



GRAU EN CIÈNCIA DE DADES



VNIVERSITAT  
DE VALÈNCIA

TREBALL FINAL DE GRAU

---

ESTIMACIÓ MITJANÇANT ANÀLISI NO  
DESTRUCTIU DELS NIVELLS  
NUTRICIONALS DE FULLES DE  
CÍTRICS APLICANT TÈCNIQUES DE  
MACHINE I DEEP LEARNING

---

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





TREBALL FINAL DE GRAU

---

ESTIMACIÓ MITJANÇANT ANÀLISI NO DESTRUCTIU  
DELS NIVELLS NUTRICIONALS DE FULLES DE CÍTRICS  
APLICANT TÈCNIQUES DE MACHINE I DEEP LEARNING

---

AUTOR: GUILLEM MIRALLES GADEA

TUTOR: JUAN GÓMEZ SANCHIS

JULIOL 2023



**Declaració d'autoria:**

Jo, Guillem Miralles Gadea, declare l'autoria del Treball Final de Grau titulat " Estimació Mitjançant Anàlisi No Destructiu Dels Nivells Nutricionals De Fulles De Cítrics Aplicant Tècniques De Machine I Deep Learning " i que el citat treball no infringeix les lleis vigents sobre propietat intel·lectual. El material no original que figura en este treball ha estat atribuït als seus legítims autors.

València, 15 de juliol de 2023.

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).

---



## **Agraïments**

Vull expressar la meua immensa gratitud cap als meus estimats pares, els quals han estat la llum que ha guiat la meua vida fins ara. Sense el seu amor inesgotable, el seu suport incondicional i la seu confiança incansable en mi, jo no seria la persona que soc avui. També a la meua germana, que ha estat una gran font d'inspiració i suport en aquest camí.

Vull donar les gràcies a tot el personal de l'Institut Valenciac d'Investigacions Agràries. Ha estat un plaer treballar amb professionals tan dedicats i apassionats i poder aprendre de la seu experiència.

Finalment, però no menys important, vull agrair als meus professors i companys de classe, especialment al meu tutor Juan Gómez. Gràcies per compartir amb mi la vostra experiència i coneixements, els quals han estat crucials per al meu creixement personal i acadèmic.

Gràcies a tots.



# ÍNDEX GENERAL

<b>1. INTRODUCCIÓ.....</b>	<b>18</b>
<b>2. OBJECTIUS DE L'ESTUDI.....</b>	<b>24</b>
<b>3. MATERIALS I MÈTODES .....</b>	<b>26</b>
<b>3.1. Creació del conjunt de dades.....</b>	<b>26</b>
<b>3.2. Anàlisi exploratòria de les dades.....</b>	<b>32</b>
<b>3.3. Neteja, divisió i oversampling de les dades:.....</b>	<b>39</b>
<b>3.4. Preprocesament espectral.....</b>	<b>42</b>
<b>3.5. Reducció de la dimensionalitat.....</b>	<b>53</b>
<b>3.6. Models.....</b>	<b>68</b>
<b>3.7. Avaluació dels models.....</b>	<b>81</b>
<b>3.8. Construcció de models i proves realitzades. ....</b>	<b>84</b>
<b>4. RESULTATS I DISCUSSIÓ .....</b>	<b>88</b>
<b>4.1. Macronutrients Primaris: .....</b>	<b>89</b>
<b>4.2. Macronutrients Secundaris.....</b>	<b>92</b>
<b>4.3. Micronutrients.....</b>	<b>96</b>
<b>5. CONCLUSIONS I FUTURA PROJECCIÓ .....</b>	<b>100</b>
<b>6. BIBLIOGARFIA.....</b>	<b>102</b>

## **ÍNDEX DE FIGURES:**

<i>Figura 1.1: Peixos i crustacis morts a la vora del Mar Menor. Fotografia: Asociación de Naturalistas del Sureste (ANSE).....</i>	<b>20</b>
<i>Figura 1.2. Espectre electromagnètic VIS-NIR. Font: Blog Digi-Key Spectral Sensors .....</i>	<b>23</b>
<i>Figura 3.1. Esquema del procés d'adquisició de mostres per a crear el conjunt de dades. Procés d'adquisició d'imatges hiperespectrals i anàlisi destructiu. Font: Elaboració pròpia.</i>	<b>28</b>
<i>Figura 3.2. Procés d'adquisició de la imatge hiperespectral d'una fulla de cítric.....</i>	<b>29</b>
<i>Figura 3.3. Preprocesament a les imatges hiperespectrals adquirides.....</i>	<b>30</b>
<i>Figura 3.4. Correcció del desplaçament i concatenació.....</i>	<b>31</b>
<i>Figura 3.5. PCA i operacions morfològiques de la imatge. ....</i>	<b>31</b>
<i>Figura 3.6. Diferència estandarditzada entre fulles Joves i Velles segons el nutrient. ....</i>	<b>33</b>
<i>Figura 3.7. Diferència estandarditzada entre les orientacions segons el nutrient. ....</i>	<b>34</b>
<i>Figura 3.8. Espectre mitjà de les mostres registrades de cada arbre .....</i>	<b>36</b>
<i>Figura 3.9. Espectre mitjà de les fulles joves i velles registrades .....</i>	<b>37</b>
<i>Figura 3.10. Espectre mitjà segons l'orientació en la que es va registrar la mostra.....</i>	<b>37</b>
<i>Variabilitat Bandes Espectrals.....</i>	<b>37</b>
<i>Figura 3.11. Bandes espectrals que registren major variabilitat .....</i>	<b>38</b>
<i>Figura 3.12. Exemple del procés K-Fold amb 3 particions.....</i>	<b>41</b>
<i>Figura 3.13. Representació gràfica de l'estandardització de 4 mostres espectrals. ....</i>	<b>44</b>
<i>Figura 3.14. Representació gràfica de 4 mostres espectrals originals.....</i>	<b>44</b>
<i>Figura 3.15. Representació gràfica de una mostra espectrals transformada amb Savinsky-Golay i la mateixa mostra espectral original.....</i>	<b>47</b>
<i>Figura 3.16. Representació de 4 mostres espectrals transformades amb Savinsky-Golay.....</i>	<b>48</b>
<i>Figura 3.17. Representació gràfica de una mostra espectrals transformada amb la primera derivada de Savinsky-Golay i la mateixa mostra espectral original.....</i>	<b>50</b>
<i>Figura 3.18. Representació gràfica de 4 mostres espectrals transformades amb Savinsky-Golay i la estandardització estàndard. .....</i>	<b>51</b>
<i>Figura 3.19. Representació gràfica de 4 mostres espectrals transformades amb la diferència respecte al espectre promig. ....</i>	<b>52</b>
<i>Figura 3.20. Variància explicada de les components de PCA. ....</i>	<b>57</b>
<i>Figura 3.21. Primera component vs. segona component de les dades estandarditzades.....</i>	<b>58</b>
<i>Figura 3.22. Parts de la neurona humana. Font: Enclopèdia Humanitats. ....</i>	<b>60</b>
<i>Figura 3.23. Parts de la neurona artificial. Font: Universidad de Murcia. ....</i>	<b>61</b>
<i>Figura 3.24. Perceptró Simple. Font: Universidad de Murcia.....</i>	<b>62</b>
<i>Figura 3.25. Red Neuronal Profunda. Font: IBM Blog "What are neural networks?".....</i>	<b>63</b>
<i>Figura 3.26. Estructura General d'un Autoencoder. Font: Fernando Sancho Caparrini Blog... </i>	<b>64</b>
<i>Figura 3.28. Proves realitzades al projecte per a cada model i cada nutrient.....</i>	<b>86</b>

<i>Figura 4.1. Comparació del RMSE obtingut calibració i prova dels diferents millors models que prediuen cada macronutrient primari.</i> .....	90
<i>Figura 4.2. Comparació del MAE obtingut calibració i prova dels diferents millors models que prediuen cada macronutrient secundari.</i> .....	93
<i>Figura 4.3. Comparació del MAE i RMSE obtingut al conjunt de prova del model K-Neighbours que utilitzqa una selecció de 6 bandes i el model Ada Boost + Bagging que utilitza totes les bandes.</i> .....	94
<i>Figura 4.4. Bandes més freqüentment classificades com a importants segons el Random Forest per a cada macronutrient que ha sigut estimat.</i> .....	95
<i>Figura 4.5. Gràfic de barres de R<sup>2</sup> obtingut pel model que ha aconseguit un millor rendiment al conjunt de prova per cada micronutrient: Na, Fe, Mn, Zn, Cu, B i Mo.</i> .....	97

## **ÍNDEX DE TAULES:**

<i>Taula 3.1. Exemple de la informació general d'algunes mostres del arbre 'I02'. Font: Elaboració pròpria.</i> .....	<b>28</b>
<i>Taula 3.2. Diferència estandarditzada entre les orientacions segons el nutrient.....</i>	<b>35</b>
<i>Taula 3.3. Principals funcions d'activació. Font: Sagar Sharma en Activation Functions in Neural Networks (Medium).....</i>	<b>61</b>
<i>Taula 3.4. Diferencies entre Gradient Boosting Regressor i AdaBoost. ....</i>	<b>80</b>
<i>Taula 3.5. Com analitzar els residus d'un model.....</i>	<b>83</b>
<i>Taula 4.1: Abreviatures que s'utilitzen per a mostrar el preprocessaments que ha emprat cada model .....</i>	<b>88</b>
<i>Taula 4.2. Resultats dels models per als macronutrients primaris N, P, K. La columna PREP mostra la millor opció de preprocesament de les dades. La columna ICR mostra la millor opció d'enginyeria de característiques, indicant el nombre de components principals o emprant totes les bandes (No). En negreta es mostra la millor metodologia per a cada macronutrient primari sobre la base del valor de R<sup>2</sup>al conjunt de prova. ....</i>	<b>89</b>
<i>Taula 4.3. Resultats dels models per als macronutrients primaris N, P i K. La columna Bandes mostra el nombre de bandes seleccionades mitjançant Random Forest. En negreta es mostra la millor selección en base del valor de R<sup>2</sup> al conjunt de prova per a cada macronutrient. ....</i>	<b>91</b>
<i>Taula 4.4. Bandes seleccionades en els models anteriors per tal de predir cada nutrient segons els resultats obtinguts amb el Random Forest. ....</i>	<b>92</b>
<i>Taula 4.7. Bandes seleccionades en els models anteriors per tal de predir cada nutrient segons els resultats obtinguts amb el Random Forest .....</i>	<b>95</b>
<i>Taula 4.8. Resultats dels models per als micronutrients Na, Fe, Mn, Zn, Cu, B i Mo. La columna PREP mostra la millor opció de preprocesament de les dades. La columna ICR mostra la millor opció d'enginyeria de característiques, indicant el nombre de components principals o emprant totes les bandes (No). En negreta es mostra la millor metodologia per a cada micronutrient sobre la base del valor de R<sup>2</sup> al conjunt de prova.....</i>	<b>96</b>
<i>Taula 4.9. Resultats dels models per als micronutrients Na, Fe, Mn, Zn, Cu, B i Mo. La columna Bandes mostra el nombre de bandes seleccionades mitjançant Random Forest. En negreta es mostra la millor selección en base del valor de R<sup>2</sup>al conjunt de prova per a cada macronutrient. ....</i>	<b>98</b>

## 1. INTRODUCCIÓ

L'aplicació de les anomenades Tecnologies de la Informació i la Comunicació (TIC) en el sector agrícola ha originat la capacitat de prendre decisions i adoptar correccions a partir de mesures preses. D'aquesta manera, s'aconsegueix millorar la productivitat dels camps i reduir l'impacte mediambiental. És per això que s'ha incrementat tant l'ús d'aquest tipus de tècniques com l'aparició de noves aplicacions, revolucionant, en conseqüència, el món agrícola.

Aquestes tecnologies es basen en la recollida de dades, que mitjançant un anàlisi adequat, proporcionen coneixement. L'aplicació d'aquestes tecnologies o tècniques en el sector agrícola rep el nom d'Agricultura de Precisió [1]. Bàsicament, d'aquesta manera s'obté, amb menors recursos, l'optimització de la productivitat alhora de mantenir els estàndards de qualitat. A més, les TIC permeten garantir de manera efectiva la seguretat dels productes alimentaris mitjançant l'ús de diverses tècniques. Entre les múltiples aplicacions destaquen el control de cultius per satèl·lit, la conducció automàtica de maquinària agrícola, el control del regadiu amb drons, etc.

### 1.1. Els cítrics: una indústria en expansió i impacte econòmic a la Comunitat Valenciana

Segons *Oxford Languages*, es pot definir un fruit cítric com: "Conjunt de fruites de sabor àcid o agredolç"; i un arbre cítric com: "Plantes que produeixen aquests fruits". Per tant, els cítrics són aquells arbres que produeixen fruites amb sabor àcid o agredolç, que estan comercialitzades, com per exemple la llima, la taronja, el pomelo o la mandarina.

Aquests cultius són àmpliament consumits a tot el món, i a més d'utilitzar-se per a usos alimentaris, també s'apliquen en medicina, farmàcia, biocombustibles, entre altres àmbits, a causa de les seues propietats. Segons l'informe estadístic de les Nacions Unides [2] de l'any 2020, es van produir 143.755.600 tones de cítrics a tot el món, i es nota una tendència positiva de creixement. Aquest informe també destaca que s'han millorat les pràctiques agrícoles en els últims seixanta anys, augmentant en un 70% el rendiment dels cultius de cítrics, amb un increment del 23% en la superfície collida i un 16% en la seu producció.

L'agricultura de precisió ha tingut un paper molt important en la millora de la producció de cítrics en els últims anys. Per exemple, la detecció ràpida de la plaga del psílid asiàtic, que afecta els cítrics, mitjançant sistemes d'intel·ligència artificial, ha permès prendre mesures anticipades i evitar la desaparició dels camps de cultiu.

Aquest cultiu té un gran valor a la Comunitat Valenciana, ja que té un impacte econòmic significatiu en el territori. Segons les dades de la Conselleria d'Agricultura de la Generalitat Valenciana, segons el seu informe de previsió de collites per a la campanya de 2021/2022 [3], es van produir un total de 3.508.051 tones de cítrics en el territori valencià durant la campanya de 2010/2021. Això converteix la Comunitat Valenciana en la principal regió productora de cítrics a nivell nacional, amb la taronja i la mandarina com a principals varietats de cítrics produïdes.

## **1.2. La importància dels nutrients en el creixement saludable dels arbres: equilibri i impacte mediambiental.**

Els arbres i les plantes requereixen nutrients essencials per a sobreviure i prosperar [4]. Quan un arbre té una baixa nutrició, es debilita i es torna vulnerable a plagues i malalties. Els arbres absorbeixen els seus nutrients a través del sòl, a més de necessitar humitat i llum solar. Una correcta nutrició permet que l'arbre creixi adequadament i, en el cas dels cítrics, produeixi aliments de més quantitat i millor qualitat.

Tenint en compte que una nutrició adequada dels arbres és fonamental per a mantenir-los sans i per a obtenir aliments de bona qualitat, és important conèixer com es poden tractar. Per solucionar la falta de nutrients, un arbre pot rebre suplements mitjançant els anomenats fertilitzants o adobs. No obstant això, l'ús excessiu de fertilitzants pot ocasionar diversos problemes. D'una banda, pot contribuir a l'alliberament de gasos d'efecte hivernacle a l'atmosfera. D'altra banda, pot causar el que es coneix com a eutrofització, que és l'addició de nutrients externs (com l'excés de nitrogen) als cursos d'aigua [5].

A la Figura 1.1 següent es mostren les conseqüències de l'eutrofització dels cursos d'aigua en el cas del Mar Menor. L'eutrofització pot provocar un creixement ràpid de microorganismes a l'aigua, els quals poden consumir tot l'oxigen disponible en aquests cursos d'aigua i crear el que es coneix com a zones mortes. A més, pot generar la proliferació d'algues que produeixen substàncies tòxiques (floracions d'algues nocives).



**Figura 1.1:** Peixos i crustacis morts a la vora del Mar Menor. Fotografia: Asociación de Naturalistas del Sureste (ANSE)

Llavors, és necessari equilibrar els beneficis de la fertilització (major aportació d'aliments) amb les conseqüències negatives de l'excés de fertilitzant (emissions de gasos d'efecte hivernacle). Per tant, conèixer adequadament les necessitats de l'arbre per evitar un ús excessiu de fertilitzants té quatre implicacions importants:

- Reduir els efectes ambientals negatius.
- Minimitzar el malbaratament de recursos innecessaris.
- Optimitzar la producció de fruita.
- Mantenir els arbres i els productes en un bon estat de salut.

## Clorofil·la

Abans d'aprofundir en els nutrients, és important entendre què és la clorofil·la i el seu paper fonamental. La clorofil·la és un pigment verd que es troba en pràcticament totes les plantes, algues i cianobacteris. La quantitat de clorofil·la en un arbre està relacionada amb la disponibilitat de nutrients en el sòl. Aquesta molècula és necessària per a la fotosíntesi [6], un procés vital per al creixement de l'arbre. Si el sòl no conté els nutrients necessaris, l'arbre no podrà produir prou clorofil·la i el seu creixement serà afectat negativament.

## Nutrients:

D'una manera general, es fa una distinció entre els macronutrients, que són els nutrients requerits en major quantitat, i els micronutrients, que són els elements necessaris en menor quantitat. Aquesta divisió no implica que un nutrient siga més important que un altre, sinó que simplement es necessiten en diferents proporcions. La combinació adequada de macronutrients i micronutrients proporciona un sòl en òptimes condicions de salut. A cada grup tenim:

**1- Macronutrients primaris:**

**1.1. Nitrogen (N):** El nitrogen ajuda els arbres i les plantes a convertir l'aigua, la llum solar i el sòl en aliments. Estimula el creixement de les fulles i contribueix al seu color verd.

**1.2. Fòsfor (P):** Responsable del correcte creixement de les arrels. També ajuda els arbres a produir llavors, flors i fruits. Enforteix l'arbre, fent-lo menys vulnerable a les malalties.

**1.3. Potassi (K):** Quan un arbre té prou potassi, produeix fruits saludables. El potassi també protegeix l'arbre de danys durant les estacions fredes. Col·labora amb el fòsfor per protegir l'arbre de malalties i ajudar-lo a recuperar-se d'aquestes.

**2- Macronutrients secundaris:**

**2.1. Calci (Ca):** Construeix les parets cel·lulars de les plantes i enforteix les tiges dels arbres.

**2.2. Magnesi (Mg):** Un altre component de la clorofil·la. També té un paper en el creixement dels arbres.

**2.3. Sofre (S):** Els arbres necessiten el sofre principalment per produir proteïnes. També intervé en la formació de clorofil·la, en la producció de llavors i en la millora de les arrels i el creixement general de la planta.

**3- Micronutrients:**

**3.1. Sodi (Na):** Auxiliar en el metabolisme i la síntesi de clorofil·la.

**3.2. Ferro (Fe):** Necessari per a la producció de clorofil·la.

**3.3. Manganès (Mn):** Suport per alsenzims en la digestió dels carbohidrats i en la conversió del nitrogen.

**3.4. Zinc (Zn):** Un component delsenzims que controlen el creixement.

**3.5. Coure (Cu):** Part essencial de la reproducció. Ajuda l'arbre en l'absorció i digestió de nutrients a través de les arrels.

**3.6. Boro (B):** Controla i equilibra altres nutrients. També és necessari per a la producció de fruits i llavors.

**3.7. Molibdè (Mo):** Un altre mineral utilitzat per a la transformació del nitrogen.

Tota aquesta informació es troba detallada a [7].

### 1.3. Anàlisi espectral per a l'estimació de nutrients en cultius: Una alternativa eficient i econòmica

Com s'ha pogut comprovar, és molt important conèixer les deficiències de nutrients en els cultius i saber quins i en quina mesura no es troben en la quantitat òptima. Per tant, actualment es disposa de diverses tècniques que ajuden a avaluar l'estat dels cultius.

En l'actualitat, l'eina més utilitzada és l'anàlisi de l'ionòmica foliar de les fulles. L'ionòmica és la mesura de l'acumulació total de metalls, metal·loides i no metalls en els organismes vius. L'ionòmica vegetal s'ha aplicat en diverses investigacions durant l'última dècada [8].

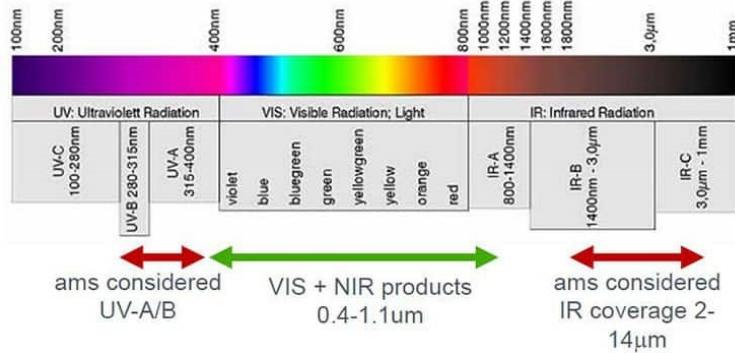
No obstant això, aquesta tècnica té un cost econòmic i temporal elevat, ja que requereix ànàlisis químiques en laboratori. És per això que s'estan investigant tècniques alternatives que permeten estimar aquests nutrients amb un menor cost econòmic i temporal, encara que puguin tenir una precisió lleugerament inferior.

### 1.4. Reflectància de llum: Anàlisi a través de l'espectre visible i infraroig proper

L'espectre electromagnètic es defineix com el conjunt de tota la radiació electromagnètica que existeix a l'univers. La freqüència es defineix per la longitud d'ona, que és la distància des del pic d'una ona fins al pic següent. Aquests dos atributs estan inversament relacionats: com més gran siga la freqüència, menor serà la longitud d'ona i viceversa.

Dins d'aquest espectre hi ha diversos tipus d'ones, des dels raigs gamma (baixa longitud d'ona/alta freqüència) fins a les ones de ràdio (gran longitud d'ona/baixa freqüència). En aquest projecte, es treballa amb les longituds d'ona que pertanyen a una part de l'espectre visible (VIS) i infraroig proper (NIR). Aquesta informació es pot apreciar gràficament a la Figura 1.2.

- **Vis:** Aquesta part correspon al segment de l'espectre electromagnètic que l'ull humà pot veure. És a dir, la llum visible. En general, l'ull humà pot detectar ones d'entre 380-700 nanòmetres (nm).
- **NIR:** Comprén radiació de baixa freqüència adjacent als tons vermellos en el visible. Aquest espectre va des de 700 fins a 2500 nm tot i que no hi ha una definició universalment acceptada.



**Figura 1.2.** Espectre electromagnètic VIS-NIR. Font: Blog Digi-Key Spectral Sensors

Així doncs, irradiant un material amb diferents longituds d'ona, és possible mesurar la seva reflectància. La reflectància es defineix com el grau de llum que és reflectit per la superfície d'un material. El grau d'aquesta reflectància dependrà del nivell d'absorció de llum per part del material. Per exemple, un material de color blanc reflectirà més la il·luminació que un de color negre, de manera que tindrà un índex de reflectivitat més alt. Aquesta informació es troba detallada a [9].

Els equips encarregats de mesurar aquesta reflectància en un material il·luminat amb una determinada longitud d'ona es coneixen com a espectròmetres de reflectància. La reflectància es mesura en una escala de 0 a 1 (o també en percentatge de 0% a 100%). Si la reflectància d'un material és 0, significa que no reflecteix cap llum i, per tant, ho absorbeix tot (és de color negre absolut). En canvi, un valor de 1 indica que el material reflecteix tota la llum (és de color blanc perfectament reflectant).

## **2. OBJECTIUS DE L'ESTUDI**

### **2.1. Estimació precisa de nutrients en cítrics mitjançant l'espectroscòpia**

Les dades espectrals s'utilitzen àmpliament en diversos camps, des de la recerca farmacèutica i mèdica fins a la ciència dels materials i l'agricultura. L'espectroscòpia VIS-NIR és una tècnica fàcil, ràpida, no destructiva i econòmica [10]; permet realitzar evaluacions de qualitat amb poca o cap preparació prèvia de la mostra. Aquesta tècnica té aplicacions en l'anàlisi de qualitat de fruites i verdures en els àmbits industrial i semiindustrial de l'agricultura [11].

Això significa que amb les dades espectrals es poden realitzar moltes tasques que solien requerir estudis o investigacions costoses en termes de temps i recursos econòmics, tot aconseguint resultats molt bons a un cost realment baix. Això ha provocat un augment de l'ús de l'espectroscòpia en els últims anys.

Dins de l'estudi del sector agrícola i alimentari, que és el camp d'inclusió d'aquest projecte, hi ha múltiples aplicacions de l'espectroscòpia. Algunes d'elles són:

- Anàlisi multielemental dels sòls.
- Detecció de malalties o danys en els aliments (control de qualitat).
- Monitoratge de contaminants orgànics.

En aquest cas concret, el projecte se centra en la creació de models de regressió d'aprenentatge automàtic capaços d'estimar els nivells de nutrients d'un arbre cítric a partir de l'espectre generat per les seves fulles. Per a això, es disposa d'una col·lecció de mostres de fulles, de les quals s'ha calculat l'espectre en un rang de longituds d'ona i també s'ha realitzat un anàlisi destructiva per determinar els nivells de diversos nutrients.

