CAMO Software, 2016).

Para el grupo de muestras de harinas libres de gluten, se creó un modelo que pudiera diferenciar entre aquellas mezclas que estaban formadas únicamente por harinas de las que contenían, además de las harinas, otros ingredientes. A las primeras se les asignó la clave +1 y a las otras, -1.

Con el grupo de los suplementos alimenticios se crearon dos modelos, uno que clasificara a las muestras según la marca comercial y otro por el sabor. El primero contenía cinco categorías, una por cada marca: Ensure®, Ensure Advance®, Ensure Plus®, Glucerna® y Pediasure®, y el segundo tres, uno por cada sabor: vainilla, fresa y chocolate. Al igual que en las harinas, a las pertenecientes a una clase se les asignó el valor de +1 y a las que no, de -1.

Los parámetros para la construcción del modelo de regresión PLS-DA fueron los siguientes:

- Número de componentes principales: 10
- Rango espectral: 1000 a 2500 nm, con un total de 1501 variables.
- Número de muestras: 20, pertenecientes al set de calibración.
- Escalado: Centrado a la media y dividido por la desviación estándar del espectro.
- Algoritmo: Kernel PLS.
- Prueba de incertidumbre: Activada, se seleccionó la opción de utilizar el número óptimo de componentes para su cálculo. Esto permitirá identificar a las longitudes de onda importantes.
- Validación: Cruzada “dejar uno fuera”.
- Detección de Anormalidades “Outliers”: Residuales-F y distribución $T^2$ de Hotelling. Los límites para los dos parámetros fueron los establecidos por el software: 5%.

El software generó distintas tablas y gráficos con los resultados del cómputo. Para la evaluación del desempeño del modelo, se utilizó el gráfico que relaciona los valores de Referencia contra los de Predicción y se verificó que las muestras pertenecientes a la clase +1 superaran la línea de decisión, es decir, fueran mayores que cero y que las de la clase -1 fueran menores que cero. Asimismo, se revisó que los valores de la desviación estándar estimados no tocaran la línea divisoria.

Si todas las muestras fueron clasificadas correctamente, tanto en la calibración como en la validación cruzada, el modelo se considera listo para ser aplicado al set de prueba para demostrar si es apto para medir muestras desconocidas, como se indica en la fig. 23 que consiste en un resumen del proceso de desarrollo y evaluación del modelo cualitativo.

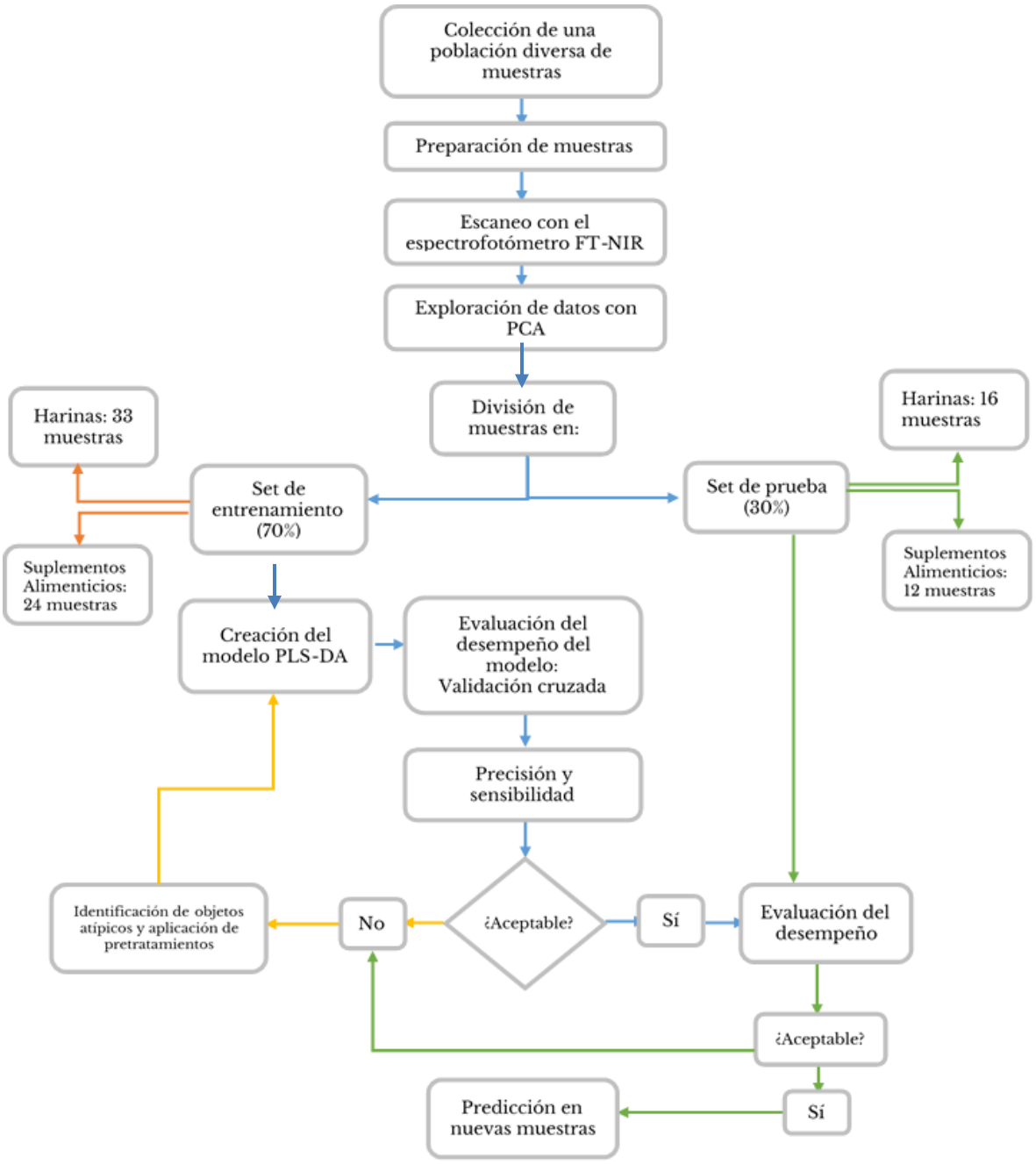



*Figura 23 Diagrama de flujo del desarrollo y evaluación del modelo cualitativo*

Fuente: autor

En el caso de que algún objeto fuera clasificado erróneamente en la calibración, en la validación cruzada o en el set de prueba, se llevó a cabo el siguiente procedimiento:

1. Revisión del gráfico de Influencia para determinar si se trataron de objetos anómalos u "outliers". De ser afirmativo, se hizo un registro de las muestras.
2. Someter a los datos espectrales a la segunda derivada Savitzky-Golay para corregir la línea base y el solapamiento de las bandas de absorción.
3. Repetir el cómputo del modelo PLS-DA y evaluarlo.

Si después de realizar el proceso anterior no mejoró la capacidad de clasificación, se realizó un registro que incluyó el número de objetos clasificados erróneamente junto con sus características físicas y composición química. De esta manera, se recabaría información importante para establecer las limitaciones del modelo PLS-DA.

### *3.5.2 Modelos cuantitativos*

Se utilizó el método de regresión de mínimos cuadrados parciales (PLS, por sus siglas en inglés) para crear modelos que permitieran determinar el contenido de proteínas, carbohidratos y lípidos en las mezclas de harinas libres de gluten y en los suplementos alimenticios. Asimismo, se desarrolló otro modelo que midiera la concentración de cada tipo de harina presente en la mezcla de harinas libres de gluten.

Los parámetros para la construcción de los modelos de regresión PLS, tanto para las harinas como para los suplementos alimenticios, fueron los siguientes:

- Número de componentes principales: 10
- Rango espectral: 1000 a 2500 nm, con un total de 1501 variables.
- Número de muestras: 20, las pertenecientes al set de calibración.
- Escalado: Centrado a la media y dividido por la desviación estándar del espectro.
- Algoritmo: Kernel PLS.
- Prueba de incertidumbre: Activada, se seleccionó la opción de usar el número óptimo de componentes para su cálculo.
- Validación: Cruzada "dejar uno fuera".

- Detección de Anormalidades “Outliers”: Residuales-F y distribución $T^2$ de Hotelling. Los límites para ambos parámetros fueron los sugeridos por el software: *5%*.

El software devolvió los resultados del cómputo en forma de tablas y gráficos. Los seleccionados para evaluar el desempeño del modelo fueron los siguientes:

- **Gráfico de Regresión.** Relaciona los valores de referencia contra los de predicción, lo que permite verificar la calidad del modelo. Lo ideal es que los objetos se encuentren sobre la línea de regresión, que ésta pase por cero y la pendiente sea próxima a 1.
- **Estadísticas de la Regresión**. Esta tabla contiene los valores de la pendiente, intercepto, error medio cuadrático (RMSE, por sus siglas en inglés) y la correlación lineal ($R^2$).
- **Gráfico de RMSE.** Permite visualizar el valor del error medio cuadrático a través de los componentes principales (PC, por sus siglas en inglés) utilizados en el cálculo. De esta manera, se corrobora que el número óptimo de PCs calculados por el software no sobrealimenta al modelo.
- **Gráfico de Residuales Y.** Relaciona los valores residuales contra los predichos. Si un modelo predice adecuadamente a Y, por ejemplo: el contenido de proteínas, los residuos de las variaciones serán próximos a cero. Si no es así, significa que el modelo no es del todo satisfactorio y que describe pobremente al objeto que se encuentra alejado de cero. Por tanto, dicha muestra es considerada como atípica u “outlier”.
- **Gráfico de Coeficientes de Regresión.** Al haber activado la Prueba de incertidumbre, este gráfico marca las longitudes de onda relevantes para el modelo.

Si el valor de $R^2$ estuvo alejado de 1 y el de RMSE de cero, tanto para la calibración, validación cruzada e independiente, significó que el modelo no fue apropiado y se siguió el siguiente procedimiento:

1. Revisión del gráfico de Influencia para determinar si los objetos alejados de la línea de regresión se trataron de objetos anómalos u “outliers”. De ser afirmativo, se hizo un registro de las muestras.
2. Someter a los datos espectrales a distintos pretratamientos.
3. Repetir el cómputo del modelo PLS y evaluarlo.

La figura 24 resume el proceso seguido para desarrollar los modelos de cuantificación para las muestras de las harinas libres de gluten y para los suplementos alimenticios.

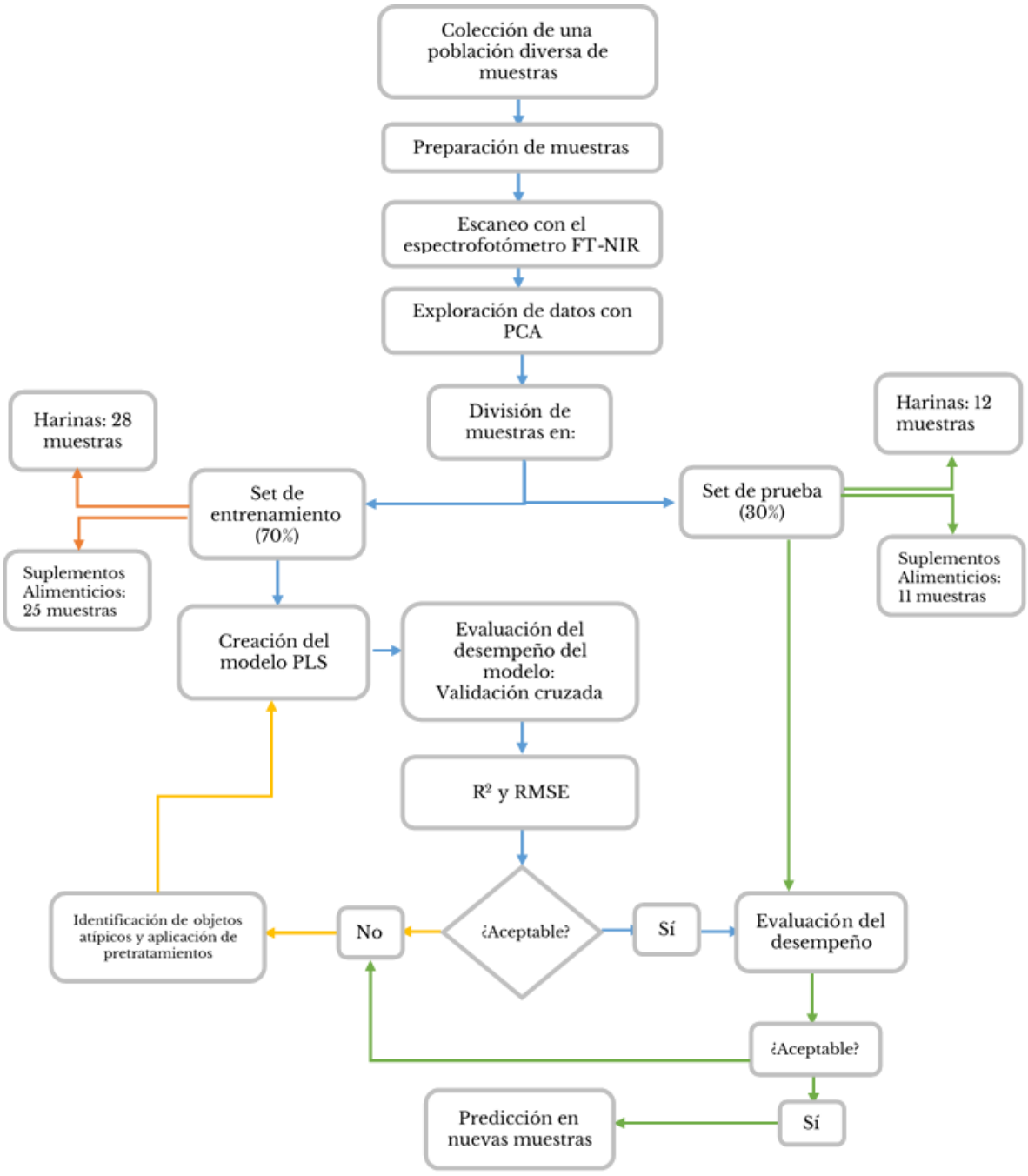



*Figura 24 Diagrama de flujo del desarrollo y evaluación de modelos cuantitativos*

Fuente: autor

*“El mundo del hombre contemporáneo se funda sobre los resultados de la ciencia: el dato reemplaza al mito, la teoría a la fantasía, la predicción a la profecía”.*

*Mario Bunge*

# 4

# Resultados

En este apartado se expone el desempeño obtenido de los distintos modelos cualitativos y cuantitativos desarrollados según los parámetros descritos en el capítulo anterior, comenzando por aquellos diseñados para el grupo de Harinas libres de gluten y finalizando con los de Suplementos alimenticios.

La descripción de los resultados inicia con la interpretación de los espectros FT-NIR, continuando con el análisis de la exploración de datos, seguido de las características de los modelos cualitativos y cuantitativos seleccionados como óptimos, así como una explicación de su desempeño auxiliado de distintos gráficos.

Cada sección finaliza con una discusión de los resultados obtenidos que incluye un resumen de los parámetros empleados para el desarrollo de los modelos junto con los indicadores del desempeño.

## 4.1 Harinas libres de gluten

### *4.1.1 Espectros NIR*

La figura 25 contiene los espectros FT-NIR sin pretratamientos matemáticos de 50 mezclas de harinas libres de gluten, cuyas características se describen en las tablas x y x, que abarcan el rango de los 1000 a 2500 nm y la intensidad en log(1/R). En ella se distinguen cinco regiones asociadas a los carbohidratos, dos correspondientes a los lípidos y tres de las proteínas.

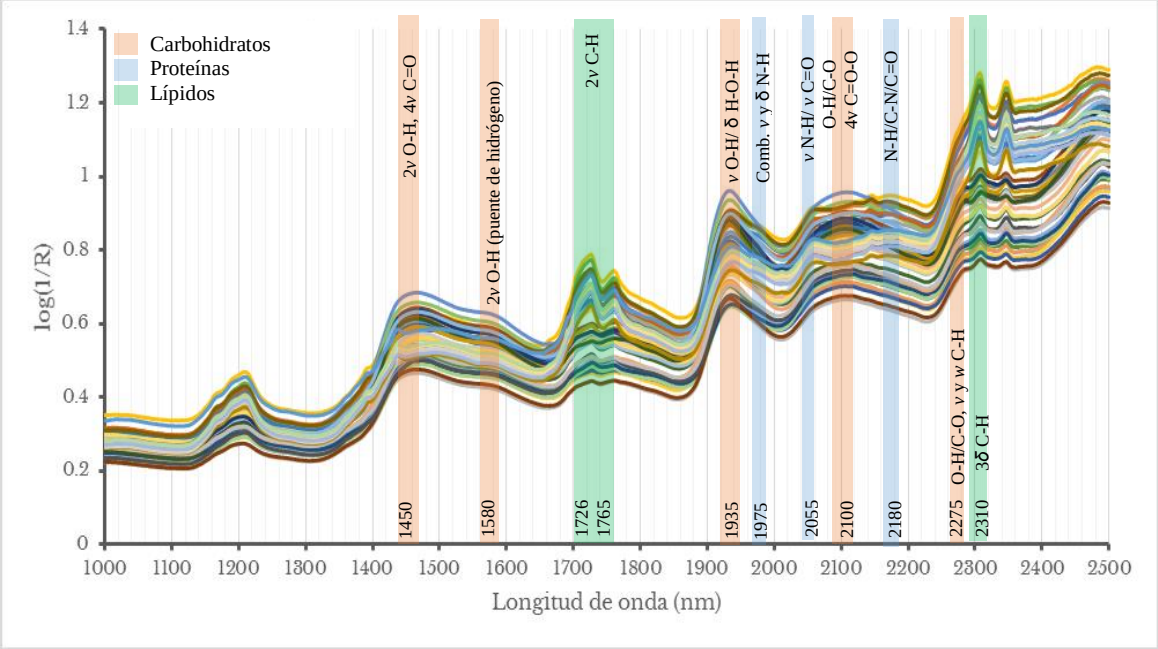



*Figura 25 Espectros FT-NIR de mezclas de Harinas libres de gluten*

*v*= tensión, *δ*= flexión, 2*v*= primer sobretono de tensión, 4*v*= tercer sobretono de tensión, 3*δ*=segundo sobretono de flexión, *w*=deformación
Fuente: autor

La intensidad de las señales varió entre los espectros de cada muestra analizada, algo esperado debido a que poseían diferente concentración de los analitos.

Al corroborarse visualmente la presencia de las biomoléculas que se pretenden cuantificar en el presente estudio, se prosiguió con la aplicación de técnicas quimiométricas para el desarrollo de los modelos de calibración.

