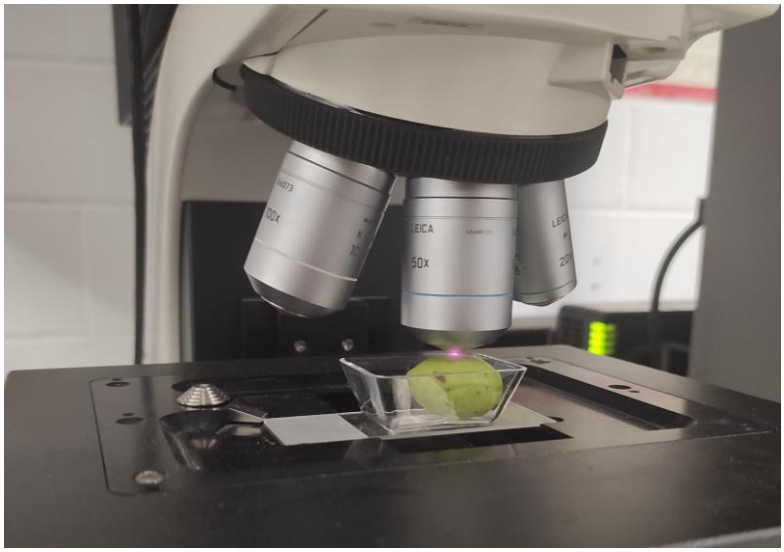
Universidad del País Vasco Euskal Herriko Unibertsitatea
BILBOKO INGENIARITZA ESKOLA
ESCUELA DE INGENIERÍA DE BILBAO

MÁSTER UNIVERSITARIO EN INGENIERÍA DE TELECOMUNICACIONES

TRABAJO FIN DE MASTER

Espectroscopia Raman y Análisis Multivariante de Datos para la detección de fitosanitarios en aceituna entera



Estudiante: Pérez Muñoz, Gorka

Director: Ayesta Ereño, Igor

Curso: 2023-2024

Fecha: Bilbao, 27 de febrero de 2024

RESUMEN

Hoy en día, existen métodos para analizar el contenido de substancias o la composición de materiales usando análisis químicos, los cuales requieren de un laboratorio especializado y un tiempo de varios días para obtener resultados. Sin embargo, actualmente está surgiendo un nuevo enfoque que permite el uso de nueva tecnología para el análisis de muestras de forma más rápida. En la industria alimentaria, por ejemplo, esta tecnología permitiría abaratar costes, ya que con ella sería posible hacer un análisis a cada una de las muestras que forman un lote potencialmente peligroso. De esta forma, solo las muestras detectadas como peligrosas serían descartadas, sin la necesidad de tener que prescindir de todo el lote. Mediante esta tecnología, se evitaría la necesidad de recurrir a especialistas químicos, siendo capaces de obtener resultados rápidamente sin la necesidad de una infraestructura como un laboratorio. En la búsqueda de esa tecnología han aparecido varias opciones que pueden resolver esos dilemas dependiendo de la industria a la que esté enfocada.

Este documento, está enfocado en mostrar el desarrollo y conclusiones obtenidas en torno al uso de la tecnología Raman como herramienta para la obtención de datos de las muestras, en este caso olivas tratadas con fitosanitario. Para la clasificación de los datos y generación de modelos capaces de predecir y clasificar nuevos datos correctamente, se ha hecho uso de análisis de datos multivariante, mediante el software SIMCA. Al final del documento se hallan las conclusiones preliminares respecto al uso de esta tecnología en torno a este proyecto. Estas conclusiones se consideran preliminares ya que el estudio del proyecto sigue en marcha.

LABURPENA

Gaur egun, analisi kimikoak erabiliz substantzien edukia edo materialen osaera aztertzeko metodoak existitzen dira jada. Baina, horretarako, laborategi espezializatuak eta egun batzuetako denbora behar da emaitzak lortzeko. Hala ere, gaur egun beste ikuspegi bat sortzen ari da, laginak azkarrago aztertzeko teknologia berria erabiltzea ahalbidetzen duena. Elikagaien industrian, adibidez, teknologia horrek kostuak merkatzeko aukera emango luke; izan ere, horren bidez, arriskutsua izan daitekeen lote oso bat osatzen duten lagin guztiak azter litezke. Horrela, arriskutsutzat jotzen diren laginak bakarrik baztertuko dira, eta ez da beharrezkoa izango lote osoa baztertzea. Teknologia horren bidez, espezialista kimikoetara jotzeko beharra saihestuko litzateke, eta emaitzak azkar lortzeko gai izango lirateke, laborategi gisako azpiegitura baten beharrik gabe. Teknologia horren bilaketan, dilema horiek konpon ditzaketen hainbat aukera agertu dira, zein industriatara bideratuta dagoenaren arabera.

Dokumentu honen bidez, Raman teknologia laginen datuak lortzeko tresna gisa erabiltzearen inguruan lortutako garapena eta ondorioak erakutsi nahi dira. Kasu honetan, laginak landare-osasunerako produktuekin tratatutako olibak. Datuak sailkatzeko eta datu berriak zuzen aurreikusi eta sailkatzeko gai diren ereduak sortzeko, aldagai anitzeko datuen analisia erabili da, SIMCA softwarearen bidez. Dokumentuaren amaieran, teknologia hori proiektu honen inguruan erabiltzeari buruzko aurretiko ondorioak daude. Ondorio horiek atarikotzat hartzen dira, proiektuaren azterketak martxan jarraitzen baitu.

ABSTRACT

Nowadays, there are methods for analyzing the content of substances or the composition of materials using chemical analysis, which require a specialized laboratory and a time of several days to obtain results. However, a new approach is now emerging that allows the use of new technology to analyze samples more quickly. In the food industry, for example, this technology would allow lower costs, since it would be possible to analyze each of the samples that make up a potentially hazardous batch. In this way, only the samples detected as hazardous would be discarded, without the need to dispense with the entire batch. Through this technology, the need to resort to chemical specialists would be avoided, being able to obtain results quickly without the need for an infrastructure such as a laboratory. In the search for such technology, several options have appeared that can solve these dilemmas depending on the industry that is focused on.

This document is focused on showing the development and conclusions obtained from the use of Raman technology as a tool for obtaining data from samples, in this case olives treated with phytosanitary products. For the classification of the data and generation of models capable of predicting and classifying new data correctly, multivariate data analysis has been used, using SIMCA software. Preliminary conclusions regarding the use of this technology in this project can be found at the end of the document. These conclusions are considered preliminary as the study of the project is still in progress.

Índice

Lista de Tablas

Lista de Ilustraciones

Lista de acrónimos

HPLC	High Performance Liquid Chromatography
GC	Gas Chromatography
MVDA	Multivariate Data Analysis
PCA	Principal Component Analysis
PLS	Partial Least Squares regression
OPLS	Orthogonal Partial Least Squares regression
PLS-DA	Partial Least Squares regression – Discriminant Analysis
OPLS-DA	Orthogonal Partial Least Squares regression – Discriminant Analysis

1. Introducción

Con el paso del tiempo, se han impuesto muchos estándares y límites que deben cumplir los alimentos para ser aceptados en el mercado [1]. Estos estándares han sido registrados legalmente, por lo que se ha convertido de carácter obligatorio el cumplimiento de los mismos. El objetivo de ello es que la comida que circula en el mercado y llega a manos del consumidor sea lo más saludable posible, además de dar los medios para que el consumidor tenga acceso a mejor información sobre la composición de dichos productos, otorgando así mejor capacidad de elección.

Es por esto que la industria alimentaria ha tenido que adaptarse para cumplir con esos estándares y han entrado organismos en juego que se aseguran de que los productos cumplan con lo establecido mediante pruebas en el laboratorio.

2. Contexto

En esta sección se tratará de explicar la importancia que tiene el sector alimentario en la salud de los consumidores. Se mencionarán posibles medidas tomadas en la actualidad para asegurar la veracidad y evitar el fraude alimenticio y, a continuación, se proseguirá hablando sobre la industria del aceite de oliva y el cultivo olivarero. Enlazado a esto, se explicará la importancia del uso de pesticidas y de los efectos que estos pueden causar en la salud. También se hablará de los métodos que existen para poder detectar dichos pesticidas o para medir la concentración hallada en los productos de origen vegetal. Por último, se explicarán el procesado y los métodos de análisis estadístico usados en el proyecto.

2.1. Sector alimentario

El sector alimentario en España, es un sector clave en la economía del país, tanto por su contribución al producto interno bruto (PIB) como por su papel en el empleo y la exportación. La industria abarca una amplia gama de actividades, desde la producción y transformación de materias primas hasta la distribución y comercialización de alimentos y bebidas. La diversidad y calidad de los productos alimentarios españoles, así como su tradición culinaria, han posicionado a España como un referente en el mercado internacional.

Esta calidad de los productos ha sido conseguida gracias a que Los gobiernos nacionales y las organizaciones regionales se encuentran en una posición única para desarrollar, supervisar, controlar y hacer cumplir las medidas contra el fraude alimentario.

2.1.1. Fraude alimentario

El fraude alimentario ocurre cuando un defraudador engaña intencionadamente a un cliente sobre la calidad y/o el contenido de los alimentos que desea comprar, y dicho acto se realiza para obtener una ventaja indebida, casi siempre económica, para el defraudador.

El fraude alimentario ha acosado a los gobiernos durante siglos, y las respuestas legales al mismo se han adaptado de forma única a las sensibilidades de la época, sin que exista actualmente ninguna definición legal reconocida internacionalmente. Hoy en día, una regulación eficaz del fraude alimentario debe tener en cuenta las complejidades modernas, como el creciente comercio mundial de variedades de productos alimentarios e ingredientes susceptibles de fraude; la prolongación de las cadenas opacas de suministro de alimentos; el alcance internacional de las amenazas para la salud pública; la creciente sofisticación en la comisión de fraudes; los rápidos avances en la comercialización en línea y el comercio electrónico, que multiplican las oportunidades de fraude alimentario; la creciente cobertura mediática de los escándalos de fraude alimentario; y el mayor interés de los consumidores por la autenticidad y la integridad de los alimentos.

Los escándalos de fraude alimentario en algunas de las principales economías productoras de alimentos en un pasado no muy lejano, como el incidente de la carne de caballo como carne de vacuno en la Unión Europea en 2013 [2], el escándalo del Fipronil en los huevos en 2017 en Asia y la Unión Europea [3], o la leche contaminada con melamina en China en 2008 [4], ponen de manifiesto las complejidades en la regulación del fraude alimentario. No menos compleja es la regulación del fraude alimentario a menor escala, que también puede incluir la sustitución de ingredientes clave por alternativas de menor calidad, la venta de alimentos convencionales como productos orgánicos, incorrecto etiquetado del peso y la sustitución de variedades caras de pescado por especies de bajo valor [5] .

Los gobiernos nacionales y las organizaciones regionales se encuentran en una posición única para desarrollar, supervisar, controlar y hacer cumplir las medidas contra el fraude alimentario. Por difícil que sea regular el fraude alimentario, es fundamental que los gobiernos nacionales actúen. Como ya se ha mencionado, la no aplicación o la aplicación insuficiente de la legislación contra el fraude alimentario socava la credibilidad de la acción gubernamental, dejando que los perjuicios económicos y para la salud pública reinen sin control.

2.1.2. Industria del aceite de oliva

Esta es una de las industrias más relevantes del país, siendo el sector del aceite de oliva un pilar fundamental en el sistema agroalimentario español.

España es líder mundial en superficie, producción, y comercio exterior gracias a la tradición olivarera de nuestro país y a una industria tecnológicamente avanzada y profesional capaz de obtener aceites de gran calidad. La producción española de aceite de oliva supone el 70% de la producción de la UE y el 45% de la mundial [6] .

El sector no solo tiene una indiscutible importancia económica, sino que también tiene una gran repercusión social, ambiental y territorial. Más de 350.000 agricultores se dedican al cultivo del olivar, el sector mantiene unos 15.000 empleos en la industria y genera más de 32 millones de jornales por campaña [6].

Asimismo, los procesos de transformación y distribución de sus producciones, incluidos sus subproductos, constituyen la principal actividad de numerosos municipios y una industria asociada que vertebra y cohesiona, en muchos casos, el medio rural donde se asienta, apoyándose en un fuerte movimiento cooperativo de base.

Es por esto que el sector necesita una forma de mostrar la calidad de sus productos para convencer a los clientes de que sus productos siguen siendo de la mejor calidad posible, como es el etiquetado del producto el cual muestra la calidad de cada uno respaldado por las pruebas pertinentes. Una parte sumamente importante a la hora de mantener esa calidad distinguida de estos productos es la prevención, el control y el tratamiento de plagas de forma adecuada.

La puesta en marcha de la gestión integrada de plagas (GIP) es un paso más en el manejo de la sanidad vegetal de las explotaciones agrícolas. Combina medidas culturales, tratamientos químicos y soluciones alternativas. Todo con el objetivo de mantener las plagas por debajo de los umbrales establecidos para garantizar la rentabilidad económica respetando el medio ambiente y la salud del agricultor y el consumidor. Existen guías de gestión integrada de plagas publicadas por el ministerio de agricultura, alimentación y medioambiente para informar a los agricultores y que estos estén debidamente formados y puedan realizar la GIP de forma adecuada. Estas guías son documentos técnicos que recogen las distintas estrategias a emplear ante problemas fitosanitarios que puedan aparecer en el cultivo.

Una opción, siempre que se disponga de esas alternativas, podría ser el control biológico o tecnológico mediante el uso de depredadores naturales de la plaga o trampas y feromonas. Otra opción, será el uso de productos fitosanitarios. En este caso, el agricultor se asesorará sobre el producto más indicado para su explotación por su localización y el tipo y la extensión de la plaga. A la hora de la realización del tratamiento, el aplicador debe respetar las indicaciones que figuran en la etiqueta del producto fitosanitario y la distancia con los puntos de agua y utilizar equipos de protección individual para la prevención de riesgos laborales debido a que estos productos pueden ser perjudiciales para la salud y el medio ambiente si no se gestionan bien.

2.1.3. Productos fitosanitarios

Los productos fitosanitarios son medios imprescindibles para la producción agrícola, tanto bajo los sistemas convencionales de agricultura, como bajo otros sistemas de agricultura, como la integrada o la ecológica, pues los estragos potenciales de las diferentes clases de plagas, determinarían la inviabilidad de muchos cultivos en las zonas de producción de mayor interés económico y social e incluso la posibilidad de mantener almacenadas las cosechas.

Sin embargo, la utilización de productos fitosanitarios puede provocar efectos adversos (esterilidad, anemia aplástica, cáncer y trastornos diversos a largo plazo en individuos directamente expuestos) y es imprescindible que estos efectos no sean en ningún modo peligrosos para la salud humana, ni tampoco que lleguen a presentar niveles de riesgo inaceptables para el medio ambiente, incluidas la flora y la fauna silvestres [7].

En consecuencia, el Estado aplica los mecanismos necesarios, los cuales muchos vienen determinados por Europa, para que sólo puedan comercializarse aquellos productos fitosanitarios que sean útiles y eficaces para combatir las plagas, pero que no comporten otros riesgos colaterales. Para que un producto pueda comercializarse debe estar autorizado previamente e inscrito necesariamente en el Registro Oficial de Productos Fitosanitarios [8].

Tan importante como la regulación de esta clase de productos, es la comprobación de que estos se están usando de forma adecuada siguiendo las indicaciones pertinentes para su uso. Dicha comprobación se efectúa mediante un análisis químico de substancias enfocado aleatoriamente en ciertos productos pertenecientes a un lote. Si el análisis resulta desfavorable, todo el lote es descartado, lo que repercute en pérdidas materiales y monetarias. Esto es debido a que dichos análisis requieren de un laboratorio y personal especializado para ser efectuados, además de que los resultados no son instantáneos. Esto se traduce en que es necesaria cierta inversión de capital y tiempo, por lo que no sería viable hacer un análisis químico a cada producto de cada lote.

Por esta razón, se están investigando nuevos métodos de análisis y obtención de datos más rentables que sean rápidos, fáciles de usar y puedan, al menos, dar la información suficiente para saber si el producto cumple con las características mínimas para no ser descartado, evitando así tener que desechar todo un lote completo sin tener conocimiento de qué productos cumplen o no con la normativa vigente.

2.2. Métodos de detección

La química analítica es una rama de la química aplicada que estudia a profundidad la composición de la materia. La química analítica separa, identifica, mide y estudia los componentes de una sustancia; también estudia, desarrolla y mejora las herramientas y los métodos para analizar muestras.

Los contaminantes orgánicos, como los fitosanitarios, son determinados en alimentos, frutas o verduras, entre otros, por medio de una serie de etapas analíticas presentes en el análisis químico también conocidas como proceso de medida químico. En este proceso, la muestra es sometida a un conjunto de operaciones que permiten aislar el compuesto de interés (analito) desde la matriz por medio de una extracción y otras etapas necesarias para su posterior cuantificación.

El análisis puede ser cualitativo, enfocado en determinar la presencia o ausencia de un compuesto en particular, pero no la masa o la concentración, o cuantitativo, centrado en medir las cantidades de constituyentes químicos particulares presentes en una sustancia. Sin embargo, la química analítica moderna está dominada por la instrumentación sofisticada, haciendo posible la detección de substancias y su concentración mediante los métodos instrumentales más avanzados.

Algunos de los métodos instrumentales más avanzados son los siguientes.

2.2.1. Cromatografía

La cromatografía es un método de separación de mezclas complejas, que es ampliamente utilizado en diversas ramas de la ciencia. Para determinar la presencia de ciertos compuestos, como los plaguicidas presentes en alimentos, se suele recurrir a las técnicas cromatográficas las cuales permiten separar los compuestos de interés presentes en la muestra. Comúnmente, dependiendo de las características químicas del analito, se utiliza cromatografía líquida de alta resolución (HPLC) o cromatografía gaseosa (GC). Posterior a la separación de estos compuestos, tiene lugar la etapa de cuantificación. Para ello, es necesario utilizar alguna propiedad química que posean los analitos, de tal forma que esta pueda ser interpretada en forma de concentración. En este proceso, es necesario confirmar la identidad del analito comparando sus características químicas utilizadas en la cuantificación con un estándar de referencia.

La cromatografía líquida de alta resolución y la cromatografía gaseosa son los métodos de análisis químico más usados actualmente, aunque lo más efectivo es usarlos en combinación con la espectrometría de masas. A pesar de ello, uno de los inconvenientes para el uso de esta técnica, es que requieren de una preparación previa de las muestras, además de necesitar también bastante tiempo para la obtención de resultados.

2.2.2. Espectroscopía de masas

La espectrometría de masas, es una técnica de análisis cualitativo, de amplia utilización para la determinación de estructuras orgánicas.

Hay que aclarar que la espectrometría de masas tiene muy poco en común con las técnicas clásicas de espectrofotometría, ya que, en sentido estricto, no es propiamente un método espectroscópico (desde el punto de vista clásico, un espectro es una información bidimensional que representa un parámetro relacionado con la emisión o absorción de una radiación con la energía de dicha radiación). En la espectrometría de masas, no se utiliza ningún tipo de radiación, por lo que básicamente, no puede ser considerada como una técnica espectroscópica. Otra diferencia esencial que presenta la espectrometría de masas con las espectroscopías clásicas es que, en estos últimos métodos, los procesos que se originan son puramente físicos, no destructivos, de forma que la muestra utilizada para la obtención del espectro no se modifica químicamente y se puede volver a recuperar. Por contra, en la espectrometría de masas, durante la obtención del espectro tienen lugar procesos químicos, con lo que la muestra utilizada se destruye y no puede recuperarse. Este hecho no es un inconveniente grave, ya que la cantidad de muestra necesaria para la obtención de un espectro de masas, es del orden de μg.

La espectrometría de masas está basada en la obtención de iones a partir de moléculas orgánicas en fase gaseosa. Una vez obtenidos estos iones, se separan de acuerdo con su masa y su carga, y

finalmente se detectan por medio de un dispositivo adecuado. Un espectro de masas será, en consecuencia, una información bidimensional que representa un parámetro relacionado con la abundancia de los diferentes tipos de iones en función de la relación masa/carga de cada uno de ellos.

Como ya se ha mencionado, los procesos que tienen lugar en un espectrómetro de masas son de naturaleza química, por lo que la presencia y abundancia en el espectro de determinados tipos de iones, identificables a partir de su masa, será función de la estructura química de cada compuesto. La información ofrecida por un espectro de masas es, de alguna forma, comparable a la obtenida mediante una gran cantidad de reacciones de las utilizadas para la determinación de estructuras por vía química, por lo que la espectrometría de masas puede ofrecer una enorme cantidad de información sobre un compuesto determinado.

A pesar de ello, esta técnica presenta diversas desventajas. Tienen diferentes tipos de interferencias, lo que puede llegar a afectar la precisión y la exactitud de los resultados. En cuanto a las muestras, solo pueden analizar elementos de uno en uno y no se pueden analizar todos los elementos del Sistema Periódico. Además, por ser una técnica de absorción, sus curvas de calibrado sólo son lineales en un corto rango de concentración.

2.2.3. Espectroscopía Raman

La espectroscopia Raman constituye una técnica fotónica de alta resolución no destructiva que proporciona información detallada acerca de la estructura química, fase y polimorfia, cristalinidad, así como las interacciones moleculares de un material. Esta técnica se fundamenta en la interacción entre la luz y los enlaces químicos presentes en el material.

La técnica Raman se basa en la dispersión de la luz, en la cual se usa una fuente láser de alta intensidad para incidir en una molécula. Esta molécula dispersa la luz incidente. Sin embargo, existen diferentes formas en las que un material dispersa la luz.

Si un haz de luz de un solo color incide y pasa a través de una sustancia, una fracción de la luz cambia de dirección respecto de la que provenía (técnicamente se dice que es dispersada) debido a la interacción con las moléculas de la sustancia. La nube de electrones que envuelve al núcleo en una molécula puede ser polarizada (y de esta forma deformada) de muy diversas maneras por un campo eléctrico. Si a una molécula se le aplica un campo eléctrico oscilante como el campo eléctrico de un haz de luz, la deformación de la nube de electrones oscilará con la frecuencia del haz de luz incidente. La oscilación de la nube produce lo que se llama como un dipolo oscilante que radia a la misma frecuencia de la luz incidente. Este proceso es llamado dispersión Rayleigh [9].

En realidad, en las moléculas no solo se realizan transiciones electrónicas (“saltos” de los electrones entre diferentes órbitas), sino que las moléculas vibran y rotan también. En esos movimientos también se realizan transiciones energéticas. Hay una pequeña probabilidad de que la radiación incidente sobre una sustancia transfiera parte de su energía a uno de los niveles energéticos de vibración o rotación de las moléculas que la conforman. Como resultado de ello la radiación dispersada tendrá una frecuencia menor comparada con la frecuencia con la que vibran las moléculas. De manera similar, hay

una pequeña probabilidad de que la molécula en algún estado energético excitado vibracional o rotacional aumente su energía luminosa, en este caso la radiación dispersada tendrá una frecuencia más alta que la del haz incidente. En ambos casos, la luz se dispersa en diversas longitudes de onda dependiendo de la estructura de la molécula o el analito. Cuando esto ocurre, se dice que ha ocurrido un intercambio de energía y a este fenómeno se le conoce como dispersión Raman.

De esta manera, con la ayuda de la ilustración 1, observamos tres líneas de la radiación dispersada: una línea correspondiente a la dispersión Rayleigh de la misma longitud de onda (amarilla) y dos líneas Raman, una de frecuencia mayor (azul) llamada línea anti-Stokes y la otra de frecuencia menor (roja) llamada línea de Stokes. Las dos líneas Raman son extremadamente débiles comparadas con la intensidad Rayleigh de la luz dispersada. Solamente $1/10000$ de la intensidad dispersada corresponde a las líneas Raman. Es por esto que la intensidad Raman es mucho menor que la intensidad incidente de la luz y, por lo tanto, más difícil de detectar [10].

