



GRADO EN CIENCIA DE DATOS



VNIVERSITAT
DE VALÈNCIA

TRABAJO FINAL DE GRADO

ESTIMACIÓN MEDIANTE ANALISIS NO
DESTRUCTIVO DE LOS NIVELES
NUTRICIONALES DE HOJAS DE
CÍTRICOS APLICANDO TECNICAS DE
MACHINE I DEEP LEARNING

AUTOR: GUILLEM MIRALLES GADEA
TUTOR: JUAN GÓMEZ SANCHIS
JULIO 2023



TRABAJO FINAL DE GRADO

ESTIMACIÓN MEDIANTE ANALISIS NO
DESTRUCTIVO DE LOS NIVELES
NUTRICIONALES DE HOJAS DE
CÍTRICOS APLICANDO TECNICAS DE
MACHINE I DEEP LEARNING

AUTOR: GUILLEM MIRALLES GADEA
TUTOR: JUAN GÓMEZ SANCHIS
JULIO 2023

Declaración de autoría:

Yo, Guillem Miralles Gadea, declaro la autoría del Trabajo Final de Grado titulado " Estimación Mediante Análisis No Destructivo De Los Niveles Nutricionales De Hojas De Cítricos Aplicando Técnicas De Machine Y Deep Learning " y que el citado trabajo no infringe las leyes vigentes sobre propiedad intelectual. El material no original que figura en este trabajo ha sido atribuido a sus legítimos autores.

Fdo: Guillem Miralles Gadea

Resum:

Una de les característiques que tots els éssers vius comparteixen és que una nutrició adequada té un impacte positiu en la nostra salut. En el cas dels arbres, i més específicament en el cas dels arbres cítrics, una nutrició adequada és essencial perquè l'arbre siga resistent a plagues i malalties, produïsca aliments de la màxima qualitat i quantitat possible, i creixà de manera saludable. Per tant, una nutrició adequada és una qüestió clau per a qualsevol agricultor.

Actualment, això s'aconsegueix principalment mitjançant l'ús d'adobs o fems. No obstant això, l'ús excessiu d'aquests productes pot ser perjudicial per al medi ambient i representar un malbaratament de recursos per als agricultors. En altres paraules, no és fàcil conéixer la quantitat i el tipus d'adob que cada arbre necessita. Per a fer-ho, és necessari identificar els nutrients que falten i en quina mesura. A hores d'ara, per determinar amb precisió els nivells nutricionals d'un arbre, s'utilitzen anàlisi destructiva de les fulles o del sòl, el que representa un cost econòmic i temporal significatiu.

En aquest treball, s'han usat imatges hiperespectrals de visió propera i infraroja (Vis-NIR) i models de regressió d'aprenentatge automàtic per a estimar macronutrients primaris (N, P i K), macronutrients secundaris (Ca, Mg, S) i micronutrients (Na, Fe, Mn, Zn, Cu, B i Mo). La metodologia ha implicat l'aplicació de diversos models de regressió d'aprenentatge automàtic i tècniques, tant de preprocessament de dades com de reducció de la dimensionalitat per a determinar la combinació òptima.

Els resultats han estat particularment satisfactoris en l'estimació dels macronutrients, especialment en el cas del nitrogen (N), i també del potassi (K) i del sofre (S). Els micronutrients han presentat un major repte, encara que s'han assolit resultats prometedors en el cas del manganés (Mn), coure (Cu) i bor (B).

Resumen:

Una de las características que todos los seres vivos comparten es que una nutrición adecuada tiene un impacto positivo en nuestra salud. En el caso de los árboles, y más específicamente en el caso de los árboles cítricos, una nutrición adecuada es esencial para que el árbol sea resistente a plagas y enfermedades, produzca alimentos de la máxima calidad y cantidad posible, y crezca de manera saludable. Por lo tanto, una nutrición adecuada es una cuestión clave para cualquier agricultor.

Actualmente, esto se logra principalmente mediante el uso de fertilizantes o abono. Sin embargo, el uso excesivo de estos productos puede ser perjudicial para el medio ambiente y representar un desperdicio de recursos para los agricultores. En otras palabras, no es fácil conocer la cantidad y el tipo de fertilizante que cada árbol necesita. Para hacerlo, es necesario identificar los nutrientes que faltan y en qué medida. En la actualidad, para determinar con precisión los niveles nutricionales de un árbol, se utilizan análisis destructivos de las hojas o del suelo, lo que representa un costo económico y temporal significativo.

En este trabajo, se han utilizado imágenes hiperespectrales de visión cercana e infrarroja (Vis-NIR) y modelos de regresión de aprendizaje automático para estimar macronutrientes primarios (N, P y K), macronutrientes secundarios (Ca, Mg, S) y micronutrientes (Na, Fe, Mn, Zn, Cu, B y Mo). La metodología ha implicado la aplicación de varios modelos de regresión de aprendizaje automático y técnicas de preprocesamiento de datos y reducción de dimensionalidad para determinar la combinación óptima.

Los resultados han sido particularmente satisfactorios en la estimación de los macronutrientes, especialmente en el caso del nitrógeno (N), y también del potasio (K) y azufre (S). Los micronutrientes han presentado un mayor desafío, aunque se han logrado resultados prometedores en el caso del manganeso (Mn), cobre (Cu) y boro (B).

Abstract:

One of the characteristics that all living beings share is that adequate nutrition has a positive impact on our health. In the case of trees, and more specifically in the case of citrus trees, adequate nutrition is essential for the tree to be resistant to pests and diseases, produce food of the highest possible quality and quantity, and grow in a healthy way. Therefore, adequate nutrition is a key issue for any farmer.

Currently, this is mainly achieved through the use of fertilizers or manure. However, the excessive use of these products can be harmful to the environment and represent a waste of resources for farmers. In other words, it is not easy to know the amount and type of fertilizer that each tree needs. To do this, it is necessary to identify the nutrients that are missing and to what extent. Currently, to accurately determine the nutritional levels of a tree, destructive analyses of the leaves or soil are used, which represents a significant economic and temporal cost.

In this study, hyperspectral images of near-infrared and visible light (Vis-NIR) and machine learning regression models were used to estimate primary macronutrients (N, P, and K), secondary macronutrients (Ca, Mg, S), and micronutrients (Na, Fe, Mn, Zn, Cu, B, and Mo). The methodology involved the application of various machine learning regression models and data preprocessing and dimensionality reduction techniques to determine the optimal combination.

The results have been particularly satisfactory in estimating macronutrients, especially in the case of nitrogen (N), as well as potassium (K) and sulfur (S). Micronutrients have presented a greater challenge, although promising results have been achieved for manganese (Mn), copper (Cu), and boron (B).

Agradecimientos

Quiero expresar mi inmensa gratitud hacia mis queridos padres, que han sido la luz que ha guiado mi vida hasta ahora. Sin su inagotable amor, su apoyo incondicional y su confianza incansable en mí, yo no sería la persona que soy hoy. También a mi hermana, que ha sido una gran fuente de inspiración y apoyo en ese camino.

Quiero dar las gracias a todo el personal del Instituto Valenciano de Investigaciones Agrarias. Ha sido un placer trabajar con profesionales tan dedicados y apasionados y poder aprender de su experiencia.

Por último, pero no menos importante, quiero agradecer a mis profesores y compañeros de clase, especialmente a mi tutor Juan Gómez. Gracias por compartir conmigo su experiencia y conocimientos, los cuales han sido cruciales para mi crecimiento personal y académico.

Gracias a todos.

ÍNDICE GENERAL

1. INTRODUCCIÓN.....	18
2. OBJECTIVOS DEL ESTUDIO	24
3. MATERIALES Y METODOS	26
3.1. <i>Creación del conjunto de datos</i>	26
3.2. <i>Análisis exploratorio de los datos</i>.....	32
3.3. <i>Limpieza, división y oversampling de los datos:</i>	39
3.4. <i>Preprocesamiento espectral</i>.....	42
3.5. <i>Reducción de la dimensionalidad</i>	53
3.6. <i>Modelos</i>.....	68
3.7. <i>Evaluación de los modelos</i>	81
3.8. <i>Construcción de modelos y pruebas realizadas.</i>	84
4. RESULTADOS Y DISCUSIÓN	88
4.1. <i>Macronutrientes Primarios:</i>	89
4.2. <i>Macronutrientes Secundarios.....</i>	92
4.3. <i>Micronutrientes</i>.....	96
5. CONCLUSIONES Y FUTURA PROYECCIÓN.....	100
6. BIBLIOGRAFÍA.....	102

ÍNDICE DE FIGURAS:

<i>Figura 1.1: Peces y crustáceos muertos a orillas del Mar Menor. Fotografía: Asociación de Naturalistas del Sureste (ANSE).....</i>	20
<i>Figura 1.2. Espectro electromagnético VIS-NIR. Fuente: Blog Digi-Key Spectral Sensores.....</i>	23
<i>Figura 3.1. Esquema del proceso de adquisición de muestras para la creación del conjunto de datos. Proceso de adquisición de imágenes hiperespectrales y análisis destructivo. Fuente: Elaboración propia.</i>	28
<i>Figura 3.2. Proceso de adquisición de la imagen hiperespectral de una hoja de cítrico</i>	29
<i>Figura 3.3. Preprocesamiento en las imágenes hiperespectrales adquiridas.</i>	30
<i>Figura 3.4. Corrección del desplazamiento y concatenación</i>	31
<i>Figura 3.5. PCA y operaciones morfológicas de la imagen</i>	31
<i>Figura 3.6. Diferencia estandarizada entre hojas Jóvenes y Viejas según el nutriente.....</i>	33
<i>Figura 3.7. Diferencia estandarizada entre las orientaciones según el nutriente.</i>	34
<i>Figura 3.8. Espectro medio de las muestras registradas de cada árbol.....</i>	36
<i>Figura 3.9. Espectro medio de las hojas jóvenes y viejas registradas.....</i>	37
<i>Figura 3.10. Espectro medio según la orientación en la que se registró la muestra.....</i>	37
<i>Figura 3.11. Bandas espectrales que registran mayor variabilidad</i>	38
<i>Figura 3.12. Ejemplo del proceso K-Fold con 3 particiones.....</i>	41
<i>Figura 3.13. Representación gráfica de la estandarización de 4 muestras espectrales.....</i>	44
<i>Figura 3.14. Representación gráfica de 4 muestras espectrales originales</i>	44
<i>Figura 3.15. Representación gráfica de una muestra espectrales transformada con Savinsky-Golay y la propia muestra espectral original.....</i>	47
<i>Figura 3.16. Representación de 4 muestras espectrales transformadas con Savitzky-Golay..</i>	48
<i>Figura 3.17. Representación gráfica de una muestra espectrales transformada con la primera derivada de Savinsky-Golay y la misma muestra espectral original.</i>	50
<i>Figura 3.18. Representación gráfica de 4 muestras espectrales transformadas con Savinsky-Golay y la estandarización estándar.....</i>	51
<i>Figura 3.19. Representación gráfica de 4 muestras espectrales transformadas con la diferencia respecto al espectro promedio.</i>	52
<i>Figura 3.20. Variancia explicada de las componentes de PCA.....</i>	57
<i>Figura 3.21. Primera componente vs. segunda componente de los datos estandarizados.....</i>	58
<i>Figura 3.22. Partes de la neurona humana. Fuente: Enciclopedia Humanidades.</i>	60
<i>Figura 3.23. Partes de la neurona artificial. Fuente: Universidad de Murcia.....</i>	61
<i>Figura 3.24. Perceptrón Simple. Fuente: Universidad de Murcia</i>	62
<i>Figura 3.25. Red Neuronal Profunda. Fuente: IBM Blog “What are neural networks?”</i>	63
<i>Figura 3.26. Estructura General de un Autoencoder. Fuente: Fernando Sancho Caparrini Blog.</i>	64

<i>Figura 3.28. Pruebas realizadas en el proyecto para cada modelo y cada nutriente.</i>	86
<i>Figura 4.1. Comparación del RMSE obtenido calibración y prueba de los diferentes mejores modelos que predicen cada macronutriente primario.</i>	90
<i>Figura 4.2. Comparación del MAE obtenido calibración y prueba de los diferentes mejores modelos que predicen cada macronutriente secundario.</i>	93
<i>Figura 4.3. Comparación del MAE y RMSE obtenido en el conjunto de prueba del modelo K-Neighbours que utilice una selección de 6 bandas y el modelo Ada Boost + Bagging que utiliza todas las bandas.</i>	94
<i>Figura 4.4. Bandas más frecuentemente clasificadas como importantes según el Random Forest para cada macronutriente que ha sido amado.</i>	95
<i>Figura 4.5. Gráfico de barras de R^2 obtenido por el modelo que ha logrado un mejor rendimiento en el conjunto de prueba por cada micronutriente: Na, Fe, Mn, Zn, Cu, B y Mo.</i>	97

ÍNDEX DE TAULES:

<i>Tabla 3.1. Ejemplo de la información general de algunas muestras del árbol I02. Fuente: Elaboración propia.</i>	28
<i>Tabla 3.2. Diferencia estandarizada entre las orientaciones según el nutriente.</i>	35
<i>Tabla 3.3. Principales funciones de activación. Fuente: Sagar Sharma en Activation Functions in Neural Networks (Medium)</i>	61
<i>Tabla 3.4. Diferencias entre Gradient Boosting Regressor y AdaBoost.</i>	80
<i>Tabla 3.5. Cómo analizar los residuos de un modelo.</i>	83
<i>Tabla 4.1: Abreviaturas que se utilizan para mostrar los preprocesos que ha utilizado cada modelo.</i>	88
<i>Tabla 4.2. Resultados de los modelos para los macronutrientes primarios N, P, K. La columna PREP muestra la mejor opción de preprocesamiento de los datos. La columna ICR muestra la mejor opción de ingeniería de características, indicando el número de componentes principales o utilizando todas las bandas (No). En negrita se muestra la mejor metodología para cada macronutriente primario sobre la base del valor de R2 en el conjunto de prueba.</i>	89
<i>Tabla 4.3. Resultados de los modelos para los macronutrientes primarios N, P y K. La columna Bandas muestra el número de bandas seleccionadas mediante Random Forest. En negrita se muestra la mejor selección en base al valor de R2 en el conjunto de prueba para cada macronutriente.</i>	91
<i>Tabla 4.4. Bandas seleccionadas en los modelos anteriores para predecir cada nutriente según los resultados obtenidos con el Random Forest.</i>	92
<i>Tabla 4.7. Bandas seleccionadas en los modelos anteriores para predecir cada nutriente según los resultados obtenidos con el Random Forest.</i>	95
<i>Tabla 4.8. Resultados de los modelos para los micronutrientes Na, Fe, Mn, Zn, Cu, B y Mo. La columna PREP muestra la mejor opción de preprocesamiento de los datos. La columna ICR muestra la mejor opción de ingeniería de características, indicando el número de componentes principales o utilizando todas las bandas (No). En negrita se muestra la mejor metodología para cada micronutriente sobre la base del valor de R2 en el conjunto de prueba.</i>	96
<i>Tabla 4.9. Resultados de los modelos para los micronutrientes Na, Fe, Mn, Zn, Cu, B y Mo. La columna Bandas muestra el número de bandas seleccionadas mediante Random Forest. En negrita se muestra la mejor selección en base al valor de R2al conjunto de prueba para cada macronutriente.</i>	98

1. INTRODUCCIÓN

La aplicación de las llamadas Tecnologías de la Información y la Comunicación (TIC) en el sector agrícola ha originado la capacidad de tomar decisiones y adoptar correcciones a partir de medidas tomadas. De esta forma, se consigue mejorar la productividad de los campos y reducir el impacto medioambiental. Por ello, se ha incrementado tanto el uso de este tipo de técnicas como la aparición de nuevas aplicaciones, revolucionando, en consecuencia, el mundo agrícola.

Estas tecnologías se basan en la recogida de datos, que mediante un análisis adecuado proporcionan conocimiento. La aplicación de estas tecnologías o técnicas en el sector agrícola recibe el nombre de Agricultura de Precisión [1]. Básicamente, de esta forma se obtiene, con menores recursos, la optimización de la productividad a la hora de mantener los estándares de calidad. Además, las TIC permiten garantizar de forma efectiva la seguridad de los productos alimenticios mediante el uso de diversas técnicas. Entre sus múltiples aplicaciones destacan el control de cultivos por satélite, la conducción automática de maquinaria agrícola, el control del regadío con drones, etc.

1.1. Los cítricos: una industria en expansión y impacto económico en la Comunidad Valenciana

Según Oxford Languages, se puede definir un fruto cítrico como: "Conjunto de frutas de sabor ácido o agridulce"; y un árbol cítrico como: "Plantas que producen estos frutos". Por tanto, los cítricos son aquellos árboles que producen frutas con sabor ácido o agridulce, que están comercializadas, como por ejemplo el limón, la naranja, el pomelo o la mandarina.

Estos cultivos son ampliamente consumidos en todo el mundo, y además de utilizarse para usos alimenticios, también se aplican en medicina, farmacia, biocombustibles, entre otros ámbitos, debido a sus propiedades. Según el informe estadístico de Naciones Unidas [2] del año 2020, se produjeron 143.755.600 toneladas de cítricos en todo el mundo, y se nota una tendencia positiva de crecimiento. Este informe también destaca que se han mejorado las prácticas agrícolas en los últimos sesenta años, aumentando en un 70 por ciento el rendimiento de los cultivos de cítricos, con un incremento del 23 por ciento en la superficie cosechada y un 16 por ciento en su producción.

La agricultura de precisión ha jugado un papel muy importante en la mejora de la producción de cítricos en los últimos años. Por ejemplo, la rápida detección de la plaga del psílido asiático, que afecta a los cítricos, mediante sistemas de inteligencia artificial, ha permitido tomar medidas anticipadas y evitar la desaparición de los campos de cultivo.

Este cultivo tiene un gran valor en la Comunitat Valenciana, al tener un impacto económico significativo en el territorio. Según los datos de la Conselleria de Agricultura de la Generalitat Valenciana, según su informe de previsión de cosechas para la campaña de 2021/2022 [3], se produjeron un total de 3.508.051 toneladas de cítricos en el territorio valenciano durante la campaña de 2010/2021. Esto convierte a la Comunidad Valenciana en la principal región productora de cítricos a nivel nacional, con la naranja y la mandarina como principales variedades de cítricos producidas.

1.2. La importancia de los nutrientes en el saludable crecimiento de los árboles: equilibrio e impacto medioambiental

Los árboles y las plantas requieren nutrientes esenciales para sobrevivir y prosperar [4]. Cuando un árbol tiene una baja nutrición, se debilita y se vuelve vulnerable a plagas y enfermedades. Los árboles absorben sus nutrientes a través del suelo, además de necesitar humedad y luz solar. Una correcta nutrición permite que el árbol crezca adecuadamente y, en el caso de los cítricos, produzca alimentos de mayor cantidad y mejor calidad.

Teniendo en cuenta que una nutrición adecuada de los árboles es fundamental para mantenerlos sanos y para obtener alimentos de buena calidad, es importante conocer cómo se pueden tratar. Para solucionar la falta de nutrientes, un árbol puede recibir suplementos mediante los llamados fertilizantes o abonos. Sin embargo, el uso excesivo de fertilizantes puede ocasionar varios problemas. Por un lado, puede contribuir a la liberación de gases de efecto invernadero en la atmósfera. Por otra parte, puede causar lo que se conoce como eutrofización, que es la adición de nutrientes externos (como el exceso de nitrógeno) en los cursos de agua [5].

En la siguiente Figura 1.1 se muestran las consecuencias de la eutrofización de los cursos de agua en el caso del Mar Menor. La eutrofización puede provocar un rápido crecimiento de microorganismos en el agua, los cuales pueden consumir todo el oxígeno disponible en estos cursos de agua y crear lo que se conoce como zonas muertas. Además, puede generar la

proliferación de algas que producen sustancias tóxicas (floraciones de algas nocivas).



Figura 1.1: Peces y crustáceos muertos a orillas del Mar Menor. Fotografía: Asociación de Naturalistas del Sureste (ANSE)

Entonces, es necesario equilibrar los beneficios de la fertilización (mayor aporte de alimentos) con las consecuencias negativas del exceso de fertilizante (emisiones de gases de efecto invernadero). Por tanto, conocer adecuadamente las necesidades del árbol para evitar un uso excesivo de fertilizantes tiene cuatro implicaciones importantes:

- Reducir los efectos ambientales negativos.
- Minimizar el desperdicio de recursos innecesarios.
- Optimizar la producción de fruta.
- Mantener los árboles y productos en buen estado de salud.

Clorofila

Antes de profundizar en los nutrientes, es importante entender qué es la clorofila y su papel fundamental. La clorofila es un pigmento verde que se encuentra en prácticamente todas las plantas, algas y cianobacterias. La cantidad de clorofila en un árbol está relacionada con la disponibilidad de nutrientes en el suelo. Esta molécula es necesaria para la fotosíntesis [6], un proceso vital para el crecimiento del árbol. Si el suelo no contiene los nutrientes necesarios, el árbol no podrá producir suficiente clorofila y su crecimiento será afectado negativamente.

Nutrientes:

De forma general, se hace una distinción entre los macronutrientes, que son los nutrientes requeridos en mayor cantidad, y los micronutrientes, que son los elementos necesarios en menor cantidad. Esta división no implica que un nutriente sea más importante que otro, sino que simplemente se necesitan

en distintas proporciones. La combinación adecuada de macronutrientes y micronutrientes proporciona un suelo en óptimas condiciones de salud. En cada grupo tenemos:

1- Macronutrientes primarios:

1.1. Nitrógeno (N): El nitrógeno ayuda a los árboles y las plantas a convertir el agua, la luz solar y el suelo en alimentos. Estimula el crecimiento de las hojas y contribuye a su color verde.

1.2. Fósforo (P): Responsable del correcto crecimiento de las raíces. También ayuda a los árboles a producir semillas, flores y frutos. Fortalece el árbol, haciéndolo menos vulnerable a las enfermedades.

1.3. Potasio (K): Cuando un árbol tiene suficiente potasio, produce frutos saludables. El potasio también protege al árbol de daños durante las estaciones frías. Colabora con el fósforo para proteger el árbol de enfermedades y ayudarle a recuperarse de éstas.

2- Macronutrientes secundarios:

2.1. Calcio (Ca): Construye las paredes celulares de las plantas y fortalece los tallos de los árboles.

2.2. Magnesio (Mg): Otro componente de la clorofila. También desempeña un papel en el crecimiento de los árboles.

2.3. Azufre (S): Los árboles necesitan el azufre principalmente para producir proteínas. También interviene en la formación de clorofila, en la producción de semillas y en la mejora de las raíces y el crecimiento general de la planta.

3- Micronutrientes:

3.1. Sodio (Na): Auxiliar en el metabolismo y la síntesis de clorofila.

3.2. Hierro (Fe): Necesario para la producción de clorofila.

3.3. Manganeso (Mn): Apoyo para las enzimas en la digestión de los carbohidratos y en la conversión del nitrógeno.

3.4. Zinc (Zn): Un componente de las enzimas que controlan el crecimiento.

3.5. Cobre (Cu): Parte esencial de la reproducción. Ayuda al árbol en la absorción y digestión de nutrientes a través de las raíces.

3.6. Boro (B): Controla y equilibra otros nutrientes. También es necesario para la producción de frutos y semillas.

3.7. Molibdeno (Mo): Otro mineral utilizado para la transformación del nitrógeno.

Toda esta información se encuentra detallada en [7].

1.3. Análisis espectral para la estimación de nutrientes en cultivos: Una alternativa eficiente i económica

Cómo se ha podido comprobar, es muy importante conocer las deficiencias de nutrientes en los cultivos y saber cuáles y en qué medida no se encuentran en la cantidad óptima. Por tanto, actualmente se dispone de diversas técnicas que ayudan a evaluar el estado de los cultivos.

En la actualidad, la herramienta más utilizada es el análisis de la ionómica foliar de las hojas. La ionómica es la medida de la acumulación total de metales, metaloides y no metales en los organismos vivos. La ionómica vegetal se ha aplicado en diversas investigaciones durante la última década [8].

Sin embargo, esta técnica tiene un coste económico y temporal elevado, ya que requiere análisis químicos en laboratorio. Por ello, se están investigando técnicas alternativas que permitan estimar estos nutrientes con un menor coste económico y temporal, aunque puedan tener una precisión ligeramente inferior.

1.4. Reflectancia de la luz: Analisis a través del espectro visible i infrarrojo cercano

El espectro electromagnético se define como el conjunto de toda la radiación electromagnética que existe en el universo. La frecuencia se define por la longitud de onda, que es la distancia desde el pico de una onda hasta el siguiente pico. Estos dos atributos están inversamente relacionados: cuanto mayor sea la frecuencia, menor será la longitud de onda y viceversa.

Dentro de este espectro existen varios tipos de ondas, desde los rayos gamma (baja longitud de onda/alta frecuencia) hasta las ondas de radio (gran longitud de onda/baja frecuencia). En este proyecto, se trabaja con las longitudes de onda que pertenecen a una parte del espectro visible (VIS) e infrarrojo cercano (NIR). Esta información puede apreciarse gráficamente en la Figura 1.2.

- **Vis:** Esta parte corresponde al segmento del espectro electromagnético que el ojo humano puede ver. Es decir, la luz visible. Por lo general, el ojo humano puede detectar ondas de entre 380-700 nanómetros (nm).

- **NIR:** Comprende radiación de baja frecuencia adyacente a los tonos rojos en lo visible. Este espectro va desde 700 hasta 2500 nm aunque no existe una definición universalmente aceptada.

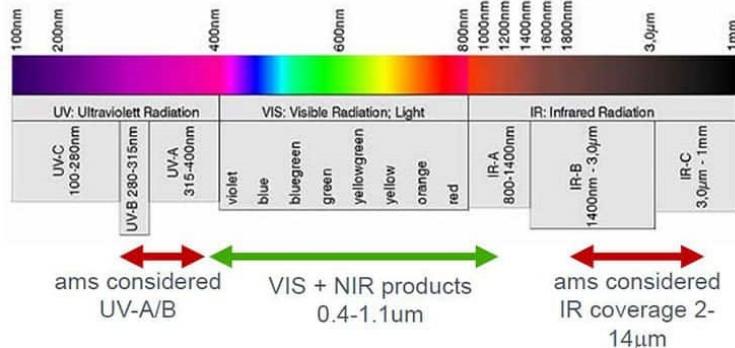


Figura 1.2. Espectro electromagnético VIS-NIR. Fuente: Blog Digi-Key Spectral Sensores

Así pues, irradiando un material con distintas longitudes de onda, es posible medir su reflectancia. La reflectancia se define como el grado de luz que se refleja por la superficie de un material. El grado de esa reflectancia dependerá del nivel de absorción de luz por parte del material. Por ejemplo, un material de color blanco reflejará más la iluminación que uno de color negro, por lo que tendrá un mayor índice de reflectividad. Esta información se encuentra detallada en [9].

Los equipos encargados de medir esta reflectancia en un material iluminado con una determinada longitud de onda se conocen como espectrómetros de reflectancia. La reflectancia se mide en una escala de 0 a 1 (o también en porcentaje de 0% a 100%). Si la reflectancia de un material es 0, significa que no refleja ninguna luz y, por tanto, lo absorbe todo (es de color negro absoluto). En cambio, un valor de 1 indica que el material refleja toda la luz (es de color blanco perfectamente reflectante).

2. OBJETIVOS DEL ESTUDIO

2.1. Estimación precisa de nutrientes en cítricos mediante la espectroscopia

Los datos espectrales se utilizan ampliamente en varios campos, desde la investigación farmacéutica y médica hasta la ciencia de los materiales y la agricultura. La espectroscopia VIS-NIR es una técnica fácil, rápida, no destructiva y económica [10]; permite realizar evaluaciones de calidad con poca o ninguna preparación previa de la muestra. Esta técnica tiene aplicaciones en el análisis de calidad de frutas y verduras en los ámbitos industrial y semiindustrial de la agricultura [11].

Esto significa que con los datos espectrales se pueden realizar muchas tareas que solían requerir estudios o investigaciones costosas en términos de tiempo y recursos económicos, consiguiendo resultados muy buenos a un coste realmente bajo. Esto ha provocado un aumento del uso de la espectroscopia en los últimos años.

Dentro del estudio del sector agrícola y alimentario, que es el campo de inclusión de este proyecto, existen múltiples aplicaciones de la espectroscopia. Algunas de ellas son:

- Análisis multielemental de los suelos.
- Detección de enfermedades o daños en los alimentos (control de calidad).
- Monitorización de contaminantes orgánicos.

En este caso concreto, el proyecto se centra en la creación de modelos de regresión de aprendizaje automático capaces de estimar los niveles de nutrientes de un árbol cítrico a partir del espectro generado por sus hojas. Para ello, se dispone de una colección de muestras de hojas, de las cuales se ha calculado el espectro en un rango de longitudes de onda y también se ha realizado un análisis destructivo para determinar los niveles de varios nutrientes.