### *4.1.2 Exploración de Datos*

El método PCA fue aplicado a los datos espectrales sin pretratamientos de las 50 muestras, según los parámetros establecidos en el capítulo anterior. Los resultados revelan que el número de componentes principales óptimo calculado fue de tres, en donde el PC1 explica el 89% de la varianza, el PC2 el 7% y el PC3 el 4% dando un total de1 100%. Esto significa que para construir los modelos de calibración a partir de los datos analizados, se requieren un mínimo de 3 variables latentes para su cálculo.

En el gráfico Puntuaciones (fig. 26a) se observa que los objetos de estudio se encuentran distribuidos sobre los ejes, lo que demuestra que éstos son representativos para el desarrollo de los modelos. Además, se aprecian cuatro agrupaciones de muestras.

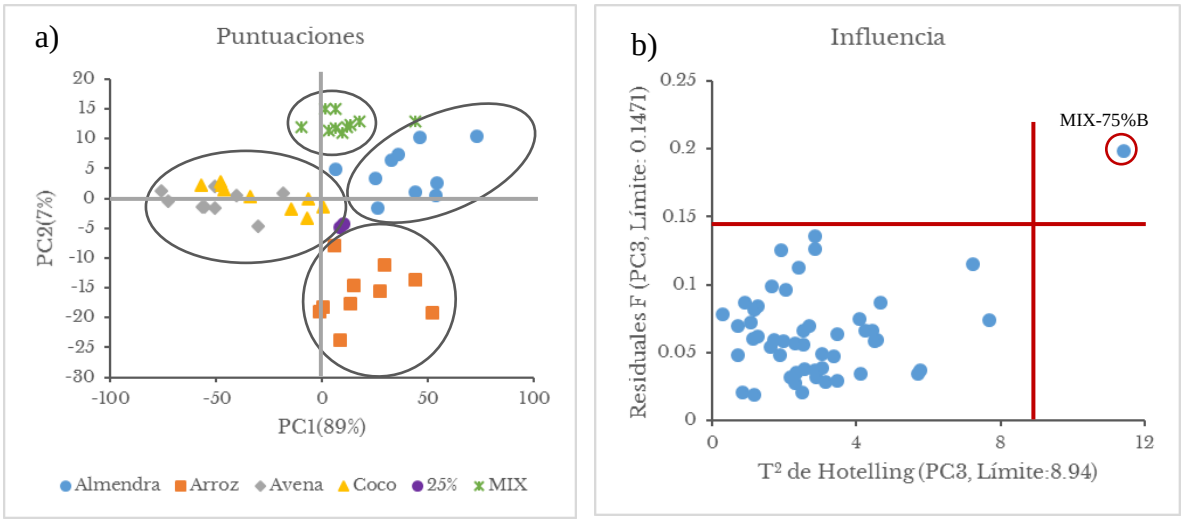



*Figura 26 Gráficos de Puntuaciones e Influencia de las Harinas libres de gluten*

Fuente: autor

Las ubicadas sobre le primer cuadrante corresponden a aquellas mezclas cuya composición porcentual mayoritaria fue de harina de almendra; las del cuarto cuadrante, de harina de arroz y las que están sobre el segundo y tercer cuadrante, las de avena y coco, confirmando que la concentración de carbohidratos de este tipo de harinas, entre 57 a 65g/100 g, provocó la semejanza de sus espectros. Finalmente, el último grupo lo conforman las mezclas de la formulación para hacer panecillos libres de gluten, etiquetados como MIX. Estos resultados revelan que existe el potencial de desarrollar modelos de clasificación.

En la búsqueda de posibles anomalías, se analizó el gráfico de Influencia (fig. 26b) y el objeto MIX-75%B se encontró fuera de los límites al presentar un alto valor residual F y de distribución $T^2$ de Hotelling en el componente principal 3. Se llevó a cabo el procedimiento descrito en el capítulo anterior para estos casos y al volver a obtener el mismo resultado, se le excluyó para el cálculo de los modelos. En la tabla 8 se describen las características de la muestra atípica para futuras referencias.

*Tabla 8 Reporte de objeto anómalo del grupo de las Harinas libres de gluten*

| Objeto anómalo "Outlier" | |
|---|---|
| **ID Muestra:** | MIX-75%B |
| **Límites excedidos:** | Residuales-F y $T^2$ de Hotelling |
| **Observaciones:** | Composición química y apariencia física similar a las otras muestras pertenecientes al grupo de harinas libres de gluten. La recolección del espectro fue en el mismo día que las otras y el vial empleado pertenece a los materiales utilizados exclusivamente para este propósito. Por tanto, se desconoce la causa de tan marcada diferencia al resto de los objetos. |

Fuente: autor

La selección de muestras que integrarían al set de entrenamiento y de prueba, se llevó a cabo con ayuda del algoritmo Kennard-Stone. Para el desarrollo de los modelos cualitativos se eligieron 33 objetos para el primer set (28 de harinas y 5 de la formulación), y 16 para el segundo (12 de harinas y 4 de la formulación).

En cuanto a los modelos cuantitativos, en la tabla 9 se especifica la estadística descriptiva de las macromoléculas analizadas (proteínas, carbohidratos, lípidos) en el grupo de las harinas, tanto en el set de calibración como en el de prueba.

*Tabla 9 Estadística descriptiva de las macromoléculas analizadas: Harinas*

| | Set de entrenamiento | | | Set de prueba | | |
|---|---|---|---|---|---|---|
| **Parámetro** | P | C | L | P | C | L |
| **Número de muestras** | 28 | 28 | 28 | 12 | 12 | 12 |
| **Media (g/100g)** | 14.46 | 18.30 | 55.91 | 14.77 | 19.05 | 55.13 |
| **Máximo (g/100g)** | 19.61 | 41.54 | 73.62 | 18.68 | 37.30 | 67.79 |
| **Mínimo (g/100g)** | 7.57 | 5.81 | 30.68 | 10.29 | 9.75 | 35.22 |
| **Rango (g/100g)** | 12.04 | 35.73 | 42.93 | 8.39 | 27.55 | 32.57 |
| **Desviación estándar** | 2.99 | 10.33 | 8.96 | 2.62 | 10.95 | 9.99 |

P: Proteínas, C: Carbohidratos, L: Lípidos
Fuente: autor

### *4.1.3 Modelo cualitativo*

Se empleó el método de análisis de discriminación con la regresión de mínimos cuadrados parciales (PLS-DA) para desarrollar un modelo que discrimine entre objetos de una formulación para producir panecillos de aquellos que formados exclusivamente de harina. Los parámetros empleados se describen en el capítulo anterior. En la tabla 10 se exhiben los datos estadísticos del modelo cualitativo.

*Tabla 10 Estadísticas del modelo PLS-DA para identificar muestras basadas en la formulación de panecillos libres de gluten*

| | | Validación cruzada | | | | | Set de prueba | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Pretratamiento** | Factor | VP | VN | FP | FN | Precisión (%) | VP | VN | FP | FN | Precisión (%) |
| Ninguno | 7 | 5 | 28 | 0 | 0 | 100 | 4 | 12 | 0 | 0 | 100 |
| **2da. Derivada** | **2** | **5** | **28** | **0** | **0** | **100** | **4** | **12** | **0** | **0** | **100** |

VP: Verdaderos Positivos, VN: Verdaderos Negativos, FP: Falsos Positivos, FN: Falsos Negativos
Fuente: autor

Ambos modelos clasificaron correctamente a los objetos, así que para seleccionar al mejor se aplicó el principio de la Navaja de Ockham. El parámetro empleado para determinar la sencillez fueron los factores requeridos. Por tanto, se eligió el modelo desarrollado con espectros preprocesados con la Segunda derivada Savitzky-Golay del segundo orden polinomial con 23 puntos totales de suavizado (fig. 27a) porque sólo requirió de dos factores para clasificar correctamente a las muestras.

Para confirmar la correcta clasificación, se revisó el gráfico de regresión (fig. 27b) y en él se aprecian las muestras de la formulación por encima de la línea de decisión y al resto de los objetos por debajo de ella, tanto para la validación cruzada como para el set de prueba. Para complementar, el gráfico Puntuaciones (fig. 27c) demuestra que las muestras pertenecientes a dicha categoría se encuentran lo suficientemente alejadas de las otras.

Finalmente, otra manera de corroborar el desempeño del modelo es con el gráfico que relaciona el valor predicho de las muestras del set de prueba junto con su desviación estándar (fig. 27d). En él se aprecia que los objetos de la formulación, identificados con el prefijo MIX, se ubican por encima de la línea de decisión próximos al 1. Asimismo, el grupo de mezclas de harinas están por debajo de cero, cercanos al -1.

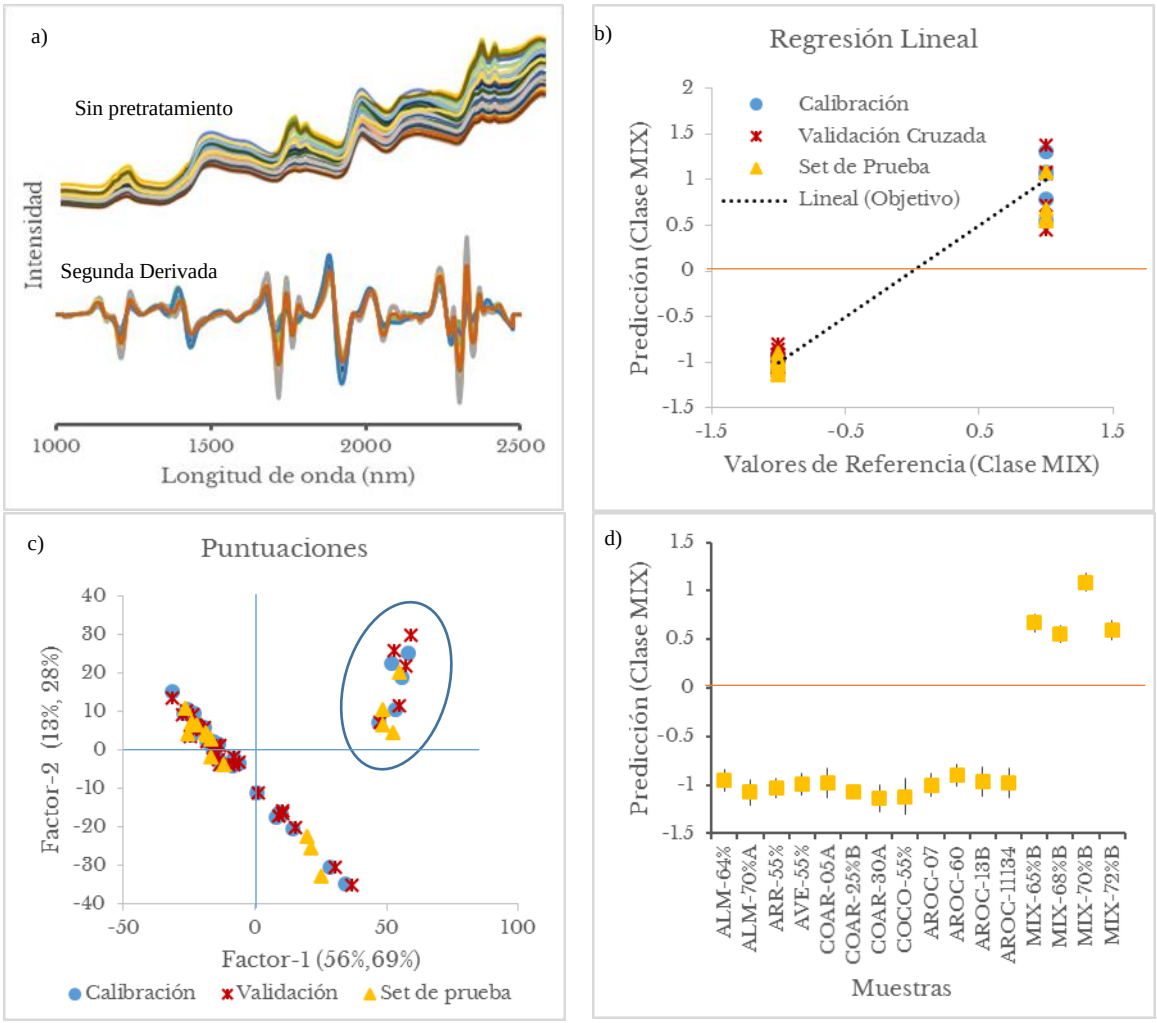



*Figura 27 Gráficos de la revisión general del modelo PLS-DA: Harinas libres de gluten*

a) Espectros NIR de harinas libres de gluten, b) Gráfico de regresión lineal, c) Gráfico Puntuaciones, d) Predicción con desviación estándar del set de prueba
Fuente: autor

### *4.1.4 Modelo cuantitativo – Proteínas*

Utilizando el algoritmo PLS se elaboraron distintos modelos de regresión para determinar la concentración de proteínas presente en una mezcla formada por cuatro tipos de harinas: almendra, arroz, avena y coco. Se emplearon 28 muestras para entrenar al modelo y 12 para evaluarlo. Los parámetros para el cálculo se especifican en el capítulo anterior.

El modelo de calibración desarrollado a partir de los datos espectrales sin pretratamientos, generó en el set de prueba una $R^2$ de 0.80 y un error de estimación de ±1.12 g/100g. Al presentar una baja correlación lineal en la predicción y al no encontrar objetos atípicos, los espectros de los 28 objetos pertenecientes al set de entrenamiento fueron sometidos a distintos preprocesamientos y se realizó de nuevo el cómputo PLS.

La aplicación de la Variable Normal Estándar (SNV, por sus siglas en inglés) junto con la segunda derivada de Savitzky-Golay del segundo orden polinomial con 23 puntos de suavizado, permitió generar un modelo en donde el valor de la incertidumbre en la predicción del contenido de proteínas en muestras desconocidas fue de ±0.756 g/100 g (tabla 11). Además, al haber obtenido una $R^2$ de 0.9107 sugiere que presenta un comportamiento aceptable en la predicción de la concentración del analito.

*Tabla 11 Datos estadísticos del modelo PLS para cuantificar Proteínas: Harinas*

| Pretratamiento | Factores | $R^2$Cal | RMSEC | $R^2$ Val | RMSECV | $R^2$Pred | RMSEP |
|---|---|---|---|---|---|---|---|
| Ninguno | 5 | 0.9432 | 0.6998 | 0.9064 | 0.9322 | 0.8024 | 1.1164 |
| SNV+DT | 5 | 0.9155 | 0.6862 | 0.8995 | 0.966 | 0.8041 | 1.1117 |
| 1RA DER | 4 | 0.9411 | 0.7133 | 0.903 | 0.9493 | 0.8772 | 0.8802 |
| 2DA DER | 4 | 0.9666 | 0.537 | 0.9124 | 0.9021 | 0.8658 | 0.9209 |
| **SNV+2DA** | **4** | **0.9869** | **0.3368** | **0.9413** | **0.7384** | **0.9107** | **0.7562** |
| OSC | 1 | 0.9658 | 0.5432 | 0.9616 | 0.5973 | 0.8405 | 1.0032 |

$R^2$: coeficiente de correlación, RMSEC: error cuadrático medio de la calibración, RMSECV: error cuadrático medio de la validación cruzada, RMSEP: error cuadrático medio de la predicción. Cal: Calibración, Val: Validación y Pred: Predicción.
Fuente: autor

Del modelo elegido, se verificó que las señales de las vibraciones encontradas en los espectros originales no hubiesen sido removidas con los pretratamientos y, como se observa en la figura 28a, las bandas de absorción fueron intensificadas al corregirse el solapamiento de las bandas. Además, dentro de las variables importantes identificadas, 627 de las 1501, se encuentran las regiones características de las proteínas: 1502, 1960-1980, 2050-2060 y 2180 nm.

Lo siguiente fue comprobar que el número de factores fuera el adecuado al revisar los valores de RMSE de la calibración y validación cruzada en cada una de las variables latentes empleadas para el cálculo. En la figura 28b, se aprecia a partir del factor 4 el valor del error en la validación no disminuye significativamente y por tanto, se confirma que éste es el número óptimo de factores.

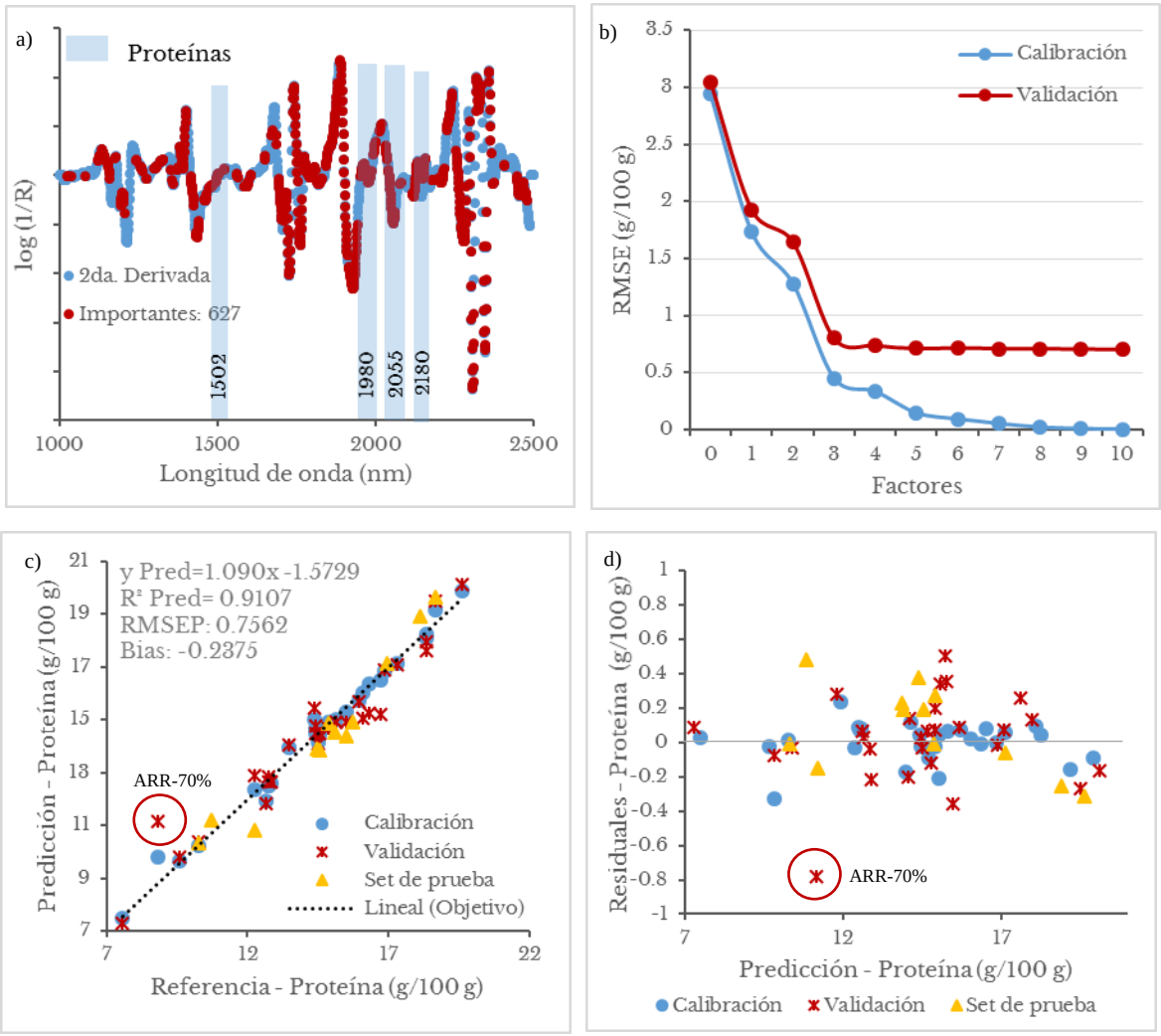



*Figura 28 Gráficos de la revisión general del modelo PLS: Proteínas en Harinas*

a) Espectro NIR promedio resaltando las variables importantes, b)Gráfico RMSE, c) Regresión Lineal, d) Residuales.
Fuente: autor

En el gráfico de regresión lineal (fig. 28c), que está basado en los resultados del factor-4, se puede ver que la mayoría de las muestras de la validación cruzada y del set de prueba se encuentran alrededor de la línea de regresión. Sin embargo, destaca una que se encuentra alejada considerablemente del resto, lo cual la cataloga como posible anomalía u "outliers". Por esta razón, se revisó el gráfico de Residuales (fig. 28d) y en él se observa que la muestra ARR-70% presenta un gran valor residual comparado con el resto, sin embargo parece no perturbar al modelo. Por lo anterior, no se le consideró como un objeto atípico.