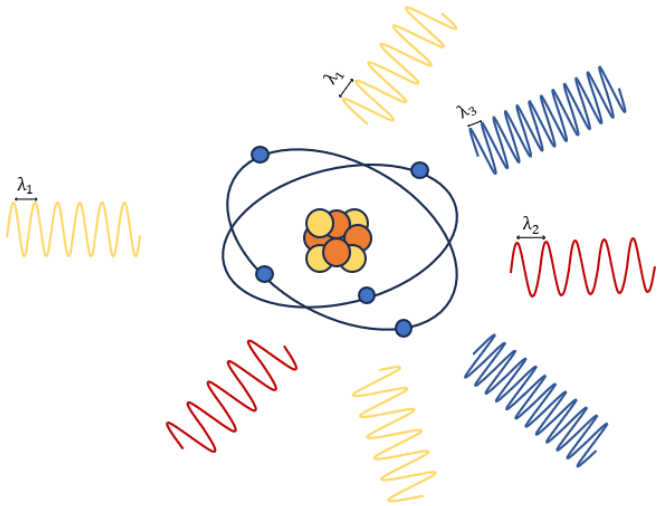


Ilustración 1: Imagen ilustrativa de como interacciona un haz de luz con una molécula. Las líneas amarillas, azules y rojas representan la línea Rayleigh, anti-Stokes y de Stokes, respectivamente.

Un espectro Raman exhibe múltiples picos que indican la intensidad y posición de la longitud de onda de la luz dispersada. Cada pico corresponde a una vibración específica de enlace molecular, incluyendo enlaces individuales. Véase el ejemplo de la ilustración 2:

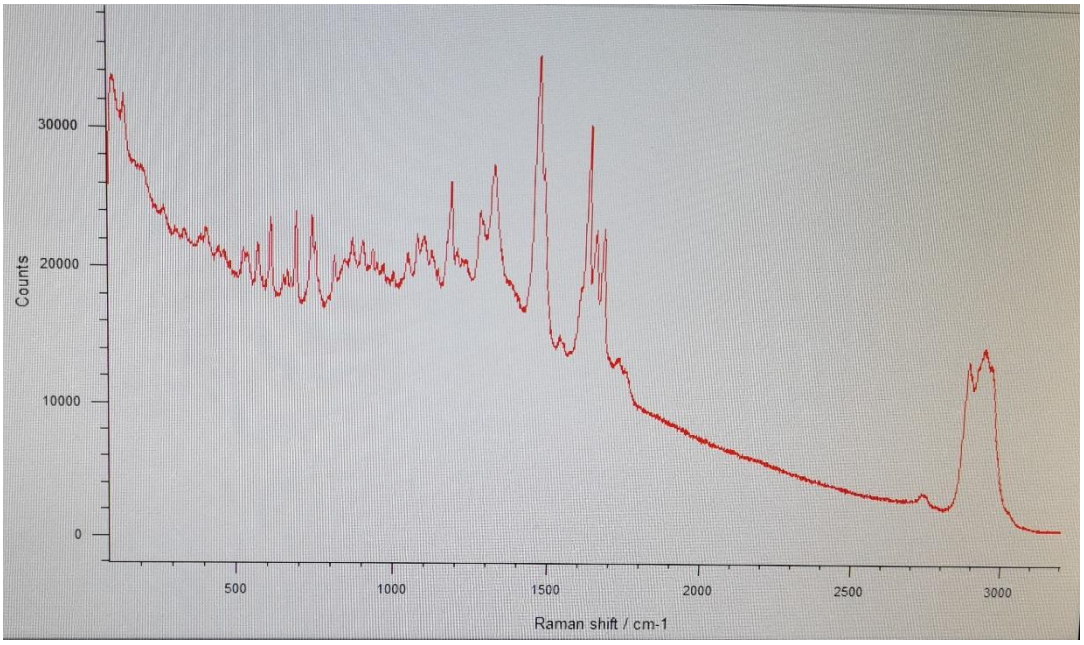


Ilustración 2: imagen de un espectro Raman tomada del software usado para la captura de datos con el microscopio Raman.

Como se puede observar, el eje de abscisas representa la unidad inversa a la longitud de onda (número de onda) de la dispersión Raman recibida, en unidades de cm^{-1}. El eje de ordenadas representa la intensidad de la radiación recibida. Hay que tener en cuenta que en el gráfico aparece un offset debido a la fluorescencia del material irradiado que más adelante hay que eliminar mediante un algoritmo en el preprocesado de los datos. Si la fluorescencia es demasiado alta, puede provocar que la información de la dispersión Raman sea completamente camuflada debido a su baja intensidad. Incluso puede llegar a saturarlo, lo que provoca que los datos sean ilegibles. Este tema se abordará y se explicará con más detalle en próximos apartados.

2.3. Funcionamiento

El comportamiento de la dispersión Raman se puede interpretar de forma diferente a nivel macroscópico y microscópico [11].

A nivel macroscópico, se entiende como la energía emitida por un dipolo oscilante producido por un campo electromagnético. Toda luz monocromática que se desplace en un eje *z* tendrá un campo eléctrico que oscila en el eje *x*. El valor de ese campo eléctrico (E_x) en cualquier instante de tiempo (t) se puede expresar de la siguiente manera:

$$E_x = E_x^0 \cos(2\pi v_0 t), \quad (2.1)$$

Donde E_x^0 es la amplitud máxima del campo eléctrico y v_0 es la frecuencia de la luz monocromática. El campo eléctrico de esta luz, es el causante de la polarización de la nube de electrones antes mencionada. Como se ha explicado en el anterior apartado, el desplazamiento de esos electrones oscilando junto con el campo eléctrico creará un dipolo eléctrico oscilante que emitirá luz dispersa a su vez. Por lo tanto, si la luz dispersada depende del dipolo oscilante eléctrico, las variaciones en la polarización del dipolo afectarán la frecuencia de la luz emitida. La polaridad del dipolo la determina el momento de su dipolo (μ).

$$\mu = \alpha E_x = \alpha E_x^0 \cos(2\pi v_0 t), \tag{2.2}$$

Donde α es la polaridad tensora y describe el comportamiento de la nube de electrones pertenecientes a un campo eléctrico externo. Así mismo, cuando una molécula oscila en su frecuencia natural (v_m), el desplazamiento (q) del núcleo de la molécula se describe de la siguiente forma:

$$q = q_0 \cos(2\pi v_m t), \tag{2.3}$$

Donde q_0 es la amplitud de la vibración. Para amplitudes de vibración pequeñas, el valor de α se puede expresar en función de q mediante la serie de Taylor:

$$\alpha = \alpha_0 + \left(\frac{\partial \alpha}{\partial q}\right)_0 q + \cdots. \tag{2.4}$$

Si combinamos la fórmula 2.2 y 2.4, podemos explicar el momento del dipolo de la siguiente manera:

$$\mu = \alpha_0 E_x^0 \cos(2\pi v_0 t) + \left(\frac{\partial \alpha}{\partial q}\right)_0 q E_x^0 \cos(2\pi v_0 t)\,.$$

Agregando la 2.3 obtenemos la siguiente expresión:

$$= \alpha_0 E_x^0 \cos(2\pi v_0 t) + \left(\frac{\partial \alpha}{\partial q}\right)_0 q_0 \cos(2\pi v_m t)\, E_x^0 \cos(2\pi v_0 t) \tag{2.5}$$

$$= \alpha_0 E_x^0 \cos(2\pi v_0 t) + \frac{1}{2}\left(\frac{\partial \alpha}{\partial q}\right)_0 q_0 E_x^0 \{\cos[2\pi(v_0 + v_m)t] + \cos[2\pi(v_0 - v_m)t]\}\,.$$

De esta última expresión, se deduce que el primer elemento hace referencia al dipolo que oscila a la misma frecuencia que el campo electromagnético involucrado en esta interacción. Este irradia en la frecuencia v_0 por lo que le corresponde a la dispersión *Rayleigh*. Por otra parte, el segundo y tercer elemento irradian en las frecuencias $(v_0 + v_m)$ y $(v_0 - v_m)$. Como estas son diferentes frecuencias a las del haz incidente, corresponden respectivamente a las líneas anti-Stokes y Stokes de la dispersión Raman, siguiendo con la explicación del apartado anterior.

Desde un punto de vista microscópico, en cambio, la dispersión Raman se puede analizar como transferencia de energía. En este caso, la luz se estudia en forma de materia centrándose en los propios fotones. La energía (E) del fotón es proporcional a la frecuencia de la luz:

$$E = hv \,,$$

2.6)

Donde h es la constante de Planck. Cuando un cúmulo de fotones inciden en la materia y soportan dispersión, las moléculas que forman la materia son excitadas a una situación de mayor energía. La cantidad de energía de esta excitación es igual a la energía de los fotones incidentes. Estas moléculas, vuelven a su estado anterior inmediatamente, emitiendo un fotón en el proceso. La energía de este fotón es la diferencia de energía que hay entre las dos situaciones energéticas. Por lo tanto, si la situación energética no ha cambiado, la energía del fotón irradiado será idéntica a la del fotón incidente. Cuando ocurre esto, se dice que se ha producido dispersión Rayleigh. Cuando ocurre lo contrario, el fotón irradiado tendrá mayor o menor energía, teniendo así una frecuencia diferente a la del fotón incidente. Este es el caso conocido como dispersión Raman.

En el caso de la línea de Stokes, la molécula ha absorbido energía, por lo que el fotón resultante es de inferior frecuencia y se genera una línea de Stokes en el lado rojo del espectro incidente. En el caso de la línea anti-Stokes, en cambio, la molécula pierde energía por lo que los fotones incidentes son desplazados a frecuencias más elevadas (azul) del espectro.

Sin embargo, con este último enfoque hay que tener en cuenta una característica importante, ya que, la frecuencia de la luz resultante de la dispersión Raman depende de la situación de las moléculas de la materia. Por lo que cualquier fenómeno que afecte a la situación de la materia, como la tensión o la temperatura, puede dar lugar a un cambio en la frecuencia de la dispersión Raman.

2.4. Equipamiento

Un espectrómetro Raman consta de varios componentes básicos. Uno de ellos, el láser que sirve como fuente de excitación, se usa normalmente en instrumentos modernos Raman con longitudes de onda de 535 nm, 785 nm, 830 nm y 1064 nm. Los láseres de longitud de onda más corta tienen secciones transversales de dispersión Raman más altas, por lo que la señal resultante es mayor. Sin embargo, la incidencia de fluorescencia también aumenta a una longitud de onda más corta. La energía del láser se transmite a la muestra, se dispersa en y es recogida de nuevo por el equipo. Se utiliza un filtro de hendidura o de borde para eliminar la dispersión de Rayleigh y anti-Stokes. La luz dispersa restante de la línea de Stokes, es transmitida a un elemento de dispersión, normalmente una rejilla holográfica. Finalmente, un detector CCD captura la luz, lo que da lugar al espectro Raman. Dado que la dispersión Raman produce una señal débil, es muy importante que en el espectrómetro Raman se usen componentes de alta calidad y ópticamente bien adaptados [12].

2.5. Análisis Multivariante (MVA)

El análisis multivariante se refiere a la utilización de técnicas estadísticas para analizar conjuntos de datos que incluyen más de una variable. Hoy en día, este tipo de análisis es de gran interés debido a que muchas de las investigaciones y aplicaciones que se realizan en diferentes campos de estudio dan lugar a bases de datos estadísticos que corresponden a muchas variables. La información que es posible obtener de estas bases de datos es más rica cuando se considera la extracción de patrones que pueden existir en conjunto entre ciertos valores de las variables.

Con métodos analíticos multivariantes adecuados, como pueden ser *Principal Component analysis* (PCA) y *Partial Least Squares regression* (PLS regression), las masas de datos de proceso pueden proporcionar información gráfica fácil de comprender sobre el estado del proceso y las relaciones entre conjuntos importantes de variables de proceso. Estos métodos multivariantes hacen un uso eficiente de todos los datos pertinentes, con poca pérdida de información.

2.5.1. Variables de proceso

Los datos medidos en un proceso suelen almacenarse en algún tipo de base de datos. Una base de datos de procesos que contiene los valores de K variables para N puntos de datos puede considerarse una tabla, o una matriz. A esta tabla se la denomina X. Cada columna de la tabla corresponde a una variable (x_k), y una fila (x_i) corresponde a los valores observados en un momento de tiempo o, en el caso de este proyecto, una medida o una muestra diferente.

2.5.2. PCA - Principal Component Analysis

El Análisis de Componentes Principales, o PCA, es un procedimiento estadístico que permite resumir el contenido informativo de grandes tablas de datos mediante un conjunto más pequeño de "índices resumen" que pueden visualizarse y analizarse más fácilmente. Los datos subyacentes pueden ser mediciones que describan propiedades de muestras de producción, compuestos o reacciones químicas, puntos de tiempo de un proceso continuo, lotes de un proceso discontinuo, individuos biológicos o ensayos, por ejemplo.

El Análisis de Componentes Principales es considerada una de las técnicas estadísticas multivariantes más populares hoy en día. Ha sido ampliamente utilizada en áreas de reconocimiento de patrones y procesamiento de señales.

El PCA constituye la base del análisis de datos multivariantes basado en métodos de proyección. Los índices resumen generados por el PCA son denominados componentes principales y se crean con el fin de observar tendencias, saltos, conglomerados y valores atípicos. Esta visión de conjunto puede descubrir las relaciones entre observaciones y variables, y entre las variables. También permite analizar conjuntos de datos que pueden contener, por ejemplo, multicolinealidad, valores perdidos, datos categóricos y mediciones imprecisas [13].

2.5.3. Funcionamiento

Para la matriz de datos, se construye un espacio de variables con tantas dimensiones como variables. Cada variable representa un eje de coordenadas. Para cada variable, la longitud se ha normalizado según un criterio de escalado, normalmente escalando a la varianza unitaria. Cada observación (fila) de la matriz X se coloca en el espacio de variables de K dimensiones. En consecuencia, las filas de la tabla de datos forman un enjambre de puntos en este espacio, tal y como se puede ver en la Ilustración 3. Una vez colocados, se hace un centrado de la media. Para eso, se sustraen de los datos las medias de las variables. El vector de medias corresponde a un punto del espacio K. Con ese punto en mente, se hace un reposicionamiento del sistema de coordenadas trasladando todos los puntos de ese espacio junto con el punto que representa la media, de modo que el punto medio es ahora el origen [13] .

Tras centrar la media y ajustar la escala a la varianza unitaria, el conjunto de datos está listo para el cálculo del primer índice resumen, el primer componente principal (PC1). Este componente es la línea en el espacio variable de K dimensiones que mejor se aproxima a los datos en el sentido de mínimos cuadrados. Esta línea pasa por el punto medio. Cada observación (punto amarillo) puede proyectarse ahora sobre esta línea para obtener un valor de coordenadas a lo largo de la línea PC. Este nuevo valor de coordenadas también se conoce como *score*.

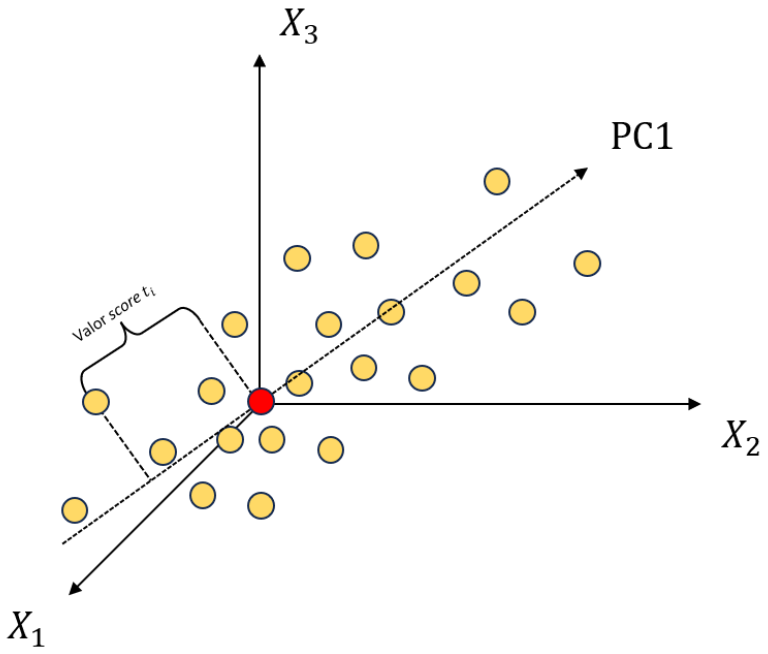


Ilustración 3: ejemplo gráfico de la formación del primer componente principal (PC1).

Normalmente, un índice resumen o componente principal es insuficiente para modelizar la variación sistemática de un conjunto de datos. Por lo tanto, se suele calcular un segundo componente principal (PC2) el cual está representado por una línea en el espacio variable de K dimensiones, que es ortogonal al primer PC. Esta línea también pasa por el punto medio y mejora la aproximación de los datos X en la medida de lo posible.

Estos dos componentes principales juntos definen un plano, una ventana en el espacio de variables de K dimensiones. Al proyectar todas las observaciones en el subespacio de baja dimensión y representar gráficamente los resultados, es posible visualizar la estructura del conjunto de datos investigado (ver Ilustración 4). Los valores de las coordenadas de las observaciones en este plano se denominan *scores*, por lo que la representación gráfica de una configuración proyectada de este tipo se conoce como *score plot*.

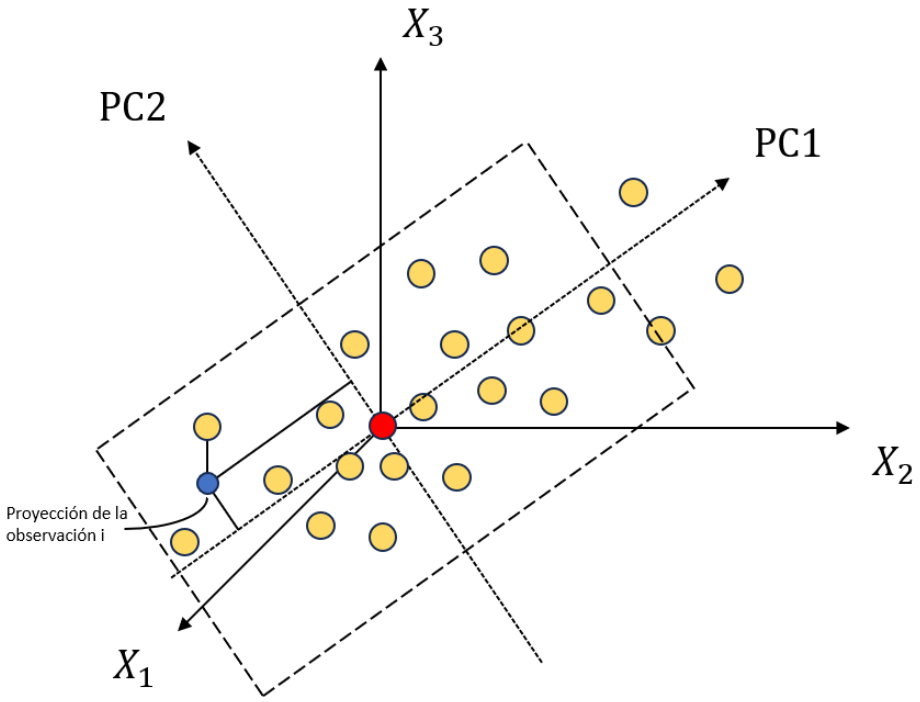


Ilustración 4: proyección del plano formado por el conjunto de PC1 y PC2.

El *score plot* es la visión de conjunto de la cual se podrá descubrir cuáles son las relaciones entre observaciones y variables, y entre las variables. De esta forma, se obtiene una mejor visión del conjunto de datos con el que se está trabajando. Se tendrá un ejemplo más claro del funcionamiento del *score plot* en el apartado donde se detalla la descripción de la solución propuesta.

2.5.4. PLS - Partial Least Squares regression

En pocas palabras, la regresión de Mínimos Cuadrados Parciales, o PLS como se conoce por sus siglas en inglés, se puede considerar una extensión del Análisis de Componentes Principales. Mientras que el PCA se utiliza para buscar un resumen o visión general de los datos, mediante PLS se trata de encontrar hiperplanos de máxima varianza entre la variable de respuesta y las variables independientes. Es por esto que PCA se utiliza mejor en casos en los que se busca un resumen o visión general de los datos [14].

La regresión de Mínimos Cuadrados Parciales se centra en encontrar una regresión lineal mediante la proyección de las variables de predicción y las variables observables a un nuevo espacio. Es una técnica

que reduce las variables a un conjunto más pequeño de componentes no correlacionados y realiza una regresión de mínimos cuadrados sobre estos componentes, en lugar de hacerlo sobre los datos originales. La regresión PLS resulta especialmente útil cuando las variables son muy colineales o cuando se tienen más variables que observaciones y la regresión de mínimos cuadrados ordinarios produce coeficientes con altos errores estándar o falla por completo. PLS no presupone que las variables sean fijas, a diferencia de la regresión múltiple. Esto significa que las variables pueden medirse con error, lo que hace que PLS sea más robusta a la incertidumbre de las mediciones.

Por lo tanto, cuando se necesite analizar y modelizar un conjunto de datos muy amplio (es decir, un conjunto de datos con muchas más variables que observaciones), es probable que se plantee utilizar un método como PLS u OPLS. El OPLS es una variante del PLS, pero mientras que ambos métodos dividen la variabilidad de un conjunto de datos en lo sistemático (estructurado) y residual (ruido), el OPLS divide, además, la variabilidad de lo sistemático (la tabla de datos X) en dos componentes: predictivo (todo lo correlacionado con la respuesta, Y) y ortogonal (todo lo no correlacionado con la respuesta). Esto mejora la interpretabilidad del modelo. En el caso de una sola variable Y, sólo hay un componente predictivo, y todos los componentes más allá del primero reflejan una variación ortogonal. Sin embargo, con múltiples variables Y puede haber más de un componente OPLS predictivo [14].

En la situación de una única variable Y, un modelo PLS y un modelo OPLS ajustados a los mismos datos tendrán el mismo poder predictivo (siempre que se comparen modelos con el mismo número total de componentes). La gran ventaja del OPLS sobre el PLS reside en la interpretabilidad simplificada del modelo que surge con el OPLS, debido a la capacidad de separar la varianza explicada en compartimentos predictivos y ortogonales del modelo. Por lo tanto, OPLS ofrece una mejor visualización cuando hay una gran cantidad de estructura ortogonal Y en X y puede ayudar a aclarar y comprender la variación correlacionada y no correlacionada [14].

2.5.5. PLS-DA/OPLS-DA

PLS-DA/OPLS-DA son métodos basados en los métodos PLS/OPLS, como su nombre indica. DA significa Análisis Discriminante, o *Discriminant Analysis* por sus siglas en inglés.

Con modelos anteriores, conceptualmente lo que se obtiene son modelos de clase para diferentes tipos de muestras y se define una envolvente alrededor de los puntos de datos - un tamaño del modelo o borde del modelo.

Luego, en la clasificación, lo que interesa es saber si la nueva muestra que aparece en el conjunto de predicciones es similar a una u otra clase, o si no encaja en ninguna clase en absoluto. Como el objetivo principal es definir los límites de las clases e inferir la pertenencia a una clase de las futuras muestras del conjunto de predicción, no se centra tanto en por qué las clases son diferentes.

Sin embargo, con el análisis discriminante lo que se plantea es la siguiente pregunta: ¿Cuál es la diferencia? Aquí en lo que se centra es en las variables. ¿Qué variables impulsan la separación entre los dos grupos? En el caso de tener un problema de dos grupos, el modelo OPLS-discriminante

resultante será muy fácil de interpretar porque sólo tendrá un componente predictivo que interpretar. Este componente se representa como el eje de abscisas en el gráfico de dispersión de *scores* resultante del modelo OPLS-DA [10].

De esa manera, la dirección horizontal del gráfico de dispersión de *scores* captará la variación entre los grupos. ¿Cuáles son las diferencias sistemáticas entre el grupo de la izquierda y el grupo de la derecha? Mientras, la dimensión vertical y cualquier componente superior de los denominados de tipo ortogonal captarán la variación dentro de los grupos.