Como se ha comentado anteriormente, la ventaja de utilizar un sistema que utiliza información espectral para estimar los niveles de nutrientes frente a un análisis destructivo es que se reducen los gastos. Esta reducción de costes permite a los agricultores que no pueden permitirse un análisis destructivo de sus árboles por razones económicas puedan acceder a sistemas que

ayudan a evitar el uso excesivo de fertilizantes ya mantener sus árboles en un buen estado de salud.

2.2. Un enfoque de regresión supervisada

El objetivo del proyecto será evaluar el funcionamiento de distintos modelos de aprendizaje automático para ver cuáles tienen un mejor funcionamiento. Existen diferentes tipos de modelos de aprendizaje automático, que se pueden clasificar a grandes rasgos en tres categorías: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo.

El aprendizaje supervisado es aquél en el que el ordenador recibe un conjunto de datos de entrenamiento, que incluye las respuestas correctas, y se utiliza el algoritmo para aprender de estos datos para ser capaz de generalizar y hacer predicciones sobre nuevos datos. El aprendizaje no supervisado es aquel en el que el ordenador recibe datos, pero no las respuestas correctas, y debe aprender de estos datos por sí mismo para encontrar patrones y relaciones. El aprendizaje por refuerzo es aquél en el que el ordenador recibe un objetivo que debe alcanzar y debe aprender por ensayo y error para averiguar su mejor manera.

Volviendo al objetivo del proyecto, la idea es predecir el nivel nutricional en base al espectro generado, por tanto, será un problema a resolver mediante aprendizaje supervisado, aunque también se utilizan técnicas de aprendizaje no supervisado con otros fines .

La regresión y clasificación son ambos tipos de problemas de aprendizaje supervisado. La regresión es un tipo de problema de aprendizaje supervisado en el que el objetivo es predecir un valor continuo. Por ejemplo, puede utilizar la regresión para predecir el precio de una casa en función de su tamaño, edad y ubicación. La clasificación es un tipo de problema de aprendizaje supervisado en el que el objetivo es predecir una etiqueta de clase. Por ejemplo, puede utilizar la clasificación para predecir si un paciente tiene cáncer en función de su historial médico.

Por tanto, el nivel nutricional, como se ha comentado anteriormente, es un dato numérico continuo. Sabiendo esto, definimos el problema como un problema de regresión.

3. MATERIALES Y MÉTODOS

3.1. Creación del conjunto de datos

Para crear un conjunto de datos para la resolución del problema, se requiere dos partes: el cálculo real de los nutrientes de las hojas y los datos espectrales. A continuación, se explica cómo se han obtenido estos datos.

- **Cálculo real de los nutrientes de las hojas:** Para determinar los niveles de nutrientes en las hojas de cítricos, se ha realizado un análisis destructivo. Esto implica tomar muestras de hojas y someterlas a procedimientos de análisis químico en laboratorio para medir las concentraciones de distintos nutrientes. Este proceso permite obtener datos precisos y confiables sobre los niveles de nutrientes presentes en las hojas.
- **Datos espectrales:** Para obtener los datos espectrales, se han utilizado espectrómetros de reflectancia. Estos dispositivos emiten luz en distintas longitudes de onda y miden la cantidad de luz reflejada por las hojas en cada longitud de onda. Esto genera un espectro que muestra los niveles de reflectancia en distintos intervalos de longitud de onda. Mediante este proceso, se han recopilado los datos espectrales para las hojas de cítricos utilizadas en el estudio.

Con estas dos partes de datos, es posible crear el conjunto de datos necesario para entrenar y evaluar los modelos de aprendizaje automático para la predicción de los niveles de nutrientes en base a los datos espectrales. Este conjunto de datos combina las medidas reales de los nutrientes con los datos espectrales correspondientes, proporcionando una base para el desarrollo y la evaluación de los modelos.

3.1.1. Desarrollo Experimental

Para llevar a cabo el proyecto, se ha realizado un ensayo en una parcela comercial de cítricos en Almenara (Castellón), cuya localidad se encuentra ubicada a 39°45'12"N 0°13'32" O. La variedad estudiada es Clemenules (*Citrus clementina*, Hort ex Tan) unos árboles que producen un cítrico híbrido entre la mandarina y la naranja amarga. La parcela cuenta con 210 árboles y el ensayo se efectuó el día 4 de mayo de 2021.

Se tomaron muestras de un total de 33 árboles, de cada árbol se tomaron 6 muestras de 30 hojas cada una: hojas jóvenes (brotación de primavera) y

3. Materiales y Métodos

hojas viejas (brotación ciclos vegetativos anteriores) de la orientación este, oeste y cenital , muestreando un total de 5940 hojas. Todas las hojas jóvenes/viejas de cada orientación pertenecían a la misma rama del árbol.

Entonces, de cada árbol tenemos un conjunto de 30 hojas jóvenes y 30 hojas viejas de cada orientación. De cada uno de ellos, dividimos la muestra en 22 hojas para calcular el análisis destructivo y, por tanto, saber los nutrientes del grupo. De esta forma, se conocen los nutrientes reales de las hojas.

Después, las 8 restantes de la muestra se utilizan para calcular el espectro. En el caso de las hojas viejas se realizan dos particiones de 4 hojas. Esto se hace debido a que las hojas viejas son grandes, y no caben todas en el ángulo de la lente de la cámara.

Entonces, de cada orientación se calcula el espectro de 8 hojas jóvenes, por un lado, y, por otro, se calcula el espectro de dos divisiones de 4 hojas viejas. Por tanto, de cada árbol y cada orientación se obtienen 3 espectros: uno de hojas jóvenes, uno de la primera partición de 4 hojas viejas, y otro de la segunda partición de otras 4 hojas viejas. Las dos divisiones de hojas viejas tendrán espectros diferentes, pero los valores de los nutrientes serán los mismos, puesto que se ha elaborado el análisis destructivo de la otra parte de la muestra.

Por último, tenemos como resultado, de cada árbol 9 espectros (representando a y b cada partición de hojas viejas):

LEJ, LEV, LEV_b, LOJ, LOV_a, LOV_b, CeJ, CeV_a y CeV_b.

Puede verse en la siguiente Figura 3.1. un mapa conceptual para poder entender cuál ha sido el proceso de adquisición de los datos:

MAPA CONCEPTUAL DEL PROCESO RECOLECCIÓN

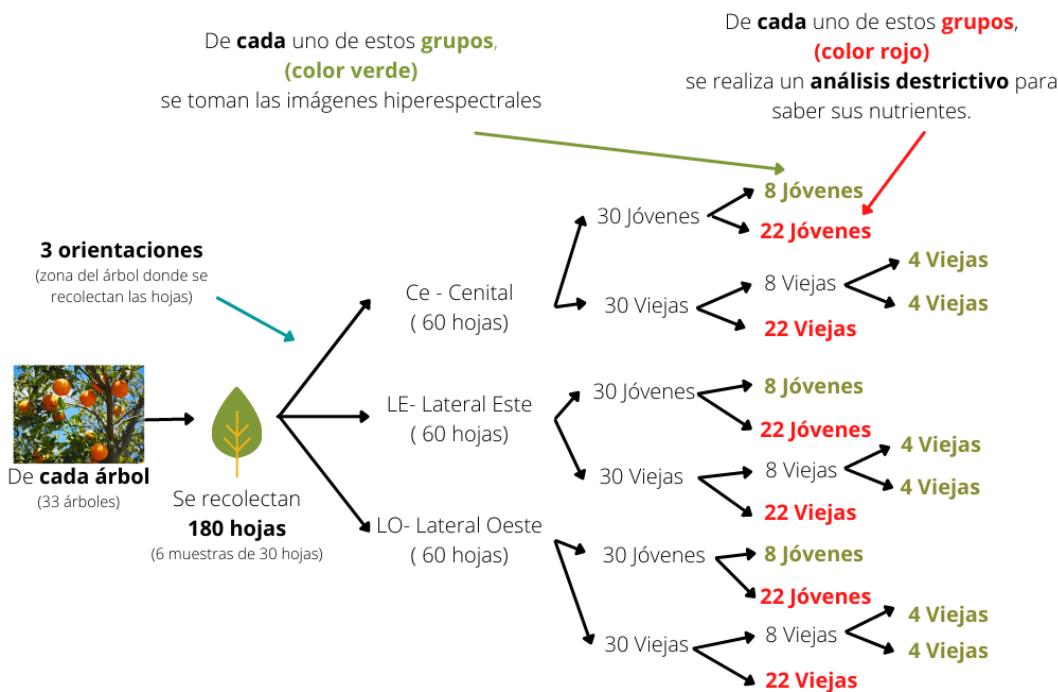


Figura 3.1. Esquema del proceso de adquisición de muestras para la creación del conjunto de datos. Proceso de adquisición de imágenes hiperespectrales y análisis destructivo. Fuente: Elaboración propia.

Lo que finalmente genera un archivo en el que se pueden diferenciar las siguientes partes:

- **Información General:** Tenemos una serie de variables que contribuyen a conocer de dónde viene el valor espectral y los nutrientes.
- **Nombre:** Junta el nombre del árbol, orientación, zona, a/b y muestra.
- **Muestra:** Mediante un identificador (desde 1 hasta 198) tenemos identificado cada grupo de hojas de los que se calculan los nutrientes.
- **Árbol:** Identifica el árbol del que provienen las hojas.
- **Zona:** Identifica la orientación y si son jóvenes o viejas las hojas de forma conjunta (LEJ, LEV, LOJ, LOV, CeJ y CeV).
- **Orientación:** Representa la orientación (LE, LO, Ce)
- **J/V:** Identifica si las hojas son jóvenes (J) o viejas (V).
- **a/b:** Representa la partición de las hojas con las que se calcula el espectro. Las Jóvenes siempre serán del grupo a (ya que no se realiza ninguna partición), y las viejas pueden ser o a/b (dependiendo de la partición).

La información general de un árbol, por ejemplo el árbol identificado por I02, puede apreciarse en la Tabla 3.1.

Nombre	Muestra	Árbol	Zona	Orientación	J/V	a/b
AI02_HLEJa_001	001	I02	LEJ	LE	J	a
AI02_HLEVa_002	002	I02	LEV	LE	V	a
AI02_HLEVb_002	002	I02	LEV	LE	V	b
AI02_HLOJa_003	003	I02	LOJ	LO	J	a
AI02_HLOVa_004	004	I02	LOV	LO	V	a
AI02_HLOVb_004	004	I02	LOV	LO	V	b
AI02_HCejJa_005	005	I02	CeJ	Ce	J	a
AI02_HCejVa_006	006	I02	CeV	Ce	V	a
AI02_HCejVb_006	006	I02	CeV	Ce	V	b

Taula 3.1. Ejemplo de la información general de algunas muestras del árbol I02. Fuente: Elaboración propia.

Seguidamente, para cada muestra, también tenemos:

- **Nutrientes:** Los nutrientes (N, P, K, Ca, Mg, Na, S, Fe, Mn, Zn, Cu, B y Mo) de cada muestra obtenidos mediante el análisis destructivo.
- **Espectro:** De cada 'nombre', es decir, de cada partición de cada muestra, tenemos tantas variables como longitudes de onda espectrales.

3.1.2. Adquisición de las Imágenes Hiperespectrales

En el laboratorio de Visión y Espectroscopía del Centro de Ingeniería Agrícola del IVIA (Instituto Valenciano de Investigaciones Agrarias), se utilizaron sistemas de imágenes hiperespectrales en el rango Vis-NIR para adquirir imágenes de las hojas de cítricos utilizadas en el trabajo. El sistema consiste en una cámara industrial y dos filtros sintonizables de cristal líquido LCTF (Liquid Crystal Tunable Filter), capaz de adquirir imágenes de 1392 x 1040 píxeles con una resolución espacial de 0,14 mm/píxel en 65 bandas distintas con una resoluciónpectral de 10 nm, en el rango spectral de 400 nm a 1050 nm.

Se utilizó un sistema basado en iluminación halógena difusa eficiente en todo el rangopectral de trabajo. Para evitar problemas de imágenes borrosas debido a la dispersión cromática de la lente, se estableció su enfoque a la longitud de onda central (730 nm) del rango de trabajo. Se adquirió una imagen hiperspectral de un blanco de referencia certificada que cubría el 90% del rango dinámico de la cámara para evitar la saturación de la imagen

y corregir la sensibilidad espectral del sistema. Aquest entorn on es van aconseguir les imatges hiperespectals es pot apreciar a la següent Figura 3.2.

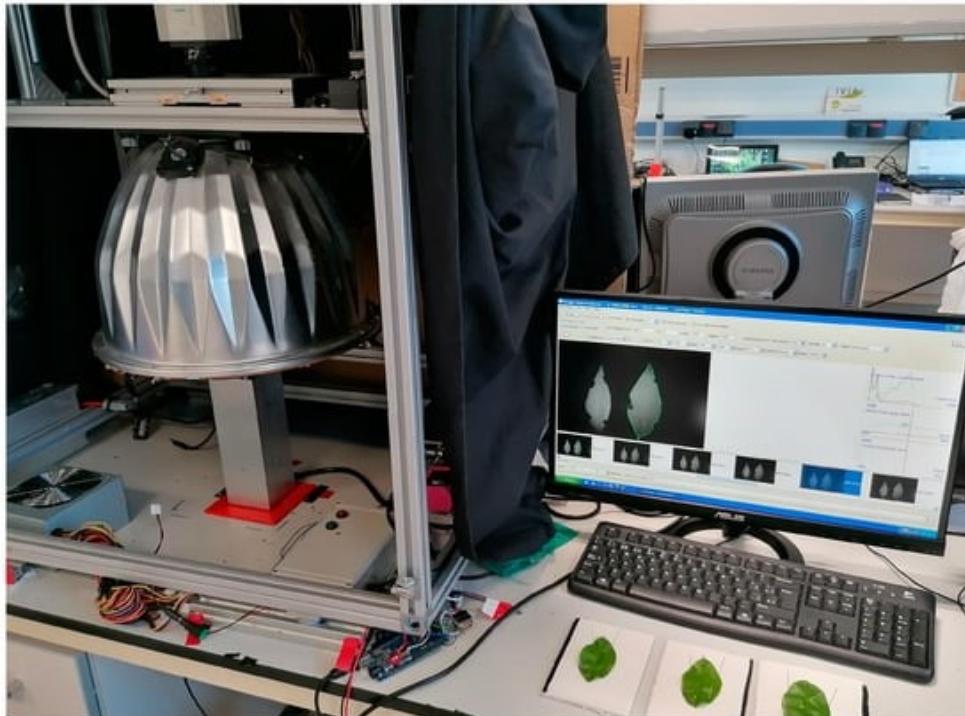


Figura 3.2. Proceso de adquisición de la imagen hiperespectral de una hoja de cítrico.

3.1.3. Adquisición del Espectro en base a las Imágenes Hiperespectrales

El espectro de cada hoja y sus propiedades se extrajeron mediante algoritmo desarrollado en MATLAB que procesó digitalmente las imágenes hiperespectrales adquiridas individualmente. Este procesamiento digital consta de tres bloques o etapas distintas.

En un primer paso, se calibraron imágenes hiperespectrales en bruto grabadas en diferentes rangos de longitud de onda, el rango visible (400 a 720 nm) y el rango de infrarrojo cercano (650 a 1050 nm), aplicando la siguiente ecuación (Ecuación 3.1):

$$R(x,y,\lambda) = R_{ref(\lambda)} * \frac{Image\ Raw(x,y,\lambda) - Image\ Dark(x,y,\lambda)}{Image\ White(x,y,\lambda) - Image\ Dark(x,y,\lambda)} * 100$$

Ecuación 3.1. Ecuación de Calibración de las imágenes:

La Image White es la imagen adquirida del blanco de referencia blanco antes mencionado y la Image Dark es la imagen adquirida de la referencia de negro. Se comprueba de forma gráfica el proceso en la Figura 3.3.

3. Materiales y Métodos

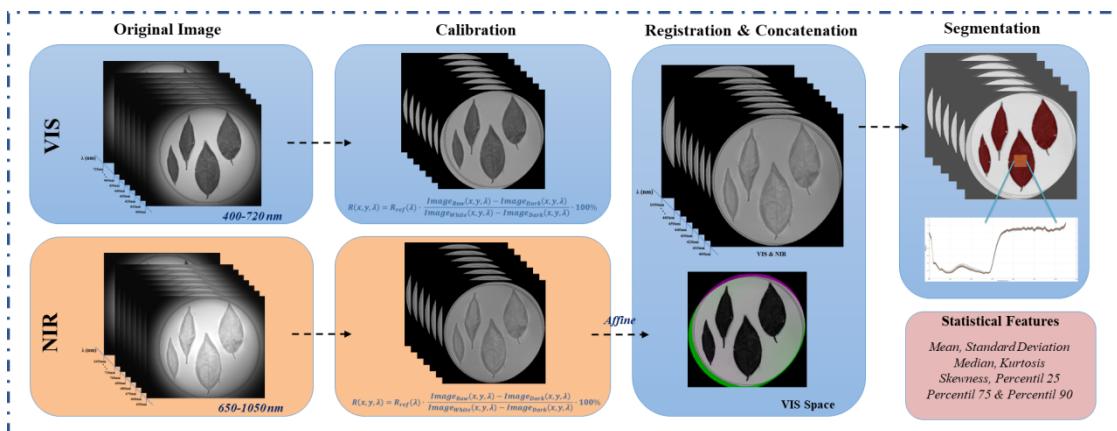


Figura 3.3. Preprocesamiento a las imágenes hiperespectrales adquiridas.

En la segunda etapa, que se ve en la Figura 3.4, debe corregirse el desplazamiento espacial existente entre la imagen NIR y la imagen VIS debido a la configuración del mismo sistema de grabación. Para corregirlo, se creó un registro en el que se calcula una transformación afín y se aplica a imágenes de la región NIR para poder alinearlas con imágenes de la región VIS.

Esta transformación consiste en mantener la imagen del rango VIS mientras se desliza la imagen NIR por encima, utilizando una matriz de transformación afín adecuada para obtener una visualización conjunta.

Lo que se está haciendo es aumentar la coherencia espacial entre las imágenes de las dos regiones capturadas, permitiéndonos utilizar un único hipercubo en vez de dos.

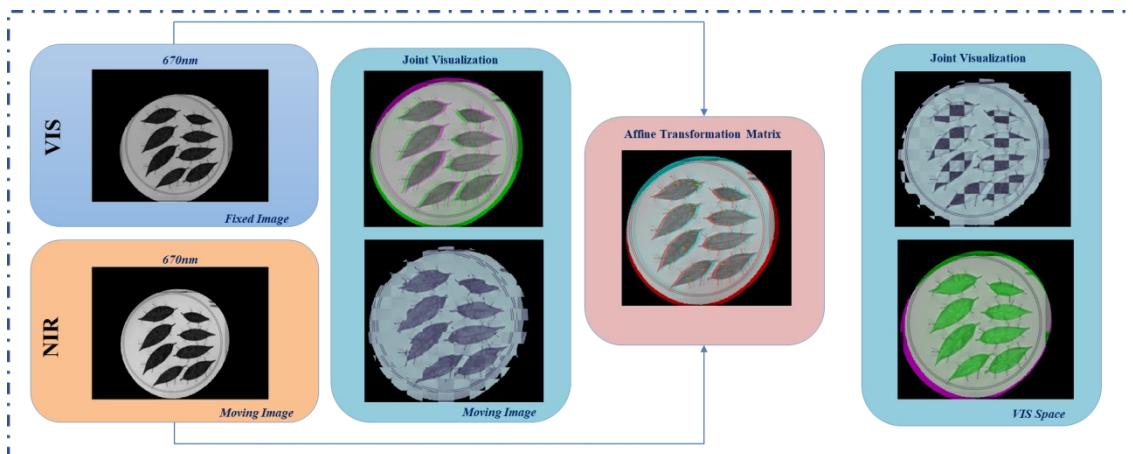


Figura 3.4. Corrección del desplazamiento y concatenación.

Finalment, es va aplicar una etapa de segmentació automàtica per extreure la regió d'interès, és a dir, la superfície de les fulles. Per a això es va realitzar un algoritme de segmentació basat en components principals (PCA) com es veu en la Figura 3.5 que hi ha a continuació. De la tècnica PCA, es parla més endavant.

Este algoritmo calculaba el PCA de todo el hipercubo y seleccionaba la primera componente por ser la que más contraste mostraba de entre los distintos componentes de la imagen. Después mediante diferentes operaciones morfológicas de imagen se eliminaron las estructuras coincidentes con los bordes de la imagen y se realizó una mascarilla binaria, eliminando los elementos que no se busca cómo puede ser el fondo. De esta forma obtenemos las zonas de especial interés, es decir, la superficie foliar.

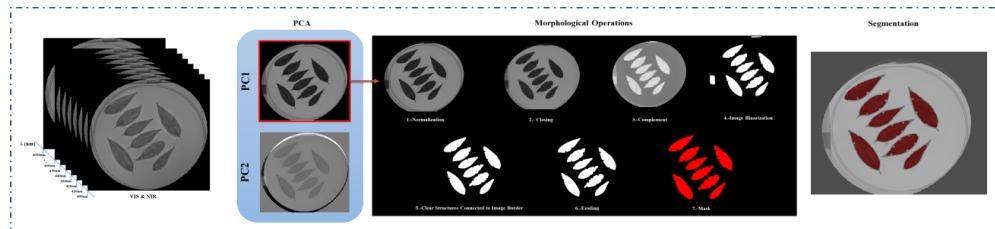


Figura 3.5. PCA y operaciones morfológicas de la imagen.

Este proceso lo encontramos detallado en [12].

3.2. Análisis exploratorio de los datos

En la ciencia de datos siempre es importante conocer los datos; saber sus principales características nos ayuda a entender su comportamiento y poder formular preguntas o hipótesis que en pasos posteriores nos serán de utilidad para resolver de una mejor manera el problema. La siguiente cita del estadístico Jhon W. Tukey puede definir a la perfección el motivo de la importancia de elaborar este análisis:

"Es mejor tener una respuesta aproximada a la pregunta correcta que una respuesta exacta a la pregunta equivocada"

~ John W. Tukey

Como se ha comentado anteriormente, tenemos tres partes en los datos: la información general, los nutrientes y el espectro. Entonces, este apartado se divide en dos subapartados, en el primero, se analizan los datos nutricionales, y en el segundo, los datos espectrales.

3.2.1. Análisis de los nutrientes

- Análisis de los nutrientes en función del tipo de hoja (J/V)**

Una vez se han estandarizado los datos nutricionales, puede procederse a la comparación de la variabilidad inherente a cada nutriente, en función de si la hoja es joven o vieja. En la Figura 3.6, se presentan las diferencias estandarizadas en la concentración de nutrientes entre hojas jóvenes y viejas.

La desviación estándar (std) se utiliza para cuantificar la variabilidad o dispersión de un conjunto de valores nutricionales. Para cada nutriente, se calcula la desviación estándar de las concentraciones estandarizadas. Este cálculo proporciona una medida de la cantidad en la que las concentraciones de un nutriente específico difieren o varían en la población de hojas, ya sean jóvenes o viejas. En la ecuación 2.0 se observa cómo se realiza este cálculo.

$$\sigma = \sqrt{\sum (xi - \mu)^2 / N}$$

Ecuación 2.0. Desviación estándar

- xi es cada valor de la variable (en este caso, la concentración estandarizada de un nutriente específico),
- μ es la media de todos los valores de la variable,
- Σ denota la sumatoria de todos los valores,
- N es el número total de valores (es decir, el número total de mediciones de la concentración del nutriente).

Por ejemplo, el calcio muestra una variabilidad significativa entre las hojas jóvenes y viejas, con una desviación estándar del 39,5% [13]. Esto significa que, en promedio, las concentraciones de calcio en una hoja específica difieren en un 39,5% de la media de las concentraciones de calcio para todas las hojas.

La Figura 3.6 ilustra visualmente estas desviaciones estándar para cada nutriente, proporcionando una representación intuitiva de la variabilidad de las concentraciones de nutrientes en las hojas jóvenes y viejas.

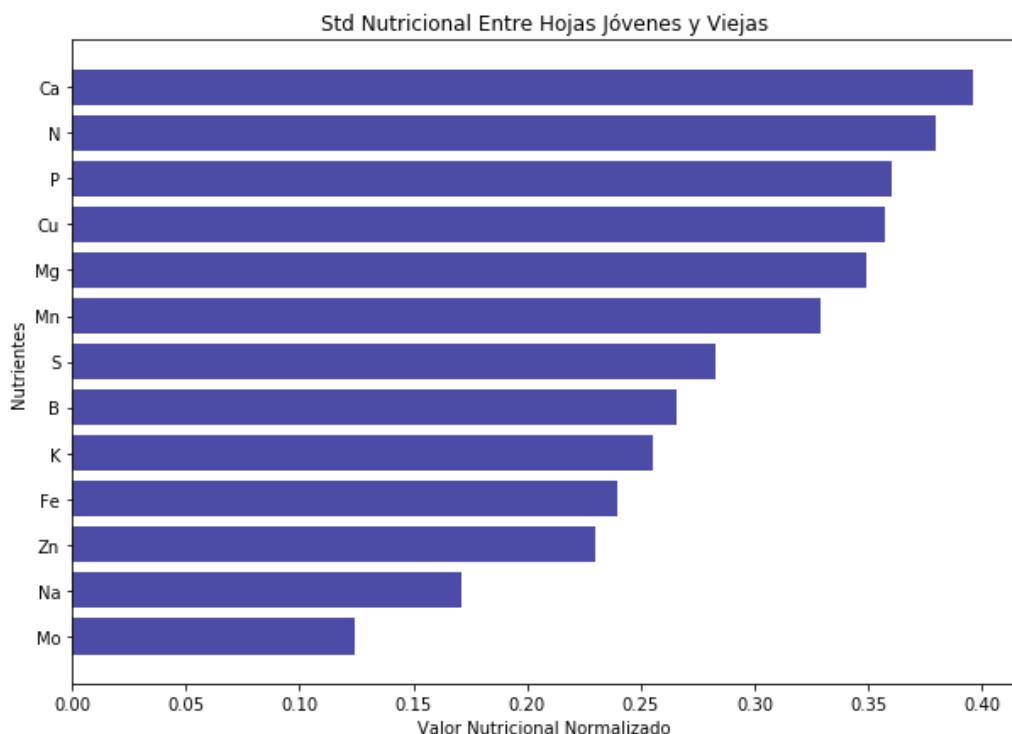


Figura 3.6. Diferencia estandarizada entre hojas Jóvenes y Viejas según el nutriente.

Esta disparidad en los niveles de calcio puede explicarse por diversas razones. En primer lugar, el calcio es un nutriente esencial para el crecimiento y el desarrollo de las plantas. A medida que las plantas crecen, absorben calcio del suelo y lo almacenan en las hojas, lo que puede resultar en una mayor concentración en las hojas viejas.

Además, el calcio es un nutriente inmóvil dentro de las plantas, lo que significa que una vez es absorbido por las raíces y transportado en las hojas, no se mueve significativamente a través de la planta. Esto hace que la concentración de calcio continúe aumentando en las hojas a medida que las raíces absorben más calcio y lo transportan hacia las hojas.

En cuanto al nitrógeno, se observa que las hojas jóvenes tienen una mayor concentración que las hojas viejas. Esto se debe a que las hojas jóvenes están en crecimiento activo y requieren un mayor suministro de nitrógeno para su desarrollo. A medida que las hojas envejecen, pueden perderse nitrógeno a través de procesos como la respiración y la transpiración.

En el caso del potasio, se encuentra que las hojas jóvenes suelen tener una mayor concentración que las hojas viejas. Esto se debe al papel esencial que el potasio juega en el crecimiento y desarrollo de las plantas. A medida que las hojas envejecen, pueden perder potasio a través de varios procesos.

Con otros nutrientes, también pueden observarse diferencias según el tipo de hoja, y estas diferencias pueden ser explicadas por diversas razones específicas.

- **Anàlisis de los nutrientes en función de la orientación**

Ahora podemos ver lo mismo que en el caso anterior, en el caso de las distintas orientaciones. Se observa que las orientaciones tienen una influencia mucho menor en los nutrientes según el tipo de hoja, tal y como se muestra en la Figura 3.7. Puede apreciarse que las hojas jóvenes y viejas tienen un impacto significativo en los niveles de calcio en comparación con los otros nutrientes, mientras que las orientaciones no presentan una diferencia destacable. Sin embargo, puede observarse una diferencia entre los niveles de hierro (Fe) y los otros nutrientes.

Según la posición de la hoja en el árbol, la luz solar puede influir en la producción de clorofila, que contribuye a la absorción de algunos nutrientes como el hierro. Sin embargo, en el caso del calcio, como se ha mencionado anteriormente, este nutriente se absorbe principalmente del suelo, por lo que la luz solar no tiene una influencia directa en la concentración de ese nutriente en las hojas.