### *4.1.5 Modelo cuantitativo – Carbohidratos*

Los datos estadísticos de los modelos entrenados para cuantificar carbohidratos en una mezcla de harinas de almendra, arroz, avena y coco, se especifican en la tabla 12.

*Tabla 12 Datos estadísticos del modelo PLS para cuantificar Carbohidratos: Harinas*

| Pretratamiento | Factores | $R^2$Cal | RMSEC | $R^2$ Val | RMSECV | $R^2$Pred | RMSEP |
|---|---|---|---|---|---|---|---|
| **Ninguno** | 3 | 0.9875 | 1.135 | 0.9836 | 1.3491 | 0.9532 | 2.2681 |
| **SNV+DT** | **1** | **0.9858** | **1.2095** | **0.9837** | **1.3441** | **0.9648** | **1.9665** |
| **1RA DER** | 2 | 0.9774 | 1.5249 | 0.9724 | 1.747 | 0.9572 | 2.1699 |
| **2DA DER** | 2 | 0.9823 | 1.3477 | 0.9767 | 1.6043 | 0.9565 | 2.1857 |
| **SNV+2DA** | 1 | 0.9757 | 1.5828 | 0.9707 | 1.7993 | 0.9726 | 1.7349 |
| **OSC** | 1 | 0.9905 | 0.9874 | 0.9895 | 1.0765 | 0.9627 | 2.0249 |

$R^2$: coeficiente de correlación, RMSEC: error cuadrático medio de la calibración, RMSECV: error cuadrático medio de la validación cruzada, RMSEP: error cuadrático medio de la predicción. Cal: Calibración, Val: Validación y Pred: Predicción. Fuente: autor

La aplicación de la Variable Normal Estándar (SNV, por sus siglas en inglés) junto con la segunda derivada de Savitzky-Golay, permitió generar el modelo con el menor RMSE: 1.7349 g/100 g. Sin embargo, presentó una $R^2$ mayor que en la validación cruzada pudiendo ser esto un indicio de sobreajuste. Por tanto, el modelo elegido para cuantificar carbohidratos en futuras muestras fue al que se le procesó con SNV y "de-trend" en donde el valor de la incertidumbre en el set de prueba fue de ±1.9665 g/100 g. Además, al haber obtenido una $R^2$ de 0.96 sugiere que presenta un comportamiento aceptable en la determinación de la concentración del analito.

Se verificó que las bandas de las vibraciones encontradas en los espectros originales no hubiesen sido removidas con el preprocesamiento y, como se observa en la figura 29a, éstos corrigieron los efectos de la dispersión de la luz y la línea base cuadrática. Además, las variables importantes fueron 1311 de las 1501 empleadas. Aunque prácticamente se usaron todas, el modelo sólo requirió de un factor para cuantificar a los carbohidratos. Para sustentar que no se necesitaron más, se revisaron los valores de RMSE de la validación cruzada en cada uno de los factores empleados para el cálculo. En la figura 29b, se observa que el factor-1 fue quien obtuvo el menor error de estimación y que de ahí en adelante, éste fue aumentando. Por tanto, se confirma que el número óptimo de componentes para el desarrollo del modelo es uno.

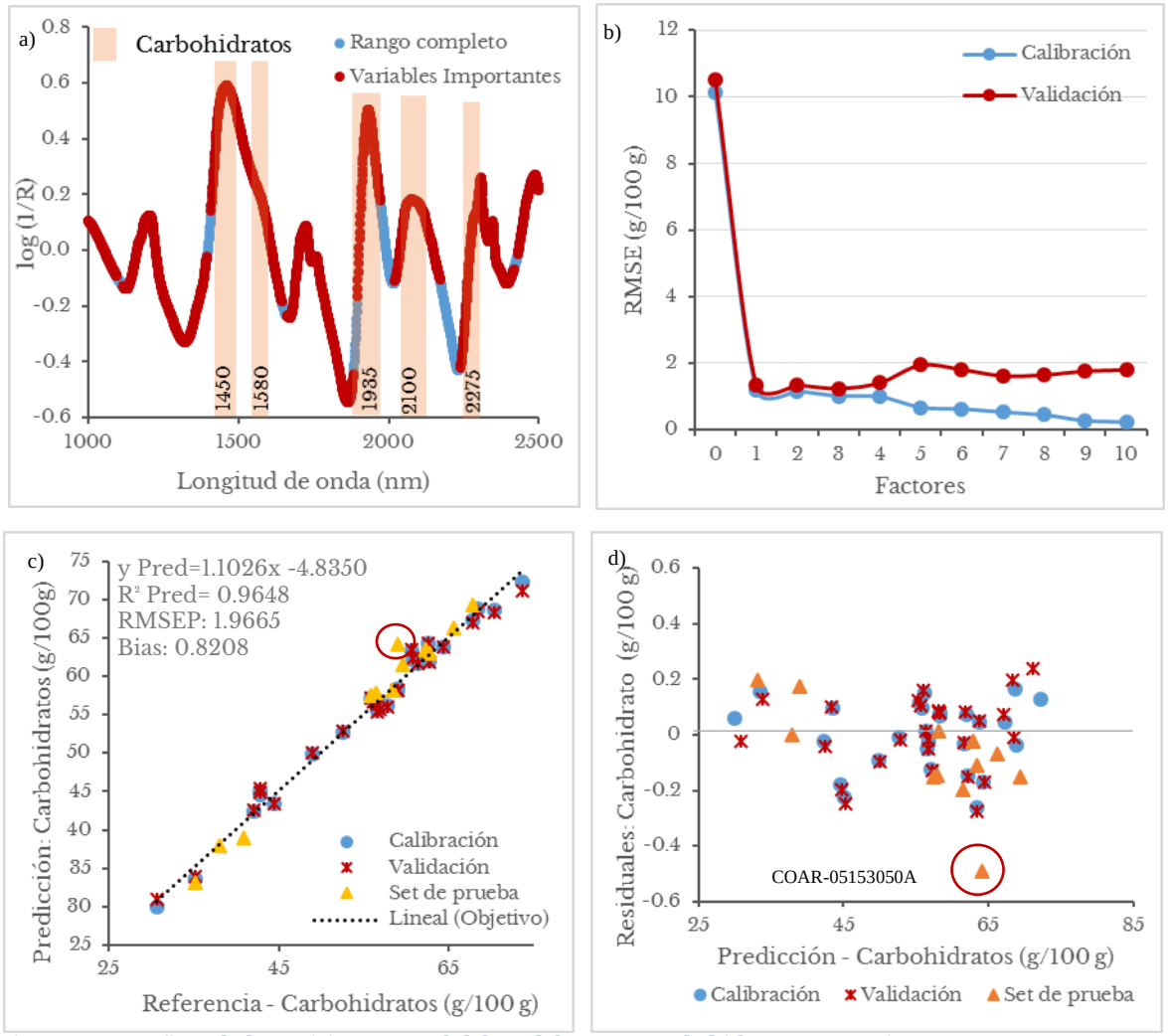



*Figura 29 Gráficos de la revisión general del modelo PLS: Carbohidratos en Harinas*

a) Espectro NIR promedio resaltando las variables importantes, b)Gráfico RMSE, c) Regresión Lineal, d) Residuales.
Fuente: autor

Para visualizar el comportamiento de las muestras en la regresión, se graficaron los valores de referencia de los carbohidratos contra los predichos (fig. 29c). En él se puede ver que la mayoría de las muestras de la validación cruzada y del set de prueba se encuentran alrededor de la línea de regresión. Sin embargo, existe una que se encuentra alejada de dicha línea, lo cual podría indicar que es un objeto atípico. Por tanto, se revisó el gráfico de Residuales (fig. 29d) en donde se aprecia que la muestra COAR-05153050A tiene un gran valor residual comparado contra el resto. Aun así, no parece influir en los demás y por esa razón, no fue considerado como atípico u "outlier".

### *4.1.5 Modelo cuantitativo – Lípidos*

Se elaboraron distintos modelos con el algoritmo PLS para determinar la concentración de lípidos presentes en una mezcla formada por cuatro tipos de harinas: almendra, arroz, avena y coco. Los parámetros para el cálculo se especifican en el capítulo anterior y los resultados, en la tabla 13.

*Tabla 13 Datos estadísticos del modelo PLS para cuantificar Lípidos: Harinas*

| Pretratamiento | Factores | $R^2$Cal | RMSEC | $R^2$ Val | RMSECV | $R^2$Pred | RMSEP |
|---|---|---|---|---|---|---|---|
| Ninguno | 3 | 0.9918 | 0.7976 | 0.9893 | 0.9424 | 0.9705 | 1.6438 |
| SNV+DT | 1 | 0.9808 | 1.219 | 0.9793 | 1.3116 | 0.9543 | 2.044 |
| 1RA DER | 2 | 0.9849 | 1.0805 | 0.9818 | 1.2292 | 0.9679 | 1.7131 |
| 2DA DER | 2 | 0.9889 | 0.9273 | 0.9855 | 1.0994 | 0.9667 | 1.7456 |
| **SNV+2DA** | **1** | **0.9919** | **0.7911** | **0.9904** | **0.8925** | **0.976** | **1.4817** |
| OSC | 1 | 0.9929 | 0.7398 | 0.9923 | 0.8022 | 0.9712 | 1.6226 |

$R^2$: coeficiente de correlación, RMSEC: error cuadrático medio de la calibración, RMSECV: error cuadrático medio de la validación cruzada, RMSEP: error cuadrático medio de la predicción. Cal: Calibración, Val: Validación y Pred: Predicción.
Fuente: autor

El modelo elegido fue quien presentó el menor valor de RMSE y éste fue el generado a partir de datos a los que se les aplicó SNV junto con la segunda derivada de Savitzky-Golay (fig. 30a). Al tener un error de estimación de ±1.48 g/100 g y una $R^2$ de 0.976 en el set de prueba, demuestra que tiene un desempeño aceptable para predecir el contenido de lípidos en futuras muestras.

Se revisó en el espectro preprocesado que las bandas vibracionales características de los lípidos no hubiesen sido removidas. Se confirmó su presencia en él y en la lista de las variables importantes (fig. 30a). Éstas corresponden a las ubicadas entre los 1725 a 1765 y 2310 nm.

Lo siguiente fue comprobar que sólo se requiere de un factor para el cálculo del modelo al estudiar los valores de RMSE generados en cada uno de los componentes principales empleados. En la figura 30b se aprecia que en el factor-1 el error de estimación fue el menor en la validación cruzada y por tanto, no se requieren más componentes en el modelo para predecir el contenido de lípidos en futuras muestras.

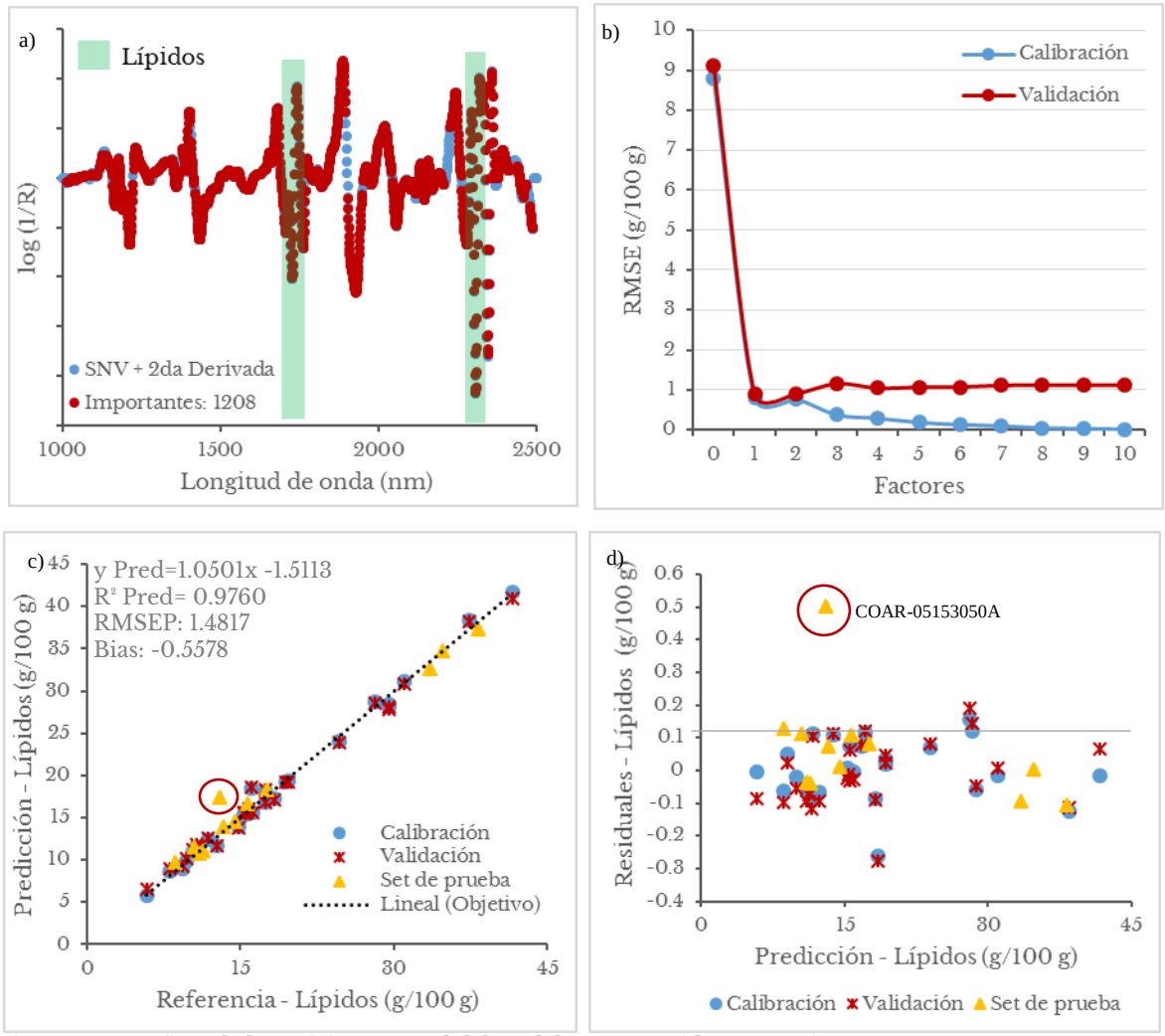



*Figura 30 Gráficos de la revisión general del modelo PLS: Lípidos en Harinas*

a) Espectro NIR promedio resaltando las variables importantes, b)Gráfico RMSE, c) Regresión Lineal, d) Residuales.
Fuente: autor

Además, se graficaron los valores de referencia de la concentración de lípidos contra los estimados (fig. 30c) para visualizar el desempeño del modelo. En el gráfico se puede observar que la mayor parte de las muestras de la validación cruzada y del set de prueba se encuentran sobre la línea de regresión. Sin embargo, existe una ubicada lejos de dicha línea que podría ser un objeto anómalo. Para comprobar lo anterior, se analizó el gráfico de Residuales (fig. 30d) y en él se determinó que el objeto COAR-05153050A es pobremente descrito por el modelo al presentar un alto valor residual. No fue considerado como anómalo porque no atrae hacia él a otras muestras.

En la tabla 14, se especifican los datos estadísticos de los modelos cualitativos y cuantitativos desarrollados para la las harinas libres de gluten empleadas en el presente estudio.

*Tabla 14 Datos estadísticos de los modelos NIR cualitativos y cuantitativos: Harinas libres de gluten*

| Parámetros | Clasificación | Proteínas | Carbohidratos | Lípidos |
|---|---|---|---|---|
| Muestras del set de calibración | 33 | 28 | 28 | 28 |
| Muestras del set de prueba | 16 | 12 | 12 | 12 |
| Algoritmo | PLS-DA | PLS | PLS | PLS |
| Rango espectral (nm) | 1000-2500 | 1000-2500 | 1000-2500 | 1000-2500 |
| Número de variables | 1501 | 1501 | 1501 | 1501 |
| Rango de Concentración (g/100 g) | - | 7.62-19.65 | 30.89-73.75 | 5.88-41.57 |
| Escala | Centrado a la media dividido entre la desviación estándar espectral. | | | |
| Pretratamientos espectrales | Segunda Derivada | SNV + 2da. Derivada | SNV + DT | SNV + 2da. Derivada |
| Factores | 2 | 4 | 1 | 1 |
| Precisión en la clasificación (%) | 100 | - | - | - |
| $R^2$ Calibración | - | 0.9869 | 0.9858 | 0.9919 |
| RMSEC | - | 0.3368 | 1.2095 | 0.7911 |
| $R^2$ Validación cruzada | - | 0.9413 | 0.9837 | 0.9904 |
| RMSECV | - | 0.7384 | 1.3441 | 08925 |
| Objetos con alto valor residual | - | 1 | 1 | 1 |
| $R^2$ Predicción | - | 0.9107 | 0.9648 | 0.9760 |
| RMSEP | - | 0.7562 | 1.9665 | 1.4817 |
| Objetos con alto valor residual | - | 1 | 1 | 1 |

$R^2$: coeficiente de correlación, RMSEC: error cuadrático medio de la calibración, RMSECV: error cuadrático medio de la validación cruzada, RMSEP: error cuadrático medio de la predicción. SNV: Variable normal estándar, DT: Detrending.
Fuente: autor

Una manera de mejorar a los modelos es agregar más muestras al set de entrenamiento en lugar de eliminar aquellas que no se ajusten. Por esta razón, el objeto COAR-05153050A del set de prueba, que después de revisar el gráfico de Residuales se determinó que no era una anomalía, fue descrito pobremente por el modelo para cuantificar carbohidratos y lípidos.