3. Objetivos y alcance del trabajo

El proyecto se centra en la obtención de datos de varios lotes de olivas y hojas de olivos, los cuales estarán diferenciados por tipo de fitosanitario y nivel de concentración aplicado en cada uno. El objetivo del proyecto no es otro que determinar si se es capaz de clasificar y diferenciar entre sí las distintas muestras en base a sus datos obtenidos mediante el desarrollo de un modelo predictivo desarrollado en base a las herramientas de Análisis Multivariante de Datos (MVDA). Para la obtención de los datos de las muestras se ha recurrido a la tecnología de espectroscopía Raman, mientras que para el desarrollo del modelo se ha hecho uso del software SIMCA, el cual permite el uso de una gran variedad de herramientas MVDA para la clasificación de los datos y la predicción usando datos nuevos.

Para cumplir con el objetivo principal, el proyecto se ha dividido en objetivos secundarios más pequeños que permitan un mejor seguimiento del mismo:

1. Mediar con el socio encargado de facilitar las muestras de las olivas y hojas de los olivos para llegar a un acuerdo en cuanto a los diferentes grupos de muestras a enviar. Estas son separadas según el fitosanitario utilizado para rociar las olivas y la concentración utilizada.
2. Tratar de obtener resultados óptimos haciendo varias mediciones con cada muestra y descartando las que se salgan de la norma dentro de ese grupo de mediciones.
3. Elección del mejor método de análisis multivariante de datos, teniendo en cuenta el tipo de datos a tratar y el objetivo buscado. En este caso, el tipo de datos a tratar consta de menos observaciones que variables.
4. Desarrollar un modelo robusto y validar su eficacia mediante la predicción de nuevos datos.

4. Beneficios

En este apartado se valorarán los beneficios que puede aportar el desarrollo de este proyecto. En este caso, el proyecto aportará beneficios en diferentes ámbitos, como son:

4.1. Beneficio tecnológico

Con el desarrollo de este proyecto se incrementa el número de estudios que usan la tecnología Raman como pilar importante en su desarrollo, indagando así más en los usos y aportaciones que está tecnología tiene en diferentes ámbitos de estudio. De esta manera, futuros proyectos podrán disponer de referencias de más trabajos, incluso teniendo la posibilidad de que futuras investigaciones se centren en mejorar dicha tecnología para un mejor uso de esta en proyectos enfocados en el mismo ámbito de este proyecto.

4.2. Beneficios en la salud

Este proyecto está enfocado en la mejora de las capacidades deterministas que puede tener la industria alimentaria en cuanto a determinar si los productos que estén por encima o por debajo de cierto umbral permitido son desechados o no. En este caso, el proyecto está dirigido a la industria de la oliva, la cual tiene ciertos estándares que se deben superar para que las olivas sean consideren aptas para la salud y el consumo humano, aun habiendo usado fitosanitarios en ellas con el fin de proteger la producción.

Con el desarrollo de este proyecto, aparece una nueva forma de detectar, con un porcentaje muy aceptable de acierto, si el producto supera un umbral determinado. Esta tecnología, a pesar de no ser tan preciso a la hora de determinar la concentración del fitosanitario como lo es el análisis químico, es capaz de determinar con bastante exactitud si el producto es potencialmente peligroso, lo que hace que se puedan tomar medidas para desecharlo.

Así mismo, su punto fuerte es la rapidez que aporta a la hora de obtener resultados frente a otro tipo de análisis. Además, también existen equipos portables y más baratos que, a pesar de tener menos potencia podrían llegar a ser una alternativa viable. Las productoras podrían valerse de esta tecnología para determinar que productos de un lote son potencialmente peligrosos sin la necesidad de tener que descartar todo un lote como hasta ahora, cuando era una única muestra la que era analizada en el laboratorio y obtenía resultados negativos.

5. Análisis de alternativas

5.1. Software

A la hora de elegir cualquier software, es importante tener en cuenta varios aspectos, como pueden ser la facilidad en su manejo (si son intuitivos), si requiere conocimientos técnicos previos para su uso o si el software necesita de licencia para su uso y si estas son muy caras.

En cuanto al software de análisis multivariante, también se recomienda probarlo primero antes de tomar decisiones. Esto es debido a que el software de análisis multivariante es un campo que avanza año tras año, por lo que existe una gran diferencia entre los distintos softwares.

A la hora de elegir un Software de análisis multivariante, dos características importantes a tener en cuenta son, si es capaz de hacer lo que se quiere conseguir y si es fácil de usar. Además, como el software se ha vuelto más complejo en los últimos años, cada vez más proveedores ofrecen servicios de mantenimiento de pago.

Además del mantenimiento, el paquete suele incluir la resolución de problemas operativos y la formación de los usuarios. En este caso, es aconsejable determinar si se requiere asistencia técnica en función de la sensación de la prueba.

5.2. Paquetes de Software de Análisis Multivariante

Los paquetes de software de análisis multivariante no requieren que el usuario conozca los métodos de cálculo específicos. En los últimos años se han publicado muchos códigos de programación para utilizar el software del que hacen uso ciertos programas, pero la ventaja del software de paquetes es que el análisis puede realizarse sin conocimientos de programación.

Sin embargo, incluso cuando se utiliza software empaquetado, es al menos necesario tener conocimientos sobre la estructura de los datos, el significado del análisis y la selección del método de análisis que mejor se adapte al propósito. Algunos proveedores ofrecen formación a los usuarios en estas áreas.

5.2.1. Unscrambler

Unscrambler destaca como una de las principales herramientas de análisis en la industria, ofreciendo capacidades líderes para modelar, predecir y optimizar mediante el uso de potentes análisis, gráficos interactivos y visualizaciones.

En la actualidad, se posiciona como una de las herramientas analíticas más completas, fáciles de usar y poderosas en el mercado empresarial. Sus opciones de capacitación flexibles se adaptan a diversas preferencias de aprendizaje, niveles de habilidades y roles de usuarios, ofreciendo soluciones personalizadas para abordar los desafíos específicos de cada negocio. Incorporando las más recientes técnicas de aprendizaje automático y análisis multivariado, eleva la calidad de procesos y productos.

Está diseñado para resolver problemas complejos utilizando un poderoso análisis multivariado, con capacidades únicas para espectroscopia y quimiometría. Da opción a elegir entre 20 métodos diferentes para analizar datos, incluido el Diseño de experimentos (DoE), el análisis exploratorio de datos, la Regresión de mínimos cuadrados parciales (PLSR), el Análisis de componentes principales (PCA) y el Modelo independiente suave de analogías de clase (SIMCA). Permite explorar y validar fácilmente modelos mediante gráficos interactivos y visualizaciones para optimizar el desarrollo del producto, mejorar la calidad del producto y la eficiencia del proceso. Algunas de las ventajas que ofrece son [15]:

- Mejor comprensión y optimización de los procesos a través del análisis de todo tipo de datos
- Mejor clasificación de productos para entregar una calidad de producto consistente
- Enfoque rentable para el desarrollo y la optimización de productos con diseño de experimentos
- Optimizar el desarrollo de productos y los procesos de fabricación para mejorar la calidad y la eficiencia
- Fácil importación de todo tipo de datos, como material, sensor, proceso y datos espectrales de más de 30 formatos diferentes y nuevos formatos que se agregan fácilmente

Entre las ventajas destacables de esta esta tecnología frente al resto, destacan su gran número de opciones disponibles en cuanto al método de análisis de datos a utilizar, su diseño específico para campos como la espectrometría, el cual usa datos de la misma naturaleza que la espectroscopía, y la formación que ofrece. Sin embargo, el uso de este software requiere de licencia para su uso de forma legítima. Aunque hay que destacar que ofrece una prueba gratuita que permite su uso para decidir si es la tecnología correcta para el desarrollo del proyecto.

5.2.2. SIMCA

SIMCA de SARTORIUS es uno de los softwares de análisis multivariante de datos más usados actualmente [16]. Es una solución avanzada de análisis de datos que transfiere los datos a información visual interpretable. Está diseñada con el objetivo de ser una herramienta muy útil en campos que requieren el tratamiento de un gran número de datos, ya sean variables u observaciones, como es el campo de la investigación científica. Esta tecnología, ayuda a visualizar fácilmente tendencias y agrupaciones mediante una interfaz gráfica intuitiva. Con SIMCA se puede analizar las variaciones del proceso, identificar los parámetros críticos y predecir la calidad del producto final.

Combina su potente motor multivariante con visualizaciones interactivas, una interfaz intuitiva y la capacidad de automatizar flujos de trabajo, para ofrecer un software realmente fácil de usar que facilita la carga de trabajo analítico. Algunas de las funcionalidades importantes que ofrece son:

- Revisar, trazar y explorar los datos de forma interactiva para identificar correlaciones importantes.
- Hacer clic en puntos de datos individuales para revelar las contribuciones subyacentes
- Examinar las relaciones entre variables
- Identificar rápidamente los factores e interacciones más importantes
- Implementar secuencias de comandos Python para automatizar los flujos de trabajo
- Investigar y diagnosticar las causas de los problemas
- Predecir el rendimiento, la calidad y el comportamiento futuro

En cuanto a las ventajas de usar esta tecnología, destacan que sea una herramienta muy intuitiva y fácil de usar, su interfaz gráfica interactiva, la cómoda visualización de los datos y resultados que ofrece, la posibilidad de usar *scripts* de ejemplo escritos en *Python*, y la gran cantidad de información y documentación que ofrece desde su página oficial de forma gratuita. Además, también dispone de

forma gratuita en su página oficial de un gran número de videotutoriales centrados en explicar paso a paso como generar los diferentes modelos y cuando se debe usar cada uno, aunque, como punto en contra, hay que destacar que el software y toda la información y documentación en su página web se encuentra en inglés.

Además de las ventajas que ofrece el uso de este software, también es importante destacar que Sartorius, la empresa detrás de SIMCA, ofrece una licencia de prueba gratuita de 30 días que permite probar el software para ver si cumple con lo que necesitas antes de comprar la licencia oficial.

Un punto muy importante a favor del uso de este software ha sido la posesión de licencias por parte de la UPV/EHU, lo que ha permitido un ahorro sustancial en el proyecto. Además, el haber tenido acceso a tanta documentación bien explicada en cuanto a la teoría sobre análisis multivariante de datos y el uso del software y todas las opciones que este ofrece, ha sido determinante a la hora de haber elegido el software como el candidato idóneo para el desarrollo del proyecto.

5.2.3. PLS-Toolbox

PLS_Toolbox es una completa colección de herramientas esenciales y avanzadas para el análisis multivariante, diseñada para su integración con el entorno computacional MATLAB.

La caja de herramientas ofrece más de 300 funciones y una interfaz gráfica unificada, abarcando una amplia gama de áreas técnicas. Su nombre deriva del método de regresión de mínimos cuadrados parciales (PLS), ampliamente reconocido como estándar en diversas aplicaciones de calibración y modelado. Dirigido a ingenieros químicos, químicos analíticos y otros científicos de datos centrados en el análisis, PLS_Toolbox proporciona todas las herramientas necesarias para aprovechar al máximo los datos y construir modelos predictivos.

En la mayoría de los casos, los usuarios prefieren la facilidad de uso mediante la interfaz gráfica para realizar tareas de edición de datos y modelado. Sin embargo, en situaciones donde es necesario incorporar funciones personalizadas o automatizar análisis, PLS_Toolbox permite a los usuarios trabajar en ambas direcciones. Además de interfaces sofisticadas para abordar diversas tareas de modelado, también facilita el acceso a toda la funcionalidad a través de la línea de comandos, gracias a su código orientado a objetos potente y bien documentado [17].

Como desventaja, esta solución requiere de MATLAB para su uso, además de una licencia que ronda los 950€. Por otra parte, La utilización de esta solución requiere más inversión de tiempo en su aprendizaje.

5.2.4. MATLAB

MATLAB es una plataforma de programación y cálculo numérico muy popular en la ingeniería y en la ciencia para analizar datos, desarrollar algoritmos y crear modelos. Combina un entorno de escritorio perfeccionado para el análisis iterativo y los procesos de diseño con un lenguaje de programación que expresa las matemáticas de matrices y *arrays* directamente. Su lenguaje de programación propio se conoce como M.

Como una de las alternativas, existe la posibilidad de desarrollar una interfaz gráfica propia con MATLAB para la creación de modelos de predicción basados en algunos de los métodos de análisis multivariante de datos. Para el desarrollo de dicha GUI, existe la herramienta *App Designer* proporcionada por MATLAB que permite la creación de apps profesionales sin la necesidad de tener conocimiento como desarrollador de software profesional.

Como gran ventaja, destacan que MATLAB está ampliamente diseñado para la manipulación de información de señales electromagnéticas, su sencillez a la hora de crear algoritmos y su página web como guía y soporte técnico con un enorme número de documentación y ejemplos de forma gratuita. Sin embargo, como gran desventaja frente a las otras alternativas está el tiempo que hay que invertir para su creación, además del aprendizaje previo necesario para hacerlo. Por otro lado, otra de sus desventajas podría ser el uso de licencia, ya que la licencia de usuario ronda los 800€. Sin embargo, la UPV/EHU tiene comprada la licencia “Total Academic Headcount Full Suite” por lo que cualquier estudiante perteneciente a dicha institución puede usar el software libremente, además de poder usar también los *Toolbox* proporcionados por MATLAB sin la necesidad de tener que comprar licencias aparte.

6. Descripción de la solución propuesta

6.1. Equipamiento empleado

Para el desarrollo de una solución adecuada, en las instalaciones donde se ha llevado a cabo el proyecto se ha hecho uso del microscopio Raman de la marca *Renishaw* y modelo *InVia*, como se puede apreciar en la Ilustración 5.

Ilustración 5: fotografía del microscopio Raman Renishaw InVia.

En cuanto a la fuente del láser, dispone de dos fuentes de luz a elegir. Uno irradia con una longitud de onda de 532 nm y de hasta 55 mW de potencia óptica. El otro, con una longitud de onda de 785 nm, es capaz de irradiar con una potencia óptica de hasta 170 mW. Además, estas potencias pueden ser reducidas usando uno de los 3 filtros integrados (de tipo $optical\ density$) de los que dispone el sistema. Por otra parte, el microscopio también dispone de dos rejillas de difracción, cada una ajustada para cada laser. En este caso, para el láser de 532 nm se usa una rejilla de 1800 $lineas/nm$ y, para el de 785 nm, una de 1200 $lineas/nm$ [18].

En cuanto a los objetivos disponibles, el microscopio puede trabajar con 4 objetivos instalados en el revólver del microscopio. Estos son de $5X$, $20X$, $50X$ y $100X$, teniendo estas una apertura numérica de 0.12, 0.4, 0.75 y 0.85, respectivamente. Se puede apreciar mejor el revólver del microscopio en la Ilustración 6.

Ilustración 6: Revólver del microscopio.

Además, como se puede apreciar en la imagen anterior, el microscopio también dispone de una plataforma motorizada que permite ajustar la muestra siguiendo los ejes XYZ, permitiendo una mejor calibración de la muestra y la zona donde el láser debe incidir.

6.2. Obtención de los datos

Para la obtención de los datos, el material usado como muestra han sido un conjunto de olivas separadas en varios lotes definidos por el fitosanitario y la concentración empleada en ellas. Las muestras fueron rociadas con los siguientes fitosanitarios: Deltametrina, Diflufenican, Lambda Cihalometrina, Oxifluorfen y Tebuconazol. Estas muestras fueron clasificadas por lotes de la siguiente manera (tabla 1).

Nombre del lote	Concentración (mg/Kg)
Deltametrina 1	37,2
Deltametrina 2	19,1
Deltametrina 3	0,77
Deltametrina 5	Muy baja (por debajo de 0.0010)
Diflufenican 1	31,8
Diflufenican 2	3,42
Diflufenican 3	0,46
Diflufenican 4	0,015
Diflufenican 5	Muy baja (por debajo de 0.0010)
Lambda Cihalometrina 1	84,8
Lambda Cihalometrina 2	11,4
Lambda Cihalometrina 3	0,95
Lambda Cihalometrina 4	0,014
Lambda Cihalometrina 5	Muy baja (por debajo de 0.0010)
Oxifluorfen 1	72,3
Oxifluorfen 2	11,5
Oxifluorfen 3	1,12
Oxifluorfen 5	Muy baja (por debajo de 0.0010)
Tebuconazol 1	116
Tebuconazol 2	19,2
Tebuconazol 3	1,29
Tebuconazol 4	Muy baja (por debajo de 0.0010)
Tebuconazol 5	Muy baja (por debajo de 0.0010)
Sin Tratar	0

Tabla 1: lotes y su concentración.

La Ilustración 7 y la Ilustración 8 sirven como ejemplos de lote de olivas recibidos por parte del distribuidor.

Ilustración 7: Lote de olivas Diflufenican 4.

Ilustración 8: Lote de olivas Diflufenican 5.

Como se puede observar en las imágenes anteriores, se puede apreciar que dentro de cada lote se encuentran olivas de diferentes características. Esto es debido a que algunas se han conservado mejor que otras desde que empezó el proceso de recolección hasta llegar a nuestras manos y es un indicativo de que el tiempo es crucial para hacer las mediciones, ya que la composición del conjunto de muestras podría verse afectada. Sin ir más lejos, había casos en los que el lote ya había llegado con alguna aceituna podrida habiendo sido afectada por el moho, como se puede apreciar en la Ilustración 9.

Ilustración 9: ejemplo de oliva podrida.

En vista de estos casos, se decidió revisar todos los lotes como primer paso para deshacerse de las olivas más afectadas como medida preventiva para que estas no acelerasen el deterioro del resto de muestras.

Teniendo en cuenta lo anterior, se decidió establecer una ventana de tiempo para el proceso de obtención de datos de una semana para todos los lotes. Debido a esto, se calculó que teniendo en cuenta el tiempo limitado y el número de muestras por lote, hacer tres medidas por muestra a tres muestras diferentes por cada lote sería suficiente para tener un conjunto de datos aceptable. Habiendo recibido un conjunto de 5 lotes de diferente concentración por cada uno de los 5 grupos diferentes de fitosanitarios más un único lote de aceitunas sin tratar, se poseían un total de 26 lotes. Contando con que se estableció la norma de hacer 3 mediciones por cada muestra y 3 muestras diferentes por cada lote, se obtiene un total de 234 mediciones a realizar en una semana. Este proceso requirió de 4 horas diarias durante esos 5 días plenamente enfocadas en la obtención de los datos.

Antes de empezar con el proceso de medición se establecieron algunas medidas de seguridad, como es el uso de guantes para la manipulación de las muestras. Esto es debido al fitosanitario con el que las olivas habían sido rociadas, pudiendo ser perjudicial para la salud. Para la medición, las muestras son colocadas en un recipiente rectangular transparente y se procede a enfocar la zona a irradiar. Para ello, se puede hacer uso de cada uno de los objetivos, enfocando con cada uno hasta lograr a un buen enfoque con el objetivo de deseado. En este caso, las mediciones se hicieron con un objetivo de $50X$. Se puede apreciar como la muestra es colocada y enfocada en la Ilustración 10.

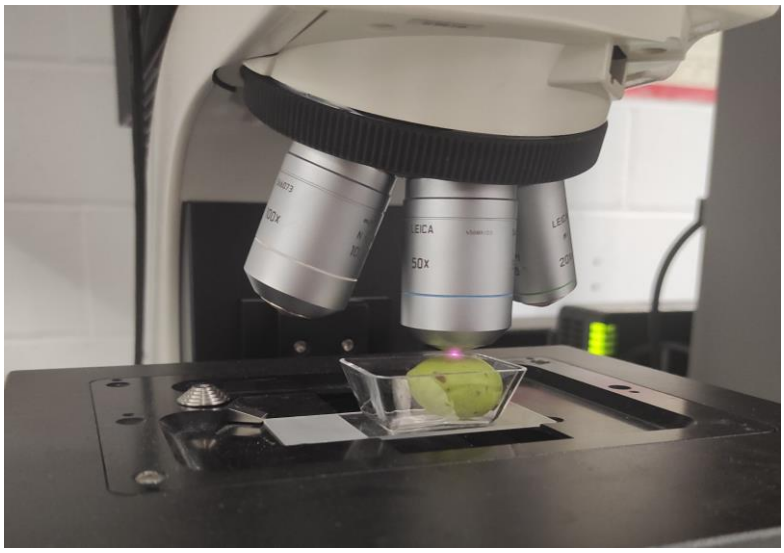


Ilustración 10: imagen de una muestra bajo el objetivo.

Las mediciones fueron efectuadas con las siguientes características:

- Laser de 785 nm de potencia.
- Rejilla de 1200 $lineas/mm$.
- 100% de intensidad del láser.
- Objetivo de $50X$.
- Tiempo de exposición $10s$.

En cuanto al número de medidas, en alguna ocasión excepcional se obtuvo alguna medida fuera de lo común por lo que se decidió repetir la medición en una zona diferente de la muestra, obteniendo resultados dentro de la norma de nuevo. Esto ocurría en casos muy aislados y se debía principalmente al hecho de haber enfocado directamente a una "piedra" de fitosanitario en la muestra. Estas "piedras" son una acumulación relativamente grande de fitosanitario cristalizado en la superficie de la muestra. A continuación, desde la Ilustración 11 hasta la Ilustración 14 se puede observar la diferencia en las lecturas cuando se efectúa una medida en una zona normal de la muestra frente a una hecha en una "piedra".

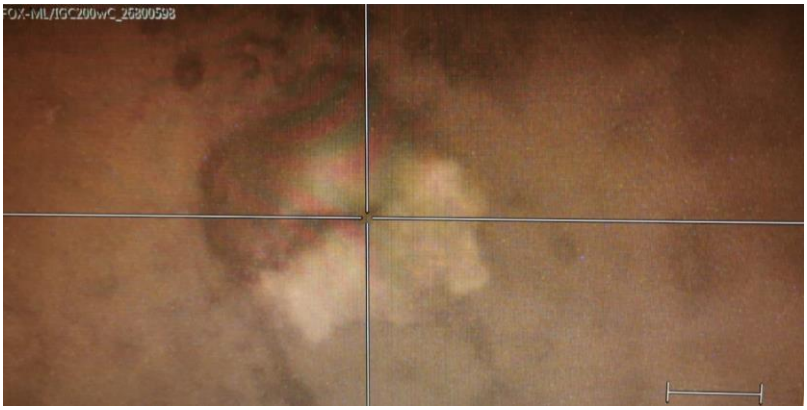


Ilustración 11: imagen del microscopio enfocando una "piedra" en la superficie de la muestra.

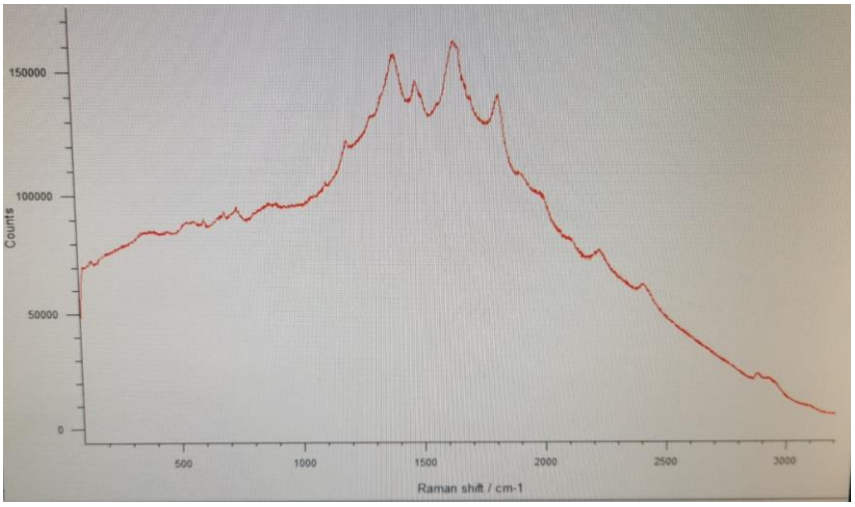


Ilustración 12: imagen del espectro obtenido como resultado de enfocar una "piedra".