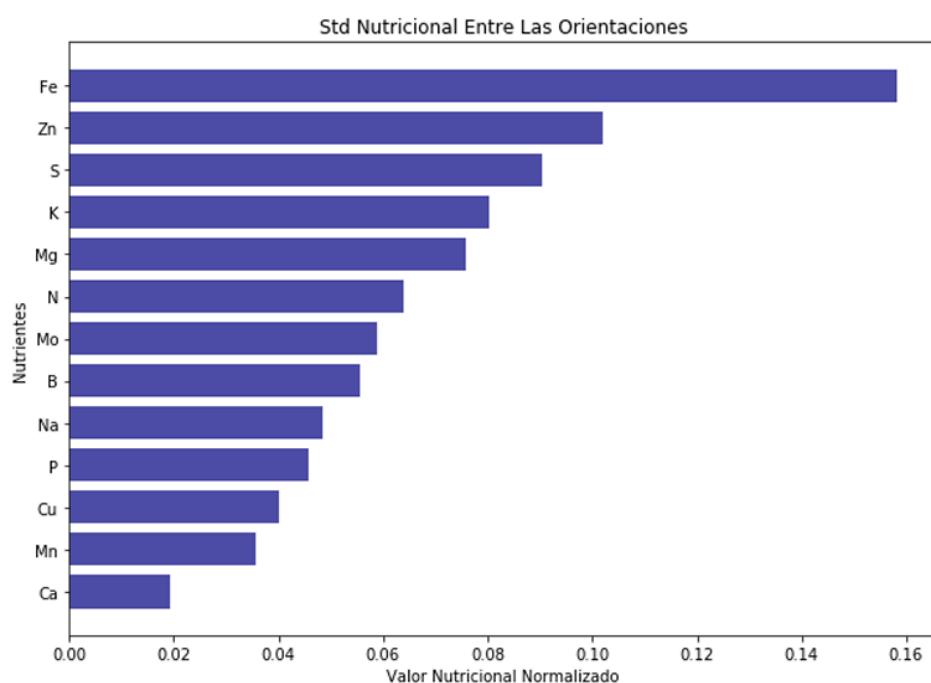


Figura 3.7. Diferencia estandarizada entre las orientaciones según el nutriente.

- **Tabla Desviación Típica según tipo de hoja y orientación**

Nutrientes	Tipo Hoja (J/V)	Orientación (Ce, LE, LO)
N	37.94 %	6.38 %
P	35.99 %	4.58 %
K	25.49 %	8.02 %
Ca	39.57 %	1.91 %
Mg	34.88 %	7.57 %
Na	17.12 %	4.83 %
S	28.27 %	9.04 %
Fe	23.99 %	15.81 %
Mn	32.88 %	3.55 %
Zn	22.97 %	10.21 %
Cu	35.73 %	4.00 %
B	26.53 %	5.54 %
Mo	12.41 %	5.88 %

Taula 3.2. Diferencia estandarizada entre las orientaciones según el nutriente.

En la Tabla 3.2 se muestra el porcentaje de variaciones comentadas anteriormente. Puede observarse que, en general, el porcentaje de variación según la orientación de la hoja es significativamente mayor en comparación con el tipo de hoja.

3.2.2. Anàlisis de los espectros:

- **Espectro Medio de cada árbol**

En esta sección se explora el espectro promedio de cada árbol. Se ha recogido un total de 180 hojas de distintas orientaciones y edades de cada uno de los 33 árboles. En la Figura 3.8 se muestra la media de todos los espectros de estos árboles, donde cada línea representa el espectro promedio de un árbol.

En el espectro promedio, se observa una elevada reflectancia en la parte violeta-azul del espectro, seguida de una reflectancia más baja en el color verde. Después, se vuelve a observar una mayor reflectancia en la parte naranja y roja del espectro. Esta reflectancia está relacionada con el color verde de las hojas de los cítricos, que es causado por la presencia de clorofila. La clorofila tiene la capacidad de absorber la luz en las partes rojas y azules del espectro, mientras que el color verde es el que más se refleja.

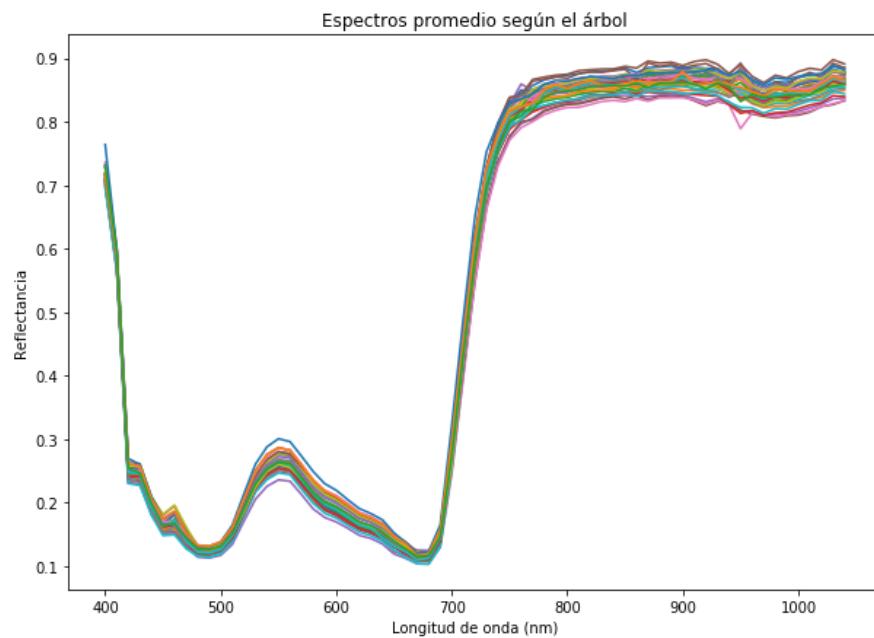


Figura 3.8. Espectro medio de las muestras registradas de cada árbol

- **Espectro según el tipo de hoja**

En la Figura 3.9, se aprecia claramente una diferencia entre la media del espectro de las hojas jóvenes y la media de las hojas viejas. Hasta las primeras bandas espectrales (primera parte del color violeta-azul del espectro visible), las medias de las hojas se mantienen similares. Sin embargo, a partir del color verde, se comienza a observar una diferencia significativa.

La diferencia más destacada entre las medias de los espectros se encuentra en las bandas espectrales de 490 a 680nm, siendo el valor de 550nm en donde se aprecia la máxima diferencia. Esta diferencia se debe a que las hojas jóvenes tienen una mayor concentración de clorofila que las hojas viejas, lo que les confiere un color verde más intenso.

En resumen, las diferencias en los espectros de las hojas jóvenes y viejas se deben a las variaciones en las cantidades de clorofila presente en cada tipo de hoja.

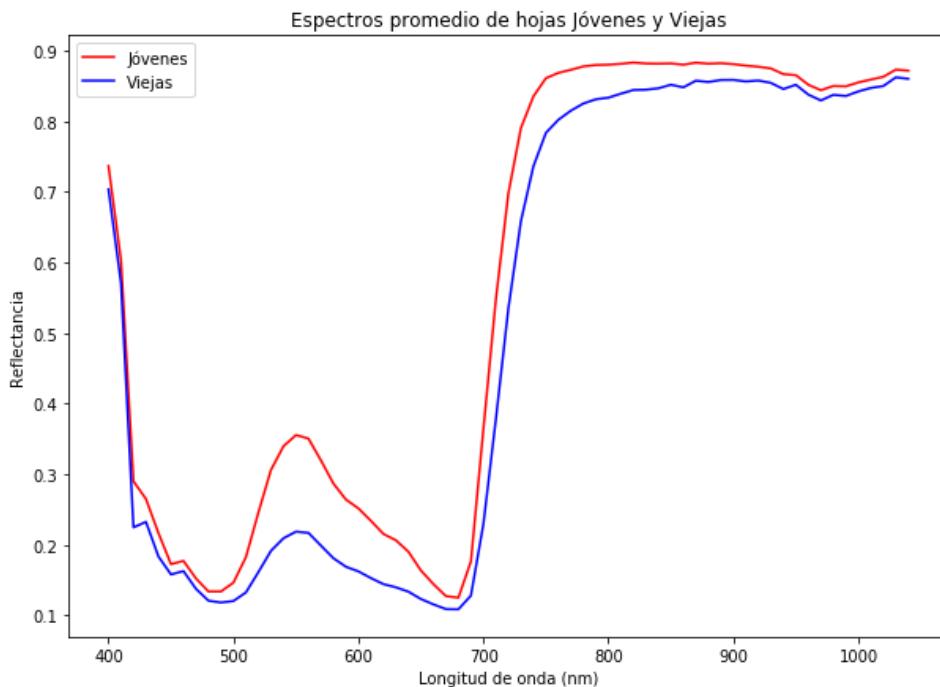


Figura 3.9. Espectro medio de las hojas jóvenes y viejas registradas

- **Espectro según la orientación**

Puede apreciarse en la Figura 3.10 que no existen diferencias evidentes entre los espectros medios según las diferentes orientaciones.

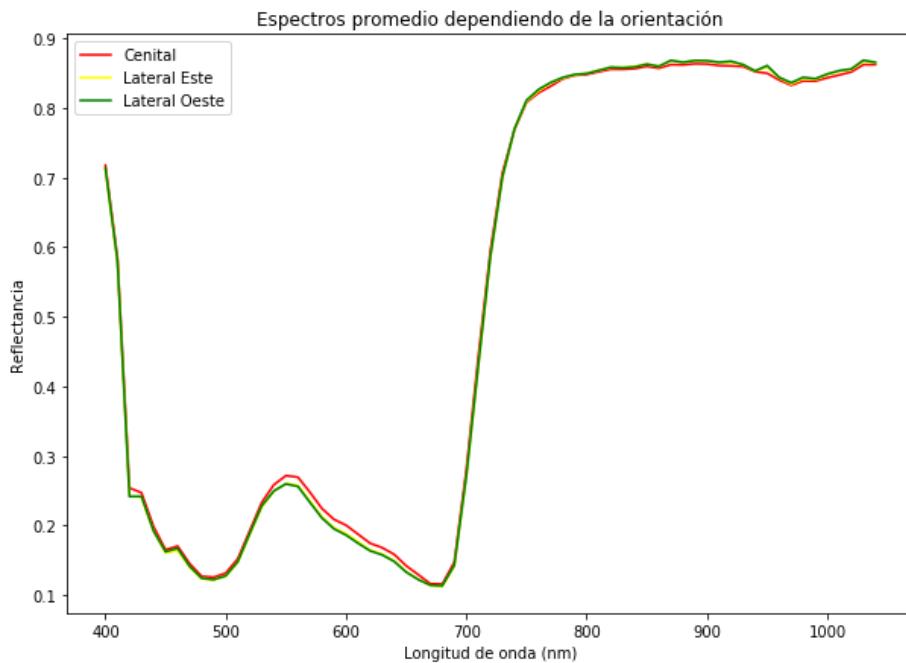


Figura 3.10. Espectro medio según la orientación en la que se registró la muestra

- **Variabilidad Bandas Espectrales**

En la Figura 3.11, se confirma lo anteriormente mencionado: las bandas espectrales entre 700 y 730nm se encuentran entre las 10 bandas con mayor variabilidad. Esta parte del espectro corresponde a la región de color rojo. Esta alta variabilidad en las bandas de color rojo puede atribuirse a diferencias en la concentración de pigmentos asociados al color rojo, como carotenoides y antocianinas.

Por otro lado, destaca la presencia de las bandas espectrales entre 530 y 580nm, que corresponden a los colores verdes del espectro visible. Esta parte del espectro también muestra cierta variabilidad, aunque en menor medida que las bandas del color rojo. Esta variabilidad en las bandas verdes puede reflejar diferencias en la concentración de clorofila, el principal pigmento responsable del color verde de las hojas.

Esta información proporciona una comprensión más detallada de las bandas espectrales que presentan mayor variabilidad en los espectros de las hojas, particularmente en las regiones del rojo y del verde.

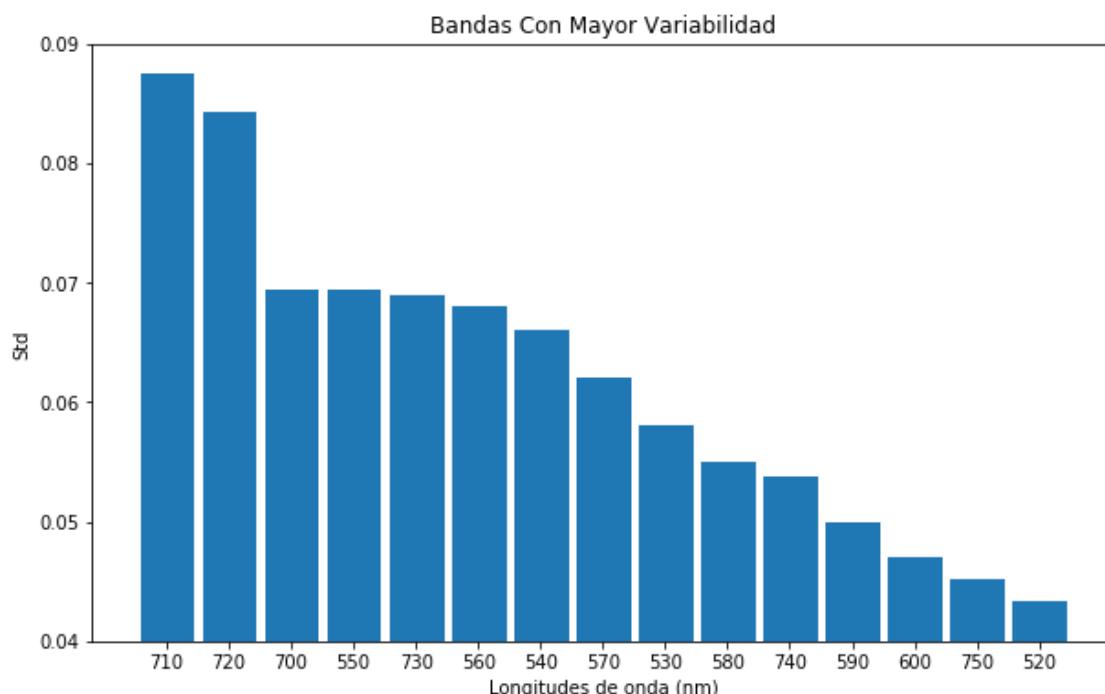


Figura 3.11. Bandas espectrales que registran mayor variabilidad

3.3. Limpieza, división y oversampling de los datos::

3.3.1. Eliminar datos faltantes

En primer lugar, se eliminan aquellos datos que no tienen un valor. En el conjunto de datos podemos identificar dos casos específicos. En el primer caso, existen instancias en las que no disponemos de ningún valor para la concentración de ningún nutriente. Estos datos no son útiles para el proceso de modelización, puesto que no aportan información para la variable respuesta. Por tanto, estos datos son eliminados completamente del estudio.

En segundo lugar, existe un grupo de datos en el que sólo falta el valor para el nutriente del nitrógeno. En este caso, estos datos serán utilizados en el proceso de modelización para todos los demás nutrientes salvo el nitrógeno. Estos datos disponibles servirán para construir modelos predictivos para los demás nutrientes basándonos en su información disponible.

Este proceso de selección de datos nos permitirá trabajar con un conjunto de datos más completo y al mismo tiempo asegurar que los datos que utilizamos son adecuados para nuestros fines de modelización.

3.3.2. Conjunto Train/Validación/Test

Ante todo, se divide el conjunto con variables explicativas (X), que serán las bandas espectrales, y variables predictoras (y), que serán los diferentes nutrientes (N, K, Ca, Fe, etc...).

Seguidamente, debe dividirse nuestro conjunto de datos en dos partes: una parte de calibración (también llamado entrenamiento) y otra de prueba (test).

- **Conjunto de calibración (*train*):** Son los datos que se utilizan para entrenar el modelo.
- **Conjunto de prueba (*test*):** Son los datos que se reservan para comprobar si el modelo que se ha generado a partir de los datos de entrenamiento funciona, y evaluar lo bien que funciona mediante métricas.

Para la división del conjunto de datos, se realiza una partición aleatoria, asignando aproximadamente el 75% de los datos al conjunto de calibración y el 25% restante al conjunto de prueba. Esta división permite garantizar que el modelo sea entrenado adecuadamente y después evaluado con datos que no ha visto durante el entrenamiento.

- **Problemas**

Un problema que surge siempre en el entrenamiento de un modelo de aprendizaje automático es el sobreajuste. El sobreajuste es un problema que puede producirse cuando los modelos de aprendizaje automático son demasiado complejos para los datos con los que se han entrenado. Esto puede conducir a un rendimiento pobre de generalización en nuevos datos, ya que el modelo ha aprendido patrones que son específicos de los datos de entrenamiento y no son generalizables a otros datos.

Hay varias formas de detectar el exceso de adaptación. Una de ellas es observar el rendimiento del modelo en los datos de entrenamiento y ver si es mucho mejor que el rendimiento en los datos de prueba. Si el modelo está sobreajustado, el rendimiento del entrenamiento será mucho mayor que el rendimiento de la prueba. Para obtener más información sobre este problema, puede consultarse la referencia [14].

Es importante tener en cuenta el sobreajuste en el entrenamiento de modelos de aprendizaje automático, ya que puede afectar negativamente a la capacidad del modelo para realizar predicciones precisas en nuevos datos. Para mitigar este problema, se pueden aplicar técnicas como la regularización, el mayor uso de conjuntos de datos, la selección de características relevantes y la validación cruzada, entre otros. Estas estrategias ayudan a controlar la complejidad del modelo y mejorar su capacidad de generalización.

- **Validación Cruzada**

Validación cruzada es la técnica escogida para evitar el problema de sobreajuste. En este proyecto concreto, se utiliza K-FOLD, que es una técnica en la que los datos se dividen en k particiones, en nuestro caso 3 (ya que no existe una gran cantidad de datos). Dos particiones se utilizan para entrenar el modelo, y una partición se utiliza para realizar la predicción (conjunto de validación), por lo que todas las particiones son utilizadas por el modelo para predecir una vez.

Posteriormente, se puede calcular la media de las métricas de evaluación del modelo conseguido con los conjuntos de validación. De esta forma, se evita el sobreajuste y también se pueden encontrar las técnicas de preprocessamiento que mejor funcionan utilizando más datos y, por tanto, obtener un funcionamiento más optimizado en un contexto más general.

Los resultados obtenidos del conjunto de validación habrán sido estimados mediante la media de la validación cruzada. En la Figura 3.12 puede observarse un ejemplo de validación cruzada con 3 particiones.

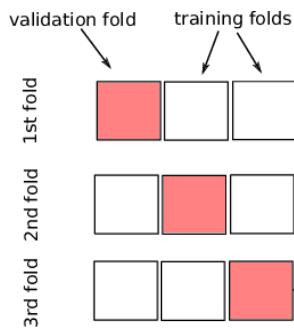


Figura 3.12. Ejemplo del proceso K-Fold con 3 particiones

Se puede encontrar este proceso de K-Fold Cross Validation detallado en [15].

3.3.3. *Oversample*

El oversampling es una técnica utilizada en aprendizaje automático para abordar el desequilibrio de clases en un conjunto de datos. En este caso, como hemos visto anteriormente, existe un mayor número de espectros de hojas viejas (ya que hay 2 grupos de hojas viejas para cada grupo de hojas jóvenes). Por tanto, aumentando el número de ejemplos de la clase de hojas jóvenes en el conjunto de datos, se puede obtener una mejora en la generalización de los modelos.

Esta técnica es especialmente útil cuando se trabaja con un número reducido de muestras, puesto que puede ayudar a mejorar el rendimiento de los modelos. Puede encontrar más información sobre cómo esta técnica puede ser beneficiosa en la referencia [16].

3.4. Preprocessament espectral

El conjunto de técnicas utilizadas antes de la aplicación de un método de Machine Learning es conocido como preprocessamiento de datos, siendo considerado una de las fases más importantes en el proceso de descubrimiento de patrones a partir de los datos. El preprocessamiento de datos consiste en la aplicación de un conjunto de técnicas para preparar adecuadamente los datos que serán utilizados como entrada por los algoritmos de minería de datos. Estas técnicas son a menudo consideradas como obligatorias, puesto que sin ellas los algoritmos de extracción de conocimiento no podrían ser ejecutados o, en otros casos, darían resultados de baja calidad.

Las técnicas de preprocessamiento de datos utilizadas en este trabajo tienen como objetivo mejorar la interpretación de las variables disponibles por los modelos. Por ejemplo, tomamos en consideración el peso de una persona (en kg) y su altura (en metros); podemos observar que una persona puede tener un peso de 80 kg y medir 1,70 m. En este caso, el valor del peso es considerablemente mayor que la altura, lo que podría afectar a un modelo de aprendizaje automático que da un peso excesivo de peso a la variable del peso en comparación con la altura.

En este trabajo, se han probado distintas técnicas de preprocessamiento de datos para evaluar cuál de ellas ofrece los mejores resultados para este problema específico. A continuación, explicaremos en qué consisten estas técnicas.

3.4.1. Standard Normal Variate (SNV)

- ***Introducción***

La primera técnica que aplicamos en las bandas espectrales es la estandarización estándar llamada también Standard Normal Variate (SNV). Una técnica de la más común, por no decir la más común, a la hora de estandarizar los datos. Esta técnica lo que hace es escalar las características para que tengan una media de 0 y una varianza de una unidad. De esta forma se normalizan las bandas espectrales estandarizándolas para uniformizar su rango de variación.

- ***Marco Teorico***

Supondremos un conjunto de datos X .

$$X = \begin{bmatrix} x_{1,1} & \dots & x_{p,1} \\ \dots & \dots & \dots \\ x_{1,n} & \dots & x_{p,n} \end{bmatrix}$$

Ecuación 3.1.1

En la ecuación 3.1.1 n son el número de observaciones, y p el nombre de características; en este proyecto, las bandas espectrales (un total de 66). Los pasos seguidos para conseguir escalar los datos son los siguientes:

En la ecuación 3.1.2 se calcula la media de cada columna.

$$\bar{X}_p = \frac{1}{n} \sum_{i=1}^n x_{p,i}$$

Ecuación 3.1.2

La ecuación 3.1.2 genera un vector en el que se tiene la media de cada columna como el de la ecuación 3.1.3.

$$\bar{X} = [\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p]$$

Ecuación 3.1.3

Se hace lo mismo para calcular la varianza:

$$\sigma_p = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{p,i} - \bar{X}_p)^2}$$

Ecuación 3.1.4

La ecuación 3.1.4, nos genera otro vector donde se tiene la varianza de cada columna (ecuación 3.1.4):

$$\sigma = [\sigma_1, \sigma_2, \dots, \sigma_n]$$

Ecuación 3.1.5

Por último, a cada valor de los datos ($x_{p,i}$) de cada columna p se le restamos la media de su columna y se divide entre la varianza de la columna para tener una media 0 y una varianza de una unidad tal y como se aprecia en la ecuación 3.1.6.

$$Z = \frac{x_{p,i} - \bar{X}_p}{\sigma_p}$$

Ecuación 3.1.6

- **Calibración y prueba**

Algo a tener en cuenta es no aplicar la estandarización sobre todo el conjunto de datos. Como tenemos un conjunto de calibración y uno de prueba, si a la hora de calcular la media y la varianza de las columnas, se incluyen los

espectros que tenemos en el conjunto de test, se está utilizando información proveniente de este conjunto lo que implica un sesgo a la hora de evaluar el funcionamiento de los modelos.

Entonces, lo que se hace es calcular estas medias y varianzas con el conjunto de calibración, y con los resultados obtenidos transformar, por separado, el conjunto de traicionero y de test.

- **Representación grafica**

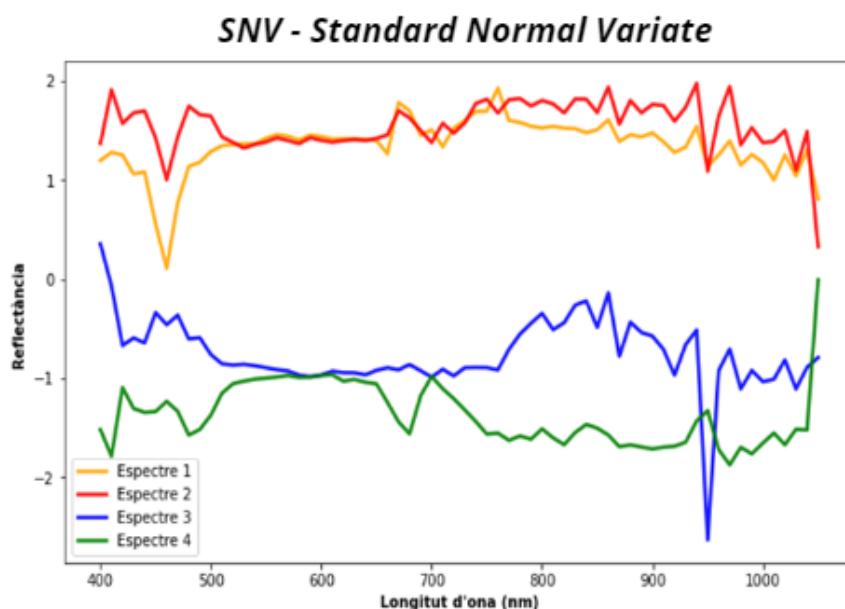


Figura 3.13. Representació grafica de SNV de 4 muestras espectrales.

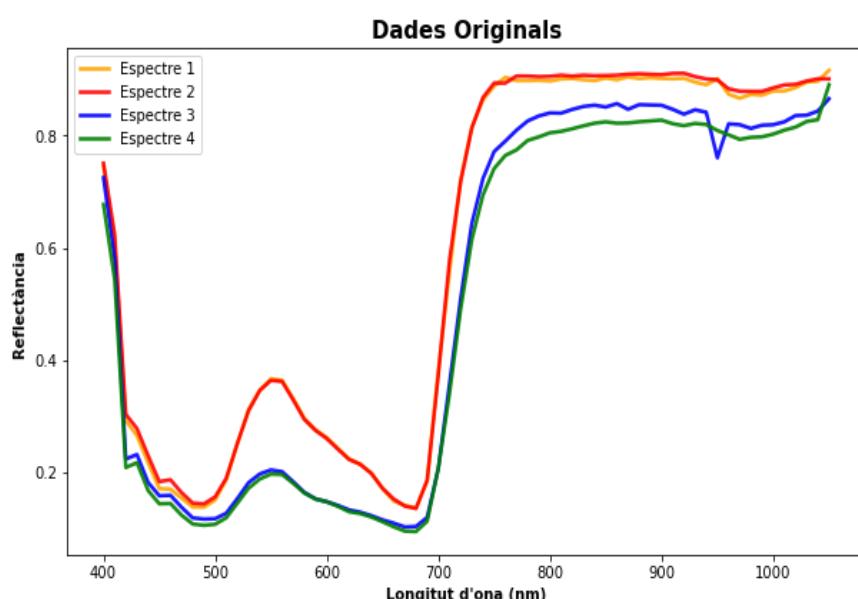


Figura 3.14. Representación grafica de 4 muestras espectrales originales.

Se puede ver en la Figura 3.13 y en la Figura 3.14 una visualización de cuatro espectros para ver cómo se transforman. Se ve en la Figura 3.13 cómo la estandarización causa que la media de todos los espectros sea 0 y, la varianza de una unidad.

Por último, se puede encontrar más información sobre SNV en la referencia [17].

3.4.2. Savitzky-Golay

- ***Introducción***

Un filtro Savitzky-Golay es un tipo de filtro que se utiliza para eliminar el ruido de los datos y suavizarlos. Las ponderaciones se calculan mediante un ajuste polinómico de los datos. Este ajuste se calcula a partir del tamaño de la ventana y del grado del polinomio. Es un filtro que puede ser de mucho interés ya que preserva las características de la distribución inicial como los máximos y mínimos relativos y los picos del espectro.

- ***Marco Teórico***

Dado un espectro con las 66 bandas tal que $x[n] = [x_1, x_2, \dots, x_j]$ on $j= 1, 2, \dots, 66$.

Se busca ir suavizando el espectro, por tanto, progresivamente se calculan los nuevos valores dados los valores de los datos. El primero en definir es el tamaño de la ventana (m), que indica el número de puntos adyacentes que se utilizan para ajustar el polinomio.

Como debe haber un punto central, del que se quiere calcular su nuevo valor, ese valor m siempre debe ser impar. Por ejemplo, si $m = 3$, se utilizan los dos puntos adyacentes al punto central para determinar el nuevo valor. Estos puntos, sabiendo que el punto central representa el 0, serían -1 y 1. Sólo se modifica este punto central y esta ventana va avanzando progresivamente en los valores de todo el espectro.

Entonces n_l denota el número de puntos a la izquierda del punto central, es decir, los negativos, y n_r denota el número de puntos a la derecha del punto central. Siendo así, el número de puntos utilizados lo definido en la ecuación 3.2.1.

$$m = n_l + n_r + 1$$

Ecuación 3.2.1

Otra condición que debe cumplir es ser inferior, o igual, al número de valores que hay en el espectro, es decir, $m \leq j$.

