POLITECNICO DI TORINO

Master's Degree in Mathematical Engineering

Master's Degree Thesis

Riduzione della dimensionalità spettroscopica per la classificazione di piante di insalata sotto stress idrico tramite CARS e MWPLS-DA

Relatori

Prof. Renato FERRERO

Dr. Nicola DILILLO

Candidato

Andrea SANNA

Luglio 2024

Abstract

L'agricoltura è da sempre fondamentale per l'uomo e numerosi fattori, come l'aumento della popolazione, i cambiamenti climatici e gli eventi geo-politici, rendono sempre più necessaria un'ottimizzazione della produzione agricola. Con l'agricoltura di precisione si è riusciti ad aumentare la resa della produzione e contemporaneamente a ridurre gli sprechi tramite l'uso di vari strumenti, come il telerilevamento, l'IoT, la computer vision e l'analisi multi-spettrale.

Lo studio effettuato approfondisce l'uso della spettroscopia, la disciplina che studia le radiazioni elettromagnetiche, per monitorare lo stato di salute delle piante. Il principio su cui ci si basa è che piante sane e sotto stress hanno differenze nei loro spettri e, tramite questi ultimi, è possibile rilevare lo stato di stress prima che questo sia visibile ad occhio nudo, in modo da poter intervenire di conseguenza tempestivamente. Utilizzare i dati spettroscopici direttamente può però essere complicato perché gli spettri contengono solitamente centinaia di variabili (corrispondenti alle lunghezze d'onda misurate).
Questo lavoro ha lo scopo di individuare un numero ridotto di lunghezze d'onda che permetta di effettuare la classificazione in modo più agevole, ma senza non comprometterne l'accuratezza. In particolare sono state considerate delle piante di insalata per effettuare una classificazione fra piante in salute o sottoposte a stress idrico, utilizzando diverse tecniche di pre-processing e applicando due tecniche per selezionare le lunghezze d'onda più utili nella classificazione: la Competitive Adaptive Reweight Sampling (CARS) e la Moving Windows Partial Least Square - Discriminant Analisys (MWPLS-DA), entrambe basate sulla Partial Least Square (PLS). Nello specifico CARS, tramite l'unione del campionamento statistico e mimando il principio evoluzionistico di selezione del più forte, individua un numero ridotto di lunghezze d'onda per effettuare la classificazione, mentre la MWPLS-DA, applicando la PLS tramite finestre mobili su tutto lo spettro, identifica delle regioni continue da usare nella classificazione.

Indice

Elenco delle tabelle

Elenco delle figure

Acronimi

CARS

Competitive Adaptive Reweight Sampling

PCA

Principal Component Analysis

PC

Principal Component

PLS

Partial Least Square

PLS-DA

Partial Least Square - Discriminant Analysis

MWPLS-DA

Moving Window Partial Least Square - Discriminant Analysis

RMSE

Root Mean Squared Error

Capitolo 1

Introduzione

1.1 Motivazioni

Le piante sono da sempre state risorse fondamentali per l'uomo, infatti offrono numerosi benefici per il corpo umano grazie ai micronutrienti e macronutrienti (vitamine, minerali, proteine, etc.) che esse contengono. [1] L'agricoltura è quindi sempre stata importante della storia umana e continuerà ad esserlo: si stima che entro il 2050 la popolazione dovrebbe raggiungere e anche superare i 9 miliardi e, associato a ciò, è previsto anche un aumento dei consumi alimentari: la FAO (Food and Agriculture Organization, agenzia delle Nazioni Unite) stima che entro il 2050 sarà richiesto, per esempio, oltre un miliardo di tonnellate aggiuntive di cereali e, più in generale, un aumento della richiesta di prodotti alimentari di circa il 60% rispetto alla media annuale registrata fra il 2005 e il 2007. Nonostante i miglioramenti della produzione agricola, eventi recenti come la pandemia di COVID-19 e la guerra in Ucraina hanno comunque messo in evidenza la fragilità della sicurezza alimentare globale.[1] La qualità dei cibi ottenuti dalle piante dipende da numerosi fattori durante il loro periodo di crescita: fattori di stress biotici (come funghi e insetti) e abiotici (come temperature estreme e siccità) giocano un ruolo cruciale nella perdita di resa e qualità, ma sono naturalmente presenti durante la crescita della pianta e, di conseguenza, potrebbero portare ad alterazioni durante la sua crescita. [1] Di conseguenza, l'individuazione precoce di problematiche relative alla crescita della pianta è cruciale nell'evitare perdite della produzione agricola e per questo motivo è fondamentale tenere traccia della salute delle piante. [2] In particolare, l'acqua ha un ruolo fondamentale in agricoltura: questo è infatti il settore economico più sensibile rispetto alla scarsità d'acqua, basti pensare che è responsabile di circa il 70% dei prelievi di acqua dolce a livello globale. Nel report della FAO "Coping with water scarcity: An action framework for agriculture and food security" ci si chiede se ci sarà acqua sufficiente per produrre abbastanza cibo

per la crescente popolazione dei 50 anni seguenti: la risposta che viene data è che i metodi attuali di produzione porteranno a crisi idriche in molte parti del mondo e solo migliorando l'uso dell'acqua in agricoltura si riuscirà a fronteggiare le sfide dovute alla sua scarsità dei prossimi decenni.

1.2 Agricoltura di precisione

Nel secolo scorso, varie regioni del mondo stavano già affrontando il problema della malnutrizione e della carestia; inoltre, la rapida crescita della popolazione non fece che acuire le preoccupazioni legate all'insicurezza alimentare. Per questi motivi, grazie anche agli avanzamenti tecnologici, negli anni Quaranta ebbe inizio quella che venne definita Rivoluzione Verde, che trasformò la produzione agricola globale. Questa, attraverso la ricerca scientifica, permise di sviluppare nuove colture più resistenti e con un rendimento più elevato rispetto a quelle tradizionali. Oltre a queste, vennero introdotte nuove tecniche di irrigazione, che permisero di fornire un apporto costante di acqua alle piante (non dipendendo più dalle precipitazioni), e si iniziò a diffondere l'uso di fertilizzanti chimici e pesticidi.
Negli anni successivi alla rivoluzione verde, l'agricoltura ha continuato ad evolversi; con gli obiettivi di migliorare la gestione agricola, la produzione che da essa deriva e mitigare alcuni dei problemi sorti con la Rivoluzione Verde (come l'inquinamento dovuto alle sostanze usate nella coltivazione), a metà degli anni Ottanta ha iniziato a svilupparsi l'agricoltura di precisione [3]: è possibile definirla come la disciplina scientifica che ha l'obiettivo di migliorare la produzione agricola assistendo il management tramite l'uso di informazioni ottenute da strumenti di analisi e sensori tecnologici e riguarda sia l'acquisizione e l'analisi dei dati, sia gli avanzamenti tecnologici necessari per effettuarle[4] ed è stata considerata una fra le dieci rivoluzioni più importanti in ambito agricolo [5]. La disciplina generalmente riguarda una migliore gestione degli input agricoli come erbicidi, fertilizzanti, semi, carburante, grazie a procedure di controllo più precise: se nell'agricoltura tradizionale gli input vengono usati in modo uniforme, grazie all'agricoltura di precisione è possibile adeguare le quantità di input alle reali necessità, ottimizzandone l'uso e riducendo anche l'impatto ambientale, grazie al minor uso di prodotti superflui (come fertilizzanti, pesticidi, benzina).[3] Ciò ha ovviamente risvolti positivi in termini di profitto, grazie a una maggiore produttività, au una maggiore qualità dei raccolti, alla riduzione degli sprechi e dei costi associati. Data la forte componente tecnologica, fra i benefici dello sviluppo di questo settore, ci sono anche quelli sociali legati agli avanzamenti tecnologici e alla creazione di posti di lavoro in ambiti hardware, software, sistemi di supporto alle decisioni, information management.

Se in passato il modo principale con cui venivano identificate le problematiche

delle piante è stata l'ispezione visiva, dispendiosa in termini di tempo e di manodopera, oltre a essere poco obiettiva,[2], oggi la produzione agricola si affida al monitoraggio del raccolto attraverso l'osservazione e la misurazione di variabili riguardanti lo stato delle piante, del suolo, l'irrigazione, l'uso di pesticidi e fertilizzanti. Fra i diversi approcci per farlo, che possono essere molto diversi fra loro, un supporto prezioso è dato dalle implementazioni di **telerilevamento**, in inglese **remote sensing**, sulle quali ci concentreremo.

Con telerilevamento vengono indicate tecniche usate per ottenere informazioni da un oggetto senza contatto fisico, ma attraverso la misurazione dell'energia elettromagnetica, riflessa o emessa dalle piante o dal suolo, o anche della fluorescenza o dell'energia termica. [3] [6] È possibile classificare le tecniche di telerilevamento in base a diversi criteri.

- Tipo di piattaforma usata dal sensore. Le prime applicazioni di telerilevamento riguardavano le rilevazioni effettuate coi satelliti, che con gli avanzamenti tecnologici hanno migliorato notevolmente la risoluzione delle rilevazione, successivamente si sono sviluppati sensori per rilevazioni aeree (con aerei e droni) e sensori al livello del terreno.

- Fonte di luce usata. È possibile distinguere i sensori fra attivi e passivi a seconda che, rispettivamente, emettano la radiazione che usano per i rilevamenti oppure usino una fonte esterna, per esempio la luce solare.[6]

- Tipologia di dati. È possibile distinguere i sensori in due classi in base alla tipologia dei dati prodotti: sensori a immagini e no. I sensori a immagini, detti imaging sensor, producono fotografie o mappe visive e sono particolarmente utili nel monitoraggio della crescita delle piante e nell'identificazione di problemi specifici, come malattie o stati di stress. Sono sensori di questo tipo le camere multi-spettrali e iper-spettrali, ossia camere che catturano immagini in diverse bande dello spettro, che possono andare da 2-3 fino a oltre una decina, nel caso delle prime, oppure a diverse centinaia di bande molto strette nel caso delle seconde. Per quanto invece riguarda i sensori non a immagini, detti non-imaging sensor, sono sensori che forniscono dati in formato numerico e vengono utilizzati per avere dati quantitativi sulle condizioni ambientali e del suolo, come ad esempio i sensori di temperatura, umidità o dei nutrienti presenti nel suolo, utili per monitorare le condizioni della produzione.

1.3 Spettroscopia

La spettroscopia è la disciplina che studia l'interazione della radiazione elettromagnetica con i materiali per studiarne le proprietà. Il suo oggetto di studio sono gli spettri, ovvero le distribuzioni dell'energia elettromagnetica in funzione delle lunghezza d'onda misurata; a seconda della tipologia della radiazione studiata si ottengono informazioni diverse sull'oggetto di interesse ed è per questo utilizzata in numerosi ambiti (ambientale, scienze dei materiali, medicina, ecc.). A seconda della lunghezza d'onda considerate è possibile dividere lo spettro in regioni: a partire da quelle con frequenza minore si hanno nell'ordine onde radio, microonde, l'infrarosso (indicato con IR), luce visibile (che è appunto quella visibile dall'occhio umano, indicata con VIS), ultravioletto, raggi X e raggi gamma. In particolare, per le analisi effettuate è di interesse la regione della luce visibile, che va da circa 380 nm a circa 750 nm, e nella regione dell'infrarosso, che va da 750 nm a 1000μm, la regione dell' infrarosso vicino, da 750nm a 2500nm. È possibile vedere la divisione dello spettro elettromagnetico nella Figura 1.1

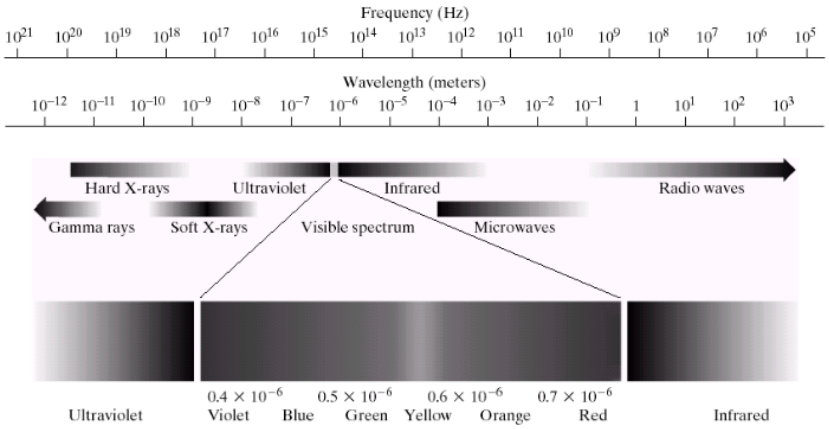

Figura 1.1: Divisione dello spettro elettromagnetico al variare della frequenza e della lunghezza d'onda

In generale, un fascio di luce emesso da una sorgente si propaga finché non colpisce un corpo e può essere scomposto in tre componenti: la luce che rimbalza sul materiale, detta riflessa, quella che passa attraverso, detta trasmessa, e quella che viene catturata dal materiale, detta assorbita; a seconda delle lunghezza d'onda che vengono riflesse o trasmesse da un corpo è possibile ottenere informazioni sulla sua composizione chimica e altre proprietà.

Per quanto riguarda l'agricoltura di precisione e, nello specifico, lo studio dello stato di salute delle piante, nonostante le foglie possano essere piuttosto diverse fra loro, sia in termini di forma che di concentrazione di acqua e nutrienti nello spazio

intercellulare, portando a risultati diversi per quanto riguarda la loro riflettanza, ma diversi studi [3] hanno comunque osservato che le piante che sottoposte a stress, dovuto per esempio alla mancanza di acqua o nutrienti, avevano dei comportamenti in comune e ciò si rifletteva in differenze negli spettri delle stesse piante fra quando sono sane o in una situazione di stress. Gli spettri delle foglie raccolti in laboratorio o sul campo vengono quindi usati per determinare le regioni spettrali o le lunghezze d'onda che possono essere usati nella rilevazione di malattie o stati di stress. In generale, la quantità di radiazioni riflesse dalla pianta è inversamente proporzionale alla quantità di radiazioni assorbite dai pigmenti e varia a seconda della lunghezza d'onda della radiazione incidente [3]; per le piante sotto stress è stato osservato un aumento della propria riflettanza nella regione dello spettro visibile e una diminuzione in quella del vicino infrarosso, a causa di una forte riduzione dell'assorbimento della luce usata per la fotosintesi dalla foglia [6]. Una pianta sana, al contrario, assorbe fortemente la luce rossa durante la fotosintesi e riflette maggiormente nell'infrarosso. Più precisamente, quando la luce colpisce una foglia in salute, è stata osservata una bassa riflettanza nella parte visibile dello spettro, con un picco nella regione del verde (circa 550 nm, che motiva la colorazione verde delle piante percepita dall'occhio umano), dovuta al forte assorbimento di questa porzione dello spettro da parte dei pigmenti interni alla foglia, mentre si ha invece un aumento della riflettanza nell'infrarosso (oltre i 700 nm). [7] [3] Un tipico spettro di riflettanza è quello nella Figura 1.2. Un limite di questo approccio è che gli studi disponibili riguardano colture specifiche e non possono essere generalizzati ad altre riuscendo ad ottenere un accuratezza simile. [6]

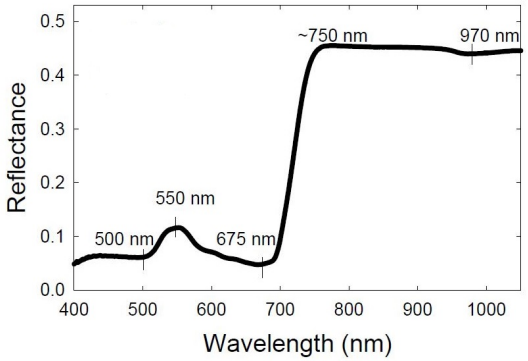

Figura 1.2: Esempio dello spettro di riflettanza di una pianta al variare delle lunghezze d'onda

Capitolo 2

Background

2.1 Preprocessing

Gli spettri di sistemi biologici eterogenei, come cellule e tessuti, che sono costituiti da un gran numero di bio-molecole (come nel caso delle piante) sono di base complessi; inoltre le differenze presenti fra diversi campioni in diverse condizioni patologiche sono abbastanza ridotte e difficili da osservare negli spettri "grezzi". Per queste ragioni, per riuscire ad ottenere informazioni utili e comprendere meglio i processi studiati è fondamentale processare e analizzare i dati. Ciascuno spettro è formato da migliaia di variabili, che corrispondono alle misurazioni effettuate nelle diverse lunghezze d'onda: lo studio delle misure spettroscopiche è quindi associato a tecniche di analisi multivariate, che permettono l'analisi di più variabili contemporaneamente, con l'obiettivo di comprendere la relazione fra di esse. Grazie all'analisi multivariata è possibile modellare le informazioni dei dati in modo da poterle utilizzare successivamente per fare predizioni su dati della stessa tipologia. Il preprocessing è la fase nella quali i dati grezzi vengono puliti e trasformati per l'analisi successiva ed è, in generale, fondamentale per riuscire ad ottenere una buona analisi, eliminando informazioni superflue, dovute magari al rumore, e per rendere confrontabili grandezze diverse, che potrebbero influenzare negativamente l'uso di alcune tecniche. In spettroscopia serve per eliminare dagli spettri gli effetti di segnali indesiderati rendere più evidenti le differenze fra i diversi campioni, ovvero le caratteristiche utili all'analisi [8]. Per questo motivo vengono utilizzate diverse trasformazioni dei dati, necessarie per rendere più omogenee e confrontabili le osservazioni. Più precisamente si considera il problema per il quale, nonostante le misurazioni avvengano nelle stesse condizioni sperimentali, queste siano inevitabilmente influenzate da fattori esterni a quelli di interesse: l'obiettivo di queste tecniche è rimuovere, o almeno ridurre,le differenze fra osservazioni dovute a fattori esterni, come l'illuminazione o le caratteristiche dello strumento utilizzato, cercando

di preservare quelle dovute al fenomeno di interesse.