Com s'ha comentat anteriorment, l'avantatge d'utilitzar un sistema que utilitza informació espectral per estimar els nivells de nutrients en comparació amb un anàlisi destructiva és que es redueixen les despeses. Aquesta reducció de costos permet que els agricultors que no poden permetre's una anàlisi destructiva dels seus arbres per raons econòmiques puguin accedir a sistemes que ajuden a evitar l'ús excessiu de fertilitzants i a mantenir els seus arbres en un bon estat de salut.

## **2.2. Un enfocament de regressió supervisada**

L'objectiu del projecte serà avaluar el funcionament de diferents models d'aprenentatge automàtic per tal de veure quins tenen un millor funcionament. Hi ha diferents tipus de models d'aprenentatge automàtic, que es poden classificar a grans trets en tres categories: aprenentatge supervisat, aprenentatge no supervisat i aprenentatge per reforç.

L'aprenentatge supervisat és aquell en què l'ordinador rep un conjunt de dades d'entrenament, que inclou les respostes correctes, i s'utilitza l'algorisme per aprendre d'aquestes dades per tal de ser capaç de generalitzar i fer prediccions sobre noves dades. L'aprenentatge no supervisat és aquell on l'ordinador rep dades, però no les respostes correctes, i ha d'aprendre d'aquestes dades per si mateix per trobar patrons i relacions. L'aprenentatge per reforç és aquell en què l'ordinador rep un objectiu que ha d'assolir i ha d'aprendre per assaig i error per esbrinar-ne la millor manera.

Tornant a l'objectiu del projecte, la idea és predir el nivell nutricional en base al espectre generat, per tant, serà un problema a resoldre mitjançant aprenentatge supervisat, tot i que, també s'utilitzen tècniques d'aprenentatge no supervisat amb altres finalitats.

La regressió i la classificació són tots dos tipus de problemes d'aprenentatge supervisat. La regressió és un tipus de problema d'aprenentatge supervisat on l'objectiu és predir un valor continu. Per exemple, podeu utilitzar la regressió per predir el preu d'una casa en funció de la seva mida, edat i ubicació. La classificació és un tipus de problema d'aprenentatge supervisat on l'objectiu és predir una etiqueta de classe. Per exemple, podeu utilitzar la classificació per predir si un pacient té càncer en funció del seu historial mèdic.

Per tant, el nivell nutricional, com s'ha comentat anteriorment, és una dada numèrica continua. Sabent açò, definim el problema com un problema de regressió.

## 3. MATERIALS I MÈTODES

### 3.1. Creació del conjunt de dades

Per a crear un conjunt de dades per a la resolució del problema, es requereixen dues parts: el càlcul real dels nutrients de les fulles i les dades espectrals. A continuació, s'explica com s'han obtingut aquestes dades.

- Càlcul real dels nutrients de les fulles: Per a determinar els nivells de nutrients en les fulles de cítrics, s'ha realitzat una ànalisi destructiva. Això implica prendre mostres de fulles i sotmetre-les a procediments d'ànalisi química en laboratori per a mesurar les concentracions de diferents nutrients. Aquest procés permet obtenir dades precises i confiables sobre els nivells de nutrients presents en les fulles.
- Dades espectrals: Per a obtenir les dades espectrals, s'han utilitzat espectòmetres de reflectància. Aquests dispositius emeten llum en diferents longituds d'ona i mesuren la quantitat de llum reflectida per les fulles en cada longitud d'ona. Això genera un espectre que mostra els nivells de reflectància en diferents intervals de longitud d'ona. Mitjançant aquest procés, s'han recopilat les dades espectrals per a les fulles de cítrics utilitzades en l'estudi.

Amb aquestes dues parts de dades, és possible crear el conjunt de dades necessari per a entrenar i avaluar els models d'aprenentatge automàtic per a la predicció dels nivells de nutrients basant-se en les dades espectrals. Aquest conjunt de dades combina les mesures reals dels nutrients amb les dades espectrals corresponents, proporcionant una base per al desenvolupament i evaluació dels models.

#### 3.1.1. Desenvolupament Experimental

Per a dur a terme el projecte, s'ha realitzat un assaig en una parcel·la comercial de cítrics a Almenara (Castelló), la localitat de la qual es troba ubicada a  $39^{\circ} 45' 12''N$   $0^{\circ} 13' 32''E$ . La varietat estudiada és *Clemenules* (*Citrus clementina*, *Hort ex Tan*) uns arbres que produïxen un cítric híbrid entre la mandarina i la taronja amarga. La parcel·la compta amb 210 arbres i l'assaig es va efectuar el dia 4 de maig de 2021.

Es van prendre mostres d'un total de 33 arbres, de cada arbre es van agafar 6 mostres de 30 fulles cadascuna: fulles joves (brotació de primavera) i fulles velles (brotació cicles vegetatius anteriors) de l'orientació est, oest i zenital,

mostrejant un total de 5940 fulles. Totes les fulles joves/velles de cada orientació pertanyien a la mateixa branca de l'arbre.

Llavors, de cada arbre tenim un conjunt de 30 fulles joves i 30 fulles velles de cada orientació. De cadascun d'ells, dividim la mostra en 22 fulls per calcular l'anàlisi destructiva i, per tant, saber els nutrients del grup. D'aquesta manera, es coneixen els nutrients reals de les fulles.

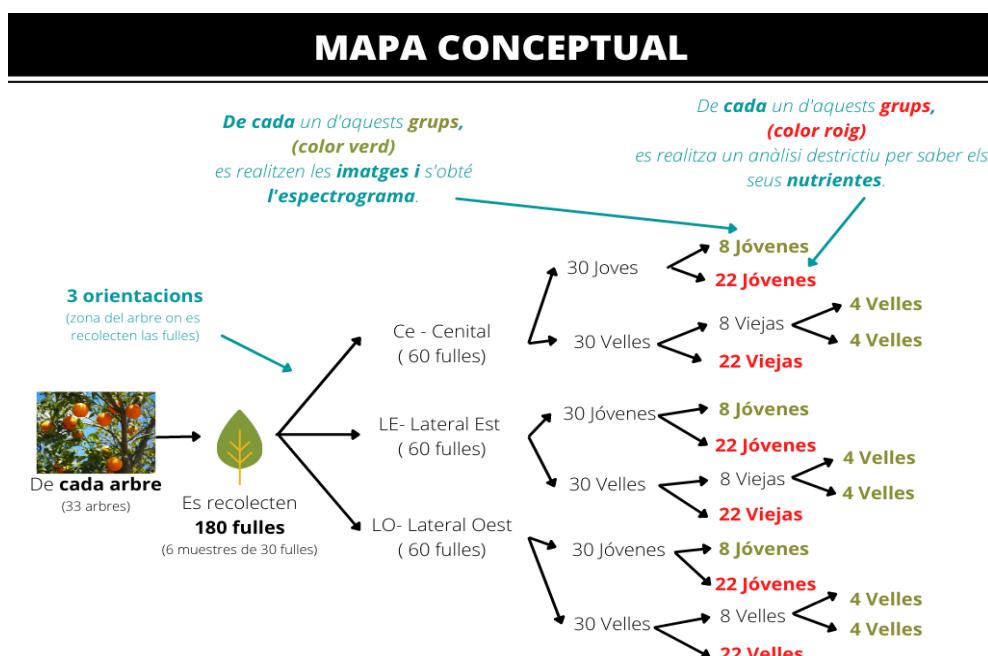
Després, les 8 restants de la mostra, s'utilitzen per calcular l'espectre. En el cas dels fulles velles es realitzen dues particions de 4 fulles. Això es fa a causa que les fulles velles són grans, i no hi caben totes dins l'angle de la lent de la càmera.

Llavors, de cada orientació es calcula l'espectre de 8 fulles joves, d'una banda, i, d'altra banda, es calcula l'espectre de dues divisions de 4 fulles velles. Per tant, de cada arbre i cada orientació s'obtenen 3 espectres: un de fulles joves, un de la primera partició de 4 fulles velles, i un altre de la segona partició d'altres 4 fulles velles. Les dues divisions de fulles velles tindran espectres diferents, però els valors dels nutrients seran els mateixos, ja que s'ha elaborat l'anàlisi destructiva de l'altra part de la mostra.

Finalment, tenim com a resultat, de cada arbre 9 espectres (representant a i b cada participació de fulles velles):

LEJ, LEV, LEV<sub>b</sub>, LOJ, LOV<sub>a</sub>, LOV<sub>b</sub>, CeJ, CeV<sub>a</sub> y CeV<sub>b</sub>.

Es pot veure en la següent Figura 3.1. un mapa conceptual per poder entendre quin ha sigut el procés d'adquisició de les dades:



**Figura 3.1.** Esquema del procés d'adquisició de mostres per a crear el conjunt de dades. Procés d'adquisició d'imatges hiperrespectrals i anàlisi destructiu. Font: Elaboració pròpia.

La qual cosa finalment genera un arxiu en el qual es pot diferenciar les parts següents:

- **Informació General:** Tenim una sèrie de variables que contribueixen a conéixer d'on ve el valor espectral i els nutrients.
  - **Nom:** Junta el nom de l'arbre, orientació, zona, a/b i mostra.
  - **Mostra:** Mitjançant un identificador (des d'1 fins a 198) tenim identificat cada grup de fulles dels quals es calculen els nutrients.
  - **Arbre:** Identifica l'arbre del qual provenen les fulles.
  - **Zona:** Identifica l'orientació i si són joves o vells els fulls de forma conjunta (LEJ, LEV, LOJ, LOV, CeJ i CeV).
  - **Orientació:** Representa l'orientació (LE, LO, Ce)
  - **J/V:** Identifica si les fulles són joves (J) o velles (V).
  - **a/b:** Representa la partició de les fulles amb què es calcula l'espectre. Les Joves sempre seran del grup a (ja que no es realitza cap partició), i les velles poden ser o a/b (depenent de la partició).
  -

La informació general d'un arbre, per exemple l'arbre identificat per I02, es pot apreciar a la Taula 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_HCevVa_006	006	I02	CeV	Ce	V	a
AI02_HCevb_006	006	I02	CeV	Ce	V	b

**Taula 3.1.** Exemple de la informació general d'algunes mostres del arbre 'I02'. Font: Elaboració pròpia.

Seguidament, per a cada mostra, també tenim:

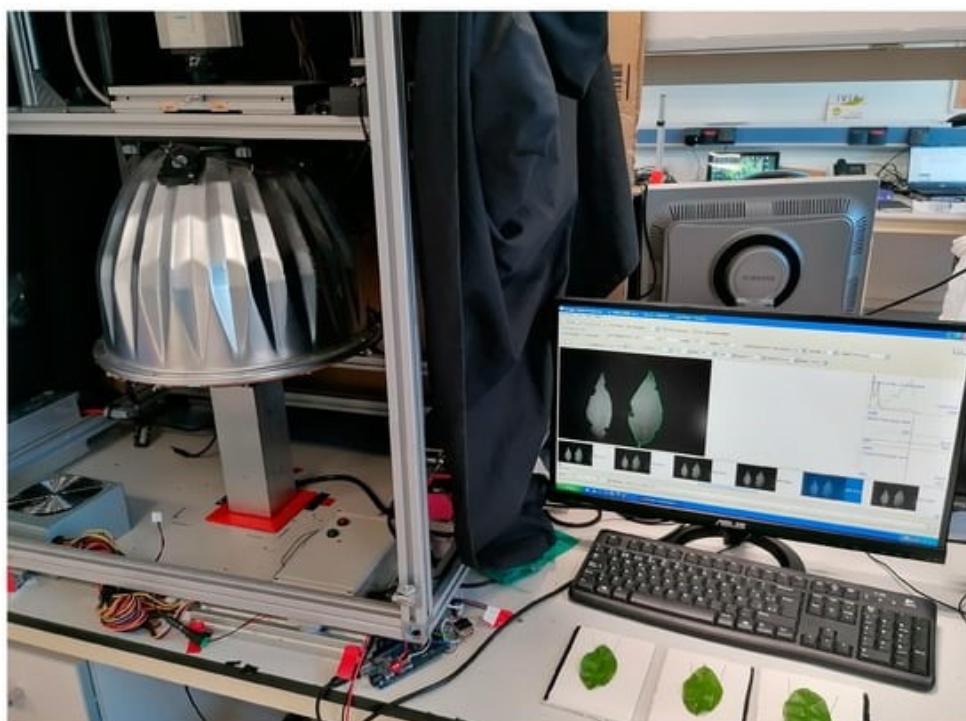
- **Nutrients:** Els nutrients (N, P, K, Ca, Mg, Na, S, Fe, Mn, Zn, Cu, B i Mo) de cada mostra obtinguts mitjançant l'anàlisi destructiva.
- **Espectre:** De cada 'nom', és a dir, de cada partició de cada mostra, tenim tantes variables com longituds d'ona espectrals.

#### 3.1.2. Adquisició de les Imatges Hiperespectrals

Al laboratori de Visió i Espectroscòpia del Centre d'Enginyeria Agrícola de l'IVIA (Institut Valencià d'Investigacions Agràries), es van utilitzar sistemes d'imatges hiperespectrals al rang Vis-NIR per adquirir imatges dels fulls de cítrics utilitzats al treball. El sistema consisteix en una càmera industrial i dos filtres sintonitzables de cristall líquid LCTF (Liquid Crystal Tunable Filter), capaç d'adquirir imatges de 1392 x 1040 píxels amb una resolució espacial de 0,14 mm/pixel en 65 bandes diferents amb una resolució espectral de 10 nm, al rang espectral de 400 nm a 1050 nm.

Es va utilitzar un sistema basat en il·luminació halògena difusa eficient a tot el rang espectral de treball. Per evitar problemes d'imatges borroses a causa de la dispersió cromàtica de la lent, se'n va establir l'enfocament a la longitud d'ona central (730 nm) del rang de treball. Es va adquirir una imatge hiperspectral d'un blanc de referència certificada que cobria el 90% del rang dinàmic de la càmera per evitar la saturació de la imatge i corregir la sensibilitat espectral del sistema.

Aquest entorn on es van aconseguir les imatges hiperespectrals es pot apreciar a la següent Figura 3.2.



**Figura 3.2.** Procés d'adquisició de la imatge hiperespectral d'una fulla de cítric.

### 3.1.3. Adquisició de l'Espectre des de les Imatges Hiperespectrals

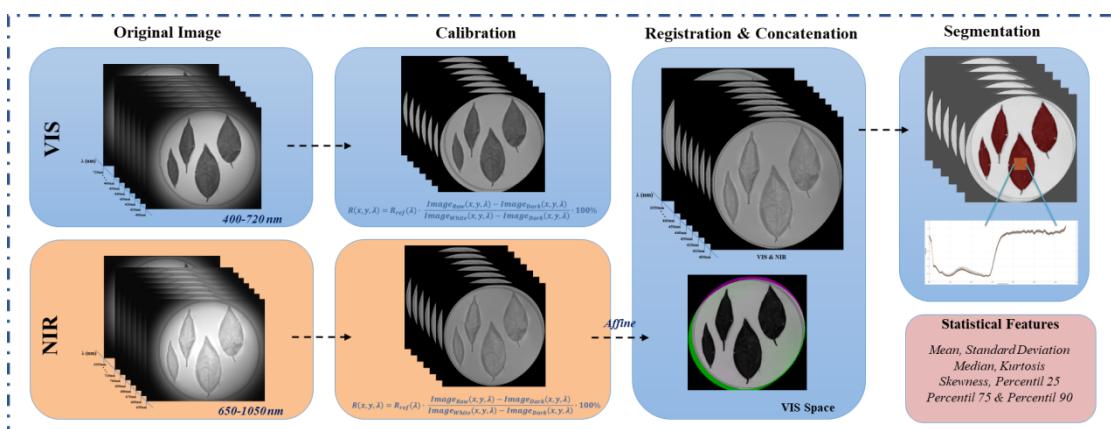
L'espectre de cada fulla i les seues propietats es van extreure mitjançant un algorisme desenvolupat a MATLAB que va processar digitalment les imatges hiperespectrals adquirides individualment. Aquest processament digital consta de tres blocs o etapes diferents.

En un primer pas, es van calibrar imatges hiperespectrals en brut enregistrades en diferents rangs de longitud d'ona, el rang visible (400 a 720 nm) i el rang d'infraroig proper (650 a 1050 nm), aplicant la següent equació (*Equació 3.1*):

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

*Equació 3.1. Equació de Calibració de es imatges:*

La *Image White* és la imatge adquirida del blanc de referència blanc abans esmentat i la *Image Dark* és la imatge adquirida de la referència de negre. Es comprova de forma gràfica el procés a la Figura 3.3.

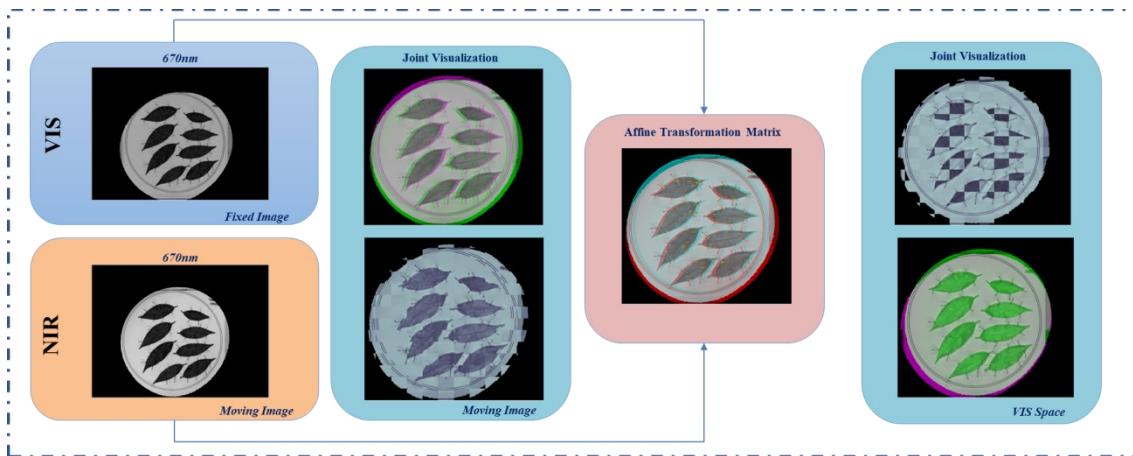


*Figura 3.3. Preprocesament a les imatges hiperespectrals adquirides.*

En la segona etapa, que es veu a la Figura 3.4, s'ha de corregir el desplaçament espacial existent entre la imatge NIR i la imatge VIS a causa de la configuració del mateix sistema d'enregistrament. Per corregir-ho, es va crear un registre en el qual es calcula una transformació afí i s'aplica a imatges de la regió NIR per tal de poder alinear-les amb imatges de la regió VIS.

Aquesta transformació consisteix a mantenir la imatge del rang VIS mentre es llisca la imatge NIR per sobre, i utilitzar una matriu de transformació afí adequada per obtenir una visualització conjunta.

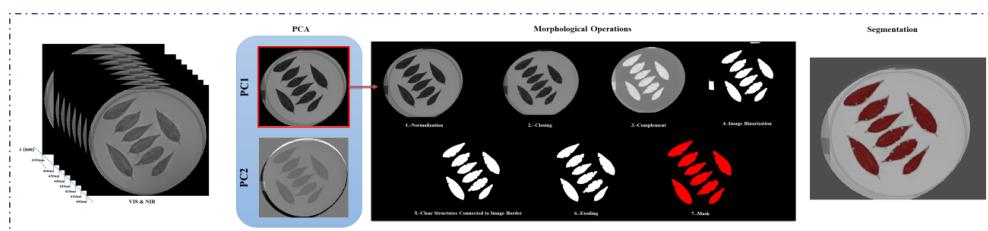
El que s'està fent és augmentar la coherència espacial entre les imatges de les dues regions capturades, permetent-nos utilitzar un únic hipercub en comptes de dos.



**Figura 3.4.** Correcció del desplaçament i concatenació.

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.

Aquest algoritme calculava el PCA de tot l'hipercub i seleccionava la primera component per ser la que més contrast mostrava d'entre els diferents components de la imatge. Després mitjançant diferents operacions morfològiques d'imatge es van eliminar les estructures coincidents amb les vores de la imatge i es va realitzar una màscara binària, eliminant els elements que no es busca com pot ser el fons. D'aquesta manera obtenim les zones d'especial interès, és a dir, la superfície foliar.



**Figura 3.5.** PCA i operacions morfològiques de la imatge.

Aquest procés el trobem detallat a [12].

### 3.2. Anàlisi exploratòria de les dades

En la ciència de dades sempre és important conéixer les dades; saber les seues característiques principals ens ajuda a entendre el seu comportament i poder formular preguntes o hipòtesis que en passos posteriors ens seran d'utilitat per a resoldre d'una millor manera el problema. La següent cita de l'estadístic Jhon W. Tukey pot definir a la perfecció el motiu de la importància de elaborar aquesta anàlisi:

*"És millor tenir una resposta aproximada a la pregunta correcta que una resposta exacta a la pregunta equivocada"*

*~ John W. Tukey*

Com s'ha comentat anteriorment, tenim tres parts en les dades: la informació general, els nutrients i l'espectre. Aleshores, aquest apartat es divideix en dos subapartats, en el primer, s'analitza les dades nutricionals, i en el segon, les dades espectrals.

#### 3.2.1. Anàlisi dels nutrients

- Anàlisis nutrients en funció del tipus de full (J/V)**

Una vegada s'han estandarditzat les dades nutricionals, es pot procedir a la comparació de la variabilitat inherent a cada nutrient, en funció de si la fulla es jove o vella. En la Figura 3.6, es presenten les diferències estandarditzades en la concentració de nutrients entre fulles joves i velles.

La desviació estàndard (*std*) s'utilitza per quantificar la variabilitat o dispersió d'un conjunt de valors nutricionals. Per a cada nutrient, es calcula la desviació estàndard de les concentracions estandarditzades. Aquest càlcul proporciona una mesura de la quantitat en què les concentracions d'un nutrient específic difereixen o varien en la població de fulles, ja siguin joves o velles. A l'equació 2.0 s'observa com es fa aquest càlcul.

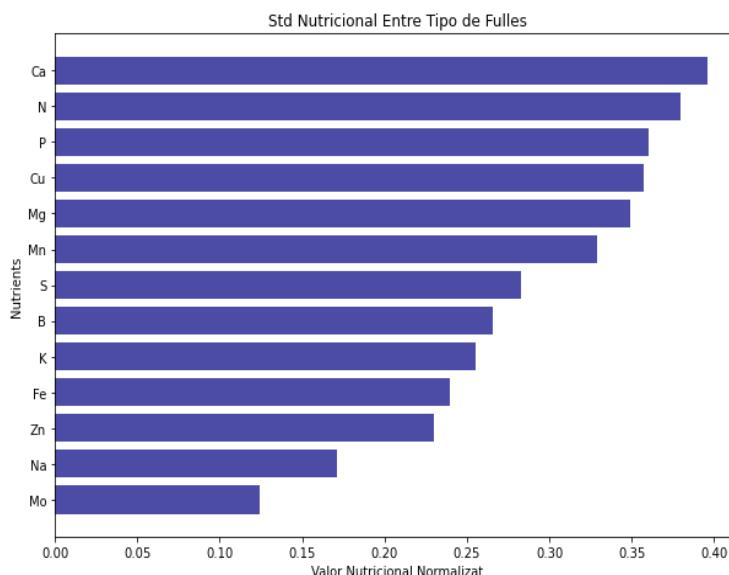
$$\sigma = \sqrt{\sum (xi - \mu)^2 / N}$$

*Equació 2.0. Desviació estàndard*

- $xi$  és cada valor de la variable (en aquest cas, la concentració estandarditzada d'un nutrient específic),
- $\mu$  és la mitjana de tots els valors de la variable,
- $\Sigma$  denota la sumatòria de tots els valors,
- $N$  és el número total de valors (és a dir, el número total de mesuraments de la concentració del nutrient).

Per exemple, el calci mostra una variabilitat significativa entre les fulles joves i velles, amb una desviació estàndard del 39,5% [13]. Això significa que, de mitjana, les concentracions de calci en una fulla específica difereixen en un 39,5% de la mitjana de les concentracions de calci per a totes les fulles.

La Figura 3.6 il·lustra visualment aquestes desviacions estàndard per a cada nutrient, proporcionant una representació intuïtiva de la variabilitat de les concentracions de nutrients en les fulles joves i velles.



**Figura 3.6.** Diferència estandarditzada entre fulles Joves i Velles segons el nutrient.

Aquesta disparitat en els nivells de calci pot explicar-se per diverses raons. En primer lloc, el calci és un nutrient essencial per al creixement i desenvolupament de les plantes. A mesura que les plantes creixen, absorbeixen calci del sòl i el emmagatzemen a les fulles, la qual cosa pot resultar en una major concentració en les fulles velles.

A més, el calci és un nutrient immòbil dins de les plantes, el que significa que una vegada és absorbit per les arrels i transportat a les fulles, no es mou significativament a través de la planta. Això fa que la concentració de calci continui augmentant a les fulles a mesura que les arrels absorbeixen més calci i el transporten cap a les fulles.

Pel que fa al nitrogen, s'observa que les fulles joves tenen una concentració més alta que les fulles velles. Això es deu al fet que les fulles joves estan en creixement actiu i requereixen un major subministrament de nitrogen per al seu desenvolupament. A mesura que les fulles envelleixen, poden perdre nitrogen a través de processos com la respiració i la transpiració.

