RELAZIONE CHEMIOMETRIA

Analisi del dataset delle Resine

Cecci Riccardo

Matr: 20023915
Università del Piemonte Orientale

Spiegazione del dataset

Le righe sono i campioni di resina, per un totale di 169 campioni

Questi campioni sono divisi in 4 famiglie, divise sulla base di:

- 3 variabili quantitative
    - NOH → numero di ossidrile
    - RS → residuo secco
    - NAC → numero di acido

Per ogni campione è fornito uno spettro NIR (sono le V, da V1 a V99), quindi è uno spettro composto da 99 lunghezze d'onda

Abbiamo 99 lunghezze d'onda perché è già stato fatto un pre-trattamento in cui è stato fatto un downsampling delle variabili (cioè si è selezionata una variabile ogni tot). Quelle riportate sono già le derivate prime

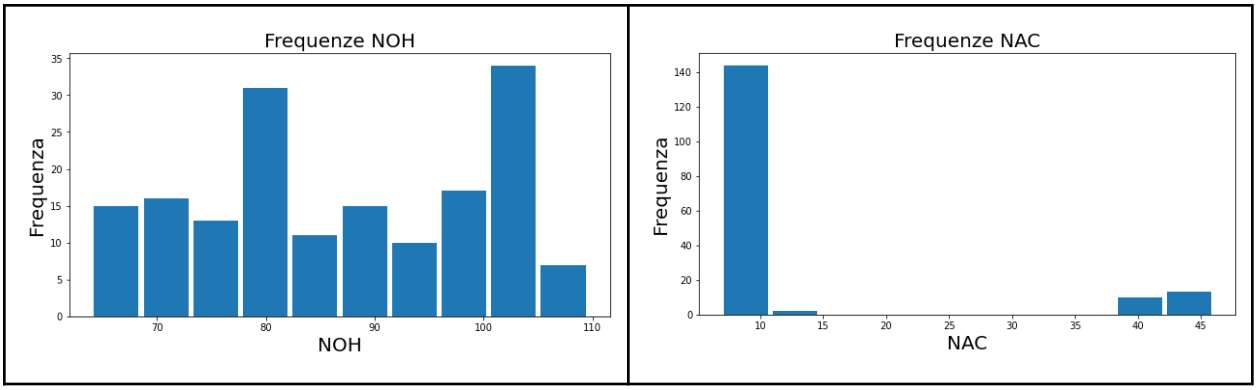
Questo dataset si presta a diversi tipi di lavori:

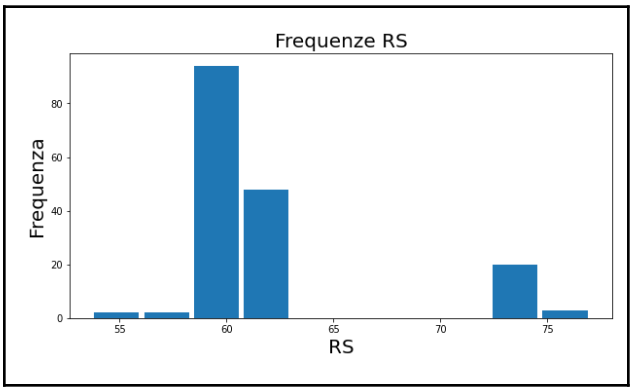
- Classificazione → abbiamo 4 famiglie e quindi potrebbe essere interessante capire che cosa le distingue (e che cosa possono avere in comune)
- Capire se si può utilizzare lo spettro per costruire un modello dei 3 parametri NOH, RS e NAC. Quindi fare regressione per stabilire un modello tra ciascuno dei 3 parametri e lo spettro
- Pattern recognition

Statistiche di NOH, RS, NAC

Variabile	Min	Max	Media	Varianza	Dev Std
NOH	64,000	109,600	87,038	164,1227	12,81104
RS	53,700	77,000	62,237	23,16972	4,813493
NAC	6,900	46,000	13,635	140,5823	11,85674