## 4.2 Suplementos alimenticios

### *4.2.1 Espectros NIR*

La figura 31 muestra los espectros NIR obtenidos del grupo de suplementos alimenticios integrado por 37 muestras, cuyas características se describen en la tabla 5, entre la región de los 1000 a 2500 nm sin preprocesamientos matemáticos.

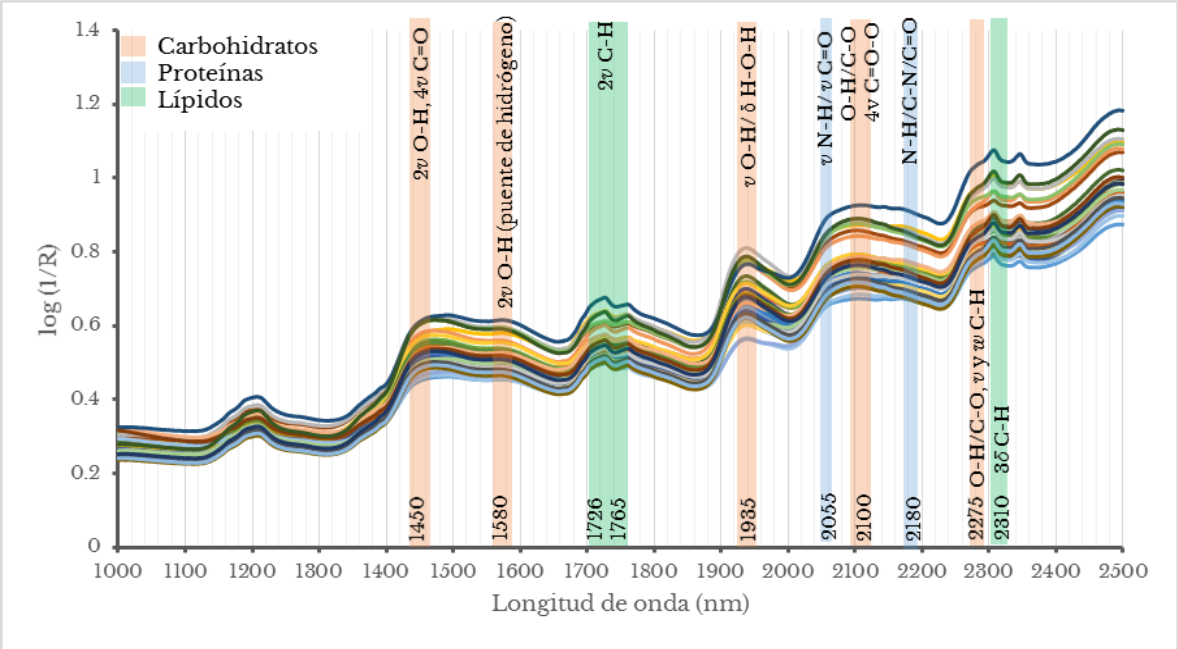



*Figura 31 Espectros NIR sin pretratamientos de los Suplementos Alimenticios*

*ν*= tensión, *δ*= flexión, 2*ν*= primer sobretono de tensión, 4*ν*= tercer sobretono de tensión, 3*δ*=segundo sobretono de flexión, *w*=deformación
Fuente: autor

Se identificaron cinco bandas de absorción asignadas a los carbohidratos, tres regiones correspondientes a los lípidos y dos señales débiles relacionadas a las proteínas en cada objeto analizado.

Al confirmar visualmente la presencia de las macromoléculas de interés en las muestras, se prosiguió con la aplicación de técnicas quimiométricas para el entrenamiento de los modelos cualitativos y cuantitativos.

### *4.2.2 Exploración de Datos*

El método PCA fue aplicado a los datos espectrales sin pretratamientos de las 37 muestras. El número de componentes principales óptimo calculado fue tres, en donde el PC1 explica el 95% de la varianza, el PC2 el 2% y el PC3 el 2% dando un total de 99%. Esto indica que para el entrenamiento del modelo PLS a partir de espectros sin preprocesamientos, se requieren un mínimo de tres variables latentes.

Se revisó el gráfico de Puntuaciones (fig. 32a) para buscar indicios de patrones en las muestras de suplementos alimenticios. Se encontró que la mayoría de los objetos estaban distribuidos sobre los cuadrantes I y IV. Sin embargo, después de etiquetarlos se identificaron tres pequeñas aglomeraciones integradas por suplementos de la misma marca comercial. A pesar de que no estaban alejadas unas de las otras, indica que existe el potencial de crear un modelo de clasificación.

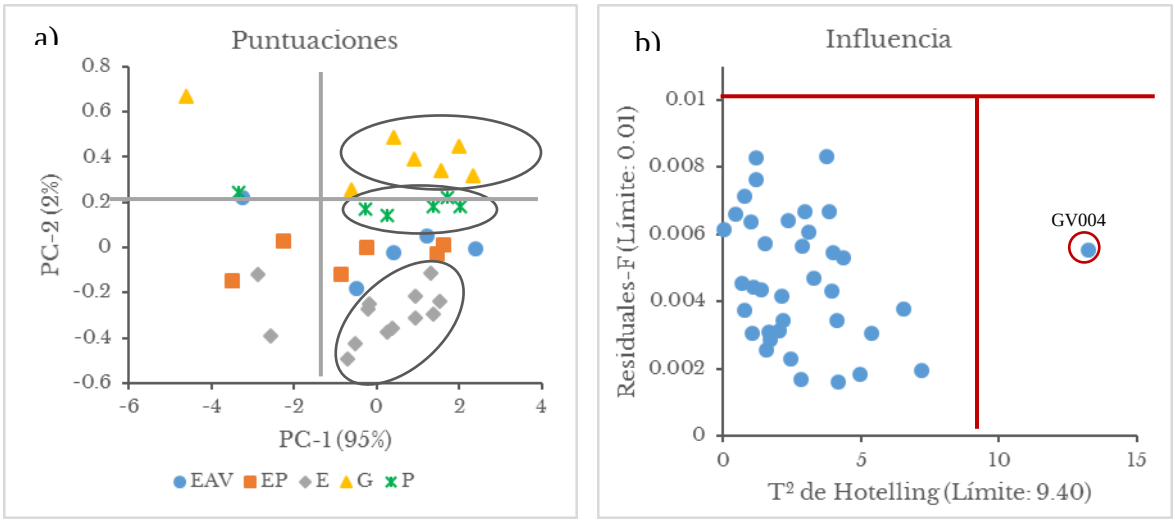



*Figura 32 Gráficos Puntuaciones e Influencia de los Suplementos alimenticios*

EAV: Ensure Advance ®, EP: Ensure Plus ®, E: Ensure®, G: Glucerna®, P: Pediasure®
Fuente: autor

Antes de realizar la selección de muestras que integrarían el set de entrenamiento, se analizó el gráfico de Influencia (fig. 32b) para identificar a objetos atípicos. En él se observa que el objeto GV004 se ubica fuera de los límites de la distribución $T^2$ de Hotelling, lo que significa que puede perturbar al modelo. Se realizó el procedimiento para estos casos, descrito en el capítulo anterior, y al volver a obtener

el mismo resultado se decidió excluirla del estudio. En la tabla 15 se describen las características de la muestra atípica para futuras referencias.

*Tabla 15 Reporte de objeto anómalo del grupo de Suplementos Alimenticios*

| Objeto anómalo "Outlier" | |
|---|---|
| **ID Muestra:** | GV004 |
| **Límites excedidos:** | $T^2$ de Hotelling |
| **Observaciones:** | Composición química y apariencia física similar a las otras muestras pertenecientes al grupo de suplementos alimenticios. La recolección del espectro se realizó bajo las mismas condiciones que los demás objetos. Sin embargo, su almacenamiento fue comprometido al ser retirado frecuentemente del refrigerador y esto podría ser la causa de la anomalía. |

Fuente: autor

La selección del set de calibración y el de prueba se llevó a cabo con la ayuda del algoritmo Kennard-Stone aplicado sobre las puntuaciones obtenidas en el componente principal 3 (PC3). En la tabla 16a se expone la distribución de las muestras para el desarrollo de los modelos cualitativos y en la 16b, la estadística descriptiva de las macromoléculas analizadas: proteínas, carbohidratos y lípidos.

*Tabla 16 Distribución de muestras y estadística descriptiva: Suplementos alimenticios*

a)

| | No. de objetos en el Set de entrenamiento | | No. de objetos en el Set de prueba | |
|---|---|---|---|---|
| **Modelo PLS-DA** | Clase +1 | Clase -1 | Clase +1 | Clase -1 |
| **Ensure Advance®** | 3 | 21 | 2 | 10 |
| **Ensure Plus®** | 4 | 20 | 2 | 10 |
| **Ensure®** | 9 | 15 | 4 | 8 |
| **Glucerna®** | 4 | 20 | 2 | 10 |
| **Pediasure®** | 4 | 20 | 2 | 10 |
| **Vainilla** | 17 | 7 | 8 | 4 |
| **Fresa** | 3 | 21 | 2 | 10 |
| **Chocolate** | 4 | 20 | 2 | 10 |

b)

| | Set de entrenamiento | | | Set de prueba | | |
|---|---|---|---|---|---|---|
| **Parámetro** | P | C | L | P | C | L |
| **Número de muestras** | 25 | 25 | 25 | 11 | 11 | 11 |
| **Media (g/100g)** | 4.46 | 15.96 | 3.66 | 4.38 | 16.09 | 3.72 |
| **Máximo (g/100g)** | 5.54 | 21.34 | 5.12 | 5.54 | 21.34 | 5.12 |
| **Mínimo (g/100g)** | 3.07 | 11.97 | 2.71 | 3.07 | 11.97 | 2.71 |
| **Rango (g/100g)** | 2.47 | 9.37 | 2.41 | 2.47 | 9.37 | 2.41 |
| **Desviación estándar** | 0.85 | 3.58 | 0.99 | 0.90 | 3.68 | 0.99 |

P: Proteínas, C: Carbohidratos, L: Lípidos
Fuente: autor

### *4.2.5 Modelo de clasificación: Según marca comercial*

Utilizando la regresión PLS-DA se entrenaron cinco modelos para clasificar a los objetos según la marca comercial del suplemento alimenticio: Ensure Advance®, Ensure Plus®, Ensure®, Glucerna®, Pediasure®. Los parámetros para su desarrollo se especifican en el capítulo anterior y los resultados en la tabla 17.

*Tabla 17 Medidas de desempeño de los modelos PLS-DA: Según su marca comercial*

| | | | Validación cruzada | | | | | Set de prueba | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Marca | Pretratamiento | Factor | VP | VN | FP | FN | Precisión (%) | VP | VN | FP | FN | Precisión (%) |
| Ensure® | Ninguno | 5 | 9 | 15 | 0 | 0 | 100 | 4 | 8 | 0 | 0 | 100 |
| Ensure Advance® | Ninguno | 5 | 3 | 21 | 0 | 0 | 100 | 2 | 10 | 0 | 0 | 100 |
| Ensure Plus® | Ninguno | 7 | 3 | 20 | 0 | 1 | 95.83 | 1 | 8 | 2 | 1 | 75 |
| Glucerna® | Ninguno | 8 | 4 | 20 | 0 | 0 | 100 | 2 | 10 | 0 | 0 | 100 |
| Pediasure® | Ninguno | 8 | 4 | 20 | 0 | 0 | 100 | 2 | 10 | 0 | 0 | 100 |
| **Ensure®** | **2da. Derivada** | **3** | **9** | **15** | **0** | **0** | **100** | **4** | **8** | **0** | **0** | **100** |
| **Ensure Advance®** | **2da. Derivada** | **4** | **3** | **21** | **0** | **0** | **100** | **2** | **10** | **0** | **0** | **100** |
| **Ensure Plus®** | **2da. Derivada** | **7** | **4** | **20** | **0** | **0** | **100** | **2** | **10** | **0** | **0** | **100** |
| **Glucerna®** | **2da. Derivada** | **4** | **4** | **20** | **0** | **0** | **100** | **2** | **10** | **0** | **0** | **100** |
| **Pediasure®** | **2da. Derivada** | **4** | **4** | **20** | **0** | **0** | **100** | **2** | **10** | **0** | **0** | **100** |

VP: Verdaderos Positivos, VN: Verdaderos Negativos, FP: Falsos Positivos, FN: Falsos Negativos
Fuente: autor

Los modelos entrenados a partir de los datos espectrales sin pretratamientos clasificaron a los objetos con una precisión del 100%, tanto en la validación cruzada como en el set de prueba, a excepción del diseñado a identificar productos de la marca Ensure Plus®. Por tanto, a los espectros se les aplicó la segunda derivada Savitzky-Golay del segundo orden polinomial con 39 puntos de suavizado, para corregir la línea base cuadrática característica de estos espectros NIR. Con esto, el modelo aumentó el porcentaje de precisión de un 75 a un 100% en el set de prueba y en los otros, disminuyó el número de factores requeridos para su cálculo. Por consiguiente, dichos modelos fueron seleccionados para clasificar a muestras futuras.

Para sustentar el desempeño de cada modelo, se realizó un análisis de cuatro gráficos: Puntuaciones, Predicción por muestra, Predicción con desviación estándar y el de variables importantes.

El modelo que identifica a los productos marca Ensure® tuvo una precisión del 100% y esto se puede comprobar con el gráfico de Puntuaciones (fig. 33a), donde se aprecia que los objetos de esta marca comercial se encuentran agrupados, lejos del resto de los suplementos alimenticios. Además, en las figuras 33b y 33c presentan cómo fueron clasificadas correctamente cada una de las muestras en el set de calibración y en el de prueba respectivamente, todos los objetos de la marca comercial de interés se ubicaron arriba de la línea de decisión como se esperaba.

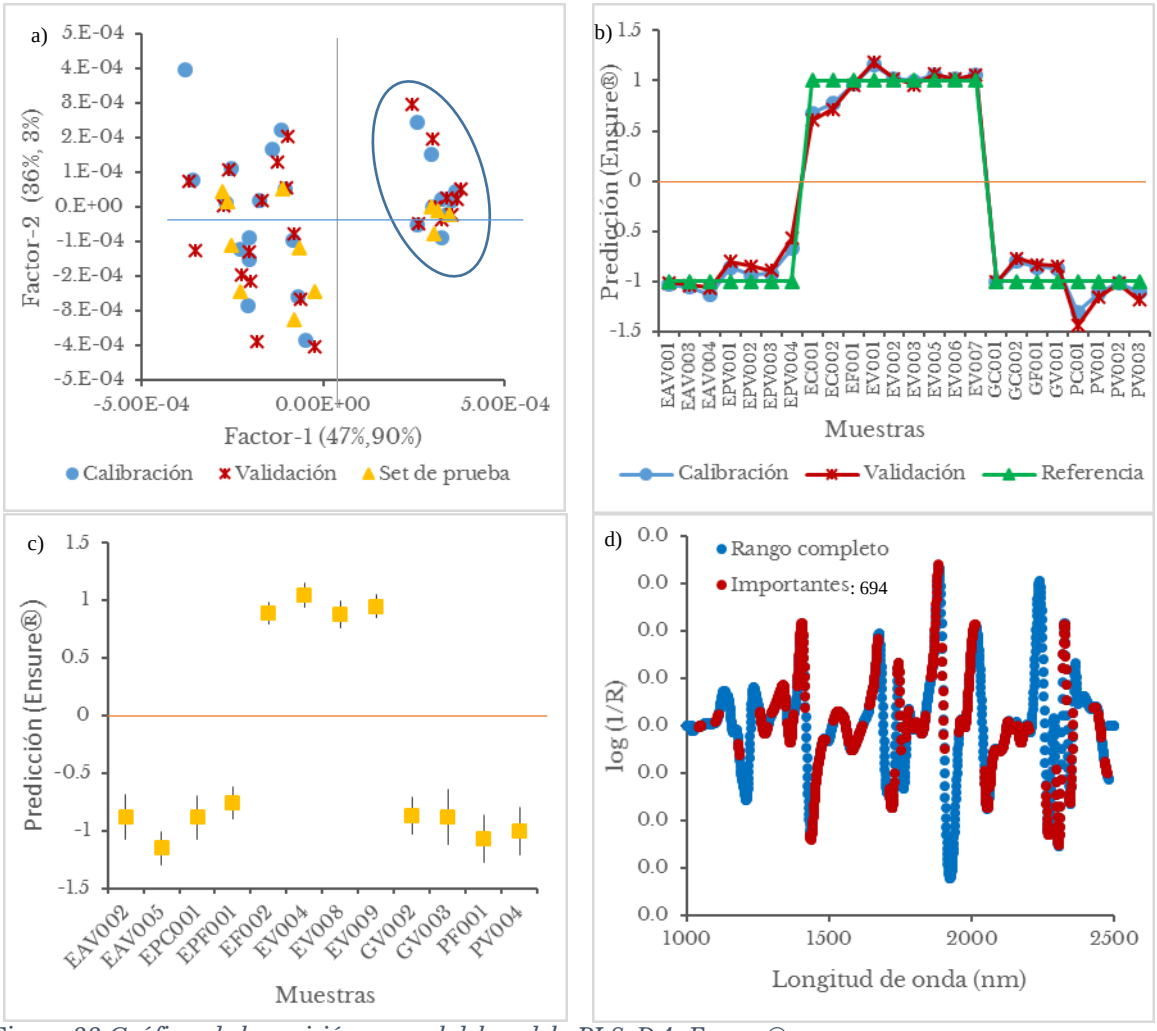



*Figura 33 Gráficos de la revisión general del modelo PLS-DA: Ensure®*

a) Gráfico Puntuaciones, b) Clasificación de cada muestra del set de entrenamiento, c) Predicción con desviación estándar del set de prueba y d) Espectro NIR promedio resaltando las variables importantes.
Fuente: autor

Por último, en el espectro NIR promedio de los suplementos alimenticios (fig. 33d) se resaltaron 694 longitudes de onda que aportaron la mayor información al modelo.