2.1.1 Multiplicative Scatter Correction (MSC)

Due importanti fonti di inaccuratezza nella campo della spettroscopia di riflettanza sono la dispersione della luce (light scatter) e la non linearità della risposta dello strumento; in casi estremi è possibile arrivare ad attribuire quasi il 99% della varianza dello spettro riflesso allo scatter noise sistematico. [9] Fra le tecniche presentate, la Multiplicative Scatter Correction (MSC) è una tecnica di preprocessing specifica della spettroscopia, che cerca di occuparsi anche di questo problema. In modo più pratico, gli obiettivi della tecnica sono la linearizzazione degli spettri e la riduzione della varianza del rumore. La considerazione alla base della tecnica è che la diffusione della luce a diverse lunghezze d'onda è diversa rispetto all'assorbimento della luce causato dalla composizione chimica: utilizzando i dati raccolti a diverse lunghezze d'onda è quindi possibile distinguere fra dispersione e assorbimento nello spettro. Nella pratica ciò che la MSC fa è stimare la dispersione di ogni campione rispetto a un campione ideale, correggendo quindi ogni spettro in modo che abbia la stessa dispersione del campione ideale.

Negli esperimenti effettuati come campione ideale è stato utilizzato lo spettro mediano calcolato fra tutti quelli del dataset ed è stato considerato un modello lineare per ogni campione. Indicizzando con $i = 1,2,\ldots,N$ i campioni e con $k = 1,2,\ldots,K$ le lunghezze d'onda, l'equazione usata è:

$$x_{ik} = a_i + b_i \bar{x}_k + e_{ik}$$

dove $\bar{x}_k$ rappresenta lo spettro del campione ideale nella lunghezza d'onda k, x_{ik} è l'osservazione considerata, a_i e b_i sono i coefficienti del modello lineare stimati per ogni campione i e rappresentano rispettivamente la componente additiva e moltiplicativa dell'errore, mentre e_{ik} è il residuo del modello, nel quale ci si aspetta sia contenuta l'informazione non modellabile con solo delle costanti additive e moltiplicative, ovvero l'informazione effettivamente di interesse.

È possibile stimare i coefficienti a_i e b_i tramite una regressione ai minimi quadrati dello spettro osservato X rispetto allo spettro ideale $\bar{X}$, effettuata su tutte le lunghezze d'onda k disponibili. Indicando con $\hat{a}_i$, $\hat{b}_i$ e $\hat{e}_{ik}$ le stime ottenute per uno specifico spettro, è possibile calcolare la correzione per questo spettro come:

$$x^*_{ik} = \frac{x_{ik} - \hat{a}_i}{\hat{b}_i} = \bar{x}_k + \frac{\hat{e}_{ik}}{\hat{b}_i}$$

In pratica ciò che l' MSC attua è una traslazione verticale, in modo da ottenere un' intercetta nulla, e una rotazione della retta fino a raggiungere la pendenza di quella ideale: in questo modo gli spettri avranno intercetta e pendenza pari a quelle dello

spettro ideale, mentre le informazioni specifiche rimangono approssimativamente invariate, a parte un riscalamento. È possibile vedere un esempio di ciò nella Figura 2.3.

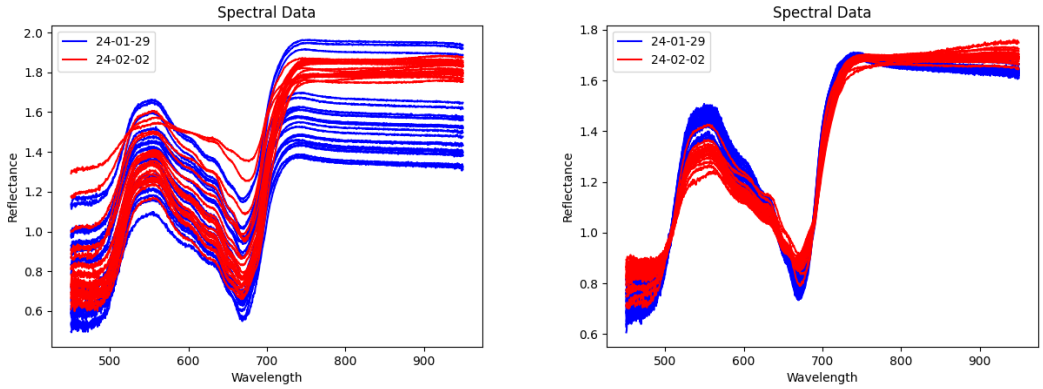

Figura 2.3: Confronto fra gli spettri grezzi (sinistra) e dopo l'applicazione della MSC (destra).

La correzione MSC, in conclusione, permette di ottenere due vantaggi: il modello viene semplificato, nel senso che le componenti principali necessarie per ottenere buone predizioni sono ridotte per i dati corretti con l' MSC rispetto ai dati non corretti, e solitamente la linearità aumenta. [10]

2.1.2 Min - Max Normalization

La normalizzazione min-max è una nota tecnica di preprocessing e consiste nel riscalare ciascuna feature nell'intervallo [0,1] tramite la trasformazione:

$$x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

dove x è la variabile considerata, $\max(x)$ e $\min(x)$ i suoi valori massimo e minimo, rispettivamente. Nella Figura 2.6 è possibile vedere un esempio di applicazione e notare come la distribuzione relativa degli spettri rimanga invariata, venendo semplicemente riscalata in [0,1]. Uno dei vantaggi della tecnica è il fatto di mantenere la proporzione fra i dati originali; dall'altra parte essa non centra i dati intorno alla media, non rimuovendo quindi le differenze di scala fra grandezze diverse, ma nel caso considerato non è un problema, dato che le variabili trattate corrispondono tutte misure della stessa grandezza fisica, ovvero la riflettanza.

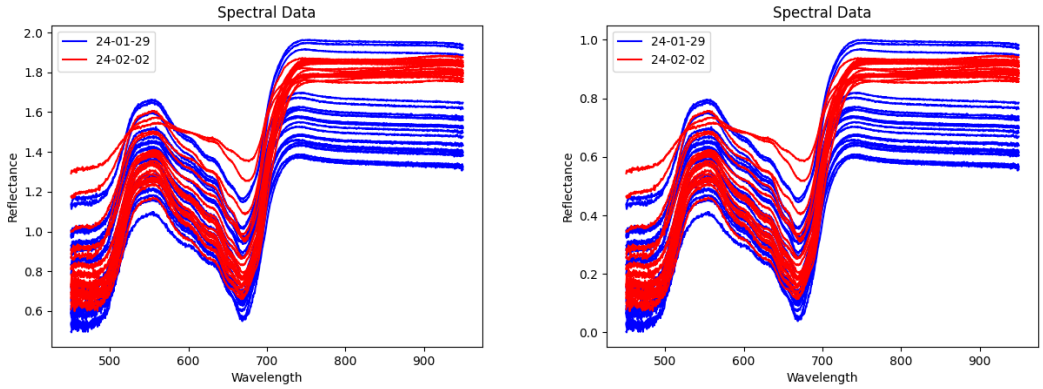

Figura 2.6: Confronto fra gli spettri grezzi (sinistra) e dopo l'applicazione della normalizzazione min-max (destra).

2.1.3 Standard Normal Variate (SVN)

L'obiettivo della tecnica è standardizzare gli spettri, ovvero centrare la loro media intorno allo 0 e far aver loro una deviazione standard unitaria. Usando la stessa notazione dell'MSC e definendo con m_i la media di tutte le K lunghezze d'onda del campione i e s_i la loro deviazione standard, la standardizzazione si ottiene usando:

$$x_{ik} = \frac{x_{ik} - m_i}{s_i}$$

La tecnica permette di eliminare le variazione sistematiche nell'intercetta dei dati spettrali, causate da fenomeni come lo scattering (di cui si è parlato nella normalizzazione MSC). Questa tecnica inoltre rimuove le variazioni di ampiezza non rilevanti, permettendo di evidenziare le variazioni che sono effettivamente utili nell'analisi

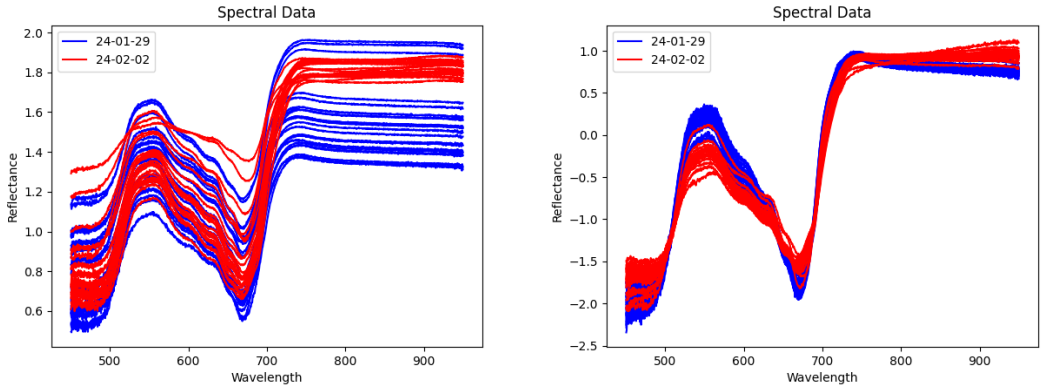

Figura 2.9: Confronto fra gli spettri grezzi (sinistra) e dopo l'applicazione della SVN (destra).

2.2 Analisi

Supponiamo di avere una variabile chiamata variabile dipendente o variabile risposta, e di volerne predire il valore utilizzando una o più variabili chiamate predittori o feature e indichiamo con la lettera Y maiuscola le prime e con la lettera X maiuscola le seconde; utilizzando un pedice nel caso in cui se ne consideri più di una. Si vuole trovare una funzione f per modellare la relazione fra X e Y, cioè $Y = f(X)$. Tranne nel caso in cui la relazione sia deterministica, non si riesce a verificarla in modo esatto per numerosi motivi (rumore, errore nelle misurazioni, ecc) quindi, per poter scrivere l'uguaglianza, viene aggiunta all'equazione una variabile casuale ϵ che rappresenti l'errore:

$$Y = f(X) + \epsilon$$

Un insieme di metodi per stimare la funzione f, per poi utilizzarla per effettuare predizioni su dati diversi ma della stessa tipologia, è il machine learning. Indicando con la lettera minuscola le osservazioni delle variabili utilizzate,il training set per addestrare il modello è rappresentabile con $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, in cui ciascun $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, dove p è il numero di feature di ciascuna osservazione. Le tecniche utilizzate sono molteplici e diverse fra loro, ma è possibile effettuare diverse classificazioni:

- se la variabile Y è **quantitativa** (continua) si parla di problemi di regressione, mentre se Y è **qualitativa** (discreta) di classificazione;
- se la variabile Y viene utilizzata per addestrare il modello si parla di apprendimento **supervisionato** (supervised), mentre se non viene utilizzata

nell'addestramento (per esempio nel caso in cui si vogliano trovare relazioni fra le osservazioni o fra le varie feature) si parla di parla di addestramento **non supervisionato** (unsupervised);

- è possibile assumere che la funzione f da stimare abbia una forma funzionale specifica: per metodi che fanno questa assunzione si parla di metodi **parametrici**, mentre senza questa assunzione di metodi **non-parametrici**

Una questione rilevante nel machine learning riguarda la capacità dei modelli ottenuti di operare con dati non visti precedentemente: un modello che riesce a ottenere buoni risultati, e dunque un errore ridotto, con i dati di training, ma che non ha buona capacità predittiva su dati dello stesso tipo, ma che non ha mai visto, è poco utile. Tale fenomeno è chiamato overfitting e si verifica quando, appunto, il modello aderisce troppo ai dati di training senza aver poi la capacità di generalizzare i risultati ottenuti e applicarli su altri dati. Ciò si verifica perché il metodo cerca di minimizzare l'errore il più possibile e ciò può portare a considerare pattern casuali come se facessero invece parte della struttura dei dati.

Un modo per ridurre questo problema è dividere i dati a disposizione in due sottoinsiemi chiamati train set e test set: l'obiettivo è avere due dataset simili, ma indipendenti fra loro, in modo da addestrare il modello sul primo e testarlo sul secondo, di dimensioni più ridotte, per valutare le capacità predittive del modello su dati non visti precedentemente

2.2.1 Cross-validation

Nel caso in cui non si disponesse di un numero sufficientemente elevato di osservazione per dividerle nei due sottoinsiemi, fra le possibili soluzioni una è la K-fold cross-validation. Il metodo consiste nella divisione casuale dei dati in K insiemi di dimensione simile: ciascuno di questi verrà usato come test set per il modello addestrato sui rimanenti K-1 insiemi, ottenendo alla fine della procedura K stime dell'errore associato al modello; gli errori trovati vengono poi combinati, per esempio facendo la media degli errori ottenuti. La tecnica ha un ovvio svantaggio: il modello dovrà essere addestrato K volte e, in alcuni casi, questa può essere una problematica rilevante. La scelta di K può arrivare fino a K=N, cioè il numero di fold è pari al numero di osservazioni, ogni test set contiene una sola osservazione e in questo parla di Leave-One-Out-Cross-Validation, ma questo potrebbe essere computazionalmente troppo oneroso, quindi spesso vengono usati valori di K come 5 o 10.
La tecnica ha un duplice utilizzo: è possibile servirsene sia per valutare le abilità predittive di un modello su dei dati indipendenti, sia per confrontare diversi modelli o un solo modello usando diversi parametri, e capire quale sia il migliore.