En el cas del potassi, es troba que les fulles joves solen tenir una concentració més gran que les fulles velles. Això és degut al paper essencial que el potassi

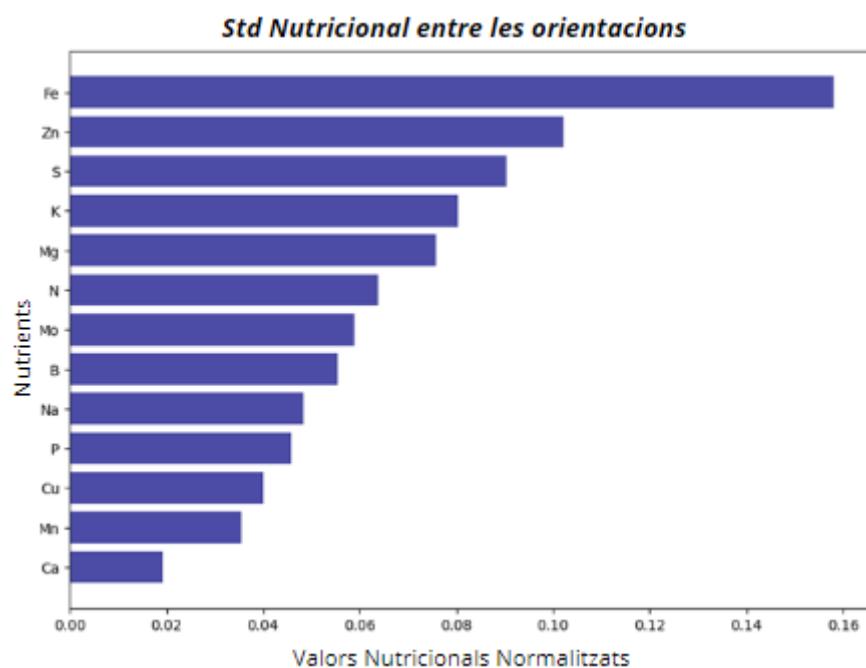
juga en el creixement i desenvolupament de les plantes. A mesura que les fulles envelleixen, poden perdre potassi a través de diversos processos.

Amb altres nutrients, també es poden observar diferències segons el tipus de fulla, i aquestes diferències poden ser explicades per diverses raons específiques.

- **Anàlisi nutrient en funció de l'orientació**

Ara podem veure el mateix que en el cas anterior, en el cas de les diferents orientacions. S'observa que les orientacions tenen una influència molt menor en els nutrients segons el tipus de fulla, tal com es mostra a la Figura 3.7. Es pot apreciar que les fulles joves i velles tenen un impacte significatiu en els nivells de calci en comparació amb els altres nutrients, mentre que les orientacions no presenten una diferència destacable. No obstant això, es pot observar una diferència entre els nivells de ferro (Fe) i els altres nutrients.

Segons la posició de la fulla en l'arbre, la llum solar pot influir en la producció de clorofil·la, la qual contribueix a l'absorció d'alguns nutrients com el ferro. No obstant això, en el cas del calci, com s'ha mencionat anteriorment, aquest nutrient s'absorbeix principalment del sòl, de manera que la llum solar no té una influència directa en la concentració d'aquest nutrient a les fulles.



**Figura 3.7.** Diferència estandarditzada entre les orientacions segons el nutrient.

- **Taula Desviació Típica segons tipus de full i orientació**

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

**Taula 3.2.** Diferència estandarditzada entre les orientacions segons el nutrient.

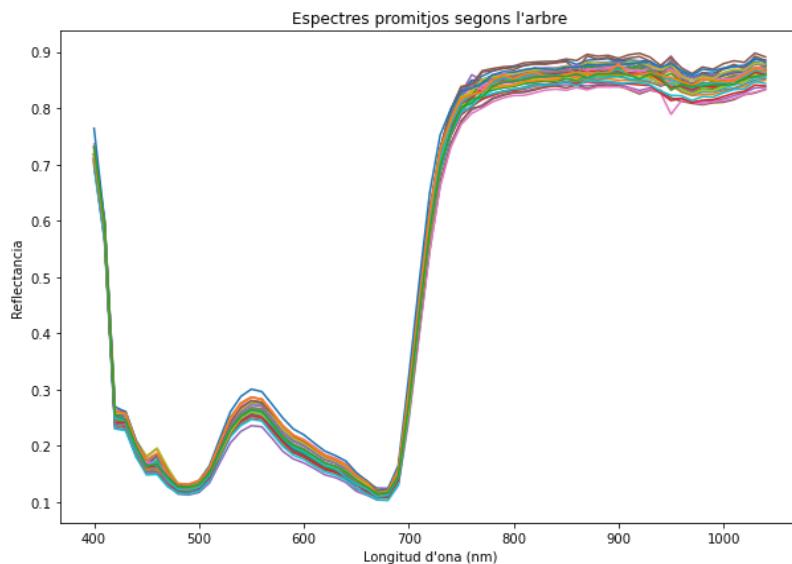
En la Taula 3.2 es mostra el percentatge de variacions comentades anteriorment. Es pot observar que, en general, el percentatge de variació segons l'orientació de la fulla és significativament major en comparació amb el tipus de fulla.

### 3.2.2. Anàlisi dels espectres:

#### - Espectre Promig de cada arbre

En aquesta secció s'explora l'espectre promig de cada arbre. S'ha recollit un total de 180 fulles de diferents orientacions i edats de cada un dels 33 arbres. A la Figura 3.8, es mostra la mitjana de tots els espectres d'aquests arbres, on cada línia representa l'espectre promig d'un arbre.

En l'espectre promig, s'observa una reflectància elevada en la part violeta-blava de l'espectre, seguida d'una reflectància més baixa en el color verd. Després, es torna a observar una major reflectància a la part taronja i vermella de l'espectre. Aquesta reflectància està relacionada amb el color verd de les fulles dels cítrics, el qual és causat per la presència de clorofil·la. La clorofil·la té la capacitat d'absorir la llum en les parts vermella i blava de l'espectre, mentre que el color verd és el que més es reflecteix.



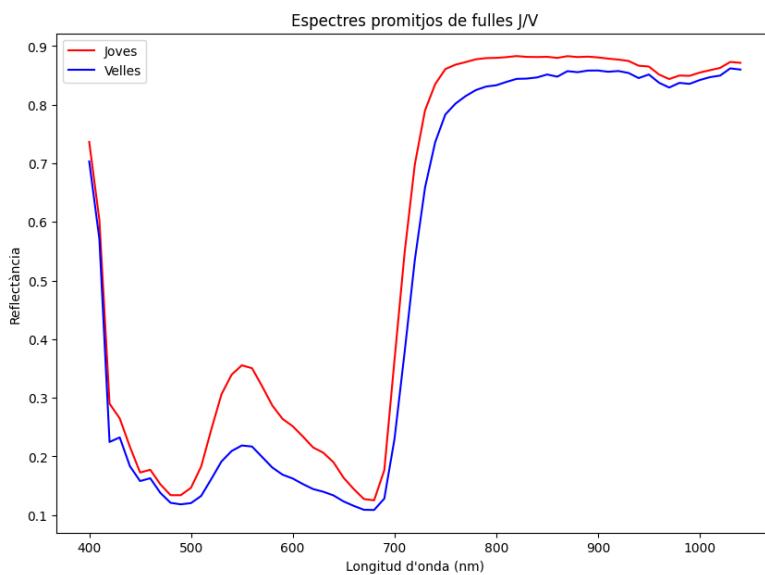
**Figura 3.8.** Espectre mitjà de les mostres registrades de cada arbre

- **Espectre Segons el tipus de fulla**

A la Figura 3.9, s'aprecia clarament una diferència entre la mitjana de l'espectre de les fulles joves i la mitjana de les fulles velles. Fins a les primeres bandes espectrals (primera part del color violeta-blau de l'espectre visible), les mitjanes de les fulles es mantenen semblants. No obstant això, a partir del color verd, es comença a observar una diferència significativa.

La diferència més destacada entre les mitjanes dels espectres es troba en les bandes espectrals de 490 a 680 nm, essent el valor de 550 nm on s'aprecia la màxima diferència. Aquesta diferència es deu al fet que les fulles joves tenen una concentració més alta de clorofil·la que les fulles velles, la qual cosa els confereix un color verd més intens.

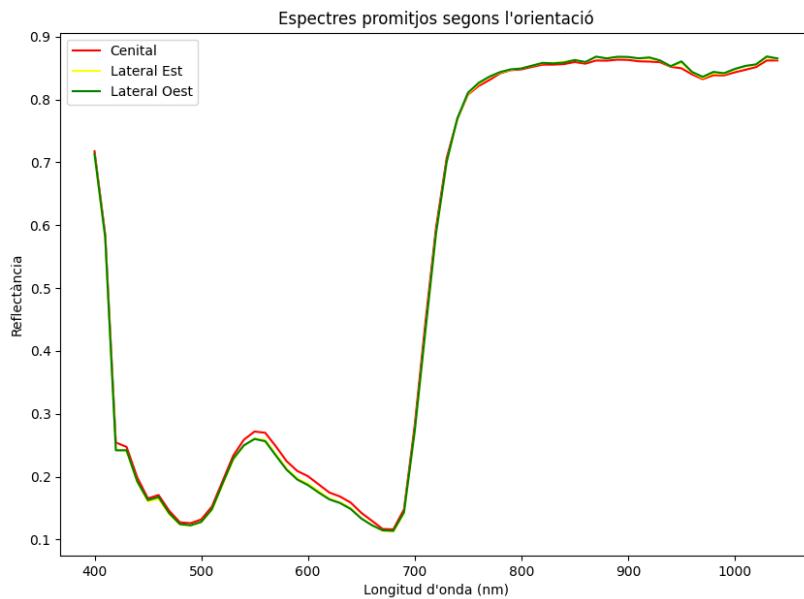
En resum, les diferències en els espectres de les fulles joves i velles es deuen a les variacions en les quantitats de clorofil·la present en cada tipus de fulla.



**Figura 3.9.** Espectre mitjà de les fulles joves i velles registrades

- **Espectre Segons l'Orientació**

Es pot apreciar a la Figura 3.10 que no existeixen diferències evidents entre els espectres mitjans segons les diferents orientacions.



**Figura 3.10.** Espectre mitjà segons l'orientació en la que es va registrar la mostra

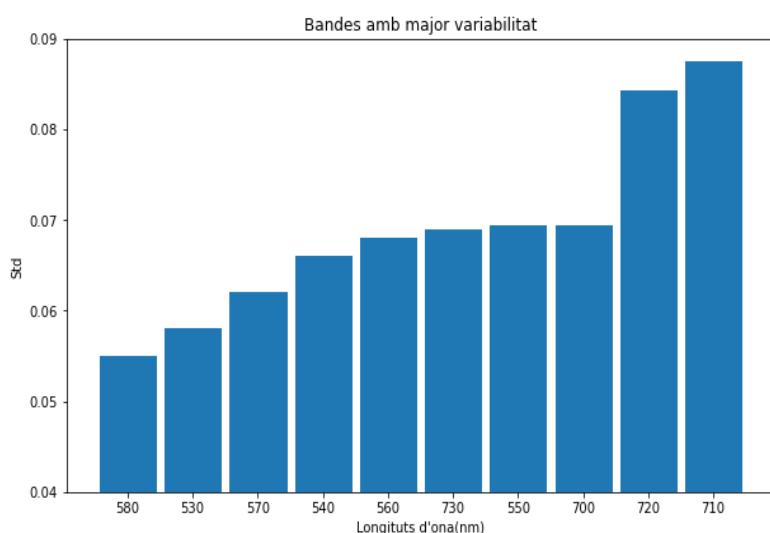
- **Variabilitat Bandes Espectrals**

A la Figura 3.11, es confirma el que s'ha mencionat anteriorment: les bandes espectrals entre 700 i 730 nm es troben entre les 10 bandes amb major variabilitat. Aquesta part de l'espectre correspon a la regió de color vermell.

Aquesta alta variabilitat en les bandes de color vermell pot ser atribuïda a diferències en la concentració de pigments associats al color vermell, com ara carotenoides i antocianines.

D'altra banda, es destaca la presència de les bandes espectrals entre 530 i 580 nm, que corresponen als colors verds de l'espectre visible. Aquesta part de l'espectre també mostra una certa variabilitat, encara que en menor mesura que les bandes del color vermell. Aquesta variabilitat en les bandes verdes pot reflectir diferències en la concentració de clorofil·la, el pigment principal responsable del color verd de les fulles.

Aquesta informació proporciona una comprensió més detallada de les bandes espectrals que presenten major variabilitat en els espectres de les fulles, particularment en les regions del vermell i del verd..



**Figura 3.11.** Bandes espectrals que registren major variabilitat

### 3.3. Neteja, divisió i *oversampling* de les dades:

#### 3.3.1. Eliminar dades faltants

Primerament, s'eliminen aquelles dades que no tenen un valor. En el conjunt de dades, podem identificar dos casos específics. En el primer cas, hi ha instàncies en les quals no disposem de cap valor per a la concentració de cap nutrient. Aquestes dades no són útils per al procés de modelització, ja que no aporten informació per a la variable resposta. Per tant, aquestes dades són eliminades completament de l'estudi.

En segon lloc, hi ha un grup de dades en què només falta el valor per al nutrient del nitrogen. En aquest cas, aquestes dades seran utilitzades en el procés de modelització per a tots els altres nutrients excepte el nitrogen. Aquestes dades disponibles serviran per a construir models predictius per als altres nutrients basant-nos en la seva informació disponible.

Aquest procés de selecció de dades ens permetrà treballar amb un conjunt de dades més complet i alhora assegurar que les dades que utilitzem són adequades per a les nostres finalitats de modelització.

#### 3.3.2. Conjunt Train/Validació/Test

Primer que res, es divideix el conjunt amb variables explicatives (X), que seran les bandes espectrals, i variables predictores (y), que seran els diferents nutrients (N, K, Ca, Fe, etc...).

Seguidament, s'ha de dividir el nostre conjunt de dades en dues parts: una part de calibració (també anomenat entrenament) i altra de prova (*test*).

- **Conjunt de calibració (*train*):** Són les dades que es fan servir per entrenar el model.
- **Conjunt de prova (*test*):** Són les dades que es reserven per comprovar si el model que s'ha generat a partir de les dades d'entrenament funciona, i avaluar com de bé funciona mitjançant mètriques.

Per a la divisió del conjunt de dades, es realitza una partició aleatòria, assignant aproximadament el 75% de les dades al conjunt de calibració i el 25% restant al conjunt de prova. Aquesta divisió permet garantir que el model sigui entrenat adequadament i després avaluat amb dades que no ha vist durant l'entrenament.

- **Problemes**

Un problema que sorgeix sempre en l'entrenament d'un model d'aprenentatge automàtic és el sobreajust. El sobreajust és un problema que es pot produir quan els models d'aprenentatge automàtic són massa complexos per a les dades amb les quals s'han entrenat. Això pot condir a un rendiment pobre de generalització en dades noves, ja que el model ha après patrons que són específics de les dades d'entrenament i no són generalitzables a altres dades.

Hi ha diverses maneres de detectar l'excés d'adaptació. Una és observar el rendiment del model en les dades d'entrenament i veure si és molt millor que el rendiment en les dades de prova. Si el model està sobreajustat, el rendiment de l'entrenament serà molt més gran que el rendiment de la prova. Per a obtenir més informació sobre aquest problema, es pot consultar la referència [14].

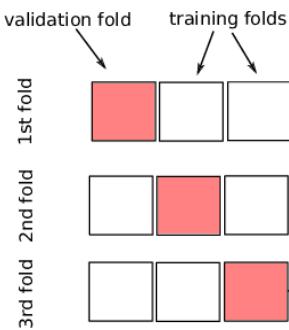
És important tindre en compte el sobreajust en l'entrenament de models d'aprenentatge automàtic, ja que pot afectar negativament la capacitat del model per a fer prediccions precises en noves dades. Per a mitigar aquest problema, es poden aplicar tècniques com la regularització, l'ús de conjunts de dades més grans, la selecció de característiques rellevants i la validació creuada, entre altres. Aquestes estratègies ajuden a controlar la complexitat del model i a millorar la seua capacitat de generalització.

- **Validació Creuada**

Validació creuada és la tècnica escollida per tal d'evitar el problema de sobreajust. En aquest projecte concret, s'utilitza *K-FOLD*, que és una tècnica en la qual les dades es divideixen en *k* particions, en el nostre cas 3 (ja que no hi ha una gran quantitat de dades). Dues particions s'utilitzen per entrenar el model, i una partió s'utilitza per realitzar la predicció (conjunt de validació), de manera que totes les particions són utilitzades pel model per predir una vegada.

Posteriorment, es pot calcular la mitjana de les mètriques d'avaluació del model aconseguit amb els conjunts de validació. D'aquesta manera, s'evita el sobreajust i també es pot trobar les tècniques de preprocessament que funcionen millor utilitzant més dades i, per tant, obtenir un funcionament més optimitzat en un context més general.

Els resultats obtinguts del conjunt de validació hauran estat estimats mitjançant la mitjana de la validació creuada. A la Figura 3.12 es pot observar un exemple de validació creuada amb 3 particions.



**Figura 3.12.** Exemple del procés *K-Fold* amb 3 particions

Es pot trobar aquest procés de *K-Fold Cross Validation* detallat a [15].

### 3.3.3. *Oversample*

L'*oversampling* és una tècnica utilitzada en aprenentatge automàtic per abordar el desequilibri de classes en un conjunt de dades. En aquest cas, com hem vist anteriorment, hi ha un major nombre d'espectres de fulles velles (ja que hi ha 2 grups de fulles velles per a cada grup de fulles joves). Per tant, augmentant el nombre d'exemples de la classe de fulles joves en el conjunt de dades, es pot obtenir una millora en la generalització dels models.

Aquesta tècnica és especialment útil quan es treballa amb un nombre reduït de mostres, ja que pot ajudar a millorar el rendiment dels models. Podeu trobar més informació sobre com aquesta tècnica pot ser beneficiosa en la referència [16].

### 3.4. Preprocessament espectral

El conjunt de tècniques utilitzades abans de l'aplicació d'un mètode de *Machine Learning* és conegut com a preprocessament de dades, i és considerat una de les fases més importants en el procés de descobriment de patrons a partir de les dades. El preprocessament de dades consisteix en l'aplicació d'un conjunt de tècniques per a preparar adequadament les dades que seran utilitzades com a entrada pels algoritmes de mineria de dades. Aquestes tècniques són sovint considerades com a obligatòries, ja que sense elles els algoritmes d'extracció de coneixement no podrien ser executats o, en altres casos, donarien resultats de baixa qualitat.

Les tècniques de preprocessament de dades utilitzades en aquest treball tenen com a objectiu millorar la interpretació de les variables disponibles pels models. Per exemple, prenem en consideració el pes d'una persona (en kg) i la seua alçada (en metres); podem observar que una persona pot tenir un pes de 80 kg i mesurar 1,70 m. En aquest cas, el valor del pes és considerablement més gran que l'alçada, el que podria afectar un model d'aprenentatge automàtic que dona un pes excessiu de pes a la variable del pes en comparació amb l'alçada.

En aquest treball, s'han provat diferents tècniques de preprocessament de dades per avaluar quina d'elles ofereix els millors resultats per a aquest problema específic. A continuació, explicarem en què consisteixen aquestes tècniques.

#### 3.4.1. Standard Normal Variate (SNV)

- **Introducció**

La primera tècnica que apliquem a les bandes espectrals és l'estandardització estàndard anomenada també *Standard Normal Variate (SNV)*. Una tècnica de la més comuna, per no dir la més comuna, a l'hora d'estandarditzar les dades. Aquesta tècnica el que fa és escalar les característiques perquè tinguen una mitja de 0, i una variància d'una unitat. D'aquesta forma es normalitzen les bandes espectrals estandarditzant-les per tal d'uniformitzar-ne el rang de variació.

- **Marc Teòric**

Suposarem un conjunt de dades  $X$ .

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

*Equació 3.1.1*

En l'equació 3.1.1  $n$  són el nombre d'observacions, i  $p$  el nombre de característiques; en aquest projecte, les bandes espectrals (un total de 66). Els passos seguits per a aconseguir escalar les dades són els següents:

En l'equació 3.1.2 es calcula la mitja de cada columna.

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

*Equació 3.1.2*

L'equació 3.1.2 genera un vector on es té la mitja de cada columna com el de l'equació 3.1.3.

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

*Equació 3.1.3*

Es fa el mateix per a calcular la variància:

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

*Equació 3.1.4*

L'equació 3.1.4, ens genera un altre vector on es té la variància de cada columna (equació 3.1.4):

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

*Equació 3.1.5*

Finalment, a cada valor de les dades ( $x_{p,i}$ ) de cada columna  $p$  se li restem la mitja de la seua columna i es divideix entre la variància de la columna per tal de tindre una mitja 0 i una variància d'una unitat tal i com s'aprecia a l'equació 3.1.6.

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

*Equació 3.1.6*

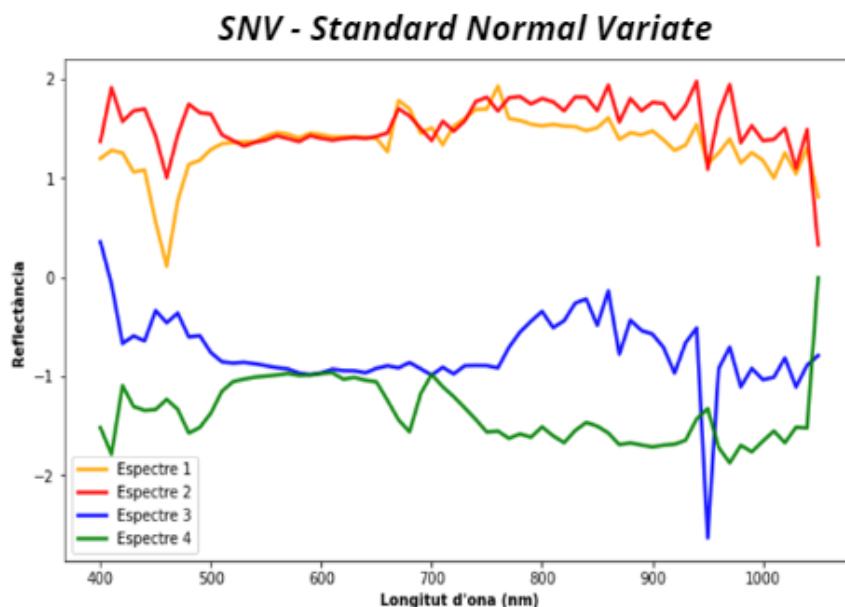
- **Calibració i prova**

Una cosa que s'ha de tenir en compte és no aplicar l'estandardització sobre tot el conjunt de dades. Com tenim un conjunt de calibració i un de prova, si a l'hora de calcular la mitja i la variància de les columnes, s'inclouen els

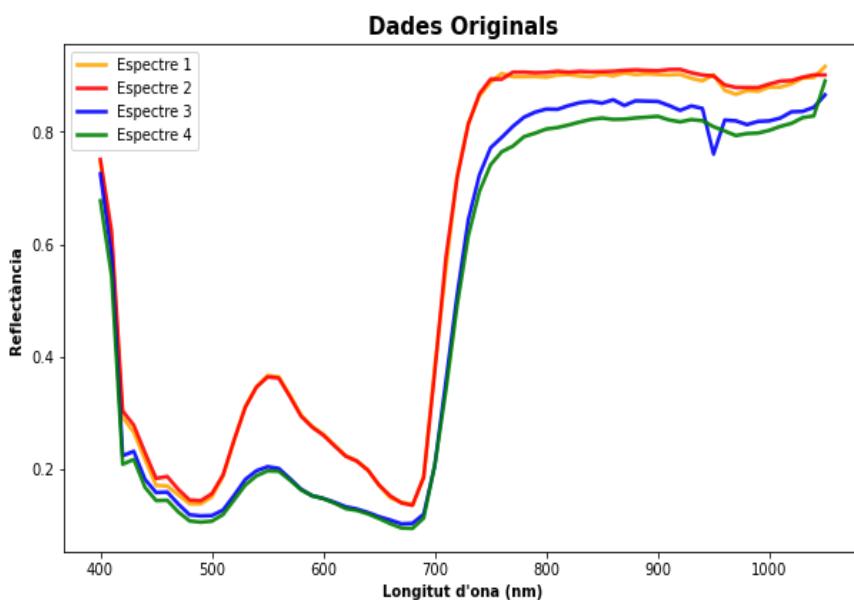
espectres que tenim a al conjunt de *test*, s'està utilitzant informació provenint d'aquest conjunt la qual cosa implica un biaix a l'hora d'avaluar el funcionament dels models.

Aleshores, el que es fa és calcular aquestes mitges i variàncies amb el conjunt de calibració, i amb els resultats obtinguts transformar, per separat, el conjunt de train i de test.

- **Representació gràfica**



**Figura 3.13.** Representació gràfica de l'estandardització de 4 mostres espectrals.



**Figura 3.14.** Representació gràfica de 4 mostres espectrals originals.

Es pot veure a la Figura 3.13 i a la Figura 3.14 una visualització de quatre espectres per tal de veure com es transformen. Es veu a la *Figura 3.13* com l'estandardització causa que la mitja de tots els espectres siga 0 i, la variància d'una unitat.

Finalment, es pot trobar més informació sobre SNV a la referència [17].