En el modelo que discrimina a los productos Ensure Advance® del resto tuvo el 100% de precisión. Se confirma lo anterior al observar el gráfico de Puntuaciones (fig. xa) donde indica que los productos están agrupados y un poco retirados del resto de los suplementos alimenticios. Asimismo, en las figuras xb y xc, se observan que todos los productos de la marca comercial de interés se ubican por encima de la línea de decisión y el resto por debajo de ella.

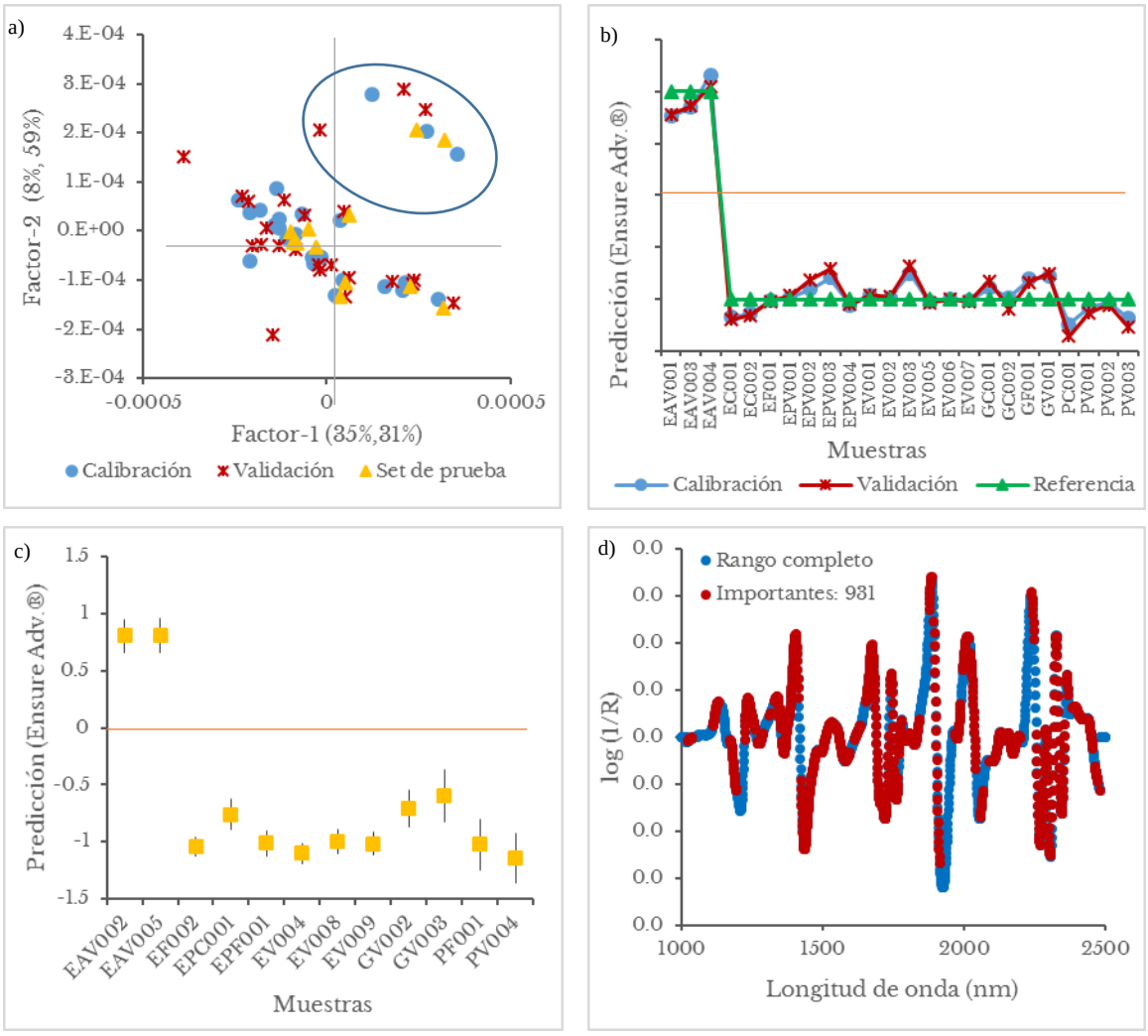



*Figura 34 Gráficos de la revisión general del modelo PLS-DA: Ensure Advance®*

a) Gráfico Puntuaciones, b) Clasificación de cada muestra del set de entrenamiento, c) Predicción con desviación estándar del set de prueba y d) Espectro NIR promedio resaltando las variables importantes.
Fuente: autor

Por otro lado, se identificaron 931 variables importantes de las 1501 que se emplearon en el cálculo. Esto significa que el modelo tiene el potencial de simplificarse en un futuro.

En contraste a los dos modelos anteriores, el gráfico Puntuaciones (fig. xa) no mostró una separación notoria entre los productos Ensure Plus® del resto. Sin embargo, esto no se relaciona con el desempeño del modelo ya que la precisión de la clasificación fue del 100%. Lo anterior se confirma en los gráficos de predicción donde se observan que los objetos de esta marca comercial, etiquetados con la clave EP, se ubican por encima de la línea de decisión en el set de calibración (fig. xb) y en el set de prueba (fig. xc). Asimismo, el número de variables importantes (fig. xd) es otra diferencia con los demás modelos debido a que en éste se identificaron 342 longitudes de onda que aportaron información relevante de las 1501 empleadas en el cálculo.

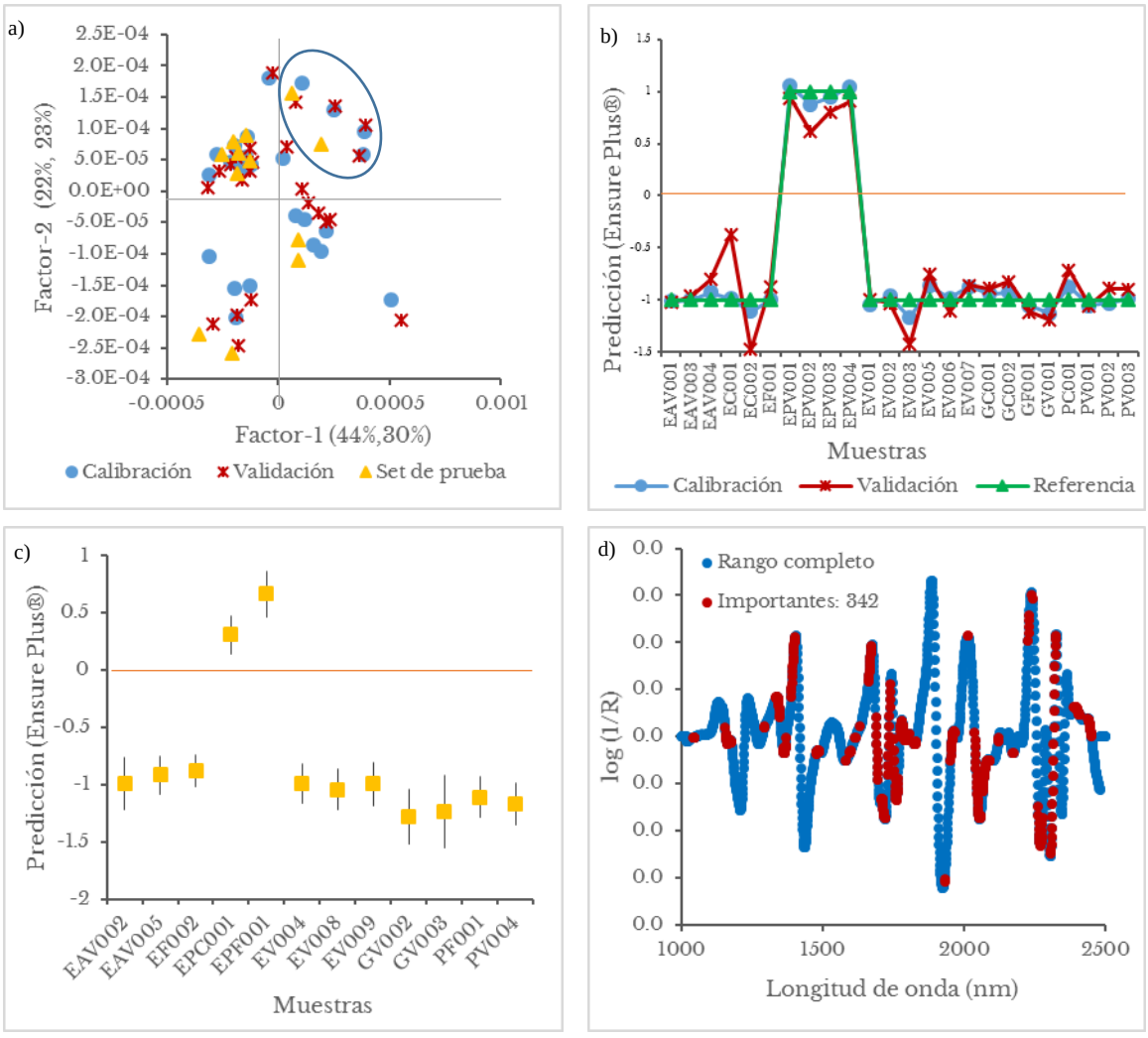



*Figura 35 Gráficos de la revisión general del modelo PLS-DA: Ensure Plus®*

a) Gráfico Puntuaciones, b) Clasificación de cada muestra del set de entrenamiento, c) Predicción con desviación estándar del set de prueba y d) Espectro NIR promedio resaltando las variables importantes.
Fuente: autor

El modelo que identificó a los productos Glucerna®, etiquetados con la letra G, tuvo una precisión del 100%. Esto se puede ver reflejado en el gráfico de Puntuaciones (fig. 36a) donde se aprecia una pequeña aglomeración correspondiente a las muestras de interés. Además, en las figuras 36b y 36c los objetos de esta categoría se ubican próximos a 1, demostrando que tiene un buen desempeño para discriminar entre una clase y otra. Sumado a lo anterior, se identificaron 624 variables importantes de las 1501 empleadas en el entrenamiento del modelo (fig. 36d), lo que significa que tiene el potencial de ser simplificado al remover las longitudes de onda que aportan ruido.

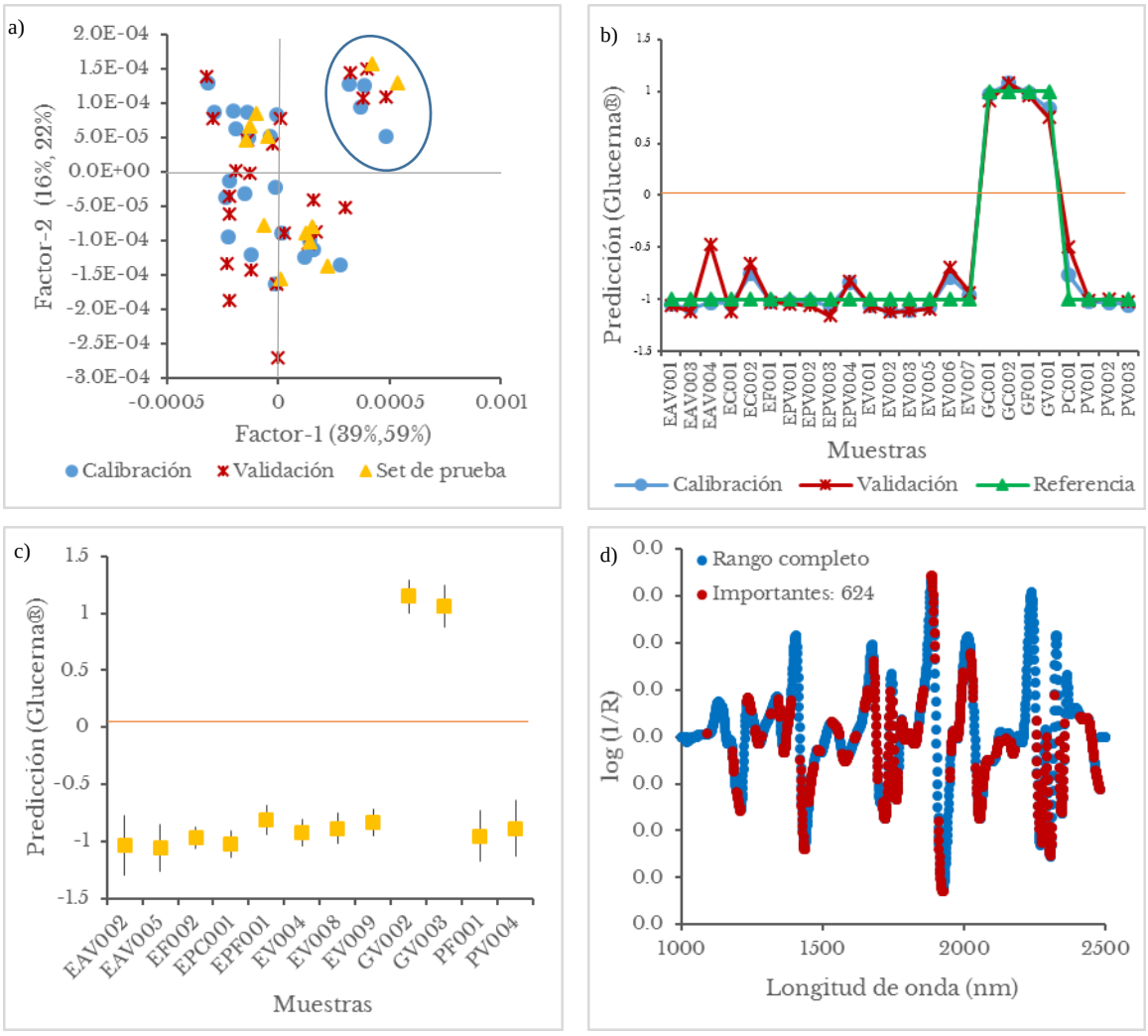



*Figura 36 Gráficos de la revisión general del modelo PLS-DA: Glucerna®*

a) Gráfico Puntuaciones, b) Clasificación de cada muestra del set de entrenamiento, c) Predicción con desviación estándar del set de prueba y d) Espectro NIR promedio resaltando las variables importantes.
Fuente: autor

Al igual que con el modelo PLS-DA: Ensure Plus®, el gráfico Puntuaciones (fig. 37a) no es muy revelador, específicamente para el caso de la validación cruzada. Aun así, la capacidad de discriminar entre la clase Pediasure® del resto de los suplementos alimenticios no se vio comprometida. Lo anterior se ratifica con las figuras 37b y 37c, donde se aprecian que los objetos de la marca comercial de interés sobrepasan la línea de decisión generando de esta manera la precisión reportada del 100%.

Para finalizar, el modelo se basó en la información de 321 variables de las 1501 utilizadas para su cálculo. Sobresale que la mayoría de ellas abarcan los rangos correspondientes a las bandas características de las proteínas: 1980 y 2050-2060 nm.

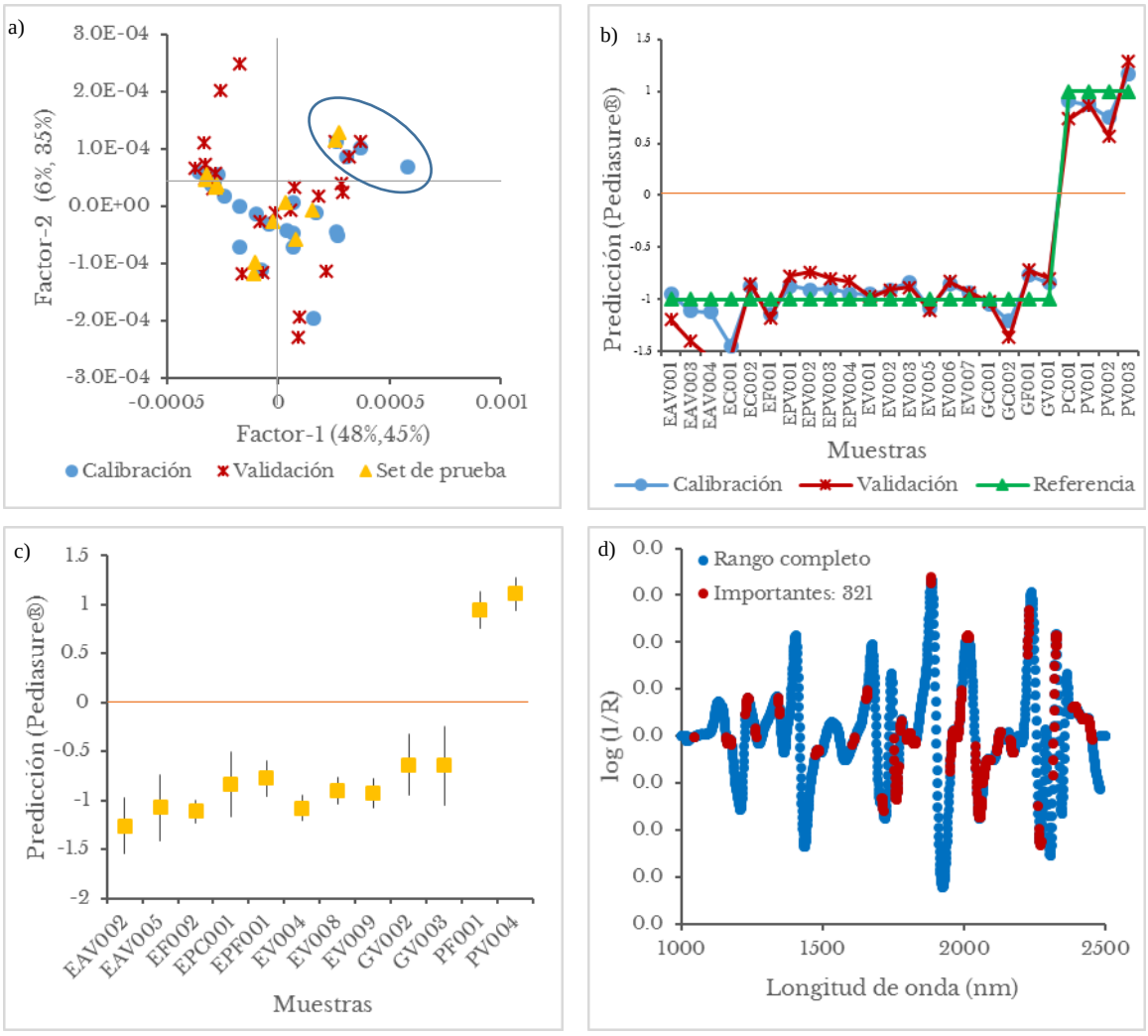



*Figura 37 Gráficos de la revisión general del modelo PLS-DA: Pediasure®*
a) Gráfico Puntuaciones, b) Clasificación de cada muestra del set de entrenamiento, c) Predicción con desviación estándar del set de prueba y d) Espectro NIR promedio resaltando las variables importantes.
Fuente: autor

### *4.2.5 Modelo de clasificación: Según el Sabor*

Empleando la regresión PLS-DA se entrenaron tres modelos para clasificar a los objetos según el sabor del suplemento alimenticio: vainilla, fresa y chocolate. Los parámetros para su desarrollo se especifican en el capítulo anterior y los resultados en la tabla 18.