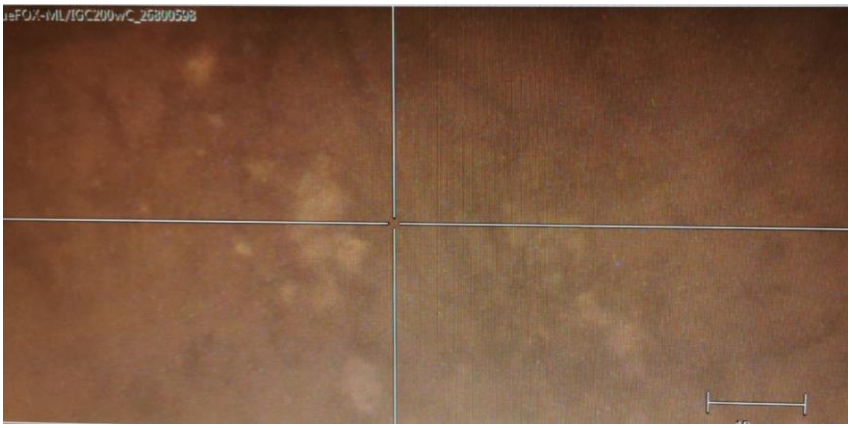


Ilustración 13: imagen del microscopio enfocando la superficie de la muestra aleatoriamente.

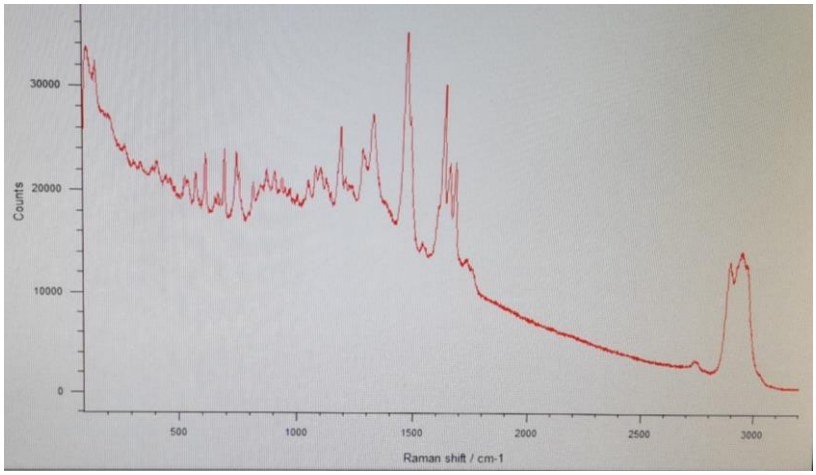


Ilustración 14: imagen del espectro obtenido como resultado de enfocar la superficie de la muestra.

Así mismo, se tomó la decisión de excluir los resultados que se salían de la norma en cuanto a la forma del espectro obtenido.

6.3. Adaptación de los datos

En esta parte del proceso es donde se tratan los datos obtenidos y se adaptan de una forma óptima para el desarrollo del modelo con el software elegido. Para ello, lo primero es eliminar el offset que aparece en el espectro obtenido por el microscopio Raman debido a la fluorescencia reflejada por la muestra. Para su eliminación se hace uso del algoritmo *4S Peak Filling* que estima la línea de base mediante supresión iterativa de la media [19] . Tras el uso de dicho algoritmo, el espectro queda de la siguiente forma (ilustración15). En comparación a como estaba antes (ilustración 14), ahora se puede obtener mejor la información del espectro, sin la interferencia de la fluorescencia.

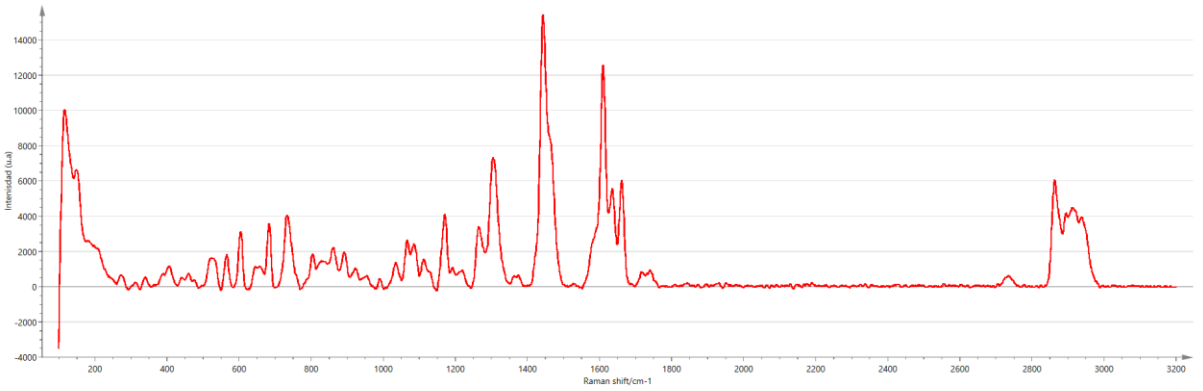


Ilustración 15: espectro suavizado de la muestra Tebuconazol 5.

Una vez tratados los datos, lo siguiente es ordenarlos de la forma en la que el programa requiera, adecuándolos a un formato que pueda interpretar de forma correcta. En este caso, se hizo uso del software Excel para conseguir el formato adecuado de los datos. De esta forma, los datos se colocan de manera en la que la primera columna ocupa el nombre dado a cada observación (nombre de fitosanitario presente en la muestra y una numeración referente al lote y número de muestra) con el título de *Primary ID*, la longitud de onda ocupa la primera fila a partir del título de las observaciones y el resto de los datos son establecidos ocupando su respectivo lugar según la observación y la longitud de onda a la que corresponden.

6.4. Preprocesado de los datos

Para este proyecto, se ha decidido el uso del software SIMCA como herramienta principal para el tratamiento de los datos. Este software permite importar los datos de un fichero directamente como puede ser una hoja de Excel en este caso. También permite la importación de datos desde una base de datos.

Una vez importada, es importante establecer el significado de cada uno de los datos de forma correcta para que SIMCA pueda interpretarlos de forma correcta tal y como deseamos. En este caso, detecta automáticamente la columna que hemos denominado como *Primary ID* y el resto de los datos como variable *X* de forma correcta. Sin embargo, en este caso son añadidas 2 columnas más a la hoja de cálculo. Estas hacen referencia al *Class ID*, estableciendo a cada observación un identificador de clase haciendo referencia a la clase de fitosanitario al que pertenece, y a la concentración de fitosanitario presente en cada una de las muestras.

Respecto a la concentración, algunas de las olivas de cada lote fueron enviadas a un laboratorio para que se determinara la concentración real de cada lote mediante un análisis químico. Sin embargo, los lotes de concentración más baja no pudieron ser correctamente analizados debido a que tenían una concentración demasiado baja para los análisis. Aunque siguiendo la regla en la que el fitosanitario fue adulterado para lograr concentraciones más bajas y los datos de concentración reales de los lotes que

sí pudieron ser analizados, se pudo estimar la concentración del resto de lotes de forma teórica, aunque aproximada.

Estas dos columnas nuevas añadidas a la hoja de cálculo, servirán para determinar si los datos de las muestras guardan alguna relación perteneciendo al grupo del mismo fitosanitario o una tendencia en caso de compartir misma concentración. La hoja de datos quedaría de la siguiente manera, tal y como se muestra en la Ilustración 16, mostrando solo las primeras filas y columnas para que una visualización correcta del ejemplo sea posible.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	Primary ID	$ClassID	Concent	100,068	101,2	102,332	103,463	104,595	105,726	106,857	107,988	109,119	110,249	111,379	112,509	113,639	114,768	115,898	117,027	118,156	119,284	120,413	121,541
2	Deltametrina 1_785nm50x100p10s_1	Delta	37,2	193,99	592,78	11,026	207,34	407,74	490,45	497,39	384,86	113,48	725,01	257,65	630,87	800,89	910,06	880,42	751,67	576,83	7370,6	075,13	752,55
3	Deltametrina 1_785nm50x100p10s_11	Delta	37,2	318,68	635,37	5,3381	316,31	586,31	734,94	805,12	5752	535,44	198,18	779,07	182,14	448,56	579,65	567,55	507,48	364,08	8125,8	841,12	489,61
4	Deltametrina 1_785nm50x100p10s_12	Delta	37,2	494,84	240,97	79,532	62,054	1912,4	771,73	565,84	267,51	852,95	347,86	775,05	063,29	206,83	6250,7	284,35	6232,5	6115,6	934,64	5730,4	509,68
5	Deltametrina 2_785nm50x100p10s_1	Delta	19,1	591,75	34,139	35,705	09,856	096,89	625,52	088,55	492,41	845,35	140,15	3368	505,19	575,85	3585,5	569,71	501,29	389,64	266,64	134,36	969,24
6	Deltametrina 2_785nm50x100p10s_11	Delta	19,1	316,57	1175,4	22,527	19,265	673,77	441,17	142,89	756,31	4261,7	680,38	5031,3	234,18	5298	338,11	303,17	5174,8	051,41	862,25	645,68	409,46
7	Deltametrina 2_785nm50x100p10s_12	Delta	19,1	748,08	91,444	9,7726	17,386	270,35	859,24	392,58	860,95	257,13	589,55	864,81	014,55	112,54	164,52	166,02	122,06	027,84	895,37	754,27	550,77
8	Deltametrina 3_785nm50x100p10s_1	Delta	0,77	783,18	08,054	99,712	31,826	297,38	897,08	2439,9	915,94	317,56	653,66	931,22	099,51	198,17	207,63	4210,7	136,73	049,35	886,36	719,03	3545,2
9	Deltametrina 3_785nm50x100p10s_11	Delta	0,77	786,82	11,936	04,872	27,996	293,83	892,76	430,13	899,68	297,39	628,51	896,45	047,63	100,39	118,92	065,73	999,49	3911,2	780,36	630,71	490,41
10	Deltametrina 3_785nm50x100p10s_12	Delta	0,77	964,42	011,24	31,133	75,604	410,12	072,57	2662,1	178,55	624,21	998,15	297,39	478,47	587,57	639,63	632,41	4545,6	4403,9	244,31	068,04	899,83
11	Deltametrina 5_785nm50x100p10s_1	Delta	0,0243	616,84	260,34	,93922	106,36	117,91	032,96	901,69	672,31	296,92	825,58	6305,1	623,27	779,66	6843,1	876,08	787,53	744,39	617,43	441,93	187,79
12	Deltametrina 5_785nm50x100p10s_11	Delta	0,0243	117,45	045,47	2,7982	837,78	650,08	384,29	061,79	659,97	159,19	580,74	943,42	197,24	336,19	360,51	353,24	327,63	5278,4	135,82	955,92	749,91
13	Deltametrina 5_785nm50x100p10s_12	Delta	0,0243	526,77	242,78	3,3321	015,52	991,09	873,59	687,37	4406,6	009,04	518,92	957,47	6249,7	423,05	517,47	532,38	478,81	386,46	250,22	6024,7	774,06
14	Diflufenican 1_785nm50x100p10s_1	Diflu	31,8	145,31	047,87	29,909	83,615	718,76	475,71	177,95	800,71	322,36	766,28	153,18	429,35	592,48	673,34	680,64	648,59	592,54	465,03	297,42	112,62
15	Diflufenican 1_785nm50x100p10s_11	Diflu	31,8	972,27	75,031	1,6607	83,708	546,19	235,95	865,82	421,82	892,76	291,39	4628,1	831,18	949,31	017,66	048,76	035,69	950,49	831,25	666,67	506,69
16	Diflufenican 1_785nm50x100p10s_12	Diflu	31,8	508,41	224,62	4,2568	029,33	000,78	880,31	699,56	4425,4	028,31	539,83	988,43	6277,1	452,35	6581,1	6580,2	515,47	415,01	321,05	116,11	871,63
17	Diflufenican 1_785nm50x100p10s_13	Diflu	31,8	025,17	013,11	77,392	70,177	542,36	239,34	871,63	427,59	898,28	294,14	623,34	806,59	906,44	964,55	964,58	924,56	803,23	680,01	535,91	333,14
18	Diflufenican 2_785nm50x100p10s_1	Diflu	3,42	029,16	017,36	1,9144	762,51	531,53	2225,3	857,23	3412,8	3880,2	272,76	601,49	782,62	913,47	946,66	927,93	833,45	737,85	4644,4	489,59	293,14
19	Diflufenican 2_785nm50x100p10s_11	Diflu	3,42	1707,2	51,858	9,9334	59,932	1317,2	912,02	451,91	928,39	335,29	680,07	968,22	126,18	252,56	335,42	338,32	288,41	171,39	4054,2	946,31	801,73
20	Diflufenican 2_785nm50x100p10s_12	Diflu	3,42	585,61	20,912	16,636	31,666	120,21	649,12	113,05	516,56	866,09	156,25	380,06	476,02	523,13	536,66	528,87	487,08	372,74	242,71	122,75	3008
21	Diflufenican 3_785nm50x100p10s_1	Diflu	0,46	770,67	31,216	57,065	63,413	219,31	810,77	325,17	774,27	171,88	505,34	760,23	909,11	992,29	038,33	985,66	928,45	811,25	643,68	493,84	303,47
22	Diflufenican 3_785nm50x100p10s_11	Diflu	0,46	714,83	01,723	53,146	43,818	176,95	746,36	238,22	665,54	043,34	357,71	593,07	731,29	798,01	3789,3	757,05	712,59	3560,1	413,68	3245,3	082,92
23	Diflufenican 3_785nm50x100p10s_12	Diflu	0,46	782,88	27,622	40,248	585,16	243,42	834,65	351,96	801,39	191,05	513,98	761,47	864,65	919,29	921,09	896,92	857,96	733,82	567,28	3446,8	289,11
24	Diflufenican 4_785nm50x100p10s_1	Diflu	0,015	344,16	1163,4	0,4568	07,732	759,22	604,21	348,13	004,23	548,95	007,62	5402,9	668,35	5797,6	867,39	5899,7	816,51	718,65	605,48	445,07	211,26

Ilustración 16: muestra parcial de la hoja de datos.

Una vez guardada la hoja de cálculo y generado el proyecto, se tiene acceso a muchas herramientas que ofrece SIMCA, por lo que se podría empezar con la generación de los modelos. Sin embargo, como último paso en el preprocesado de los datos, se decide recurrir a las herramientas que permiten la manipulación de los datos, concretamente a la herramienta que permite generar un espectro basado en los datos. Mediante esa herramienta se obtiene el siguiente espectro, manteniendo la misma forma con la que se han visualizado los datos de las muestras en el espectrómetro, pero sin el offset.

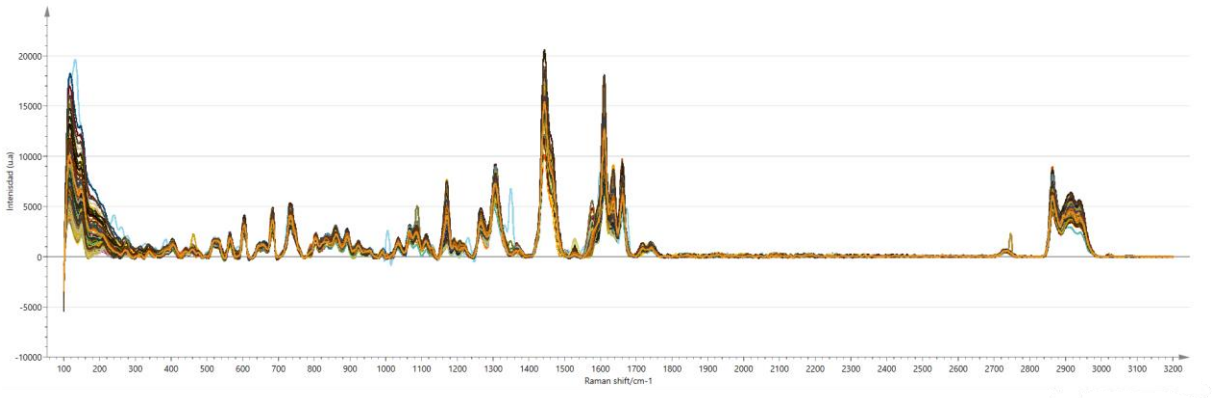


Ilustración 17: Espectros obtenidos de la totalidad de las muestras.

Como se puede apreciar en la ilustración 17, existen dos regiones de las cuales no se puede obtener información. Estas regiones se encuentran entre las longitudes de onda 100 nm y 350 nm, y 1800 nm y 2600 nm, aproximadamente. Por lo tanto, se determina que esas regiones no son determinantes para la generación de los modelos y se toma la decisión de excluir dichas zonas en los modelos a generar, tal y como se observa en la Ilustración 18.

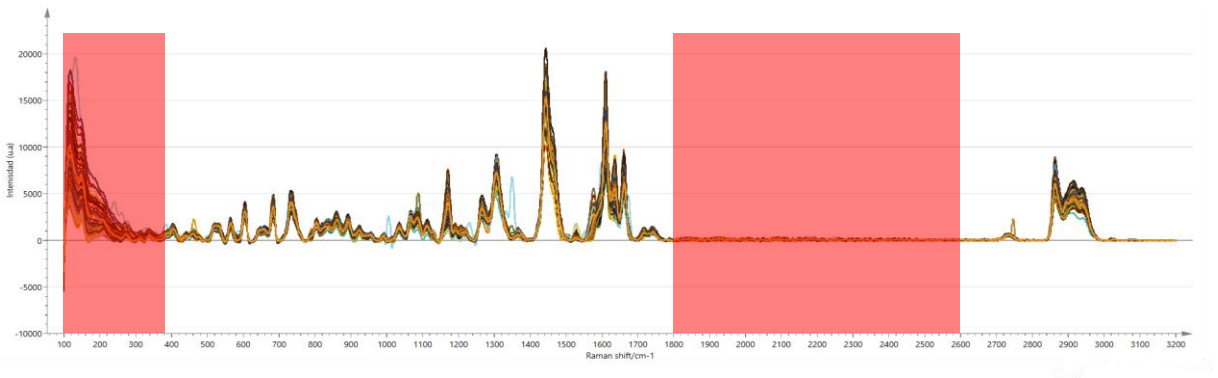


Ilustración 18: Espectros obtenidos de la totalidad de las muestras con zonas a excluir.

6.5. Generación del modelo

Una vez terminado el preprocesado de los datos, lo siguiente es la generación del modelo [20]. Pero antes, es importante tener claros los objetivos a alcanzar con el desarrollo del proyecto, ya que este punto determinará el tipo de modelo predictivo a desarrollar.

Tras varias pruebas y, teniendo en cuenta los datos disponibles, se decidió usar dichos datos para generar un modelo que tenga la capacidad de predecir si una nueva muestra ha sido tratada o no, independientemente de la concentración del tratamiento. De esta forma, toda aceituna no tratada será considerada saludable para el usuario, mientras que las catalogadas como tratadas podrán ser consideradas saludables o no según la legislación vigente, siendo descartadas para el consumo humano en caso de no serlo. En este punto, siguiendo la teoría sobre los modelos de análisis multivariante de datos explicada en apartados anteriores, se puede llegar a la conclusión de que el modelo más acertado para este proyecto es el OPLS-DA. Sin embargo, en este apartado se demostrará con resultados prácticos los pasos seguidos hasta llegar a dicha conclusión, confirmando así la teoría explicada.

Como primer paso se decide generar un modelo PCA-X, ya que lo interesante al principio es ver un resumen o una visión general de todos los datos. Como se ha explicado con anterioridad, este modelo permite analizar conjuntos de datos que pueden contener, por ejemplo, multicolinealidad, valores perdidos, datos categóricos y mediciones imprecisas. De esta manera, se genera el siguiente *score plot* teniendo en cuenta todas las muestras salvo las pertenecientes a la clase de las olivas sin tratar. De esta forma, se intenta obtener una visión general de similitudes entre los diferentes fitosanitarios y concentraciones.

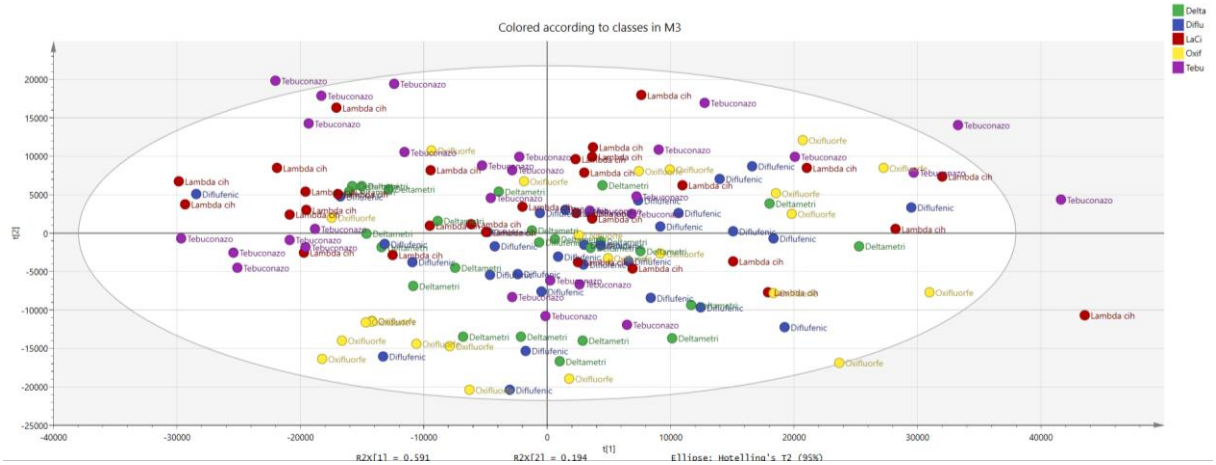


Ilustración 19: *score plot* del modelo PCA-X de todas las clases salvo Sin Tratar.

Sin embargo, como se puede observar en la Ilustración 19, no existe ningún tipo de relación aparente entre fitosanitarios o diferentes concentraciones a simple vista, ya que no se percibe que sigan ningún patrón de organización y solo se muestran como una nube aleatoria de datos. Esto es una conclusión alcanzada teniendo en cuenta que las clases parecen estar distribuidas por toda la nube, a pesar de haber pequeñas agrupaciones. Además, estas agrupaciones no guardan relación alguna con la concentración, ya que están agrupadas a pesar de tener diferente concentración, así como cerca de otras agrupaciones de diferente clase y concentración aleatoria.

Este modelo ha sido generado con 4 componentes principales y se han visualizado todas las combinaciones posibles de los componentes sin éxito. También se puede observar cómo algunas de las observaciones se salen fuera de lo considerado la norma dentro de esta nube, aunque este modelo no es lo suficientemente determinante como para excluir muestras.

En vista de estos resultados, se puede llegar a la conclusión de que las variables son bastante colineales entre sí. Por lo tanto, siguiendo con la teoría anteriormente explicada, se valora la generación de un modelo PLS u OPLS como solución al problema. La regresión PLS u OPLS resulta especialmente útil cuando las variables son muy colineales o cuando se tienen más variables que observaciones.

Extrayendo un fragmento de teoría anteriormente explicado, entendemos que en el caso de una sola variable Y solo hay un componente predictivo, y todos los componentes más allá del primero reflejan una variación ortogonal. Sin embargo, con múltiples variables Y puede haber más de un componente OPLS predictivo. En la situación de una única variable Y, un modelo PLS y un modelo OPLS ajustados a los mismos datos tendrán el mismo poder predictivo. Por lo tanto, se puede llegar a la conclusión de que, en este caso, no es relevante cuál de los dos modelos sea elegido para generar el modelo.

Sin embargo, una parte importante que se puede pasar por alto y es determinante a la hora de elegir la mejor opción, es la gran ventaja que tiene OPLS frente a PLS. Volviendo a la teoría anteriormente explicada, esta reside en la interpretabilidad simplificada del modelo que surge con el OPLS, debido a la capacidad de separar la varianza explicada en compartimentos predictivos y ortogonales del modelo. Por lo tanto, OPLS ofrece una mejor visualización cuando hay una gran cantidad de estructura

ortogonal Y en X y puede ayudar a aclarar y comprender la variación correlacionada y no correlacionada.

Siguiendo lo mencionado, se genera un modelo OPLS con los datos de todas las clases salvo la clase sin tratar (ilustración 20).