### 3.4.2. Savitzky-Golay

- ***Introducció***

Un filtre Savitzky-Golay és un tipus de filtre que s'utilitza per eliminar el soroll de les dades i suavitzar-les. Les ponderacions es calculen mitjançant un ajustament polinòmic de les dades. Aquest ajust es calcula a partir del grandària de la finestra, i del grau del polinomi. És un filtre que pot ser de molt d'interès ja que preserva les característiques de la distribució inicial com els màxims i mínims relatius, i els pics de l'espectre.

- ***Marc Teòric***

Donat un espectre que seria  $x[n] = [x_1, x_2, \dots, x_j]$  on  $j= 1, 2, \dots, 66$ .

Es busca anar suavitzant l'espectre, per tant, progressivament es calculen els nous valors donats els valors de les dades. El primer a definir és el grandària de la finestra ( $m$ ), que indica el nombre de punts adjacents que s'utilitzen per ajustar el polinomi.

Com ha d'haver-hi un punt central, del qual es vol calcular el seu nou valor, aquest valor  $m$  sempre ha de ser impari. Per exemple, si  $m = 3$ , s'utilitzen els dos punts adjacents al punt central per a determinar el nou valor. Aquests punts, sabent que el punt central representa el 0, serien el -1 i l'1. Sols es modifica aquest punt central, i aquesta finestra va avançant progressivament en els valors de tot l'espectre.

Aleshores  $n_l$  denota el nombre de punts a l'esquerra del punt central, és a dir, els negatius, i  $n_r$  denota el nombre de punts a la dreta del punt central. Sent així, el nombre de punts utilitzats el que es defineix a l'equació 3.2.1.

$$m = n_l + n_r + 1$$

*Equació 3.2.1*

Una altra condició que ha de complir és ser inferior, o igual, al nombre de valors que hi ha a l'espectre, és a dir,  $m \leq j$ .

En segon lloc, busquem aquella funció que, per als punts de la finestra, realitza un millor ajust mitjançant el mètode de mínims quadrats polinomial.

Aleshores, el segon paràmetre a decidir el seu valor és el grau del polinomi,  $k$ . Aquest, ha de ser menor que el grandària de la finestra, és a dir,  $n < m$ .

Una vegada escollits els valors de  $m$  i de  $n$  que es volen utilitzar, com hem dit, aquest filtre tracta d'encontrar la funció de mínims quadrats polinomial,  $p(x)$ , que millor s'ajusta als punts,  $x$  tal i com s'ha definit a l'equació 3.2.2.

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

Equació 3.2.2

Sent  $b_0, b_1, \dots, b_k$  els coeficients del polinomi definits a l'equació 3.2.3.

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

Equació 3.2.3

Aleshores, el punt central, és a dir  $x=0$ , l'obtindrem a partir de la funció polinòmica definida en l'equació 3.2.4.

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

Equació 3.2.4

Aleshores, el que volem és trobar aquells coeficients que minimitzen l'error de mínims quadrats de la funció  $p(x)$  anterior per als valors que tenim dins de la finestra. Aquesta funció de mínims quadrats està definida a l'equació 3.2.5.

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

Equació 3.2.5

El qual es trobada resolent la següent equació 3.2.6 que es la derivada parcial respecte a  $b_0$ .

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

Equació 3.2.6

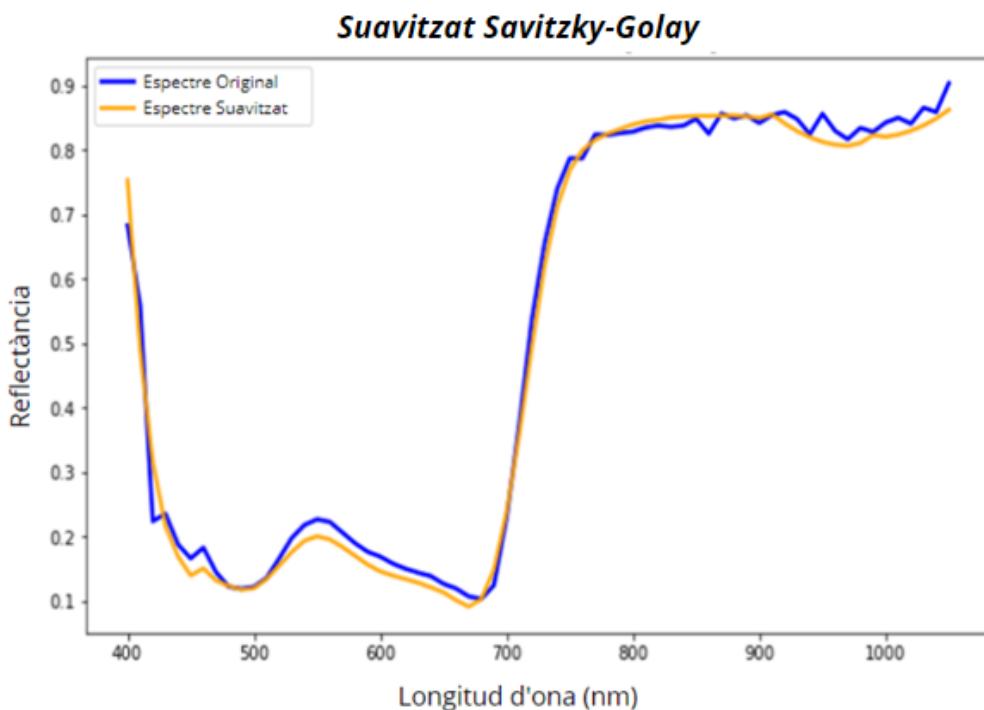
- **Calibració i prova**

En aquest cas, aquest filtre transforma cada espectre independentment dels següents, és a dir, a tots els espectres els aplica la mateixa transformació. En aquest cas, no s'ha de fer cap distinció entre el conjunt d'entrenament i prova sempre que els valors de  $m$  i  $n$  siguin els mateixos per als dos conjunts.

- **Selecció de valors**

Per a definir l'ordre del polinomi i la longitud de la finestra, primer que res, hem de tenir clara quina és la nostra finalitat. En aquest cas, el que es busca és encontrar un espectre suavitzat, és a dir, simplificar aquest espectre eliminant les variacions mantenint la distribució.

Veiem a la Figura 3.15 un espectre i el mateix espectre suavitzat. S'observa que en algunes bandes, sobretot les finals, l'espectre original (color blau) té xicotetes alteracions en els seus valors.

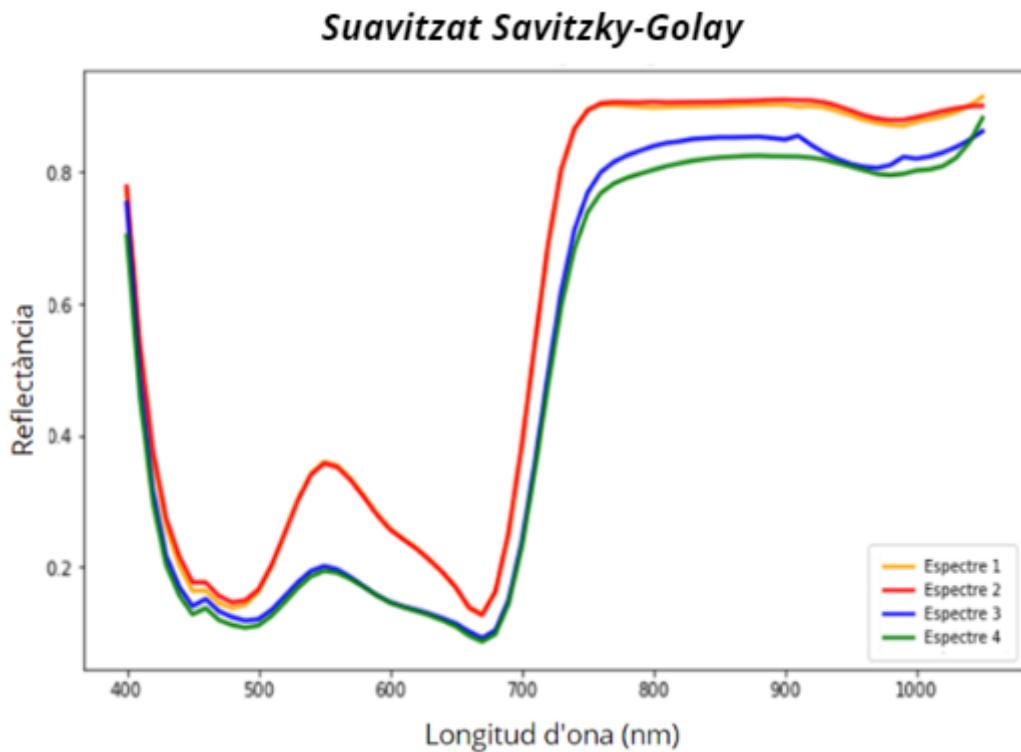


**Figura 3.15.** Representació gràfica de una mostra espectral transformada amb Savitzky-Golay i la mateixa mostra espectral original.

Aleshores, s'han realitzat diferents proves de selecció de valors, veient quins s'adaptaven en major mesura al que buscàvem. Finalment, s'ha decidit que el mida de la finestra siga de longitud  $n = 9$ , i l'ordre del polinomi siga de  $k = 3$ . Aquest criteri ha sigut definit de forma subjectiva valorant quin suavitzat, tant en espectres de fulles joves, com de fulles velles, mantenia la distribució original i eliminava els pics.

- **Representació gràfica**

A continuació, a la Figura 3.16 podem observar com aquests valors modifiquen el nostre espectre cap a un espectre més suavitzat en comparació al que trobem a la Figura 3.14 on tenim els espectres originals.



**Figura 3.16.** Representació de 4 mostres espectrals transformades amb Savitsky-Golay.

Es pot indagar més sobre Savitsky-Golay a la referencia [18].

### 3.4.3. Savitzky Golay Primera Derivada

- **Introducció**

El filtre de Savitzky Golay utilitzant derivades el que fa és augmentar l'ordre de la derivada a calcular. En el cas anterior, on no s'utilitza la derivada, aquest ordre és 0; en cas d'utilitzar la primera derivada, l'ordre de la derivada parcial és 1.

- **Marc Teòric**

Com ja sabem, que per a ajustar la funció polinòmica utilitzem l'equació 3.2.2. Aleshores, en aquest cas on s'utilitzen les derivades, la funció polinomi anterior és substituïda per la derivada de la seu funció segons l'ordre de la derivada tal i com es veu a l'equació 3.3.1.

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

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

.....

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

*Equació 3.3.1*

Aleshores, en cas d'utilitzar la primera derivada, el valor central serà definit per l'equació 3.3.2.

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

*Equació 3.3.2*

El que es fa, per tant, és trobar el mínim de l'error quadràtic entre les dades i la funció polinòmica com es veu en l'equació 3.3.3.:

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

*Equació 3.3.3*

Aquest mínim es troba en funció de la derivada de la funció polinomi, i del coeficient  $b_1$  com mostra l'equació 3.3.4.

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

*Equació 3.3.4*

- ***Calibració i prova***

En aquest cas, tampoc es fa cap distinció entre el conjunt de calibració i prova, sempre que els valors de  $m$  i  $n$  siguin els mateixos per als dos conjunts.

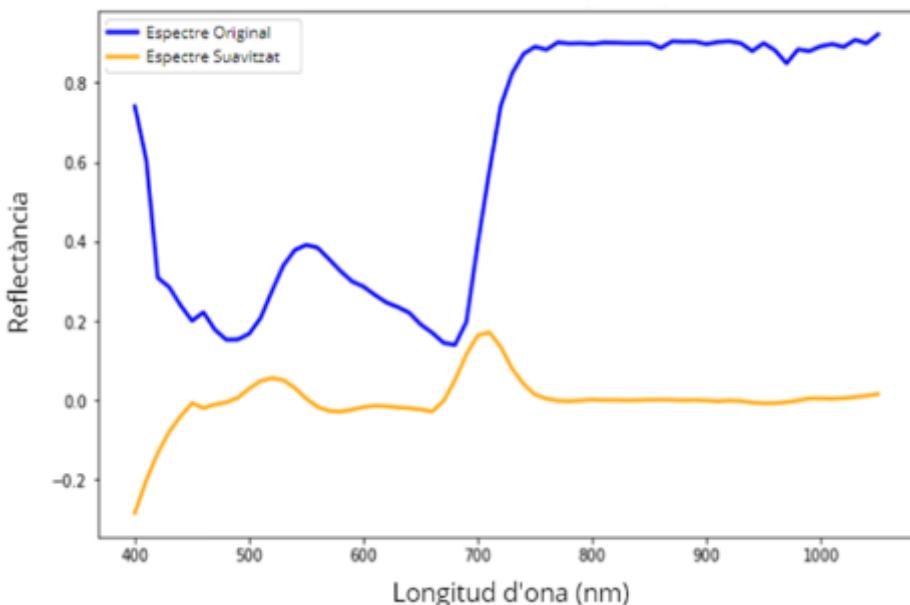
- ***Selecció de valors***

En el cas del filtre Savitzky Golay aplicant la primera derivada, utilitzem els mateixos valors que en el cas anterior, on aconseguien suavitzar l'espectre.

- ***Representació gràfica***

Es veu a la Figura 3.17 que, utilitzant la primera derivada, l'espectre és modificat per complet. A més, observant les bandes finals, es veu com el filtre elimina les xicotetes variacions que té l'espectre original.

### **Suavitzat Primera Derivada Savitzky-Golay**



**Figura 3.17.** Representació gràfica de una mostra espectral transformada amb la primera derivada de Savinsky-Golay i la mateixa mostra espectral original.

#### **3.4.4. Savitzky Golay + SNV**

- **Introducció**

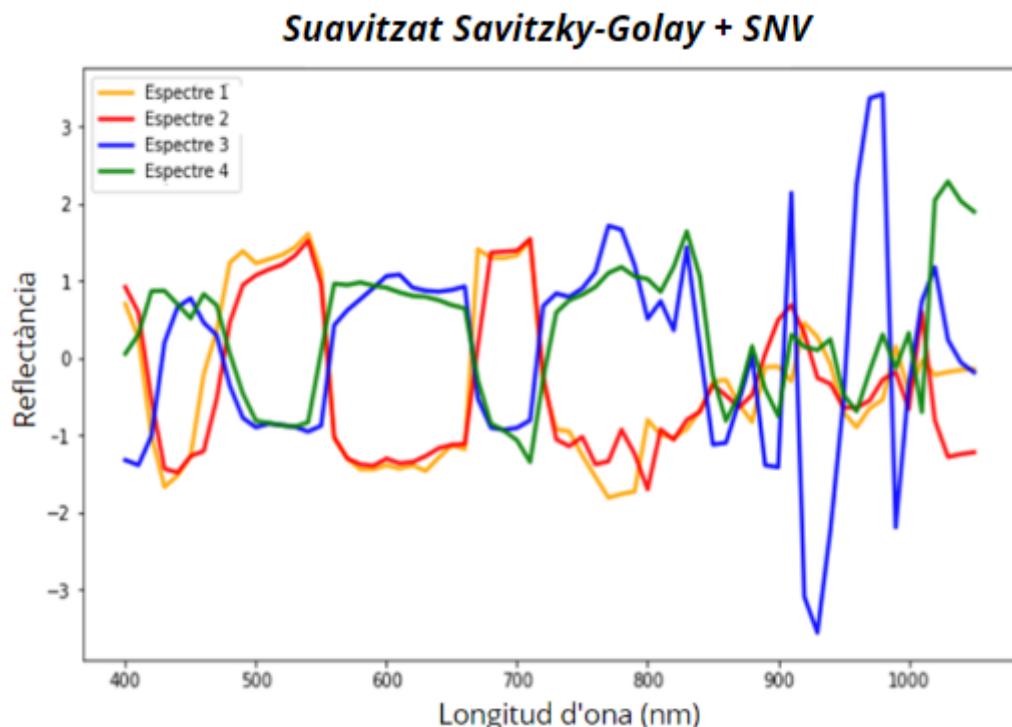
En aquest cas, el que fem és aplicar l'estandardització comentada anteriorment, als resultats obtinguts a partir del filtre suavitzat.

- **Calibració i prova**

En aquest cas, primer s'aplica el filtre Savitzky Golay, amb els valors que s'han comentat anteriorment, a les dades originals. Com s'ha dit, no es fa cap divisió entre calibració i prova (sempre que els valors aplicats siguin els mateixos per a ambdós conjunts).

Seguidament, el que es fa és aplicar l'estandardització. En aquest cas, sí que es realitza aquesta divisió entre els conjunts. Es calculen els valors amb el conjunt de calibració, i posteriorment, es transformen els dos conjunts.

- **Representació Gràfica**



**Figura 3.18.** Representació gràfica de 4 mostres espectrals transformades amb Savitsky-Golay i la estandardització estàndard.

### 3.4.5. Terme mitjà del espectre

- **Introducció**

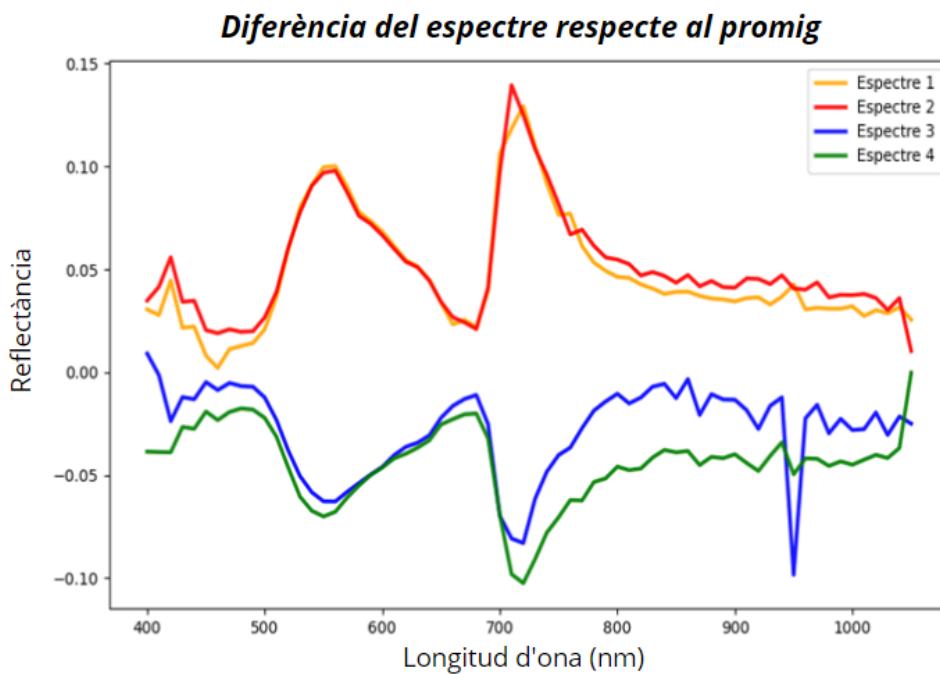
La transformació mitjana de l'espectre és un tipus d'escala de característiques que implica centrar la distribució de bandes espectrals al voltant de zero. Aquesta transformació implica restar l'espectre mitjà de tot el conjunt de dades a cada espectre individual.

- **Calibració i prova**

L'espectre mitjà s'obté calculant l'espectre mitjà de totes les mostres del conjunt de calibració, i una volta ho tenim calculat, es resta de cada espectre individual del de les dades. Això dona lloc a un nou conjunt de dades que consisteix en les diferències entre cada espectre individual a l'espectre mitjà.

Per tant, bandes espectrals es desplacen per tenir un valor mitjà de zero i una distribució més uniforme. Això pot ajudar a reduir l'efecte del soroll i la variabilitat en les bandes espectrals i fer que els models siguin més robusts als canvis en els valors d'entrada.

- **Representació Gràfica**



**Figura 3.19.** Representació gràfica de 4 mostres espectrals transformades amb la diferència respecte al espectre promig.

#### 3.4.6. Sense Transformació

També s'utilitzen les dades originals per tal de comprovar si aquests processaments són útils i en quins casos poden funcionar millor les dades sense realitzar cap transformació.

### 3.5. Reducció de la dimensionalitat

Cada columna es pot entendre com una variable. La reducció de dimensionalitat implica disminuir el nombre de variables. Aquesta reducció pot simplificar la informació i fer-la més comprensible. En general, en presència d'informació redundant, això pot ajudar al bon funcionament del model d'aprenentatge automàtic.

Hi ha diverses tècniques per aconseguir aquest objectiu de reduir el nombre de variables mantenint la major quantitat d'informació possible. En aquest estudi, s'han utilitzat les tècniques que es descriuen a continuació.

#### 3.5.1. Anàlisis de components principals (PCA)

- ***Introducció***

L'anàlisi de components principals (PCA) és una tècnica estadística que es fa servir per reduir la dimensionalitat de les dades. És un mètode que tracta de trobar els components o les variables que maximitzen la variància en un conjunt de dades. La PCA té diferents aplicacions, una d'elles és per a reduir la dimensionalitat de les dades, però altres poden ser eliminar el soroll, analitzar tendències o, tal com també hem vist, encontrar clúster o grups en un conjunt de dades. La PCA té una gran aplicació en el món científic i s'utilitza en multitud de projectes. Trobarem aquest mètode detallat a [19].

La PCA és una transformació lineal que projecta les dades sobre un nou conjunt, que en són els components principals. El primer component principal és la recta que maximitza la variància de les dades, i el segon component principal és la recta que maximitza la variància de les dades seguint ortogonal al primer component principal.

- ***Marc Teòric***

Tenint un conjunt de dades  $X$ , que és una matriu de  $m \times n$  tal i com es veu a l'equació 3.4.1.

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

Equació 3.4.1

Tenim que  $m$  són el nombre d'observacions, i  $p$  el nombre de característiques; en aquest cas, les bandes espectrals (un total de 66). El que volem, és transformar aquesta matriu  $X$  en altra matriu que anomenarem  $Y$ , mitjançant una matriu de  $m \times m$  anomenada  $P$ , tal i com s'aprecia a l'equació 3.4.2.

$$Y = PX$$

## Equació 3.4.2

Si es consideren les files de  $P$  com un vector  $p_1, p_2, \dots, p_m$  i les columnes de  $X$  com vectors  $x_1, x_2, \dots, x_n$  tal i com apareix a l'equació 3.4.3.

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

## Equació 3.4.3

Llavors, aquestes files de  $P$  es convertiran en les direccions dels components principals. Quins han de ser els valors de  $P$  per a re-expressar  $X$  de forma òptima? Cal recordar que PCA tracta de descorrelacionar les dades originals encontrant les direccions on la variància és maximitzada per a definir aquests valors de  $P$ .

Aleshores, es calcula la matriu de covariància. El primer pas és centrar les dades en 0. Com s'ha vist abans, es calcula la mitja de cada variable de les nostres dades com a l'equació 3.4.4.

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

## Equació 3.4.4

Es genera el següent vector representat a l'equació 3.4.5.

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

## Equació 3.4.5

Posteriorment, s'obté la mitja de les dades com a l'equació 3.4.6.

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

## Equació 3.4.6

Aleshores, donades les nostres dades originals centrades en mitja 0 que tenim a  $B$  i, definint la següent matriu  $X$  que tenim a l'equació 3.4.7.

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

## Equació 3.4.7

Es pot representar com a l'equació 3.4.8 cada observació com un vector  $x_1, x_2, \dots, x_n$ . Per exemple,  $x_i$  és un vector de n mostres per a la  $i$  variable.

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

Equació 3.4.8

Ara bé, el que es vol és calcular la matriu de covariància. La covariància entre dos vectors ve definida a partir de la següent equació 3.4.9.

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

Equació 3.4.9

Aleshores, per a calcular la covariància entre totes les mostres que tenim al conjunt de dades, tenim la següent equació 3.4.10.

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

Equació 3.4.10

Com es pot veure, la matriu anterior és una matriu simètrica, ja que la variància entre dos vectors  $x$  e  $y$  és igual a la variància entre  $y$  e  $x$ . Aquesta matriu descriu totes les relacions entre els possibles pars de mesuraments a les nostres dades.

Co s'està parlant de matrius de covariància, cal recordar que, al ser unes matrius simètriques, aquestes són diagonalitzables (teorema de Schur). Això significa que mitjançant una transformació lineal, la matriu pot reduir-se a una forma diagonal. De forma matemàtica s'expressa tal i com es veu en l'equació 3.4.11.

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

Equació 3.4.11

En aquesta expressió, la matriu  $A$  pot descompondre'se mitjançant una matriu invertible  $P$ , on els vectors columna són vectors propis de  $A$ , i  $D$  és una matriu diagonal formada pels valors propis de  $A$ . Els vectors propis són vectors que no canvien la direcció del vector original quan apliquem una transformació lineal, el resultat d'aquests vectors propis són els valors propis, que representarem amb  $\lambda$ .

Aleshores, tornant al objectiu, es vol reduir la redundància a les dades, és a dir, es busca que cada variable es correlacione el mínim possible amb les altres variables. Com hem dit al principi, volem aconseguir la següent transformació:  $Y = PX$  (equació 3.4.2).

Doncs llavors, el que s'ha de definir per aconseguir l'objectiu és que la matriu  $C_Y$  tots els termes no diagonals són tan pròxims com siga possible a 0, definit a l'equació 3.4.12.

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

Equació 3.4.12

Podem veure que s'ha definit a l'equació 3.4.11 que  $A = XX^T$  on  $A$  és simètrica, per tant, està diagonalitzada per la matriu de vectors propis. Els components principals de  $X$  seran els vectors propis de  $XX^T$ , és a dir, els vectors propis de la matriu de covariància  $C_X$ . Ho representem a la següent equació 3.4.13.