*Tabla 18 Estadísticas de modelos PLS-DA para clasificar a Suplementos Alimenticios según su sabor*

| | | | Validación cruzada | | | | | | Set de prueba | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sabor | Pretratamiento | Factor | VP | VN | FP | FN | Precisión (%) | Sensibilidad (%) | VP | VN | FP | FN | Precisión (%) | Sensibilidad (%) |
| Vainilla | Ninguno | 4 | 15 | 4 | 3 | 2 | 79.17 | 90 | 1 | 0 | 4 | 7 | 8.33 | 13 |
| Fresa | Ninguno | 1 | 0 | 21 | 0 | 3 | 87.5 | 0 | 0 | 6 | 4 | 2 | 50 | 0 |
| Chocolate | Ninguno | 5 | 4 | 20 | 0 | 0 | 100 | 100 | 1 | 10 | 0 | 1 | 91.67 | 50 |
| **Vainilla** | **SNV + OSC** | **9** | **17** | **7** | **0** | **0** | **100** | **100** | **1** | **8** | **2** | **1** | **75** | **50** |
| **Fresa** | **OSC** | **9** | **3** | **21** | **0** | **0** | **100** | **100** | **6** | **1** | **3** | **2** | **58.33** | **75** |
| **Chocolate** | **SNV+DT** | **7** | **4** | **20** | **0** | **0** | **100** | **100** | **2** | **10** | **0** | **0** | **100** | **100** |

VP: Verdaderos Positivos, VN: Verdaderos Negativos, FP: Falsos Positivos, FN: Falsos Negativos. OSC: Corrección Ortogonal de la Señal, SNV: Variable Normal Estándar, DT: Detrending
Fuente: autor

Los modelos desarrollados con datos sin preprocesamientos tuvieron un pobre desempeño para clasificar a las muestras según el sabor. El más aceptable fue el de chocolate cuya precisión en el set de prueba fue del 91.67% con una sensibilidad del 50%. Por tanto, los datos fueron sometidos a distintos pretratamientos para corregir los efectos de dispersión de la luz y la línea base cuadrática con SNV y "De-trend" y otros para optimizar las señales que responden al predictor con OSC. De esta manera, los modelos de fresa y vainilla mejoraron su desempeño mientras que el de chocolate aumentó su precisión al 100%.

La confirmación visual se llevó a cabo con una revisión general de cuatro gráficos: Puntuaciones, el de Predicción de clases del set de entrenamiento y el del set de prueba y el cuarto es el espectro NIR promedio de los suplementos alimenticios con el pretratamiento utilizado, en donde además se resaltan las variables importantes identificadas.

La precisión en la clasificación del modelo PLS-DA: Vainilla reportada del 100% en la validación cruzada se demuestra con el gráfico de predicción del set de entrenamiento ya que en él se visualiza la correcta discriminación entre clases. Sin embargo, la sensibilidad del modelo en el set de prueba fue del 50% al clasificar erróneamente a 5 objetos (fig. 38c) a pesar de haber utilizado la mayoría de las variables del rango espectral (fig. 38d). Un indicio de lo anterior podría deberse a que para el modelo, los objetos sabor vainilla y fresa son muy similares como se observa en el gráfico de Puntuaciones (fig. 38a) y por tanto, no presenta un buen desempeño para realizar la clasificación.

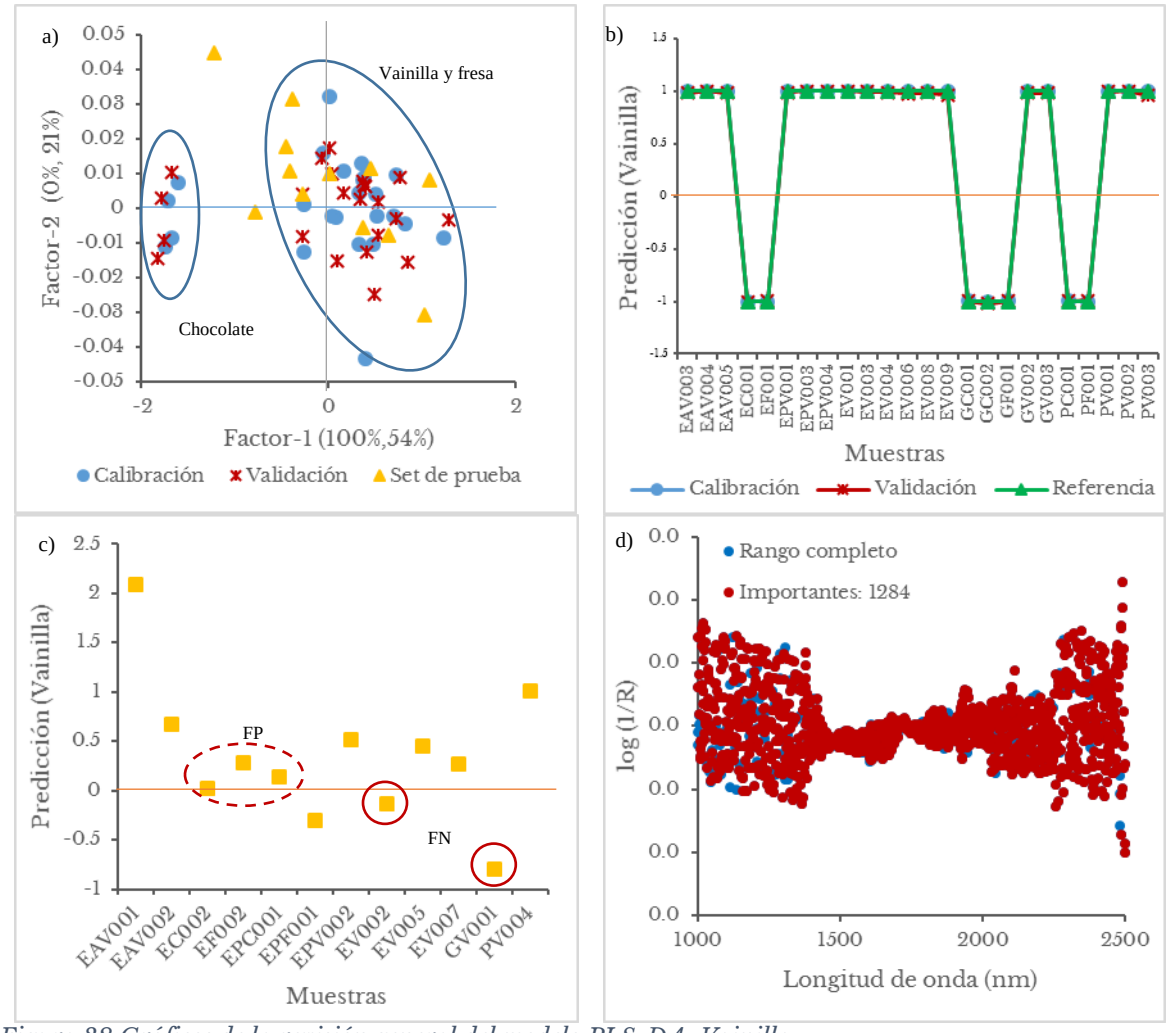



*Figura 38 Gráficos de la revisión general del modelo PLS-DA: Vainilla*

a) Gráfico Puntuaciones, b) Clasificación de cada muestra del set de entrenamiento, c) Predicción con desviación estándar del set de prueba y d) Espectro NIR promedio resaltando las variables importantes. FP: Falsos positivos, FN: Falsos negativos.
Fuente: autor

El modelo para identificar a los productos sabor fresa se comportó de manera similar al de vainilla ya que en el set de calibración obtuvo una precisión del 100% (fig. 39b), contrario al set de prueba (fig. 39c) que fue del 58%.

A pesar de haber usado casi todo el rango espectral para el desarrollo del modelo (fig. 39d), éste no se desempeñó de manera aceptable probablemente porque no existía información suficiente para hacer la distinción entre una clase y otra. Lo anterior se confirma con el gráfico Puntuaciones (fig. 39a) porque existe una ausencia de agrupaciones, indicando que todos los objetos son parecidos.

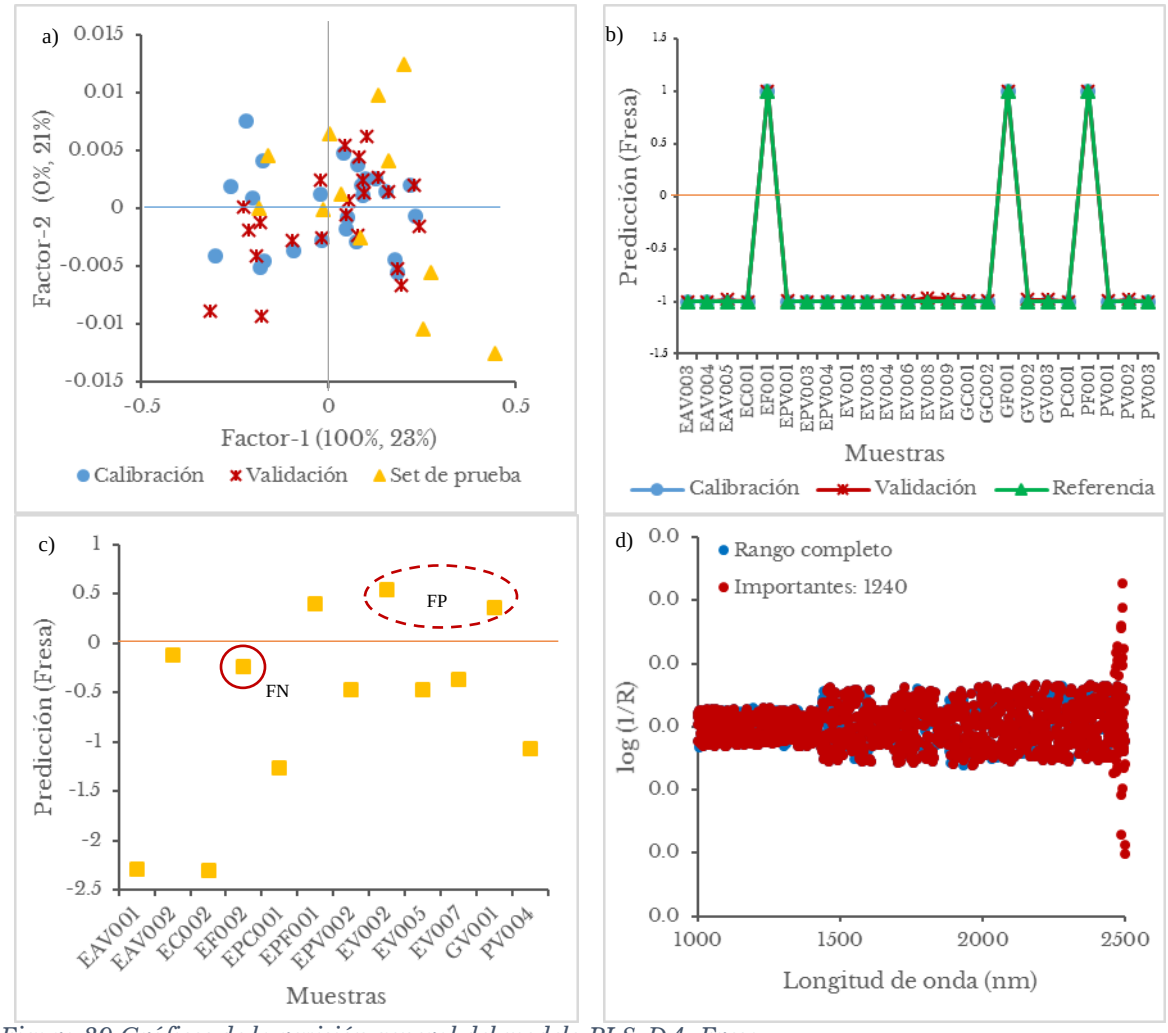



*Figura 39 Gráficos de la revisión general del modelo PLS-DA: Fresa*

a) Gráfico Puntuaciones, b) Clasificación de cada muestra del set de entrenamiento, c) Predicción con desviación estándar del set de prueba y d) Espectro NIR promedio resaltando las variables importantes. FP: Falsos positivos, FN: Falsos negativos. Fuente: autor

El modelo para identificar muestras sabor chocolate demostró tener un buen desempeño para discriminar entre una clase y otra. Clasificó correctamente a los objetos de interés del resto, tanto en el set de entrenamiento (fig. 40b) como en el de prueba (fig. 40c). Además, en el gráfico de Puntuaciones (fig. 40a) se aprecian dos agrupaciones, indicando que el modelo tiene la capacidad de hacer la distinción entre categorías. Finalmente, sólo requirió la información de 466 variables de 1501 empleadas para el cálculo, haciéndolo el más sencillo de interpretar (fig. 40d).

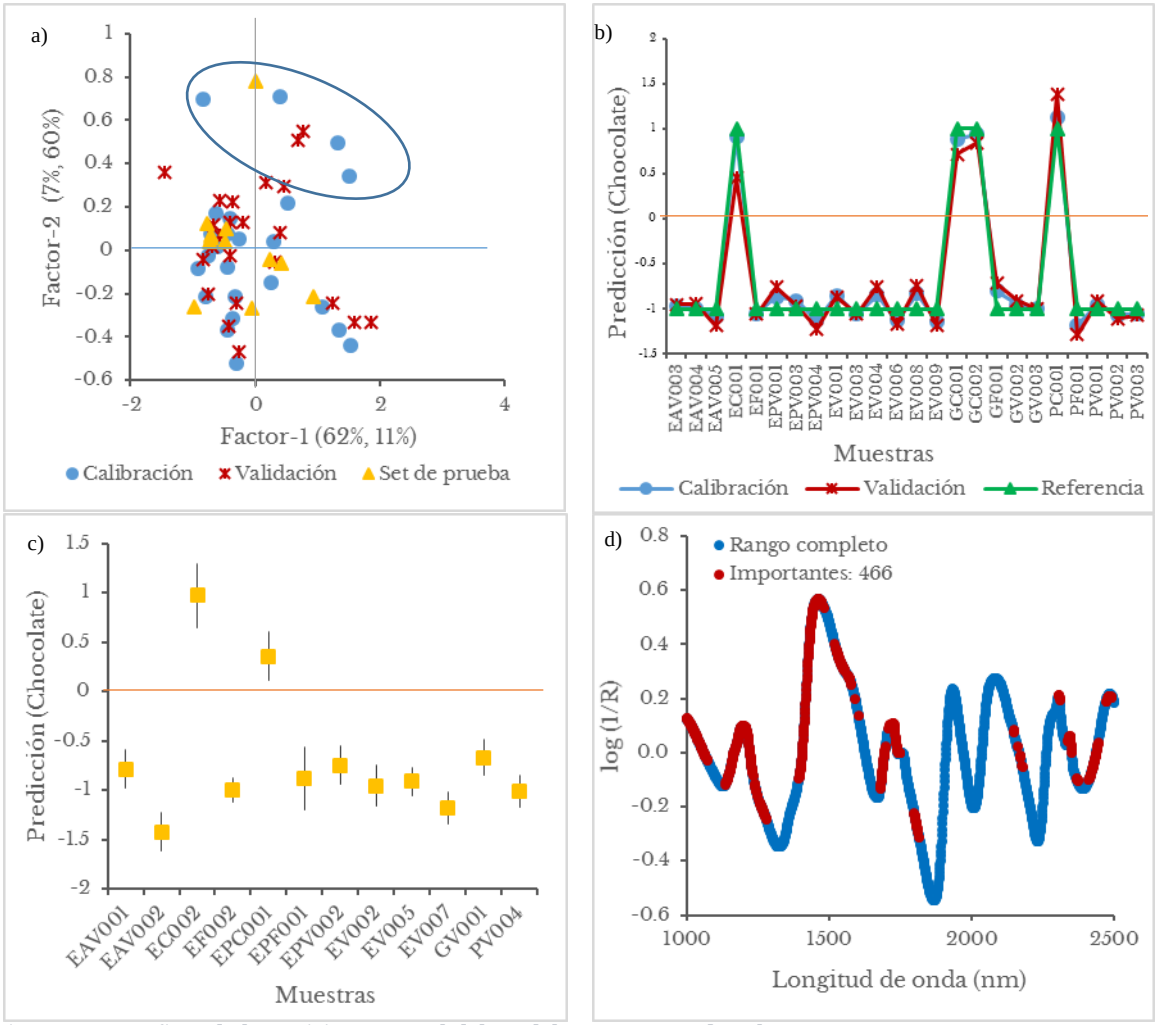



*Figura 40 Gráficos de la revisión general del modelo PLS-DA: Chocolate*

a) Gráfico Puntuaciones, b) Clasificación de cada muestra del set de entrenamiento, c) Predicción con desviación estándar del set de prueba y d) Espectro NIR promedio resaltando las variables importantes. FP: Falsos positivos, FN: Falsos negativos.
Fuente: autor

### *4.2.6 Modelos cuantitativos - Proteínas*

Usando el método de regresión de mínimos cuadrados parciales (PLS) se entrenaron modelos para cuantificar proteínas en suplementos alimenticios de distinta marca comercial. Los parámetros para su desarrollo se especifican en el capítulo anterior y los resultados en la tabla 19.

*Tabla 19 Datos estadísticos de los modelos PLS para cuantificar proteínas en suplementos alimenticios*

| Pretratamiento | Factores | $R^2$Cal | RMSEC | $R^2$ Val | RMSEV | $R^2$Pred | RMSEP |
|---|---|---|---|---|---|---|---|
| Ninguno | 10 | 0.9929 | 0.0717 | 0.9142 | 0.2589 | 0.8230 | 0.3546 |
| SNV+2DA | 7 | 0.9971 | 0.0456 | 0.9760 | 0.137 | 0.9316 | 0.2205 |
| **2da Derivada** | **8** | **0.9943** | **0.0638** | **0.9534** | **0.1908** | **0.9465** | **0.1965** |
| SNV+DT | 10 | 0.9986 | 0.0319 | 0.9581 | 0.1809 | 0.9358 | 0.2135 |

$R^2$: coeficiente de correlación, RMSEC: error cuadrático medio de la calibración, RMSECV: error cuadrático medio de la validación cruzada, RMSEP: error cuadrático medio de la predicción. SNV: Variable normal estándar, DT: Detrending. Fuente: autor

La corrección de la línea base cuadrática del espectro NIR con la segunda derivada de Savitzky-Golay del segundo orden polinomial con un total de 39 puntos de suavizado, permitió generar un modelo en donde el valor de la incertidumbre en la predicción de proteínas en muestras desconocidas fue de ±0.1965 g/100 mL. Además, al tener un valor de correlación lineal en la calibración de 0.98 y en las validaciones de 0.95, sugiere que presenta un comportamiento estable en la predicción de la concentración del analito.