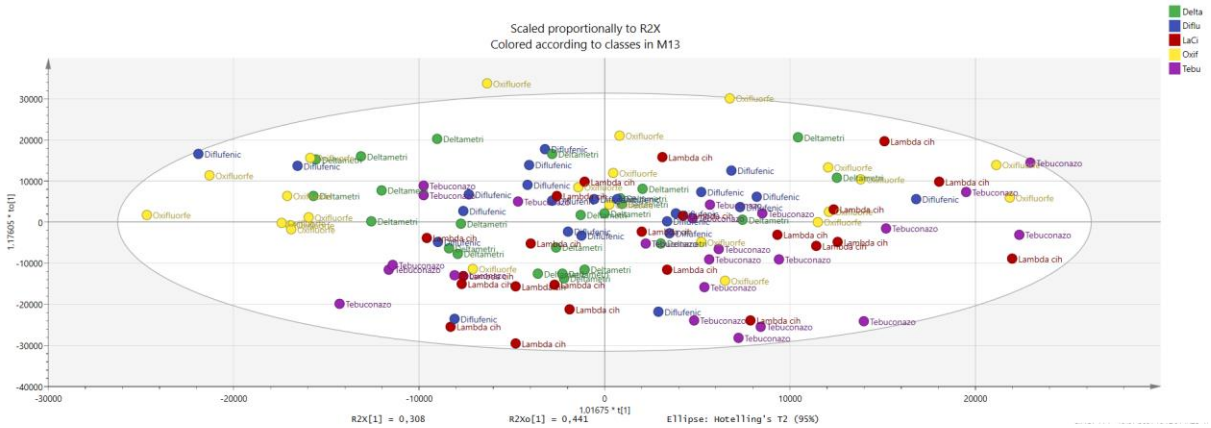


Ilustración 20: *score plot* del modelo OPLS sin la clase Sin Tratar.

Como se puede observar en la Ilustración 20, el modelo sigue sin mostrar agrupaciones concluyentes. Esto es debido a que, hasta ahora, siguiendo la teoría de nuevo, lo que se ha estado haciendo con la generación de estos modelos es determinar si las clases guardan algún tipo de relación entre sí, definiendo los límites de las clases con el objetivo de poder así inferir la pertenencia a una clase con las futuras muestras del conjunto de predicción. De esta manera, no se está centrando tanto en las diferencias que las clases tienen entre sí.

Sin embargo, lo que interesa en este caso teniendo en cuenta la similitud que guardan las diferentes clases, es separar los grupos centrándose en las variables que impulsan la separación de estas clases. Por lo que lo más acertado es usar un modelo de análisis discriminante. En base a esta deducción, se genera el siguiente modelo basado en el modelo OPLS-DA con las mismas clases que hasta ahora.

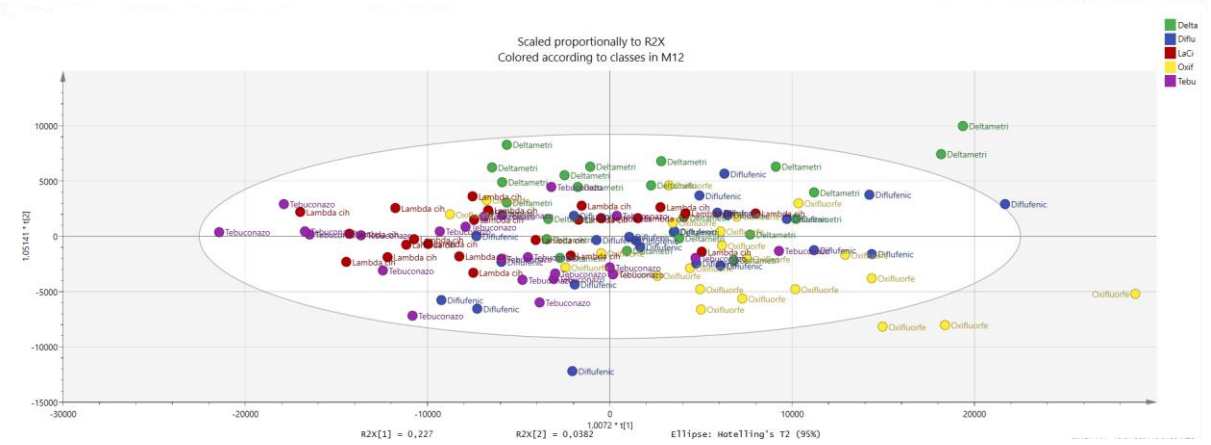


Ilustración 21: *score plot* del modelo OPLS-DA sin la clase Sin Tratar.

En este modelo, tal y como se aprecia en la Ilustración 21, se aprecian mejor las agrupaciones de cada clase, ya que se ve que estas tienen mayor tendencia a agruparse en alguno de los cuatro cuadrantes. Sin embargo, la distribución por concentración sigue siendo aleatoria.

Hay que recalcar que este tipo de modelo está pensado para la clasificación de dos clases y, aunque permite generar modelos con más clases, estos modelos no son tan precisos. La razón de esto es que, siguiendo la teoría de nuevo, en el caso de tener un problema de dos grupos el modelo OPLS-discriminante resultante será muy fácil de interpretar porque sólo tendrá un componente predictivo que interpretar. Este componente se representa como el eje de abscisas en el gráfico de dispersión de *scores* resultante del modelo OPLS-DA.

De esa manera, la dirección horizontal del gráfico de dispersión de *scores* captará la variación entre los grupos. Mientras, la dimensión vertical y cualquier componente superior de los denominados de tipo ortogonal captarán la variación dentro de los grupos.

Por lo tanto, resulta interesante pensar en generar un modelo que disponga únicamente de dos clases, diferenciando entre muestras tratadas y sin tratar. A pesar de ello, debido a que cuando se organizaron las mediciones se disponía de poco tiempo por la degeneración de las muestras, no se pudo tener en cuenta el tipo de modelo a generar y, por lo tanto, se tomó el mismo número de muestras con todos los lotes. Teniendo esto en cuenta, al disponer solo de un lote de muestras del grupo de muestras no tratadas frente a los 5 lotes que componen el resto de grupos, se llega rápidamente a la conclusión de que se dispone de un menor número de datos de la clase sin tratar. Por lo que sería muy poco preciso un modelo generado con casi el 5% de todas las muestras componiendo una clase frente al 95% del resto de muestras que formarían parte de la otra clase.

Es por esta razón que, con los datos de los que se disponía, se decidió generar un modelo por cada fitosanitario, enfrentando las muestras de cada uno contra las muestras no tratadas. De esta forma, las muestras de la clase sin tratar ascienden hasta ser casi el 20% de las muestras totales del modelo en el peor de los casos. Sigue sin ser el escenario más óptimo, pero es el mejor al que se ha podido aspirar en este caso.

Es importante recordar que, para generar un modelo predictivo funcional, para validar la capacidad predictiva hay que usar nuevas muestras no usadas para para la generación de dicho modelo. Es por esto que se decidió prescindir del 20% de las muestras en la generación del modelo con el objetivo de usar éstas para comprobar la robustez del modelo.

6.6. Resultados obtenidos

Para la validación del modelo, al ser tan dispar el número de muestras por clase, se decidió calcular el 20% de cada una de las dos clases que componen el modelo y seleccionar así acorde muestras para usarlas en la validación. En el caso de la clase sin tratar, al haber solo 7 muestras de esta clase se

tomaron solo 2 para la posterior predicción. Estas muestras fueron seleccionadas aleatoriamente en todos los casos.

Los modelos generados se pueden observar en la ilustración 22:

No.	Model	Type	A	N	R2X(cum)	R2Y(cum)	Q2(cum)	Date	Title
6	M6	OPLS-DA	1+3+0	25	0,923	0,823	0,66	04/01/2024	80%(DeltaTodas VS SinTodas)
7	M7	OPLS-DA	1+3+0	30	0,887	0,585	0,0338	03/02/2024	80%(DifluTodas VS SinTodas)
8	M8	OPLS-DA	1+3+0	30	0,955	0,751	0,481	08/01/2024	80%(LaciTodas VS SinTodas)
9	M9	OPLS-DA	1+3+0	24	0,943	0,721	0,339	08/01/2024	80%(OxifTodas VS SinTodas)
10	M10	OPLS-DA	1+3+0	29	0,9	0,65	0,315	03/02/2024	80%(DifluTodas VS SinTodas) - 1 muestra
11	M11	OPLS-DA	1+2+0	28	0,927	0,658	0,514	08/01/2024	80%(TebuTodasVsSinTodas)

Ilustración 22: tabla de los modelos generados y sus características.

6.6.1. Deltametrina

Como primer modelo OPLS-DA, tenemos el modelo de la clase *Delta* (compuesto por las muestras del fitosanitario Deltametrina) frente a la clase *Sin Tratar* (compuesto por las muestras no tratadas). En la Ilustración 23 se puede ver el *score plot* de este modelo generado con el 80% de las muestras.

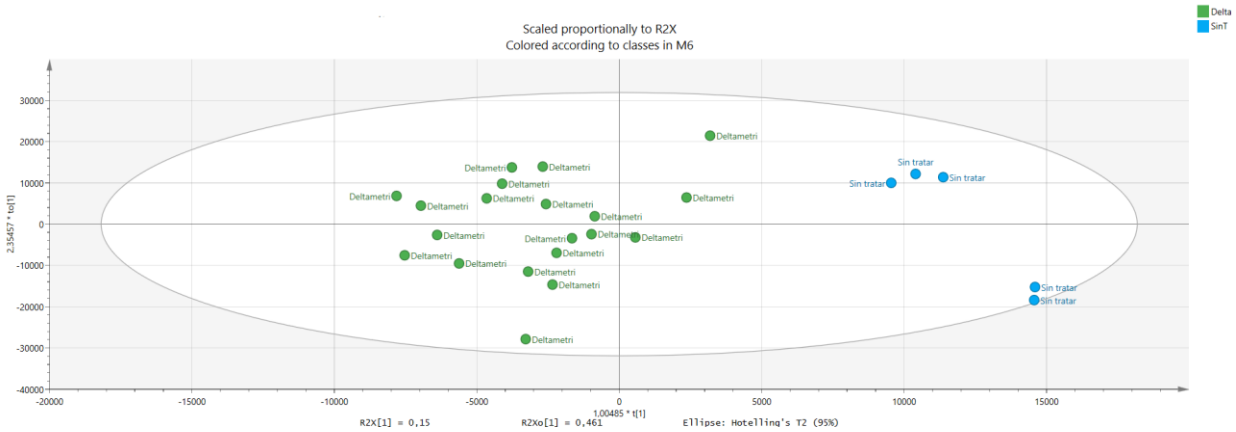


Ilustración 23: *score plot* del modelo OPLS-DA Delta vs Sin Tratar.

Como se puede observar en la ilustración 23, claramente existe una diferenciación entre ambas clases, teniendo la clase *Delta* tendencia a posicionarse hacia la izquierda mientras que la clase *Sin Tratar* está claramente posicionada a la derecha. Se entiende que la diferenciación no es perfecta al no ser las clases separadas horizontalmente en la mitad por la falta de muestras no tratadas. Esto también provoca que estas muestras no tratadas se vean desplazadas hasta el límite de la elipse y se puedan apreciar pequeños grupos diferenciados entre ellas. Es importante destacar que, en este caso, la diferenciación vertical, la cual se centra en buscar diferencias dentro de cada clase, no está clasificando las muestras por concentración, ya que estas vuelven a estar mezcladas. Este modelo ha sido ajustado con una componente ortogonal y 3 componentes, tal y como se puede observar en la Ilustración 24.

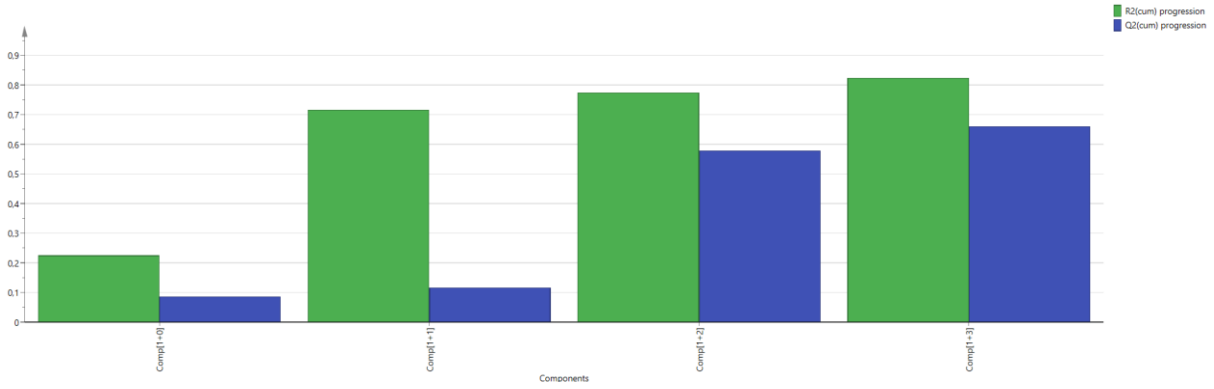


Ilustración 24: *summary of fit* del modelo OPLS-DA Delta vs Sin Tratar.

Resulta interesante fijarse también en el gráfico de los coeficientes, el cual muestra a qué partes del espectro Raman está dando más peso cada modelo. Este se puede observar en la ilustración 25.

Ilustración 25: *coefficients* del modelo OPLS-DA Delta vs Sin Tratar

En cuanto a la predicción, comprobamos que el modelo es bastante robusto. Al usar para la validación las 6 muestras descartadas en la generación del modelo, comprobamos que solo comete un fallo al clasificar una muestra no tratada como tratada. Esto se considera un falso positivo (tomando como positivas las muestras clasificadas como tratadas), lo que es considerado más aceptable que detectar falsos negativos. Esto es debido a que, en este caso, clasificar una muestra no tratada como tratada es menos peligroso para el consumidor final que haber clasificado una muestra tratada como no tratada, ya que en este último caso el consumidor podría consumir algo potencialmente peligroso para la salud. En la Ilustración 26 se pueden observar los datos de la tabla de clasificación o *misclassification table* como se conoce en inglés.

	1	2	3	4	5	6
1		Members	Correct	Delta	SinT	No class (YPred <= 0)
2	Delta	4	100%	4	0	0
3	SinT	2	50%	1	1	0
4	No class	0		0	0	0
5	Total	6	83,33%	5	1	0
6	Fisher's prob.	0,33				

Ilustración 26: *misclassification table* de la validación modelo OPLS-DA Delta vs Sin Tratar.

El peso con el que el modelo ha clasificado cada muestra se puede observar en la siguiente tabla (ilustración 27).

	1	2	3	4	5	6
1	Obs ID (Primary)	Obs ID ($ClassID)	M6.YVarPS($M6.DA(Delta))	M6.YPredPS[1]($M6.DA(Delta))	M6.YVarPS($M6.DA(SinT))	M6.YPredPS[1]($M6.DA(SinT))
2	Deltametrina 1_785nm50x100p10s_12	Delta	1	1,05738	0	-0,0573842
3	Deltametrina 5_785nm50x100p10s_12	Delta	1	1,25588	0	-0,25588
4	Sin tratar_785nm50x100p10s_13	SinT	0	-0,22878	1	1,22878
5	Deltametrina 2_785nm50x100p10s_22	Delta	1	1,15679	0	-0,156787
6	Deltametrina 3_785nm50x100p10s_2	Delta	1	1,19613	0	-0,19613
7	Sin tratar_785nm50x100p10s_21	SinT	0	0,672024	1	0,327976

Ilustración 27: lista de clasificación del modelo OPLS-DA Delta vs Sin Tratar.

Aquí, se puede observar que las muestras no tratadas usadas para predecir son las muestras en las que el modelo tiene más duda en cuál de las dos clases clasificarlas. Esto no sorprende, ya que uno de las principales limitantes del proyecto ha sido la falta de muestras, sobre todo con el lote de muestras no tratadas. En el caso de que hubiera habido mejor paridad entre ambas clases a la hora de generar el modelo, se podría haber considerado un modelo mucho más robusto. Aun así, los resultados obtenidos son muy aceptables. Algo similar ocurre con los modelos generados usando el resto de fitosanitarios.

6.6.2. Diflufenican

En el caso de Diflufenican, sin embargo, se detectó que había 2 muestras que podían ser consideraras "*outliers*". Estas muestras suelen haber sido generadas por la influencia de algo externo al sistema y su rápida detección y posterior exclusión es importante. Son muestras que se salen de la norma del resto de muestras y que deben ser excluidas de la generación del modelo ya que pueden afectar a la precisión del modelo al tener esas muestras en consideración.

La primera, fue detectada en el propio espectro, ya que era fácil de ver al salirse bastante fuera de la norma de la media de los espectros de todas las muestras. Esto era debido a que presentaba picos donde el resto de muestras no, por lo que se decidió excluirla por precaución. Se puede observar con claridad la muestra en la Ilustración 28.

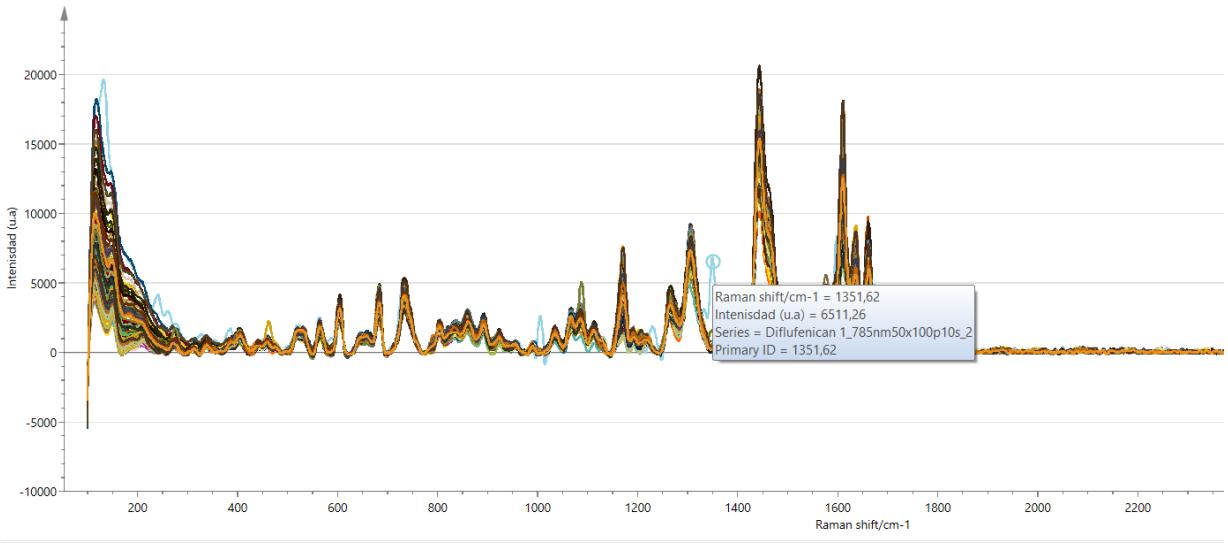


Ilustración 28: espectro de la muestra de Diflufenican excluida.

La segunda muestra a excluir se detectó a la hora de generar el modelo con el 80% de las muestras. La decisión fue tomada debido a que, según el modelo generado, esta podría ser clasificada perfectamente como una muestra no tratada. Puede que a la hora de hacer la medición se hiciera en un punto de la superficie en el que no hubiera casi rastro del fitosanitario y eso afectase a la lectura, pero por cautela se decidió excluir esta muestra también ya que afectó a la precisión del modelo. Se puede ver el gráfico de componentes generado por el modelo en la Ilustración 29, con dicha muestra incluida.

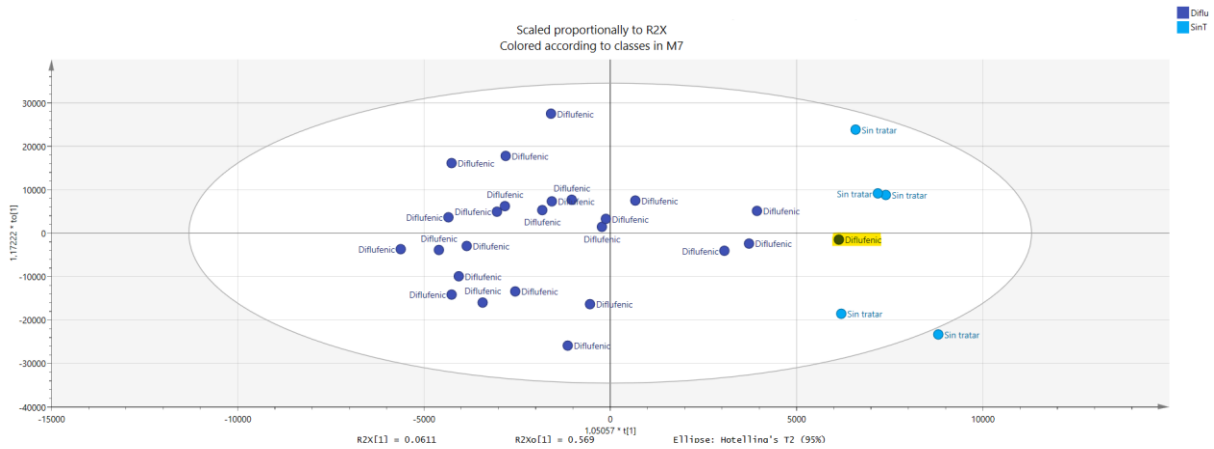


Ilustración 29: *score plot* del modelo OPLS-DA Diflu vs Sin Tratar con *outlier* incluida.

Este modelo ha sido generado con una componente ortogonal y 3 componentes, tal y como se puede observar en la Ilustración 30.

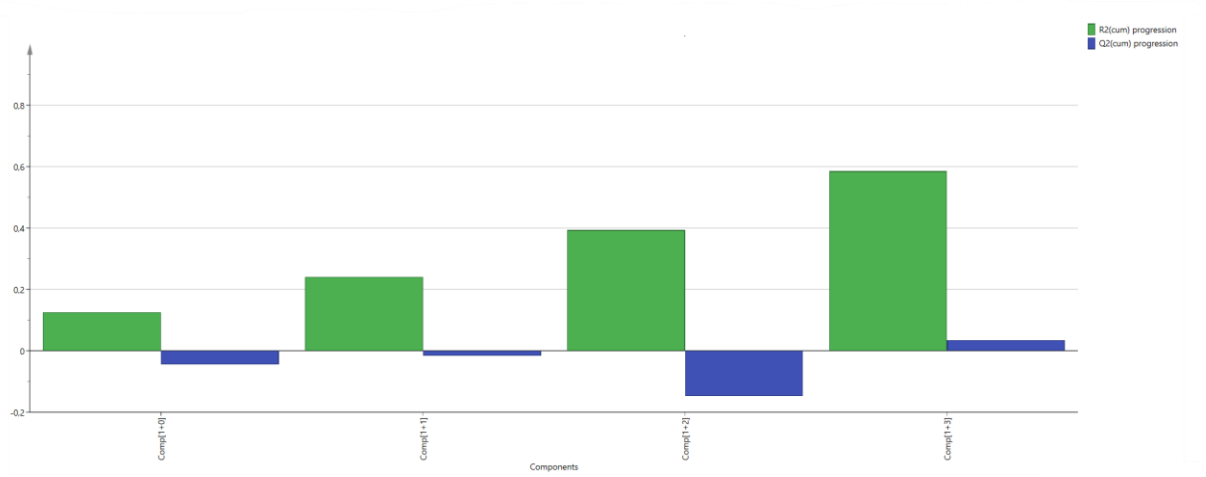


Ilustración 30: *summary of fit* del modelo OPLS-DA Diflu vs Sin Tratar con *outlier* incluida.

El gráfico *coefficients* se puede observar en la ilustración 31.