En segundo lugar, buscamos esa función que, para los puntos de la ventana, realiza un mejor ajuste mediante el método de mínimos cuadrados polinomial. Entonces, el segundo parámetro en decidir su valor es el grado del polinomio, k . Éste, debe ser menor que el tamaño de la ventana, es decir, $n < m$.

Una vez escogidos los valores de m y de n que se quieren utilizar, como hemos dicho, este filtro trata de encontrar la función de mínimos cuadrados polinomial, $p(x)$, que mejor se ajusta a los puntos, x tal y cómo se ha definido en la ecuación 3.2.2,

$$p_i(x) = \sum_{k=0}^{k=n} b_k \cdot x^k$$

Ecuación 3.2.2

Siendo b_0, b_1, \dots, b_k los coeficientes del polinomio definidos en la ecuación 3.2.3.

$$p_i(x) = b_0 + b_1 * x + b_2 * x^2 + \dots + b_n * x^n$$

Ecuación 3.2.3

Entonces, el punto central, es decir $x=0$, lo obtendremos a partir de la función polinómica definida en la ecuación 3.2.4.

$$p_i(x = 0) = b_0$$

Ecuación 3.2.4

Entonces, lo que queremos es encontrar aquellos coeficientes que minimicen el error de mínimos cuadrados de la función $p(x)$ anterior para los valores que tenemos dentro de la ventana. Esta función de mínimos cuadrados está definida en la ecuación 3.2.5.

$$\min \sum_{j=i-n_l}^{i+n_r} (p_i(x_j) - y_j)^2$$

Ecuación 3.2.5

El cual se halla resolviendo la siguiente ecuación 3.2.6 que es la derivada parcial respecto a b_0 .

$$\frac{\partial}{\partial b_0} \left[\sum_{j=i-n_l}^{i+n_r} [p_i(x_j) - y_j]^2 \right] = 0$$

Ecuación 3.2.6

- **Calibración y prueba**

En ese caso, este filtro transforma cada espectro independientemente de los siguientes, es decir, a todos los espectros les aplica la misma transformación. En este caso, no se hará distinción alguna entre el conjunto de entrenamiento y prueba siempre que los valores de m y n sean los mismos para ambos conjuntos.

- **Selección de valores**

Para definir el orden del polinomio y la longitud de la ventana, ante todo, debemos tener clara cuál es nuestra finalidad. En este caso, lo que se busca es encontrar un espectro suavizado, es decir, simplificar ese espectro eliminando las variaciones manteniendo la distribución.

Vemos en la Figura 3.15 un espectro y el mismo espectro suavizado. Se observa que en algunas bandas, sobre todo las finales, el espectro original (color azul) tiene pequeñas alteraciones en sus valores.

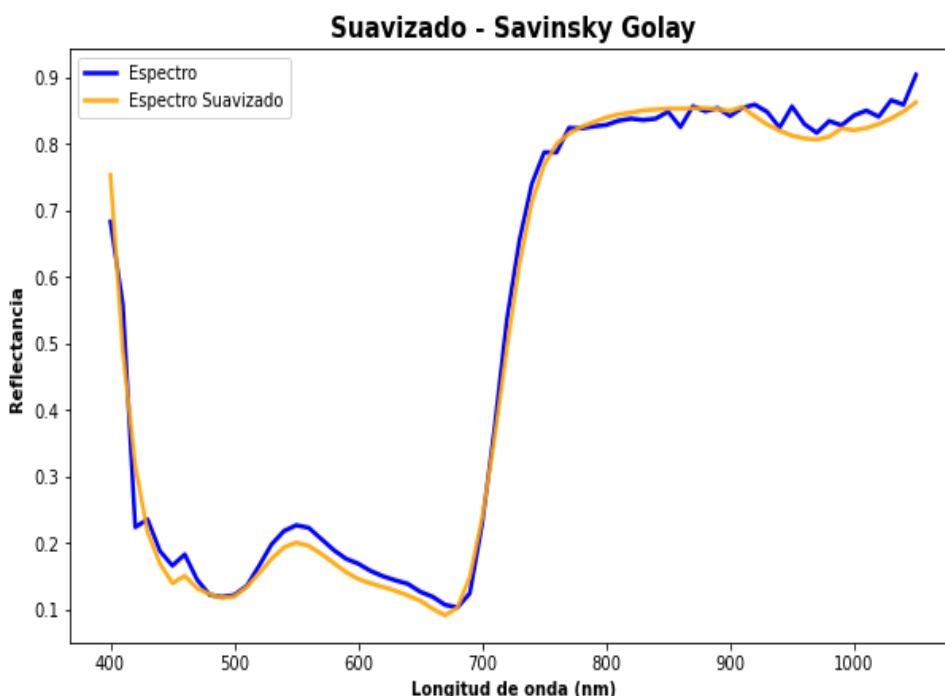


Figura 3.15. Representación gráfica de una muestra espectral transformada con Savitsky-Golay y la propia muestra espectral original.

Entonces, se han realizado diferentes pruebas de selección de valores, viendo cuáles se adaptaban en mayor medida a lo que buscábamos. Por último, se ha decidido que el tamaño de la ventana sea de longitud $n = 9$, y el orden del polinomio sea de $k = 3$. Este criterio ha sido definido de forma subjetiva valorando cuál suavizado, tanto en espectros de hojas jóvenes, como de hojas viejas, mantenía la distribución original y eliminaba los picos.

- **Representación grafica**

A continuación, en la Figura 3.16 podemos observar cómo estos valores modifican nuestro espectro hacia un espectro más suavizado en comparación con lo que encontramos en la Figura 3.14 donde tenemos los espectros originales.

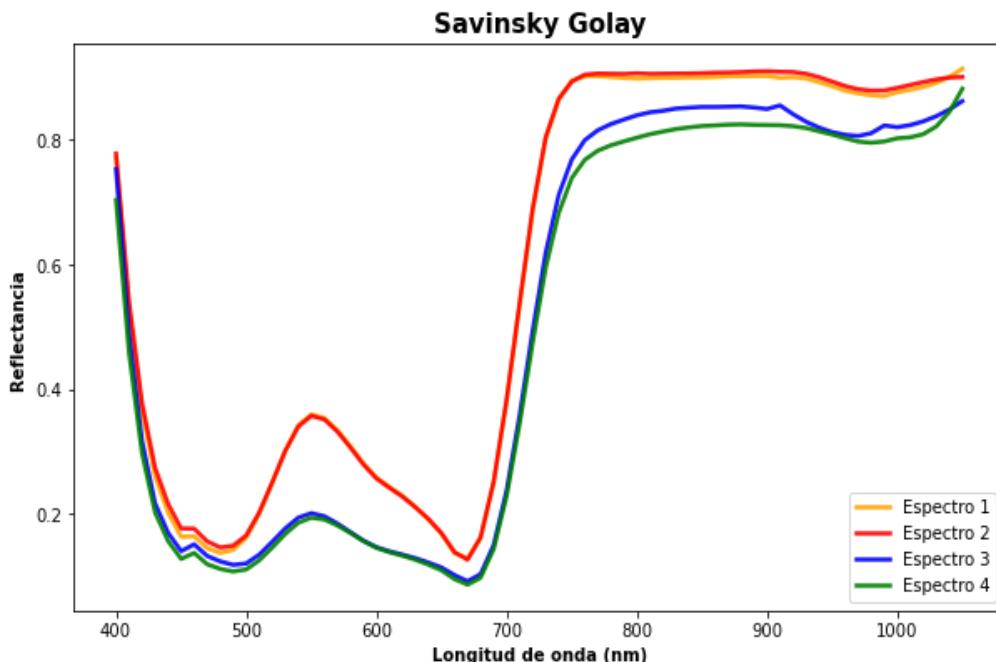


Figura 3.16. Representación de 4 muestras espectrales transformadas con Savitsky-Golay.

Se puede indagar más sobre Savitzky-Golay en la referencia [18].

3.4.3. Savitzky Golay Primera Derivada

- **Introducción**

El filtro de Savitzky Golay utilizando derivadas lo que hace es aumentar el orden de la derivada a calcular. En el caso anterior, en el que no se utiliza la derivada, este orden es 0; en caso de utilizar la primera derivada, el orden de la derivada parcial será 1.

- **Marco Teorico**

Como sabemos, que para ajustar la función polinómica utilizamos la ecuación 3.2.2. Entonces, en este caso donde se utilizan las derivadas, la función polinomio anterior es sustituida por la derivada de su función según el orden de la derivada tal y como se ve en la ecuación 3.3.1.

$$\frac{d p_i}{d x} = b_1 + 2b_2 \cdot x_2 + 3b_3 \cdot x^2 + \dots + nb_n \cdot x^{n-1}$$

$$\frac{d^2 p_i}{d x^2} = 2b_2 + 3 \cdot 2b_3 \cdot x + \dots + (n-1) \cdot nb_n \cdot x^{n-2}$$

.....

$$\frac{d^n p_i}{d x^n} = n! \cdot b_n$$

Ecuación 3.3.1

Entonces, en caso de utilizar la primera derivada, el valor central será definido por la ecuación 3.3.2.

$$\frac{d p_i(0)}{d x} = b_1$$

Ecuación 3.3.2

Lo que se hace, por tanto, es encontrar el mínimo del error cuadrático entre los datos y la función polinómica como se ve en la ecuación 3.3.3:

$$\min \sum_{j=i-n_l}^{i+n_r} [\frac{d p_i}{d x} - y_j]^2$$

Ecuación 3.3.3

Este mínimo se encuentra en función de la derivada de la función polinomio, y del coeficiente b_1 como muestra la ecuación 3.3.4.

$$\frac{\partial}{\partial b_1} \left[\sum_{j=i-n_l}^{i+n_r} (\frac{d p_i}{d x} - y_j)^2 \right] = 0$$

Ecuación 3.3.4

- ***Calibración y prueba***

En este caso, tampoco se hace distinción alguna entre el conjunto de calibración y prueba, siempre que los valores de m y n sean los mismos para ambos conjuntos.

- ***Selección de valores***

En el caso del filtro Savitzky Golay aplicando la primera derivada, utilizamos los mismos valores que en el caso anterior, donde conseguían suavizar el espectro.

- **Representación gráfica**

Se ve en la Figura 3.17 que, utilizando la primera derivada, el espectro es modificado por completo. Además, observando las bandas finales, se ve cómo el filtro elimina las pequeñas variaciones que tiene el espectro original.

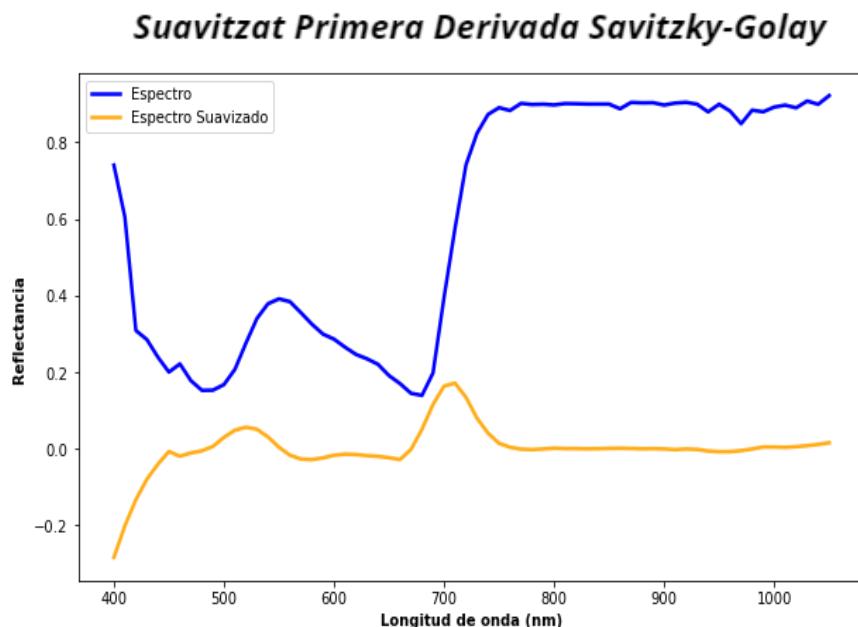


Figura 3.17. Representación gráfica de una muestra espectral transformada con la primera derivada de Savinsky-Golay y la propia muestra espectral original.

3.4.4. Savitzky Golay + SNV

- **Introducción**

En este caso, lo que hacemos es aplicar la estandarización comentada anteriormente a los resultados obtenidos a partir del filtro suavizado.

- **Calibración y prueba**

En este caso, primero se aplica el filtro Savitzky Golay, con los valores comentados anteriormente, a los datos originales. Como se ha dicho, no se realiza ninguna división entre calibración y prueba (siempre que los valores aplicados sean los mismos para ambos conjuntos).

Seguidamente, lo que se hace es aplicar la estandarización. En este caso, sí se realiza esa división entre los conjuntos. Se calculan los valores con el conjunto de calibración, y posteriormente se transforman los dos conjuntos.

- **Representación Gráfica**

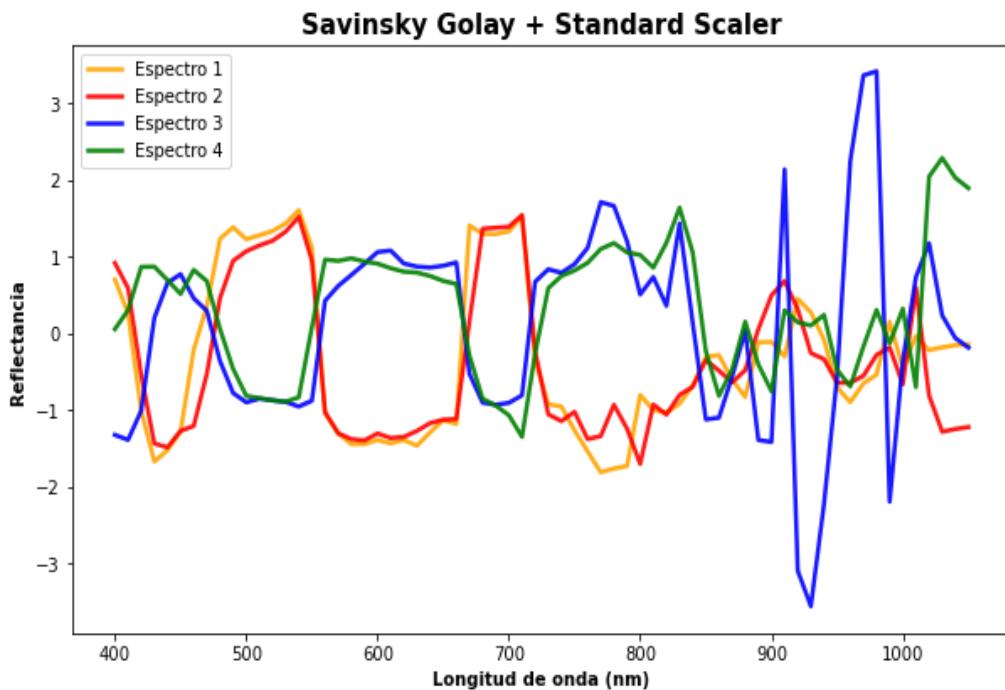


Figura 3.18. Representación gráfica de 4 muestras espetrales transformadas con Savitsky-Golay y la estandarización estándar.

3.4.5. Promedio espectral

- **Introducción**

La transformación media del espectro es un tipo de escala de características que implica centrar la distribución de bandas espetrales en torno a cero. Esta transformación implica restar el espectro medio de todo el conjunto de datos a cada espectro individual.

- **Calibración y prueba**

El espectro medio se obtiene calculando el espectro medio de todas las muestras del conjunto de calibración, y una vez lo tenemos calculado, se resta de cada espectro individual del de los datos. Esto da lugar a un nuevo conjunto de datos que consiste en las diferencias entre cada espectro individual en el espectro medio.

Por tanto, bandas espetrales se desplazan por tener un valor medio de cero y una distribución más uniforme. Esto puede ayudar a reducir el efecto del ruido y la variabilidad en las bandas espetrales y hacer que los modelos sean más robustos en los cambios en los valores de entrada.

- **Representación Grafica**

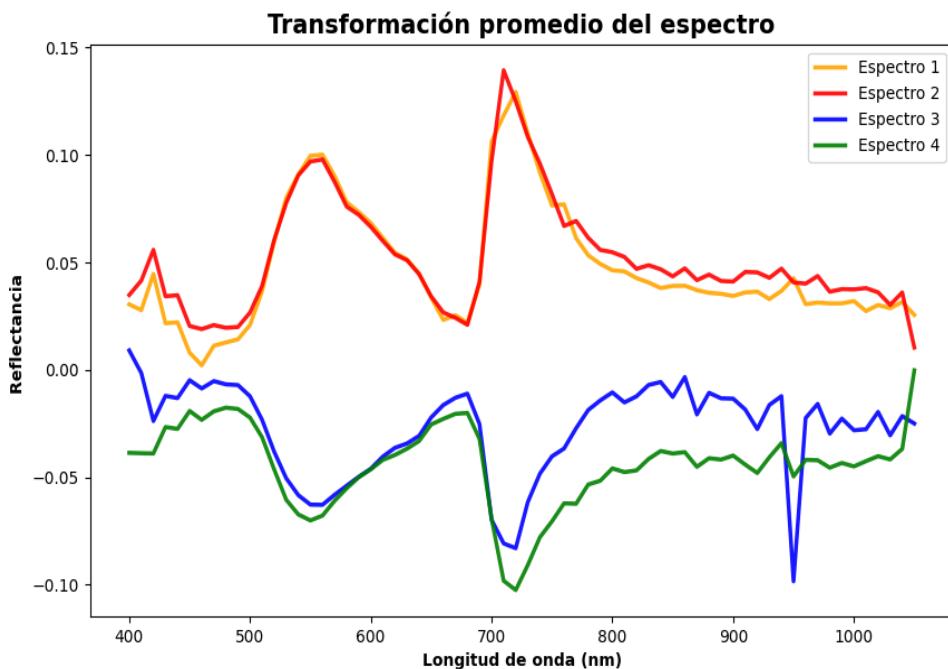


Figura 3.19. Representación gráfica de 4 muestras espectrales transformadas con la diferencia respecto al espectro promedio.

3.4.6. Sin Transformación

También se utilizan los datos originales para comprobar si estos procesamientos son útiles y en qué casos pueden funcionar mejor los datos sin realizar ninguna transformación.

3.5. Reducción de la dimensionalidad

Cada columna puede entenderse como una variable. La reducción de dimensionalidad implica disminuir el número de variables. Esta reducción puede simplificar la información y hacerla más comprensible. Por lo general, en presencia de información redundante, esto puede ayudar al buen funcionamiento del modelo de aprendizaje automático.

Existen diversas técnicas para conseguir este objetivo de reducir el número de variables manteniendo la mayor cantidad de información posible. En este estudio, se han utilizado las técnicas que se describen a continuación.

3.5.1. Análisis de componentes principales (PCA)

- ***Introducción***

El análisis de componentes principales (PCA) es una técnica estadística que se utiliza para reducir la dimensionalidad de los datos. Es un método que trata de encontrar los componentes o variables que maximizan la varianza en un conjunto de datos. La PCA tiene diferentes aplicaciones, una de ellas es para reducir la dimensionalidad de los datos, pero otras pueden ser eliminar el ruido, analizar tendencias o, como hemos visto, encontrar cluster o grupos en un conjunto de datos. La PCA tiene una gran aplicación en el mundo científico y se utiliza en multitud de proyectos. Encontramos este método detallado en [19].

La PCA es una transformación lineal que proyecta los datos sobre un nuevo conjunto, que son sus principales componentes. El primer componente principal es la recta que maximiza la varianza de los datos, siendo el segundo componente principal la recta que maximiza la varianza de los datos siendo ortogonal al primer componente principal.

- ***Marco Teórico***

Teniendo un conjunto de datos X , que es una matriz de $m \times n$ tal y como se ve en la ecuación 3.4.1.

$$X_{m \times n} = \begin{bmatrix} x_{1,1} & \dots & x_{m,1} \\ \dots & \dots & \dots \\ x_{1,n} & \dots & x_{m,n} \end{bmatrix}$$

Ecuación 3.4.1

Tenemos que m son el número de observaciones, i p el número de características; en este caso, las bandas espectrales (un total de 66). Lo que queremos es transformar esta matriz X en otra matriz que llamaremos Y , mediante una matriz de $m \times m$ llamada P , tal y como se aprecia en la ecuación 3.4.2.

$$Y = PX$$

Ecuación 3.4.2

Si se consideran las filas de P como un vector p_1, p_2, \dots, p_m y las columnas de X como vectores x_1, x_2, \dots, x_n tal y como aparece a la ecuación 3.4.3.

$$PX = (Px_1, Px_2, \dots, Px_n) = \begin{bmatrix} p_1 \cdot x_1 & p_1 \cdot x_2 & \dots & p_1 \cdot x_n \\ \dots & \dots & \dots & p_2 \cdot x_n \\ p_m \cdot x_1 & p_m \cdot x_2 & \dots & p_m \cdot x_n \end{bmatrix} = Y$$

Ecuación 3.4.3

Entonces, estas filas de P se convertirán en las direcciones de los componentes principales. ¿Cuáles deben ser los valores de P para re-expresar X de forma óptima? Cabe recordar que PCA trata de descorrelacionar los datos originales encontrando las direcciones donde la varianza es maximizada para definir estos valores de P.

Entonces, se calcula la matriz de covarianza. El primer paso es centrar los datos en 0. Como se ha visto antes, se calcula la media de cada variable de nuestros datos como en la ecuación 3.4.4.

$$\bar{x}_m = \frac{1}{n} \sum_{i=1}^n x_{m,i}$$

Ecuación 3.4.4

Se genera el siguiente vector representado en la ecuación 3.4.5.

$$\bar{X} = [\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p]$$

Ecuación 3.4.5

Posteriormente, se obtiene la media de los datos como en la ecuación 3.4.6.

$$B = X - \bar{X}$$

Ecuación 3.4.6

Entonces, dados nuestros datos originales centrados en media 0 que tenemos en B y, definiendo la siguiente matriz X que tenemos en la ecuación 3.4.7.

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ x_{2,1} & x_{2,2} & \dots & x_{2,n} \\ \dots & \dots & \dots & \dots \\ x_{m,1} & x_{m,2} & \dots & x_{m,n} \end{pmatrix}$$

Ecuación 3.4.7

Se puede representar como en la ecuación 3.4.8 cada observación como un vector x_1, x_2, \dots, x_n . Por ejemplo, x_i es un vector de n muestras para la i variable.

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Ecuación 3.4.8

Ahora bien, lo que se quiere es calcular la matriz de covarianza. La covarianza entre dos vectores viene definida a partir de la siguiente ecuación 3.4.9.

$$\sigma_{x_1, x_2}^2 = \frac{1}{n-1} x_1 x_2^T$$

Ecuación 3.4.9

Entonces, para calcular la covariancia entre todas las muestras que tenemos en el conjunto de datos, tenemos la siguiente ecuación 3.4.10.

$$C_X = \frac{1}{n-1} X X^T = \frac{1}{n-1} \begin{pmatrix} x_1 x_1^T & x_1 x_2^T & \cdots & x_1 x_m^T \\ x_2 x_1^T & x_2 x_2^T & \cdots & x_2 x_m^T \\ \cdots & \cdots & \cdots & \cdots \\ x_m x_1^T & x_m x_2^T & \cdots & x_m x_m^T \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Equació 3.4.10

Como puede verse, la matriz anterior es una matriz simétrica, ya que la varianza entre dos vectores x e y es igual a la varianza entre y e x . Esta matriz describe todas las relaciones entre los posibles pares de mediciones en nuestros datos.

Como se está hablando de matrices de covarianza, cabe recordar que, al ser unas matrices simétricas, éstas son diagonalizables (teorema de Schur). Esto significa que mediante una transformación lineal, la matriz puede reducirse a una forma diagonal. De forma matemática se expresa tal y como se ve en la ecuación 3.4.11.

$$A = P D P^{-1}$$

Ecuación 3.4.11

En esta expresión, la matriz A puede descomponerse mediante una matriz invertible P , donde los vectores columna son vectores propios de A, i D éEs una matriz diagonal formada por los valores propios de A. Los vectores propios son vectores que no cambian la dirección del vector original cuando aplicamos una transformación lineal, el resultado de estos vectores propios son los valores propios, que representaremos con λ .

Entonces, volviendo al objetivo, se quiere reducir la redundancia a los datos, es decir, se busca que cada variable se correlacione lo menos posible con las demás variables. Como hemos dicho al principio, queremos conseguir la siguiente transformación: $Y = PX$ (ecuación 3.4.2).

Pues entonces, lo que debe definirse para conseguir el objetivo es que la matriz C_Y todos los términos no diagonales son lo más próximos posible a 0, definido en la ecuación 3.4.12.

$$C_Y = \frac{1}{n-1} YY^T = \frac{1}{n-1} PX(PX)^T = \frac{1}{n-1} P(XX^T)P^T = \frac{1}{n-1} PAP^T$$

Ecuación 3.4.12

Podemos ver que se ha definido en la ecuación 3.4.11 que $A = XX^T$ donde A es simétrica, por tanto, está diagonalizada por la matriz de vectores propios. Los componentes principales de X serán los vectores propios de XX^T , es decir, los vectores propios de la matriz de covariancia C_X . Lo representamos en la siguiente ecuación 3.4.13.

$$(X^T X)v = \lambda v$$

$$C_X v = \lambda v$$

Ecuación 3.4.13

Como hemos dicho, λ son los valores propios asociados al vector propio v de la matriz de covariancia C_X . Los valores propios de C_X son las raíces de la ecuación característica (ecuación 3.4.14).

$$|C_X - \lambda I| = 0$$

Ecuación 3.4.14

Una vez se ha resuelto la expresión anterior, se pueden obtener los vectores propios. Llegado a este punto, se puede calcular el vector propio de cada valor propio tal y como vemos en la ecuación 3.4.15.

$$(C_X - \lambda I) v = 0$$

Equació 3.4.15

De esta forma, ya se tienen los valores propios y los correspondientes vectores propios. En términos generales, los vectores propios con los valores propios más altos contienen la mayor cantidad de información (mayor varianza capturada) sobre la distribución de los datos, y estos son los que se buscan para reducir la dimensionalidad. El enfoque común es clasificar los vectores propios de mayor a menor valor propio correspondiente y elegir los que vectores propios superiores.

Lo que buscamos es proyectar los datos originales en las direcciones descritas por los componentes principales. Debido a que tenemos la relación $v = P^T$, esto es simplemente lo que se puede ver en la ecuación 3.4.16.

$$Y = v^T X$$

Ecuación 3.4.16

Por tanto, lo que hay que hacer es escoger el número de componentes principales. Cuanto mayor sea el número de componentes principales, mayor varianza explicada de los datos pero menos simplificada se tiene la información.

- ***Calibración y prueba***

Cuando se aplica la PCA, debe entrenarse con el conjunto de calibración, ya que no podemos calcular los vectores propios de todos los datos, debido a que estamos incorporando información del conjunto de prueba. Lo que hacemos es, calcular vectores y valores propios de los datos de calibración, transformando los dos conjuntos a partir de estos vectores y valores propios que han sido obtenidos.

- ***Número de Componentes***

Como se ha comentado anteriormente, a mayor número de componentes escogemos, mayor será la varianza explicada de los datos. Como podemos ver en la siguiente Figura 3.20, con siete componentes superamos ya el 99% de la varianza de los datos.



Figura 3.20. Varianza explicada de las componentes de PCA.

Para realizar pruebas para ver que técnicas de procesamiento son mejores estimando los diferentes nutrientes, probaremos diferentes números de componentes que tengan diferentes varianza explicada sobre los datos.

- **Representación Gráfica**

Se observa en la Figura 3.21 las dos primeras componentes de la PCA del conjunto de calibración, donde pueden apreciarse dos grupos claramente diferenciados, correspondientes a las hojas nuevas y las hojas viejas. Esta representación gráfica nos permite identificar visualmente la presencia de dos tipos de hojas, aun sin tener previa información sobre su clasificación. La PCA nos proporciona una forma efectiva de visualizar y comprender las relaciones y variaciones en los datos, facilitando la detección de patrones y agrupaciones.