Para sustentar que es apropiado, se realizó una revisión de cuatro gráficos que proporcionan información relevante del modelo de regresión comenzando con el espectro promedio de los suplementos alimenticios corregidos con el pretratamiento.

En la figura 41a se aprecia que las bandas de absorción se intensificaron al componer el solapamiento y se ratificó que las señales características de las proteínas no fueron removidas. Además, se identificaron 381 variables que aportaron información relevante al modelo y entre ellas se encuentran dos regiones asignadas a la macromolécula, las ubicadas entre los 2050-2060 y 2180 nm.

Lo siguiente fue comprobar que el número de factores fuera el adecuado al revisar los valores de RMSE de la calibración y validación cruzada en cada una de las variables latentes empleadas para el cálculo. En la figura 41b se aprecia que el factor 8

obtuvo el menor error en la validación y que a partir de aquí, este ya no disminuye. Así, se corrobora que no se requieren emplear más componentes en el modelo para evitar el sobreajuste.

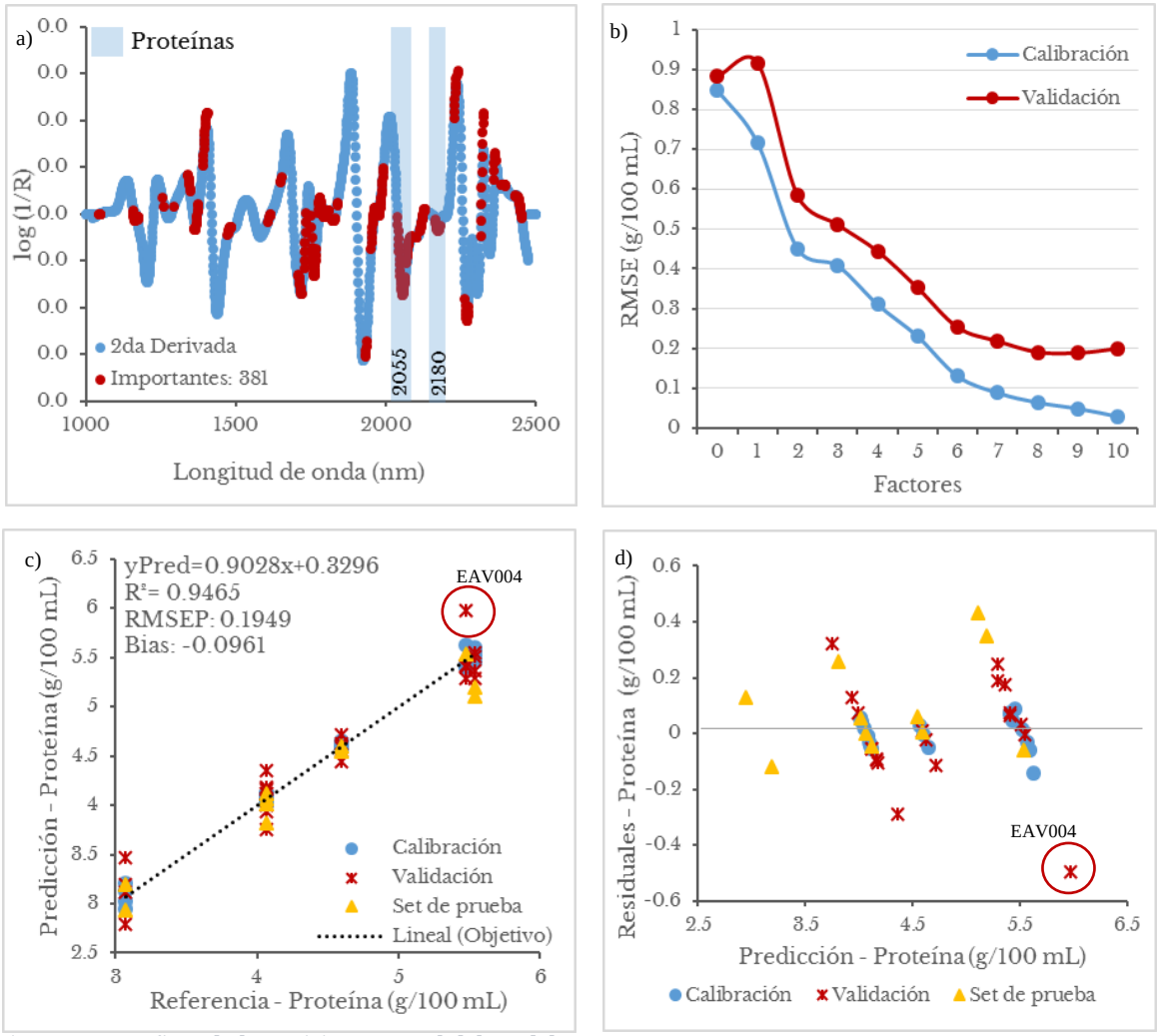



*Figura 41 Gráficos de la revisión general del modelo PLS: Proteínas*

a) Espectro NIR promedio resaltando las variables importantes, b)Gráfico RMSE, c) Regresión Lineal, d) Residuales.
Fuente: autor

La linealidad fue corroborada con el gráfico de regresión (fig. 41c) donde se observa que la mayoría de las muestras del set de entrenamiento y de prueba se ubican sobre o próximos a la línea objetivo. Sin embargo, existieron algunas que el modelo no pudo describir bien lo que generó una $R^2$ en la predicción de 0.9465. Lo anterior fue confirmado con la figura 41d al identificar que dichos objetos manifiestan grandes valores residuales.

### *4.2.4 Modelos de calibración - Carbohidratos*

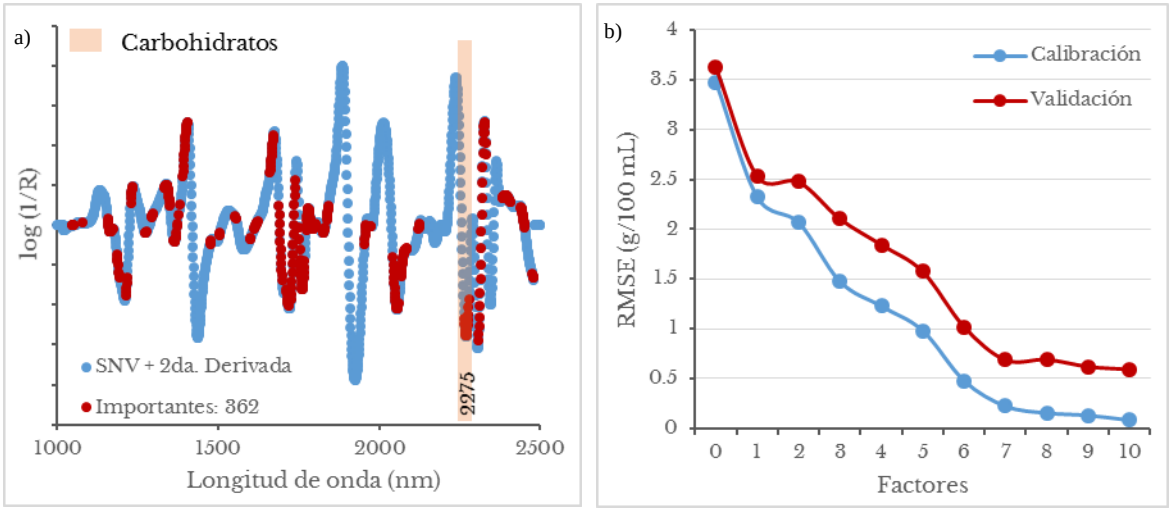
Los datos estadísticos de los distintos modelos PLS entrenados para determinar la concentración de carbohidratos en suplementos alimenticios se especifican en la tabla 20. Los parámetros para su desarrollo y selección se encuentran en el capítulo anterior.

*Tabla 20 Datos estadísticos de los modelos PLS para cuantificar carbohidratos en suplementos alimenticios*

| Pretratamiento | Factores | $R^2$Cal | RMSEC | $R^2$ Val | RMSEV | $R^2$Pred | RMSEP |
|---|---|---|---|---|---|---|---|
| Ninguno | 10 | 0.9924 | 0.3031 | 0.8401 | 1.4481 | 0.8925 | 1.1799 |
| **SNV+2da. Derivada** | **7** | **0.9959** | **0.2232** | **0.9640** | **0.6870** | **0.9438** | **0.8530** |
| 2da. Derivada | 8 | 0.9966 | 0.2019 | 0.9697 | 0.6307 | 0.9402 | 0.8801 |
| SNV+DT | 10 | 0.9975 | 0.1729 | 0.9372 | 0.9077 | 0.9214 | 1.0094 |

$R^2$: coeficiente de correlación, RMSEC: error cuadrático medio de la calibración, RMSECV: error cuadrático medio de la validación cruzada, RMSEP: error cuadrático medio de la predicción. SNV: Variable normal estándar, DT: Detrending.
Fuente: autor

El modelo elegido para cuantificar carbohidratos fue el que se entrenó con los datos espectrales corregidos con la variable normal estándar (SNV) junto con la segunda derivada de Savitzky-Golay del segundo orden polinomial con un total de 39 puntos de suavizado, ya que el valor del error en la predicción de carbohidratos en muestras desconocidas fue de ±0.853 g/100 mL. Se verificó que las señales de las vibraciones características de los carbohidratos no hubiesen sido removidas con el pretratamiento y, como se observa en la figura 42a, siguen presentes. Además, dentro de las variables importantes se encuentra la región de los 2275 nm asociada a los carbohidratos.



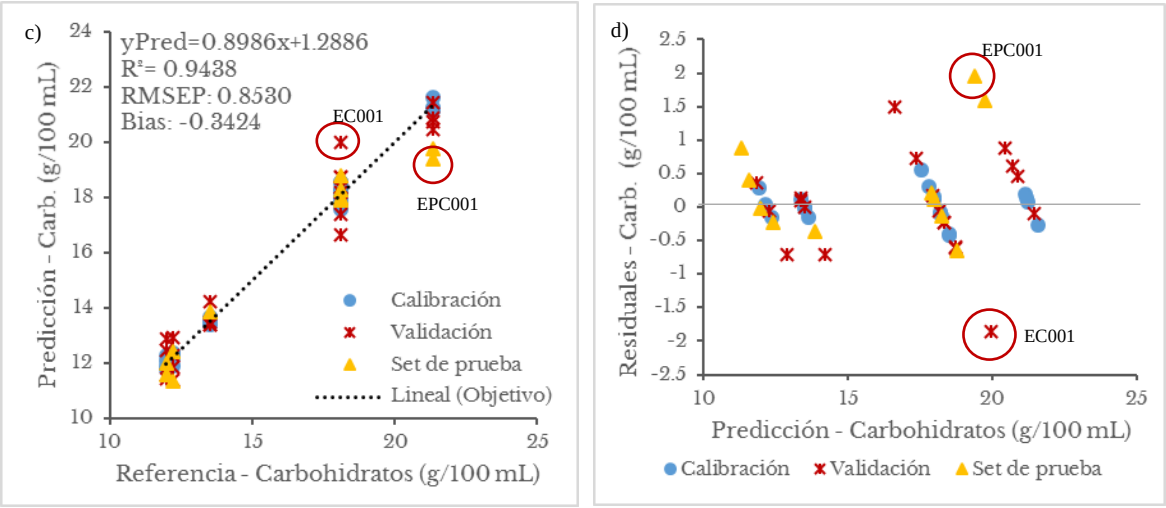



*Figura 42 Gráficos de la revisión general del modelo PLS: Carbohidratos*

a) Espectro NIR promedio resaltando las variables importantes, b)Gráfico RMSE, c) Regresión Lineal, d) Residuales.
Fuente: autor

El número de factores óptimos fue 7 y con el gráfico de RMSE (fig. 42b) se confirma lo anterior al observarse que en el octavo componente el valor del error en la validación cruzada aumenta ligeramente. Esto significa que emplear más variables latentes generaría un sobreajuste en el modelo.

Finalmente, para visualizar el desempeño del modelo se revisó el gráfico de regresión (fig. 42c). En él se miran que los objetos, en su mayoría, se ajustan a la línea objetivo. Aun así, existieron algunos que el modelo describió pobremente y esto se corroboró al estudiar el gráfico Residuales (fig. 42d) en donde dichas muestras estuvieron alejadas del resto sin manifestar un comportamiento atípico.

### *4.2.5 Modelos de calibración - Lípidos*

Los resultados del desempeño de los modelos entrenados para cuantificar lípidos se encuentran en la tabla 21. El que generó el menor valor de error en el set de prueba fue el que empleó datos corregidos con la segunda derivada Savitzky-Golay, sin embargo los valores de $R^2$ y de RMSE en la validación cruzada y en la prueba fueron iguales, lo que es un signo de sobreajuste. Por tanto, dicho modelo fue no se consideró como opción.

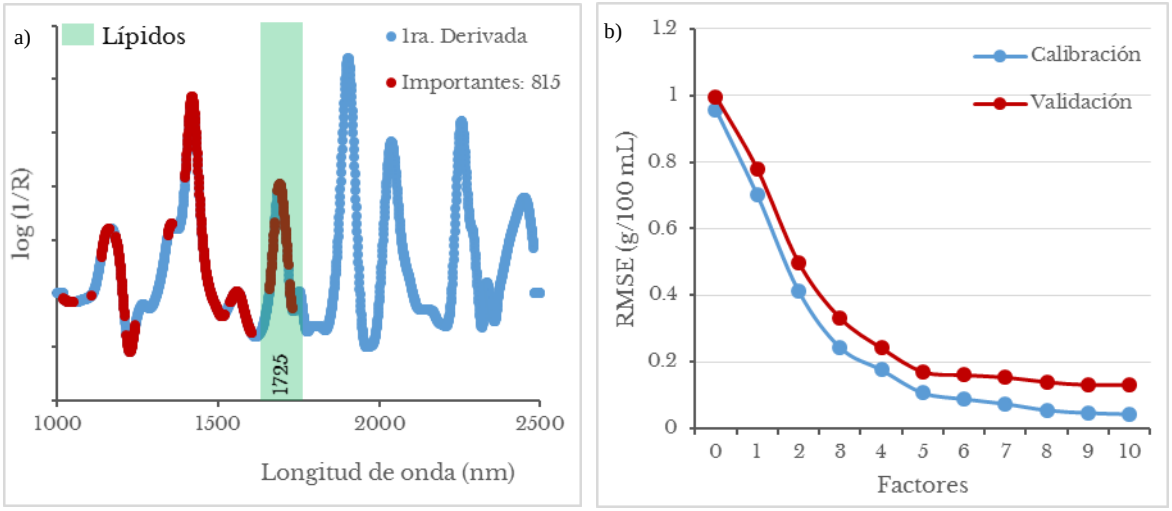
*Tabla 21 Datos estadísticos de los modelos PLS para cuantificar lípidos en suplementos alimenticios*

| Pretratamiento | Factores | $R^2$Cal | RMSEC | $R^2$ Val | RMSECV | $R^2$Pred | RMSEP |
|---|---|---|---|---|---|---|---|
| Ninguno | 6 | 0.9863 | 0.112 | 0.9585 | 0.2029 | 0.9448 | 0.2336 |
| **Ira. Derivada** | **5** | **0.9876** | **0.1062** | **0.971** | **0.1698** | **0.9638** | **0.1895** |
| 2da. Derivada | 6 | 0.9931 | 0.0793 | 0.9771 | 0.1507 | 0.977 | 0.1506 |
| SNV+DT | 4 | 0.9872 | 0.1081 | 0.9734 | 0.1625 | 0.9597 | 0.1996 |

$R^2$: coeficiente de correlación, RMSEC: error cuadrático medio de la calibración, RMSECV: error cuadrático medio de la validación cruzada, RMSEP: error cuadrático medio de la predicción. SNV: Variable normal estándar, DT: Detrending.
Fuente: autor

El modelo elegido para determinar el contenido de lípidos fue al que se le entrenó a partir de datos preprocesados con la primera derivada Savitzky-Golay (fig. xa) del segundo orden polinomial y con 39 puntos de suavizado, ya que obtuvo un error de estimación de ±0.1895 g/100 mL. Además, requirió la información de 815 variables de las 1501 utilizadas lo que significa que tiene el potencial de ser simplificado.

El número de factores óptimos fue cinco, ya que a partir de él la reducción del valor de RMSE en la validación cruzada no es significativo (fig. 43b). Asimismo, el desempeño del modelo fue aceptable y esto se corrobora al observar a la mayoría de las muestras ajustarse a la línea de regresión (fig. 43c). Las que no lo hicieron fue porque tuvieron un alto valor residual como se indica en la figura 43d, específicamente el objeto EPC001. Éste último no fue catalogado como atípico porque no perturba al modelo, simplemente fue descrito pobremente por él.



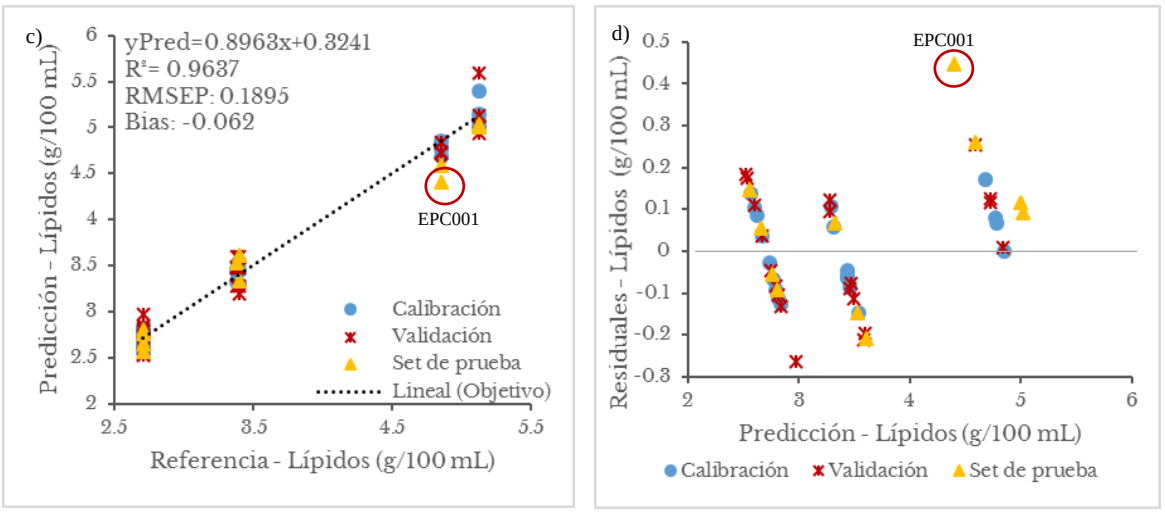



*Figura 43 Gráficos de la revisión general del modelo PLS: Lípidos*

$R^2$: coeficiente de correlación, RMSEC: error cuadrático medio de la calibración, RMSECV: error cuadrático medio de la validación cruzada, RMSEP: error cuadrático medio de la predicción. SNV: Variable normal estándar, DT: Detrending.
Fuente: autor

En la tabla 22 se resumen los parámetros del entrenamiento y los datos estadísticos de los modelos cuantitativos.