2.2.2 Metriche di valutazione

Per quanto riguarda la valutazione dei risultati sono state usate essenzialmente due tecniche:

- *Root Mean Squared Error (RMSE)*: è una metrica di valutazione utilizzata in modelli di regressione che misura la differenza fra i valori $\hat{y}_i$ previsti dal modello e i valori effettivamente osservaty y_i. Esso è definito come la radice quadrata della media dei quadrati degli errori; in formula:

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

  ricordando che N è la dimensione del campione considerato. Esso rappresenta quindi la distanza delle predizioni del modello dai dati osservati, quindi 0 rappresenta il valore minimo ottenibile e, in generale, valori più bassi rappresentano risultati migliori.

- *Accuracy*: è una metrica di valutazione utilizzata in modelli di classificazione che quantifica il numero di classificazioni corrette sul numero totale di predizioni; assume quindi valori in [0,1], dove il valore 1 rappresenta la totale correttezza delle predizioni, quindi valori più elevati rappresentano risultati migliori.

2.2.3 Principal Component Analysis (PCA)

La PCA è una tecnica di machine learning non supervisionato e non parametrico, viene solitamente usata per l'analisi di dataset contenenti un gran numero di feature. PC1 è vettore che definisce la retta che sia il più vicino possibile ai dati o, equivalentemente, definisce la direzione rispetto alla quale i dati hanno la massima varianza possibile, nel senso che se proiettassimo le osservazioni su questa retta, esse avrebbero la maggiore varianza possibile, mentre proiettandole su qualsiasi altra retta le osservazioni proiettate avrebbero minore varianza[11]. Le PC vengono determinate come combinazioni lineari delle feature iniziali e e sono calcolate in modo sequenziale per essere non correlate tra loro; quest'ultima condizione è equivalente a richiedere che le componenti principali siano ortogonali fra loro. Dato il modo in cui queste PC vengono determinate, la prima componente sarà la più informativa, quella che contiene più variabilità nei dati rispetto alle altre, seguita dalla seconda e così via; è quindi possibile riuscire a spiegare gran parte della variabilità dei dati con un basso numero di componenti, che varia in base alla struttura dei dati, ma sicuramente molto ridotto rispetto al numero di feature iniziali.

Date le sue proprietà, la tecnica può essere usata con vari scopi. Data la sua capacità di "riassumere" i dati, eliminando le componenti meno significative e probabilmente associate al rumore, è possibile usarla per fare "dimensionality reduction", ovvero descrivere le osservazioni con un numero ridotto di nuove variabili (le componenti principali) rispetto a quello iniziale: questo è utile sia come pre-processing, facilitando l'individuazione di pattern nei dati con altre tecniche, sia perché rende possibile la rappresentazione delle osservazione nello spazio usando le prime due o tre componenti principali, mentre sarebbe impossibile rappresentare efficacemente punti con centinaia o migliaia di dimensioni. Inoltre, valutando i coefficienti di ciascuna feature originale nella definizione delle prime componenti principali è possibile capire quali sono le feature originali che hanno più peso nella variabilità dei dati.

Un limite della PCA è che, essendo una tecnica non supervisionata, essa identifica la variabilità dei dati ma non distingue fra la variabilità all'interno dello stesso gruppo e quella fra gruppi diversi. Nonostante questo, viene comunque usata nell'ambito della classificazione perché in buona parte dei casi la variabilità fra gruppi è dominante rispetto a quella interna ai gruppi; nei casi in cui ciò non avviene è preferibile fare ricorso ad altre tecniche per raggiungere lo scopo, come ad esempio la Partial Least-Square Regression.[12]

2.2.4 Regressione lineare

La regressione lineare è una delle tecniche fondamentali della statistica ed è considerabile come tecnica di apprendimento parametrico e supervisionato. Dato un insieme di N osservazioni, chiamato campione, da una popolazione, l'obiettivo è trovare una relazione lineare fra delle variabili indipendenti $X_1, \ldots, X_K$, chiamate predittori, e una variabile dipendente y. Geometricamente, ciò corrisponde a cercare la retta che minimizzi la distanza deipunti dalla retta stessa. È possibile scrivere la relazione in forma matriciale:

$$y = X\beta + \epsilon$$

in cui y è un vettore di dimensione $N \times 1$, X una matrice $N \times (K+1)$, β un vettore di dimensione $(K+1) \times 1$ contenente i coefficienti della retta di regressione, oltre all'intercetta, e ϵ un vettore $N \times 1$. In particolare X è detta design matrix e contiene tutte le osservazioni come righe, con l'aggiunta di un 1 nella prima colonna. Ogni colonna sarà quindi associata a una diversa variabile, tranne la prima, costituita da soli 1, che sarà associata all'intercetta β_0 della retta di regressione. Possiamo quindi scrivere la relazione per l' i-esima osservazione come:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \epsilon_i \quad \text{con } i = 1, \ldots, N$$

indicando con x_{ik} l'osservazione i-esima della k-esima variabile.
Per $k = 1$, quindi una regressione con un unico predittore, si parla di regressione

semplice, mentre per $k > 1$ di regressione multipla; in questo secondo caso, ciascun β_i rappresenta l'effetto medio sulla risposta y al variare di un'unità del predittore X_i, mantenendo fissi tutti gli altri [11] e, geometricamente, la retta considerata nella regressione semplice diventa un iper-piano (se non si considerano le interazioni fra i predittori, altrimenti si ottengono superfici più complicate) .

È possibile dimostrare, per esempio tramite la massimizzazione della funzione di verosimiglianza [13] che i valori dei coefficienti nel vettore β sono ottenibili da:

$$\hat{\beta} = (X'X)^{-1}X'y$$

avendo indicato con X' la trasposizione della matrice X, con X^{-1} la sua inversione. Utilizzeremo il cappuccio sulle variabili quando indicheremo le stime (in questo caso quella di β). Utilizzando questo risultato è possibile effettuare delle predizioni $\hat{y}$ usando:

$$\hat{y} = X\hat{y} = X(X'X)^{-1}X'y$$

Una delle problematiche principali della regressione multipla è la collinearità dei predittori: se due o più delle variabili di regressione, quindi le colonne della matrice X, sono fortemente correlate, la matrice $X'X$ è singolare e non è quindi possibile calcolare la matrice inversa $(X'X)^{-1}$. Una possibile soluzione è l'eliminazione delle variabili correlate oppure si possono utilizzare direttamente altre tecniche che permettono di lavorare con dati di questo tipo.

2.2.5 Partial Least Square (PLS)

Per comprendere le basi della tecnica PLS sono stati consultati diverse fonti, in particolare [14], [15], [10]. La regressione PLS (PLSR) è una tecnica di regressione multivariata supervisionata, utile nel caso di dati in cui i predittori sono in numero molto superiore rispetto alle osservazioni disponibili oppure le variabili sono correlate fra loro, rendendo inutilizzabile la regressione multivariata standard (senza l'aggiunta di modifiche o selezione di variabili). I dati spettroscopici sono quindi un buon esempio per l'applicazione della PLS,in quanto costituiti da un numero elevato di variabili (corrispondenti alle lunghezze d'onda misurate), che sono solitamente molto più numerose degli spettri misurati; inoltre è solitamente presente un'elevata correlazione fra le variabili associate a frequenze d'onda vicine.

L'idea alla base di questo metodo è cercare di descrivere i dati originali X tramite delle nuove variabili, dette variabili latenti e ottenute tramite combinazioni lineari di quelle originali, per poi utilizzare queste nuove coordinate nella stima di una regressione lineare: in questo modo è possibile eliminare parte del rumore e delle informazioni meno stabili dei dati, usando nella regressione solo le informazioni più rilevanti. Questo consente inoltre di risolvere i problemi di collinearità dei dati, che è un problema rilevante in spettroscopia che rende problematico l'uso della

regressione semplice multivariata, ottenendo equazioni di regressione e predizioni più stabili.

Il criterio secondo il quale vengono generate le variabili latenti è ottenere la massima riduzione della covarianza dei dati X^TY. Come nella PCA, le nuove variabili vengono generate sequenzialmente e vengono calcolate considerando la variabilità non rimossa dalle variabili precedenti.

In modo più rigoroso, definiamo la prima componente della PLS $\hat{w}_1$ come la direzione rispetto alla quale la covarianza fra i predittori X e la variabile risposta y è massimizzata; questo è un vettore unitario ed è chiamato loading weight. Utilizzando $\hat{w}_1$, è possibile calcolare la proiezione dei dati X in questa direzione, ottenendo l' X-score t_1 associato come $\hat{t}_1 = X\hat{w}_1$. Questo rappresenta la prima variabile latente e viene a sua volta utilizzato per calcolare la proiezione dei dati originali X nella nuova direzione (quella di $\hat{t}_1$), ovvero $X \approx \hat{t}_1\hat{p}_1$. p_1 è detto loading vector e, rappresentando i dati nelle vecchie coordinate nelle nuove, permette di identificare quali sono le variabili più importanti nella determinazione della variabile $\hat{t}_1$, quindi quelle che più influenzano la variabilità dei dati. Per il modo in cui è stato definito, possiamo ottenere $\hat{p}_1$ facendo la regressione di X rispetto a t_1: dalla teoria della regressione è noto che la proiezione di X su $\hat{t}_1$ che minimizzi la distanza è data da

$$\hat{p}_1 = (\hat{t}_1'\hat{t}_1)^{-1}X'\hat{t}_1$$

Gli X-scores $\hat{t}$, moltiplicati dai loading $\hat{p}$ descrivono bene X, nel senso che i residui associati alla regressione $X = TP' + E$ sono piccoli.
In modo analogo, dopo aver calcolato i pesi $\hat{c}_1$ massimizzando la correlazione fra $\hat{t}_1$ e la variabile risposta y, si calcola lo score $\hat{u}_1$ proiettando y su $\hat{c}_1$ come $\hat{u}_1 = y\hat{c}_1$ e i loading per y facendo la regressione sullo score ottenendo, come prima:

$$\hat{q}_1 = (\hat{t'}_1\hat{t}_1)^{-1}y'\hat{t}_1$$

A questo punto i prodotti $\hat{t}_1\hat{p}_1$ e $\hat{t}_1\hat{q}_1$ vengono sottratti rispettivamente da X e y e si procede con il calcolo della seconda componente come fatto per la prima, fino al raggiungimento del numero desiderato di variabili latenti.
È possibile quindi scrivere le equazioni del modello in forma matriciale come

$$\begin{cases} T = XW \\ X = TP' + E \\ Y = TQ' + F \end{cases}$$

Il modello può avere un numero massimo di variabili latenti pari al minimo fra il numero di variabili osservate e il numero di osservazioni, ovvero $\min\{K, N\}$. Nel caso in cui venga usato il numero massimo di componenti, la PLS è uguale alla regressione lineare multipla.

È possibile notare un parallelelismo fra la PLS e l'applicazione della una regressione alle componenti principali ottenute con la PCA (questa tecnica viene chiamata Principal Component Regression, PCR), ma con la differenza fondamentale che la PLS è una tecnica di apprendimento supervisionato, mentre la PCA no: questo si riflette nel modo in cui le nuove variabili (rispettivamente variabili latenti o componenti principali) vengono determinate, infatti nella PCA vengono considerate solo le informazioni delle feature, mentre nella PLS si considera anche la variabile risposta, quindi le nuove variabili non approssimano solo le vecchie feature, ma sono correlate anche alla risposta [11]

2.2.6 Partial Least Square - Discriminant Analysis (PLS-DA)

La PLS-DA è la tecnica che usa le proprietà della PLS con lo scopo di effettuare una classificazione. È di interesse effettuare una classificazione binaria, quindi verranno usate delle etichette discrete come {1,0} o $\{1, -1\}$, ma è possibile estendere la tecnica per applicarla a più classi utilizzando delle dummy variable. La classificazione può essere effettuata in diversi modi, usando dei classificatori sui dati processati con la PLS; negli esperimenti effettuati si è deciso di separare le due classi usando l'output della PLS-R e una soglia fissa, data dal punto medio fra le due etichette, per assegnare ciascuna osservazione a una classe. I risultati suggeriscono che la PLS dovrebbere essere superiore alla PCA per ridurre la dimensionalità con l'obiettivo di ottenere una separazione di classe. [12]

Capitolo 3

Metodologia

Le tecniche di selezione delle variabili hanno acquisito una crescente importanza negli anni. In particolare, nell'ambito della spettroscopia ciò è importante perché i dati spettroscopici hanno solitamente dimensioni elevate, ma solo parte di questi è davvero utile per l'analisi, dato che molte lunghezze d'onda possono contenere rumore essere ridondanti. Questa ridondanza porta spesso alla collinearità fra le variabili, rendendo inutilizzabili tecniche come la regressione multipla, che richiede variabili non correlate per ottenere stime accurate. Dal punto di vista dell'ottimizzazione è possibile vedere la selezione delle lunghezze d'onda come un processo di ottimizzazione che massimizza le performance predittive del modello [16], infatti è stato dimostrato sia a livello teorico che sperimentale che l'applicazione di una selezione delle lunghezze d'onda, invece dell'uso dell'intero spettro, riesca a migliorare la performance predittive.[17] Un altro vantaggio di queste tecniche è la riduzione della complessità computazionale, infatti lavorare con un numero ridotto di variabile consente di velocizzare il processo di modellazione e, allo stesso tempo, utilizzare modelli più facilmente interpretabili. L'utilizzo di un numero ridotto di lunghezze d'onda rende inoltre possibile l'utilizzo di strumenti spettroscopici più semplici ed economici. Nello specifico, nel lavoro sono state considerate due di queste tecniche: Moving Window PLS-DA (MWPLS-DA) e la Competitive Adaptive Reweight Sampling (CARS), che come risultato producono appunto dei modelli basati su un numero di lunghezze d'onda molto ridotto rispetto allo spettro iniziale.

3.1 Moving Window PLS-DA (MWPLS-DA)

La MWPLS-DA è una tecnica nella quale vengono esaminati e selezionati degli intervalli spettrali, invece che delle singole lunghezze d'onda, che possono essere più utili nella costruzione del classificatore. Questa scelta è motivata dalla **continuità**

caratteristica della maggioranza degli spettri: ciò significa che se una lunghezza d'onda è informativa per il modello, probabilmente anche l'intervallo spettrale a cui appartiene conterrà informazioni utili. In modo analogo, se una lunghezza d'onda contiene rumore, probabilmente lo conterranno anche quelle vicine.[17]
A queste osservazioni se ne lega un'altra: le lunghezze d'onda contenenti rumore sono associate a una maggiore incertezza, dovuta alla difficoltà di modellazione di quest'ultimo, che causa un aumento di complessità del modello e nei modelli con variabili latenti questo si traduce nella necessità di un maggior numero di variabili nel modello; d'altra parte, è possibile identificare i canali con minore incertezza come quelli che necessitano di un numero ridotto di variabili latenti per raggiungere dei buoni risultati. La tecnica permette quindi lo sviluppo di un modello più robusto usando un numero ridotto di intervalli, grazie all'eliminazione di quelli poco informativi, e che possiede quindi una maggiore accuratezza predittiva, oltre alla riduzione dei costi computazionali. [18] È possibile strutturare l'algoritmo come segue:

- **Selezione della dimensione delle finestre**.
  Per applicare il modello si considera una finestra mobile di ampiezza H; per comodità ciascuna finestra verrà identificata dalla prima lunghezza d'onda considerata: la finestra che associata al canale j-esimo sarà costituita dall'intervallo $[j, j + H - 1]$. Associata a ciascuna finestra avremo una sottomatrice di dimensione $N \times H$, contenente tutte le osservazioni e le lunghezze d'onda considerate.