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

$$C_X v = \lambda v$$

Equació 3.4.13

Com hem dit,  $\lambda$  són els valors propis associats al vector propi  $v$  de la matriu de covariància  $C_X$ . Els valors propis de  $C_X$  són les arrels de l'equació característica (equació 3.4.14).

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

Equació 3.4.14

Una volta s'ha resolt l'anterior expressió, es poden obtenir els vectors propis. Arribat a aquest punt, es pot calcular el vector propi de cada valor propi tal i com veiem en l'equació 3.4.15.

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

Equació 3.4.15

D'aquesta forma ja es tenen els valors propis i els corresponents vectors propis. En termes generals, els vectors propis amb els valors propis més alts contenen la quantitat més gran d'informació (major variància capturada) sobre la distribució de les dades, i aquests són els que es busquen per tal de reduir la dimensionalitat. L'enfocament comú és classificar els vectors propis de més a menys valor propi corresponent i triar els que vectors propis superiors.

El que busquem és projectar les dades originals a les direccions descrites pels components principals. Com que tenim la relació  $v = P^T$ , això és simplement el que es pot veure a l'equació 3.4.16.

$$Y = v^T X$$

Equació 3.4.16

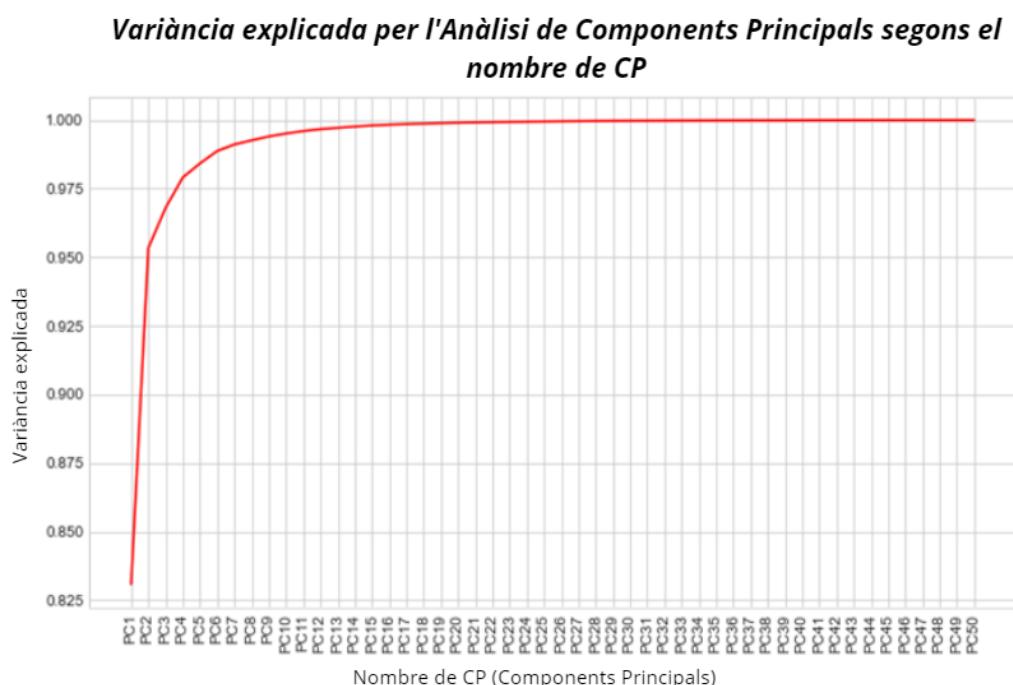
Per tant, el que s'ha de fer és escollir el nombre de components principals. Quant major siga el nombre de components principals, major variància explicada de les dades però menys simplificada es té la informació.

- ***Calibració i prova***

Quan s'aplica la PCA, s'ha d'entrenar amb el conjunt de calibració, ja que no podem calcular els vectors propis de totes les dades, degut a que estem incorporant informació del conjunt de prova. El que fem és, calcular vectors i valors propis de les dades de calibració, i transformem els dos conjunts a partir d'aquests vectors i valors propis que han sigut obtinguts.

- ***Nombre de Components***

Com s'ha comentat anteriorment, a més nombre de components escollim, major serà la variància explicada de les dades. Com podem veure a la següent Figura 3.20, amb set components ja superem el 99% de la variància de les dades.

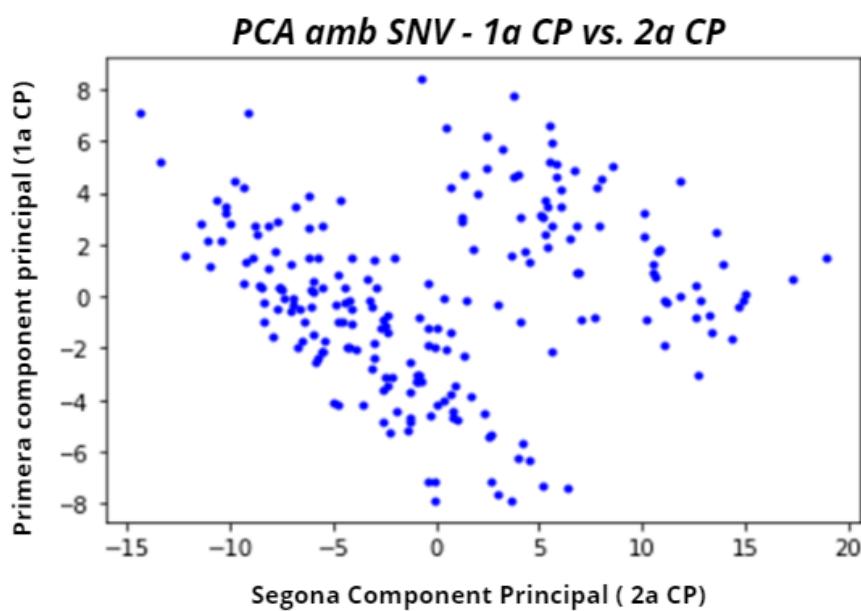


**Figura 3.20.** Variància explicada de les components de PCA.

Per tal de fer proves per a veure que tècniques de processament són millors estimant els diferents nutrients, provarem diferents nombres de components que tinguen diferents variància explicada sobre les dades.

- **Representació Gràfica**

S'observa a la Figura 3.21 les dues primeres components de la PCA del conjunt de calibració, on es poden apreciar dos grups clarament diferenciats, corresponents a les fulles noves i les fulles velles. Aquesta representació gràfica ens permet identificar visualment la presència de dos tipus de fulles, fins i tot sense tenir prèvia informació sobre la seva classificació. La PCA ens proporciona una manera efectiva de visualitzar i comprendre les relacions i variacions en les dades, facilitant la detecció de patrons i agrupacions.



**Figura 3.21.** Primera component vs. segona component de les dades estandarditzades.

### 3.5.2. Autoencoder

- ***Introducció***

Les xarxes neuronals tenen una gran rellevància en l'actualitat i estan transformant la intel·ligència artificial en diversos àmbits, com la sanitat, l'educació, la indústria i l'agricultura, gràcies a les seves múltiples aplicacions i la seva versatilitat. Les xarxes neuronals són capaces d'abordar problemes complexos com el reconeixement d'imatges, el processament del llenguatge natural o la predicció de sèries temporals [20].

Un *autoencoder* és una arquitectura de xarxa neuronal que té com a objectiu comprimir la informació d'entrada i reproduir-la amb precisió a la sortida, utilitzant una representació reduïda de la informació a l'interior. Bàsicament, l'*autoencoder* consta de dues parts: un *encoder* i un *decoder*. L'*encoder* transforma l'entrada original en una representació latent de baixa dimensió, mentre que el *decoder* intenta reconstruir l'entrada original a partir de la representació latent.

Per entendre com funciona un *autoencoder*, és important tenir en compte alguns conceptes bàsics de les xarxes neuronals. Continuem amb una breu explicació sobre aquests conceptes.

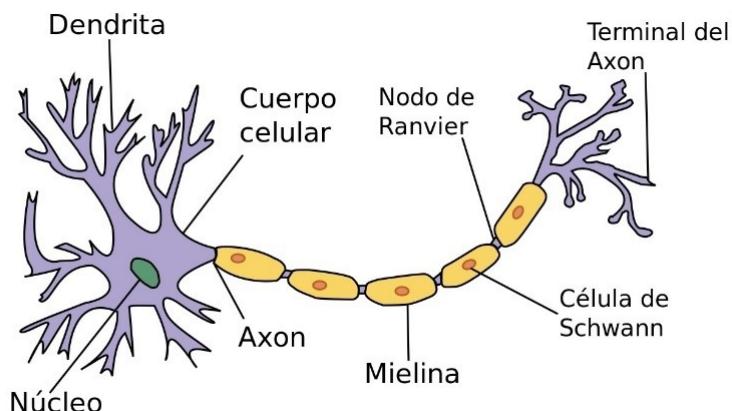
- ***Neurona***

Per a comprendre el funcionament d'una neurona artificial i d'una xarxa neuronal, és útil conèixer una breu introducció biològica sobre les neurones. A finals del segle XIX i principis del segle XX, els científics van realitzar nombrosos descobriments importants sobre les neurones.

Un dels científics més destacats d'aquesta època va ser Santiago Ramon y Cajal, un investigador espanyol conegut principalment pels seus treballs sobre el sistema nerviós. Va publicar un article detallat que descriu l'estructura de les neurones, el qual va contribuir a establir el camp de la neurociència i obrir camí a futurs descobriments sobre el cervell. El treball de Cajal va ser innovador, ja que va demostrar que les neurones són cèl·lules individuals i diferenciades, una desviació important de la visió predominant a l'època, que considerava que les neurones formaven part d'una xarxa contínua.

Aquest treball va contribuir a establir la visió moderna del cervell com un sistema complex de cèl·lules interconnectades. Avui dia, sabem que el sistema nerviós és increïblement complex i que les neurones són les unitats

fonamentals que processen i transmeten la informació a través del sistema nerviós.



**Figura 3.22.** Parts de la neurona humana. Font: Enclopèdia Humanitats.

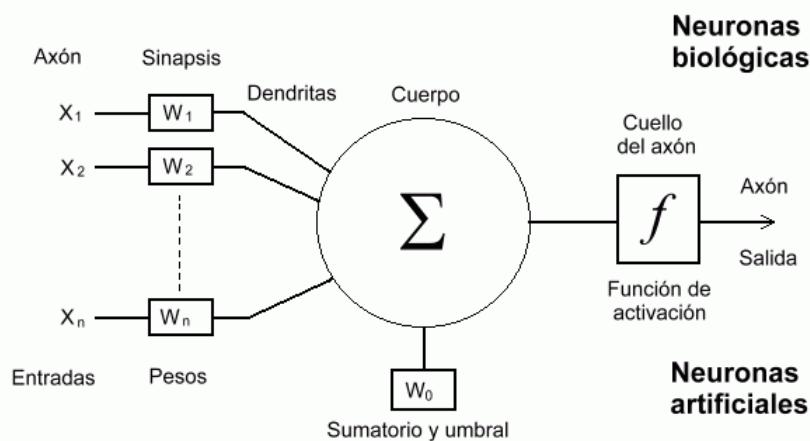
Les neurones són les unitats bàsiques del sistema nerviós i juguen un paper fonamental en la transmissió de la informació al llarg del cos. Com es pot apreciar a la Figura 3.22, una neurona està composta per un cos cel·lular, un axó i dendrites.

Les neurones es comuniquen entre elles mitjançant senyals. Quan un senyal elèctric arriba a l'axó, desencadena l'alliberament de neurotransmissors, que són substàncies químiques. Aquests neurotransmissors viatgen a través de les sinapsis i s'uneixen als receptors de les dendrites de la neurona següent. Aquest procés permet la transmissió del senyal elèctric d'una neurona a l'altra.

El cervell humà conté milers de milions de neurones que estan organitzades en una complexa xarxa. Aquesta estructura intricada permet el processament i el tractament de la informació al llarg del sistema nerviós, jugant un paper essencial en les funcions cognitives, sensorials i motores dels éssers humans.

### • **Neurona artificial**

En 1943, Warren McCulloch i Walter Pitts van publicar un article titulat "A Logical Calculus of the Ideas Immanent in Nervous Activity" en el qual van proposar un model de neurona artificial. La neurona de McCulloch i Pitts és un model matemàtic simple que imita el comportament d'una neurona.



**Figura 3.23.** Parts de la neurona artificial. Font: Universidad de Murcia.

Com es mostra a la Figura 3.23, els pesos i la funció d'activació treballen conjuntament en una neurona artificial. Els pesos determinen la contribució de cada entrada a la sortida, i la funció d'activació determina el valor de sortida de la neurona. Els pesos poden ser positius o negatius, i es poden ajustar per canviar la sortida de la neurona. La sortida de la neurona es calcula mitjançant la suma del producte de cada entrada pel seu pes. Aquesta suma és passada a través de la funció d'activació, que transforma la suma en un valor específic. Hi ha diverses funcions d'activació que funcionen millor o pitjor en funció del problema que es vol resoldre. A la Taula 3.3 es mostren algunes de les més populars

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) <sup>[11]</sup>		$\begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } x > 0 \end{cases} = \max\{0, x\} = x \mathbf{1}_{x>0}$	$\begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x > 0 \\ \text{undefined} & \text{if } x = 0 \end{cases}$	$[0, \infty)$

**Taula 3.3.** Principals funcions d'activació. Font: Sagar Sharma en Activation Functions in Neural Networks (Medium)

- **Perceptró Simple**

Ara que s'ha explicat el funcionament de les parts d'una neurona, és important entendre com aquesta aprèn els pesos necessaris per resoldre el problema. En aquest punt, entra en joc el disseny del Perceptró Simple.

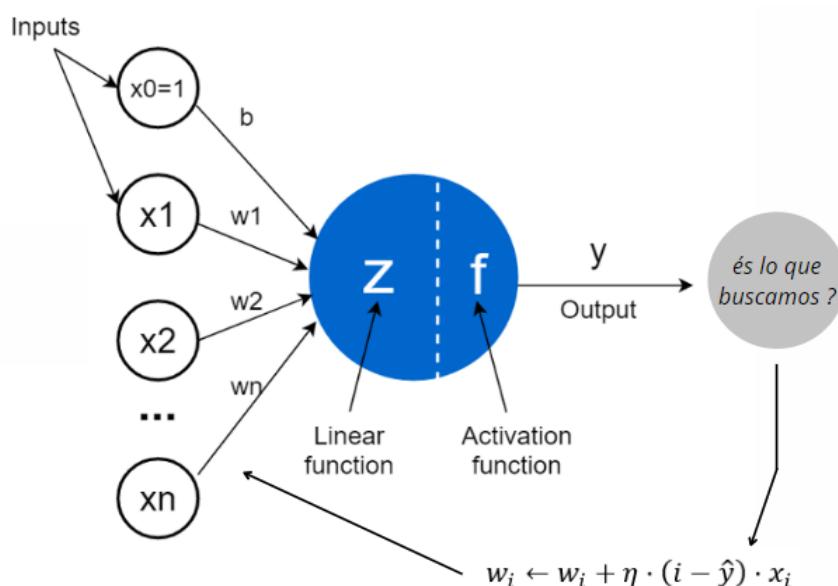
El perceptró va ser inventat a principis dels anys 50 per Frank Rosenblatt, qui es va inspirar en els treballs anteriors de Warren McCulloch i Walter Pitts. Rosenblatt va dissenyar una regla senzilla perquè el perceptró ajuste els seus pesos: considerant un conjunt d'entrenament format per exemples etiquetats, la regla d'aprenentatge actualitza els pesos del perceptró per minimitzar el nombre d'exemples mal classificats.

L'actualització és expressada per l'equació 3.5.1.

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

Equació 3.5.1

on  $w_i$  és el pes i-èsim,  $\eta$  és una taxa d'aprenentatge,  $i$  és l'etiqueta veritable de l'exemple, i  $\hat{y}$  és l'etiqueta predicta.



**Figura 3.24.** Perceptró Simple. Font: Universidad de Murcia

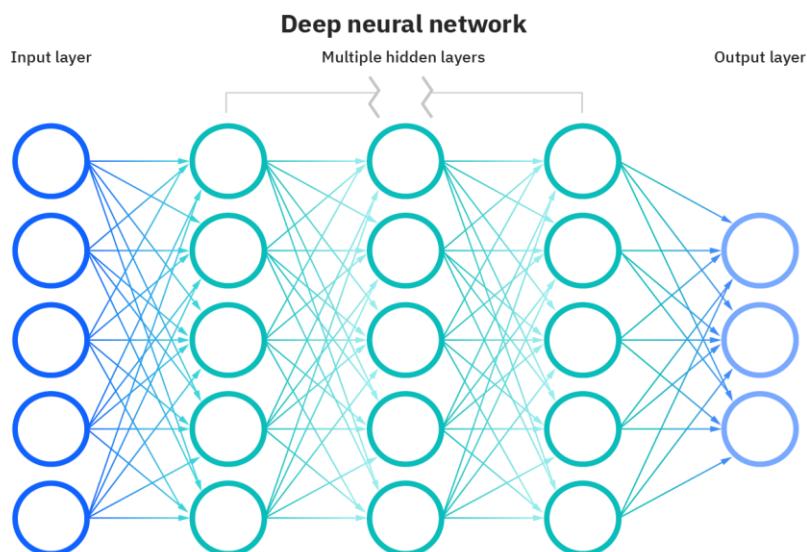
Es pot veure, a la Figura 3.24, que si el valor predit correspon al valor real, és a dir,  $i - \hat{y} = 0$ , aleshores no s'actualitzaran els pesos. En cas que aconseguim minimitzar l'error, o fins i tot, que aquest siga 0, direm que el perceptró (o la xarxa neuronal) ha convergit.

- **Xarxa Neuronal**

El perceptró té una limitació inherent: només pot resoldre problemes linealment separables. Per afrontar problemes linealment no separables, és necessari utilitzar múltiples perceptrons connectats en una xarxa neuronal. Això va ser descrit en el llibre "*Perceptrons*" de Minsky i Papert el 1969. No obstant això, la creació d'aquesta xarxa planteja el problema conegut com a "assignació de crèdit", ja que no sabem en quina proporció assignar l'error a cada perceptró.

Posteriorment, es va descobrir que utilitzant funcions d'activació diferenciables es podien actualitzar els coeficients mitjançant la regla de la cadena. També es van incorporar les funcions de cost. Hi ha diverses funcions de cost diferents que es poden utilitzar en una xarxa neuronal, com l'error quadràtic mitjà (MSE) que mesura la diferència quadràtica mitjana entre les prediccions i els valors reals, i la funció de cost d'entropia creuada (Cross Entropy) que s'utilitza en problemes de classificació, entre altres.

Per actualitzar els coeficients de la xarxa neuronal, també s'utilitzen diferents optimitzadors. Alguns dels optimitzadors més coneguts són Adam, RMSprop, Adadelta, SGD, entre d'altres. Aquests optimitzadors ajuden a trobar els valors òptims dels coeficients per minimitzar la funció de cost i millorar el rendiment de la xarxa neuronal.



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

Hi ha diversos tipus de xarxes neuronals, però totes comparteixen una estructura bàsica comuna. A nivell general, una xarxa neuronal està formada per nodes d'entrada, nodes de sortida i nodes ocults. Els nodes d'entrada reben les dades d'entrada i les processen a través dels nodes ocults. La

sortida dels nodes ocults es transmet als nodes de sortida, que produeixen la sortida final de la xarxa neuronal.

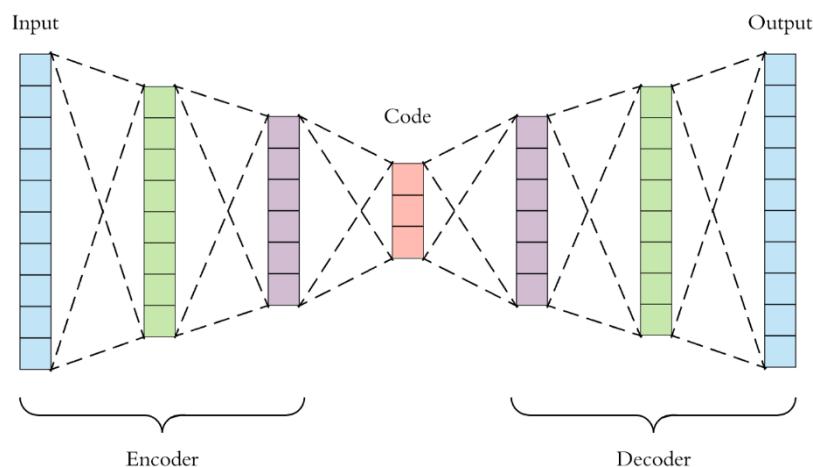
Durant el procés d'entrenament de la xarxa neuronal, les connexions entre els nodes s'ajusten en funció dels resultats obtinguts. Això es realitza mitjançant una tècnica anomenada retropropagació, on la xarxa ajusta els pesos i els biaixos per minimitzar l'error en les prediccions.

Malgrat els seus èxits, les xarxes neuronals tenen algunes limitacions. Sovint són considerades models de caixa negra, ja que pot ser difícil comprendre el raonament que utilitzen per arribar a les seves prediccions. A més, les xarxes neuronals poden requerir una gran capacitat computacional i una gran quantitat de dades d'entrenament per aprendre de manera eficient.

Per aprofundir en aquests conceptes, es pot consultar la referència [22], on s'analitzen en detall la neurona biològica, la neurona artificial, el perceptró i les xarxes neuronals, així com l'estat actual de la investigació en aquest àmbit.

- **Autoencoder**

Un *autoencoder* és una xarxa neuronal utilitzada per aprendre representacions eficients de les dades. L'objectiu principal d'un *autoencoder* és transformar les dades d'entrada en un espai latent de menor dimensió que les dades originals. Per aconseguir-ho, l'*autoencoder* aprèn a comprimir les dades d'entrada en un espai latent i després reconstrueix les dades originals a partir d'aquest espai latent.



**Figura 3.26.** Estructura General d'un Autoencoder. Font: Fernando Sancho Caparrini Blog.

Els *autoencoders* tenen diverses aplicacions, com la reducció de la dimensionalitat de les dades, la eliminació de soroll en les dades, la generació de noves dades o la detecció d'anomalies. Són un tipus d'algorisme

d'aprenentatge no supervisat, ja que no requereixen etiquetes per entrenar el model.

Encara que hi ha diferents tipus d'*autoencoders*, tots comparteixen una estructura bàsica similar. Tenen una capa d'entrada, una capa oculta i una capa de sortida. La capa oculta sol ser de menor mida que les capes d'entrada i sortida, el que permet que l'*autoencoder* funcione com a algoritme de compressió. A més, una capa oculta més petita facilita l'aprenentatge de vectors latents que capturen les característiques essencials de les dades d'entrada.

Cal tenir en compte que el nombre de neurones seleccionades a l'espai latent determina el nombre de variables amb què es representaran les dades originals. Per exemple, si hi ha tres neurones a l'espai latent, això significa que les dades s'hauran transformat en tres variables.

Per a obtenir més informació sobre els diferents tipus d'*autoencoders*, les seves aplicacions, casos d'ús i el seu funcionament, es pot consultar la referència [23].

- ***Autoencoder del projecte***

Amb l'*autoencoder* dissenyat per a aquest projecte, s'ha buscat aconseguir una reducció no lineal de la dimensionalitat amb la millor convergència possible. A continuació es presenta el diagrama de l'*autoencoder* dissenyat:

- ***Capa d'entrada (Input layer)*:** Aquesta capa constarà de 66 neurones que tractaran d'agafar les variables que representen cada una de les bandes espectrals d'un espectre.
- ***Capes ocultes (hidden layers)*:** Tenim dues capes ocultes de 30 neurones. Després d'aquesta capa oculta de 30 neurones, tenim l'espai latent on, el nombre no serà sempre el mateix, sinó que farem diverses proves amb diferents reduccions.
- ***Capa de sortida (output layer)*:** La capa de sortida, igual que la d'entrada, serà de 66 neurones que tractaran de reproduir l'espectre a partir dels valors de les capes ocultes.

Per a la funció de pèrdua, s'ha utilitzat el *mean squared error (mse)*, que s'obté mitjançant la mitjana dels errors al quadrat, tal i com s'aprecia a la següent equació 3.6.1.

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

*Equació 3.6.1*

La funció d'activació que ha proporcionat una millor convergència per a l'autoencoder és la ReLU (*Rectified Linear Unit*), la formula la es troba a l'equació 3.6.2.

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

*Equació 3.6.2*

Per a l'optimitzador, s'ha utilitzat l'algorisme Adam, que és popular en l'entrenament de models d'aprenentatge automàtic i xarxes neuronals. Adam és una combinació del mètode de descens de gradient estocàstic (SGD) amb moments de primer i segon ordre.

Per obtenir més informació sobre les diferents parts de les xarxes neuronals, es pot consultar la referència [24].

### **3.5.3. Random Forest**

- **Problema PCA i Autoencoder**

Com s'ha dit, es busca reduir la dimensionalitat de les dades, però les dues tècniques anteriors necessiten l'espectre obtingut amb totes les seues bandes espectrals (400-1050 nm), és a dir, totes les variables que hi ha a les dades.

Aleshores, també es busca reduir la dimensionalitat mantenint les variables originals agafant diferents longituds d'ona del espectre. Aquesta tècnica pot ser important ja que si, per exemple, seleccionem 5 longituds d'ona per tal de realitzar la investigació de trobar una bona estimació, facilitem l'adquisició de les dades.