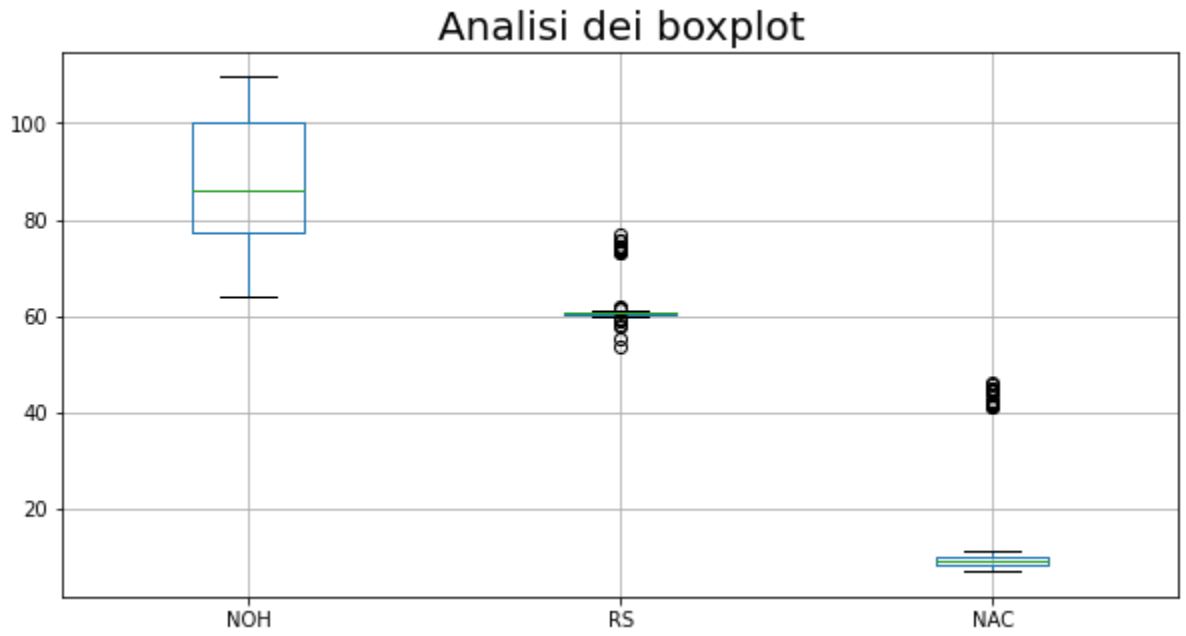
Istogrammi di frequenza





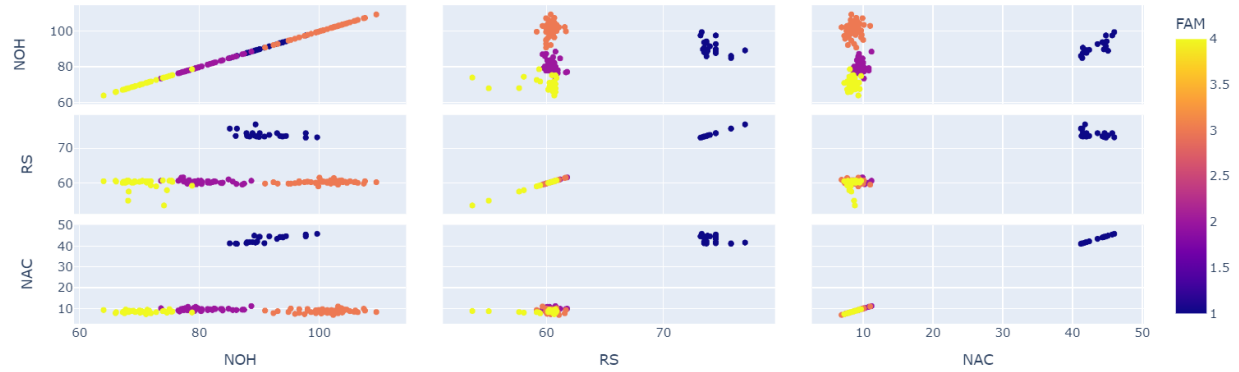
Analizzando i valori nella tabella delle statistiche e i grafici degli istogrammi, possiamo osservare chiaramente la presenza di effetti di scala nelle variabili. Infatti, i range dei dati differiscono notevolmente e si possono notare delle distribuzioni non normali e asimmetriche. La figura relativa a NOH evidenzia la presenza di due massimi, indicando due comportamenti distinti per la variabile in questione. Tuttavia, per le restanti due variabili si può notare un unico massimo.

Boxplots



Attraverso l'esame dei boxplot presenti nella figura, siamo in grado di confrontare e osservare il centro e la distribuzione dei dati relativi alle variabili, oltre a individuare eventuali valori anomali. Basandoci su questa rappresentazione visiva, possiamo dedurre la necessità di applicare una procedura di autoscalatura al fine di centrare i dati e normalizzarli in modo da ottenere una varianza unitaria.