- **Costruzione dei modelli locali.**
  La finestra mobile viene fatta scorrere lungo l'intero spettro (l'ultima finestra sarà quella che inizia in K-H+1, ricordando che K è il numero totale di lunghezze d'onda), considerando finestre che si sovrappongono parzialmente, in modo da riuscire ad esaminare ogni regione dello spettro. In ciascuna finestra viene applicata la PLS usando la cross-validation, calcolando l'errore al variare del numero di variabili latenti utilizzate, utilizzandone comunque un numero non troppo elevato per evitare che vengano modellate informazioni inutili.

- **Valutazione dei modelli locali.**
  Dato che al crescere del numero di componenti della PLS l'errore dei modelli si riduce, per quanto detto prima, le finestre considerate informative saranno quelle capaci di raggiungere un errore relativamente basso con un numero ridotto di variabili latenti; per valutare questo, sono state calcolate la media e la mediana degli errori per ciascuna finestra.

- **Costruzione del modello finale.**
  Usando il risultato precedente, vengono selezionate le finestre associate agli

errori più ridotti e vengono utilizzate per costruire il classificatore finale con la PLS-DA.

3.2 Competitive Adaptive Reweight Sampling (CARS)

Un altro metodo usato per trattare la collinearità nei dati (quindi gli spettri nel caso studiato) è l'algoritmo Competitive Adaptive Reweight Sampling, detto CARS. Questa tecnica permette di valutare quali sono le lunghezze d'onda più importanti e eliminare quelle che non lo sono, secondo il principio evoluzionistico di selezione del più forte, tramite campionamenti l'uso ripetuto di campionamenti Monte Carlo e l'applicazione della PLS; l'utilità di questa selezione è data dal fatto che è stato dimostrato, sia da un punto di vista teoriche che sperimentale, un miglioramento delle performance del modello usando solo un insieme ristretto di lunghezze d'onda informative rispetto all'uso dello spettro completo.
L'algoritmo viene ripetuto per un numero R di run, negli esperimenti è stato scelto il valore R=500, e ciascuna run è strutturata come segue:

- **Selezione casuale delle osservazioni.**
  Ad ogni iterazione non viene utilizzata la totalità delle osservazioni disponibili, ma solo una parte, selezionata in modo casuale, solitamente circa l' 80-90% delle osservazioni disponibili: questo perché l'obiettivo è trovare un insieme di variabili che siano utili per descrivere il fenomeno in modo indipendente rispetto alle osservazioni del training set.

- **PLS e pesi delle variabili.**
  La tecnica della PLS è già stata spiegata in generale in 2.2.5. L'obiettivo della tecnica in questa fase dell'algoritmo è determinare l'importanza delle variabili (lunghezze d'onda) che si stanno considerando tramite la definizione di pesi associati. Calcolata la PLS-R si ottengono i coefficienti di regressione associati alle lunghezze d'onda considerate: si può dire che il valore assoluto di questi coefficienti rappresenti quanto ciascuna variabile contribuisce alla risposta y. Dunque, per valutare l'importanza di ciascuna variabile, definiamo un peso normalizzato associato a ciascuna:

  $$w_i = \frac{|\beta_i|}{\sum_{i=1}^{p} |\beta_i|} \qquad \text{con} \qquad i = 1, \ldots, p$$

  Una volta che una variabile viene eliminata dall'algoritmo, nelle iterazioni successive avrà un peso associato pari a zero.

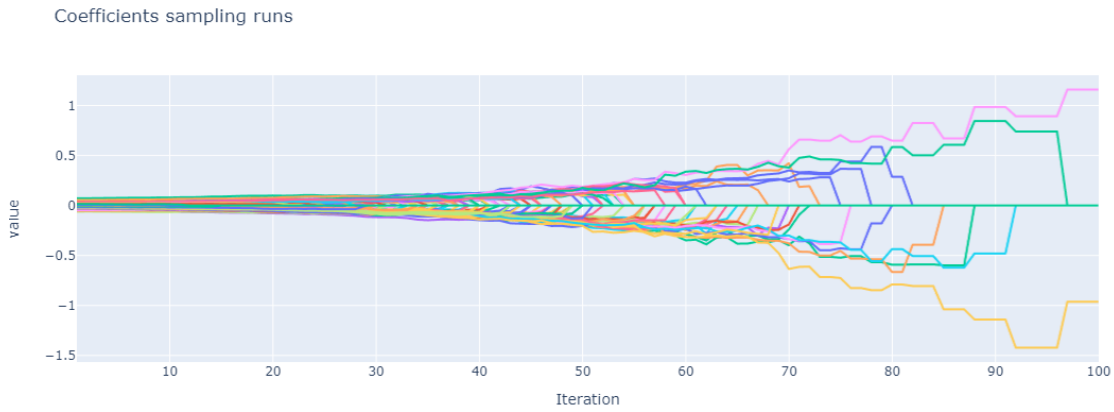

Figura 3.1: Andamento dei coefficienti al variare delle iterazioni

Dalla Figura 3.1 è possibile vedere un tipico andamento dei coefficienti β_i al variare delle iterazioni: nelle prime, tutti i coefficienti hanno un valore vicino allo zero, dovuto alla presenza di un numero elevato di variabili, che porta la singola variabile ad avere un peso minore, ma al procedere delle iterazioni,con sempre più coefficienti che valgono zero (a causa dell'eliminazione delle frequenze d'onda associate), le lunghezze d'onda che sopravvivono avranno valori più distanti da zero per la loro crescente importanza.

- **Calcolo del numero di variabili da mantenere.**
  In questa parte viene calcolata la frazione di variabili da mantenere nell'eliminazione che avverrà al passo successivo. Questa quantità viene determinata secondo una funzione esponenziale descrescente e all'iterazione i-esima è pari a:

$$r_i = \alpha e^{-ki}$$

I termini α e k sono delle costanti, che vengono determinate in modo che la funzione soddisfi due condizioni:

1. alla prima iterazione si vuole che vengano considerate tutte le p variabili; ciò si traduce nella condizione $r_1 = 1$
2. all'ultima iterazione, si vuole che rimangano solo 2 variabili, quindi $r_N = \frac{2}{p}$

Risolvendo il semplice sistema a due incognite associato alle due condizioni, si ottengono per le costanti i valori di:

$$\alpha = (\frac{p}{2})^{\frac{1}{N-1}}, \qquad k = \frac{ln(\frac{p}{2})}{N-1}$$

In particolare,questi valori sono stati calcolati per far compiere 100 iterazioni per ciascuna run.
Osservando l'andamento della funzione nella Figura 3.2 è possibile dividere la selezione in due fasi: nei primi punti la funzione ha una pendenza negativa piuttosto elevata, che si traduce nell'eliminazione di molte variabili in ciascuna delle prime iterazioni (si parla di "fast selection"), che va riducendosi man mano nelle iterazioni successive, operando quindi una selezione più ridotta (chiamata "refined selection").

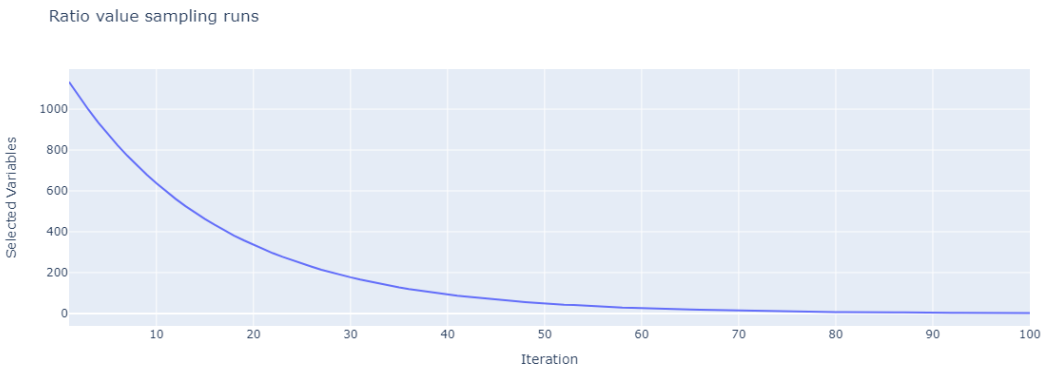

Figura 3.2: Plot della funzione esponenziale: viene mostrato il numero di variabili mantenute al variare delle iterazioni

- **Eliminazione frequenza con adaptive reweight sampling (ARS).**
  In questa fase si procede all'effettiva eliminazione delle variabili in modo competitivo: si scelgono le variabili da mantenere tramite un campionamento nel quale ciascuna ha come peso quello ottenuto nella fase precedente; una variabile con peso più elevato avrà quindi probabilità maggiore di essere estratta, venendo probabilmente selezionata un numero maggiore di volte.

Nella Figura 3.3 è presentato il diagramma di flusso che riassume i passaggi dell'algoritmo.

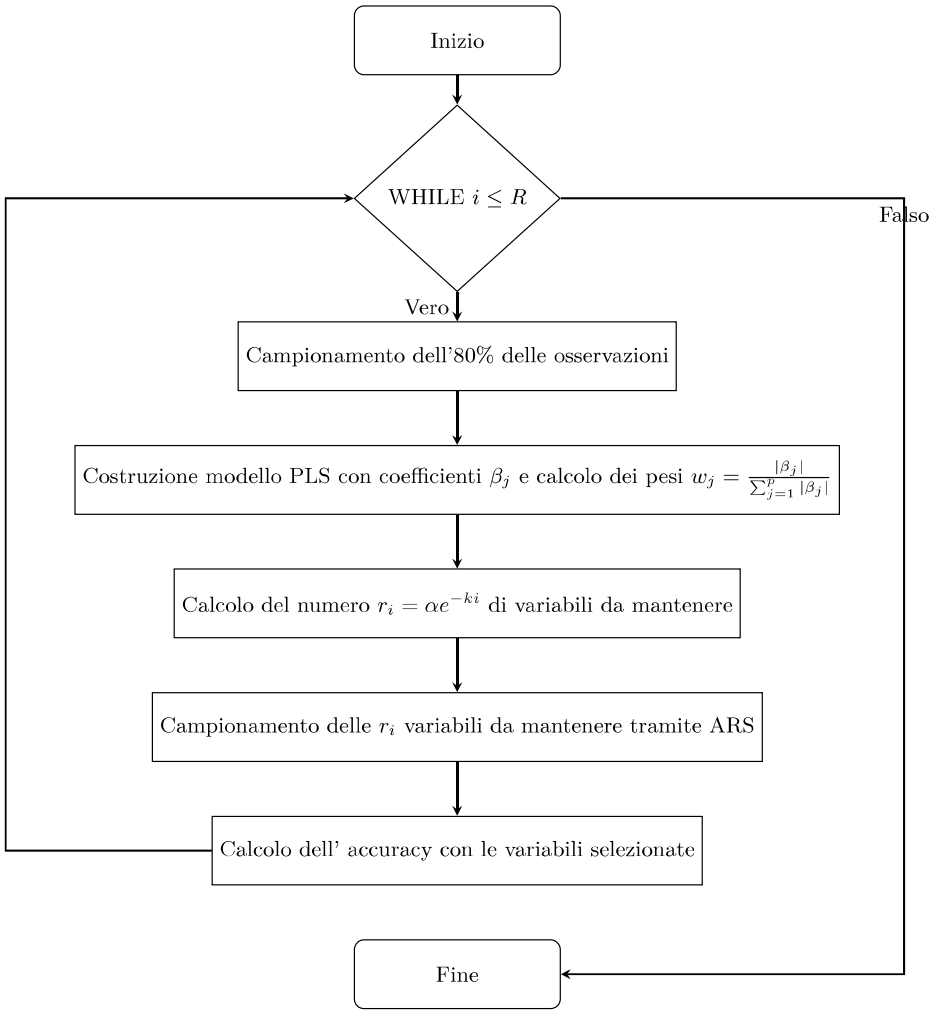

Figura 3.3: Diagramma di flusso di CARS

3.3 Materiali

3.3.1 Piante

Sono state usate piante delle piante di insalata. Le piante sono state annaffiate solo precedentemente rispetto all'inizio dei rilevamenti per provocare lo stress idrico. Sono state utilizzate rilevazioni effettuate a 5 giorni di distanza, il 29 gennaio e il 2 febbraio 2024, effettuando più rilevazioni per ciascuna pianta in ciascuna sessione. Nelle analisi nelle rilevazioni del primo giorno la pianta è stata considerata come sana mentre il quinto giorno è stata considerata sotto stress idrico. Sono stati usati nell'analisi un totale di 43 spettri, divisi in 21 spettri per la prima classe e 22 per la seconda avendo quindi un bilanciamento nel dataset fra le due classi di interesse.

Figura 3.4: Foto delle piante di insalata utilizzate nello studio

3.3.2 Spettrometro

Lo spettrometro usato per i rilevamenti è OCEAN-HDX-XR che permette di rilevare da 194nm a 1115nm; per gli è esperimenti è stato usato la regione compresa fra 450nm e 950nm, in quanto è stato indicato da esperti del settore che al di fuori di questo intervallo i rilevamenti effettuati con questo strumento non sono affidabili. Le lunghezze d'onda considerate sono state quindi 1133.

Figura 3.5: Spettrometro OCEAN HDX-XR

Capitolo 4

Risultati

4.1 PCA

Nel preprocessing sono state applicate le tre diverse normalizzazioni elencate nella sezione 2.1 per avere una prima valutazione della loro efficacia nella compressione dei dati e della separabilità delle due classi di interesse, oltre ad averne una visualizzazione, visto che per la loro dimensionalità questi dati non possono essere rappresentati se non con tecniche del genere, è stata applicata la PCA. In particolare, abbiamo utilizzato la Principal Component Analysis (PCA) per analizzare quanta varianza potesse essere spiegata dalle prime componenti principali (PC) per ciascun metodo di preprocessing applicato. Per ciascuna applicazione della tecnica vengono mostrati lo scree plot, ovvero il grafico della proporzione di varianza spiegata da ciascuna delle singole componenti principali ottenute con la PCA, e lo score plot, ovvero il grafico che mostra le osservazioni proiettate nello spazio generato dalle componenti principali.

4.1.1 PCA con normalizzazione SVN

Per i dati trattati con la normalizzazione SVN, lo scree nella Figura 4.1 questo mostra come la PCA riesca a descrivere i dati con l'uso di un numero molto ridotto di componenti; si osservi che la prima componente riesce a spiegare il 74% della varianza, mentre la seconda il 24.4%, per un totale di 98.4% con due sole componenti.