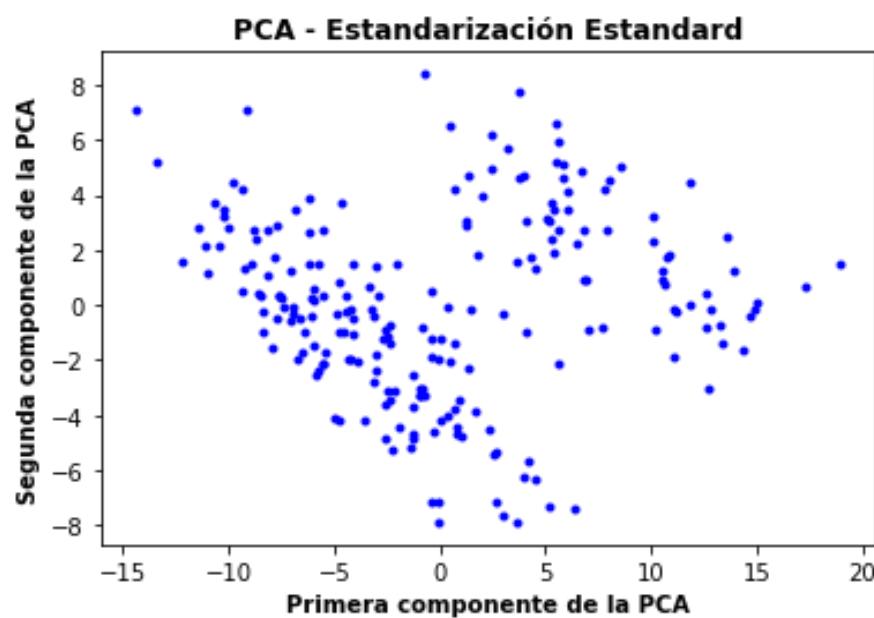


Figura 3.21. Primera componente vs. segunda componente de los datos estandarizados.

3.5.2. Autoencoder

- ***Introducción***

Las redes neuronales tienen gran relevancia en la actualidad y están transformando la inteligencia artificial en diversos ámbitos, como la sanidad, la educación, la industria y la agricultura, gracias a sus múltiples aplicaciones y su versatilidad. Las redes neuronales son capaces de abordar problemas complejos como el reconocimiento de imágenes, el procesamiento del lenguaje natural o la predicción de series temporales [20].

Un autoencoder es una arquitectura de red neuronal que tiene como objetivo comprimir la información de entrada y reproducirla con precisión en la salida, utilizando una representación reducida de la información en el interior. Básicamente, el autoencoder consta de dos partes: un encoder y un decoder. El encoder transforma la entrada original en una representación latente de baja dimensión, mientras que el decoder intenta reconstruir la entrada original a partir de la representación latente.

Para entender cómo funciona un autoencoder, es importante tener en cuenta algunos conceptos básicos de las redes neuronales. Continuamos con una breve explicación sobre estos conceptos.

- ***Neurona***

Para comprender el funcionamiento de una neurona artificial y una red neuronal, es útil conocer una breve introducción biológica sobre las neuronas. A finales del siglo XIX y principios del siglo XX, los científicos realizaron numerosos descubrimientos importantes sobre las neuronas.

Uno de los científicos más destacados de esa época fue Santiago Ramón y Cajal, un investigador español conocido principalmente por sus trabajos sobre el sistema nervioso. Publicó un artículo detallado que describe la estructura de las neuronas, que contribuyó a establecer el campo de la neurociencia y abrir camino a futuros descubrimientos sobre el cerebro. El trabajo de Cajal fue innovador, puesto que demostró que las neuronas son células individuales y diferenciadas, una desviación importante de la visión predominante en la época, que consideraba que las neuronas formaban parte de una red continua.

Este trabajo contribuyó a establecer la visión moderna del cerebro como un complejo sistema de células interconectadas. Hoy en día sabemos que el sistema nervioso es increíblemente complejo y que las neuronas son las unidades fundamentales que procesan y transmiten la información a través del sistema nervioso.

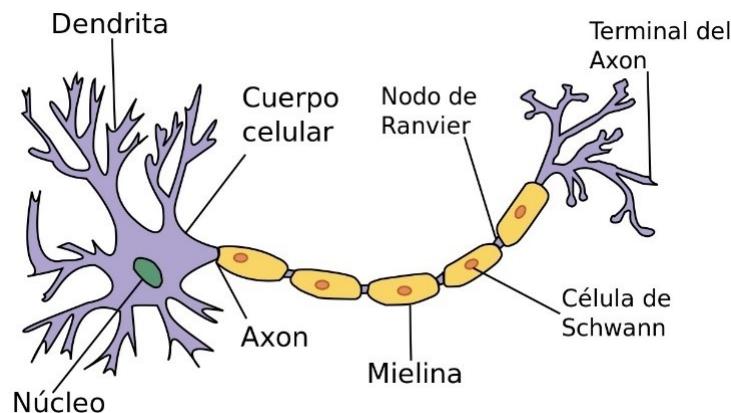


Figura 3.22. Partes de la neurona humana. Fuente: Enciclopedia Humanidades.

Las neuronas son las unidades básicas del sistema nervioso y juegan un papel fundamental en la transmisión de la información a lo largo del cuerpo. Como puede apreciarse en la Figura 3.22, una neurona está compuesta por un cuerpo celular, un axón y dendritas.

Las neuronas se comunican entre sí mediante señales. Cuando una señal eléctrica llega al axón, desencadena la liberación de neurotransmisores, que son sustancias químicas. Estos neurotransmisores viajan a través de las sinapsis y se unen a los receptores de las dendritas de la siguiente neurona. Este proceso permite la transmisión de la señal eléctrica de una neurona a otra.

El cerebro humano contiene miles de millones de neuronas que están organizadas en una compleja red. Esta estructura intrincada permite el procesamiento y el tratamiento de la información a lo largo del sistema nervioso, jugando un papel esencial en las funciones cognitivas, sensoriales y motoras de los seres humanos.

- **Neurona artificial**

En 1943, Warren McCulloch y Walter Pitts publicaron un artículo titulado "A Logical Calculus of the Ideas Inmanent in Nervous Activity" en el que propusieron un modelo de neurona artificial. La neurona de McCulloch y Pitts es un modelo matemático simple que imita el comportamiento de una neurona.

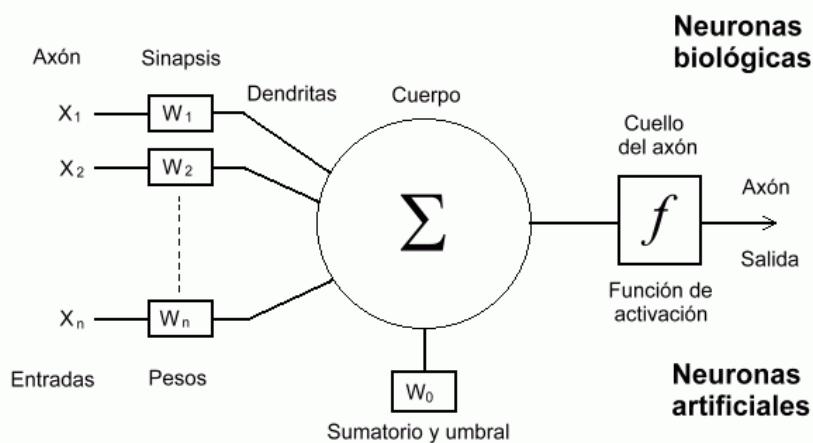


Figura 3.23. Partes de la neurona artificial. Fuente: Universidad de Murcia.

Como se muestra en la Figura 3.23, los pesos y función de activación trabajan conjuntamente en una neurona artificial. Los pesos determinan la contribución de cada entrada en la salida, y la función de activación determina el valor de salida de la neurona. Los pesos pueden ser positivos o negativos, y pueden ajustarse para cambiar la salida de la neurona. La salida de la neurona se calcula mediante la suma del producto de cada entrada por su peso. Esta suma se pasa a través de la función de activación, que transforma la suma en un valor específico. Existen varias funciones de activación que funcionan mejor o peor en función del problema que se desea resolver. En la Tabla 3.3 se muestran algunas de las más populares.

Name	Plot	Function, $f(x)$	Derivative of f , $f'(x)$	Range
Identity		x	1	$(-\infty, \infty)$
Binary step		$\begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$	$\begin{cases} 0 & \text{if } x \neq 0 \\ \text{undefined} & \text{if } x = 0 \end{cases}$	$\{0, 1\}$
Logistic, sigmoid, or soft step		$\sigma(x) = \frac{1}{1 + e^{-x}}$ [1]	$f(x)(1 - f(x))$	$(0, 1)$
tanh		$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$	$1 - f(x)^2$	$(-1, 1)$
Rectified linear unit (ReLU) ^[11]		$\begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } x > 0 \end{cases} = \max\{0, x\} = x1_{x>0}$	$\begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x > 0 \\ \text{undefined} & \text{if } x = 0 \end{cases}$	$[0, \infty)$

Tabla 3.3. Principales funciones de activación. Fuente: Sagar Sharma en Activation Functions in Neural Networks (Medium)

- **Perceptrón Simple**

Ahora que se ha explicado el funcionamiento de las partes de una neurona, es importante entender cómo ésta aprende los pesos necesarios para resolver el problema. En este punto entra en juego el diseño del Perceptrón Simple.

El perceptrón fue inventado a principios de los años 50 por Frank Rosenblatt, quien se inspiró en los trabajos anteriores de Warren McCulloch y Walter Pitts. Rosenblatt diseñó una regla sencilla para que el perceptrón ajuste sus pesos: considerando un conjunto de entrenamiento formado por ejemplos etiquetados, la regla de aprendizaje actualiza los pesos del perceptrón para minimizar el número de ejemplos mal clasificados.

La actualización está expresada por la ecuación 3.5.1.

$$w_i \leftarrow w_i + \eta \cdot (i - \hat{y}) \cdot x_i$$

Ecuación 3.5.1

donde w_i es el peso i -ésimo, η es una tasa de aprendizaje, i es la etiqueta verdadera del ejemplo, y \hat{y} es la etiqueta predicha.

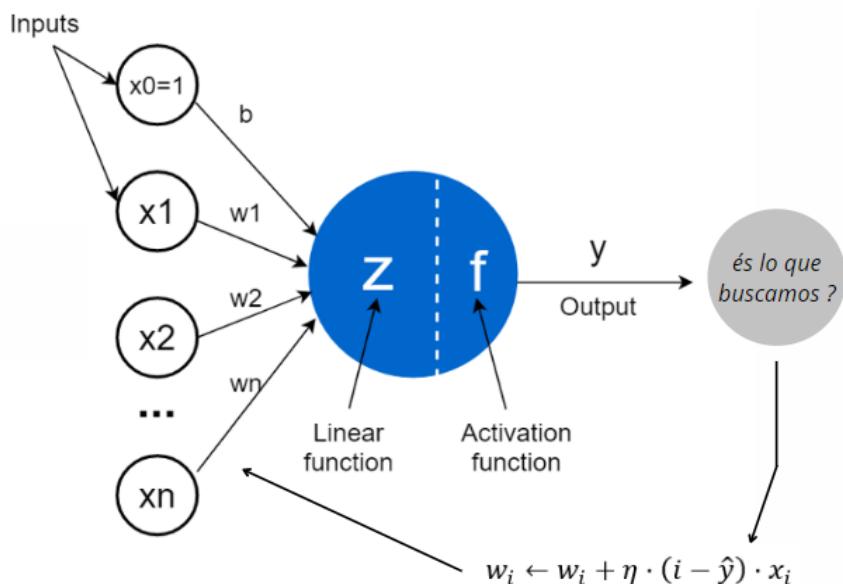


Figura 3.24. Perceptrón Simple. Fuente: Universidad de Murcia

Puede verse, en la Figura 3.24, que si el valor predicho corresponde al valor real, es decir, $i - \hat{y} = 0$, entonces no se actualizarán los pesos. En caso de que logremos minimizar el error, o incluso, que éste sea 0, diremos que el perceptrón (o la red neuronal) ha convergido.

- **Red Neuronal**

El perceptrón tiene una limitación inherente: sólo puede resolverse problemas linealmente separables. Para afrontar problemas linealmente no separables, es necesario utilizar múltiples perceptrones conectados a una red neuronal. Esto fue descrito en el libro "Perceptrons" de Minsky y Papert en 1969. Sin embargo, la creación de esta red plantea el problema conocido como "asignación de crédito", puesto que no sabemos en qué proporción asignar el error a cada perceptrón.

Posteriormente, se descubrió que utilizando funciones de activación diferenciables se podían actualizar los coeficientes mediante la regla de la cadena. También se incorporaron las funciones de coste. Existen varias funciones de coste distintas que se pueden utilizar en una red neuronal, como el error cuadrático medio (MSE) que mide la diferencia cuadrática media entre las predicciones y los valores reales, y la función de coste de entropía cruzada (Cross Entropy) que se utiliza en problemas de clasificación, entre otros.

Para actualizar los coeficientes de la red neuronal, también se utilizan distintos optimizadores. Algunos de los optimizadores más conocidos son Adam, RMSprop, Adadelta, SGD, entre otros. Estos optimizadores ayudan a encontrar los valores óptimos de los coeficientes para minimizar la función de coste y mejorar el rendimiento de la red neuronal.

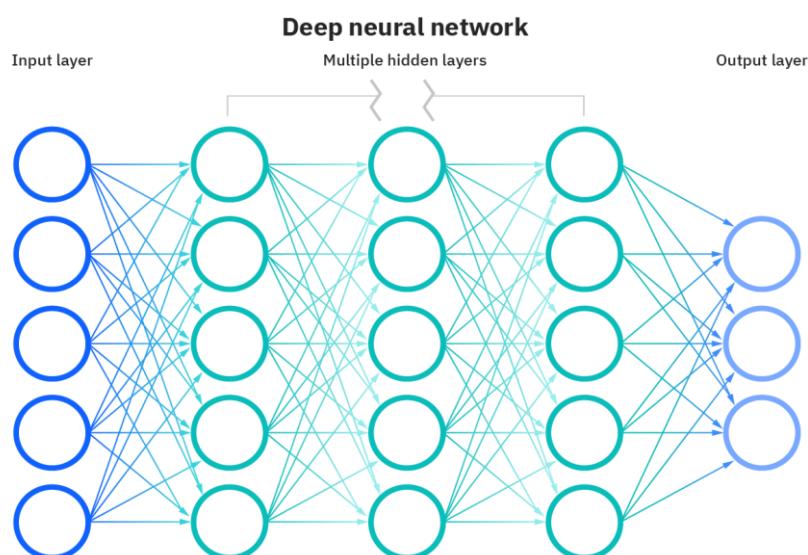


Figura 3.25. . Red Neuronal Profunda. Fuente: IBM Blog "What are neural networks?"

Existen varios tipos de redes neuronales, pero todas comparten una estructura básica común. A nivel general, una red neuronal está formada por nodos de entrada, nodos de salida y nodos ocultos. Los nodos de entrada reciben los datos de entrada y los procesan a través de los nodos ocultos. La

salida de los nodos ocultos se transmite a los nodos de salida que producen la salida final de la red neuronal.

Durante el proceso de entrenamiento de la red neuronal, las conexiones entre los nodos se ajustan en función de los resultados obtenidos. Esto se realiza mediante una técnica llamada retropropagación, donde la red ajusta pesos y sesgos para minimizar el error en las predicciones.

A pesar de sus logros, las redes neuronales tienen algunas limitaciones. A menudo se consideran modelos de caja negra, ya que puede ser difícil comprender el razonamiento que utilizan para llegar a sus predicciones. Además, las redes neuronales pueden requerir una gran capacidad computacional y una gran cantidad de datos de entrenamiento para aprender de forma eficiente.

Para profundizar en estos conceptos, se puede consultar la referencia [22], donde se analizan detalladamente la neurona biológica, la neurona artificial, el perceptrón y las redes neuronales, así como el estado actual de la investigación en este ámbito.

- ***Autoencoder***

Un autoencoder es una red neuronal utilizada para aprender representaciones eficientes de los datos. El objetivo principal de un autoencoder es transformar los datos de entrada en un espacio latente de menor tamaño que los datos originales. Para ello, el autoencoder aprende a comprimir los datos de entrada en un espacio latente y después reconstruye los datos originales a partir de ese espacio latente.

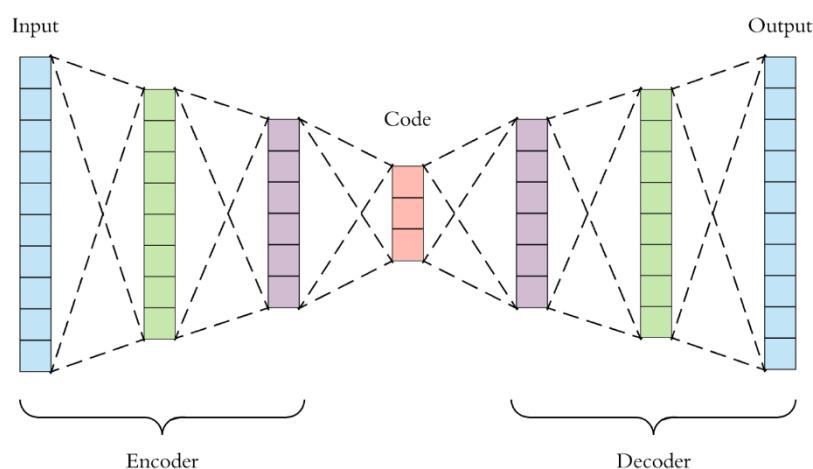


Figura 3.26. Estructura General de un Autoencoder. Fuente: Fernando Sancho Caparrini Blog.

Los autoencoders poseen diversas aplicaciones, como la reducción de la dimensionalidad de los datos, la eliminación de ruido en los datos, la generación de nuevos datos o la detección de anomalías. Son un tipo de

algoritmo de aprendizaje no supervisado, puesto que no requieren etiquetas para entrenar el modelo.

Aunque existen diferentes tipos de autoencoders, todos comparten una estructura básica similar. Tienen una capa de entrada, una capa oculta y una capa de salida. La capa oculta suele ser de menor tamaño que las capas de entrada y salida, lo que permite que el autoencoder funcione como algoritmo de compresión. Además, una capa oculta más pequeña facilita el aprendizaje de vectores latentes que capturan las características esenciales de los datos de entrada.

Hay que tener en cuenta que el número de neuronas seleccionadas en el espacio latente determina el número de variables con las que se representarán los datos originales. Por ejemplo, si hay tres neuronas en el espacio latente, esto significa que los datos se transformarán en tres variables.

Para obtener más información sobre los distintos tipos de autoencoders, sus aplicaciones, casos de uso y su funcionamiento, se puede consultar la referencia [23].

- ***Autoencoder del proyecto***

Con el autoencoder diseñado para este proyecto, se ha buscado conseguir una reducción no lineal de la dimensionalidad con la mejor convergencia posible. A continuación se presenta el diagrama del autoencoder diseñado:

- **Capa de entrada (*Input layer*):** Esta capa constará de 66 neuronas que tratarán de tomar las variables que representan cada una de las bandas espectrales de un espectro.
- **Capas ocultas (*hidden layers*):** Tenemos dos capas ocultas de 30 neuronas. Después de esta capa oculta de 30 neuronas, tenemos el espacio latente donde, el número no será siempre el mismo, sino que realizaremos varias pruebas con diferentes reducciones.
- **Capa de salida (*output layer*):** La capa de salida, al igual que la de entrada, será de 66 neuronas que tratarán de reproducir el espectro a partir de los valores de las capas ocultas.

Para la función de pérdida, se ha utilizado el *mean squared error (mse)*, que se obtiene mediante la media de los errores al cuadrado, tal y como se aprecia en la siguiente ecuación 3.6.1.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Ecuación 3.6.1

La función de activación que ha proporcionado una mejor convergencia para el autoencoder es la ReLU (*Rectified Linear Unit*), cuya fórmula se encuentra en la ecuación 3.6.2.

$$\text{relu}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Ecuación 3.6.2

Para el optimizador, se ha utilizado el algoritmo Adam, popular en el entrenamiento de modelos de aprendizaje automático y redes neuronales. Adam es una combinación del método de descenso de gradiente estocástico (SGD) con momentos de primer y segundo orden.

Para obtener más información sobre las diferentes partes de las redes neuronales, puede consultarse la referencia [24].

3.5.3. Random Forest

- Problema PCA y Autoencoder

Como se ha dicho, se busca reducir la dimensionalidad de los datos, pero las dos técnicas anteriores necesitan el espectro obtenido con todas sus bandas espectrales (400-1050 nm), es decir, todas las variables existentes en los datos.

Entonces, también se busca reducir la dimensionalidad manteniendo las variables originales cogiendo distintas longitudes de onda del espectro. Esta técnica puede ser importante puesto que si, por ejemplo, seleccionamos 5 longitudes de onda para realizar la investigación de encontrar una buena estimación, facilitamos la adquisición de los datos.

Pero la pregunta es, cómo saber qué variables (longitudes de onda) deben escogerse para que los modelos funcionen mejor. La respuesta simplificada estaría basada en la importancia de la variable en la predicción de un modelo.

- Selección de características – Random Forest (RF)

La selección de las características más relevantes en este caso se ha realizado utilizando un modelo de Random Forest (RF). En un modelo Random Forest, se construye una colección de árboles de decisión y se calcula la importancia de cada característica mediante la medida de la disminución de la precisión del modelo cuando esta característica se permuta aleatoriamente.

3. Materiales y Métodos

La importancia de las características puede cuantificarse utilizando métricas como la impureza de Gini. Las características con mayores puntuaciones de importancia se consideran más relevantes para el rendimiento del modelo y es más probable que se seleccionen durante el proceso de selección de características.

En este caso concreto, se ha entrenado un modelo Random Forest en todo el conjunto de datos y se han extraído las puntuaciones de importancia de sus características. A continuación, se han seleccionado las 3, 6 y 12 principales características en función de sus puntuaciones de importancia. Estas características seleccionadas se han utilizado posteriormente para entrenar a los diferentes modelos de aprendizaje automático que se muestran más adelante.

3.6. Modelos

En los últimos años, los modelos de aprendizaje automático han adquirido una importancia creciente. Esto se debe a que son capaces de aprender y mejorar automáticamente a partir de la experiencia. Esto significa que pueden utilizarse para resolver una serie de tareas que antes eran difíciles o imposibles para los programas informáticos tradicionales.

Por lo general, los modelos de aprendizaje automático han cobrado cada vez más importancia en los últimos años debido a su capacidad para aprender automáticamente y mejorar a partir de la experiencia. Esto ha dado lugar a una serie de importantes aplicaciones en varios campos [25].

3.6.1. Modelos Lineales

Los modelos lineales se utilizan ampliamente en diversas aplicaciones y son fáciles de utilizar e interpretar. Los modelos lineales se basan en una relación lineal entre las variables de entrada y la variable de salida. El modelo se entrena utilizando un conjunto de pares de entradas-salida y los parámetros del modelo se aprenden a partir de los datos. A continuación, se puede utilizar el modelo para predecir la salida para nuevos valores de entrada.

También son fáciles de interpretar, lo que les hace valiosos para la toma de decisiones. Sin embargo, los modelos lineales están limitados en su capacidad para capturar las relaciones no lineales entre las variables. Estos modelos presentados a continuación, se encuentran más detallados en [26].

- ***Regressión Lineal***

La regresión lineal es el modelo de aprendizaje automático más básico. Se basa en el supuesto de que existe una relación lineal entre la variable dependiente y la o las variables independientes. Para calcular la regresión lineal, el primer paso es encontrar la ecuación de la recta que mejor representa los datos. Esto se realiza encontrando los valores de la pendiente y el intercepto que minimizan la suma de los errores al cuadrado.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} + \epsilon_i$$

Ecuación 4.1.1

A la ecuación 4.1.1 β_0 corresponde con la ordenada en el origen (intercepto), el valor medio de la variable respuesta y cuando todos los predictores son cero. β_j corresponde con el efecto medio que tiene sobre la variable respuesta el incremento en una unidad de la variable predictora x_j ,

mantiéndose constantes el resto de variables. Son los coeficientes de la regresión. ϵ_i corresponde con el residuo o error, la diferencia entre el valor observado y el estimado por el modelo.

Una vez hallada la ecuación de la recta, puede utilizarse la ecuación 4.1.2 para predecir los valores futuros de la variable dependiente, basándose en los valores conocidos de la variable independiente.

$$\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_n x_{in}$$

Ecuación 4.1.2

- **Ridge Regresion**

La regresión Ridge es un tipo de regresión lineal que se utiliza para modelar datos susceptibles de multicolinealidad

La regresión Ridge es similar a la regresión por mínimos cuadrados, pero utiliza una función de coste diferente. En lugar de minimizar la suma de residuos al cuadrado, la regresión Ridge minimiza la suma de los residuos al cuadrado más un término de penalización. El término de penalización es una función de la magnitud de los coeficientes. El término de penalización, λ , se utiliza para evitar que los coeficientes sean demasiado grandes, lo que puede provocar un exceso de ajuste. Esta penalización es conocida como L_2 . Partiendo de la ecuación anterior 4.1.1 encontramos esta ecuación 4.2.1:

$$\text{suma residus al cuadrat} = \sum_i^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j \cdot x_{ij})^2$$

Ecuación 4.2.1

$$RSS = \text{suma residus al cuadrat} + \lambda \sum_{j=1}^p \beta_j^2$$

Ecuación 4.2.2

En la ecuación 4.2.2 se ve que a medida que λ aumenta, mayor es la penalización y menor el valor de los predictores. Cuando $\lambda=0$, la penalización es nula y el resultado es el mismo que una regresión lineal.

Utilizamos la regresión Ridge, ya que notamos que existe una gran correlación entre las diferentes bandas espectrales (variables predictoras), lo que puede ocasionar coeficientes muy grandes. Lo que puede llevar a un sobreajuste, porque el modelo se ajustará al ruido de los datos en lugar de la relación.

- ***Bayesian Ridge Regression***

La regresión bayesiana es un tipo de regresión que utiliza la inferencia bayesiana para estimar los parámetros del modelo. La inferencia bayesiana es un método de inferencia estadística basado en el teorema de Bayes. El teorema de Bayes se utiliza para calcular la probabilidad de que se produzca un evento determinado, basándose en datos anteriores. El teorema de Bayes se puede escribir como se muestra en la ecuación 4.3.1:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Ecuación 4.3.1

dónde:

- $P(A|B)$ es la probabilidad posterior de que ocurra el suceso A, dado que ha ocurrido el suceso B.
- $P(B|A)$ es la probabilidad de que ocurra el suceso B, dado que ha ocurrido el suceso A.
- $P(A)$ es la probabilidad a priori de que ocurra el suceso A.
- $P(B)$ es la probabilidad a priori de que se produzca el suceso B.

La Regresión bayesiana Ridge es un método de inferencia estadística en el que utilizamos el teorema de Bayes para actualizar las probabilidades de los eventos a medida que observamos datos nuevos.

En el enfoque bayesiano, los parámetros no sólo se estiman en base a los datos, sino también en el conocimiento previo sobre los parámetros. Se supone que la distribución a priori es una distribución normal con una media de cero y una varianza de la matriz de identidad. Siguiendo la regresión lineal, podemos extraer esta ecuación 4.3.2:

$$y_i = x_i^T \beta + \epsilon_i \text{ on } \epsilon_i \sim N(0, \sigma^2)$$

Ecuación 4.3.2

Podemos formular la función de verosimilitud describe la relación entre los parámetros y los datos. Esta función es conocida y se expresa como euación 4.3.3.

$$p(y|X, \beta, \sigma^2) \propto (\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta)\right)$$

Equació 4.3.3

Si aplicamos el teorema de Bayes a esta distribución de los parámetros β dada la observación tal y como vemos en la ecuación 4.3.4:

$$P(\beta|y, X) = \frac{P(y, X|\beta)P(\beta)}{P(y, X)}$$

Ecuación 4.3.4

donde nos referimos a:

- $P(\beta|y, X)$ como la distribución posterior.
 - $P(y, X|\beta)$ como la función de verosimilitud.
 - $P(\beta)$ la distribución a priori.
 - $P(y, X)$ probabilidad de las propias observaciones (evidencia).
- $$P(y, X) = \int_{-\infty}^{\infty} P(y, X|\beta) P(\beta) d\beta$$

Para las distribuciones a priori de los coeficientes se utilizan distribuciones no informativas. Esto significa que es una distribución de probabilidad que no se basa en ninguna información o conocimiento previo. Podemos imponer, por ejemplo, una distribución uniforme.

La función de verosimilitud es una distribución condicional de las características de la respuesta y del predictor dado el modelo. A medida que aumenta el número de datos de la muestra, el parecido se hará más preciso, su incertidumbre se reducirá y superará la distribución a priori en algún momento.

La distribución posterior, al contrario de la de verosimilitud, es una distribución condicional de los parámetros del modelo dadas las características de la respuesta y del predictor, y viene basada en la similitud de los datos y la distribución de los datos. Dado un número de iteraciones, iremos aproximándonos hacia la distribución a posteriori que se adapte a los datos.

Hay que añadir que todo este proceso, incluimos la norma l2, es decir, la regularización comentada anteriormente. Entonces, tenemos otro parámetro λ que penalizará a los coeficientes altos tal y como hemos visto.

• ***SGD Regressor***

El descenso de gradiente estocástico (SGD) es un algoritmo de optimización utilizado para encontrar los valores de los parámetros que minimizan una función de coste. Es una versión estocástica del descenso de gradiente, lo que significa que utiliza la aleatoriedad para seleccionar los ejemplos de entrenamiento en cada paso.