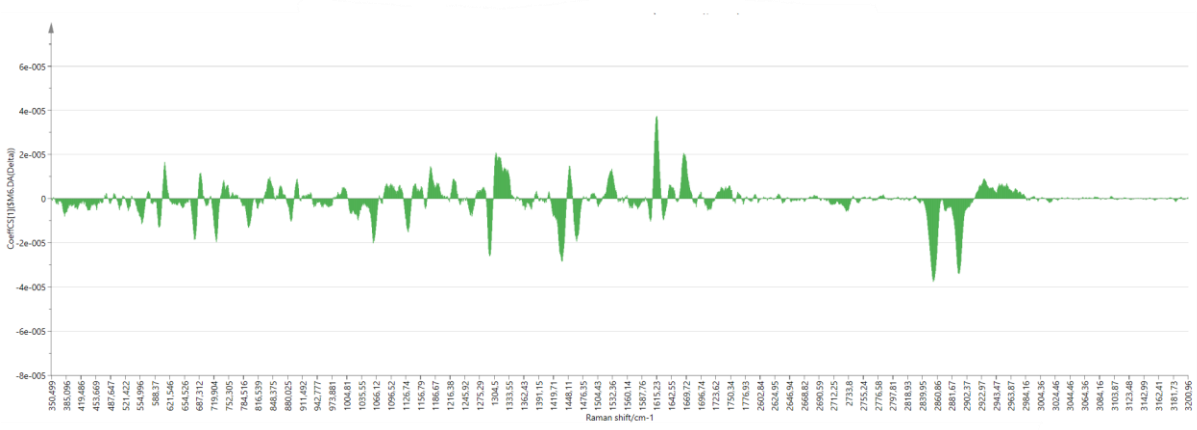


Ilustración 31: *coefficients* del modelo OPLS-DA Diflu vs Sin Tratar con *outlier* incluida.

La *misclassification table* y la lista de clasificación que acompañan a dicho modelo son las siguientes (ilustración 32 y 33).

	1	2	3	4	5	6
1		**Members**	**Correct**	**Diflu**	**SinT**	**No class (YPred <= 0)**
2	**Diflu**	6	100%	6	0	0
3	**SinT**	2	50%	1	1	0
4	**No class**	0		0	0	0
5	Total	8	87,5%	7	1	0
6	Fisher's prob.	0,25				

Ilustración 32: *misclassification table* del modelo OPLS-DA Diflu vs Sin Tratar con *outlier* incluida.

	1	2	3	4	5	6
1	**Obs ID (Primary)**	**Obs ID ($ClassID)**	**M7.YVarPS($M7.DA(Diflu))**	**M7.YPredPS[1]($M7.DA(Diflu))**	**M7.YVarPS($M7.DA(SinT))**	**M7.YPredPS[1]($M7.DA(SinT))**
2	**Diflufenican 1_785nm50x100p10s_13**	Diflu	1	1,0358	0	-0,0358039
3	**Diflufenican 3_785nm50x100p10s_12**	Diflu	1	0,757958	0	0,242042
4	**Diflufenican 4_785nm50x100p10s_1**	Diflu	1	0,749657	0	0,250343
5	**Sin tratar_785nm50x100p10s_12**	SinT	0	0,482522	1	0,517478
6	**Diflufenican 2_785nm50x100p10s_2**	Diflu	1	0,856583	0	0,143417
7	**Diflufenican 3_785nm50x100p10s_2**	Diflu	1	1,18101	0	-0,181008
8	**Diflufenican 5_785nm50x100p10s_21**	Diflu	1	0,998809	0	0,00119099
9	**Sin tratar_785nm50x100p10s_22**	SinT	0	0,567209	1	0,432791

Ilustración 33: lista de clasificación del modelo OPLS-DA Diflu vs Sin Tratar con *outlier* incluida.

En comparación, tenemos los datos del modelo generado sin dicha *outlier*. El *score plot*, summary of fit, *misclassification table* y lista de clasificación de dicho modelo son las siguientes (ilustración 34, 35 36, 37 y 38 respectivamente).

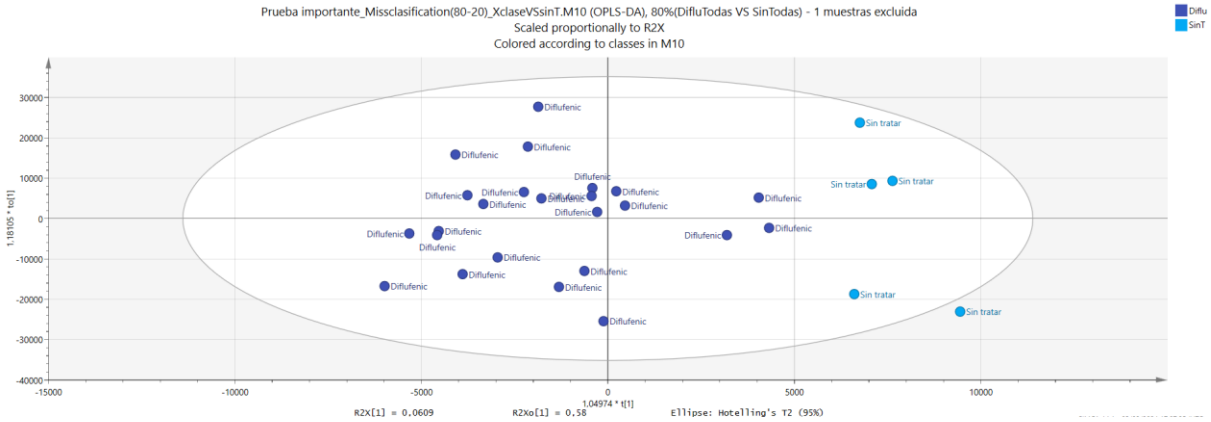


Ilustración 34: *score plot* del modelo OPLS-DA Diflu vs Sin Tratar con *outlier* excluida.

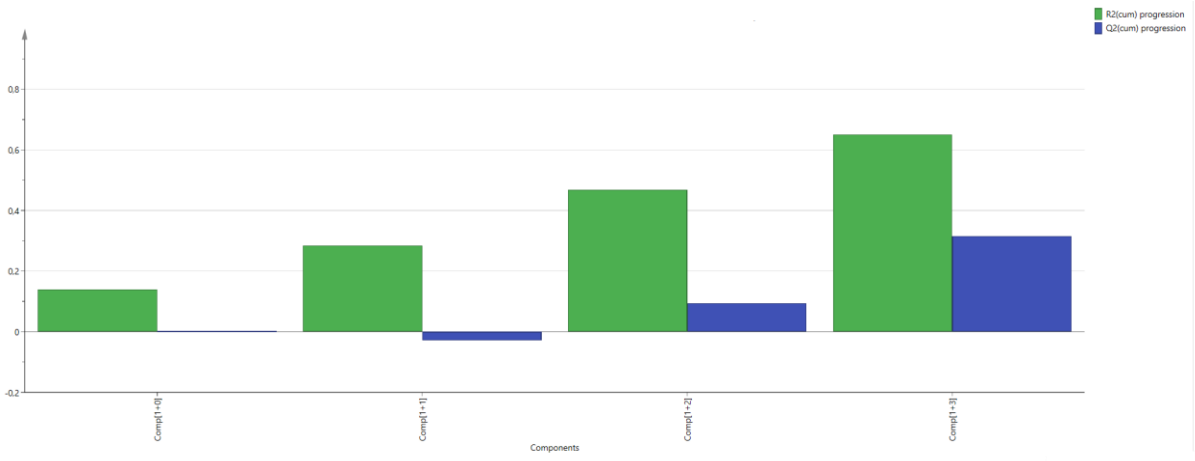


Ilustración 35: *summary of fit* del modelo OPLS-DA Diflu vs Sin Tratar con *outlier* excluida.

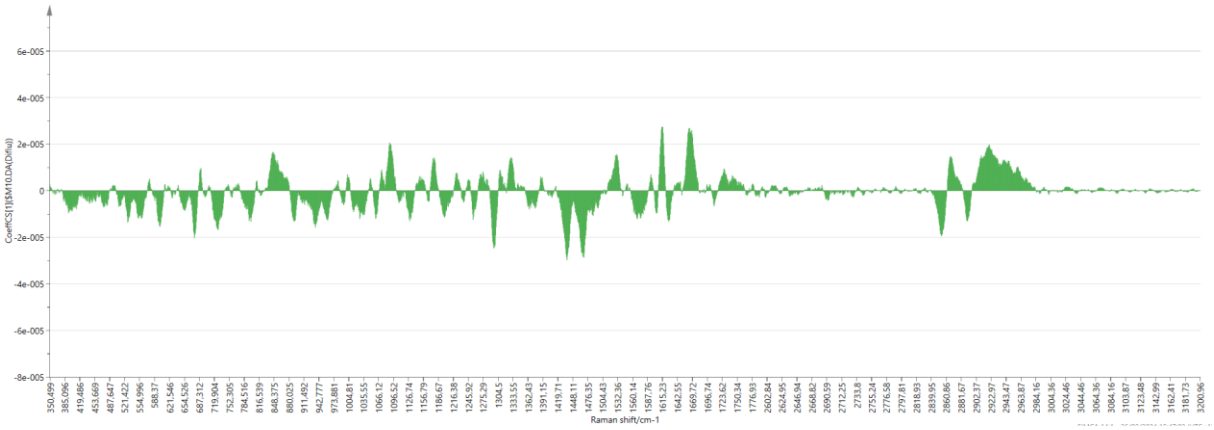


Ilustración 36: *coefficients* del modelo OPLS-DA Diflu vs Sin Tratar con *outlier* excluida.

	1	2	3	4	5	6
1		Members	Correct	Diflu	SinT	No class (YPred <= 0)
2	Diflu	6	100%	6	0	0
3	SinT	2	50%	1	1	0
4	No class	0		0	0	0
5	Total	8	87,5%	7	1	0
6	Fisher's prob.	0,25				

Ilustración 37: *misclassification table* del modelo OPLS-DA Diflu vs Sin Tratar con *outlier* excluida.

	1	2	3	4	5	6
1	Obs ID (Primary)	Obs ID ($ClassID)	M10.YVarPS($M10.DA(Diflu))	M10.YPredPS[1]($M10.DA(Diflu))	M10.YVarPS($M10.DA(SinT))	M10.YPredPS[1]($M10.DA(SinT))
2	Diflufenican 1_785nm50x100p10s_13	Diflu	1	0,889325	0	0,110675
3	Diflufenican 3_785nm50x100p10s_12	Diflu	1	0,782778	0	0,217222
4	Diflufenican 4_785nm50x100p10s_1	Diflu	1	0,767935	0	0,232065
5	Sin tratar_785nm50x100p10s_12	SinT	0	0,379931	1	0,620069
6	Diflufenican 2_785nm50x100p10s_2	Diflu	1	0,671093	0	0,328907
7	Diflufenican 3_785nm50x100p10s_2	Diflu	1	1,26739	0	-0,26739
8	Diflufenican 5_785nm50x100p10s_21	Diflu	1	0,89293	0	0,10707
9	Sin tratar_785nm50x100p10s_22	SinT	0	0,625553	1	0,374447

Ilustración 38: lista de clasificación del modelo OPLS-DA Diflu vs Sin Tratar con *outlier* excluida.

Como se puede apreciar en las ilustraciones 34 y 35, el modelo sí puede considerarse ligeramente superior después de excluir dicho *outlier*, ya que ambas clases están algo más separadas y su capacidad predictiva parece haber mejorado según los datos. Se puede observar también algo de mejoría en la lista de clasificación, concretamente en el ligero incremento del peso que tiene la primera muestra no tratada para ser clasificada como no tratada. Sin embargo, el peso de la siguiente muestra no tratada incrementa hacia la clasificación como tratada. Como conclusión, la predicción no ha mejorado mucho a pesar de una mejor separación de ambas clases, y esto sigue siendo debido a que al tener pocas muestras no tratadas el modelo no es lo suficientemente fuerte, y menos con este modelo ya que se

puede observar que las muestras de Diflufenican son bastante más similares a las no tratadas que otros fitosanitarios.

En la ilustración 39 se puede apreciar como encajan las muestras de predicción en el modelo y por qué el modelo ha clasificado una muestra erróneamente.

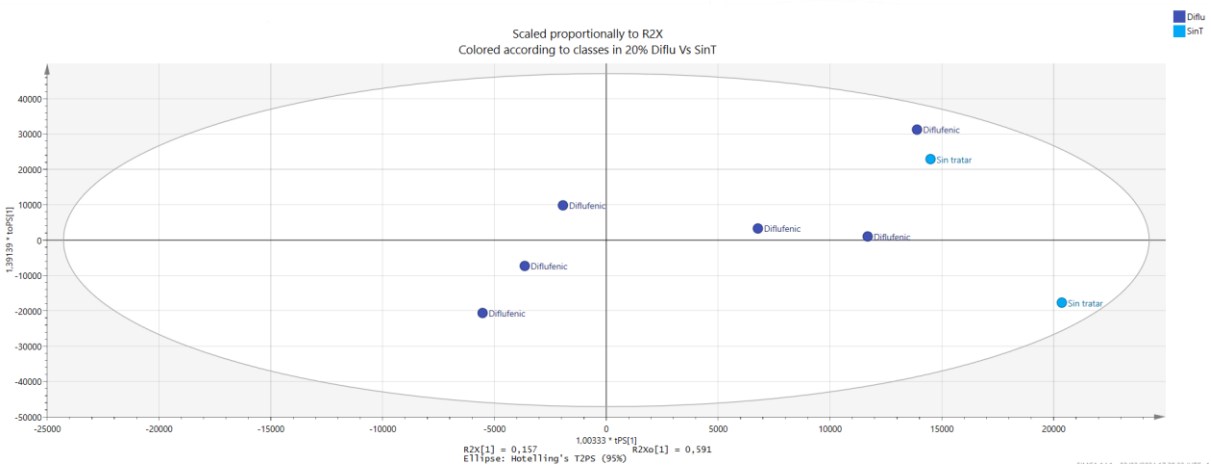


Ilustración 39: *score plot* predictivo del modelo OPLS-DA Diflu vs Sin Tratar con *outlier* excluida.

6.6.3. Lambda Cihalometrina

En cuanto al modelo OPLS-DA generado con muestras de Lambda Cihalometrina y no tratadas, se obtienen mejores resultados. Dichos resultados se pueden observar en las ilustraciones de la 40 a la 45.

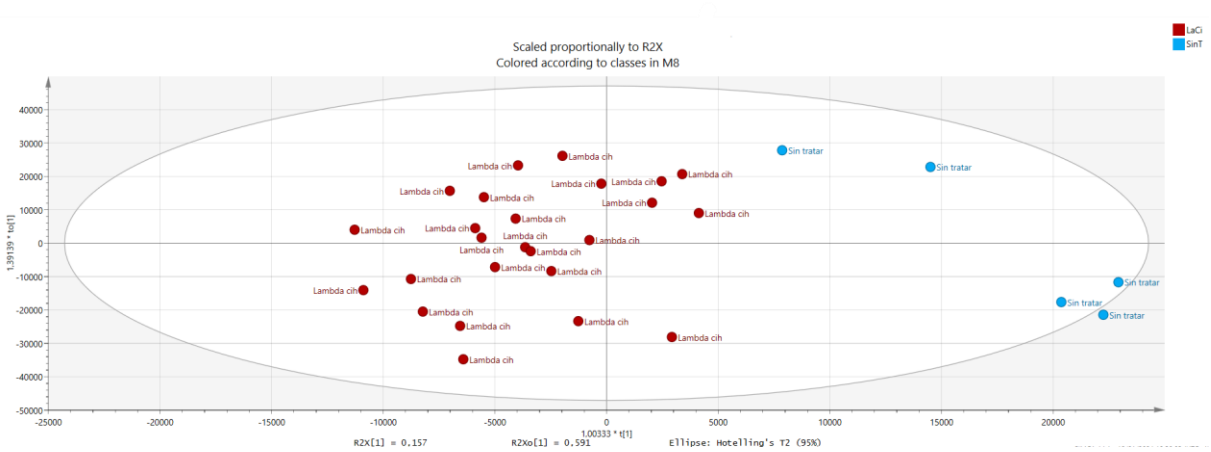


Ilustración 40: *score plot* del modelo OPLS-DA Laci vs Sin Tratar.

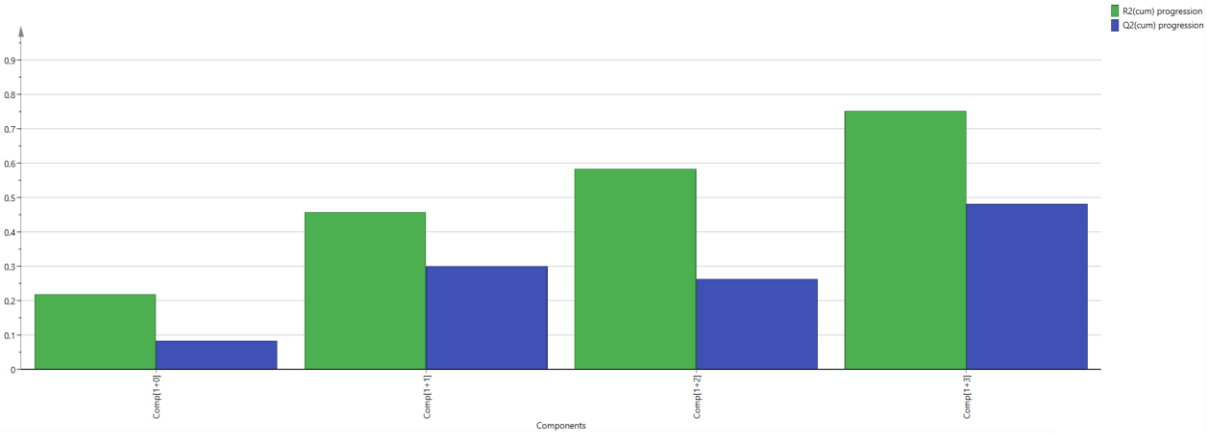


Ilustración 41: *summary of fit* del modelo OPLS-DA Laci vs Sin Tratar.

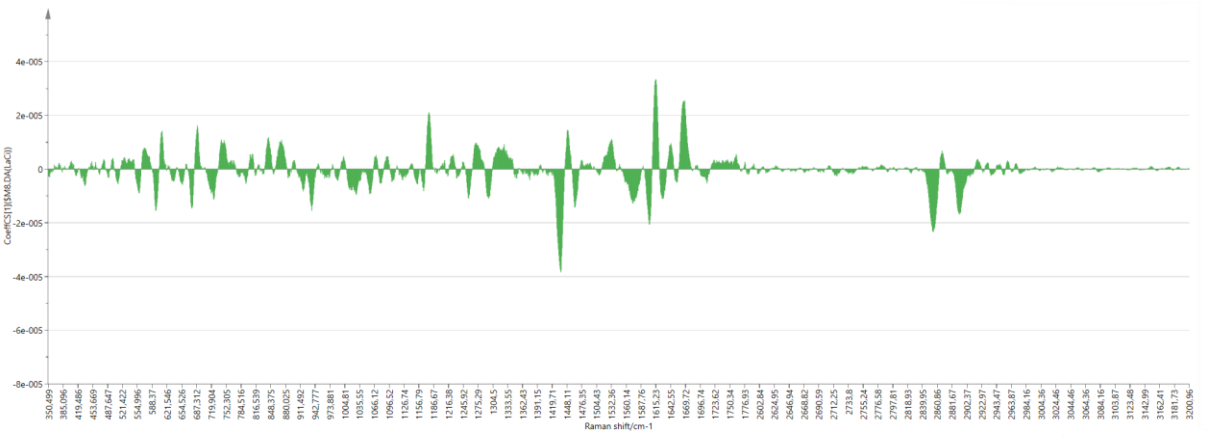


Ilustración 42: *coefficients* del modelo OPLS-DA Laci vs Sin Tratar.

	1	2	3	4	5	6
1		Members	Correct	LaCi	SinT	No class (YPred <= 0)
2	LaCi	6	100%	6	0	0
3	SinT	2	100%	0	2	0
4	No class	0		0	0	0
5	Total	8	100%	6	2	0
6	Fisher's prob.	0,036				

Ilustración 43: *misclassification table* del modelo OPLS-DA LAci vs Sin Tratar.

	1	2	3	4	5	6
1	Obs ID (Primary)	Obs ID ($ClassID)	M8.YVarPS($M8.DA(LaCi))	M8.YPredPS[1]($M8.DA(LaCi))	M8.YVarPS($M8.DA(SinT))	M8.YPredPS[1]($M8.DA(SinT))
2	Lambda cihalotrina 1_785nm50x100p10s_1	LaCi	1	0,794676	0	0,205324
3	Lambda cihalotrina 3_785nm50x100p10s_12	LaCi	1	0,687964	0	0,312036
4	Lambda cihalotrina 4_785nm50x100p10s_11	LaCi	1	0,727605	0	0,272395
5	Sin tratar_785nm50x100p10s_1	SinT	0	0,378217	1	0,621783
6	Lambda cihalotrina 1_785nm50x100p10s_2	LaCi	1	1,23287	0	-0,23287
7	Lambda cihalotrina 2_785nm50x100p10s_22	LaCi	1	0,9918	0	0,00819974
8	Lambda cihalotrina 4_785nm50x100p10s_21	LaCi	1	0,871743	0	0,128257
9	Sin tratar_785nm50x100p10s_2	SinT	0	0,371199	1	0,628801

Ilustración 44: lista de clasificación del modelo OPLS-DA Laci vs Sin Tratar.

Como se puede observar, en este caso la predicción es del 100%, por lo que se puede entender que hace una mejor clasificación de los datos. Esto puede ser debido a que los datos de las muestras del fitosanitario Lambda Cihalometrina no son tan parecidos a las muestras no tratadas como pasaba con el caso anterior. Aun así, se aprecia en la lista de clasificación que dicha clasificación, a pesar de ser certera, no tiene pesos tan altos a la hora de determinar la clasificación correcta con las muestras no tratadas. Volvemos a destacar la importancia de tener un número alto y equitativo de muestras para mejorar la calidad de la clasificación.

6.6.4. Oxifluorfen

En cuanto al modelo OPLS-DA generado con muestras de Oxifluorfen y no tratadas, se obtienen los siguientes resultados observables en las ilustraciones de la 45 a la 49.

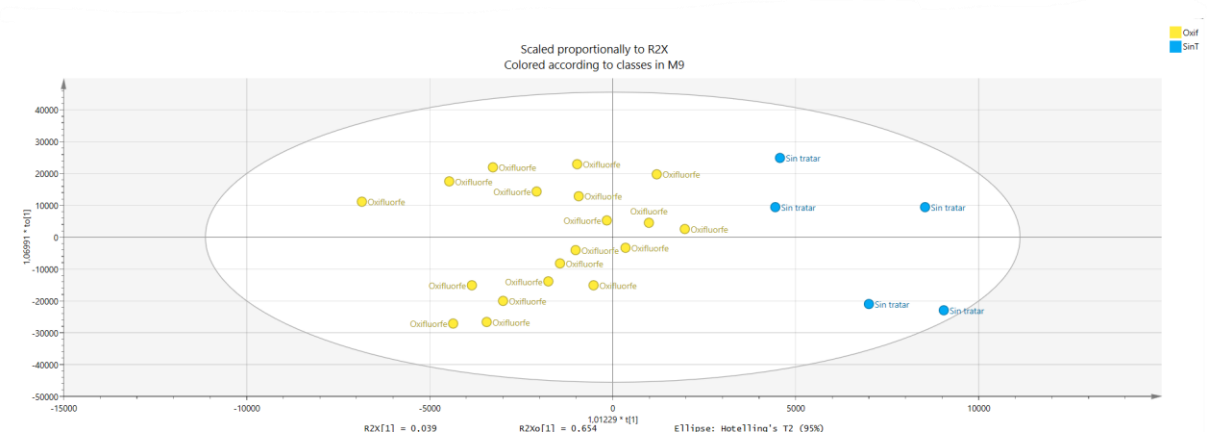


Ilustración 45: *score plot* del modelo OPLS-DA Oxi vs Sin Tratar.