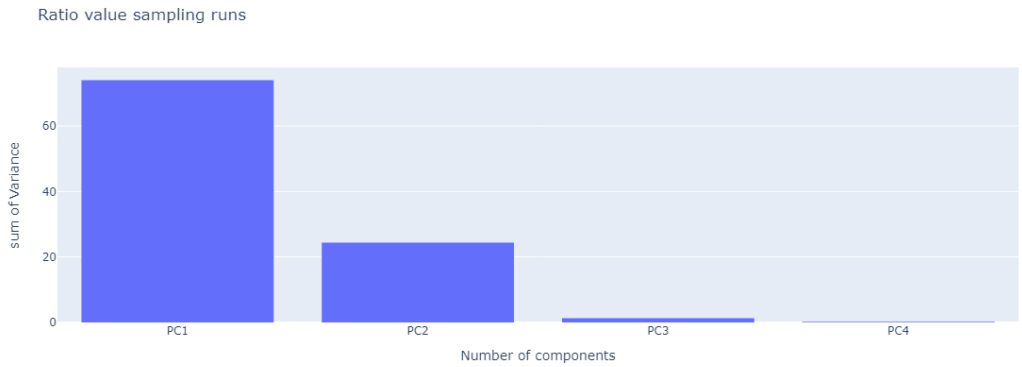

Figura 4.1: Varianza spiegata dalle prime componenti principali

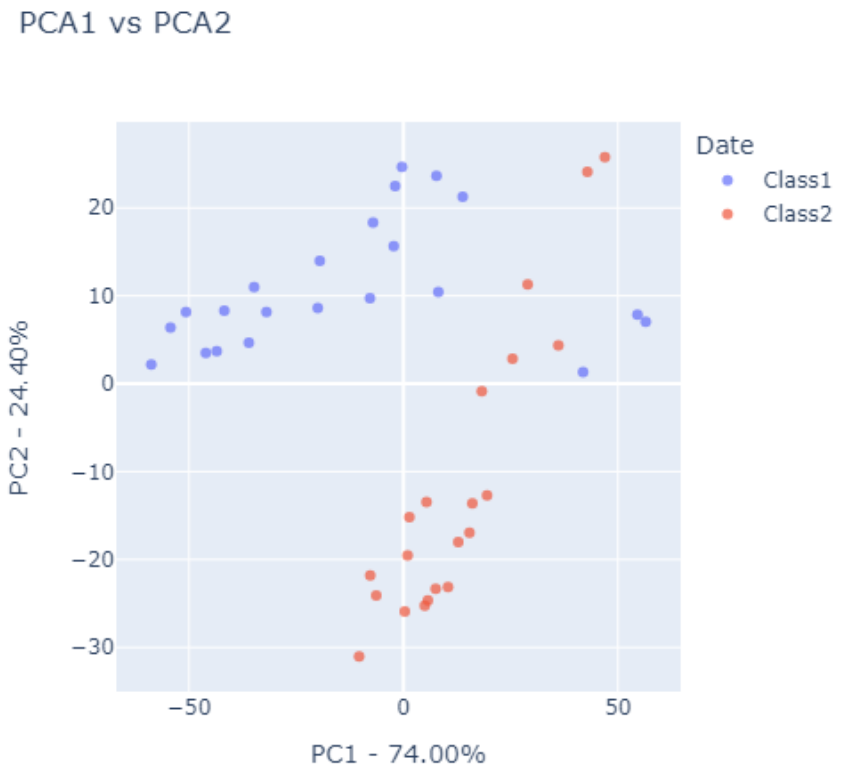

Figura 4.2: Scores delle osservazioni nello spazio generato dalle prime due PC in seguito alla normalizzazione SVN

Per quanto riguarda gli score plot, anche in questo caso le osservazioni delle due classi appaiono abbastanza separate negli score plot della Figura 4.2 delle prime due componenti principali, mentre risultano meno separate negli altri due score plot nella Figura 4.3, ma ciò potrebbe essere spiegato dal fatto che il primo plot spiega quasi la totalità della varianza del dataset. È però importante ricordare che la PCA è una tecnica di apprendimento non supervisionato, quindi le etichette di classe non vengono usate nel training e ottenere una separazione delle classi non è uno degli obiettivi della tecnica.

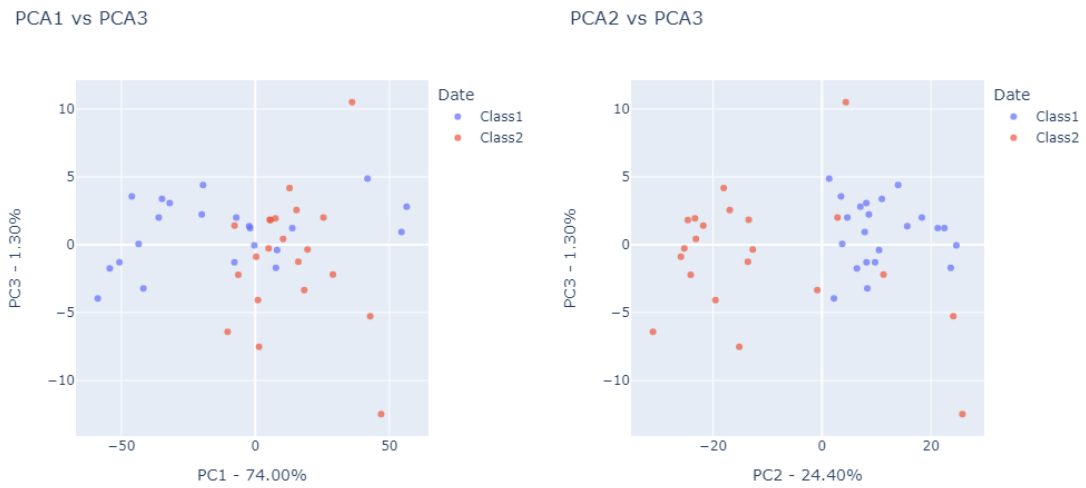

Figura 4.3: Osservazioni nello spazio delle componenti principali

4.1.2 PCA con normalizzazione MSC

Per i dati trattati con la normalizzazione MSC, nella Figura 4.4 è presente lo scree plot, nel quale è possibile vedere che le prime due componenti principali spiegano il 93% della varianza totale (81.10% per PC1 e 12.90% per PC2).

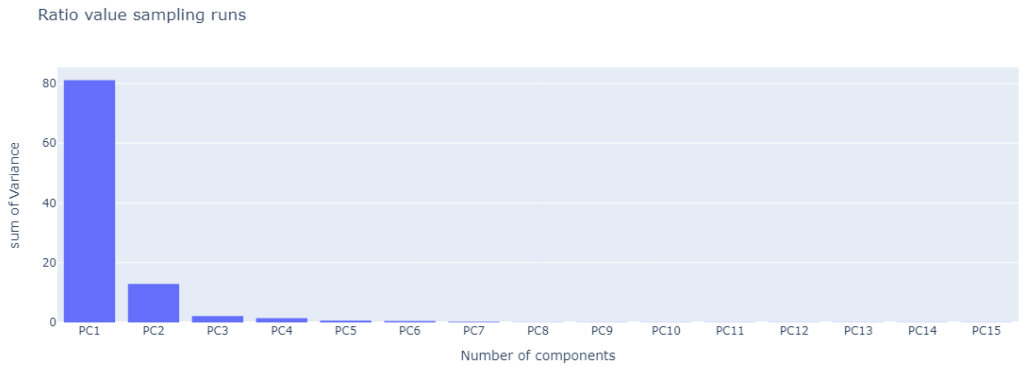

Figura 4.4: Varianza spiegata dalle prime componenti principali

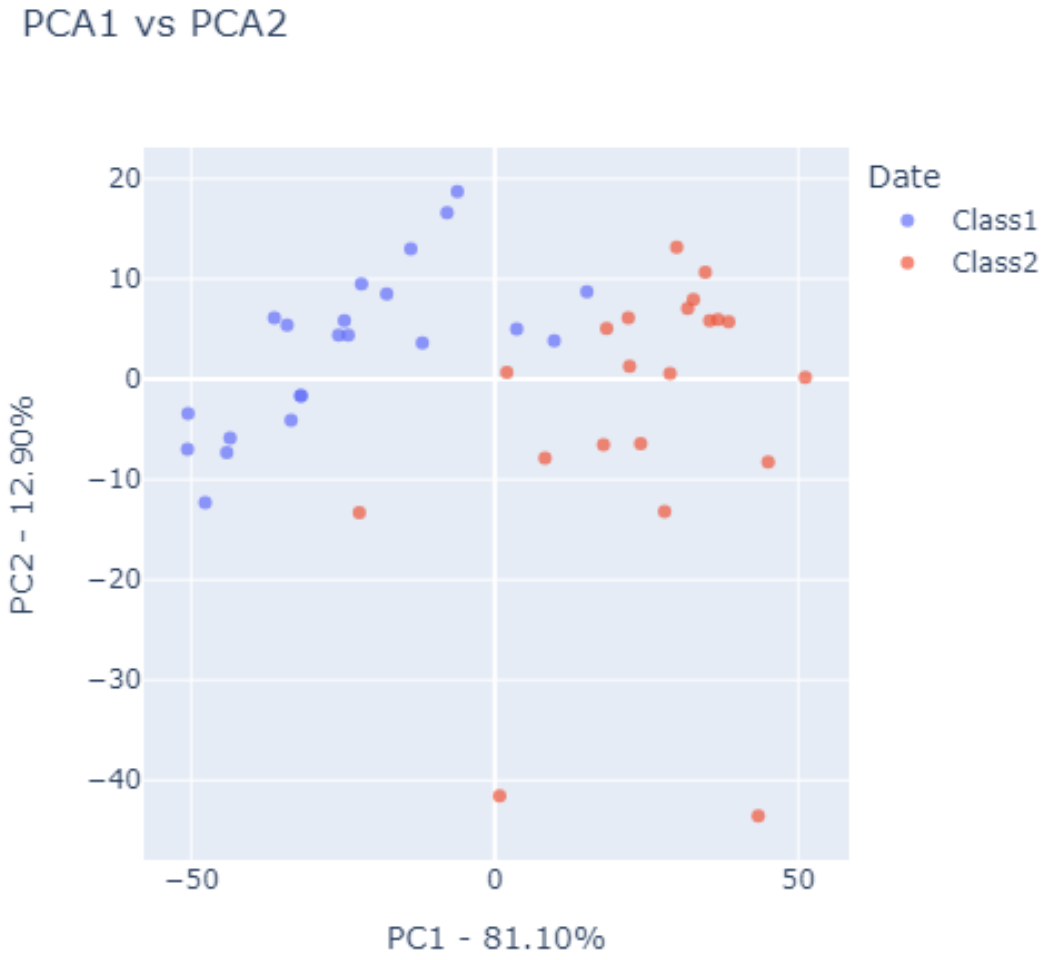

Figura 4.5: Scores delle osservazioni nello spazio generato dalle prime due PC in seguito alla normalizzazione MSC

Dagli score plot nella Figura 4.5 e nella Figura 4.6 è possibile osservare una certa separabilità fra le classi, anche se un più ridotta nello score plot delle PC2 e PC3, ma ciò potrebbe essere spiegato dalla percentuale di varianza che le due componenti spiegano, pari solo al 15%.

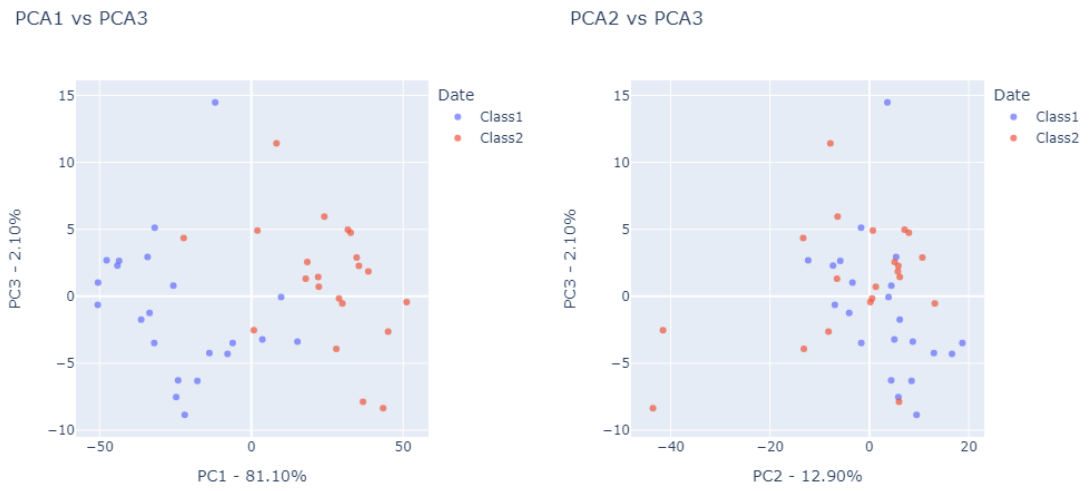

Figura 4.6: Osservazioni nello spazio delle componenti principali

4.1.3 PCA con normalizzazione min-max

Per quanto riguarda i risultati ottenuti con la normalizzazione min-max, di nuovo gran parte della varianza viene spiegata dalle prime due componenti, si ha una buona separazione nello score plot delle prime due componenti nella Figura 4.2, che però si riduce negli altri, quelli della Figura 4.3.

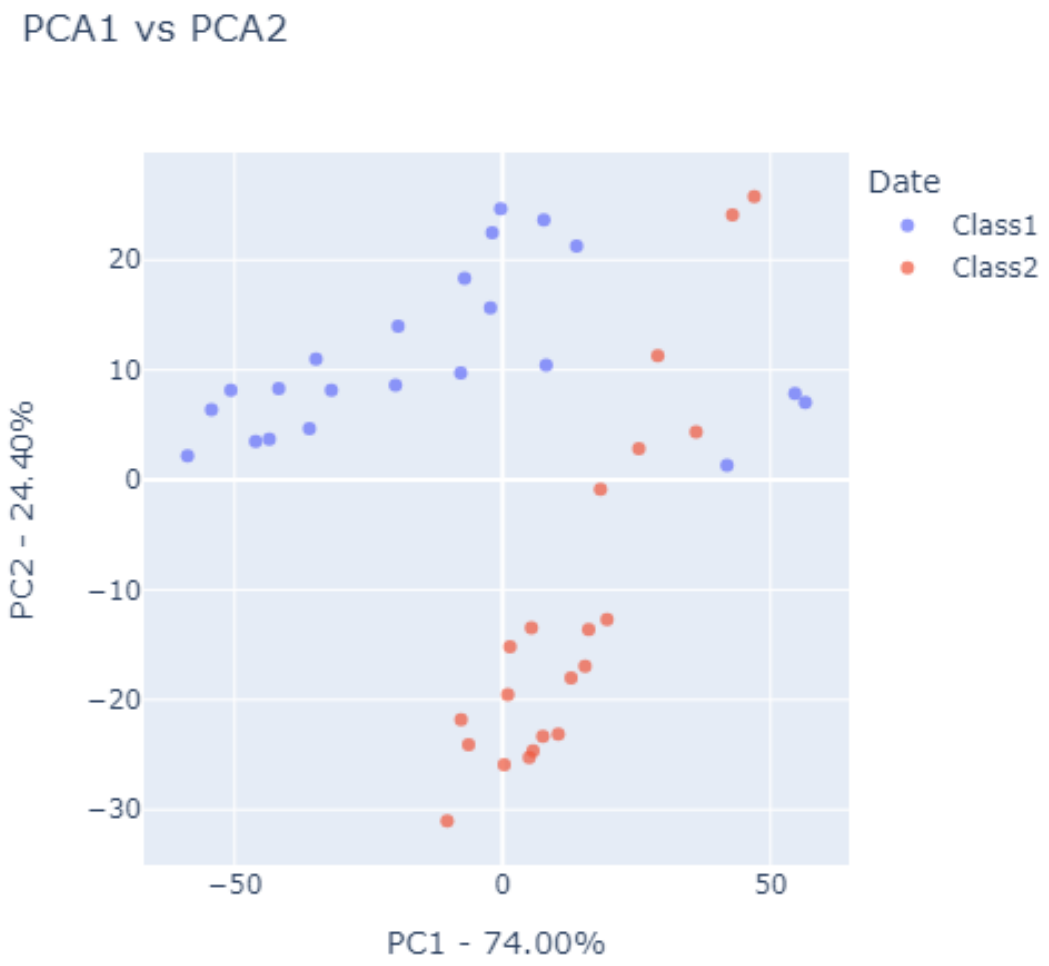

Figura 4.7: Scores delle osservazioni nello spazio generato dalle prime due PC in seguito alla normalizzazione SVN

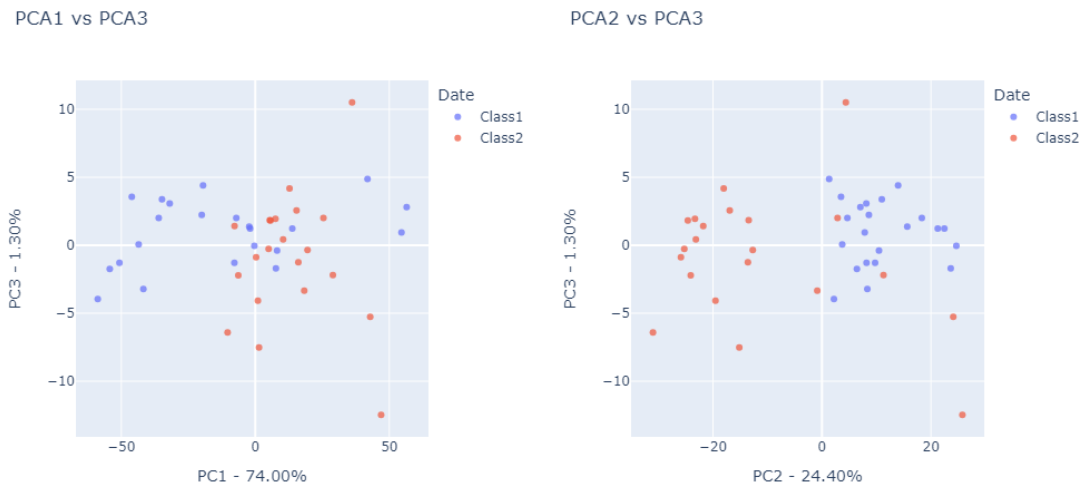

Figura 4.8: Osservazioni nello spazio delle componenti principali

È possibile osservare una forte somiglianza fra i risultati della PCA rispetto alla normalizzazione SVN e quella min-max: ciò è dovuto al fatto che, essendo i risultati della PCA basati sullo studio della matrice di covarianza, questa non viene alterata da trasformazioni lineari di traslazione e riscalamento, come nel caso delle due normalizzazioni citate, perché queste non cambiano la varianza relativa fra le variabili.