Però la pregunta és, com saber quines variables (longituds d'ona) s'han d'escollir per tal de que els models funcionen millor. La resposta simplificada seria en base a la importància de la variable en la predicción d'un model.

- **Selecció de característiques – Random Forest (RF)**

La selecció de les característiques més rellevants en aquest cas s'ha realitzat utilitzant un model de *Random Forest* (RF). En un model *Random Forest*, es construeix una col·lecció d'arbres de decisió i es calcula la importància de cada característica mitjançant la mesura de la disminució de la precisió del model quan aquesta característica es permuta aleatoriament.

La importància de les característiques es pot quantificar utilitzant mètriques com la impuresa de Gini. Les característiques amb puntuacions d'importància més altes es consideren més rellevants per al rendiment del model i és més probable que es seleccionin durant el procés de selecció de característiques.

En aquest cas concret, s'ha entrenat un model *Random Forest* en tot el conjunt de dades i s'han extret les puntuacions d'importància de les característiques. A continuació, s'han seleccionat les 3, 6 i 12 característiques principals en funció de les seves puntuacions d'importància. Aquestes característiques seleccionades s'han utilitzat posteriorment per entrenar els diferents models d'aprenentatge automàtic que es mostren més endavant.

### 3.6. Models

Els darrers anys, els models d'aprenentatge automàtic han adquirit una importància creixent. Això és perquè són capaços d'aprendre i millorar automàticament a partir de l'experiència. Això significa que es poden utilitzar per resoldre una sèrie de tasques que abans eren difícils o impossibles per als programes informàtics tradicionals.

En general, els models d'aprenentatge automàtic han cobrat cada vegada més importància els darrers anys a causa de la seva capacitat per aprendre automàticament i millorar a partir de l'experiència. Això ha donat lloc a una sèrie d'importants aplicacions a diversos camps [25].

#### 3.6.1. Models Lineals

Els models lineals s'utilitzen àmpliament en diverses aplicacions i són fàcils de fer servir i interpretar. Els models lineals es basen en una relació lineal entre les variables d'entrada i la variable de sortida. El model s'entrena utilitzant un conjunt de parells d'entrades-sortida i els paràmetres del model s'aprenen a partir de les dades. A continuació, es pot utilitzar el model per predir la sortida per a nous valors d'entrada.

També són fàcils d'interpretar, cosa que els fa valuosos per a la presa de decisions. Tot i això, els models lineals estan limitats en la seva capacitat per capturar les relacions no lineals entre les variables. Aquests models presentats a continuació, es troben més detallats a [26].

- ***Regressió Lineal***

La regressió lineal és el model d'aprenentatge automàtic més bàsic. Es basa en el supòsit que hi ha una relació lineal entre la variable dependent i la o les variables independents. Per calcular la regressió lineal, el primer pas és trobar l'equació de la recta que representa millor les dades. Això es fa trobant els valors del pendent i l'intercepte que minimitzen la suma dels errors al quadrat.

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

Equació 4.1.1

A l'equació 4.1.1  $\beta_0$  correspon amb l'ordenada a l'origen (intercepte), el valor mitjà de la variable resposta i quan tots els predictors són zero.  $\beta_j$  correspon amb l'efecte mitjà que té sobre la variable resposta l'increment en una unitat de la variable predictora  $x_j$ , mantenint-se constants la resta de variables. Són

els coeficients de la regressió.  $\epsilon_i$  correspon amb el residu o error, la diferència entre el valor observat i l'estimat pel model.

Una vegada trobada l'equació de la recta, es pot utilitzar l'equació 4.1.2 per predir els valors futurs de la variable dependent, basant-se en els valors coneguts de la variable independent.

$$\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1}x_{i1} + \widehat{\beta_2}x_{i2} + \cdots + \widehat{\beta_n}x_{in}$$

Equació 4.1.2

- **Ridge Regresion**

La regressió Ridge és un tipus de regressió lineal que es fa servir per modelar dades susceptibles de multicol·linealitat

La regressió Ridge és semblant a la regressió per mínims quadrats, però utilitza una funció de cost diferent. En comptes de minimitzar la suma dels residus al quadrat, la regressió Ridge minimitza la suma dels residus al quadrat més un terme de penalització. El terme de penalització és una funció de la magnitud dels coeficients. El terme de penalització,  $\lambda$ , s'utilitza per evitar que els coeficients siguin massa grans, cosa que pot provocar un excés d'ajust. Aquesta penalització és coneguda com l2. Partint de l'equació anterior 4.1.1 trobem aquesta equació 4.2.1:

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

Equació 4.2.1

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

Equació 4.2.2

A l'equació 4.2.2 es veu que a mesura que  $\lambda$  augmenta, més gran és la penalització i menor el valor dels predictors. Quan  $\lambda=0$ , la penalització és nul·la i el resultat és el mateix que una regressió lineal.

Fem servir la regressió Ridge, ja que notem que hi ha una gran correlació entre les diferents bandes espectrals (variables predictores), la qual cosa pot ocasionar coeficients molt grans. Cosa que pot portar a un sobreajust, perquè el model s'ajustarà al soroll de les dades en lloc de la relació.

- ***Bayesian Ridge Regression***

La regressió bayesiana és un tipus de regressió que utilitza la inferència bayesiana per estimar els paràmetres del model. La inferència bayesiana és un mètode d'inferència estadística basat en el teorema de Bayes. El teorema de Bayes s'utilitza per calcular la probabilitat que es produïsca un esdeveniment determinat, basant-se en dades anteriors. El teorema de Bayes es pot escriure com es mostra a l'equació 4.3.1:

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

Equació 4.3.1

on:

- $P(A|B)$  és la probabilitat posterior que passi el succés A, atès que ha passat el succés B.
- $P(B|A)$  és la probabilitat que passi el succés B, atès que ha passat el succés A.
- $P(A)$  és la probabilitat a priori que passi el succés A.
- $P(B)$  és la probabilitat a priori que es produueixi el succés B.

La Regressió bayesiana *Ridge* és un mètode d'inferència estadística en què utilitzem el teorema de Bayes per actualitzar les probabilitats dels esdeveniments a mesura que observem dades noves.

A l'enfocament bayesià, els paràmetres no només s'estimen basant-se en les dades, sinó també en el coneixement previ sobre els paràmetres. Es suposa que la distribució a priori és una distribució normal amb una mitjana de zero i una variància de la matriu d'identitat. Seguint la regressió lineal, podem extraure aquesta equació 4.3.2:

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

Equació 4.3.2

Podem formular la funció de versemblança que descriu la relació entre els paràmetres i les dades. Aquesta funció es conegeuda i s'expressa com a l'euació 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 apliquem el teorema de Bayes a aquesta distribució dels paràmetres  $\beta$  donada l'observació tal i com veiem a l'equació 4.3.4:

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

*Equació 4.3.4*

on ens referim a:

- $P(\beta|y, X)$  com la distribució posterior.
  - $P(y, X|\beta)$  com la funció de versemblança.
  - $P(\beta)$  la distribució a priori.
  - $P(y, X)$  probabilitat de les pròpies observacions (evidència).
- $$P(y, X) = \int_{-\infty}^{\infty} P(y, X|\beta) P(\beta) d\beta$$

Per a les distribucions a priori dels coeficients s'utilitzen distribucions no informatives. Açò significa que és una distribució de probabilitat que no es basa en cap informació o coneixement previ. Podem imposar, per exemple, una distribució uniforme.

La funció de versemblança és una distribució condicional de les característiques de la resposta i del predictor donat el model. A mesura que augmenta el nombre de dades de la mostra, la semblança es farà més precisa, la seva incertesa es reduirà i superarà la distribució a priori en algun moment.

La distribució posterior, al contrari de la de versemblança, és una distribució condicional dels paràmetres del model ateses les característiques de la resposta i del predictor, i ve basada en la similitud de les dades i la distribució de les dades. Donat un nombre d'iteracions, anirem aproximant-nos cap a la distribució a posteriori que s'adapte a les dades.

Cal afegir que tot aquest procés, incloem la norma l2, és a dir, la regularització comentada anteriorment. Aleshores, tenim altre paràmetre  $\lambda$  que penalitzarà els coeficients alts tal com hem vist.

- ***SGD Regressor***

El descens de gradient estocàstic (SGD) és un algorisme d'optimització utilitzat per trobar els valors dels paràmetres que minimitzen una funció de cost. És una versió estocàstica del descens de gradient, cosa que significa que utilitza l'aleatorietat per seleccionar els exemples d'entrenament a cada pas.

El SGD comença amb un conjunt de valors de paràmetres (normalment inicialitzats a 0) i després ajusta aquests valors de forma iterativa per minimitzar la funció de cost. A cada iteració, SGD selecciona aleatoriament un exemple d'entrenament i calcula el gradient de la funció de cost respecte als paràmetres. El gradient és un vector que indica la direcció on han de canviar els paràmetres per minimitzar la funció de cost. A continuació, el SGD ajusta els paràmetres a la direcció del gradient i repeteix el procés.

L'equació bàsica de la SGD és definida per l'equació 4.4.1.

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

Equació 4.4.1

on  $\widehat{\theta}_t$  és el vector de paràmetres actual,  $\alpha$  és la taxa d'aprenentatge, i  $\nabla_{\theta} L(\widehat{\theta}_t)$  és el gradient de la funció de pèrdua  $L(\widehat{\theta}_t)$  pel que fa a  $\theta$ .

El gradient es fa servir llavors per actualitzar els paràmetres. Aquest procés es repeteix fins que la funció de cost convergeix un mínim. La funció de cost mesura la diferència entre el valor previst i el valor real. La taxa d'aprenentatge ( $\eta$ ) és un hiperparàmetre que controla la mida dels passos fets pel SGD.

Per últim, el SGDRegressor una tècnica d'optimització iterativa utilitzada per trobar el conjunt òptim de paràmetres que minimitze una funció de cost determinada. SGDR és un algorisme eficient amb una complexitat computacional baixa.

### 3.6.2. Cross-Descomposition

- **PLS Regressor**

El model de regressió de mínims quadrats parcials (PLS) és un tipus d'anàlisi de regressió que s'utilitza per predir els valors d'una variable de resposta basant-se en els valors d'una o més variables predictores. És un model d'aprenentatge automàtic molt utilitzat en el sector de l'agroenginyeria.

El model de regressió PLS es basa en la idea de projectar les variables predictores (X) en un espai de dimensions inferiors, anomenat espai latent, per tal de trobar la millor combinació lineal de les variables predictores que pugui explicar la variància de la variable de resposta (Y).

L'espai latent és definit per un conjunt de vectors latents ortogonals, representats pel símbol T, que es calculen a partir de les variables predictores (X) mitjançant descomposició de valor singular (SVD), la qual ja hem comentat anteriorment.

Els valors predictius de la variable de resposta (Y) es calculen multiplicant els vectors latents (T) pels corresponents pesos o coeficients (B), que s'estimen minimitzant la suma residual de quadrats (RSS) mitjançant un algorisme d'optimització com el descens gradient.

En termes matemàtics, el model de regressió PLS es pot escriure com segueix la equació 4.5.1:

$$I = X * B + E$$

Equació 4.5.1

on I són els valors predictius de la variable de resposta, X és la matriu de variables predictores, B és la matriu de pesos o coeficients i E és el terme d'error.

Un dels principals avantatges del PLS és que pot manejar variables predictores molt col·lineals sense perdre interpretabilitat. A més, PLS és menys sensible a les observacions perifèriques que altres mètodes de regressió, cosa que pot millorar la precisió del model. A més, PLS és capaç de manejar dades amb valors perduts, cosa que pot ser útil en moltes aplicacions del món real. En general, PLS és una eina flexible i potent per modelitzar relacions complexes entre variables de resposta i predictores.

Podem trobar detallat en major mesura els conceptes i mètodes d'aquest model, i també les seues aplicacions, a la referencia [27].

### 3.6.3. Nearest Neighbours

- ***K – Neighbors Regressor***

El regressor k-veïns més propers és un algorisme d'aprenentatge automàtic que s'utilitza per a tasques de regressió. Funciona fent prediccions basades en els valors dels "k" punts més propers de les dades. Això significa que examina els k punts de les dades més properes al punt per al qual es fa la predició i utilitza els seus valors per fer la predició.

K-veïns més propers és un algorisme senzill i intuïtiu, fàcil d'aplicar i interpretar. Sovint es fa servir en casos en què la relació entre els predictors i la variable de resposta és complexa i no es capta fàcilment amb un model lineal.

Per entrenar el model, primer cal proporcionar a l'algorisme un conjunt de dades d'entrenament. A continuació, l'algorisme utilitza aquestes dades per aprendre la relació entre els predictors i la resposta. Trobarem més informació en [28].

### 3.6.4. Support Vector Machine (SVM)

Tots els models que s'expliquen a continuació en aquest apartat, es troben més detallats en [29].

- **Support Vector Regressor (SVR)**

SVR és un tipus de màquina de vectors de suport (SVM) que s'utilitza per a l'anàlisi de regressió. Igual que altres màquines de vectors suport, SVR és un algoritme d'aprenentatge supervisat que utilitza dades d'entrenament per aprendre la relació entre una variable de resposta i variables predictores. Un cop entrenat el model, es pot utilitzar per fer prediccions sobre la variable de resposta per a noves dades.

SVR és un mètode potent i flexible aplicat a una àmplia gamma de problemes. Un dels principals avantatges de SVR és que pot manejar dades amb relacions no lineals entre les variables de resposta i predictores, cosa que sol passar en les dades del món real. A més, s'ha demostrat que SVR funciona bé amb dades d'alta dimensionalitat, cosa que pot suposar un repte per a altres mètodes de regressió. En general, SVR és una eina valuosa per modelar i comprendre relacions complexes a les dades.

Les equacions del model *Support Vector Regressor* (SVR) es poden derivar de les equacions generals de les màquines de vectors de suport (SVM). En un problema SVR típic, l'objectiu és trobar una funció que pugua predir una variable de resposta,  $y$ , basant-se en un conjunt de variables predictores,  $X$ . El valor predit de  $y$  per a un conjunt donat de valors predictors,  $X$  ve donat per la següent equació 4.6.1:

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

Equació 4.6.1

on  $w$  és un vector de ponderacions,  $X$  és un vector de valors predictors i  $b$  és un terme de biaix.

Per trobar els valors òptims de  $w$  i  $b$ , el model SVR ha de resoldre el problema d'optimització següent:

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

Equació 4.6.2

Subjecte a:

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

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

Equació 4.6.3

On, en l'equació 4.6.2  $\|w\|$  és la norma euclidiana del vector de pesos, i a l'equació 4.6.3  $\epsilon$  és un paràmetre definit per l'usuari que controla la mida del marge d'error, i  $y$  és el veritable valor de la variable de resposta.

Un dels principals avantatges de SVR, com hem dit, és que pot manejar dades amb relacions no lineals entre les variables de resposta i predictores, cosa que sol passar en les dades del món real. A més, s'ha demostrat que SVR funciona bé amb dades d'alta dimensionalitat, cosa que pot suposar un repte per a altres mètodes de regressió. A més, SVR es pot utilitzar amb una varietat de funcions de *kernel*, cosa que permet a l'usuari especificar diferents tipus de relacions entre les variables de resposta i predictores. En general, SVR és una eina flexible i potent per a l'anàlisi de regressió.

- **Linear SVR (LSVR)**

*Linear SVR* és una variant del model SVR que s'utilitza per a l'anàlisi de regressió. A diferència d'altres màquines vectorials de suport, el model LSVR utilitza una funció del *kernel* lineal, el que significa que només pot modelar relacions lineals entre la resposta i les variables predictores. Això simplifica el problema d'optimització i fa que el model LSVR sigui més eficient computacionalment que altres màquines vectorials de suport. Un cop trobats els valors òptims de  $w$  i  $b$ , el model Linear SVR es pot utilitzar per fer prediccions sobre la variable de resposta per a dades noves.

- **NuSVR**

NuSVR és un mètode flexible d'anàlisi de regressió que sol utilitzar quan la relació entre les variables de resposta i predictores és complexa o desconeguda. S'anomena així, ja que té un paràmetre  $\nu$ . Amb aquest model es poden utilitzar una varietat de funcions de nucli.

Un avantatge de NuSVR és que es pot gestionar tasques de regressió lineal i no lineal i és altament escalable, cosa que significa que pot manejar grans conjunts de dades de manera eficient.

Com s'ha dit, altre avantatge de NuSVR és que té un paràmetre incorporat,  $\nu$ , que permet a l'usuari especificar un límit inferior en el nombre de vectors de suport que el model ha d'utilitzar. Això pot ser útil per evitar el sobreajust, ja que garanteix que el model no intente ajustar-se al soroll o les variacions aleatòries de les dades. Per tant, NuSVR proporciona un rendiment robust fins i tot en presència de punts de dades sorollosos o atípics, gràcies al seu ús del paràmetre  $\nu$ .

En resum, NuSVR és un algorisme potent i versàtil que es pot utilitzar per a diverses tasques de regressió. La seva capacitat per manejar relacions lineals i no lineals, així com la seva robustesa en presència de dades sorolloses, el converteixen en una eina valuosa.

#### 3.6.5. Decision Trees and Ensemble Models

L'explicació més detallada d'aquests mètodes es troba a [31].

- **Decision Tree Regressor**

Un arbre de decisió és un tipus de model d'aprenentatge automàtic que s'utilitza per a tasques de regressió i classificació. Funciona creant una estructura en forma d'arbre on cada node intern representa una decisió o divisió basada en el valor d'una determinada característica, i cada node fulla representa un valor predit.

L'arbre es construeix dividint les dades d'entrenament en subconjunts cada cop més petits en funció dels valors de determinades característiques. Aquest procés es repeteix recursivament fins que cada subconjunt conté un únic valor objectiu.

Per dividir les dades a cada node, el model utilitza un criteri de divisió com l'error quadràtic mitjà o l'error absolut mitjà. Per exemple, si intentem predir el preu d'una casa en funció de la mida, el model podria dividir les dades en un node determinat en funció de si la mida està per sobre o per davall de la mida mitjana de les cases del subconjunt.

El valor previst per a un punt de dades nou es determina recorrent l'arbre i arribant a un node fulla. El valor al node fulls es pren com el valor predit per al nou punt de dades.

Un dels principals avantatges d'utilitzar un arbre de decisió és la seva interpretabilitat. Atès que el model crea una estructura en forma d'arbre, les decisions i les divisions preses pel model poden ser fàcilment compreses i interpretades pels humans.

- **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 predicción 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 predicción 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* és un tipus de model d'aprenentatge automàtic que s'utilitza per a tasques de regressió. És similar a un *Random Forest*, però els arbres de decisió individuals s'entrenen utilitzant un enfocament més aleatori i menys cobejós.

En un *Random Forest*, cada arbre s'entrena en un subconjunt diferent de dades i amb un subconjunt diferent de característiques. Això ajuda a reduir el sobreajustament i millorar la precisió global del model. En canvi, els arbres individuals d'un *Extra Tree Regressor* s'entrenen utilitzant llindars aleatoris per a cada característica, en lloc dels llindars òptims trobats mitjançant un criteri

de divisió. Això fa que el model sigui menys sensible a les dades específiques utilitzades per a l'entrenament i pot donar lloc a prediccions més precises sobre dades noves.

Aleshores *l'Extra Tree Regressor* inclou llindars aleatoris. A banda, consta de les parts anteriorment ja descriptes: un valor del node full, una estructura de l'arbre i d'un mètode de combinació.

- ***Ada Boost Regressor:***

*AdaBoost*, que és l'abreviatura d'*Adaptive Boosting*, és un altre algorisme d'aprenentatge supervisat que pertany a la família de mètodes d'assemblatge. L'algorisme *AdaBoost* es basa en la idea de millorar iterativament un conjunt de regressors febles ponderant i combinant els seus resultats.

El procés general de l'algorisme *AdaBoost* per a un problema de regressió es detalla a continuació:

- **Inicialització:** s'assigna un pes uniforme a cada observació del conjunt d'entrenament. En el cas de la regressió, aquests pesos s'utilitzen per calcular l'error ponderat de cada model candidat.
- **Iteració:** l'algoritme itera a través d'un nombre definit d'etapes, cadascuna de les quals fa el següent:
  - **Ajustar un regressor feble:** un regressor feble, com ara un arbre de decisions poc profund, s'entrena utilitzant les variables independents i la variable objectiu.
  - **Calcula l'error ponderat:** el rendiment del regressor ajustat s'avalua mitjançant l'error ponderat, que té en compte els pesos de les observacions.
  - **Calcula el pes del regressor:** el pes del regressor recentment ajustat es determina en funció del seu error ponderat. Els regressors amb menys error tenen més pes en el muntatge final.
  - **Actualitza els pesos de les observacions:** s'incrementa el pes de les observacions mal previstes i es redueix el pes de les observacions predites correctament, per tal de prioritzar les observacions difícils de predir en les iteracions següents.
  - **Normalitzar els pesos:** els pesos de les observacions es normalitzen de manera que sumin 1.
- **Combinació:** un cop s'ha repetit a través de totes les etapes, els regressors ponderats es combinen per formar el model final d'*AdaBoost*.

- **Gradient Boosting Regressor:**

El *Gradient Boosting Regressor* funciona construint seqüencialment un conjunt d'arbres de decisió febles, on cada arbre es crea per corregir els errors residuals dels arbres anteriors. En altres paraules, l'algorisme intenta minimitzar la funció de pèrdua a cada pas, ajustant el model en la direcció del gradient negatiu.

Per il·lustrar el procés, considereu les següents etapes de l'algorisme:

- **Inicialització:** S'estableix un model base, que és una constant que minimitza la funció de pèrdua. En el cas de la regressió, aquest sol ser el valor mitjà de la variable objectiu en el conjunt d'entrenament.
- **Iteració:** l'algoritme itera a través d'un nombre definit d'etapes, cadascuna de les quals fa el següent:
  - **Calcula els residus:** els errors residuals s'obtenen comparant les prediccions del model actual amb els valors reals de la variable objectiu.
  - **Ajusta un arbre de decisió:** s'entrena un arbre de decisió feble utilitzant les variables independents i els residus com a variable objectiu.
  - **Calcula el pes òptim:** el pes òptim de l'arbre nou instal·lat es determina mitjançant tècniques d'optimització basades en el descens de gradients, per tal de minimitzar la funció de pèrdua.
  - **Actualitza el model:** l'arbre ponderat s'afegeix al model actual.
- **Combinació:** un cop s'ha repetit a través de totes les etapes, els arbres ponderats es combinen per formar el model final de Gradient Boosting Regressor.

Un cop finalitzat el procés d'entrenament, al igual que *Ada Boost*, el *Gradient Boosting* utilitza les prediccions combinades de tots els models individuals per fer una predicció final. Aquesta predicció final és una mitjana ponderada de les prediccions de cada model individual.

A la següent Taula 3.4 es poden veure quins són els aspectes que diferencien aquests dos mètodes que, a simple vista, poden resultar molt similars.

Aspecte	Gradient Boosting Regressor	AdaBoost
<b>Enfocament d'optimització</b>	Utilitza el descens de gradient per minimitzar la funció de pèrdua a cada iteració.	Es basa en l'adaptació dels pesos de les observacions i els models febles.
<b>Actualització del model</b>	A cada iteració, afegeix un arbre ponderat que corregeix els errors residuals.	Combina models febles ponderats segons el seu rendiment en la predicció d'observacions difícils.
<b>Funció de pèrdua</b>	Permet diferents funcions derivables .	Utilitza una funció de pèrdua exponencial específica.
<b>Control de complexitat</b>	Ofereix un major nombre d'hiperparàmetres ajustables.	Compta amb menys hiperparàmetres ajustables.
<b>Sensibilitat al soroll</b>	És menys sensible al soroll i a les observacions atípiques-	És més sensible al soroll i a les observacions atípiques.

**Taula 3.4.** Diferències entre Gradient Boosting Regressor i AdaBoost.

### 3.6.6. Bagging (Ada Boost i Gradient Boosting)

El *bagging*, que prové de "Bootstrap Aggregating", és un mètode que pretén millorar l'estabilitat i la precisió dels models de predicció. Aquest enfocament es basa en la combinació de múltiples models entrenats amb diferents subconjunts de dades generades per mostreig amb substitució (bootstrap) del conjunt de dades original [31].

Utilitzar amb *AdaBoost* i *Gradient Boosting*, tot i que ja tenen mecanismes interns per millorar la precisió i reduir l'error de predicció, *bagging* pot oferir avantatges addicionals.

Quan es fa servir aquest mètode cada model feble s'entrena amb un subconjunt de dades generades pel mostreig amb substitució en lloc d'utilitzar el conjunt de dades complet. Això pot augmentar la diversitat entre els models individuals, donant lloc a un model més robust i estable. Però, per altra banda, també pot augmentar la complexitat i el temps d'entrenament dels models.

### 3.7. Avaluació dels models

Per mesurar el funcionament dels models, es disposar d'una sèrie de mètriques. Com el problema és de regressió, s'han utilitzat els següents errors:

#### Mètriques numèriques d'error

- **Coeficient de determinació ( $R^2$ ):** Es calcula com apareix a la següent Equació 5.1. Es calcula el quadrat del coeficient de correlació de Pearson.  $R^2$  tracta d'explicar la variació de la variable resposta (en el nostre cas el nutrient que es vol predir), en relació amb una o més variables predictores. El  $R^2$  va des de 0 fins a 1, sent 1 la millor predicció possible.

$$R^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Equació 5.1. Equació del Coeficient de determinació ( $R^2$ )

- **Mean Absolute Error (MAE):** Aquest error mesura la mitjana de l'error absolut entre el valor real i les prediccions tal i com es veu a l'Equació 5.2. Atorga el mateix error a errors petits que a errors grans.