*Tabla 22 Parámetros y datos estadísticos de los modelos NIR cuantitativos: Suplementos alimenticios*

| Parámetros | Proteínas | Carbohidratos | Lípidos |
|---|---|---|---|
| Muestras del set de calibración | 25 | 25 | 25 |
| Muestras del set de prueba | 11 | 11 | 11 |
| Algoritmo | PLS | PLS | PLS |
| Rango espectral (nm) | 1000-2500 | 1000-2500 | 1000-2500 |
| Número de variables | 1501 | 1501 | 1501 |
| Rango de Concentración (g/100 mL) | 3.07-5.54 | 11.97-21.34 | 2.71-5.12 |
| Escala | Centrado a la media | Centrado a la media | Centrado a la media |
| Pretratamientos espectrales | Segunda Derivada | SNV+2da. Derivada | 1ra. Derivada |
| Factores | 8 | 7 | 5 |
| $R^2$ Calibración | 0.9943 | 0.9959 | 0.9876 |
| RMSEC | 0.0638 | 0.2232 | 0.1062 |
| $R^2$ Validación cruzada | 0.9534 | 0.9640 | 0.971 |
| RMSECV | 0.1908 | 0.6870 | 0.1698 |
| Objetos con alto valor residual | 1 | 1 | 0 |
| $R^2$ Predicción | 0.9465 | 0.9438 | 0.9638 |
| RMSEP | 0.1965 | 0.8530 | 0.1895 |
| Objetos con alto valor residual | 0 | 1 | 1 |

$R^2$: coeficiente de correlación, RMSEC: error cuadrático medio de la calibración, RMSECV: error cuadrático medio de la validación cruzada, RMSEP: error cuadrático medio de la predicción. PLS: Mínimos cuadrados parciales, SNV: Variable normal estándar.
Fuente: autor

*“La ciencia está hecha de errores; pero son errores que son útiles de cometer, porque guían poco a poco a la virtud”.*

*Julio Verne*

# 5

# Conclusiones

En este capítulo se analizan los resultados obtenidos de la aplicación de la espectroscopía NIR para determinar la composición nutrimental de un alimento, comenzando con un planteamiento de cuáles fueron los aspectos favorables y los obstáculos durante el desarrollo de los modelos cualitativos y cuantitativos. Asimismo, contempla la discusión, conclusiones, las recomendaciones y continuidad.

Se logró optimizar la infraestructura del Instituto de Ingeniería mediante el arranque y puesta en marcha del espectrómetro infrarrojo cercano. Para ello fue necesario solicitar el cambio total del accesorio NIRA así como la actualización del

software del equipo, permitiendo así establecer las condiciones óptimas de operación tanto para la obtención de espectros como el para el procesamiento de los mismos.

Se realizaron distintos modelos cualitativos y cuantitativos aplicando quimiometría, los primeros para discriminar entre las muestras y los segundos, para determinar el contenido nutricional de las mezclas de harinas libres de gluten y de los suplementos alimenticios.

Tras la exploración previa de los datos con PCA, fue posible mostrar gráficamente la existencia de patrones en el grupo de muestras analizadas y determinar que existe el potencial de realizar modelos de clasificación. Además, permitió identificar a los objetos anómalos antes de entrenar al modelo, disminuyendo así, la existencia de muestras que pudieran perturbarlo (Miller, 2010).

En el entrenamiento de los modelos cualitativos se empleó el algoritmo PLS-DA y éste demostró ser útil para identificar a un objeto perteneciente a una clase de otros que no lo eran, además de que arrojó resultados sencillos de interpretar. En el caso de las harinas, tuvo una precisión del 100% para discriminar entre una formulación para hacer panecillos de una mezcla formada únicamente de harinas.

En cuanto a los suplementos alimenticios, confirmó su precisión al clasificar a los objetos según la marca comercial. Sin embargo, tuvo un pobre desempeño para catalogar a las muestras por sabor; con las de vainilla presentó un 75% de precisión, mientras que con las de fresa, un 58% con el set de prueba. La excepción fueron las de sabor chocolate, cuya precisión fue del 100%. El deficiente rendimiento pudo deberse al hecho de que no existía una proporción uniforme entre las muestras de vainilla y fresa.

Para los modelos cuantitativos, se empleó el algoritmo PLS. El desempeño obtenido para determinar el contenido de carbohidratos, lípidos y proteínas fue aceptable, al generar valores de $R^2$ en el set de prueba entre 0.94-0.96 para los suplementos alimenticios. En el grupo de harinas, los modelos de calibración para cuantificar carbohidratos y lípidos fueron buenos, 0.96 y 0.97 respectivamente; sin embargo, en el de las proteínas sólo fue aceptable al tener una correlación de 0.91.

Probablemente, esto se debió a que el rango de concentración de ésta biomolécula fue menor comparada con las otras.

Otro indicador del desempeño fue el error cuadrático medio (RMSE) y éste osciló entre 0.19 a 0.85 g/100 mL para los suplementos y de 0.72 a 1.96 g/100g para las harinas. Éste parámetro se asocia al valor de los residuales, así que mientras existan objetos pobremente descritos por los modelos, el error tenderá a ser alto.

El pretratamiento más utilizado para el desarrollo de los modelos fue la segunda derivada Savitzky-Golay. Permitió mejorar la linealidad, el error y emplear un menor número de factores para el cálculo del modelo. Lo que no corrigió fue el espectro de los objetos con alto valor residual. Si en el modelo desarrollado a partir de datos sin pretratamientos, se identificó a un posible objeto anómalo, éste no cambiará su comportamiento aunque se le apliquen distintos preprocesamientos espectrales.

La presente investigación ha demostrado que la espectroscopía por infrarrojo cercano (NIR) junto con las técnicas quimiométricas adecuadas, permiten desarrollar modelos con un buen desempeño para discriminar entre clases, siempre y cuando exista una proporción adecuada de la categoría a analizar (Ziegler, Leitenberger, Longin, Würschum, Carle y Schweiggert, 2016). Asimismo, permite cuantificar carbohidratos, lípidos y proteínas en productos alimenticios al obtener coeficientes de determinación ($R^2$) superiores a 0.9 y errores estándares bajos en el set de prueba (Cascant, Garrigues y de la Guardia, 2015).

Esta técnica analítica secundaria, tiene el potencial de sustituir a las bromatológicas en pruebas rutinarias siempre y cuando se demuestre que la exactitud de la predicción es igual o mejor que las antes mencionadas (Burns, 2008).

A pesar de que presenta muchas bondades la espectroscopía NIR, como ser rápida y no destructiva, el entrenamiento del modelo es tardado y requiere de una gran cantidad de objetos para garantizar que la población cubra todas las variaciones posibles. Se requiere demás conocimientos avanzados en estadística y en modelado de datos, ya que si un modelo no es óptimo, el software ofrece ninguna recomendación de cómo mejorarlo, por lo que el procesamiento e interpretación quedará a criterio del usuario (Miller, 2010).

## 5.1 Recomendaciones

Para mejorar el desempeño de los modelos cualitativos y cuantitativos se requiere incluir más muestras al set de entrenamiento que aporten variaciones que probablemente, estuvieron presentes en menor proporción en el estudio. Hacer lo contrario sólo mejora la linealidad y no el RMSE (Burns, 2008).

No excederse en el uso y/o combinación de los preprocesamientos espectrales. Éstos son de gran utilidad en el incremento de los indicadores de desempeño, pero el abuso de los mismos generaría un sobreajuste en el modelo y, por tanto, predicciones imprecisas.

Los modelos cualitativos y cuantitativos creados se limitan a las marcas comerciales y a la presentación de las harinas y suplementos alimenticios utilizados en este estudio. Por tanto, se aconseja evitar su uso en muestras distintas a éstas ya que los resultados no serían confiables.

## 5.2 Continuidad

Los modelos desarrollados pueden optimizarse y abarcar un rango más amplio de productos, es decir, incluir marcas comerciales que no fueron tomadas en cuenta para así tener un modelo general que permita cuantificar carbohidratos, lípidos y proteínas en todo tipo de harinas y en cualquier suplemento alimenticio. Además, se podrían entrenar nuevos modelos para otro tipo de alimento como carnes, lácteos, frutos, entre otros.

Asimismo, se tiene la inquietud de desarrollar modelos que permitan evaluar la transformación que sufren los alimentos durante su proceso de elaboración, para registrar los cambios químicos y físicos que éste presenta.

## REFERENCIAS


(AOAC), T. A. (2016). *Official Methods and Recommended Practices of the AOCS*. (6ta. Ed.). Estados Unidos. Disponible en: http://www.aocs.org

AACC International. *Approved Methods of Analysis* (11va Ed.). AACC International, St. Paul, MN, U.S.A.

Anderson, S. (2007). Determination of fat, moisture, and protein in meat and meat products by using the FOSS FoodScan Near-Infrared Spectrophotometer with FOSS Artificial Neural Network Calibration Model and Associated Database: collaborative study. *Journal Of AOAC International*, 90(4), 1073-1083.

Azzouz, T., Puigdoménech, A., Aragay, M. y Tauler, R. (2003). Comparison between different data pre-treatment methods in the analysis of forage samples using near-infrared diffuse. *Analytica Chimica Acta 484* (1), 121-134.

Badui, S. (2006). Química de los alimentos (4a. ed.). México: Pearson.

Bhunia A. K., Kim M. S., y Taitt C. R. (2015). 7. Vibrational Spectroscopy for Food Quality and Safety. En *High Throughput Screening for Food Safety Assessment: Biosensor Technologies, Hyperspectral Imaging and Practical Applications*. Elsevier.

Burns, D. A. y Ciurczak, E. W. (2008). *Handbook of Near-infrared Analysis*. Boca Raton, FL: CRC Press.

CAMO Software AS (2016). *The Unscrambler® X* (Versión 10.4). Disponible en: http://www.camo.com/rt/Products/Unscrambler/unscrambler.html

Cascant, M. M., Garrigues, S. y de la Guardia, M. (2015). Direct determination of major components in human diets and baby foods. *Analytical and bioanalytical chemistry*, 407(7), 1961-1972.

Cheung, P. C., y Mehta, B. M. (2015). *Handbook of Food Chemistry*. Berlin: Springer.

Gobierno del Estado de Baja California-GEBC (2015). *Nuestro Estado*. Recuperado de: http://www.bajacalifornia.gob.mx/portal/nuestro_estado/municipios/mexicali/sectorprod.jsp

International Organization for Standardization ISO (2016). *ISO Standards catalogue.*

Irudayaraj, J. y Reh, C. (2008). *Nondestructive Testing of Food Quality*. Ames, Iowa: Wiley-Blackwell.

Jha, S. N. (2010). *Nondestructive Evaluation of Food Quality: Theory and Practice*. Springer Berlin Heidelberg.

Kim, Y., Singh, M. y Kays, S. E. (2007). Near-infrared spectroscopic analysis of macronutrients and energy in homogenized meals. *Food chemistry*, 105(3), 1248-1255.

Li Vigni, M., Durante, C. y Cocchi, M. (2013). Chapter 3: Exploratory Data Analysis. *Data Handling In Science And Technology, 28*(Chemometrics in Food Chemistry), 55-126. doi:10.1016/B978-0-444-59528-7.00003-X

Manley, M. (2014). Near-infrared spectroscopy and hyperspectral imaging: non-destructive analysis of biological materials. *Chemical Society Reviews*, 43(24), 8200-8214. doi:10.10.39/c4cs00062e

Miller, J. N., & Miller, J. C. (2010). *Statistics and Chemometrics for Analytical Chemistry*. Harlow, England: Prentice Hall.

Moros, J., Garrigues, S., y De la Guardia, M. l. (2010). Vibrational spectroscopy provides a green tool for multi-component analysis. *Trends In Analytical Chemistry*, 29(Green analytical chemistry), 578-591. doi:10.1016/j.trac.2009.12.012

Mucherino, A., Papajorgji, P. J. y Pardalos, P. M. (2009). *Data Mining in Agriculture*. New York, NY: Springer.

Nelson, D. L. y Cox, M. M. (2005). *Lehninger: Principles of Biochemistry* (4a. ed.). W.H. Freeman and Company Nelson.

Nielsen, S. S. (2010). *Food Analysis*. Boston, MA: Springer US.

Phan-Thien, K., Golic, M., Wright, G. C., y Lee, N. A. (2011). Feasibility of estimating peanut essential minerals by near infrared reflectance spectroscopy. *Sensing And Instrumentation For Food Quality And Safety*, (1), 43.

Picó, Y. (2012). *Chemical Analysis of Food: Techniques and Applications*. Amsterdam: Academic Press.

Secretaría de Economía. (2010). Norma Oficial Mexicana NOM-051-SCFI/SSA1-2010, Especificaciones generales de etiquetado para alimentos y bebidas no alcohólicas preenvasados-Información comercial y sanitaria. *Diario Oficial de la Federación*, 5 de abril de 2010.

Skoog, D., Holler, F. y Nieman, T. (2001). *Principios de Análisis Instrumental*. Madrid, España: McGraw-Hill.

Swarbrick, B. (2016) Near-Infrared spectroscopy and its role in scientific and engineering applications. En M. Kutz. *Handbook of measurement in Science and Engineering* (vol 3, pp.2583-2656). Chichester, West Sussex: Wiley Blackwell.

The American Oil Chemists' Society [AOCS] (2016). *Official Methods and Recommended Practices of the AOCS* (7ma. Ed.). Estados Unidos.

U.S. Department of Agriculture, Agricultural Research Service. (2016). USDA National Nutrient Database for Standard Reference, Release 28. Nutrient Data Laboratory Disponible en: http://www.ars.usda.gov/nutrientdata

U.S. Department of Agriculture, Agricultural Research Service. (2016). USDA Branded Food Products Database. Nutrient Data Laboratory. Disponible en: http://ndb.nal.usda.gov

Wang, W. y Paliwal, J. (2007). Near-infrared spectroscopy and imaging in food quality and safety. *Sensing And Instrumentation For Food Quality And Safety*, (4), 193.

Workman, J. y Weyer, L. (2008). *Practical Guide to Interpretive Near-infrared Spectroscopy*. Boca Raton, FL: CRC Press.

Yukihiro, O. (2012). Near-Infrared Spectroscopy: Its Versatility in Analytical Chemistry. *Analytical Sciences*, (6), 545-563.

Ziegler, J. U., Leitenberger, M., Longin, C. F. H., Würschum, T., Carle, R., y Schweiggert, R. M. (2016). Near-infrared reflectance spectroscopy for the rapid discrimination of kernels and flours of different wheat species. *Journal of Food Composition and Analysis*, 51, 30-36.

Zou, X. y Zhao, J. (2015). *Nondestructive Measurement in Food and Agroproducts*. Dordrecht: Springer.

# Apéndice

*Listado de objetos pertenecientes al set de calibración y validación del grupo Harinas*

| Descripción | ID Muestras |
|---|---|
| Set de entrenamiento | ALM-40%, ALM-55%, ALM-70% B, ALM-80%, AROC-11354511, AROC-13602007, AROC-20071360, AROC-35451111, AROC-45111135 ARR-40%, ARR-64%, ARR-70%, ARR-80%, AVE-40%, AVE-64%, AVE-70%, AVE-80%, COAR-05153050B, COAR-15305005A, COAR-15305005B, COAR-25%A, COAR-30500515B ,COAR-50051530A, COAR-50051530B, COCO-40%, COCO-64%, COCO-70%, COCO-80%.<br>MIX-65%A*, MIX-68%A*, MIX-70%A*, MIX-72%A*, MIX-75%A* |
| Set de prueba | ALM-64%, ALM-70%A, ARR-55%, AVE-55%, COAR-05153050A, COAR-25%B, COAR-30500515A, COCO-55%, AROC-07136020, AROC-60200713, AROC-13602007B, AROC-11113545.<br>MIX-65%B*, MIX-68%B*, MIX-70%B*, MIX-72%B* |

*: Excluidas de los modelos cuantitativos

Fuente: autor

Set de entrenamiento y de prueba: Clasificación de Suplementos Alimenticios según la marca comercial

| | | Marca Comercial | | | | | Total |
|---|---|---|---|---|---|---|---|
| Descripción | ID Muestras | EAV | EP | E | G | P | |
| Set de entrenamiento | EAV001, EAV003, EAV004, EPV001, EPV002, EPV003, EPV004, EC001, EC002, EF001, EV001, EV002, EV003, EV005, EV006, EV007, GC001, GC002, GF001, GV001, PC001, PV001, PV002, PV003, | 3 | 4 | 9 | 4 | 4 | 24 |
| Set de prueba | EAV002, EAV005, EF002, EPF001, EPC001, EV004, EV008, EV009, GV002, GV003, PF001, PV004. | 2 | 2 | 4 | 2 | 2 | 12 |

EAV: Ensure Advance ®, EP: Ensure Plus ®, E: Ensure®, G: Glucerna®, P: Pediasure®.

Fuente: autor.

Set de entrenamiento y de prueba: Clasificación de Suplementos Alimenticios según el sabor

| | | Sabor | | | Total |
|---|---|---|---|---|---|
| Descripción | ID Muestras | V | F | C | |
| Set de entrenamiento | EAV003, EAV004, EAV005, EC001, EF001, EPV001, EPV003, EPV004, EV001, EV003, EV004, EV006, EV008, EV009, GC001, GC002, GF001, GV002, GV003, PC001, PF001, PV001, PV002, PV003. | 17 | 3 | 4 | 24 |
| Set de prueba | EAV001, EAV002, EC002, EF002, EPC001, EPF001, EPV002, EV002, EV005, EV007, GV001, PV004. | 8 | 2 | 2 | 12 |

V: vainilla, F: Fresa, C: Chocolate

Fuente: autor

Set de entrenamiento y de prueba para los modelos de cuantificación: Suplementos Alimenticios

| Descripción | ID Muestras | Total |
|---|---|---|
| Set de entrenamiento | EAV003, EAV004, EAV005, EC001, EF001, EF002, EPV001, EPV003, EPV004, EV001, EV003, EV004, EV005, EV006, EV009, GC001, GC002, GF001, GV002, GV003, PC001, PF001, PV001, PV002, PV003. | 25 |
| Set de prueba | EAV001, EAV002, EC002, EPC001, EPF001, EPV002, EV002, EV007, EV008, GV001, PV004. | 11 |

Fuente: autor