Nonostante la capacità della PCA di spiegare una parte significativa della varianza, per le analisi successive i dati non sono stati descritti tramite le componenti principali ottenute perchè entrambe le tecniche principali dello studio (CARS e MWPLS-DA) si basano sull'uso della PLS, che non solo riduce la dimensionalità, ma ottimizza anche la covarianza tra i predittori e la variabile dipendente, caratteristica utile nell'ambito della classificazione.

4.2 CARS

Dai risultati ottenuti con CARS per ciascuna normalizzazione sono state considerate sia le frequenze delle lunghezze d'onda associate alla migliore accuracy con il numero più ridotto possibile di variabili, sia le frequenze delle lunghezze d'onda che sono arrivate ad essere nelle ultime 2, nelle ultime 3 e nelle ultime 4 variabili sopravvissute nelle diverse esecuzioni dell'algoritmo. Per ciascuno dei risultati elencati sono state

considerate diverse combinazioni delle lunghezze d'onda per la regressione finale, testando anche lunghezza d'onda aggiuntive nel caso di variabili che avessero raggiunto frequenze di sopravvivenza simili.

4.2.1 CARS con normalizzazione MSC

- *Miglior accuracy con il numero minore di variabili (normalizzazione MSC).* Per quanto riguarda i risultati associati alla miglior accuracy col numero minore di variabili, dalla Figura 4.9 sono state selezionate le migliori 3 frequenze, sopravvissute in almeno 300 delle 500 run, è stata testata l'aggiunta della quarta, sopravvissuta 259 volte, e la quinta, sopravvissuta 236 volte.

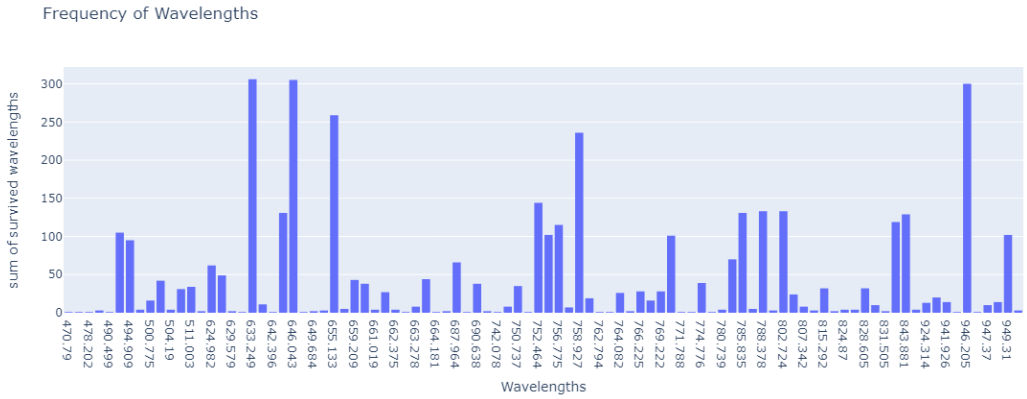

Figura 4.9: Frequenza delle lunghezze d'onda relative alla migliore accuracy con il numero minore di variabili con normalizzazione MSC

Lunghezze d'onda (nm)					Accuracy
633.249	646.043	-	-	946.205	0.907
633.249	646.043	655.133	-	946.205	0.930
633.249	646.043	655.133	758.927	946.205	0.930

Tabella 4.1: Accuracy dei modelli con le combinazioni delle variabili selezionate fra quelle associate alla migliore accuracy col numero minore di variabili con normalizzazione MSC

I risultati delle regressioni finali, elencati nella Tabella 4.1, mostrano che l'algoritmo riesce a raggiungere un'accuracy del 90.7% con sole 3 lunghezze d'onda, raggiungendo lo 93% con 4 e non ottenendo nessun miglioramento con l'aggiunta della quinta.

- *Ultime due variabili sopravvissute (normalizzazione MSC).*
  Per quanto riguarda le frequenze delle ultime due lunghezze d'onda sopravvissute, si osserva dalle frequenze nella Figura 4.10 che le due lunghezze d'onda con maggiore frequenza (rispettivamente 115 e 109) hanno frequenze inferiori rispetto al caso precedente (pari a circa un terzo). Oltre a queste due variabili, sono state testate anche la terza e la quarta, avendo ottenuto frequenze pari a 100 e 94.

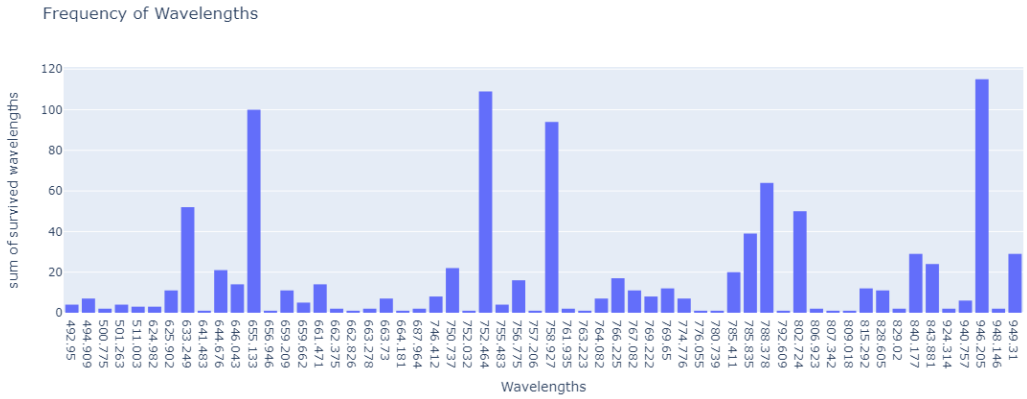

Figura 4.10: Frequenza delle migliori due lunghezze d'onda sopravvissute con la normalizzazione MSC

Come mostrato nella Tabella 4.2, le regressioni finali hanno ottenuto il 90.7% di accuracy con sole due variabili, nessun miglioramento con l'aggiunta della terza e un miglioramento al 93% di accuracy con l'uso della quarta.

Lunghezze d'onda (nm)				Accuracy
-	752.464	-	946.205	0.907
655.133	752.464	-	946.205	0.907
655.133	752.464	758.927	946.205	0.930

Tabella 4.2: Accuracy dei modelli con le combinazioni delle variabili selezionate fra le migliori due sopravvissute con normalizzazione MSC

- *Ultime tre variabili sopravvissute (normalizzazione MSC).*
  Per quanto riguarda le frequenze delle ultime tre lunghezze d'onda sopravvissute, si osserva dalla Figura 4.11 che le tre più frequenti hanno frequenze rispettivamente 161, 124 e 118; sono state dunque considerate anche la quarta, con frequenza 116, e la quinta, con frequenza 100.

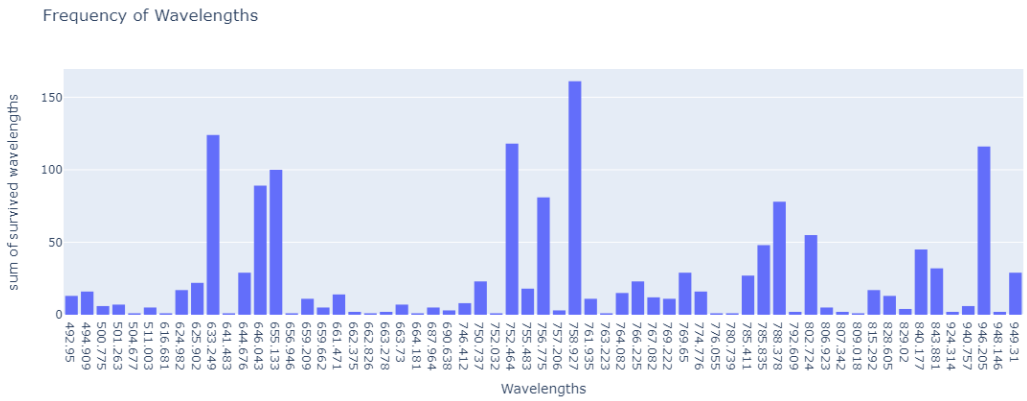

Figura 4.11: Frequenza delle migliori tre lunghezze d'onda sopravvissute con la normalizzazione MSC

Calcolando la regressione con queste lunghezze d'onda, dalla Tabella 4.3 si osserva un'accuracy del 90.7% con tre frequenze, con quattro frequenze si ha un leggero aumento dell'accuracy al 93%, mentre l'aggiunta della quinta frequenza non porta a un miglioramento dell'indice.

Lunghezze d'onda (nm)					Accuracy
633.249	-	752.464	758.927	-	0.907
633.249	-	752.464	758.927	946.205	0.930
633.249	655.133	752.464	758.927	946.205	0.930

Tabella 4.3: Accuracy dei modelli con le combinazioni delle variabili selezionate fra le migliori tre sopravvissute con normalizzazione MSC

- *Ultime quattro variabili sopravvissute (normalizzazione MSC).*
  Per quanto riguarda le frequenze delle ultime quattro lunghezze d'onda sopravvissute si osserva dalle frequenze nella Figura 4.12 che le quattro più frequenti sono sopravvissute in oltre 150 run; oltre a queste sono state testate anche la quinta e la sesta, essendo sopravvissute rispettivamente in 129 e 121 run.

  Calcolando la regressione con queste lunghezze d'onda si ottiene un'accuracy del 93% in tutti i casi, non ottenendo quindi nessun miglioramento con l'aggiunta della quinta e della sesta lunghezza d'onda.

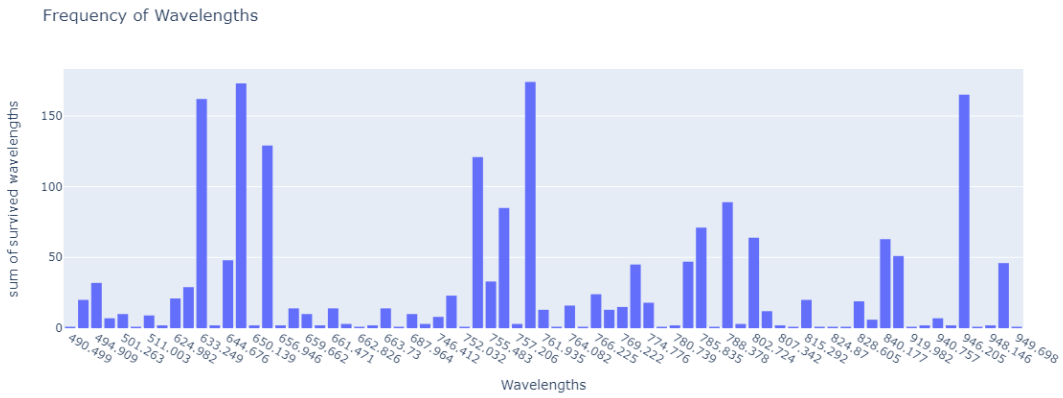

Figura 4.12: Frequenza delle migliori quattro lunghezze d'onda sopravvissute con la normalizzazione MSC

Lunghezze d'onda (nm)						Accuracy
633.249	646.043	-	-	758.927	946.205	0.930
633.249	646.043	655.133	-	758.927	946.205	0.930
633.249	646.043	655.133	752.464	758.927	946.205	0.930

Tabella 4.4: Accuracy dei modelli con le combinazioni delle variabili selezionate fra le migliori quattro sopravvissute con normalizzazione MSC

4.2.2 CARS con normalizzazione SVN

- *Miglior accuracy con il numero minore di variabili (normalizzazione SVN).* Per quanto riguarda i risultati associati alla miglior accuracy col numero minore di variabili, dalla Figura 4.13 è possibile osservare 4 frequenze che sono state selezionate in 300 o più run, che sono state testate; oltre a queste sono state testate altre due frequenze, che sono state selezionate rispettivamente in 273 e 243 run.

Lunghezze d'onda (nm)						Accuracy
489.518	-	-	663.73	664.181	665.535	0.930
489.518	498.821	-	663.73	664.181	665.535	0.930
489.518	498.821	646.043	663.73	664.181	665.535	0.930

Tabella 4.5: Accuracy dei modelli con le combinazioni delle variabili selezionate fra quelle associate alla migliore accuracy col numero minore di variabili con normalizzazione SVN

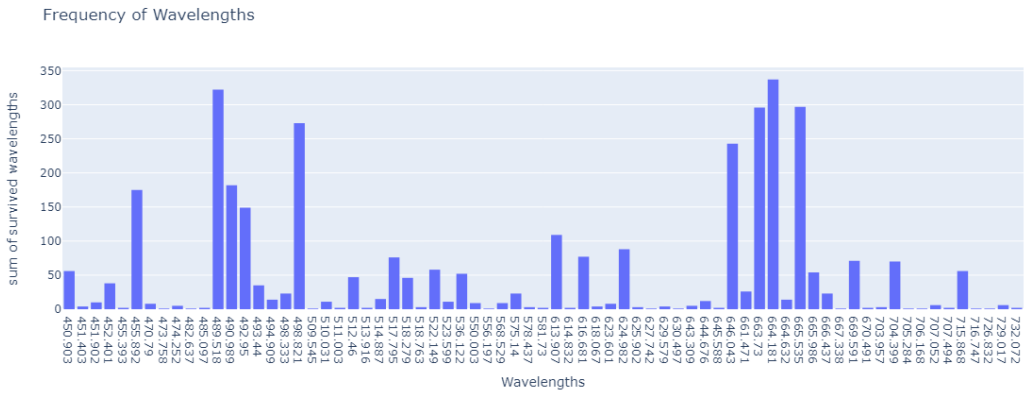

Figura 4.13: Frequenza delle lunghezze d'onda relative alla migliore accuracy con il numero minore di variabili con normalizzazione SVN

Testando queste lunghezze d'onda, dalla Tabella 4.5 si osserva che in tutti e 3 i casi si ottiene un accuracy pari al 93%, non ottenendo quindi un miglioramento con l'aggiunta di lunghezze d'onda aggiuntive rispetto alla regressione con 4 frequenze.

- *Ultime due variabili sopravvissute (normalizzazione SVN).* Per quanto riguarda i risultati associati alle frequenze delle ultime due lunghezze d'onda sopravvissute si osserva che le migliori due lunghezze d'onda sono sopravvissute 149 e 148 volte; è stata testata anche la terza lunghezza d'onda più frequente, essendo sopravvissuta 143 volte. Osserviamo che tutte e tre le frequenze selezionate risultano vicine, essendo collocate a 663.73 nm, 664.181nm, 665.535 nm. In questo caso è possibile osservare dalla Tabella 4.6 un accuracy pari a 90.7% in entrambi casi, non avendo quindi un miglioramento con l'aggiunta della terza frequenza.