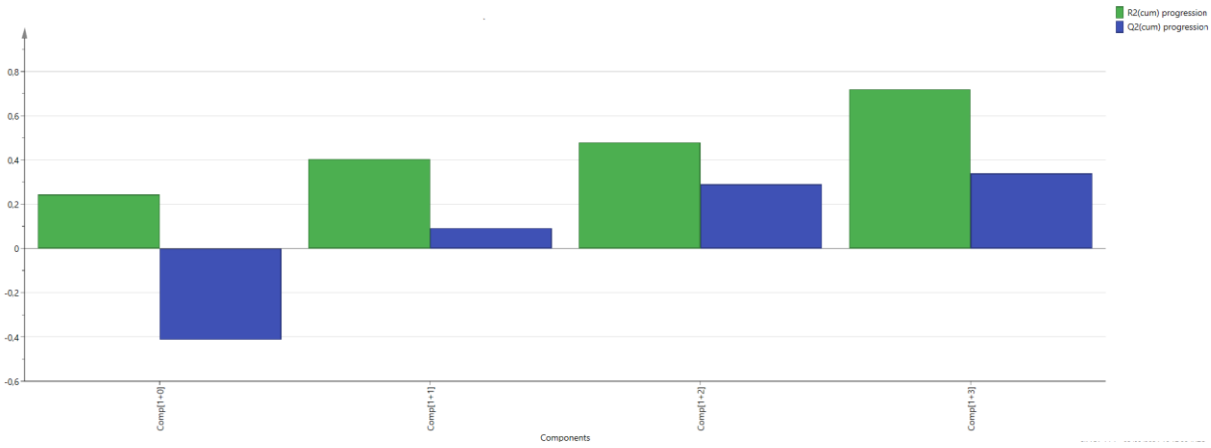


Ilustración 46: *summary of fit* del modelo OPLS-DA Oxi vs Sin Tratar.

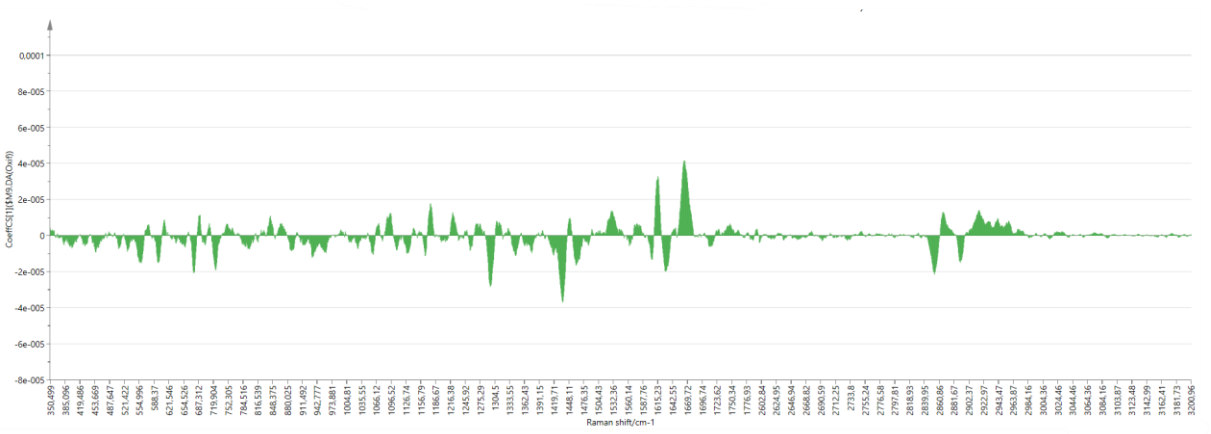


Ilustración 47: *coefficients* del modelo OPLS-DA Oxi vs Sin Tratar.

	1	2	3	4	5	6
1		Members	Correct	Oxif	SinT	No class (YPred <= 0)
2	Oxif	5	100%	5	0	0
3	SinT	2	50%	1	1	0
4	No class	0		0	0	0
5	Total	7	85,71%	6	1	0
6	Fisher's prob.	0,29				

Ilustración 48: *misclassification table* del modelo OPLS-DA Oxi vs Sin Tratar.

	1	2	3	4	5	6
1	Obs ID (Primary)	Obs ID ($ClassID)	M9.YVarPS($M9.DA(Oxif))	M9.YPredPS[1]($M9.DA(Oxif))	M9.YVarPS($M9.DA(SinT))	M9.YPredPS[1]($M9.DA(SinT))
2	Oxifluorfen 2_785nm50x100p10s_1	Oxif	1	0,824559	0	0,175441
3	Oxifluorfen 5_785nm50x100p10s_11	Oxif	1	0,782988	0	0,217012
4	Sin tratar_785nm50x100p10s_13	SinT	0	0,208463	1	0,791537
5	Oxifluorfen 1_785nm50x100p10s_22	Oxif	1	1,2315	0	-0,231499
6	Oxifluorfen 3_785nm50x100p10s_2	Oxif	1	0,871685	0	0,128315
7	Oxifluorfen 5_785nm50x100p10s_22	Oxif	1	0,80733	0	0,19267
8	Sin tratar_785nm50x100p10s_22	SinT	0	0,652617	1	0,347383

Ilustración 49: lista de clasificación del modelo OPLS-DA Oxi vs Sin Tratar.

En este caso, el modelo parece diferenciar bien ambas clases, pero a la hora de clasificar falla de nuevo en una de las muestras no tratadas. Sin embargo, en esta ocasión la clasificación la hace con un peso ligeramente mayor que en anteriores casos, para bien y para mal.

6.6.5. Tebuconazol

En cuanto al modelo OPLS-DA generado con muestras de Tebuconazol y no tratadas, se obtienen los siguientes resultados observables en las ilustraciones de la 50 a la 54.

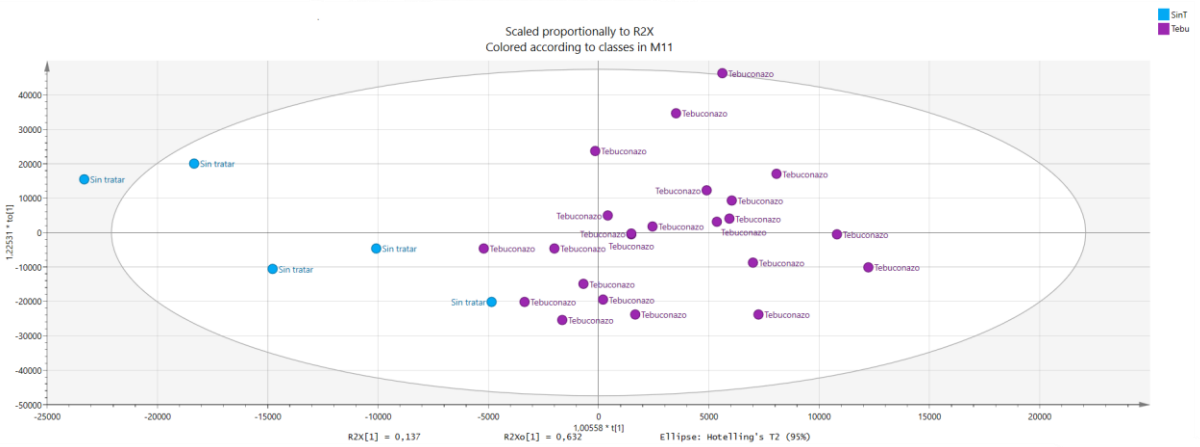


Ilustración 50: *score plot* del modelo OPLS-DA Tebu vs Sin Tratar.

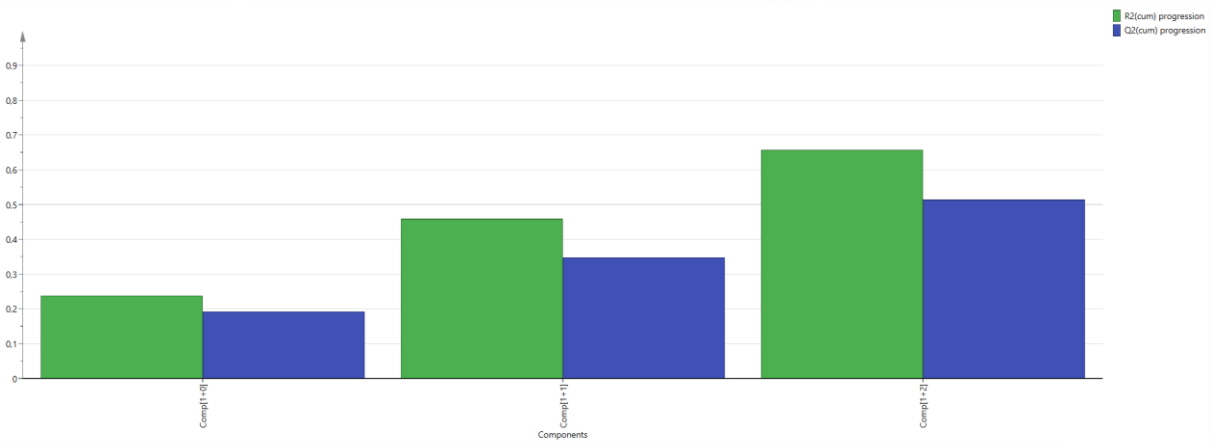


Ilustración 51: *summary of fit* del modelo OPLS-DA Tebu vs Sin Tratar.

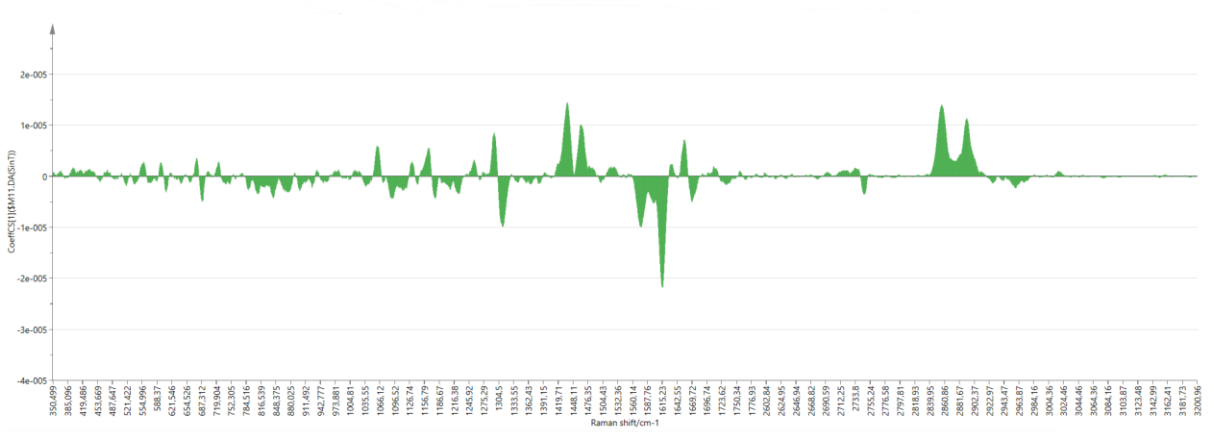


Ilustración 52: *coefficients* del modelo OPLS-DA Tebu vs Sin Tratar.

	1	2	3	4	5	6
1		Members	Correct	SinT	Tebu	No class (YPred <= 0)
2	SinT	2	100%	2	0	0
3	Tebu	6	100%	0	6	0
4	No class	0		0	0	0
5	Total	8	100%	2	6	0
6	Fisher's prob.	0,036				

Ilustración 53: *misclassification table* del modelo OPLS-DA Tebu vs Sin Tratar.

	1	2	3	4	5	6
1	Obs ID (Primary)	Obs ID ($ClassID)	M11.YVarPS($M11.DA(SinT))	M11.YPredPS[1]($M11.DA(SinT))	M11.YVarPS($M11.DA(Tebu))	M11.YPredPS[1]($M11.DA(Tebu))
2	Sin tratar_785nm50x100p10s_11	SinT	1	0,760715	0	0,239285
3	Tebuconazol 3_785nm50x100p10s_12	Tebu	0	0,282074	1	0,717926
4	Tebuconazol 4_785nm50x100p10s_1	Tebu	0	-0,211706	1	1,21171
5	Tebuconazol 4_785nm50x100p10s_12	Tebu	0	-0,0856026	1	1,0856
6	Sin tratar_785nm50x100p10s_2	SinT	1	0,59158	0	0,40842
7	Tebuconazol 1_785nm50x100p10s_22	Tebu	0	-0,13511	1	1,13511
8	Tebuconazol 2_785nm50x100p10s_2	Tebu	0	0,0772945	1	0,922706
9	Tebuconazol 5_785nm50x100p10s_2	Tebu	0	0,129264	1	0,870736

Ilustración 54: lista de clasificación del modelo OPLS-DA Tebu vs Sin Tratar.

En este caso, La clasificación vuelve a ser certera como en el caso de Lambda Cihalometrina. Pero, al igual que en ese caso, el peso de algunas muestras vuelve a ser bajo para determinarlo como una clasificación precisa para todos los casos futuros. Por lo tanto, se vuelve a destacar la importancia de contar con una mayor cantidad y mayor equidad entre las muestras.

6.7. Resumen de resultados

Para tener una visión general de los resultados obtenidos existen parámetros calculables que los resume de manera efectiva, de manera que se puede percibir la precisión de los modelos fijándose solamente en estos tres parámetros. Los parámetros son los siguientes.

- Sensibilidad o razón de verdaderos positivos:

$$VPR = \frac{Verdadero\ Positivo\ (VP)}{Positivos\ totales\ (P)} = \frac{VP}{VP + Falso\ Negativo\ (FN)} \quad 6.1)$$

- Especificidad:

$$Especificidad\ (SPC) = \frac{Verdadero\ Negativo\ (VN)}{Negativos\ totales\ (N)} = \frac{VN}{Falsos\ Positivos\ (FP) + VN} \quad 6.2)$$

- Exactitud:

$$Exactitud\ (ACC) = \frac{VP + VN}{P + N} \quad 6.3)$$

Fitosanitario	Sensibilidad (%)	Especificidad (%)	Exactitud	Calidad
Deltametrina	100	50	83,33	Muy buena
Diflufenican	100	50	87,5	Muy buena
Diflufenican (menos 1 outlier)	100	50	87,5	Muy buena
Lambda Cihalometrina	100	100	100	Excelente
Oxifluorfen	100	50	85,71	Muy buena
Tebuconazol	100	100	100	Excelente

Tabla 2: tabla resumen de resultados**.**

Se considera una clasificación perfecta cuando la sensibilidad y la especificidad están ambas en un 100%. Por el contrario, una clasificación totalmente aleatoria es definida cuando ambos parámetros marcan el 50%. Por lo tanto, valores por debajo determinan que el modelo es incapaz de predecir. En este caso, como se puede observar en la tabla 2, a pesar de disponer de pocos datos y una falta de equidad entre las clases a comparar, se han obtenido unos resultados de muy buena calidad. Incluso hay 2 modelos que llegan a tener una calidad excelente, considerados como una clasificación perfecta, por lo que se puede considerar un desarrollo de proyecto exitoso. En la imagen 55 se puede apreciar el espacio ROC y las parcelas de predicción de los 6 modelos desarrollados.

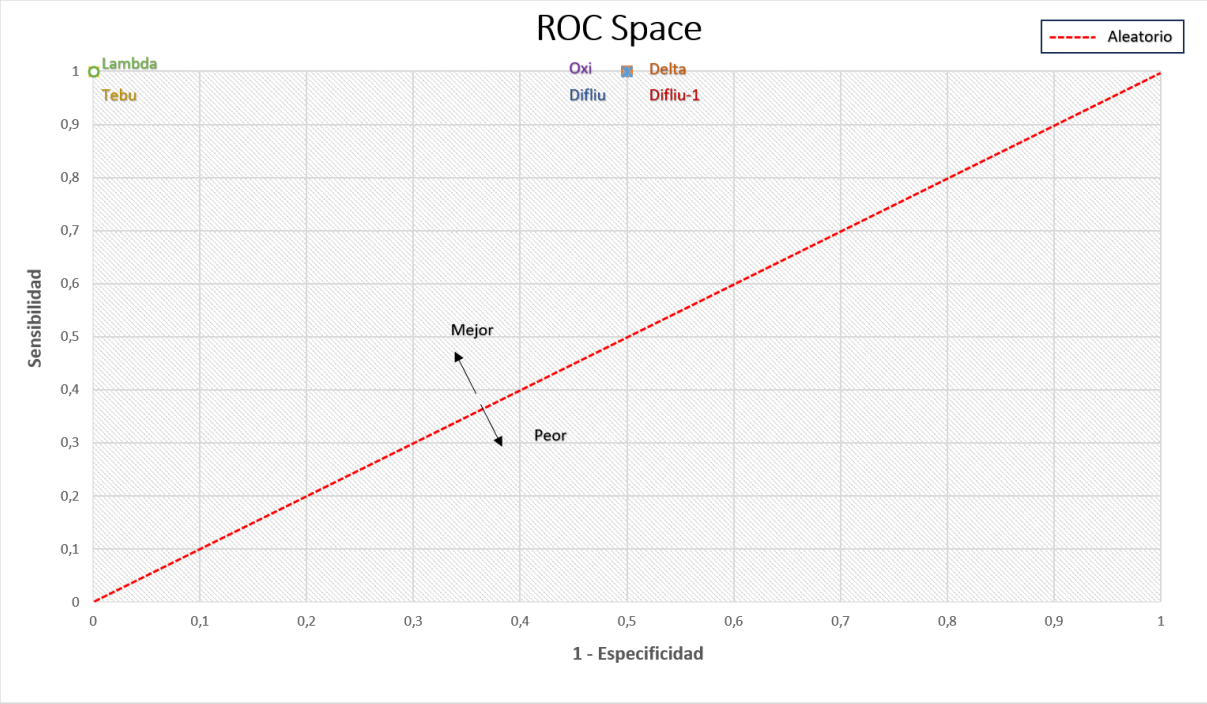


Ilustración 55: espacio ROC y las parcelas de predicción de los 6 modelos desarrollados

7. Metodología seguida en el desarrollo del trabajo

En este apartado se describen las tareas y fases que ha n sido necesarias para el desarrollo del proyecto. Es de vital importancia hacer una planificación previa al comienzo del desarrollo de cualquier proyecto, con el fin de desarrollar el proyecto de una forma ordenada, eficiente y controlada. Para ello, se describen a continuación los equipos de trabajo establecidos y los paquetes de tareas de cada equipo, así como los recursos utilizados durante el desarrollo. Al final de este apartado, se muestra el diagrama de Gantt referente al desarrollo del proyecto, con el fin de poder mostrar de forma global el trascurso del mismo.

7.1. Equipo de trabajo

En la siguiente tabla, se muestra el equipo de trabajo que ha participado en el desarrollo del proyecto.

Código	Nombre y apellidos	Cargo	Función
C1	Igor Ayesta	Director de proyecto	Propone el proyecto, define pasos a seguir en su desarrollo y se encarga del tutelaje del ingeniero a su cargo.
C2	Gorka Pérez	Ingeniero junior	Se encarga del desarrollo técnico del proyecto y de la escritura de la documentación.

Tabla 3: equipo de trabajo.

7.2. Paquetes de trabajo

En las siguientes tablas se muestran los paquetes de tareas definidos para un desarrollo competente del proyecto. En cada paquete se explican las tareas correspondientes a realizar, las fechas de inicio y finalización de la tarea y su duración.

7.2.1. PT1 Supervisión y administración del proyecto

En este paquete de trabajo se incluyen la planificación y coordinación de todo el proyecto. El director se encarga de asignar las tareas y de buscar soluciones a problemas enfrentados durante el desarrollo.

PT1	Descripción	Fecha inicio	Fecha fin	Duración
PT1.1	**Propuesta de proyecto:** se propone y se decide el proyecto a desarrollar.	9/10/23	9/10/23	1 día
H1	**Elección de proyecto:** el proyecto a desarrollar es elegido.	27/10/23	27/10/23	0 días
PT1.2	**Supervisión y administración del proyecto:** desarrollo del proyecto bajo supervisión del director de proyecto	27/10/23	27/3/24	109 días

Tabla 4: PT1 Supervisión y administración del proyecto.

7.2.2. PT2 Documentación en Espectroscopía Raman

Durante este paquete de trabajo se investiga en la documentación previa necesaria para el desarrollo del proyecto, esto incluye la documentación referente a la teoría sobre el funcionamiento de la espectroscopía Raman y la documentación sobre el manejo del microscopio Raman.

PT2	Descripción	Fecha inicio	Fecha fin	Duración
PT2.1	**Investigar acerca del funcionamiento teórico de la espectrometría Raman:** se investiga toda la documentación accesible sobre la tecnología de espectroscopía Raman.	27/10/23	3/11/23	6 días
PT2.2	**Conocer el funcionamiento del microscopio Raman:** familiarizarse con el funcionamiento del hardware de la espectroscopía Raman.	3/11/23	10/11/23	6 días

Tabla 5: PT2 Documentación en Espectroscopía Raman.

7.2.3. Analizar muestras recibidas con el Espectrómetro

Este paquete incluye todas las tareas necesarias durante el proceso de toma de medidas y análisis de muestras.

PT3	Descripción	Fecha inicio	Fecha fin	Duración
PT3.1	**Determinar el número de análisis por muestra y número de muestras a analizar:** se decide el cómputo total de análisis necesarios para disponer de una buena base de datos.	13/11/23	13/11/23	1 día
PT3.2	**Selección de muestras óptimas para su posterior análisis:** se descartan muestras que han sido degradadas por el paso del tiempo y el moho.	13/11/23	13/11/23	1 día
PT3.3	**Análisis de todas muestras:** se analizan las todas las muestras que se han definido en la subtarea anterior.	13/11/23	17/11/23	5 días
H2	**Tanda de muestras finalizada:** se han obtenido los datos necesarios para continuar con el desarrollo del proyecto.	17/11/23	17/11/23	0 días

Tabla 6: PT3 Analizar muestras recibidas con el Espectrómetro.

7.2.4. Instalar e investigar el software a utilizar

En esta tarea se incluyen la obtención del software de análisis multivariante a utilizar y la familiarización con su interfaz y herramientas.

PT4	Descripción	Fecha inicio	Fecha fin	Duración
PT4.1	**Obtener licencia e instalar el software a utilizar:** se logra obtener una licencia válida para poder usar dicho software	20/11/23	22/11/23	3 días
PT4.2	**Investigar el funcionamiento de la interfaz y herramientas del software:** se familiariza con el uso de la interfaz y herramientas de las que dispone el software.	23/11/23	19/11/23	5 días

Tabla 7: PT4 Instalar e investigar el software a utilizar.

7.2.5. Estudiar el funcionamiento del Análisis Multivariante de Datos

Durante el desarrollo de esta tarea se centra en buscar y analizar información teórica referente a cada modelo con el objetivo de entender mejor su funcionamiento y generación.

PT5	Descripción	Fecha inicio	Fecha fin	Duración
PT5.1	**Entender el funcionamiento del modelo PCA:** se busca y se investiga la documentación sobre la teoría de la generación de modelos PCA.	30/11/23	1/12/23	2 días
PT5.2	**Entender el funcionamiento del modelo PLS:** se busca y se investiga la documentación sobre la teoría de la generación de modelos PLS.	5/12/23	6/12/23	2 días
PT5.3	**Entender el funcionamiento del modelo PLS-DA:** se busca y se investiga la documentación sobre la teoría de la generación de modelos PLS-DA.	7/12/23	8/12/23	2 días
PT5.4	**Entender las diferencias del modelo PLS-DA y OPLS-DA:** se analiza la documentación teórica sobre las diferencias entre los modelos PLS-DA y OPLS-DA.	11/12/23	11/12/23	1 día

Tabla 8: PT5 Estudiar el funcionamiento del Análisis Multivariante de Datos.

7.2.6. Preprocesado de los datos

Esta tarea se centra en la mejora de los datos obtenidos mediante el análisis para una mejor interpretación por parte del software y sus herramientas.

PT6	Descripción	Fecha inicio	Fecha fin	Duración
PT6.1	**Adaptar los datos al formato especificado por el Software:** se adaptan los datos al formato necesario para un correcto uso de los mismos por el software.	12/12/23	12/12/23	1 día
PT6.2	**Exclusión de datos no determinantes:** se determinan datos considerados fuera de la norma y se excluyen del conjunto de datos utilizado.	12/12/23	14/12/23	3 días
H3	**Preprocesado de los datos realizado:** se ha completado el preprocesado de los datos.	14/12/23	14/12/23	0 días

Tabla 9: PT6 Preprocesado de los datos.

7.2.7. Realización de modelos de prueba

Durante esta tarea, se desarrollan varios modelos en los que se pone a prueba lo aprendido en cuanto a la teoría y se familiariza con la gestión de los datos y los resultados obtenidos.