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

Equació 5.2.. Equació de l'error mitjà absolut

- **Root Mean Squared Error (RMSE):** És l'arrel quadrada de la mitjana de les diferències al quadrat entre els valors reals i les prediccions. Com que la resta dels errors s'eleva al quadrat, aquesta mètrica penalitza els errors grans. Es pot veure a la següent equació (Equació 5.3.).

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

Equació 5.3. Equació de l'arrel quadrada de la mitjana de l'error

## **Residus del model**

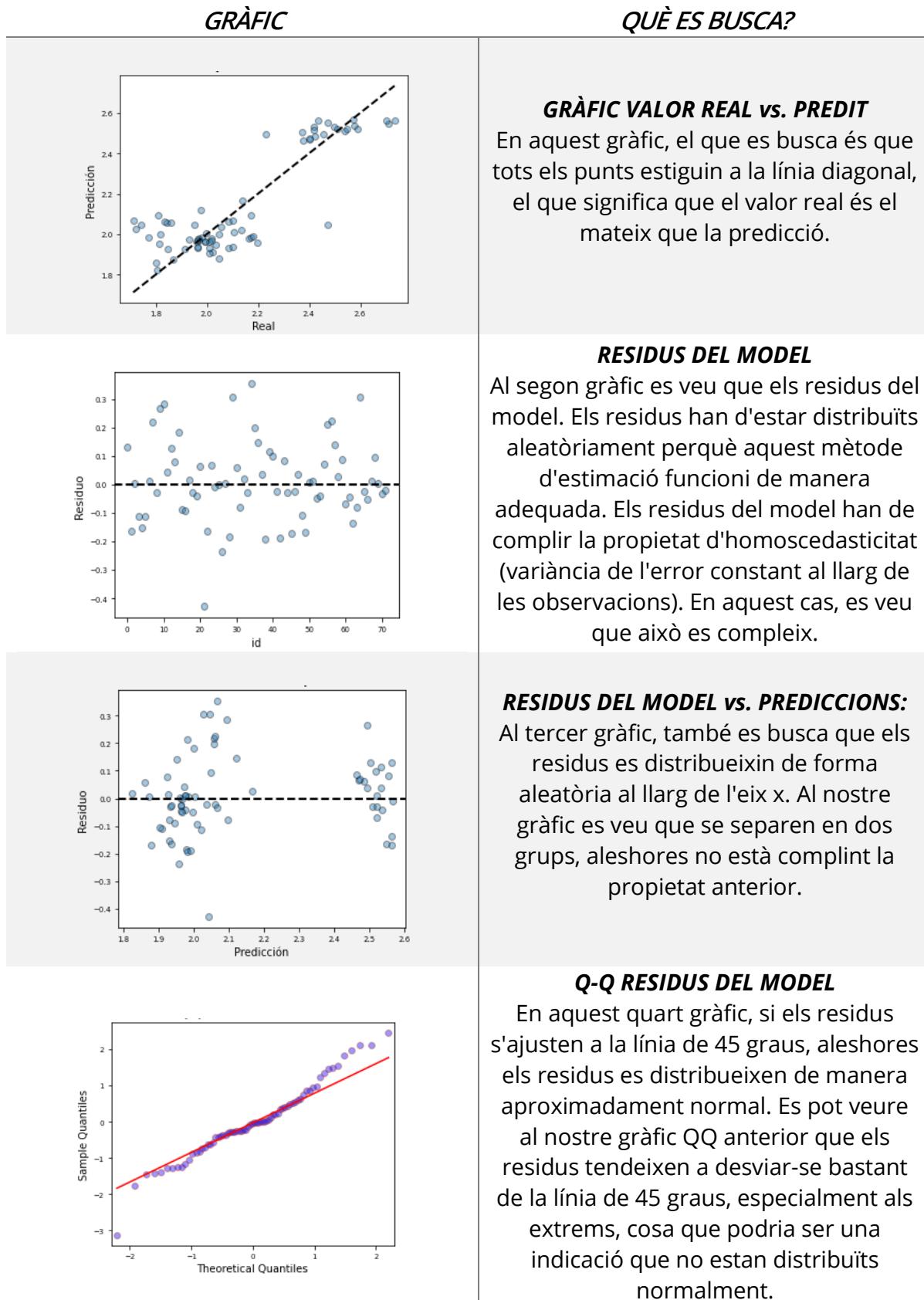
També es visualitzen gràficament els residus en el cas de que es vulga més informació sobre el model. Els residus indiquen quina és la part de la variable que es busca predir no està explicada pel model. Si són nuls, llavors no hi ha cap error en l'estimació, ja que els valors observats coincideixen amb els valors estimats. Si és positiu aleshores el valor observat és més gran que el valor estimat el que implica que s'infraestima la variable  $y$ , si per contra, el residu és negatiu aleshores el valor observat és menor que el seu valor estimat. Per tant, es sobreestima la variable  $y$ .

Aquest tipus de gràfics et permet avaluar 3 qüestions principalment:

- Si s'utilitza el tipus de relació adequada. Si el tipus de model que estem utilitzant no és adequat es troben biaixos o tendències en els residus.
- Si la variància és constant o, per contra, tenim problemes de dispersió irregular. Un dels supòsits del model de regressió lineal és que la variància dels residus és constant, és a dir, que els residus es distribueixen a l'atzar al voltant de zero.
- Si hi ha valors anormals que puguen pertorbar i invalidar el model.

Aleshores, es busca que els residus s'aproximen el màxim possible a 0, per a saber-ho, el millor que es pot fer, és representar-ho gràficament. Existeixen diferents gràfics per tal d'avaluar el funcionament del model, en aquest projecte se n'han utilitzat quatre. Es poden veure exemples de gràfics, i s'explica que és el que es busca trobar en eixos gràfics per tal de saber si el model funciona correctament.

En la següent Taula 3.5 es pot veure els diferents gràfics utilitzats al projecte i com s'han d'interpretar per poder saber com funciona el model.



**Taula 3.5.** Com analitzar els residus d'un model.

### 3.8. Construcció de models i proves realitzades.

#### Construcció de models

La cerca dels millors paràmetres d'un model d'aprenentatge automàtic és un pas essencial en el procés de construcció i entrenament d'un model. Els paràmetres són els valors que determinen el comportament del model, i trobar els valors òptims és essencial per aconseguir un bon rendiment en la tasca objectiu.

Aquest procés implica provar diferents valors per als paràmetres del model, avaluar el rendiment del model en un conjunt de validació i, tot seguit, ajustar els valors dels paràmetres en funció dels resultats. Aquest procés es pot repetir fins a trobar els valors òptims dels paràmetres.

Per tal de trobar aquests valors òptims, s'han realitzat proves amb l'anomenat *Grid Search* [32]. El mètode *Grid Search* és una tècnica molt utilitzada per cercar els millors paràmetres d'un model d'aprenentatge automàtic. Aquest mètode consisteix a definir una quadrícula de valors de paràmetres i, a continuació, entrenar i avaluar un model per a cada combinació de valors de paràmetres de la quadrícula. A continuació, se selecciona la millor combinació de valors de paràmetres segons el rendiment del model en un conjunt de calibració.

Els rang de paràmetres utilitzat per a cada model han sigut el següent:

- **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].

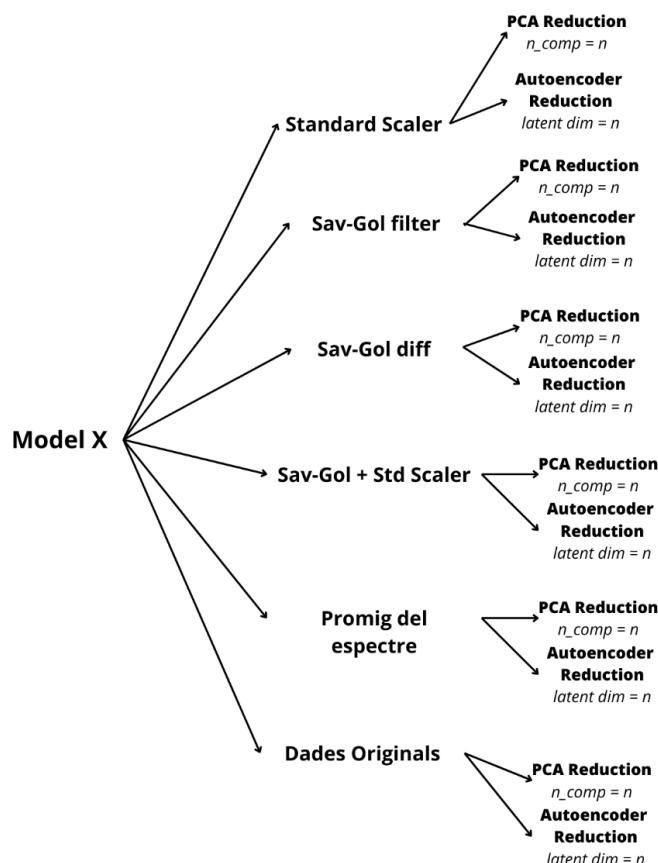
- **LinearSVR:** *epsilon* [0.1, 0.01, 0.001], *C* [0.1, 1, 10, 100], *loss* ['epsilon\_insensitive', 'squared\_epsilon\_insensitive'], *dual* [True, False], *fit\_intercept* [True, False], *intercept\_scaling* [0.1, 1, 10, 100], *max\_iter* [1000-10000], *tol* [1e-4, 1e-5, 1e-6].
- **NuSVR:** *nu* [0.1, 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].

### Proves realitzades utilitzant totes les bandes

Com aquest treball té un enfocament investigador, l'objectiu és trobar la millor manera d'estimar els valors nutricionals utilitzant dades espectrals. La idea és provar una àmplia gamma de tècniques, tant en termes de preprocessament de dades com de models d'aprenentatge automàtic, per tal de determinar les combinacions més eficaces.

En aquest sentit, s'han realitzat diverses proves amb tècniques de reducció de dimensionalitat que requereixen totes les bandes espectrals, com ara PCA (Anàlisi de Components Principals) i l'*autoencoder*.

A continuació, es detallen les proves realitzades amb aquestes tècniques:



**Figura 3.28.** Proves realitzades al projecte per a cada model i cada nutrient.

Com es pot observar a la Figura 3.25, per a cada model específic (Model X), s'han realitzat proves utilitzant les diferents transformacions mencionades anteriorment. Per a cada transformació, s'ha reduït la dimensionalitat utilitzant les dues tècniques de reducció explicades, PCA i autoencoder, amb un nombre específic de components (n).

- n per a PCA: 5, 7, 12, 15, 18, 20, 25 i 30 components
- n per a autoencoder: 5, 7, 12, 15, 18, 20, 25 i 30 components.
- Sense reduir la dimensionalitat.

Per a cada model que intenta estimar un nutrient específic, s'han realitzat proves amb 6 diferents preprocessaments de les dades. Per a cada un d'aquests preprocessaments, s'han provat 18 formes diferents de reducció de la dimensionalitat, a més de les dades originals.

### Proves realitzades seleccionant les bandes

D'altra banda, també hem realitzat una selecció de bandes utilitzant el model *Random Forest*. En aquest procés, primerament hem seleccionat el preprocessament que millor funciona per al model *Random Forest*, que és el preprocés Savitzky-Golay juntament amb l'estandardització SNV. Aquest

preprocessament ha demostrat obtenir resultats d'error més baixos per a la majoria dels nutrients en els quals s'ha provat.

Un cop entrenat aquest model, hem ordenat les bandes espectrals per a cada nutrient segons la seva importància segons el model. A continuació, hem provat els models seleccionant 3, 6 i 12 bandes específiques. En aquest cas, no hem provat més preprocessaments ja que les dades ja havien sigut preprocessades abans de seleccionar les bandes amb el *Random Forest*. Per tant, una vegada seleccionades aquestes bandes, hem entrenat els diferents models que ja hem presentat prèviament.

Aquesta selecció de bandes basada en la importància assignada pel model *Random Forest* ens permet reduir la dimensionalitat de les dades, seleccionant només les bandes més rellevants per a cada nutrient en particular.

## 4. RESULTATS I DISCUSSIÓ

L'aprenentatge automàtic és una eina potent que s'ha utilitzat àmpliament en diversos camps per resoldre problemes complexos. En aquest treball, presentem els resultats dels diferents models d'aprenentatge automàtic que hem descrit anteriorment. Aquests resultats ofereixen una visió del potencial de l'aprenentatge automàtic i de la seva capacitat per impulsar el progrés en diverses àrees de la investigació i la indústria.

En aquest article, es presenten els resultats per als macronutrients primaris, macronutrients secundaris i micronutrients. Per a cada secció, mostrem:

- Una taula amb els tres millors models utilitzant PCA o *autoencoder* (reducció de dimensionalitat amb l'ús de totes les bandes espectrals).
- Una taula que mostra els resultats utilitzant la reducció de dimensionalitat amb selecció de bandes.

A més dels millors models, també incloem una regressió lineal en cada taula. Comparar els resultats dels models amb una regressió lineal és rellevant, ja que la regressió lineal és un dels mètodes més senzills i amplament utilitzats per modelar la relació entre dues variables. Això ens permet veure la millora que proporcionen els models proposats en comparació amb aquest model simple.

Per a simplificar els termes s'han utilitzat les abreviatures que veurem a la següent Taula 4.1.

<b>Abreviatures</b>	<b>Abreviatures</b>	<b>Significat</b>
<b>PREP</b>	<b>SNV</b>	Standard Normal Variate
	<b>SavGol</b>	Suavitzat Savitsky Golay
	<b>SavGol Diff</b>	Savitsky Golay Primera derivada
	<b>SavGol SNV</b>	Savitsky Golay + SNV
	<b>Prom</b>	Promig
	<b>Original</b>	Dades originals, sense transformació
<b>ICR</b>	<b>PCA</b>	Reducció PCA
	<b>ACC</b>	Reducció amb <i>autoencoder</i>
<b>MODELO</b>	<b>Modelo + Bag</b>	Model amb <i>Bagging</i>

**Taula 4.1:** Abreviatures que s'utilitzen per a mostrar el preprocessaments que ha emprat cada model

## 4.1. Macronutrients Primaris:

- Totes les bandes

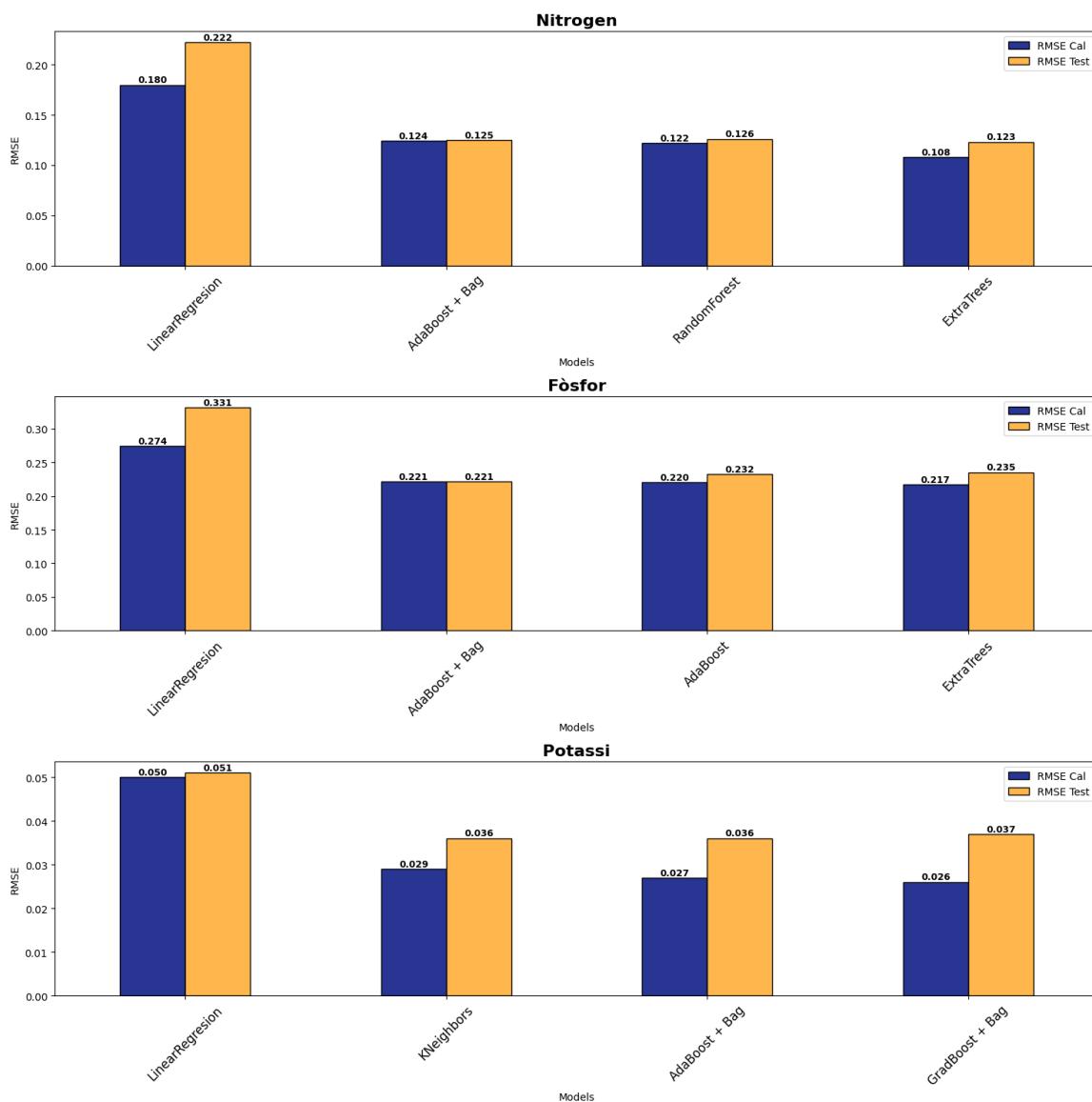
	MODEL	PREP	ICR	R <sup>2</sup> Cal	MAE Cal	RMSE Cal	R <sup>2</sup> Test	MAE Test	RMSE Test
N	Linear Regresion	SNV	PCA 5	0,584	0,132	0,1796	0,334	0,144	0,222
	<b>Extra Trees</b>	<b>Original</b>	<b>PCA 20</b>	<b>0,853</b>	<b>0,078</b>	<b>0,108</b>	<b>0,794</b>	<b>0,098</b>	<b>0,123</b>
	Ada Boost + Bag	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
	<b>Ada Boost + Bag</b>	<b>SavGol Diff</b>	<b>No</b>	<b>0,762</b>	<b>0,017</b>	<b>0,027</b>	<b>0,601</b>	<b>0,021</b>	<b>0,036</b>
	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
	<b>Ada Boost + Bag</b>	<b>SavGol Diff</b>	<b>No</b>	<b>0,651</b>	<b>0,140</b>	<b>0,221</b>	<b>0,672</b>	<b>0,144</b>	<b>0,221</b>
	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

**Taula 4.2.** Resultats dels models per als macronutrients primaris N, P, K. La columna PREP mostra la millor opció de preprocesament de les dades. La columna ICR mostra la millor opció d'enginyeria de característiques, indicant el nombre de components principals o emprant totes les bandes (No). En negreta es mostra la millor metodologia per a cada macronutrient primari sobre la base del valor de R<sup>2</sup> al conjunt de prova.

La Taula 4.2 presenta els resultats de l'estimació dels macronutrients primaris: nitrogen (N), fòsfor (P) i potassi (K). Es mostren diferents mètriques de rendiment dels models, incloent el coeficient de determinació R<sup>2</sup>, l'error absolut mitjà (MAE) i l'error quadràtic mitjà (RMSE) tant per al conjunt de calibració com per al conjunt de prova. Per al nutrient nitrogen, tots els models presenten una millora respecte a la regressió lineal, la qual cosa indica una millor capacitat d'estimació. La metodologia basada en *Extra Trees* utilitzant totes les bandes d'entrada presenta els millors resultats en el conjunt de prova per al nitrogen (R<sup>2</sup> de 0,794). Pel que fa al fòsfor, la millor metodologia és la primera derivada del filtre Savitzky-Golay utilitzant totes les bandes com a entrada, amb la combinació d'*AdaBoost* i *bagging* obtenint un R<sup>2</sup> de 0,60.

En general, les millors metodologies d'estimació per als macronutrients presenten coeficients de determinació superiors a 0,6, la qual cosa indica una bona capacitat d'estimació. Aquestes conclusions es mantenen quan s'analitzen altres mètriques de rendiment com el MAE i el RMSE. Els valors mínims dels errors d'estimació coincideixen amb els valors màxims de R<sup>2</sup>, demostrant la robustesa dels resultats en l'estimació del rendiment dels models.

Aquesta informació ens permet avaluar i comparar l'eficàcia dels models en l'estimació dels macronutrients primaris i concloure sobre la seva capacitat d'estimació i rendiment.



**Figura 4.1.** Comparació del RMSE obtingut calibració i prova dels diferents millors models que prediuen cada macronutrient primari.

La Figura 4.1 mostra, en un gràfic de barres, la comparació entre el RMSE dels conjunts de calibració i prova per a cada macronutrient primari amb cadascuna de les metodologies emprades. La pauta general és que l'error produït als conjunts de calibració i prova són molt similars (dins de cada metodologia). Aquest fet posa de manifest que no hi ha hagut sobreajust a la construcció dels models i que la capacitat de generalització dels models és óptima.

- Selecció de bandes

**Taula 4.3.** Resultats dels models per als macronutrients primaris N, P i K. La columna Bandes mostra el nombre de bandes seleccionades mitjançant Random Forest. En negreta es mostra la millor selecció en base del valor de  $R^2$  al conjunt de prova per a cada macronutrient.

	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
	<b>K-Neighbors</b>	<b>6</b>	<b>0,811</b>	<b>0,102</b>	<b>0,133</b>	<b>0,753</b>	<b>0,102</b>	<b>0,131</b>
	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
	<b>GradBoost + Bag</b>	<b>12</b>	<b>0,759</b>	<b>0,017</b>	<b>0,028</b>	<b>0,530</b>	<b>0,025</b>	<b>0,038</b>
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
	<b>Random Forest</b>	<b>12</b>	<b>0,654</b>	<b>0,147</b>	<b>0,232</b>	<b>0,621</b>	<b>0,155</b>	<b>0,243</b>

A la taula 4.3 es mostren els resultats dels models que estimen els macronutrients primaris amb una selecció de característiques mitjançant *Random Forest*. Els resultats es mostren a través dels índex de rendiment tal i com abans. Es pot observar que els resultats dels models empitjoren respecte als models que utilitzen totes les bandes per tal d'estimar els nivells dels macronutrients primaris.

Es pot observar com, en el cas del nitrogen, el model *K-Neighbors* que utilitza 6 bandes espectrals es el que millors resultats obté amb un  $R^2$  al conjunt de proves de 0,75. Aquest model, tal i com s'observa a la Taula 4.4, utilitza les bandes 590, 740, 730, 600, 560 i 530, que corresponen a la part verda-groga (700-740 nm) i verda-blava (530 – 600nm) de l'espectre visible. Aquestes dues parts de l'espectre, seran repetidament seleccionades com a les més importants per a estimar els macronutrients. Pel que fa al fòsfor, el millor model és el *Gradient Boosting* amb *bagging*, el qual obté un  $R^2$  de 0,53. Finalment, en el cas del potassi, amb un *Random Forest* i utilitzant les 12 bandes s'obté un  $R^2$  de 0,62.

Com podem veure, en el cas del fòsfor, els resultats no milloren gairebé un  $R^2$  de 0,5. En canvi, en el cas del nitrogen i el potassi, podem destacar uns resultats que superen un  $R^2$  de 0,6. Cal recordar que aquests models sols utilitzen un nombre limitat de bandes, cosa que causa que el procés d'adquisició siga molt més fàcil.

	BANDES MÉS IMPORTANTS
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.** Bandes seleccionades en els models anteriors per tal de predir cada nutrient segons els resultats obtinguts amb el Random Forest.