Lunghezze d'onda (nm)			Accuracy
663.73	664.181	-	0.907
663.73	664.181	665.535	0.907

Tabella 4.6: Accuracy dei modelli con le combinazioni delle variabili selezionate fra le migliori due sopravvissute con normalizzazione SVN

- *Ultime tre variabili sopravvissute (normalizzazione SVN).*
  Per quanto riguarda le ultime tre varibili sopravvissute, sono state testate dalle migliori tre alle migliori cinque, dato che la terza ha una frequenza pari a 182, mentre la quinta pari a 167, quindi di poco inferiore. Si osservi le prime

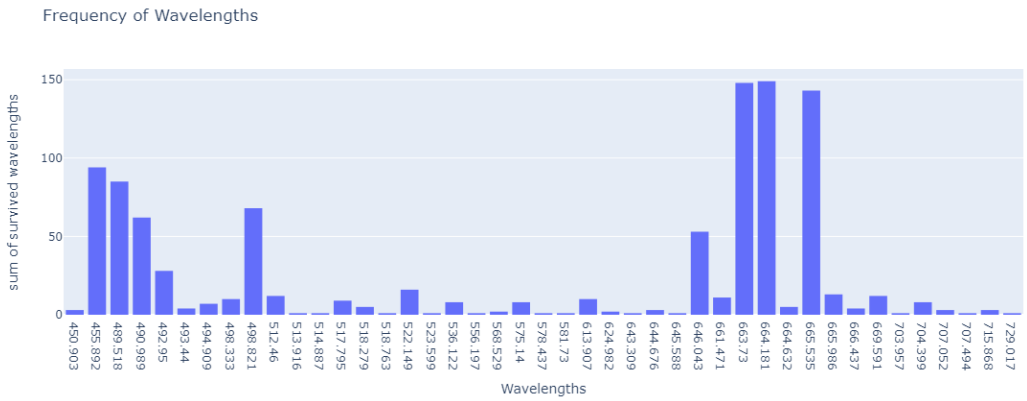

Figura 4.14: Frequenza delle migliori due lunghezze d'onda sopravvissute con la normalizzazione SVN

due frequenze si trovano a 489.518 nm e 498.821 nm, superando le prime tre frequenze del caso precedente, che sono in questo caso dalla terza alla quinta posizione. Dalla Tabella 4.7 osserviamo che utilizzando le prime tre frequenze

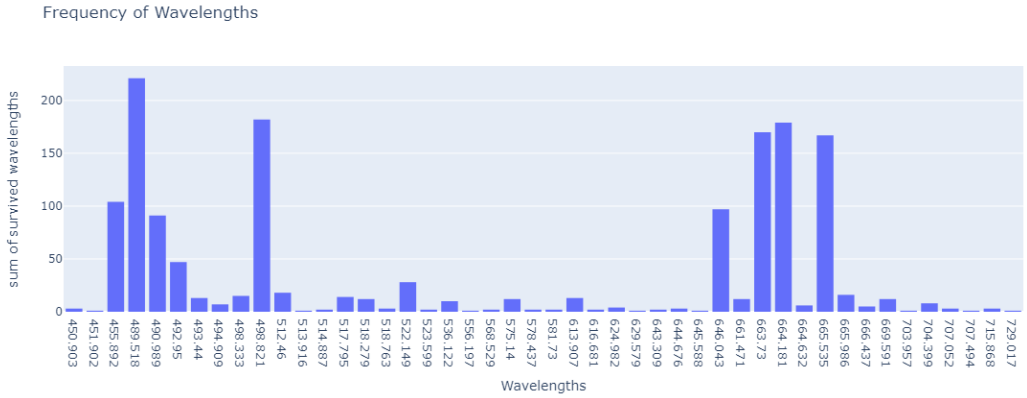

Figura 4.15: Frequenza delle migliori tre lunghezze d'onda sopravvissute con la normalizzazione SVN

si ottiene un'accuracy del 90.7% e si raggiunge il 93% aggiungendo la quarta frequenza, mentre non si ottiene un miglioramento aggiungendo la quinta.

Lunghezze d'onda (nm)					Accuracy
489.518	498.821	-	664.181	-	0.907
489.518	498.821	663.73	664.181	-	0.930
489.518	498.821	663.73	664.181	665.535	0.930

Tabella 4.7: Accuracy dei modelli con le combinazioni delle variabili selezionate fra le migliori tre sopravvissute con normalizzazione SVN

- *Ultime quattro variabili sopravvissute (normalizzazione SVN).*
  Per quanto riguarda le frequenze delle ultime quattro variabili sopravvissute, oltre alle migliori quattro sono state testate anche la quinta che è sopravvissuta solo una volte in meno rispetto alla quarta (rispettivamente 200 e 199 volte), come è possibile vedere dalla Figura 4.16. Oltre queste sono state considerate anche le due lunghezze d'onda successive per le loro frequenze simili. Dalla Tabella 4.8 è possibile osservare che effettuando la regressione

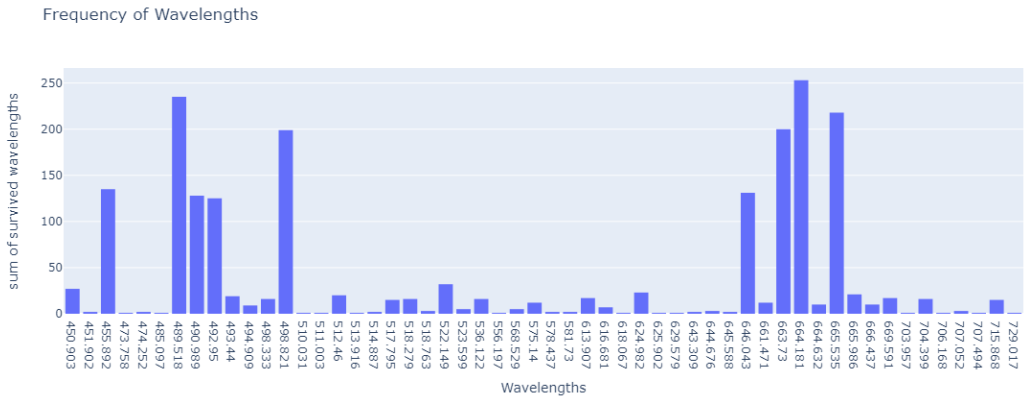

Figura 4.16: Frequenza delle migliori quattro lunghezze d'onda sopravvissute con la normalizzazione SVN

con le quattro frequenze si ottiene un'accuracy pari al 90.3%. Aggiungendo la quinta frequenza non si ottiene un miglioramento (si osservi che questo caso era stato ottenuto anche con le frequenze selezionate nella Tabella 4.7). Infine, con l'aggiunta delle altre due frequenze si riesce a ottenere un'accuracy del 97.7%.

Lunghezze d'onda (nm)							Accuracy
-	489.518	-	-	663.73	664.181	665.535	0.903
-	489.518	498.821	-	663.73	664.181	665.535	0.903
455.892	489.518	498.821	646.043	663.73	664.181	665.535	0.977

Tabella 4.8: Accuracy dei modelli con le combinazioni delle variabili selezionate fra le migliori quattro sopravvissute con normalizzazione SVN

4.2.3 CARS con normalizzazione min-max

- *Miglior accuracy con il numero minore di variabili (normalizzazione min-max).*
  Per quanto riguarda i risultati associati alla miglior accuracy col numero minore di variabili, dalla Figura 4.17 è possibile notare che le lunghezze d'onda più frequenti sono simili a quelle ottenute allo stesso modo ma con la normalizzazione SVN nella Figura4.5, anche se in quel caso le lunghezze d'onda selezionate avevano frequenze più elevate rispetto a quelle ottenute con la normalizzazione min-max.

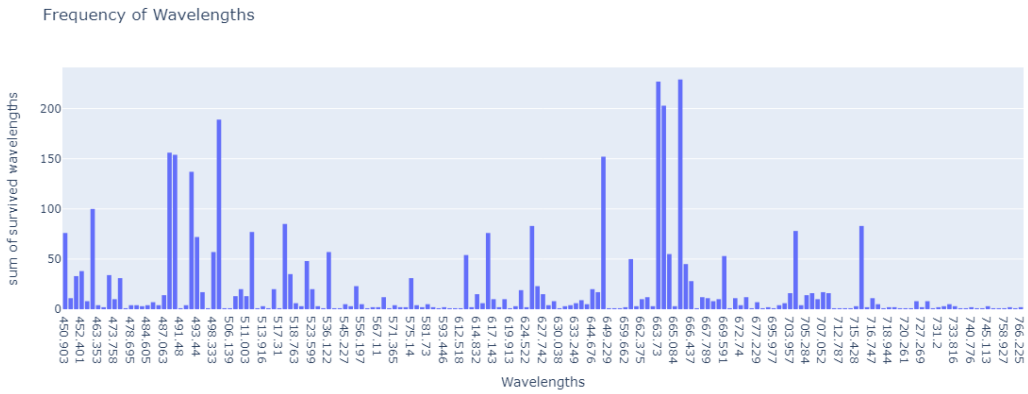

Figura 4.17: Frequenza delle lunghezze d'onda relative alla migliore accuracy con il numero minore di variabili con normalizzazione min-max

Nella Tabella 4.9 è possibile osservare, per quanto detto, risultati simili alla 4.5, ma in questo caso nell'ultimo esperimento è stata testata l'aggiunta della lunghezza d'onda 646.043 nm, che porta un aumento dell'accuracy al 97.7%

- *Ultime due variabili sopravvissute (normalizzazione min-max).* Come nel caso precedente, anche per quello in cui vengono selezionate le frequenze delle ultime due variabili sopravvissute le lunghezze d'onda che nella Figura 4.17 hanno

Lunghezze d'onda (nm)							Accuracy
-	-	-	-	663.73	664.181	665.535	0.907
-	-	498.821		663.73	664.181	665.535	0.930
489.518	-	498.821	-	663.73	664.181	665.535	0.930
489.518	490.989	498.821	-	663.73	664.181	665.535	0.930
489.518	490.989	498.821	646.043	663.73	664.181	665.535	0.977

Tabella 4.9: Accuracy dei modelli con le combinazioni delle variabili selezionate fra quelle associate alla migliore accuracy col numero minore di variabili con normalizzazione minmax

frequenza maggiore sono 663.73 nm, 664.181 nm, 665.535 nm, come nel caso della normalizzazione SVN, ma con frequenza di sopravvivenza leggermente inferiore (circa 100 contro le circa 150 nel caso precedente)

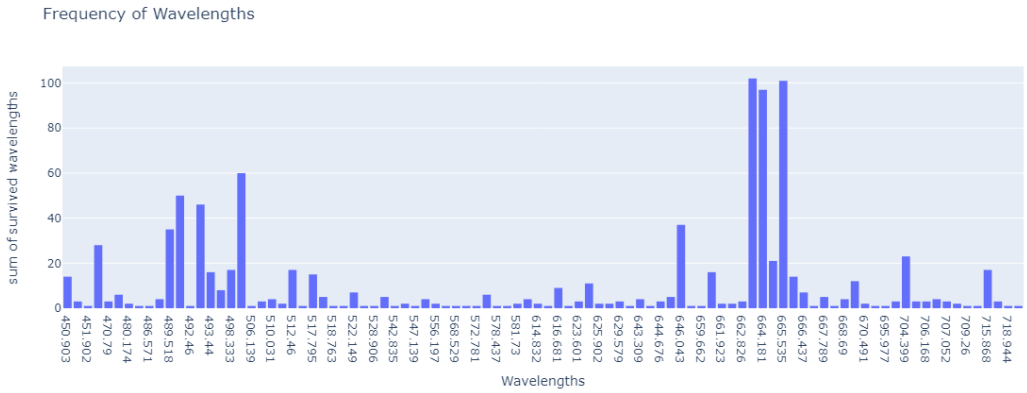

Figura 4.18: Frequenza delle migliori due lunghezze d'onda sopravvissute con la normalizzazione min-max

Dalla Tabella 4.10 si osserva un'altra differenza data dal fatto che la lunghezza 665.535 nm venga selezionata prima della 664.181 nm, anche se questo non comporta differenze per quanto riguarda l'accuracy. Neanche l'aggiunta della terza frequenza comporta un aumento dell'accuracy, come già visto nella Tabella4.6

Lunghezze d'onda (nm)			Accuracy
663.73	-	665.535	0.907
663.73	664.181	665.535	0.907

Tabella 4.10: Accuracy dei modelli con le combinazioni delle variabili selezionate fra le migliori due sopravvissute con normalizzazione min-max

- *Ultime tre variabili sopravvissute (normalizzazione min-max).* Per quanto riguarda le ultime tre variabili sopravvissute, è possibile osservare dalla Figura 4.19, rispetto al caso precedente, l'aggiunta della lunghezza d'onda 498.821, con una frequenza simile alle prime tre. Dalla Tabella 4.11 è possibile osservare che

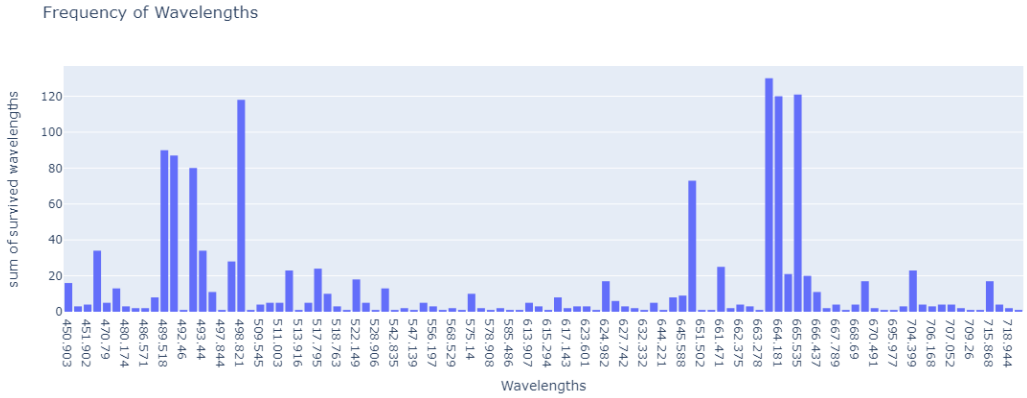

Figura 4.19: Frequenza delle migliori tre lunghezze d'onda sopravvissute con la normalizzazione min-max

l'aggiunta di questa variabile porti a un aumento dell'accuracy, raggiungendo il 93%.

Lunghezze d'onda (nm)				Accuracy
-	663.73	664.181	665.535	0.907
498.821	663.73	664.181	665.535	0.930

Tabella 4.11: Accuracy dei modelli con le combinazioni delle variabili selezionate fra le migliori tre sopravvissute con normalizzazione min-max

- *Ultime quattro variabili sopravvissute (normalizzazione min-max).*

  Per quanto riguarda le ultime quattro variabili sopravvissute, dalla Figura 4.20 sono state selezionate tutte le variabili con frequenza superiore a 150,

limitandosi a 7 nonostante la presenza altre frequenze con risultati di poco inferiori per non selezionare un numero eccessivo di variabili. Questo ha portato delle leggere differenze rispetto al caso della miglior accuracy, portando alla selezione della lunghezza d'onda 646.043 nm rispetto alla 492.95 nm.