El SGD comienza con un conjunto de valores de parámetros (normalmente inicializados a 0) y después ajusta estos valores de forma iterativa para minimizar la función de coste. En cada iteración, SGD selecciona aleatoriamente un ejemplo de entrenamiento y calcula el gradiente de la función de coste respecto a los parámetros. El gradiente es un vector que

indica la dirección donde deben cambiar los parámetros para minimizar la función de coste. A continuación, el SGD ajusta los parámetros a la dirección del gradiente y repite el proceso.

La ecuación básica de la SGD está definida por la ecuación 4.4.1.

$$\widehat{\theta_{t+1}} = \widehat{\theta_t} - \alpha \nabla_{\theta} L(\widehat{\theta_t})$$

Ecuación 4.4.1

dónde $\widehat{\theta_t}$ es el vector de parámetros actual, α es la tasa de aprendizaje, y $\nabla_{\theta} L(\widehat{\theta_t})$ es el gradiente de la función de pérdida $L(\widehat{\theta_t})$ en cuanto a θ .

El gradiente se utiliza entonces para actualizar los parámetros. Este proceso se repite hasta que la función de coste converge un mínimo. La función de coste mide la diferencia entre el valor previsto y el valor real. La tasa de aprendizaje (η) es un hiperparámetro que controla el tamaño de los pasos dados por el SGD.

Por último, el SGDRegresor una técnica de optimización iterativa utilizada para encontrar el conjunto óptimo de parámetros que minimice una función de coste determinada. SGDR es un algoritmo eficiente con una complejidad computacional baja.

3.6.2. Cross-Descomposition

- ***PLS Regressor***

El modelo de regresión de mínimos cuadrados parciales (PLS) es un tipo de análisis de regresión que se utiliza para predecir los valores de una variable de respuesta en base a los valores de una o más variables predictoras. Es un modelo de aprendizaje automático muy utilizado en el sector de la agroingeniería.

El modelo de regresión PLS se basa en la idea de proyectar las variables predictoras (X) en un espacio de dimensiones inferiores, llamado espacio latente, a fin de encontrar la mejor combinación lineal de las variables predictoras que pueda explicar la varianza de la variable de respuesta (Y).

El espacio latente está definido por un conjunto de vectores latentes ortogonales, representados por el símbolo T , que se calculan a partir de las variables predictoras (X) mediante descomposición de valor singular (SVD), que ya hemos comentado anteriormente.

Los valores predichos de la variable de respuesta (Y) se calculan multiplicando los vectores latentes (T) por los correspondientes pesos o

coeficientes (B), que se estiman minimizando la suma residual de cuadrados (RSS) mediante un algoritmo de optimización como el descenso gradiente.

En términos matemáticos, el modelo de regresión PLS puede escribirse como sigue la ecuación 4.5.1:

$$I = X * B + E$$

Ecuación 4.5.1

donde I son los valores predichos de la variable de respuesta, X es la matriz de variables predictoras, B es la matriz de pesos o coeficientes y E es el término de error.

Una de las principales ventajas del PLS es que puede manejar variables predictoras muy colineales sin perder interpretabilidad. Además, PLS es menos sensible a las observaciones periféricas que otros métodos de regresión, lo que puede mejorar la precisión del modelo. Además, PLS es capaz de manejar datos con valores perdidos, lo que puede resultar útil en muchas aplicaciones del mundo real. Por lo general, PLS es una herramienta flexible y potente para modelizar relaciones complejas entre variables de respuesta y predictoras.

Podemos encontrar detallado en mayor medida los conceptos y métodos de este modelo, así como sus aplicaciones, en la referencia [27].

3.6.3. Nearest Neighbours

- ***K - Neighbors Regressor***

El regresor k-vecinos más cercanos es un algoritmo de aprendizaje automático que se utiliza para tareas de regresión. Funciona haciendo predicciones basadas en los valores de los "k" puntos más cercanos de los datos. Esto significa que examina los k puntos de los datos más cercanos al punto para el que se realiza la predicción y utiliza sus valores para realizar la predicción.

K-vecinos más cercanos es un algoritmo sencillo e intuitivo, fácil de aplicar e interpretar. A menudo se utiliza en casos en los que la relación entre los predictores y la variable de respuesta es compleja y no se capta fácilmente con un modelo lineal.

Para entrenar el modelo, primero es necesario proporcionar al algoritmo un conjunto de datos de entrenamiento. A continuación, el algoritmo utiliza estos datos para aprender la relación entre los predictores y la respuesta. Encontramos más información en [28].

3.6.4. Support Vector Machine (SVM)

Todos los modelos que se explican a continuación en este apartado, se encuentran más detallados en [29].

- **Support Vector Regressor (SVR)**

SVR es un tipo de máquina de vectores de soporte (SVM) que se utiliza para el análisis de regresión. Al igual que otras máquinas de vectores soporte, SVR es un algoritmo de aprendizaje supervisado que utiliza datos de entrenamiento para aprender la relación entre una variable de respuesta y variables predictoras. Una vez entrenado el modelo, se puede utilizar para realizar predicciones sobre la variable de respuesta para nuevos datos.

SVR es un método potente y flexible aplicado a una amplia gama de problemas. Una de las principales ventajas de SVR es que puede manejar datos con relaciones no lineales entre las variables de respuesta y predictoras, lo que suele ocurrir en los datos del mundo real. Además, se ha demostrado que SVR funciona bien con datos de alta dimensionalidad, lo que puede suponer un reto para otros métodos de regresión. Por lo general, SVR es una herramienta valiosa para modelar y comprender relaciones complejas en los datos.

Las ecuaciones del modelo Support Vector Regressor (SVR) pueden derivarse de las ecuaciones generales de las máquinas de vectores de soporte (SVM). En un problema SVR típico, el objetivo es encontrar una función que pueda predecir una variable de respuesta, y , basándose en un conjunto de variables predictoras, X . El valor predicho de y para un conjunto dado de valores predictores, X viene dado por la siguiente ecuación 4.6.1:

$$f(X) = w^T * X + b$$

Ecuación 4.6.1

donde w es un vector de ponderaciones, X es un vector de valores predictores y b es un término de inglete.

Para encontrar los valores óptimos de w y b , el modelo SVR debe resolver el siguiente problema de optimización:

$$\frac{1}{2} \|w\|^2$$

Ecuación 4.6.2

Sujeto a:

$$y - (w^T * X + b) \leq \epsilon$$

$$y - (w^T * X + b) \geq -\epsilon$$

Ecuación 4.6.3

Donde, en la ecuación 4.6.2 $|w|$ es la norma euclídea del vector de pesos, y en la ecuación 4.6.3 ϵ es un parámetro definido por el usuario que controla el tamaño del margen de error, y y es el verdadero valor de la variable de respuesta.

Una de las principales ventajas de SVR, como hemos dicho, es que puede manejar datos con relaciones no lineales entre las variables de respuesta y predictoras, lo que suele ocurrir en los datos del mundo real. Además, se ha demostrado que SVR funciona bien con datos de alta dimensionalidad, lo que puede suponer un reto para otros métodos de regresión. Además, SVR puede utilizarse con una variedad de funciones de kernel, lo que permite al usuario especificar diferentes tipos de relaciones entre las variables de respuesta y predictoras. Por lo general, SVR es una herramienta flexible y potente para el análisis de regresión.

- **Linear SVR (LSVR)**

Linear SVR es una variante del modelo SVR que se utiliza para el análisis de regresión. A diferencia de otras máquinas vectoriales de soporte, el modelo LSVR utiliza una función del kernel lineal, lo que significa que sólo puede moldear relaciones lineales entre la respuesta y las variables predictoras. Esto simplifica el problema de optimización y hace que el modelo LSVR sea más eficiente computacionalmente que otras máquinas vectoriales de soporte. Una vez encontrados los valores óptimos de w y b , el modelo Linear SVR se puede utilizar para realizar predicciones sobre la variable de respuesta para datos nuevos.

- **NuSVR**

NuSVR es un método flexible de análisis de regresión que suele utilizarse cuando la relación entre las variables de respuesta y predictoras es compleja o desconocida. Se llama así, puesto que tiene un parámetro desnudo. Con este modelo se pueden utilizar una variedad de funciones de kernel.

Una ventaja de NuSVR es que se puede gestionar tareas de regresión lineal y no lineal y es altamente escalable, lo que significa que puede manejar grandes conjuntos de datos de forma eficiente.

Como se ha dicho, otra ventaja de NuSVR es que tiene un parámetro incorporado, desnudo, que permite al usuario especificar un límite inferior en el número de vectores de soporte que el modelo debe utilizar. Esto puede

ser útil para evitar el sobreajuste, ya que garantiza que el modelo no intente ajustarse al ruido o variaciones aleatorias de los datos. Por tanto, NuSVR proporciona un rendimiento robusto incluso en presencia de puntos de datos ruidosos o atípicos, gracias a su uso del parámetro desnudo.

En resumen, NuSVR es un algoritmo potente y versátil que puede utilizarse para diversas tareas de regresión. Su capacidad para manejar relaciones lineales y no lineales, así como su robustez en presencia de datos ruidosos, le convierten en una herramienta valiosa.

3.6.5. Decision Trees and Ensemble Models

La explicación más detallada de estos métodos se encuentra en [31].

- ***Decision Tree Regressor***

Un árbol de decisión es un tipo de modelo de aprendizaje automático que se utiliza para tareas de regresión y clasificación. Funciona creando una estructura en forma de árbol donde cada nodo interno representa una decisión o división basada en el valor de una determinada característica, y cada nodo hoja representa un valor predicho.

El árbol se construye dividiendo los datos de entrenamiento en subconjuntos cada vez menores en función de los valores de determinadas características. Este proceso se repite recursivamente hasta que cada subconjunto contiene un único valor objetivo.

Para dividir los datos en cada nodo, el modelo utiliza un criterio de división como el error cuadrático medio o el error absoluto medio. Por ejemplo, si intentamos predecir el precio de una casa en función del tamaño, el modelo podría dividir los datos en un nodo determinado en función de si el tamaño está por encima o por debajo del tamaño medio de las casas del subconjunto.

El valor previsto para un nuevo punto de datos se determina recorriendo el árbol y llegando a un nodo hoja. El valor en el nodo hojas se toma como el valor predicho para el nuevo punto de datos.

Una de las principales ventajas de utilizar un árbol de decisión es su interpretabilidad. Dado que el modelo crea una estructura en forma de árbol, las decisiones y las divisiones tomadas por el modelo pueden ser fácilmente comprendidas e interpretadas por los humanos.

- ***Random Forest Regressor:***

El *Random Forest Regressor* és un tipus de model d'aprenentatge automàtic que s'utilitza per a tasques de regressió. És un model de conjunt, el que significa que està format per múltiples arbres de decisió individuals que treballen junts per fer prediccions.

Els arbres de decisió individuals del *Random Forest* s'entrenen a diferents subconjunts de dades i amb diferents subconjunts de característiques. Això crea un conjunt divers de models que poden capturar una àmplia gamma de patrons a les dades.

En fer una predicció per a un nou punt de dades, el *Random Forest* combina les prediccions de tots els arbres de decisió individuals utilitzant un vot majoritari o algun altre mètode de combinació. Això pot conduir a prediccions més precises que les realitzades per qualsevol dels arbres individuals.

Cada arbre de decisió s'entrena a un subconjunt diferent de dades i amb un subconjunt diferent de característiques, i les prediccions de tots els arbres es combinen per fer una predicció final per a un nou punt de dades.

Els arbres de decisió individuals del *Random Forest* es construeixen utilitzant les mateixes parts que els arbres de decisió. Això inclou, com hem comentat anteriorment; un criteri de divisió, com l'error quadràtic mitjà o l'error absolut mitjà, per determinar com dividir les dades a cada node, i un valor de node full, que es pren com el valor predict. Aleshores, la novetat d'aquest mètode respecte a l'anterior és que s'utilitza una combinació de molts arbres de decisió que, mitjançant el vot majoritari, aconsegueixen predir un valor.

Un dels principals avantatges d'utilitzar-lo és la seva capacitat per aconseguir una gran precisió en conjunts de dades complexes. Atès que el model és un conjunt de múltiples arbres de decisió individuals, és capaç de capturar una àmplia gamma de patrons a les dades i fer prediccions més precises que qualsevol arbre individual.

Altre avantatge del és la seva robustesa davant del sobreajustament. Com cada arbre s'entrena en un subconjunt diferent de dades i amb un subconjunt diferent de característiques, és menys probable que el model en conjunt s'ajuste excessivament a les dades d'entrenament i més probable que es generalitze bé a noves dades.

- ***Extra Tree Regressor:***

Extra Tree Regressor es un tipo de modelo de aprendizaje automático que se utiliza para tareas de regresión. Es similar a un Random Forest, pero los árboles de decisión individuales se entrenan utilizando un enfoque más aleatorio y menos codicioso.

En un Random Forest, cada árbol se entrena en un subconjunto distinto de datos y con un subconjunto distinto de características. Esto ayuda a reducir el sobreajuste y mejorar la precisión global del modelo. Por el contrario, los árboles individuales de un Extra Tree Regressor se entrenan utilizando umbrales aleatorios para cada característica, en lugar de los umbrales óptimos encontrados mediante un criterio de división. Esto hace que el modelo sea menos sensible a los datos específicos utilizados para el entrenamiento y puede dar lugar a predicciones más precisas sobre datos nuevos.

Entonces el Extra Tree Regressor incluye umbrales aleatorios. Aparte, consta de las partes anteriormente ya descritas: un valor del nodo hoja, una estructura del árbol y de un método de combinación.

- ***Ada Boost Regressor:***

AdaBoost, que es la abreviatura de Adaptive Boosting, es otro algoritmo de aprendizaje supervisado que pertenece a la familia de métodos de ensamblaje. El algoritmo AdaBoost se basa en la idea de mejorar iterativamente un conjunto de regresores débiles ponderando y combinando sus resultados.

El proceso general del algoritmo AdaBoost para un problema de regresión se detalla a continuación:

- **Inicialización:** se asigna un peso uniforme a cada observación del conjunto de entrenamiento. En el caso de regresión, estos pesos se utilizan para calcular el error ponderado de cada modelo candidato.
- **Iteración:** el algoritmo itera a través de un número definido de etapas, cada una de las cuales realiza lo siguiente:
 - **Ajustar un regresor débil:** un regresor débil, como un árbol de decisiones poco profundo, se entrena utilizando las variables independientes y la variable objetivo.
 - **Calcula el error ponderado:** el rendimiento del regresor ajustado se evalúa mediante el error ponderado, que tiene en cuenta los pesos de las observaciones.

- **Calcula el peso del regresor:** el peso del regresor recientemente ajustado se determina en función de su error ponderado. Los regresores con menos error tienen mayor peso en el montaje final.
- **Actualiza los pesos de las observaciones:** se incrementa el peso de las observaciones mal previstas y se reduce el peso de las observaciones predichas correctamente, con el fin de priorizar las observaciones difíciles de predecir en las siguientes iteraciones.
- **Normalizar los pesos:** los pesos de las observaciones se normalizan de forma que sumen 1.
- **Combinación:** una vez repetida a través de todas las etapas, los regresores ponderados se combinan para formar el modelo final de AdaBoost.
- ***Gradient Boosting Regressor:***

El Gradient Boosting Regresor funciona construyendo secuencialmente un conjunto de árboles de decisión débiles, en los que cada árbol se crea para corregir los errores residuales de los árboles anteriores. En otras palabras, el algoritmo intenta minimizar la función de pérdida en cada paso, ajustando el modelo en la dirección del gradiente negativo.

Para ilustrar el proceso, considere las siguientes etapas del algoritmo:

- **Inicialización:** Se establece un modelo base, que es una constante que minimiza la función de pérdida. En el caso de la regresión, éste suele ser el valor medio de la variable objetivo en el conjunto de entrenamiento.
- **Iteración:** el algoritmo itera a través de un número definido de etapas, cada una de las cuales realiza lo siguiente:
 - **Calcula los residuos:** los errores residuales se obtienen comparando las predicciones del modelo actual con los valores reales de la variable objetivo.
 - **Ajusta un árbol de decisión:** se entrena un árbol de decisión débil utilizando las variables independientes y los residuos como variable objetivo.
 - **Calcula el peso óptimo:** el peso óptimo del árbol nuevo instalado se determina mediante técnicas de optimización basadas en el descenso de gradientes, para minimizar la función de pérdida.

- **Actualizar el modelo:** el árbol ponderado se añade al modelo actual.
- **Combinación:** una vez repetida a través de todas las etapas, los árboles ponderados se combinan para formar el modelo final de Gradient Boosting Regressor.

Una vez finalizado el proceso de entrenamiento, al igual que Ada Boost, el Gradient Boosting utiliza las predicciones combinadas de todos los modelos individuales para realizar una predicción final. Esta predicción final es una media ponderada de las predicciones de cada modelo individual.

En la siguiente Tabla 3.4 se pueden ver cuáles son los aspectos que diferencian a estos dos métodos que, a simple vista, pueden resultar muy similares.

Aspecto	<i>Gradient Boosting Regressor</i>	<i>AdaBoost</i>
Enfoque de optimización	Usa el descenso de gradiente para minimizar la función de pérdida en cada iteración.	Se basa en la adaptación de los pesos de las observaciones y modelos débiles.
Actualización del modelo	En cada iteración, añade un árbol ponderado que corrige los errores residuales.	Combina modelos débiles ponderados en función de su rendimiento en la predicción de observaciones difíciles.
Función de pérdida	Permite diferentes funciones derivables.	Use una función de pérdida exponencial específica.
Control de complejidad	Ofrece mayor número de hiperparámetros ajustables.	Cuenta con menos hiperparámetros ajustables.
Sensibilidad al ruido	Es menos sensible al ruido ya las observaciones atípicas-	Es más sensible al ruido y observaciones atípicas.

Taula 3.4. Diferencias entre Gradient Boosting Regressor y AdaBoost.

3.6.6. Bagging (Ada Boost i Gradient Boosting)

El bagging, que proviene de Bootstrap Aggregating, es un método que pretende mejorar la estabilidad y la precisión de los modelos de predicción. Este enfoque se basa en la combinación de múltiples modelos entrenados con distintos subconjuntos de datos generados por muestreo con sustitución (bootstrap) del conjunto de datos original [31].

Utilizar con AdaBoost y Gradient Boosting, aunque ya tienen mecanismos internos para mejorar la precisión y reducir el error de predicción, bagging puede ofrecerle ventajas adicionales.

Cuando se utiliza este método, cada modelo débil se entrena con un subconjunto de datos generados por el muestreo con sustitución en lugar de utilizar el conjunto de datos completo. Esto puede aumentar la diversidad entre los modelos individuales, dando lugar a un modelo más robusto y estable. Pero, por otra parte, también puede aumentar la complejidad y tiempo de entrenamiento de los modelos.

3.7. Evaluación de los modelos

Para medir el funcionamiento de los modelos, se dispuso de una serie de métricas. Como el problema es de regresión, se han utilizado los siguientes errores:

Métricas numéricas de error

- **Coeficiente de determinación (R^2):** Se calcula cómo aparece en la siguiente Ecuación 5.1. Se calcula el cuadrado del coeficiente de correlación de Pearson. R^2 trata de explicar la variación de la variable respuesta (en nuestro caso el nutriente que se quiere predecir), en relación a una o más variables predictoras. El R^2 va desde 0 hasta 1, siendo 1 la mejor predicción posible.

$$R^2 = \frac{\sum_{i=1}^n (\mathbf{y}_i - \mathbf{z}_i)^2 - \sum_{i=1}^n (\mathbf{x}_i - \mathbf{y}_i)^2}{\sum_{i=1}^n (\mathbf{y}_i - \mathbf{z}_i)^2}$$

Ecuación 5.1. Ecuación del Coeficiente de determinación (R^2)

- **Mean Absolute Error (MAE):** Este error mide la media del error absoluto entre el valor real y las predicciones tal y como se ve en la Ecuación 5.2. Otorga el mismo error a pequeños errores que a errores grandes.

- $MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$

Ecuación 5.2. Ecuación del error medio absoluto

- **Root Mean Squared Error (RMSE):** Es la raíz cuadrada de la media de las diferencias en el cuadrado entre los valores reales y las predicciones. Dado que el resto de los errores se eleva al cuadrado, esta métrica penaliza los errores grandes. Puede verse en la siguiente ecuación (Ecuación 5.3.).

$$\bullet \quad RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Equació 5.3. Ecuación de la raíz cuadrada de la media del error

Residuos de los modelos

También se visualizan gráficamente los residuos en caso de que se quiera más información sobre el modelo. Los residuos indican cuál es la parte de la variable que se busca predecir no está explicada por el modelo. Si son nulos, entonces no existe ningún error en la estimación, ya que los valores observados coinciden con los valores estimados. Si es positivo entonces el valor observado es mayor que el valor estimado lo que implica que se infraestima la variable y, por el contrario, el residuo es negativo entonces el valor observado es menor que su valor estimado. Por tanto, se sobreestima la variable y.

Este tipo de gráficos te permite evaluar 3 cuestiones principalmente:

- Si se utiliza el tipo de relación adecuada. Si el tipo de modelo que estamos utilizando no es adecuado se encuentran sesgos o tendencias en los residus.
- Si la varianza es constante o, por el contrario, tenemos problemas de racimo irregular. Uno de los supuestos del modelo de regresión lineal es que la varianza de los residuos es constante, es decir, que los residuos se distribuyen al azar en torno a cero.
- Si existen valores anormales que puedan perturbar e invalidar el modelo.

Entonces, se busca que los residuos se aproximen lo máximo posible a 0, para saberlo, lo mejor que se puede hacer, es representarlo gráficamente. Existen diferentes gráficos para evaluar el funcionamiento del modelo, en este proyecto se han utilizado cuatro. Pueden verse ejemplos de gráficos, y se explica que es lo que se busca encontrar en esos gráficos para saber si el modelo funciona correctamente.

En la siguiente Tabla 3.5 se pueden ver los diferentes gráficos utilizados en el proyecto y cómo deben interpretarse para poder saber cómo funciona el modelo.

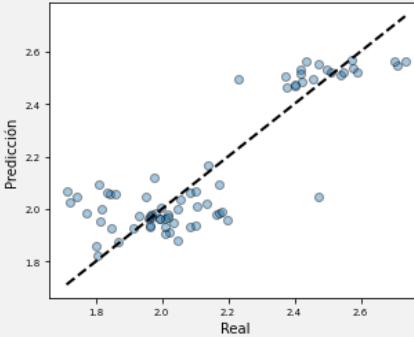
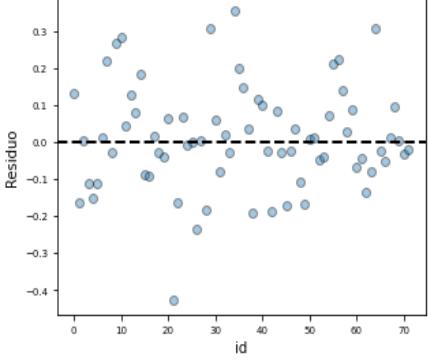
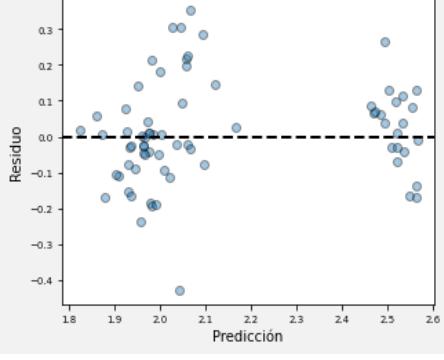
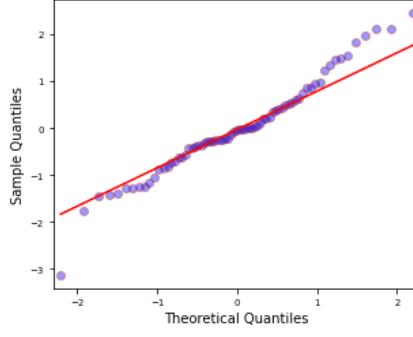
GRAFICO	¿QUÉ SE BUSCA?
	<p>GRÁFICO VALOR REAL vs. PREDIDO En este gráfico, lo que se busca es que todos los puntos estén en la línea diagonal, lo que significa que el valor real es el mismo que la predicción.</p>
	<p>RESIDUOS DEL MODELO En el segundo gráfico se ve que los residuos del modelo. Los residuos deben estar distribuidos aleatoriamente para que este método de estimación funcione de forma adecuada. Los residuos del modelo deben cumplir la propiedad de homocedasticidad (variancia del error constante a lo largo de las observaciones). En este caso, se ve que esto se cumple.</p>
	<p>RESIDUOS DEL MODELO vs. PREDICCIONES: En el tercer gráfico también se busca que los residuos se distribuyan de forma aleatoria a lo largo del eje x. En nuestro gráfico se ve que se separan en dos grupos, entonces no está cumpliendo con la propiedad anterior.</p>
	<p>Q-Q RESIDUOS DEL MODELO En este cuarto gráfico, si los residuos se ajustan a la línea de 45 grados, entonces los residuos se distribuyen de forma aproximadamente normal. Puede verse en nuestro gráfico QQ anterior que los residuos tienden a desviarse bastante de la línea de 45 grados, especialmente en los extremos, lo que podría ser una indicación de que no están distribuidos normalmente.</p>

Tabla 3.5. Cómo analizar los residuos de un modelo.

3.8. 3.8. Construcción de modelos y pruebas realizadas

Construcción de modelos

La búsqueda de los mejores parámetros de un modelo de aprendizaje automático es un paso esencial en el proceso de construcción y entrenamiento de un modelo. Los parámetros son los valores que determinan el comportamiento del modelo, y encontrar los valores óptimos es esencial para conseguir un buen rendimiento en su tarea objetivo.

Este proceso implica probar distintos valores para los parámetros del modelo, evaluar el rendimiento del modelo en un conjunto de validación y, a continuación, ajustar los valores de los parámetros en función de los resultados. Este proceso puede repetirse hasta encontrar los valores óptimos de los parámetros.

Para encontrar estos valores óptimos, se han realizado pruebas con el llamado Grid Search [32]. El método Grid Search es una técnica muy utilizada para buscar los mejores parámetros de un modelo de aprendizaje automático. Este método consiste en definir una cuadrícula de valores de parámetros y, a continuación, entrenar y evaluar un modelo para cada combinación de valores de parámetros de la cuadrícula. A continuación, se selecciona la mejor combinación de valores de parámetros según el rendimiento del modelo en un conjunto de calibración.

Los rango de parámetros utilizado para cada modelo han sido los siguientes:

- **LinearRegression:** Sense paràmetres
- **RidgeRegression:** *tol* [1e-06, 1e-03, 1e-01, 1], *solver* ['auto', 'svd', 'cholesky', 'lsqr', 'sparse_cg', 'sag', 'saga', 'lbfgs'].
- **BayesianRidge:** *alpha_1* [1e-3, 1e-6], *alpha_2* [1e-3, 1e-6], *lambda_1* [1e-3, 1e-6], *lambda_2* [1e-6, 1e-7], *fit_intercept* [True, False].
- **SGDRegressor:** *loss* ['squared_loss', 'huber', 'epsilon_insensitive', 'squared_epsilon_insensitive'], *alpha* [0.01, 0.1, 1, 10], *learning_rate* ['constant', 'optimal', 'invscaling', 'adaptive'], *eta0* [0.01, 0.1, 1], *fit_intercept* [True, False].
- **PLSRegression:** *n_components* [1-25], *scale* [False, True], *max_iter* [100-1000], *tol* [0.1-0.00001].
- **KNeighbors:** *n_neighbors* [1-25], *weights* ['uniform', 'distance'], *algorithm* ['auto', 'ball_tree', 'kd_tree', 'brute'], *leaf_size* [10, 20, 30, 40, 50], *p* [1, 2].
- **SVR:** *kernel* ['linear', 'rbf', 'sigmoid'], *gamma* ['scale', 'auto'], *C* [0.1, 1, 10, 50, 75], *epsilon* [0.1, 0.01, 0.001].
- **LinearSVR:** *epsilon* [0.1, 0.01, 0.001], *C* [0.1, 1, 10, 100], *loss* ['epsilon_insensitive', 'squared_epsilon_insensitive'], *dual* [True, False],

- fit_intercept* [True, False], *intercept_scaling* [0.1, 1, 10, 100], *max_iter* [1000-10000], *tol* [1e-4, 1e-5, 1e-6].
- **NuSVR:** *nu* [0.1, 0.3, 0.5, 0.7, 0.9], *kernel* ['linear', 'rbf', 'sigmoid'], *gamma* ['scale', 'auto'], *C* [0.1, 1, 10, 100].
 - **DecisionTree:** *criterion* ['mse', 'friedman_mse', 'mae'], *max_depth* [10, 50, 100, None], *min_samples_split* [2, 5, 10], *min_samples_leaf* [1, 2, 4], *max_features* ['auto', 'sqrt', 'log2', None].
 - **RandomForest:** *n_estimators* [5, 10, 50, 75, 100, 150, 200], *max_depth* [10, 50, 100, None], *min_samples_split* [2, 3, 5, 10], *min_samples_leaf* [1, 2, 4], *max_features* ['auto', 'sqrt', 'log2', None], *bootstrap* [True, False].
 - **ExtraTrees:** *n_estimators* [5, 10, 50, 75, 100, 150, 200], *max_depth* [10, 50, 100, None], *min_samples_split* [2, 3, 5, 10], *min_samples_leaf* [1, 2, 4], *max_features* ['auto', 'sqrt', 'log2', None], *bootstrap* [True, False].
 - **AdaBoost:** *n_estimators* [5, 10, 50, 75, 100, 150, 200], *learning_rate* [0.001, 0.01, 0.1, 1], *loss* ['linear', 'square', 'exponential'].
 - **GradientBoosting:** *learning_rate* [0.001, 0.01, 0.1, 1], *n_estimators* [5, 10, 50, 75, 100, 150, 200], *max_depth* [3, 5, 10], *min_samples_split* [2, 3, 5, 10], *min_samples_leaf* [1, 2, 4], *max_features* ['auto', 'sqrt', 'log2', None].