## 4.2. Macronutrients Secundaris

- Totes les bandes

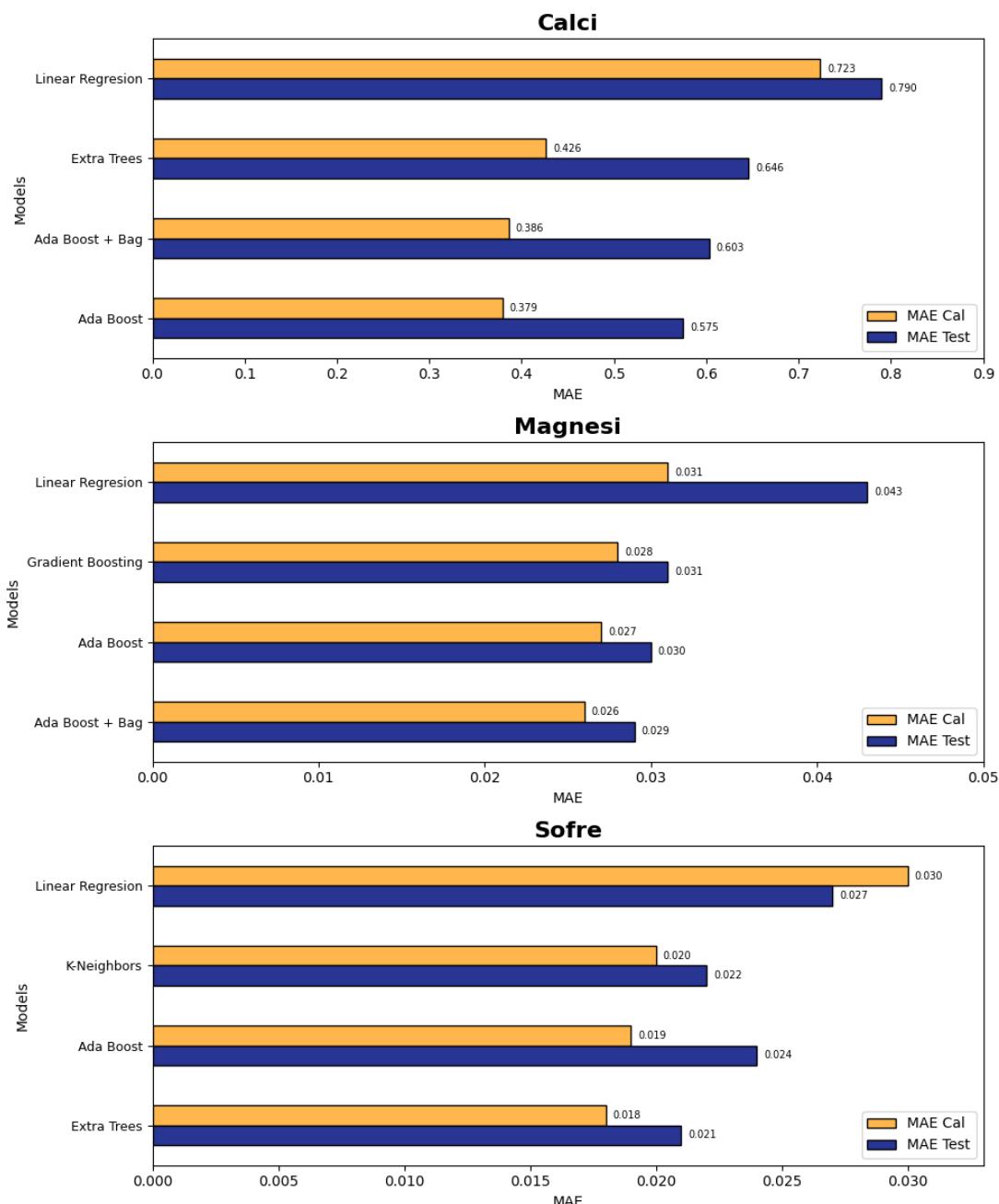
	MODEL	PREP	ICR	R <sup>2</sup> Cal	MAE Cal	RMSE Cal	R <sup>2</sup> Test	MAE Test	RMSE Test
Ca	Linear Regresion	SNV	PCA 5	0,339	0,723	1,214	0,528	0,790	1,123
	<b>Ada Boost</b>	<b>Original</b>	<b>PCA 15</b>	<b>0,769</b>	<b>0,379</b>	<b>0,686</b>	<b>0,645</b>	<b>0,575</b>	<b>0,974</b>
	Ada Boost + Bag	Original	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
	<b>Ada Boost + Bag</b>	<b>SavGol SNV</b>	<b>PCA 15</b>	<b>0,607</b>	<b>0,026</b>	<b>0,037</b>	<b>0,552</b>	<b>0,030</b>	<b>0,039</b>
	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
	<b>Extra Trees</b>	<b>SavGol Diff</b>	<b>No</b>	<b>0,661</b>	<b>0,018</b>	<b>0,027</b>	<b>0,670</b>	<b>0,021</b>	<b>0,028</b>
	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.** Resultats dels models per als macronutrients secundaris N, P, K. La columna PREP mostra la millor opció de preprocessament de les dades. La columna ICR mostra la millor opció d'enginyeria de característiques, indicant el nombre de components principals o emprant totes les bandes (No). En negreta es mostra la millor metodologia per a cada macronutrient secundari sobre la base del valor de R<sup>2</sup> al conjunt de prova.

A la Taula 4.5 observem els resultats relatius a l'estimació dels macronutrients secundaris Calcí, Magnesi i Sofre. Es mostren els mateixos índexs de rendiment que en els casos anteriors.

S'observa que, el calci amb 15 components de PCA i un estimador *Ada Boost* s'obtenen els millors resultats amb un coeficient de determinació R<sup>2</sup> de 0,64 al conjunt de prova. En el cas del magnesi, es el macronutrient secundari que majors dificultats presenta en l'estimació. Aquest nutrient es predit per una combinació d'*Ada Boost* i *Bagging* que, amb un preprocessament Savitsky Golay SNV i 15 components principals obté un 0,55 de R<sup>2</sup> al conjunt de proves. Per últim, en el sofre, obtenim els millors resultats amb un coeficient de determinació R<sup>2</sup> de 0,67 mitjançant l'estimador *Extra Trees* que utilitza un preprocessament de la primera derivada de Savinsky Golay.

Igual que amb els macronutrients primaris, es pot veure que els valors mínims de MAE i RMSE per a cada macronutrient coincideixen amb els valors màxims de R<sup>2</sup> la qual cosa indica que els models son robustos.



**Figura 4.2.** Comparació del MAE obtingut calibració i prova dels diferents millors models que predueixen cada macronutrient secundari.

La Figura 4.2 mostra, en un gràfic de barres, la comparació entre el MAE dels conjunts de calibració i prova per a cada macronutrient secundari en base a les metodologies comentades anteriorment. Veiem que al calci hi ha una lleugera diferència entre l'error als dos conjunts respecte als altres dos macronutrient secundaris, no obstant, els errors són molt similars en general. Al igual que en els macronutrients primaris, aquest fet demostra que els models no presenten sobreajust.

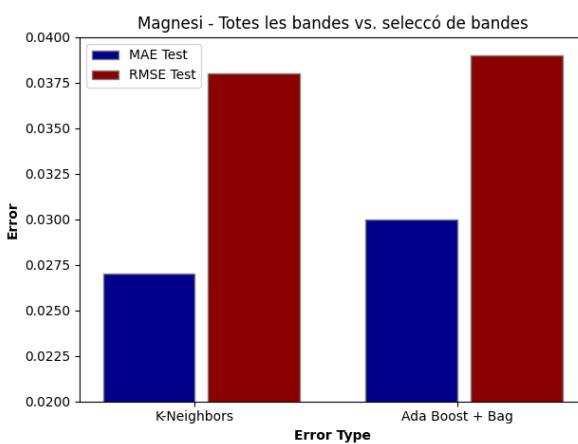
- Selecció de bandes

**Taula 4.6.** Resultats dels models per als macronutrients secundaris Ca, Mg i S. La columna Bandes mostra el nombre de bandes seleccionades mitjançant Random Forest. En negreta es mostra la millor selecció en base del valor de  $R^2$  al conjunt de prova per a cada macronutrient.

	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
	<b>PLS Regression</b>	<b>12</b>	<b>0,809</b>	<b>0,397</b>	<b>0,712</b>	<b>0,595</b>	<b>0,606</b>	<b>1,059</b>
Mg	Random Forest	3	0,658	0,025	0,037	0,547	0,029	0,040
	<b>K-Neighbors</b>	<b>6</b>	<b>0,613</b>	<b>0,025</b>	<b>0,037</b>	<b>0,573</b>	<b>0,027</b>	<b>0,038</b>
	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
	<b>Ada Boost + Bag</b>	<b>6</b>	<b>0,590</b>	<b>0,022</b>	<b>0,033</b>	<b>0,648</b>	<b>0,020</b>	<b>0,029</b>
	Linear SVR	12	0,607	0,024	0,034	0,653	0,022	0,030

A la taula 4.6 es mostren els resultats dels models que estimen els macronutrients secundaris amb una selecció de característiques mitjançant Random Forest.

En el cas del calci, es pot veure que la regressió PLS amb 12 bandes obté un  $R^2$  de 0,59. En aquest cas, els resultats empitjoren respecte als models que utilitzen totes les bandes com era el Ada Boost amb PCA de 15 components que obtenia un  $R^2$  de 645. En canvi, com es pot observar, el magnesi millora els resultats utilitzant una selecció de 6 bandes, amb un coeficient  $R^2$  de 0,57 en comparació amb el 0,55 obtingut amb l'utilització de totes les bandes. Els errors de MAE i RMSE també ho demostren tal i com es veu a la següent Figura 4.3.



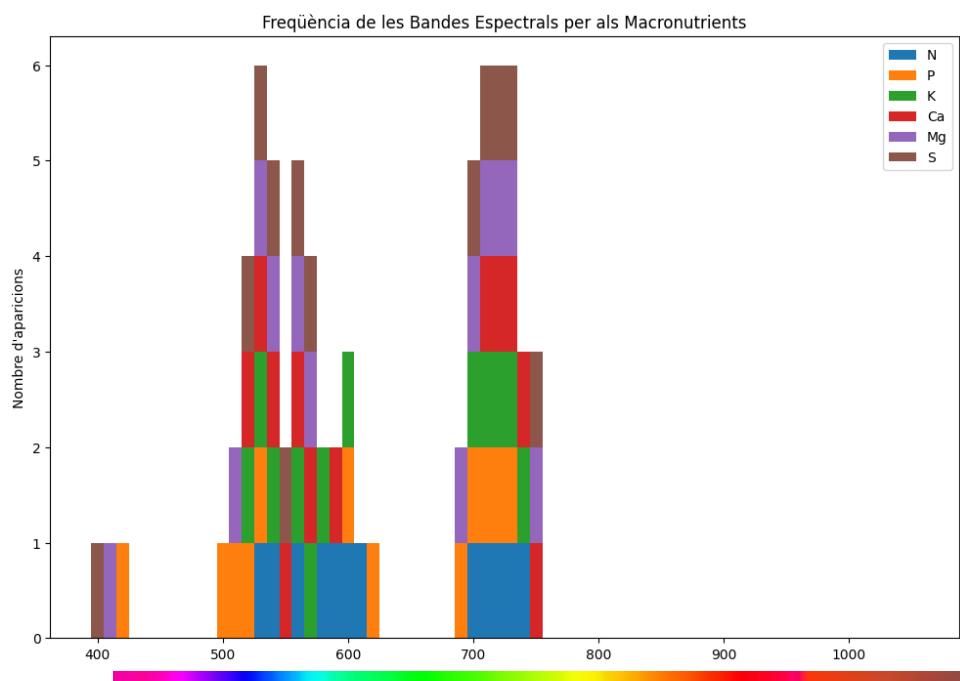
**Figura 4.3.** Comparació del MAE i RMSE obtingut al conjunt de prova del model K-Neighbours que utilitza una selecció de 6 bandes i el model Ada Boost + Bagging que utilitza totes les bandes.

Per últim, el model de 6 bandes *d'Ada Boost i bagging*, obté un coeficient de  $R^2$  de 0,648 per al sofre el qual no consegueix millorar el rendiment que el model anterior que utilitzava totes les bandes.

	BANDES MÉS IMPORTANTS
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.

Com s'observa a la Taula 4.7, la selecció de bandes més importants per als macronutrients secundaris, al igual que els primaris, també es centra en les dues parts comentades anteriorment de l'espectre electromagnètic.



**Figura 4.4.** Bandes més freqüentment classificades com a importants segons el Random Forest per a cada macronutrient que ha sigut estimat.

Notem a la Figura 4.4 com hi ha dos grans zones espectrals que resulten d'utilitat per tal d'estimar el nivell dels macronutrients. Aquestes son des de 500 fins a 600 nm i des de 690 fins a 760 nm.

## 4.3. Micronutrients

- Totes les bandes

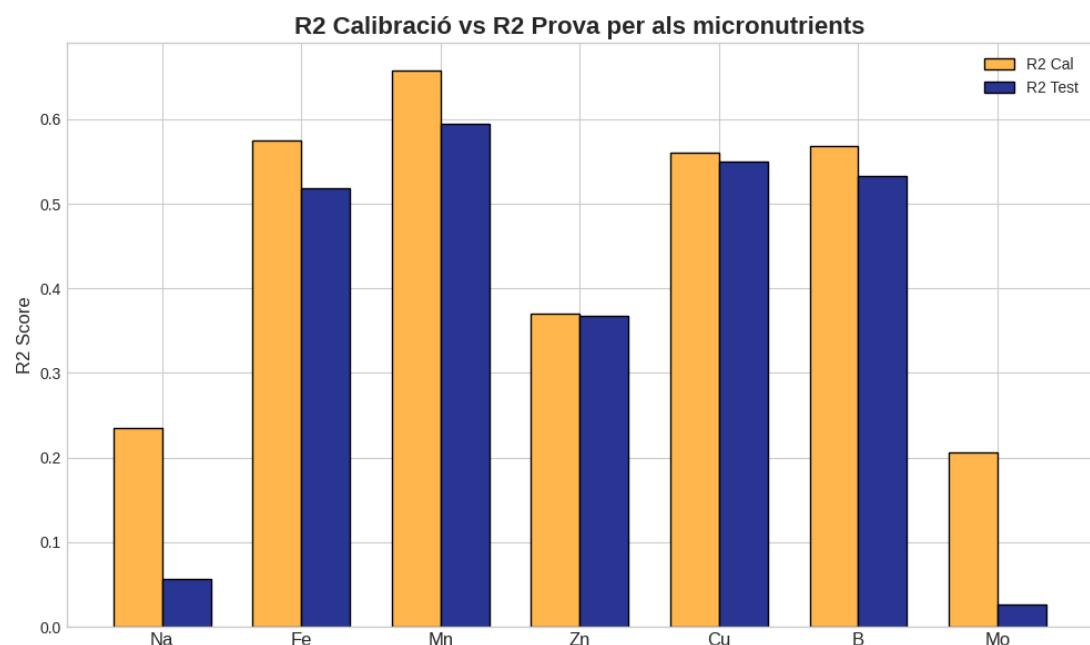
	MODEL	PREP	ICR	R <sup>2</sup> Cal	MAE Cal	RMSE Cal	R <sup>2</sup> Test	MAE Test	RMSE Test
Na	Linear Regresion	SNV	PCA 5	0,094	0,014	0,018	0,002	0,018	0,024
	<b>SVR</b>	<b>Original</b>	<b>PCA 12</b>	<b>0,235</b>	<b>0,012</b>	<b>0,015</b>	<b>0,057</b>	<b>0,013</b>	<b>0,017</b>
	Linear SVR	SavGol Diff	PCA 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
	<b>Extra Trees</b>	<b>SavGol Diff</b>	<b>No</b>	<b>0,575</b>	<b>11,973</b>	<b>15,091</b>	<b>0,518</b>	<b>14,298</b>	<b>17,917</b>
	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
	<b>Ada Boost + Bag</b>	<b>Original</b>	<b>PCA 20</b>	<b>0,657</b>	<b>3,202</b>	<b>4,629</b>	<b>0,594</b>	<b>4,128</b>	<b>5,480</b>
	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
	<b>Extra Trees</b>	<b>SavGol SNV</b>	<b>PCA 20</b>	<b>0,370</b>	<b>5,757</b>	<b>7,726</b>	<b>0,367</b>	<b>5,986</b>	<b>7,607</b>
	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
	<b>AdaBoost + Bag</b>	<b>SNV</b>	<b>PCA 15</b>	<b>0,560</b>	<b>11,364</b>	<b>14,129</b>	<b>0,550</b>	<b>11,378</b>	<b>15,080</b>
	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
	<b>GradBoost + Bag</b>	<b>SavGol Diff</b>	<b>PCA 9</b>	<b>0,568</b>	<b>4,596</b>	<b>6,079</b>	<b>0,533</b>	<b>4,945</b>	<b>6,692</b>
	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
	<b>AdaBoost</b>	<b>Original</b>	<b>PCA 25</b>	<b>0,206</b>	<b>0,030</b>	<b>0,038</b>	<b>0,026</b>	<b>0,034</b>	<b>0,042</b>
	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.** Resultats dels models per als micronutrients Na, Fe, Mn, Zn, Cu, B i Mo. La columna PREP mostra la millor opció de preprocesament de les dades. La columna ICR mostra la millor opció d'enginyeria de característiques, indicant el nombre de components principals o emprant totes les bandes (No). En negreta es mostra la millor metodologia per a cada micronutrient sobre la base del valor de R<sup>2</sup> al conjunt de prova.

A la Taula 4.8 s'observa que, en el cas del sodi, utilitzant totes les bandes, aconseguim amb SVR els millors resultats, encara que notem que és un micronutrient de difícil predicción amb dades espectrals ja que, amb PCA de 12 components aconseguim un R<sup>2</sup> de 0,0518 al conjunt de prova. En el cas

del ferro es pot observar que amb la primera derivada Savinsky Golay s'obté en *test* un  $R^2$  de 0,518. El manganés és el micronutrient en el que millors resultats s'aconsegueixen, obtenint un  $R^2$  de 0,59 al conjunt de prova utilitzant un estimador *Ada Boost* amb *bagging* i 20 components principals. En el cas del Zinc, *l'Extra Trees* amb Savinsky Golay, SNV i 20 components principals, obté un  $R^2$  de 0,36. Seguidament, el coure aconsegueix el segon millor  $R^2$  al conjunt de prova amb un 0,55 utilitzant *Ada Boost* i *bagging*. De preprocessament, aquest model utilitza SNV i 15 components principals. En el cas del bor, aconseguim un  $R^2$  de 0,53 al conjunt de prova utilitzant d'estimador *Gradient Boosting* amb *bagging* i un preprocessament primera derivada de Savinsky Golay i 9 components principals. Finalment, amb el molibdè, aconseguim els pitjors resultats amb un  $R^2$  de 0,02 al conjunt de prova utilitzant *Ada Boost*.

Tal i com s'ha vist anteriorment amb els macronutrients, els valors mínims de MAE i RMSE tornen a coincidir amb els valors màxims de  $R^2$ . Ja hem dit que aquest fet indica que els models son robustos.



**Figura 4.5.** Gràfic de barres de  $R^2$  obtingut pel model que ha aconseguit un millor rendiment al conjunt de prova per cada micronutrient: Na, Fe, Mn, Zn, Cu, B i Mo.

S'observen la Figura 4.5 els  $R^2$  dels millors models obtinguts estimant els micronutrients. Es pot apreciar que no hi ha cap model que supere un 0,6 de  $R^2$  al conjunt de proves, cosa que, excepte en el cas del magnesi, tots els macronutrients han superat. Aquest fet ens indica que els micronutrients, que estan presents en menors quantitats en les fulles del cítric, son de més difícil estimació.

No obstant, veiem que els conjunts calibració i *test* son prou similars en els models dels nutrients que superen un  $R^2$  de 0,5, cosa que ens indica que els models no es sobreajusten i, per tant, la capacitat de generalització és òptima.

- Selecció de bandes

	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
	<b>K-Neighbors</b>	<b>6</b>	<b>0,381</b>	<b>0,010</b>	<b>0,014</b>	<b>0,055</b>	<b>0,013</b>	<b>0,018</b>
	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
	<b>Linear SVR</b>	<b>12</b>	<b>0,430</b>	<b>13,167</b>	<b>17,202</b>	<b>0,438</b>	<b>15,056</b>	<b>18,875</b>
Mn	Ridge Regresion	3	0,522	4,076	5,664	0,460	4,325	5,820
	<b>Ada Boost</b>	<b>6</b>	<b>0,590</b>	<b>3,614</b>	<b>5,041</b>	<b>0,578</b>	<b>4,003</b>	<b>5,409</b>
	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
	<b>Extra Trees</b>	<b>6</b>	<b>0,216</b>	<b>6,930</b>	<b>8,638</b>	<b>0,225</b>	<b>7,590</b>	<b>9,794</b>
	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
	<b>Nu SVR</b>	<b>12</b>	<b>0,551</b>	<b>10,984</b>	<b>17,164</b>	<b>0,475</b>	<b>13,215</b>	<b>18,838</b>
B	SVR	3	0,498	4,981	6,874	0,533	4,912	6,559
	<b>K-Neighbors</b>	<b>6</b>	<b>0,525</b>	<b>4,536</b>	<b>6,506</b>	<b>0,544</b>	<b>4,620</b>	<b>6,487</b>
	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
	<b>Ridge Regresion</b>	<b>12</b>	<b>0,073</b>	<b>0,035</b>	<b>0,043</b>	<b>0,010</b>	<b>0,036</b>	<b>0,043</b>

**Taula 4.9.** Resultats dels models per als micronutrients Na, Fe, Mn, Zn, Cu, B i Mo. La columna Bandes mostra el nombre de bandes seleccionades mitjançant Random Forest. En negreta es mostra la millor selecció en base del valor de  $R^2$  al conjunt de prova per a cada macronutrient.

Es pot veure a la Taula 4.9 en el cas de la selecció de bandes, que els per al sodi i per al molibdè, s'obtenen valors de coeficient de determinació molt baixos, amb un  $R^2$  al conjunt de prova inferiors a 0,1. Posteriorment, s'observa que en el cas del zinc el  $R^2$  de *test* es de 0,225, cosa que ens indica que el model, al igual que utilitzant totes les bandes, tampoc aconsegueix sobrepassar un  $R^2$  de 0,5.

El mateix ocorre amb el coure i el ferro, on s'observa un  $R^2$  de 0,475 i 0,438 respectivament al conjunt de proves. Per altra banda tenim el manganés, on els  $R^2$  de prova si que supera el 0,5 amb un 0,578, que s'apropa al  $R^2$  del model

que utilitzava totes les bandes ( $R^2$  test de 0,594). Per últim, es pot apreciar que el bor millora els resultats ( $R^2$  test de 0,544) respecte a quan s'utilitzen totes les bandes ( $R^2$  test de 0,533).

## **5. CONCLUSIONS I FUTURA PROJECCIÓ**

La metodologia proposada, basada en l'ús d'un sistema de visió hiperspectral Vis-NIR i tècniques de regressió d'aprenentatge automàtic, permet estimar el contingut del nitrogen amb un  $R^2$  de 0,79. Per al potassi, sofre i calci, es van obtenir valors de  $R^2$  de 0,67, 0,67 i 0,64 respectivament. En el cas del fòsfor, manganés i magnesi s'ha obtingut uns  $R^2$  de 0,60, 0,59 i 0,57 respectivament al conjunt de proves. Seguidament, s'ha obtingut un  $R^2$  de 0,55, 0,54 i 0,51 per al coure, bor i ferro respectivament. Ja per davall de  $R^2$  de 0,5 podem trobar el zinc, amb un  $R^2$  de 0,36. Per últim, en el sodi i molibdè s'han obtingut  $R^2$  de 0,05 i 0,02 respectivament.

Es comprova que, en el cas dels macronutrients, l'estimació és molt més precisa amb un  $R^2$  mitjà de 0,66, en comparació amb els micronutrients on el  $R^2$  mitjà és de 0,48. Es troben especials dificultats per tal de predir el sodi i el molibdè. En canvi, podem trobar el nitrogen, fòsfor, potassi, calci i sofre amb un  $R^2$  superior a 0,6, cosa que ens indica, que es possible realitzar aquesta estimació.

A més, trobem que molts models on es seleccionen bandes poden ser de molta utilitat poder tal d'estimar molts dels nutrients estudiat ja que, el procés d'adquisició podria ser molt menys costos.

La utilització d'un sistema de visió hiperspectral Vis-NIR i els models de regressió d'aprenentatge automàtic s'ha evidenciat com una alternativa útil per estimar els nivells de nutrients a les fulles de cítrics. Aquests mètodes ofereixen una solució no destructiva, més econòmica, ràpida i precisa en comparació amb les tècniques destructives, costoses i lentes, complint així el nostre objectiu principal deficiència i reducció de l'impacte ambiental.

Per millorar, en el futur, seria beneficiós reunir més dades que cobreixin un rang més ampli, amb una major varietat de condicions ambientals i nivells de nutrients. A més, la incorporació d'altres formes de dades, com podrien ser dades meteorològiques i propietats del sòl, podria millorar encara més la precisió dels models sobretot en aquells nutrients on la estimació ha sigut més tediosa.

Pel que fa a les tècniques d'aprenentatge automàtic, explorar mètodes més avançats com l'aprenentatge profund podria ser una via interessant per a futures investigacions. Es coneix que aquests models són més eficients a l'hora de manejar dades d'alta dimensió i poden millorar la precisió dels models de predicción nutricional.

En el camp de l'enginyeria agrícola, com hem comentat anteriorment, és de vital importància conèixer els dèficits d'aquests nutrients per poder millorar la producció. Aleshores, aquests models suposen una poderosa eina per a conèixer de millor manera el estat dels arbres. Un punt fonamental és la investigació de com integrar aquests models per a que puguen ser de fàcil ús per a els agricultors; podent saber, amb un cost reduït i amb una funcionalitat simple, quin és l'estat dels arbres. En última instància, això pot conduir a millors rendiments dels cultius, una millora de la qualitat dels cultius i un impacte ambiental reduït.

A més, és important tenir en compte la infraestructura i els recursos disponibles per als agricultors, especialment a les zones rurals on l'accés a la tecnologia i l'experiència pot ser limitat. Els models s'han de dissenyar per funcionar amb dispositius àmpliament utilitzats, com ara telèfons intel·ligents, i s'ha de proporcionar suport per ajudar els agricultors a interpretar els resultats i prendre decisions informades.

## **6. BIBLIOGRAFIA**

- 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](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/](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](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](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/](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. Bibliografia

---

- 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](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](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](https://doi.org/10.1007/978-3-030-053185_1)