PT7	Descripción	Fecha inicio	Fecha fin	Duración
PT7.1	**Realización de modelos PCA globales:** se realizan varios modelos PCA con el fin de poner a prueba los datos obtenidos de las muestras y corroborar con la teoría con los resultados conseguidos.	15/12/23	19/12/23	3 días
PT7.2	**Realización de modelos PLS globales:** se realizan varios modelos PLS con el fin de poner a prueba los datos obtenidos de las muestras y corroborar con la teoría con los resultados conseguidos.	20/12/23	25/12/23	4 días
PT7.3	**Realización de modelos discriminantes:** se realizan varios modelos de análisis discriminante (DA) con el fin de poner a prueba los datos obtenidos de las muestras y corroborar con la teoría con los resultados conseguidos.	8/1/24	9/1/24	2 días
PT7.4	**Realización de modelos PLS-DA más específicos:** se realizan modelos PLS-DA más concretos con el fin de acercarse a la obtención de modelos más precisos.	10/1/24	11/1/24	2 días
PT7.5	**Realización de modelos OPLS-DA específicos:** se realizan modelos OPLS-DA más concretos con el fin de acercarse a la obtención de modelos más precisos.	12/1/24	15/1/24	2 días

Tabla 10: PT7 Realización de modelos de prueba.

7.2.8. Realización de modelos determinantes

Esta tarea incluye la realización y obtención de modelos más precisos cercanos a los resultados finales.

PT8	Descripción	Fecha inicio	Fecha fin	Duración
PT8.1	**Realización de modelos OPLS-DA finales:** se realizan modelos cercanos a la solución final.	15/1/24	18/1/24	4 días
H4	**modelos finales creados:** se logran crear 5 modelos OPLS-DA con clasificación óptima.	18/1/24	18/1/24	0 días

Tabla 11: PT8 Realización de modelos determinantes.

7.2.9. Comprobación de efectividad de los modelos

Durante el desarrollo de esta tarea, se ponen a prueba los 5 modelos OPLS-DA generados en la tarea anterior en busca de mejorar y obtener una precisión alta de predicción.

PT9	Descripción	Fecha inicio	Fecha fin	Duración
PT9.1	**Comprobar capacidad predictiva del modelo 1:** se comprueba la capacidad predictiva del modelo según el porcentaje de acierto obtenido durante la clasificación de nuevos datos.	19/1/24	19/1/24	1 día
PT9.2	**Adaptar modelo 1 según datos recopilados:** se adapta el modelo con el fin de mejorar dicho porcentaje de predicción y lograr un porcentaje de acierto aceptable.	22/1/24	22/1/24	1 día
PT9.3	**Comprobar capacidad predictiva del modelo 2:** se comprueba la capacidad predictiva del modelo según el porcentaje de acierto obtenido durante la clasificación de nuevos datos.	23/1/24	23/1/24	1 día
PT9.4	**Adaptar modelo 2 según datos recopilados:** se adapta el modelo con el fin de mejorar dicho porcentaje de predicción y lograr un porcentaje de acierto aceptable.	24/1/24	24/1/24	1 día
PT9.5	**Comprobar capacidad predictiva del modelo 3:** se comprueba la capacidad predictiva del modelo según el porcentaje de acierto obtenido durante la clasificación de nuevos datos.	25/1/24	25/1/24	1 día
PT9.6	**Adaptar modelo 3 según datos recopilados:** se adapta el modelo con el fin de mejorar dicho porcentaje de predicción y lograr un porcentaje de acierto aceptable.	26/1/24	26/1/24	1 día
PT9.7	**Comprobar capacidad predictiva del modelo 4:** se comprueba la capacidad predictiva del modelo según el porcentaje de acierto obtenido durante la clasificación de nuevos datos.	29/1/24	29/1/24	1 día
PT9.8	**Adaptar modelo 4 según datos recopilados:** se adapta el modelo con el fin de mejorar dicho porcentaje de predicción y lograr un porcentaje de acierto aceptable.	30/1/24	30/1/24	1 día
PT9.9	**Comprobar capacidad predictiva del modelo 5:** se comprueba la capacidad predictiva del modelo según el porcentaje de acierto obtenido durante la clasificación de nuevos datos.	31/1/24	31/1/24	1 día
PT9.10	**Adaptar modelo 5 según datos recopilados:** se adapta el modelo con el fin de mejorar dicho porcentaje de predicción y lograr un porcentaje de acierto aceptable.	1/2/24	1/2/24	1 día

Tabla 12: PT9 Comprobación de efectividad de los modelos.

7.2.10. Obtención de resultados óptimos

Esta tarea incluye últimos ajustes en los modelos y obtención de resultados considerados óptimos.

PT10	Descripción	Fecha inicio	Fecha fin	Duración
PT10.1	**Realizar últimas pruebas para conseguir resultados óptimos:** se realizan últimas comprobaciones para corroborar que los modelos logran resultados considerados lo suficientemente óptimos.	2/2/24	5/2/24	2 días
H5	**Resultados óptimos obtenidos:** se consiguen resultados óptimos en todos los modelos generados y se da por terminado el desarrollo del práctico del proyecto.	5/2/24	5/2/24	0 días

Tabla 13: PT10 Obtención de resultados óptimos.

7.2.11. Documentación y entrega del proyecto

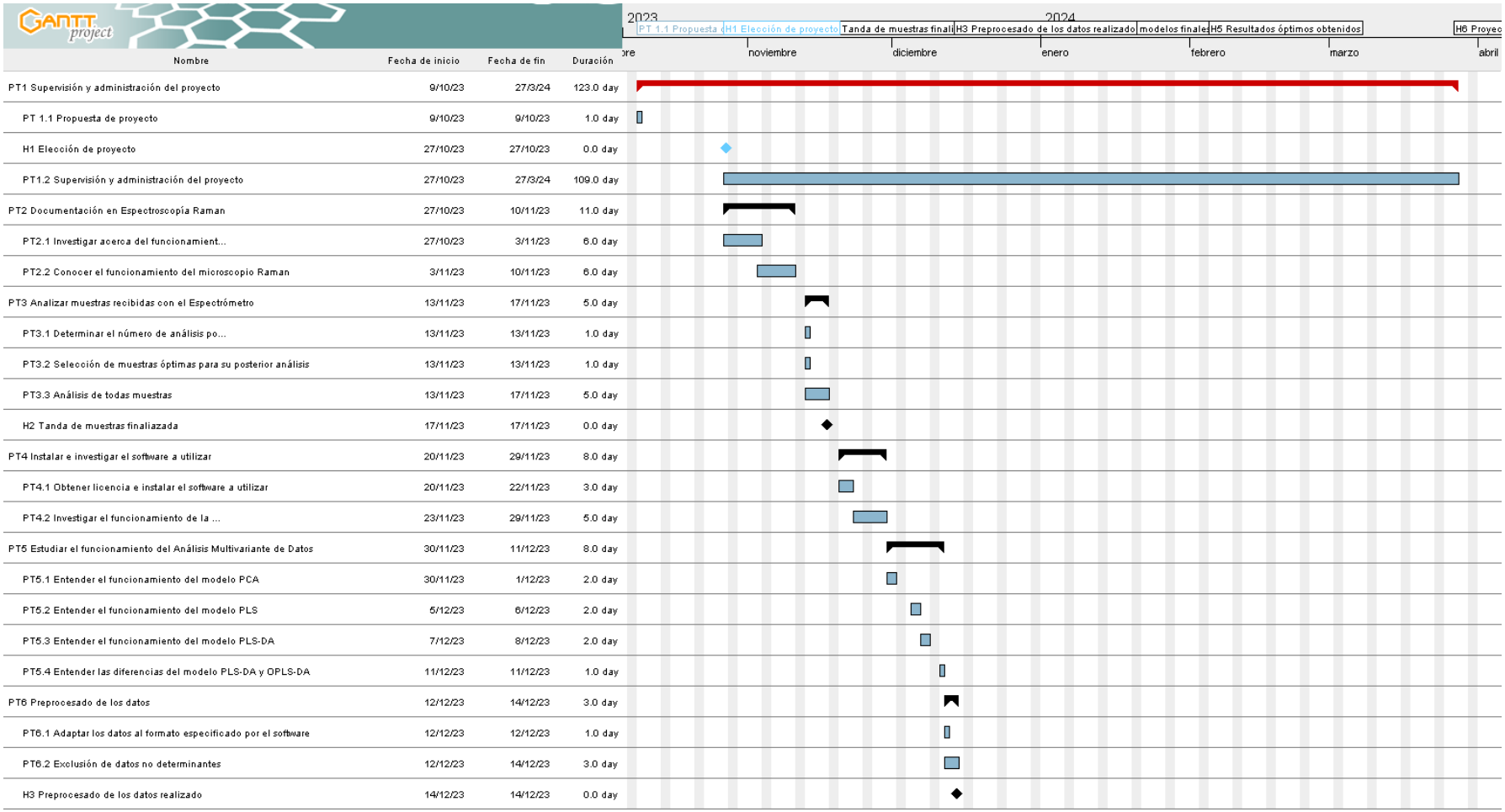
Esta tarea incluye la elaboración de toda la documentación referente al proyecto, su entrega y la posterior presentación oral ante un jurado.

PT11	Descripción	Fecha inicio	Fecha fin	Duración
PT11.1	**Documentación:** se elabora la documentación que engloba el desarrollo completo del proyecto y se entrega para su revisión y evaluación.	5/2/24	1/3/24	20 días
PT11.2	**Presentación oral del proyecto:** se prepara la presentación oral del proyecto y se presenta delante de un jurado	4/3/24	27/3/24	3 días
H6	**Proyecto finalizado:** el proyecto ha finalizado con éxito.	27/3/24	27/3/24	0 días

Tabla 14: PT11 Documentación y entrega del proyecto.

7.3. Diagrama de Gantt/Cronograma

Con el fin de tener una visión más global y esquemática del proyecto, se ha hecho uso del diagrama de Gantt en el cual se visualizan las tareas realizadas y los hitos conseguidos. Se ha separado en 2 imágenes para poder tener mejor visión de los detalles. Estas son la ilustración 55 y 56.



Nombre	Fecha de inicio	Fecha de fin	Duración
PT1 Supervisión y administración del proyecto	9/10/23	27/3/24	123.0 day
PT 1.1 Propuesta de proyecto	9/10/23	9/10/23	1.0 day
H1 Elección de proyecto	27/10/23	27/10/23	0.0 day
PT1.2 Supervisión y administración del proyecto	27/10/23	27/3/24	109.0 day
PT2 Documentación en Espectroscopía Raman	27/10/23	10/11/23	11.0 day
PT2.1 Investigar acerca del funcionament...	27/10/23	3/11/23	6.0 day
PT2.2 Conocer el funcionamiento del microscopio Raman	3/11/23	10/11/23	6.0 day
PT3 Analizar muestras recibidas con el Espectrómetro	13/11/23	17/11/23	5.0 day
PT3.1 Determinar el número de análisis po...	13/11/23	13/11/23	1.0 day
PT3.2 Selección de muestras óptimas para su posterior análisis	13/11/23	13/11/23	1.0 day
PT3.3 Análisis de todas muestras	13/11/23	17/11/23	5.0 day
H2 Tanda de muestras finaliazada	17/11/23	17/11/23	0.0 day
PT4 Instalar e investigar el software a utilizar	20/11/23	29/11/23	8.0 day
PT4.1 Obtener licencia e instalar el software a utilizar	20/11/23	22/11/23	3.0 day
PT4.2 Investigar el funcionamiento de la ...	23/11/23	29/11/23	5.0 day
PT5 Estudiar el funcionamiento del Análisis Multivariante de Datos	30/11/23	11/12/23	8.0 day
PT5.1 Entender el funcionamiento del modelo PCA	30/11/23	1/12/23	2.0 day
PT5.2 Entender el funcionamiento del modelo PLS	5/12/23	6/12/23	2.0 day
PT5.3 Entender el funcionamiento del modelo PLS-DA	7/12/23	8/12/23	2.0 day
PT5.4 Entender las diferencias del modelo PLS-DA y OPLS-DA	11/12/23	11/12/23	1.0 day
PT6 Preprocesado de los datos	12/12/23	14/12/23	3.0 day
PT6.1 Adaptar los datos al formato especificado por el software	12/12/23	12/12/23	1.0 day
PT6.2 Exclusión de datos no determinantes	12/12/23	14/12/23	3.0 day
H3 Preprocesado de los datos realizado	14/12/23	14/12/23	0.0 day

Ilustración 56: Diagrama de Gantt (parte 1).

PT7 Realización de modelos de prueba	15/12/23	15/1/24	22.0 day
PT7.1 Realización de modelos PCA globales	15/12/23	19/12/23	3.0 day
PT7.2 Realización de modelos PLS globales	20/12/23	25/12/23	4.0 day
PT7.3 Realización de modelos discriminantes	8/1/24	9/1/24	2.0 day
PT17.4 Realización de modelos PLS-DA más específicos	10/1/24	11/1/24	2.0 day
PT17.5 Realización de modelos OPLS-DA específicos	12/1/24	15/1/24	2.0 day
PT8 Realización de modelos determinantes	15/1/24	18/1/24	4.0 day
PT8.1 Realización de modelos OPLS-DA finales	15/1/24	18/1/24	4.0 day
H4 modelos finales creados	18/1/24	18/1/24	0.0 day
PT9 Comprobación de efectividad de los modelos	19/1/24	1/2/24	10.0 day
PT9.1 Comprobar capacidad predictiva del modelo 1	19/1/24	19/1/24	1.0 day
PT9.2 Adaptar modelo 1 según datos recopilados	22/1/24	22/1/24	1.0 day
PT9.3 Comprobar capacidad predictiva del modelo 2	23/1/24	23/1/24	1.0 day
PT9.4 Adaptar modelo 2 según datos recopilados	24/1/24	24/1/24	1.0 day
PT9.5 Comprobar capacidad predictiva del modelo 3	25/1/24	25/1/24	1.0 day
PT9.6 Adaptar modelo 3 según datos recopilados	26/1/24	26/1/24	1.0 day
PT9.7 Comprobar capacidad predictiva del modelo 4	29/1/24	29/1/24	1.0 day
PT9.8 Adaptar modelo 4 según datos recopilados	30/1/24	30/1/24	1.0 day
PT 9.9 Comprobar capacidad predictiva del modelo 5	31/1/24	31/1/24	1.0 day
PT9.10 Adaptar modelo 5 según datos recopilados	1/2/24	1/2/24	1.0 day
PT10 Obtención de resultados óptimos	2/2/24	5/2/24	2.0 day
PT10.1 Realizar últimas pruebas para conseguir resultados óptimos	2/2/24	5/2/24	2.0 day
H5 Resultados óptimos obtenidos	5/2/24	5/2/24	0.0 day
PT11 Documentación y entrega del proyecto	5/2/24	27/3/24	38.0 day
PT11.1 Documentación	5/2/24	1/3/24	20.0 day
PT11.2 Presentación oral del proyecto	4/3/24	27/3/24	18.0 day
H6 Proyecto finalizado	27/3/24	27/3/24	0.0 day

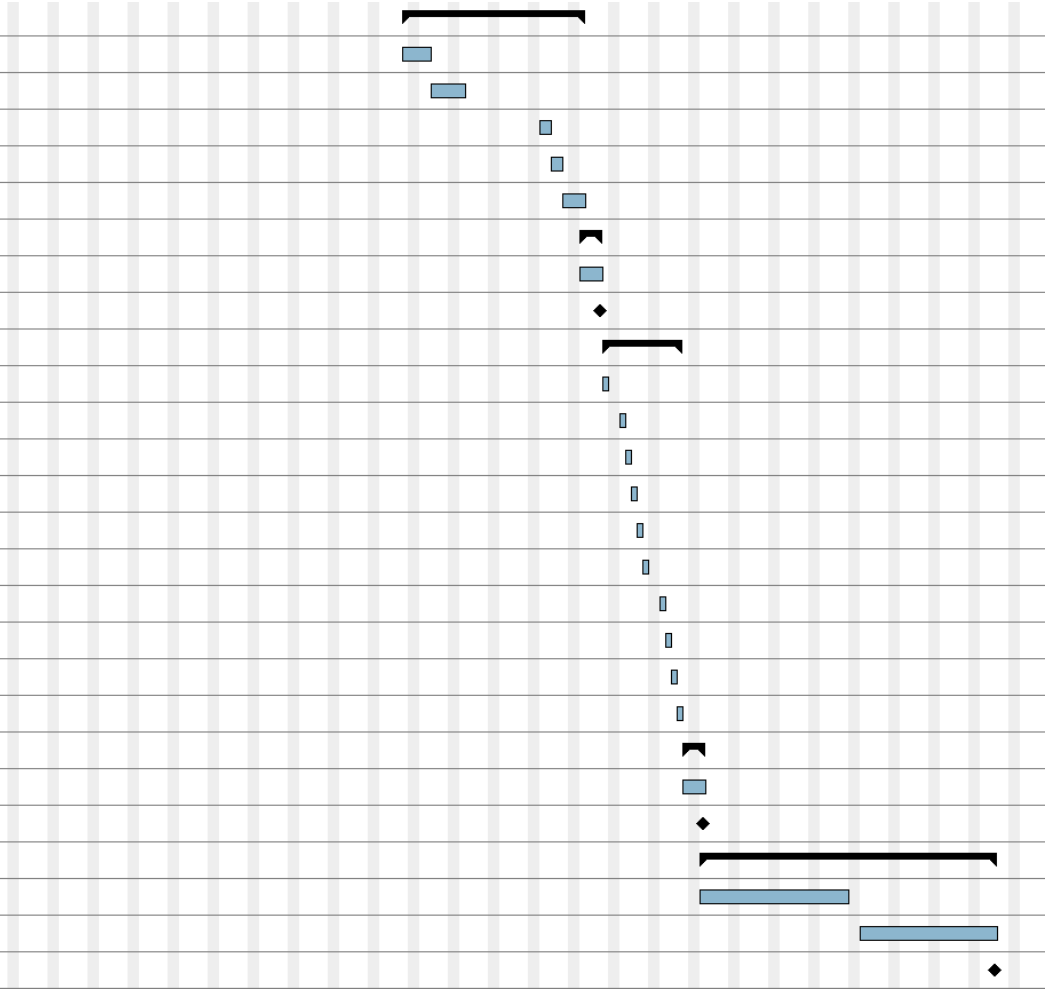


Ilustración 57: Diagrama de Gantt (parte 2).

8. Aspectos económicos

En este apartado se desglosa el coste del proyecto, el cual se hace una vez acabado el proyecto y se conocen los costes necesarios para el desarrollo del mismo. Dichos costes se han diferenciado en horas internas, amortizaciones y gastos.

8.1. Horas internas

Las horas internas se muestran en la siguiente tabla:

Horas internas			
Trabajador	**Horas empleadas**	**Coste/h**	**Coste total**
Ingeniero senior	74 h	50 €	3.700 €
Ingeniero junior	492 h	35 €	17.220 €
Coste total de las horas internas			20.920 €

Tabla 15: costes de las horas internas.

8.2. Amortizaciones

En este apartado, se muestran los recursos utilizados para el desarrollo del proyecto. Estos aparecen listados en la siguiente tabla.

Amortizaciones					
Producto	**Coste/Unidad**	**Cantidad**	**Vida útil**	**Uso**	**Coste total**
Ordenador	700 €	2	60 meses	4 meses	93,33 €
Licencia SIMCA	150 €	2	12 meses	4 meses	100 €
Espectrómetro Renishaw Invia Raman	84700 €	1	120 meses	4 meses	2.823,33 €
Costo total de las amortizaciones					3.016,66 €

Tabla 16: amortizaciones.

8.3. Gastos

En este apartado se muestran los costes de los recursos los cuales no pueden ser aprovechados una vez acaba el proyecto, por lo que son considerados gastos.

Gastos	
Concepto	**Coste**
Factura de energía	100 €
Material de laboratorio	40 €
Disco duro	25 €
Coste total	165 €

Tabla 17: gastos.

8.4. Presupuesto total

Para obtener el presupuesto total del proyecto se deben sumar los costes totales de los apartados que conforman el presupuesto. Dichos apartados se muestran en la siguiente tabla.

Presupuesto total	
Concepto	**Coste**
Horas internas	20.920 €
Amortización	3.016,66 €
Gastos	165 €
Subtotal1	24.101,66 €
Gastos indirectos	4%
Subtotal 2	25.065,72 €
Imprevistos	5%
Presupuesto total	26.319 €

Tabla 18: presupuesto total.

BIBLIOGRAFÍA

[1] Comisión Europea, "Regulation (EC) no 1333/2008 of the european parliament and of the council of 16 december 2008 on food additives (text with EEA relevance)," 2023.

[2] 20MINUTOS.ES, "Las claves del escándalo de la carne de caballo: qué productos contiene y cómo es el etiquetado," 2013. Available: https://www.20minutos.es/noticia/1736928/0/claves-hallazgo/carne-caballos/productos-europa/

[3] BBC Mundo, "El escándalo de los huevos contaminados con pesticida: la alerta alimentaria que afecta a 17 países y millones de consumidores en Europa y Asia," 2017.
Available: https://www.bbc.com/mundo/noticias-internacional-40893413

[4] elmundo.es, "La contaminación de leche en polvo en China afecta ya a 69 marcas diferentes," 2008.
Available: https://www.elmundo.es/elmundosalud/2008/09/16/medicina/1221575585.html

[5] R. Michael T., "International and national regulatory strategies to counter food fraud," School of Law Resnick Center for Food Law & Policy, 2022.

[6] (s.f.). *Aceite de oliva y aceituna de mesa*.
Available: https://www.mapa.gob.es/es/agricultura/temas/producciones-agricolas/aceite-oliva-y-aceituna-mesa/aceite.aspx

[7] (s.f.). *Registro de Productos Fitosanitarios*.
Available: https://www.mapa.gob.es/es/agricultura/temas/sanidad-vegetal/productos-fitosanitarios/registro-productos/.

[8] E. Nº *et al*, "Importaciones paralelas de productos fitosanitarios en españa," 2024.

[9] J. C. Ramos, A. E. Villanueva and C. M. Ortiz Lima, "Raman spectroscopy and its applications," *Opt. Pura Apl.*, vol. 46, *(1)*, pp. 83, 2013. DOI: 10.7149/opa.46.1.83.

[10] (s.f.). *OPLS vs PCA: Explaining Differences or Grouping Data?*.
Available: https://www.sartorius.com/en/knowledge/science-snippets/explaining-differences-or-grouping-data-opls-da-vs-pca-data-analysis-507204%20

[11] M. Azkune Ulla, "Design and Development of Polymer Optical Fiber Based Platforms for Glucose Detection." 2019.

[12] (s.f.). *Espectroscopía Raman*.
Available: https://www.mt.com/es/es/home/applications/L1_AutoChem_Applications/Raman-Spectroscopy.html%20

[13] (s.f.). *What Is Principal Component Analysis (PCA) and How It Is Used?*.
Available: https://www.sartorius.com/en/knowledge/science-snippets/what-is-principal-component-analysis-pca-and-how-it-is-used-507186%20

[14] (s.f.). *OPLS vs. PLS Modeling to Improve Bioprocess Yields of Batch Processes*.
Available: https://www.sartorius.com/en/knowledge/science-snippets/opls-vs-pls-modeling-to-improve-bioprocess-yields-of-batch-processes-602762%20

[15] (s.f.). *UNSCRAMBLER*. Available: https://www.tecnilab.es/analisis-multivariante/#1557202990039-4884034d-9a64

[16] (s.f.). *Multivariate Data Analysis Software That Turns Data Into Growth*. Available: https://www.sartorius.com/en/products/process-analytical-technology/data-analytics-software/mvda-software/simca.

[17] (s.f.). *PLS_toolbox*. Available: https://es.mathworks.com/products/connections/product_detail/pls-toolbox.html%20

[18] (). *inVia™ confocal Raman microscope*. Available: https://www.renishaw.com/en/invia-confocal-raman-microscope--6260%20

[19] K. H. Liland, "4S Peak Filling – baseline estimation by iterative mean suppression," 2015. Available: http://hdl.handle.net/11250/2576468. DOI: 10.1016/j.mex.2015.02.009.

[20] RF Fernández, AD Morales and EL Dennes, "Desarrollo De Un Modelo SIMCA De Reconocimiento De Patrones Para La Clasificación De Combustible Diesel." , Universidad de Oriente, Santiago de Cuba, Cuba, 2009.