Pruebas realizadas utilizando todas las bandas

Como este trabajo tiene un enfoque investigador, el objetivo es encontrar la mejor forma de estimar los valores nutricionales utilizando datos espectrales. La idea es probar una amplia gama de técnicas, tanto en términos de preprocesamiento de datos como de modelos de aprendizaje automático, a fin de determinar las combinaciones más eficaces.

En este sentido, se han realizado diversas pruebas con técnicas de reducción de dimensionalidad que requieren de todas las bandas espectrales, como PCA (Análisis de Componentes Principales) y el autoencoder.

A continuación, se detallan las pruebas realizadas con estas técnicas:

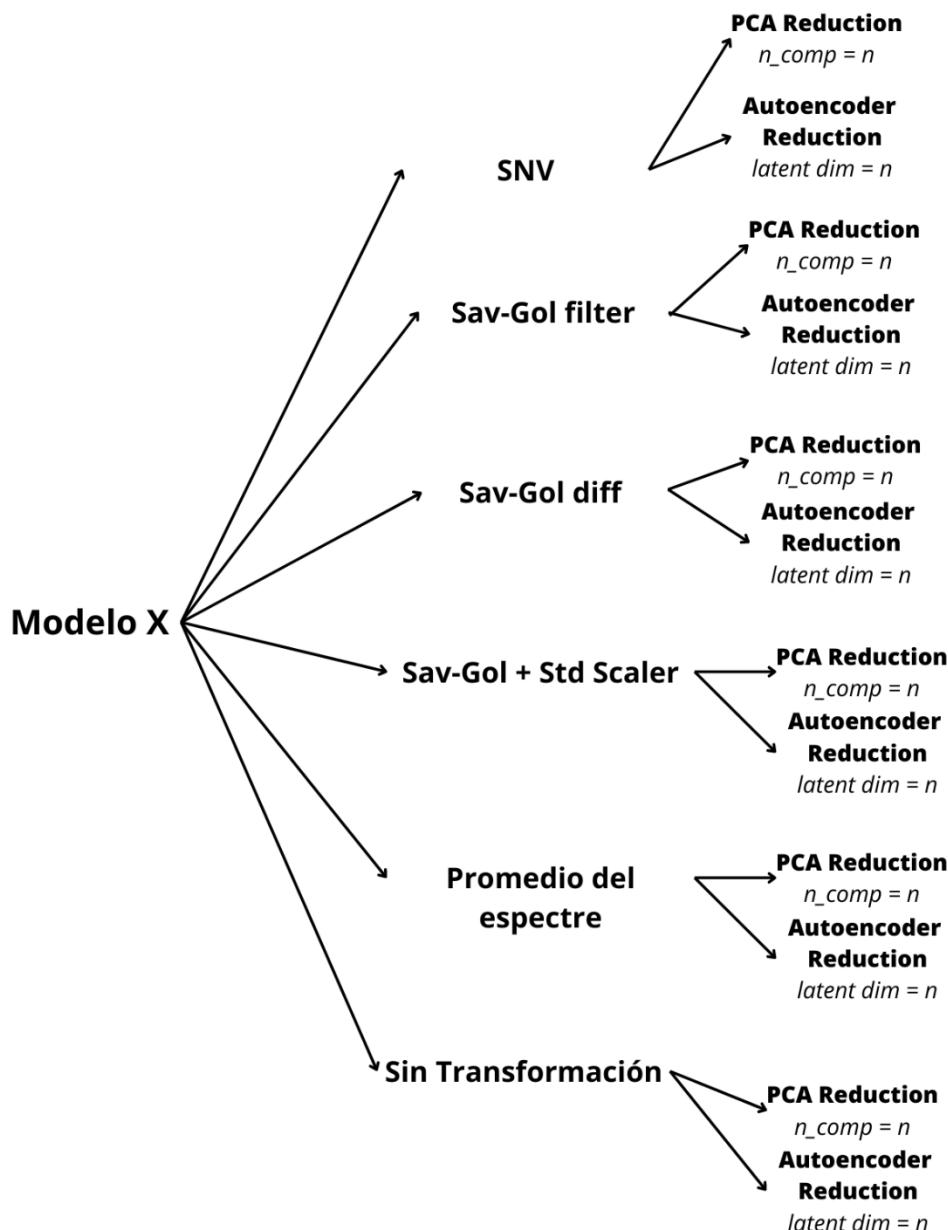


Figura 3.28. Pruebas realizadas en el proyecto para cada modelo y cada nutriente.

Como puede observarse en la Figura 3.25, para cada modelo específico (Modelo X), se han realizado pruebas utilizando las diferentes transformaciones mencionadas anteriormente. Para cada transformación, se ha reducido su dimensionalidad utilizando las dos técnicas de reducción explicadas, PCA y autoencoder, con un número específico de componentes (n).

- n per a PCA: 5, 7, 12, 15, 18, 20, 25 i 30 componentes.
- n per a autoencoder: 5, 7, 12, 15, 18, 20, 25 i 30 componentes.
- Sin reducir la dimensionalidad.

Para cada modelo que intenta estimar un nutriente específico, se han realizado pruebas con 6 distintos preprocesamientos de los datos. Para cada uno de estos preprocesamientos, se han probado 18 formas distintas de reducción de la dimensionalidad, además de los datos originales.

Pruebas realizadas seleccionando las bandas

Por otra parte, también hemos realizado una selección de bandas utilizando el modelo Random Forest. En este proceso, primero hemos seleccionado el preprocesamiento que mejor funciona para el modelo Random Forest, que es el preproceso Savitzky-Golay junto con la estandarización SNV. Este preprocesamiento ha demostrado obtener resultados de error más bajos para la mayoría de los nutrientes en los que se ha probado.

Una vez entrenado este modelo, hemos ordenado las bandas espectrales para cada nutriente en función de su importancia según el modelo. A continuación, hemos probado los modelos seleccionando 3, 6 y 12 bandas específicas. En este caso, no hemos probado más preprocesamientos, ya que los datos ya habían sido preprocesados antes de seleccionar las bandas con el Random Forest. Por tanto, una vez seleccionadas estas bandas, hemos entrenado los diferentes modelos que ya hemos presentado previamente.

Esta selección de bandas basada en la importancia asignada por el modelo Random Forest nos permite reducir la dimensionalidad de los datos, seleccionando sólo las bandas más relevantes para cada nutriente en particular.

4. RESULTADOS Y DISCUSIÓN

El aprendizaje automático es una potente herramienta que se ha utilizado ampliamente en diversos campos para resolver problemas complejos. En este trabajo, presentamos los resultados de los distintos modelos de aprendizaje automático que hemos descrito anteriormente. Estos resultados ofrecen una visión del potencial del aprendizaje automático y de su capacidad para impulsar el progreso en distintas áreas de la investigación y la industria.

En este artículo, se presentan los resultados para los macronutrientes primarios, macronutrientes secundarios y micronutrientes. Para cada sección, mostramos:

- Una tabla con los tres mejores modelos utilizando PCA o autoencoder (reducción de dimensionalidad con el uso de todas las bandas espectrales).
- Una tabla que muestra los resultados utilizando la reducción de dimensionalidad con selección de bandas.

Además de los mejores modelos, también incluimos una regresión lineal en cada mesa. Comparar los resultados de los modelos con una regresión lineal es relevante, puesto que la regresión lineal es uno de los métodos más sencillos y ampliamente utilizados para modelar la relación entre dos variables. Esto nos permite ver la mejora que proporcionan los modelos propuestos frente a este modelo simple. Para simplificar los términos se han utilizado las abreviaturas que veremos en la siguiente Tabla 4.1.

Abreviaturas	Abreviaturas	Significado
PREP	SNV	Standard Normal Variate
	SavGol	Suavizado Savitsky Golay
	SavGol Diff	Savitsky Golay Primera derivada
	SavGol SNV	Savitsky Golay + SNV
	Prom	Promedio
	Original	Datos originales, sin transformación
ICR	PCA	Reducción PCA
	ACC	Reducción amb <i>autoencoder</i>
MODELO	Modelo + Bag	Modelo con <i>Bagging</i>

Taula 4.1: Abreviaturas que se utilizan para mostrar el preprocessamiento que debe empleado cada modelo

4.1. Macronutrients Primaris:

- Todas las bandas

	MODEL	PREP	ICR	R ² Cal	MAE Cal	RMSE Cal	R ² Test	MAE Test	RMSE Test
N	Linear Regresion	SNV	PCA 5	0,584	0,132	0,1796	0,334	0,144	0,222
	Extra Trees	Original	PCA 20	0,853	0,078	0,108	0,794	0,098	0,123
	Ada Boost + Bag	SavGol Diff	PCA 20	0,807	0,098	0,124	0,788	0,100	0,125
	Random Forest	SavGol Diff	PCA 20	0,812	0,093	0,122	0,785	0,096	0,126
P	Linear Regresion	SNV	PCA_5	0,119	0,031	0,050	0,196	0,038	0,051
	Ada Boost + Bag	SavGol Diff	No	0,762	0,017	0,027	0,601	0,021	0,036
	K-Neighbors	SavGol Diff	PCA 25	0,718	0,020	0,029	0,593	0,022	0,036
	Grad Boost + Bag	SavGol Diff	No	0,777	0,018	0,026	0,583	0,023	0,037
K	Linear Regresion	SNV	PCA_5	0,482	0,195	0,274	0,267	0,256	0,331
	Ada Boost + Bag	SavGol Diff	No	0,651	0,140	0,221	0,672	0,144	0,221
	Ada Boost	SavGol Diff	No	0,659	0,142	0,220	0,639	0,162	0,232
	Extra Trees	SNV	PCA 9	0,674	0,136	0,217	0,629	0,155	0,235

Tabla 4.2. Resultados de los modelos para los macronutrientes primarios N, P, K. La columna PREP muestra la mejor opción de preprocesamiento de los datos. La columna ICR muestra la mejor opción de ingeniería de características, indicando el número de componentes principales o utilizando todas las bandas (No). En negrita se muestra la mejor metodología para cada macronutriente primario sobre la base del valor de R² al conjunto de prueba.

La Tabla 4.2 presenta los resultados de la estimación de los macronutrientes primarios: nitrógeno (N), fósforo (P) y potasio (K). Se muestran diferentes métricas de rendimiento de los modelos, incluyendo el coeficiente de determinación R², error absoluto medio (MAE) y error cuadrático medio (RMSE) tanto para el conjunto de calibración como para el conjunto de prueba. Para el nutriente nitrógeno, todos los modelos presentan una mejora respecto a la regresión lineal, lo que indica una mejor capacidad de estimación. La metodología basada en Extra Trees utilizando todas las bandas de entrada presenta los mejores resultados en el conjunto de prueba para el nitrógeno (R² de 0,794). Por lo que respecta al fósforo, la mejor metodología es la primera derivada del filtro Savitzky-Golay utilizando todas las bandas como entrada, con la combinación de AdaBoost y bagging obteniendo un R² de 0,60.

Por lo general, las mejores metodologías de estimación para los macronutrientes presentan coeficientes de determinación superiores a 0,6, lo que indica una buena capacidad de estimación. Estas conclusiones se mantienen cuando se analizan otras métricas de rendimiento como MAE y RMSE. Los valores mínimos de los errores de estimación coinciden con los valores máximos de R², demostrando la robustez de los resultados en la estimación del rendimiento de los modelos.

Esta información nos permite evaluar y comparar la eficacia de los modelos en la estimación de los macronutrientes primarios y concluir sobre su capacidad de estimación y rendimiento.

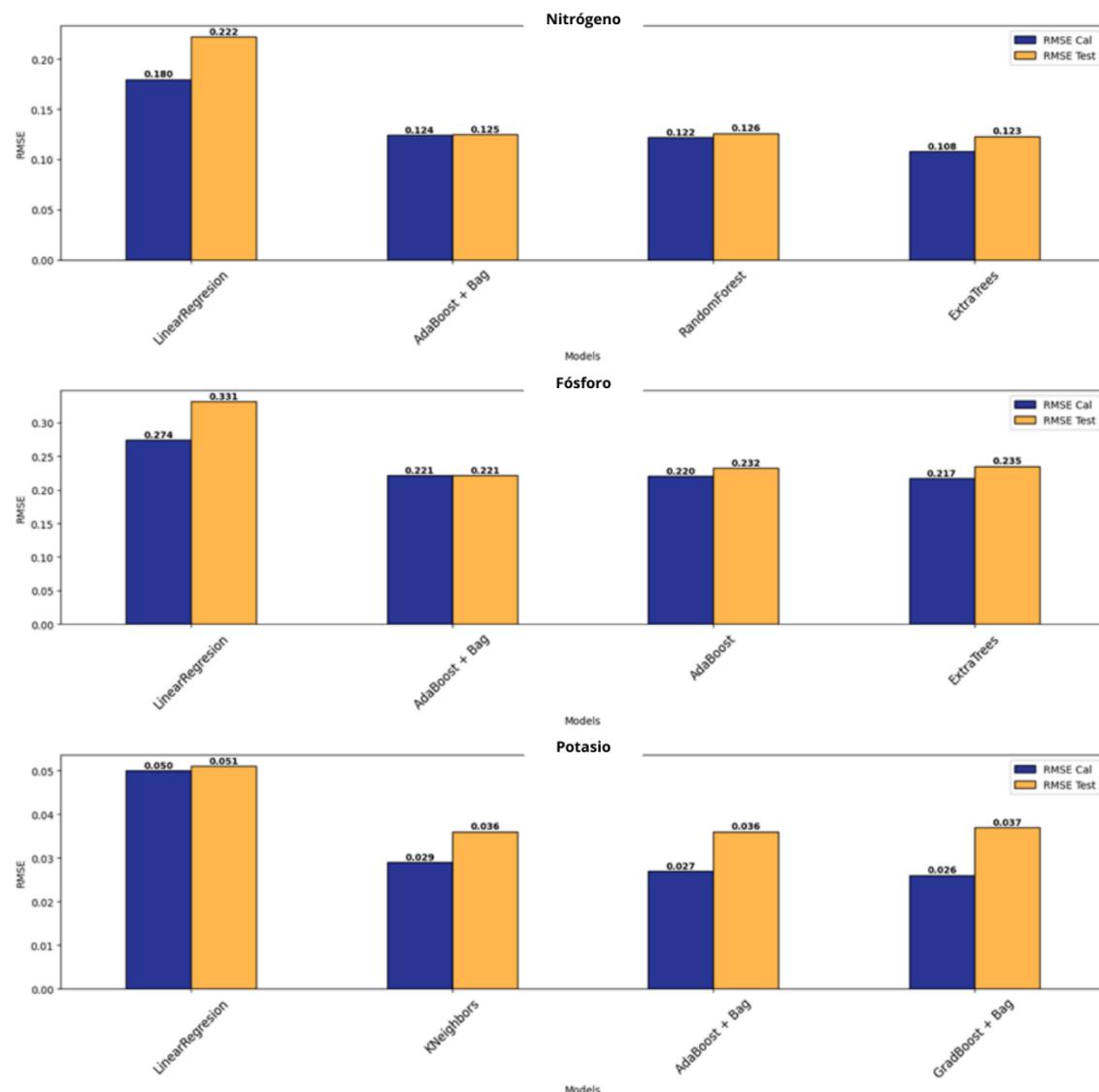


Figura 4.1. Comparació del RMSE obtengut durant la calibració i prova de les diferents millors models que predicen cada macronutriente primari.

La Figura 4.1 muestra, en un gráfico de barras, la comparación entre el RMSE de los conjuntos de calibración y prueba para cada macronutriente primario con cada una de las metodologías utilizadas. La pauta general es que el error producido en los conjuntos de calibración y prueba son muy similares (dentro de cada metodología). Este hecho pone de manifiesto que no ha habido sobreajuste en la construcción de los modelos y que la capacidad de generalización de los modelos es óptima.

- Selección de bandas

Taula 4.3. Resultados de los modelos para los macronutrientes primarios N, P y K. La columna Bandas muestra el número de bandas seleccionadas mediante Random Forest. En negrita se muestra la mejor selección en base al valor de R^2 en el conjunto de prueba para cada macronutriente.

	MODEL	BANDES	R^2 Cal	MAE Cal	RMSE Cal	R^2 Test	MAE Test	RMSE Test
N	Nu SVR	3	0,815	0,115	0,136	0,675	0,116	0,163
	K-Neighbors	6	0,811	0,102	0,133	0,753	0,102	0,131
	Extra Trees	12	0,828	0,098	0,131	0,750	0,103	0,132
P	Bayesian Ridge	3	0,680	0,024	0,033	0,488	0,029	0,043
	PLS Regression	6	0,756	0,019	0,029	0,529	0,026	0,041
	GradBoost + Bag	12	0,759	0,017	0,028	0,530	0,025	0,038
K	Extra Trees	3	0,693	0,147	0,222	0,631	0,142	0,240
	K-Neighbors	6	0,693	0,138	0,227	0,628	0,140	0,241
	Random Forest	12	0,654	0,147	0,232	0,621	0,155	0,243

En la tabla 4.3 se muestran los resultados de los modelos que estiman los macronutrientes primarios con una selección de características mediante Random Forest. Los resultados se muestran a través de los índices de rendimiento tal y como antes. Se puede observar que los resultados de los modelos empeoran respecto a los modelos que utilizan todas las bandas para estimar los niveles de los macronutrientes primarios.

Se puede observar cómo, en el caso del nitrógeno, el modelo K-Neighbors que utiliza 6 bandas espectrales es el que mejores resultados obtiene con un R2 en el conjunto de pruebas de 0,75. Este modelo, tal y como se observa en la Tabla 4.4, utiliza las bandas 590, 740, 730, 600, 560 y 530, que corresponden a la parte verde-amarilla (700-740 nm) y verde-azul (530 - 600nm) del espec. Estas dos partes del espectro, serán repetidamente seleccionadas como las más importantes para estimar a los macronutrientes. En cuanto al fósforo, el mejor modelo es el Gradient Boosting con bagging, que obtiene un R2 de 0,53. Por último, en el caso del potasio, con un Random Forest y utilizando las 12 bandas se obtiene un R2 de 0,62.

Como puede verse, en el caso del fósforo, los resultados no mejoran prácticamente un R2 de 0,5. Por el contrario, en el caso del nitrógeno y el potasio, podemos destacar unos resultados que superan un R2 de 0,6. Cabe recordar que estos modelos sólo utilizan un número limitado de bandas, lo que causa que el proceso de adquisición sea mucho más fácil.

	BANDAS MÁS IMPORTANTES
N	590, 740, 730, 600, 560, 530, 720, 700, 710, 610, 540, 580
P	710, 720, 500, 420, 510, 700, 520, 530, 690, 620, 730, 600
K	710, 720, 740, 730, 580, 560, 540, 700, 520, 530, 600, 570

Taula 4.4. Bandas seleccionadas en los modelos anteriores para predecir cada nutriente según los resultados obtenidos con el Random Forest.

4.2. Macronutrientes Secundarios

- Todas las bandas

	MODEL	PREP	ICR	R ² Cal	MAE Cal	RMSE Cal	R ² Test	MAE Test	RMSE Test
Ca	Linear Regresion	SNV	PCA 5	0,339	0,723	1,214	0,528	0,790	1,123
	Ada Boost	Original	PCA 15	0,769	0,379	0,686	0,645	0,575	0,974
	Ada Boost + Bag	Original	PCA 30	0,762	0,386	0,702	0,613	0,603	1,017
	Extra Trees	Original	PCA 25	0,769	0,426	0,708	0,604	0,646	1,029
Mg	Linear Regresion	SNV	PCA_5	0,457	0,031	0,042	0,079	0,043	0,056
	Ada Boost + Bag	SavGol SNV	PCA 15	0,607	0,026	0,037	0,552	0,030	0,039
	Ada Boost	SavGol SNV	PCA 20	0,601	0,027	0,040	0,538	0,030	0,040
	Gradient Boosting	SavGol SNV	ACC 10	0,587	0,028	0,041	0,514	0,031	0,041
S	Linear Regresion	SNV	PCA_5	0,199	0,030	0,042	0,543	0,027	0,034
	Extra Trees	SavGol Diff	No	0,661	0,018	0,027	0,670	0,021	0,028
	K-Neighbors	SavGol	PCA 12	0,608	0,020	0,028	0,649	0,022	0,030
	Ada Boost	SavGol	PCA 12	0,595	0,019	0,029	0,606	0,024	0,031

Taula 4.5. Resultados de los modelos para los macronutrientes secundarios N, P, K. La columna PREP muestra la mejor opción de preprocessamiento de los datos. La columna ICR muestra la mejor opción de ingeniería de características, indicando el número de componentes principales o utilizando todas las bandas (No). En negrita se muestra la mejor metodología para cada macronutriente secundario sobre la base del valor de R² en el conjunto de prueba.

En la Tabla 4.5 observamos los resultados relativos a la estimación de los macronutrientes secundarios Calcio, Magnesio y Azufre. Se muestran los mismos índices de rendimiento que en casos anteriores.

Se observa que el calcio con 15 componentes de PCA y un estimador Ada Boost se obtienen los mejores resultados con un coeficiente de determinación R² de 0,64 en el conjunto de prueba. En el caso del magnesio, es el macronutriente secundario que mayores dificultades presenta en la estimación. Este nutriente es predicho por una combinación de Ada Boost y Bagging que, con un preprocessamiento Savitsky Golay SNV y 15 componentes principales obtiene un 0,55 de R² en el conjunto de pruebas. Por último, en el azufre obtenemos los mejores resultados con un coeficiente de determinación R² de 0,67 mediante el estimador Extra Trees que utiliza un preprocessamiento de la primera derivada de Savinsky Golay.

Al igual que con los macronutrientes primarios, se puede ver que los valores mínimos de MAE y RMSE para cada macronutriente coinciden con los valores máximos de R² lo que indica que los modelos son robustos.

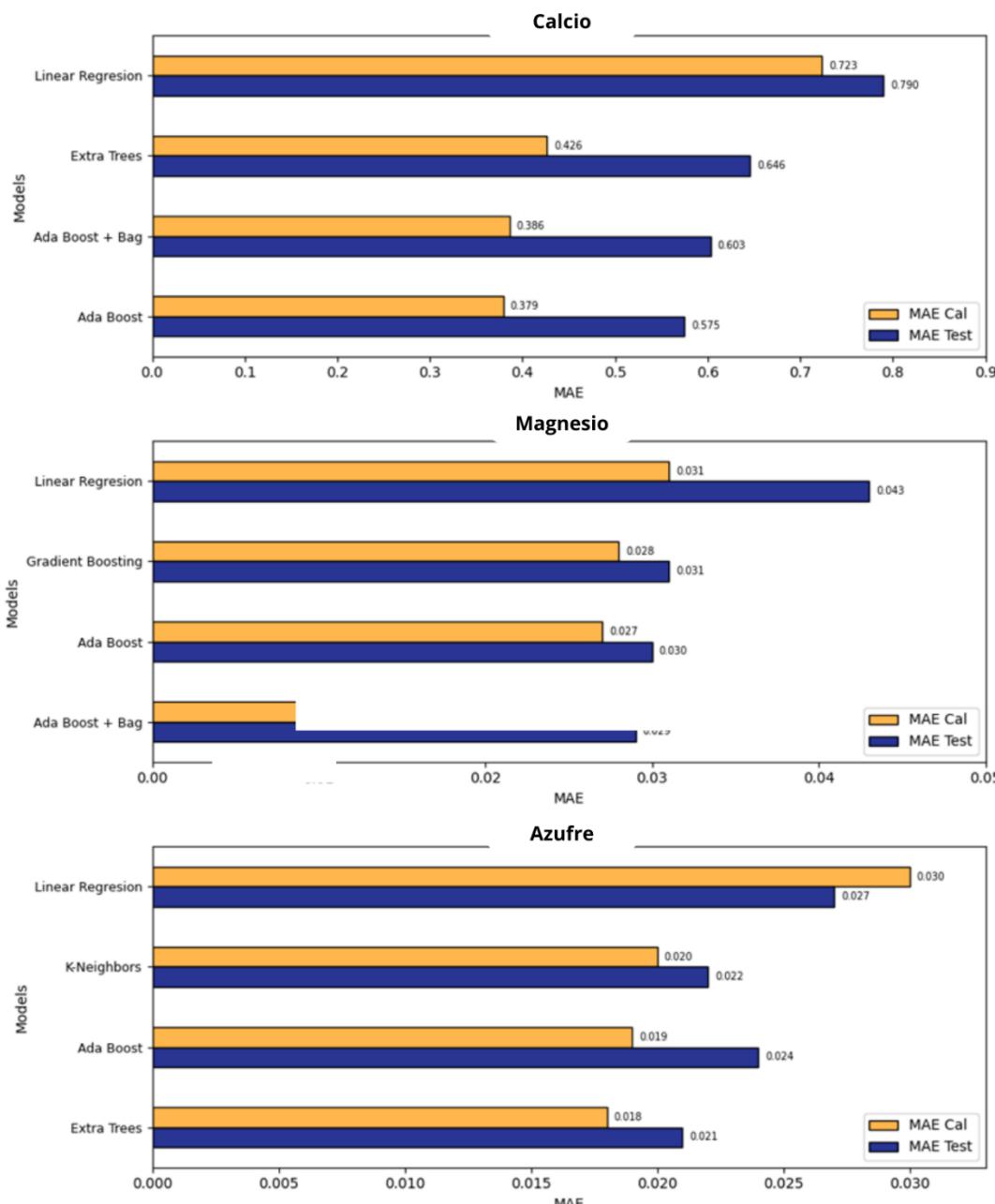


Figura 4.2. Comparación del MAE obtenido calibración y prueba de los diferentes mejores modelos que predicen cada macronutriente secundario.

La Figura 4.2 muestra, en un gráfico de barras, la comparación entre el MAE de los conjuntos de calibración y prueba para cada macronutriente secundario en base a las metodologías comentadas anteriormente. Vemos que en el calcio existe una ligera diferencia entre el error en ambos conjuntos respecto a los otros dos macronutrientes secundarios, sin embargo, los errores son muy similares en general. Al igual que en los macronutrientes primarios, este hecho demuestra que los modelos no presentan sobreajuste.

- Selección de bandas

Taula 4.6. Resultados de los modelos para los macronutrientes secundarios Ca, Mg y S. La columna Bandas muestra el número de bandas seleccionadas mediante Random Forest. En negrita se muestra la mejor selección en base al valor de R^2 al conjunto de prueba para cada macronutriente.

	MODEL	BANDES	R^2 Cal	MAE Cal	RMSE Cal	R^2 Test	MAE Test	RMSE Test
Ca	SVR	3	0,791	0,368	0,740	0,578	0,568	1,068
	PLS Regression	6	0,778	0,457	0,742	0,587	0,665	1,057
	PLS Regression	12	0,809	0,397	0,712	0,595	0,606	1,059
Mg	Random Forest	3	0,658	0,025	0,037	0,547	0,029	0,040
	K-Neighbors	6	0,613	0,025	0,037	0,573	0,027	0,038
	K-Neighbors	12	0,566	0,027	0,039	0,572	0,028	0,039
S	NuSVR	3	0,539	0,023	0,034	0,635	0,023	0,032
	Ada Boost + Bag	6	0,590	0,022	0,033	0,648	0,020	0,029
	Linear SVR	12	0,607	0,024	0,034	0,653	0,022	0,030

En la tabla 4.6 se muestran los resultados de los modelos que estiman los macronutrientes secundarios con una selección de características mediante Random Forest.

En el caso del calcio, puede verse que la regresión PLS con 12 bandas obtiene un R^2 de 0,59. En este caso, los resultados empeoran respecto a los modelos que utilizan todas las bandas como era el Ada Boost con PCA de 15 componentes que obtenía un R^2 de 645. En cambio, como puede observarse, el magnesio mejora los resultados utilizando una selección de 6 bandas, con un coeficiente R^2 de 0,57 en comparación con las 0,57. Los errores de MAE y RMSE también lo demuestran tal y como se ve en la siguiente Figura 4.3.