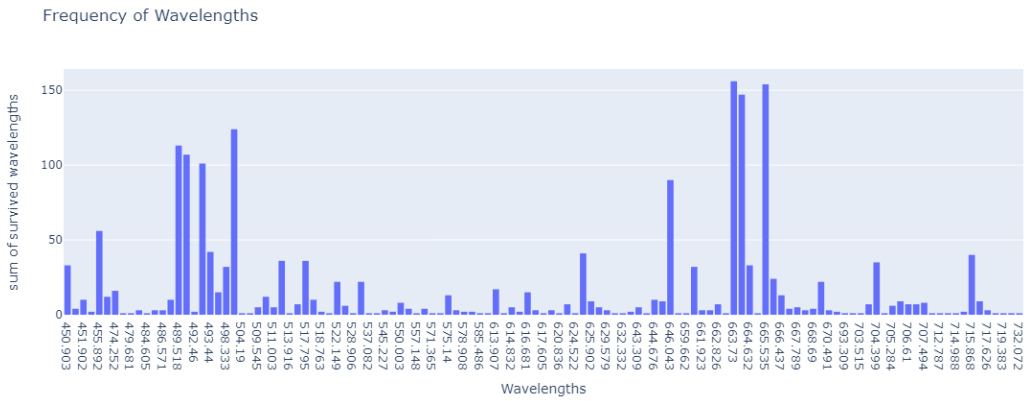

Figura 4.20: Frequenza delle migliori quattro lunghezze d'onda sopravvissute con la normalizzazione min-max

Nella Tabella 4.12 si osserva che i primi tre casi sono già stati esaminati; per quanto riguarda l'ultimo caso, l'aggiunta della lunghezza d'onda 492.95 nm porta a un aumento dell'accuracy a 97.7%.

Lunghezze d'onda (nm)							Accuracy
-	-	-	498.821	663.73	664.181	665.535	0.930
489.518	-	-	498.821	663.73	664.181	665.535	0.930
489.518	490.989	-	498.821	663.73	664.181	665.535	0.930
489.518	490.989	492.95	498.821	663.73	664.181	665.535	0.977

Tabella 4.12: Accuracy dei modelli con le combinazioni delle variabili selezionate fra le migliori quattro sopravvissute con normalizzazione min-max

4.3 MWPLS-DA

Per quanto riguarda gli esperimenti con la MWPLS-DA, sono state testate finestre formate da 15, 20, 30 e 40 lunghezze d'onda, corrispondenti a intervalli con ampiezza da circa 6 nm a circa 15 nm. Per le finestre da 15 e 20 variabili stati valutati i risultati della selezione di una e due finestre, mentre per le finestre di dimensioni 30 e 40 è stata utilizzata solo una finestra, al fine di non aumentare eccessivamente il numero di variabili selezionate.

4.3.1 MWPLS-DA con normalizzazione MSC

Nella Tabella 4.13 sono riportati i risultati ottenuti applicando la normalizzazione MSC. Considerando l'errore mediano è possibile osservare che usando una finestra da 15 variabili si ottiene un'accuracy pari all'88.4%, che aumenta al 90.7%, se nel risultato vengono considerate due finestre; aumentando ulteriormente la dimensione della finestra, si osserva che quasi in tutti i casi gli intervalli selezionati si trovano nella stessa regione, ma non si osserva un miglioramento dell'accuracy. Per quanto riguarda l'errore medio, i due risultati associati alle finestre da 15 variabili hanno ottenuto l'accuracy migliore, pari a 95.3%; anche in questo caso l'aumento della dimensione della finestra non si rifletti in un aumento dell'accuracy. Si osservi che nei due esperimenti migliori l'algoritmo ha selezionato l'intervallo [658.757,665.084], che non è invece presente negli altri esperimenti.

Tipo	Dimensione finestra	Numero finestre	Numero variabili	Accuracy
Mediana	15	1	15	0.884
	Intervallo: [810.693, 816.544]			
Media	15	1	15	0.953
	Intervallo: [658.757,665.084]			
Mediana	15	2	30	0.907
	Intervallo: [810.693,816.544] ∪ [901.752,907.32]			
Media	15	2	30	0.953
	Intervallo: [658.757, 665.84] ∪ [901.752,907.32]			
Mediana	20	1	20	0.907
	Intervallo: [902.15, 909.701]			
Media	20	1	20	0.907
	Intervallo: [864.731, 872.439]			
Mediana	20	2	21	0.907
	Intervallo: [901.752, 909.701]			
Media	20	2	40	0.907
	Intervallo: [864.731, 872.439] ∪ [902.15, 909.701]			
Mediana	30	1	30	0.907
	Intervallo: [704.842,717.626]			
Media	30	1	30	0.907
	Intervallo: [863.511,875.27]			
Mediana	40	1	40	0.907
	Intervallo: [901.354,916.825]			
Media	40	1	40	0.907
	Intervallo: [901.354,916.825]			

Tabella 4.13: Risultati dell'applicazione della MWPLS-DA ai dati con normalizzazione MSC

4.3.2 MWPLS-DA con normalizzazione SVN

Per quanto riguarda i dati trattati con la normalizzazione SVN i risultati sono riportati nella tabella 4.14. È possibile osservare che non sono presenti differenze negli intervalli selezionati considerando la media o la mediana degli errori. Usando una finestra da 15 variabili in tutti i casi viene selezionata la finestra [901.752, 907.32], con l'aggiunta della [866.357,872.034] quando vengono selezionate due finestre. A partire dalla finestra di dimensione 20 viene sempre selezionato l'intervallo [916.034, 923.527], anche se ovviamente viene allargato per le finestre di dimensione superiore. Si osservi che l'accuracy non cambia in nessun caso ed è sempre pari al 93%.

Tipo	Dimensione finestra	Numero finestre	Numero variabili	Accuracy
Mediana	15	1	15	0.930
	Intervallo: [901.752, 907.32]			
Media	15	1	15	0.930
	Intervallo: [901.752, 907.32]			
Mediana	15	2	30	0.930
	Intervallo: [866.357,872.034] ∪ [901.752,907.32]			
Media	15	2	30	0.930
	Intervallo: [866.357,872.034] ∪ [901.752,907.32]			
Mediana	20	1	20	0.930
	Intervallo: [916.034,923.527]			
Media	20	1	20	0.930
	Intervallo: [916.034,923.527]			
Mediana	20	2	38	0.930
	Intervallo: [916.034,930.596]			
Media	20	2	38	0.930
	Intervallo: [916.034,930.596]			
Mediana	30	1	30	0.930
	Intervallo: [916.034,927.457]			
Media	30	1	30	0.930
	Intervallo: [916.034,927.457]			
Mediana	40	1	40	0.930
	Intervallo: [915.244,930.204]			
Media	40	1	40	0.930
	Intervallo: [915.244,930.204]			

Tabella 4.14: Accuracy MWPLS-DA SVN

4.3.3 MWPLS-DA con normalizzazione min-max

Nella Tabella 4.15 sono riportati i risultati ottenuti applicando la normalizzazione min-max; è possibile osservare che questi sono uguali a quelli della Tabella 4.14, ottenuti con la normalizzazione SVN.

Tipo	Dimensione finestra	Numero finestre	Numero variabili	Accuracy
Mediana	15	1	15	0.930
	Intervallo: [901.752, 907.32]			
Media	15	1	15	0.930
	Intervallo: [901.752, 907.32]			
Mediana	15	2	30	0.930
	Intervallo: [866.357, 872.034] ∪ [901.752, 907.32]			
Media	15	2	30	0.930
	Intervallo: [866.357,872.034] ∪ [901.752,907.32]			
Mediana	20	1	20	0.930
	Intervallo: [916.034,923.527]			
Media	20	1	20	0.930
	Intervallo: [916.034,923.527]			
Mediana	20	2	38	0.930
	Intervallo: [916.034,930.596]			
Media	20	2	38	0.930
	Intervallo: [916.034,930.596]			
Mediana	30	1	30	0.930
	Intervallo: [916.034,927.457]			
Media	30	1	30	0.930
	Intervallo: [916.034,927.457]			
Mediana	40	1	40	0.930
	Intervallo: [915.244,930.204]			
Media	40	1	40	0.930
	Intervallo: [915.244,930.204]			

Tabella 4.15: Accuracy MWPLS-DA min-max

Capitolo 5

Conclusioni

La spettroscopia è un'utile strumento nell'agricoltura di precisione, ma le analisi spettroscopiche effettuate su uno spettro completo, formato da migliaia di lunghezze d'onda, richiedono l'uso di strumenti costosi per le rilevazioni. In questo lavoro sono stati testati due algoritmi per effettuare una selezione delle lunghezze d'onda a partire da uno spettro formato da 1133 variabili: la Competitive Adaptive Reweight Sampling (CARS) e la Moving Windows Partial Least Square Discriminant Analysis (MWPLS-DA). Abbinati a questi algoritmi sono state utilizzate tre diverse tecniche di pre-processing: la Multiplicative Scatter Correction (MSC), la Standard Normal Variate (SVN) e la normalizzazione min-max.
Dalle esecuzioni di CARS con le tre tecniche di pre-processing citate sono state selezionate un numero ridotto di lunghezze d'onda, fra le 2 e le 7, valutando quali fossero le lunghezze d'onda più frequenti associate alla migliore accuracy con il numero più basso di frequenze e quali fossero le lunghezze d'onda che più frequentemente riuscivano a sopravvivere fino alle ultime iterazioni di ogni run dell'algoritmo. Sono stati ottenuti risultati con accuracy superiore al 90% in tutti gli esperimenti, simile a quella ottenuta utilizzando lo spettro completo anche nei casi nei quali le variabili usate nella classificazione finale sono state solo 2, ed è stato osservato che spesso le lunghezze d'onda selezionate secondo i vari criteri in esperimenti diversi corrispondevano, che potrebbe confermare l'effettiva utilità delle lunghezze d'onda selezionate.
Anche per quanto riguarda i risultati della MWPLS-DA con le tre tecniche di pre-processing i risultati in termini di accuracy sono simili, anche se in questo caso non sono associati alle singole frequenze, ma a delle finestre costituite da un numero variabile di lunghezze d'onda, a partire da un minimo di 15 fino a un massimo di 40. Anche in questo caso molti dei risultati ottenuti con le diverse normalizzazioni confermavano la selezione delle stesse lunghezze d'onda. In molti casi queste non erano le stesse ottenute con CARS, ma questo potrebbe essere dovuto agli obiettivi differenti dei due algoritmi, ovvero la selezione di singole lunghezze d'onda per

CARS e la selezioni di regioni dello spettro per la MWPLS-DA.
Questi risultati mostrano il potenziale delle due tecniche nella riduzione del numero di variabili spettroscopiche da utilizzare nell'analisi.

In conclusione, il lavoro ha dimostrato il potenziale degli algoritmi CARS e MWPLS-DA nella riduzione delle variabili spettroscopiche per la costruzione di classificatori efficaci nella classificazione fra piante di insalata sane e sotto stress idrico. Questi risultati aprono nuove prospettive per l'uso di tecniche avanzate di elaborazione dati in agricoltura, contribuendo a migliorare la sostenibilità e l'efficienza delle pratiche agricole.

5.1 Lavori Futuri

Futuri lavori potrebbero analizzare piante diverse e porzioni più ampie dello spettro, includendo una porzione maggiore della regione dello spettro del vicino infrarosso (NIR): questa parte dello spettro viene ampiamente utilizzata anche in altri campi per ottenere utili informazioni sulla composizione chimica della pianta. Un altro utile sviluppo potrebbe essere l'estensione della classificazione binaria a una classificazione multi-classe, in grado di determinare non solo la presenza di stress idrico della pianta, ma anche la severità. Infine, a questi si potrebbe aggiungere l'effettiva implementazione dei risultati ottenuti, utilizzando dei sensori che sfruttino le lunghezze d'onda selezionate per il monitoraggio delle piante.

Bibliografia

[1] Reza Adhitama Putra Hernanda, Junghyun Lee e Hoonsoo Lee. «Spectroscopy Imaging Techniques as In Vivo Analytical Tools to Detect Plant Traits». In: *Applied Sciences* 13 (2023) (cit. a p. 1).

[2] Rathnaprabha Dharavath Srividya Attaluri. «Novel plant disease detection techniques-a brief review». In: *Molecular Biology Reports* 50 (2023), pp. 9677–9690 (cit. alle pp. 1, 3).

[3] David J. Mulla. «Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps». In: *Biosystems Engineering 114* (2013), pp. 358–371 (cit. alle pp. 2, 3, 5).

[4] Prachi Singh, Prem Chandra Pandey, George P. Petropoulos, Andrew Pavlides, Prashant K. Srivastava1, Nikos Koutsias, Khidir Abdala Kwal Deng e Yangson Bao. *Hyperspectral remote sensing in precision agriculture: present status, challenges, and future trends*. Elviser, 2020, pp. 121–146 (cit. a p. 2).

[5] Crookston R. K. «A Top 10 List of Developments and Issues Impacting Crop Management and Ecology During the Past 50 Years». In: *Crop Science* 46 (5 2006), pp. 2253–2262 (cit. a p. 2).

[6] Federico Martinelli et al. «Advanced methods of plant disease detection. A review». In: *Agronomy for Sustainable Development* (2014) (cit. alle pp. 3, 5).

[7] Edward B. Knipling. «Physical and Physiological Basis for the Reflectance of Visible and Near-Infrared Radiation from Vegetation». In: *Remote Sensing of Environment* 1 (Summer 1970), pp. 155–159 (cit. a p. 5).

[8] Rekha Gautam, Sandeep Vanga, Freek Ariese e Siva Umapathy. «Review of multidimensional data processing approaches for Raman and infrared spectroscopy». In: *EPJ Techniques and Instrumentation* (2015) (cit. a p. 6).

[9] P. Geladi, D. MacDougall e H. Martens. «Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat». In: *Applied Spectroscopy* 39 (mag. 1985), pp. 491–500 (cit. a p. 7).

[10] T. Naes, T. Isaksson, T. Fearn e T. Davies, cur. *A user-friendly guide to Multivariate Calibration and Classification*. New York: Nir Publications, 2004 (cit. alle pp. 8, 14).

[11] Gareth James, Daniela Witten e Trevor Hastie Robert Tibshirani. *An Introduction To Statistical Learning*. 2017 (cit. alle pp. 12, 14, 16).

[12] Matthew Barker e William Rayens. «Partial least squares for discrimination». In: *Journal of chemometrics* 17 (2003), pp. 166–173 (cit. alle pp. 13, 16).

[13] Mauro Gasparini. *Modelli probabilistici e statistici*. 2014 (cit. a p. 14).

[14] Svente Wold, Michael Sjöström e Lennart Eriksson. «PLS-regression: a basic tool of chemometrics». In: *Chemometrics and intelligent laboratory systems* 58 (2001), pp. 109–130 (cit. a p. 14).

[15] Agnar Höskuldsson. «PLS regression method». In: *Journal of Chemometrics* 2 (1988), pp. 211–228 (cit. a p. 14).

[16] Hongdong Li, Yizeng Lianga, Qingsong Xub e Dongsheng Caoa. «Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration». In: *Analytica Chimica Acta* 648 (1 2009), pp. 77–84 (cit. a p. 17).

[17] Jian-Hui Jiang, R. James Berry, Heinz W. Siesler e Yukihiro Ozaki. «Wavelength Interval Selection in Multicomponent Spectral Analysis by Moving Window Partial Least-Squares Regression with Applications to Mid-Infrared and Near-Infrared Spectroscopic Data». In: *Analytical Chemistry* 74 (2002) (cit. alle pp. 17, 18).

[18] Hai-Yan Fu, Shuang-Yan Huan, Lu Xu, Li-Juan Tang, Jian-Hui Jiang, Hai-Long Wu, Guo-Li Shen e Ru-Qin Yu. «Moving window partial least-squares discriminant analysis for identification of different kinds of bezoar samples by near infrared spectroscopy and comparison of different pattern recognition». In: *Journal of Near Infrared Spectroscopy* 15 (15 2007), pp. 291–297 (cit. a p. 18).