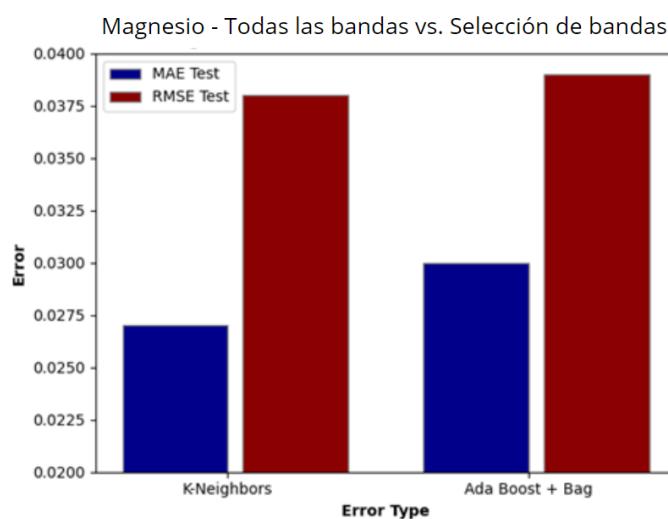


Figura 4.3. Comparación del MAE y RMSE obtenido en el conjunto de prueba del modelo K-Neighbours que utilice una selección de 6 bandas y el modelo Ada Boost + Bagging que utiliza todas las bandas.

Por último, el modelo de 6 bandas de Ada Boost y bagging, obtiene un coeficiente de R² de 0,648 para el azufre que no consigue mejorar el rendimiento que el modelo anterior que utilizaba todas las bandas.

BANDAS MÁS IMPORTANTES	
Ca	720, 730, 710, 550, 560, 540, 570, 520, 530, 740, 590, 750
Mg	710, 720, 700, 530, 750, 570, 540, 410, 560, 690, 510, 730
S	710, 720, 540, 550, 530, 570, 520, 560, 750, 400, 730, 700

Taula 4.7. Bandes seleccionades en els models anteriors per tal de predir cada nutrient segons els resultats obtinguts amb el Random Forest.

Como se observa en la Tabla 4.7, la selección de bandas más importantes para los macronutrientes secundarios, al igual que los primarios, también se centra en las dos partes comentadas anteriormente del espectro electromagnético.

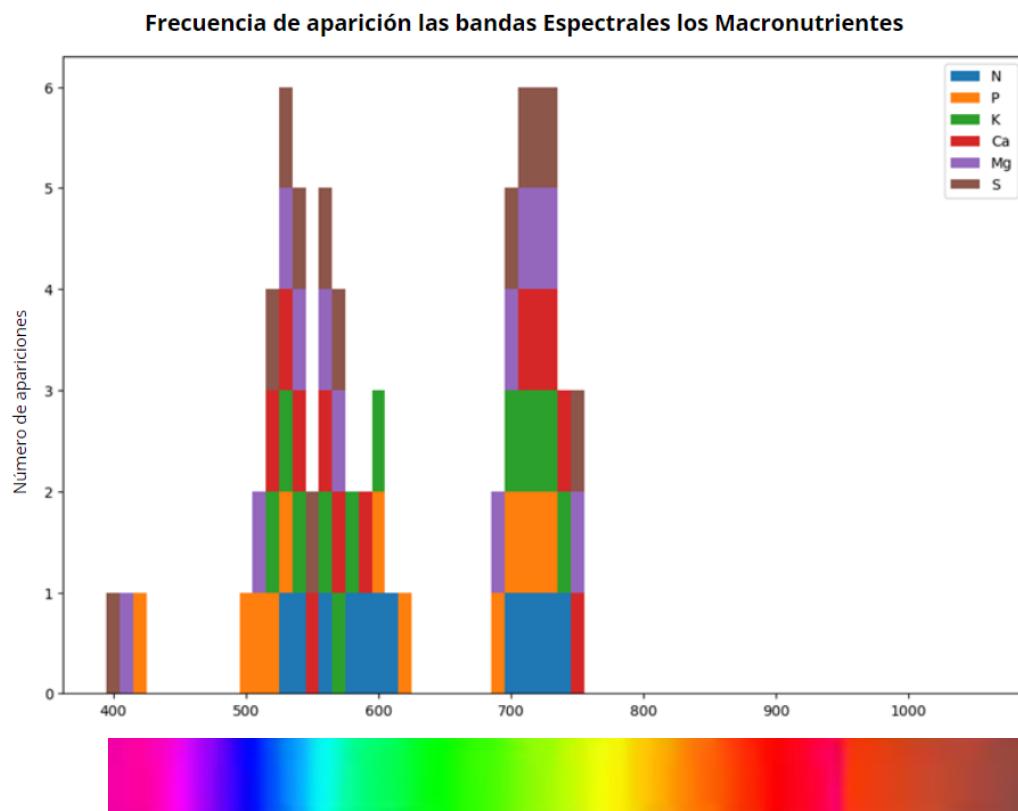


Figura 4.4. Bandas más frecuentemente clasificadas como importantes según el Random Forest para cada macronutriente que ha sido analizado.

Note en la Figura 4.4 cómo existen dos grandes zonas espectrales que resultan de utilidad para estimar el nivel de los macronutrientes. Éstas son desde 500 hasta 600 nm y desde 690 hasta 760 nm.

4.3. Micronutrientes

- Todas las bandas

	MODEL	PREP	ICR	R ² Cal	MAE Cal	RMSE Cal	R ² Test	MAE Test	RMSE Test
Na	Linear Regresion	SNV	PCA 5	0,094	0,014	0,018	0,002	0,018	0,024
	SVR	Original	PCA 12	0,235	0,012	0,015	0,057	0,013	0,017
	Linear SVR	SavGol Diff	PCA 9	0,253	0,011	0,015	0,047	0,013	0,017
	Ridge Regresion	Promig	PCA 6	0,238	0,012	0,015	0,037	0,013	0,017
Fe	Linear Regresion	SNV	PCA 5	0,419	13,615	17,496	0,225	16,811	20,602
	Extra Trees	SavGol Diff	No	0,575	11,973	15,091	0,518	14,298	17,917
	GradBoost + Bag	SavGol Diff	No	0,531	12,351	15,816	0,496	14,675	18,326
	Ada Boost	SavGol Diff	No	0,553	12,231	15,304	0,434	15,218	18,601
Mn	Linear Regresion	SNV	PCA 5	0,423	4,021	5,432	0,381	5,023	6,231
	Ada Boost + Bag	Original	PCA 20	0,657	3,202	4,629	0,594	4,128	5,480
	K- Neighbors	SavGol	PCA 5	0,634	3,499	5,024	0,584	4,254	5,544
	Ada Boost	Promig	PCA 20	0,615	3,707	4,989	0,578	4,349	5,586
Zn	Linear Regresion	SNV	PCA_5	0,253	8,941	13,995	0,193	9,986	12,813
	Extra Trees	SavGol SNV	PCA 20	0,370	5,757	7,726	0,367	5,986	7,607
	Gradient Boosting	Original	PCA 30	0,360	5,774	7,770	0,310	6,282	7,991
	Random Forest	SNV	PCA 12	0,334	6,017	7,941	0,270	6,451	8,246
Cu	Linear Regresion	SNV	PCA 5	0,122	15,515	21,362	0,052	19,265	33,268
	AdaBoost + Bag	SNV	PCA 15	0,560	11,364	14,129	0,550	11,378	15,080
	K-Neighbors	Original	PCA 12	0,603	11,733	14,458	0,547	11,971	15,148
	Ada Boost	SavGol Diff	No	0,611	11,558	14,235	0,537	12,355	15,333
B	Linear Regresion	SNV	PCA 5	0,017	5,828	8,679	0,245	6,232	8,505
	GradBoost + Bag	SavGol Diff	PCA 9	0,568	4,596	6,079	0,533	4,945	6,692
	Extra Trees	SavGol SNV	ACC 30	0,519	4,726	6,319	0,459	5,128	6,730
	Random Forest	SavGol SNV	ACC 30	0,489	4,812	6,505	0,454	5,041	6,757
Mo	Linear Regresion	SNV	PCA 5	0,004	0,033	0,042	0,001	0,040	0,046
	AdaBoost	Original	PCA 25	0,206	0,030	0,038	0,026	0,034	0,042
	Extra Trees	SavGol	PCA 25	0,253	0,031	0,038	0,025	0,034	0,043
	SGD Regressor	Promig	No	0,007	0,034	0,043	0,006	0,036	0,043

Taula 4.8. Resultados de los modelos para los micronutrientes Na, Fe, Mn, Zn, Cu, B y Mo. La columna PREP muestra la mejor opción de preprocessamiento de los datos. La columna ICR muestra la mejor opción de ingeniería de características, indicando el número de componentes principales o utilizando todas las bandas (No). En negrita se muestra la mejor metodología para cada micronutriente sobre la base del valor de R² en el conjunto de prueba.

En la Tabla 4.8 se observa que, en el caso del sodio, utilizando todas las bandas, conseguimos con SVR los mejores resultados, aunque notamos que es un micronutriente de difícil predicción con datos espectrales ya que, con PCA de 12 componentes conseguimos un R² de 0,0518 en el conjunto de prueba. En el caso del hierro puede observarse que con la primera derivada Savinsky Golay se obtiene en test un R² de 0,518. El manganeso es el

micronutriente en el que mejores resultados se consiguen, obteniendo un R² de 0,59 en el conjunto de prueba utilizando un estimador Ada Boost con bagging y 20 componentes principales. En el caso del Zinc, la Extra Trees con Savinsky Golay, SNV y 20 componentes principales, obtiene un R² de 0,36. Seguidamente, el cobre consigue el segundo mejor R² en el conjunto de prueba con un 0,55 utilizando Ada Boost y bagging. De preprocessamiento, este modelo utiliza SNV y 15 componentes principales. En el caso del boro, conseguimos un R² de 0,53 en el conjunto de prueba utilizando de estimador Gradient Boosting con bagging y un preprocessamiento primera derivada de Savinsky Golay y 9 componentes principales. Por último, con el molibdeno, conseguimos los peores resultados con un R² de 0,02 en el conjunto de prueba utilizando Ada Boost.

Tal y como se ha visto anteriormente con los macronutrientes, los valores mínimos de MAE y RMSE vuelven a coincidir con los valores máximos de R². Ya hemos dicho que este hecho indica que los modelos son robustos.

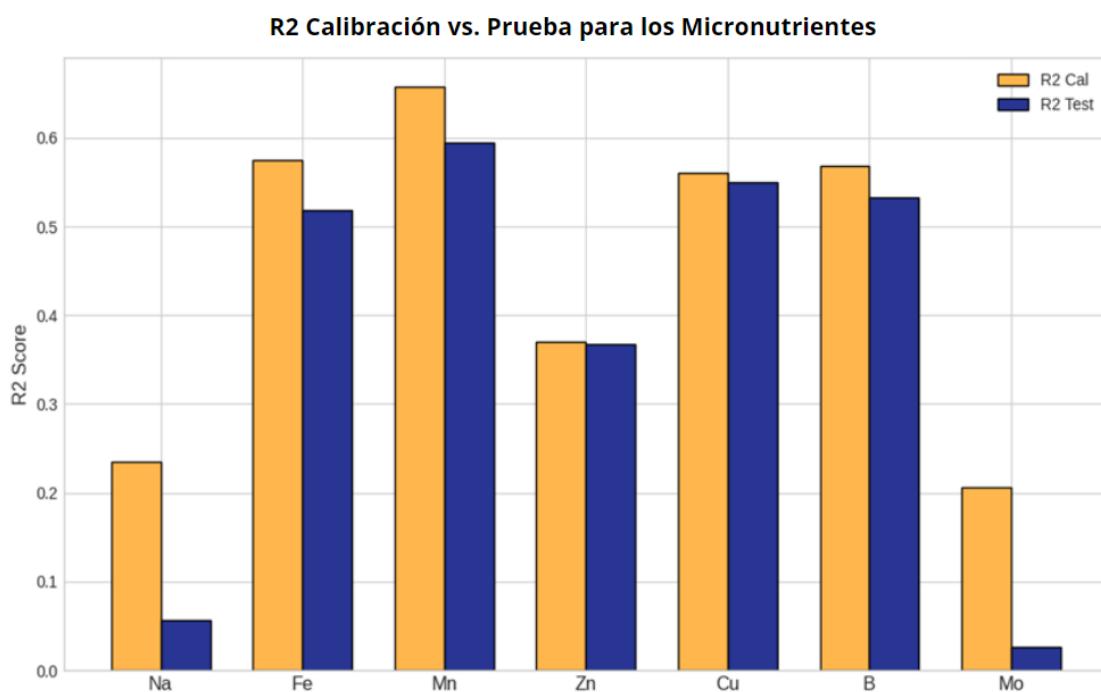


Figura 4.5. Gráfico de barras de R² obtenido por el modelo que ha logrado un mejor rendimiento en el conjunto de prueba por cada micronutriente: Na, Fe, Mn, Zn, Cu, B y Mo.

Se observan la Figura 4.5 los R² de los mejores modelos obtenidos estimando los micronutrientes. Puede apreciarse que no existe ningún modelo que supere un 0,6 de R² en el conjunto de pruebas, lo que, salvo en el caso del magnesio, todos los macronutrientes han superado. Este hecho nos indica que los micronutrientes, que están presentes en menores cantidades en las hojas del cítrico, son de mayor difícil estimación.

Sin embargo, vemos que los conjuntos calibración y test son bastante similares en los modelos de los nutrientes que superan un R² de 0,5, lo que nos indica que los modelos no se sobreajustan y, por tanto, la capacidad de generalización es óptima.

- Selección de bandas

	MODEL	BANDES	R ² Cal	MAE Cal	RMSE Cal	R ² Test	MAE Test	RMSE Test
Na	GradBoost + Bag	3	0,345	0,012	0,015	0,054	0,012	0,016
	K-Neighbors	6	0,381	0,010	0,014	0,055	0,013	0,018
	PLS Regression	12	0,377	0,011	0,014	0,025	0,013	0,017
Fe	Nu SVR	3	0,388	14,345	17,798	0,422	14,899	19,136
	Nu SVR	6	0,420	13,031	17,239	0,418	15,149	19,198
	Linear SVR	12	0,430	13,167	17,202	0,438	15,056	18,875
Mn	Ridge Regresion	3	0,522	4,076	5,664	0,460	4,325	5,820
	Ada Boost	6	0,590	3,614	5,041	0,578	4,003	5,409
	Ada Boost + Bag	12	0,522	4,059	5,388	0,564	4,148	5,493
Zn	SGD Regressor	3	0,201	7,430	9,392	0,138	7,262	8,916
	Extra Trees	6	0,216	6,930	8,638	0,225	7,590	9,794
	Nu SVR	12	0,291	6,858	8,832	0,167	7,034	8,762
Cu	PLS Regression	3	0,513	13,444	17,875	0,432	14,526	19,591
	Nu SVR	6	0,5620	11,714	16,957	0,441	13,965	19,429
	Nu SVR	12	0,551	10,984	17,164	0,475	13,215	18,838
B	SVR	3	0,498	4,981	6,874	0,533	4,912	6,559
	K-Neighbors	6	0,525	4,536	6,506	0,544	4,620	6,487
	PLS Regression	12	0,564	4,626	6,353	0,534	4,843	6,554
Mo	SGD Regressor	3	0,063	0,035	0,044	0,009	0,036	0,043
	SGD Regresion	6	0,073	0,035	0,043	0,010	0,036	0,043
	Ridge Regresion	12	0,073	0,035	0,043	0,010	0,036	0,043

Taula 4.9. Resultados de los modelos para los micronutrientes Na, Fe, Mn, Zn, Cu, B y Mo. La columna Bandas muestra el número de bandas seleccionadas mediante Random Forest. En negrita se muestra la mejor selección en base al valor de R²al conjunto de prueba para cada macronutriente.

Puede verse en la Tabla 4.9 en el caso de la selección de bandas, que los para el sodio y para el molibdeno, se obtienen valores de coeficiente de determinación muy bajos, con un R² en el conjunto de prueba inferiores a 0,1. Posteriormente, se observa que en el caso del zinc el R² de maceta es de 0,225, lo que nos indica que el modelo, al igual que utilizando todas las bandas, tampoco consigue sobrepasar un R² de 0,5.

Lo mismo ocurre con el cobre y el hierro, donde se observa un R² de 0,475 y 0,438 respectivamente en el conjunto de pruebas. Por otro lado tenemos

el manganeso, donde los R² de prueba sí supera el 0,5 con un 0,578, que se acerca al R² del modelo que utilizaba todas las bandas (R² test de 0,594). Por último, se puede apreciar que el boro mejora los resultados (R² test de 0,544) respecto a cuándo se utilizan todas las bandas (R² test de 0,533).

5. CONCLUSIONES Y FUTURA PROYECCIÓN

La metodología propuesta, basada en el uso de un sistema de visión hiperspectral Vis-NIR y técnicas de regresión de aprendizaje automático, permite estimar el contenido del nitrógeno con un R² de 0,79. Para el potasio, azufre y calcio, se obtuvieron valores de R² de 0,67, 0,67 y 0,64 respectivamente. En el caso del fósforo, manganeso y magnesio se ha obtenido unos R² de 0,60, 0,59 y 0,57 respectivamente en el conjunto de pruebas. Seguidamente, se ha obtenido un R² de 0,55, 0,54 y 0,51 para el cobre, boro y hierro respectivamente. Ya por debajo de R² de 0,5 podemos encontrar el zinc, con un R² de 0,36. Por último, en el sodio y molibdeno se han obtenido R² de 0,05 y 0,02 respectivamente.

Se comprueba que, en el caso de los macronutrientes, la estimación es mucho más precisa con un R² medio de 0,66, en comparación con los micronutrientes donde el R² medio es de 0,48. Se encuentran especiales dificultades para predecir el sodio y el molibdeno. En cambio, podemos encontrar el nitrógeno, fósforo, potasio, calcio y azufre con un R² superior a 0,6, lo que nos indica que es posible realizar esta estimación.

Además, encontramos que muchos modelos donde se seleccionan bandas pueden ser de mucha utilidad poder estimar muchos de los nutrientes estudiado ya que, el proceso de adquisición podría ser mucho menos costes.

La utilización de un sistema de visión hiperspectral Vis-NIR y los modelos de regresión de aprendizaje automático se ha evidenciado como una alternativa útil para estimar los niveles de nutrientes en las hojas de cítricos. Estos métodos ofrecen una solución no destructiva, más económica, rápida y precisa en comparación con las técnicas destructivas, costosas y lentas, cumpliendo así nuestro principal objetivo de deficiencia y reducción del impacto ambiental.

Para mejorar, en el futuro, sería beneficioso reunir más datos que cubran un rango más amplio, con mayor variedad de condiciones ambientales y niveles de nutrientes. Además, la incorporación de otras formas de datos, como podrían ser datos meteorológicos y propiedades del suelo, podría mejorar aún más la precisión de los modelos sobre todo en aquellos nutrientes en los que la estimación ha sido más tediosa.

En lo que se refiere a las técnicas de aprendizaje automático, explorar métodos más avanzados como el aprendizaje profundo podría ser una vía interesante para futuras investigaciones. Se conoce que estos modelos son

más eficientes a la hora de manejar datos de alta dimensión y pueden mejorar la precisión de los modelos de predicción nutricional.

En el campo de la ingeniería agrícola, como hemos comentado anteriormente, es de vital importancia conocer los déficits de estos nutrientes para mejorar la producción. Entonces, estos modelos suponen una poderosa herramienta para conocer mejor el estado de los árboles. Un punto fundamental es la investigación de cómo integrar estos modelos para que puedan ser de fácil manejo para los agricultores; pudiendo saber, con un coste reducido y con una funcionalidad simple, cuál es el estado de los árboles. En última instancia, esto puede conducir a mejores rendimientos de los cultivos, una mejora de la calidad de los cultivos y un reducido impacto ambiental.

Además, es importante tener en cuenta la infraestructura y los recursos disponibles para los agricultores, especialmente en las zonas rurales en las que el acceso a la tecnología y la experiencia puede ser limitado. Los modelos deben diseñarse para funcionar con dispositivos ampliamente utilizados, tales como teléfonos inteligentes, y debe proporcionarse soporte para ayudar a los agricultores a interpretar los resultados y tomar decisiones informadas.

6. BIBLIOGRAFÍA

- 1- Geovanny Rambauth-Ibarra, "Agricultura de Precisión: La integración de las TIC en la producción Agrícola", J. Comput. Electron. Sci.: Theory Appl., vol. 3 no. 1 pp. 34-38. January - June, 2022, doi: <http://dx.doi.org/10.17981/cesta.03.01.2022.04>
- 2- Nations, Food and Agriculture Organization of the United, "CITRUS FRUIT FRESH AND PROCESSED Statistical bulletin 2020". Rome, 2021.
- 3- Conselleria de Agricultura, Desarrollo Rural, Emergencia Climática y Transición Ecológica, BALANCE 2020/2021 en "Previsión de cosecha de Cítricos para 2021/2022". Servicio Documentación, Publicaciones y Estadística Departamental Septiembre 2021.
- 4- Petra Marschner , "Marschner's Mineral Nutrition of Higher Plants", (3rd ed.), Elsevier, 2012, doi: <https://doi.org/10.1016/C2009-0-63043-9>
- 5- Jiao Chen, Shaoyu Lü, Zhe Zhang, Xuxia Zhao, Xinming Li, Piao Ning, Mingzhu Liu, "Environmentally friendly fertilizers: A review of materials used and their effects on the environment", Science of The Total Environment, Volumes 613–614, 2018, Pages 829-839, ISSN 0048-9697, doi: <https://doi.org/10.1016/j.scitotenv.2017.09.186>
- 6- Leegood, R.C. and Sharkey, T.D. and von Caemmerer, S., "The Role of Chlorophyl in Photosynthesis" in Photosynthesis: Physiology and Metabolism, Springer Netherlands, 2000, doi: <https://doi.org/10.1007/0-306-48137-5>
- 7- Vashisth, T., & Kadyampakeni, D,"Diagnosis and management of nutrient constraints in citrus." in Fruit Crops: Diagnosis and Management of Nutrient Constraints (pp. 723-737). Elsevier. 2020, doi: <https://doi.org/10.1016/B978-0-12-818732-6.00049-6>
- 8- Singh, V. P., & Siddiqui, M. H, "Ionomics in Citrus Trees." in Plant Ionomics: Sensing, Signaling, and Regulation, (pp. 1-10), John Wiley & Sons Ltd. 2023, doi: <https://doi.org/10.1002/9781119803041>
- 9- Zwinkels, J. , "Light, Electromagnetic Spectrum" in: Luo, R. (eds) Encyclopedia of Color Science and Technology. Springer, Berlin, Heidelberg, 2021, doi: https://doi.org/10.1007/978-3-642-27851-8_204-1
- 10- Yuxin Chen; Yuejie Chi; Jianqing Fan; Cong Ma, "Spectral Methods for Data Science: A Statistical Perspective", 2021.doi: <https://doi.org/10.48550/arXiv.2012.08496>
- 11- Dheeraj Kumar Singh, Manik Pradhan, Arnulf Materny," Modern Techniques of Spectroscopy. Basics, Instrumentation, and Applications", Springer Singapore, 1st ed. 2021, doi: <https://doi.org/10.1007/978-981-33-6084-6>
- 12- Antonio Fazari, Oscar J. Pellicer-Valero, Juan Gómez-Sanchis, Bruno Bernardi, Sergio Cubero, Souraya Benalia, Giuseppe Zimbalatti, Jose Blasco, "Application of deep convolutional neural networks for the detection of anthracnose in olives using VIS/NIR hyperspectral images", Computers and Electronics in Agriculture, Volume 187, 2021, 106252, ISSN 0168-1699, doi: <https://doi.org/10.1016/j.compag.2021.106252>
- 13- Quinones, A., Martínez-Alcántara, B., Legaz, F. & Bermejo, A. (2015). Fraccionamiento del calcio en los distintos órganos de plantas jóvenes de cítricos

- cultivadas en distintas condiciones de aporte de calcio. Levante Agrícola, 425, 41-46. doi: <http://hdl.handle.net/20.500.11939/7156>
- 14- Emilio Soria Olivas, Manuel Antonio Sánchez-Montaños Isla, Ruth Gamero Cruz, Borja Castillo Caballero, "3.2.3. Sobreajuste" en "Sistemas de Aprendizaje Automático", Grupo Editorial ra-ma, 2023, https://www.ra-ma.es/libro/sistemas-de-aprendizaje-automatico_147454/
- 15- Refaeilzadeh, P., Tang, L., Liu, H., "K-FOLD" in "Cross-Validation " In: LIU, L., ÖZSU, M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA, doi: https://doi.org/10.1007/978-0-387-39940-9_565
- 16- Sharma, S., Gosain, A., Jain, S. (2022). A Review of the Oversampling Techniques in Class Imbalance Problem. In: Khanna, A., Gupta, D., Bhattacharyya, S., Hassanien, A.E., Anand, S., Jaiswal, A. (eds) International Conference on Innovative Computing and Communications. Advances in Intelligent Systems and Computing, vol 1387. Springer, Singapore. doi: https://doi.org/10.1007/978-981-16-2594-7_38
- 17- Emily Grisanti, Maria Totska, Stefan Huber, Christina Krick Calderon, Monika Hohmann, Dominic Lingenfelser, Matthias Otto, "Dynamic Localized SNV, Peak SNV, and Partial Peak SNV: Novel Standardization Methods for Preprocessing of Spectroscopic Data Used in Predictive Modeling", Journal of Spectroscopy, vol. 2018, Article ID 5037572, 14 pages, 2018, doi: <https://doi.org/10.1155/2018/5037572>
- 18- Savitzky, A.; Golay, M.J.E. (1964). "Smoothing and Differentiation of Data by Simplified Least Squares Procedures". Analytical Chemistry. 36 (8): 1627–1639. doi: <http://doi.org/10.1021/ac60214a047>
- 19- I. T. Jolliffe (2001). "Principal Component Analysis". Springer New York, NY. doi: <https://doi.org/10.1007/b98835>
- 20- Baraniuk, R., Donoho, D., and Gavish, M. (2020). "The science of deep learning". Proc. Natl. Acad. Sci. U.S.A. 117, 30029–30032. doi: <http://doi.org/10.1073/pnas.2020596117>
- 21- Emilio Soria Olivas, Pablo Rodríguez Belenguer, Quique García Vidal, Fran Vaquer Estarlich, Juan Vicent Camisón, Jorge Vila Tomás, "Cap 2. Modelos Neuronales Multifunción" en "Inteligencia Artificial. Casos prácticos con Aprendizaje Profundo", Grupo Editorial ra-ma, 2022, https://www.ra-ma.es/libro/inteligencia-artificial_139032/
- 22- Arbib, M.A. "The Handbook of Brain Theory and Neural Networks". MIT Press. 2002. <https://mitpress.mit.edu/9780262511025/the-handbook-of-brain-theory-and-neural-networks/>
- 23- Dor Bank, Noam Koenigstein, Raja Giryes. "Autoencoders". 2021, doi: <https://doi.org/10.48550/arXiv.2003.05991>
- 24- Buduma, N., Locascio, N. "Fundamentals of Deep Learning". O'Reilly Media. 2017, doi: <https://www.oreilly.com/ai/free/files/fundamentals-of-deep-learning-sampler.pdf>
- 25- Alpaydin E. "Machine Learning". MIT Press. 2016. <https://mitpress.mit.edu/9780262529518/machine-learning/>

6. Bibliografía

- 26- Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, Brian D. Marx. "Regression. Models, Methods and Applications". Springer Berlin. 2022. doi: <https://doi.org/10.1007/978-3-662-63882-8>
- 27- Vinzi, V. E., Chin, W. W., Henseler, J., & Wang, H, "Handbook of Partial Least Squares: Concepts, Methods and Applications", Springer, 2010, doi: <https://doi.org/10.1007/9783-540-32827-8>
- 28- Ciaburro, G., & Joshi, P. "Constructing a k-nearest neighbors regressor" in "Python Machine Learning Cookbook ". Second Edition. O'Reilly Media . 2019. doi: <https://www.oreilly.com/library/view/machine-learning-with/9781491989371/>
- 29- Naiyang Deng, Yingjie Tian, Chunhua Zhang, "Support Vector Machines Optimization Based Theory, Algorithms, and Extensions". CRS Press. 1st Edition. 2012.
- 30- Halawi, L., Clarke, A., George, K. "Decision Trees and Ensemble. In: Harnessing the Power of Analytics". Springer, Cham. 2022. doi: https://doi.org/10.1007/978-3-030-89712-3_5
- 31- Sarang, P. (2023). Ensemble: Bagging and Boosting. In: Thinking Data Science. The Springer Series in Applied Machine Learning. Springer, Cham. doi: https://doi.org/10.1007/978-3-031-02363-7_5
- 32- Feurer, M., Hutter, F, "Hyperparameter Optimization". In: Hutter, F., Kotthoff, L., Vanschoren, J. (eds) Automated Machine Learning. The Springer Series on Challenges in Machine Learning. Springer, Cham, 2019, doi: https://doi.org/10.1007/978-3-030-053185_1