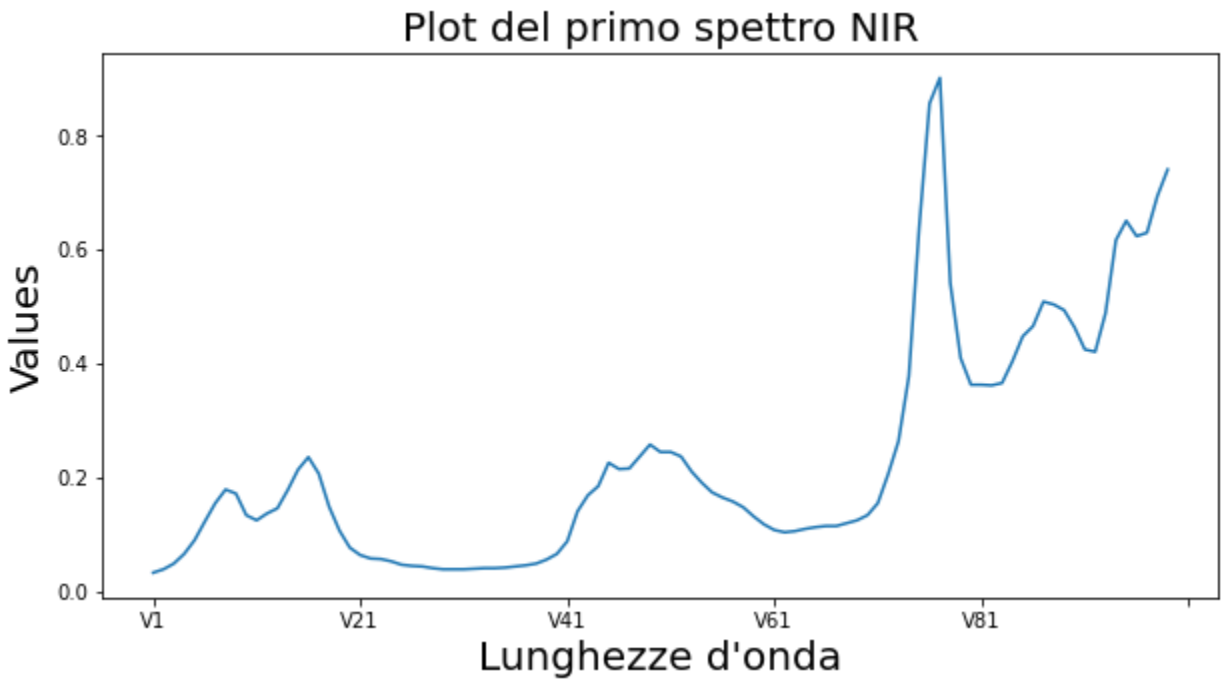
Matrix Plot



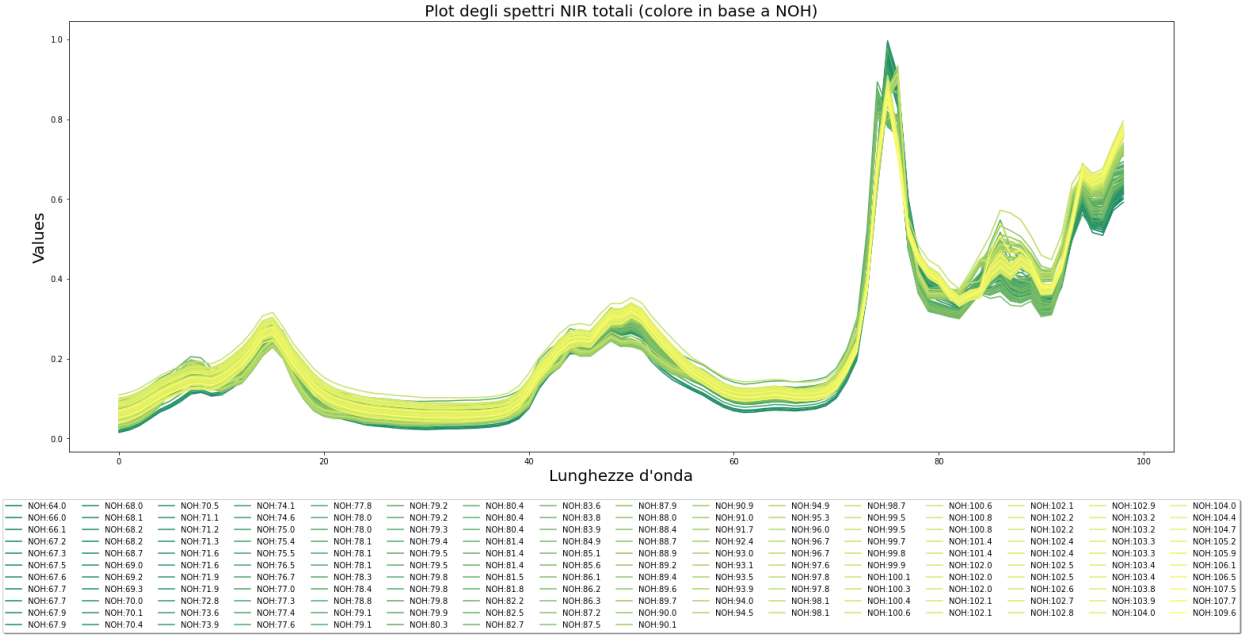
Attraverso l'analisi degli scatterplot presenti nella matrice dei grafici, siamo in grado di individuare eventuali raggruppamenti o correlazioni nei dati. Negli scatterplot relativi alle variabili RS e NAC, possiamo osservare la presenza di raggruppamenti distinti, con una chiara separazione da un singolo tipo di famiglia. Se tracciassimo una retta per collegare i centri di questi raggruppamenti, otterremmo un andamento lineare positivo. Tuttavia, è evidente che i punti sono molto distanti tra loro e non seguono tale andamento, suggerendo l'assenza di una correlazione. Inoltre, non sono presenti correlazioni tra la variabile NOH e le variabili RS e NAC. Gli scatterplot relativi a NOH mostrano una grande somiglianza tra di loro e le diverse famiglie si distinguono chiaramente.

Spettri NIR

Spettro NIR del primo campione



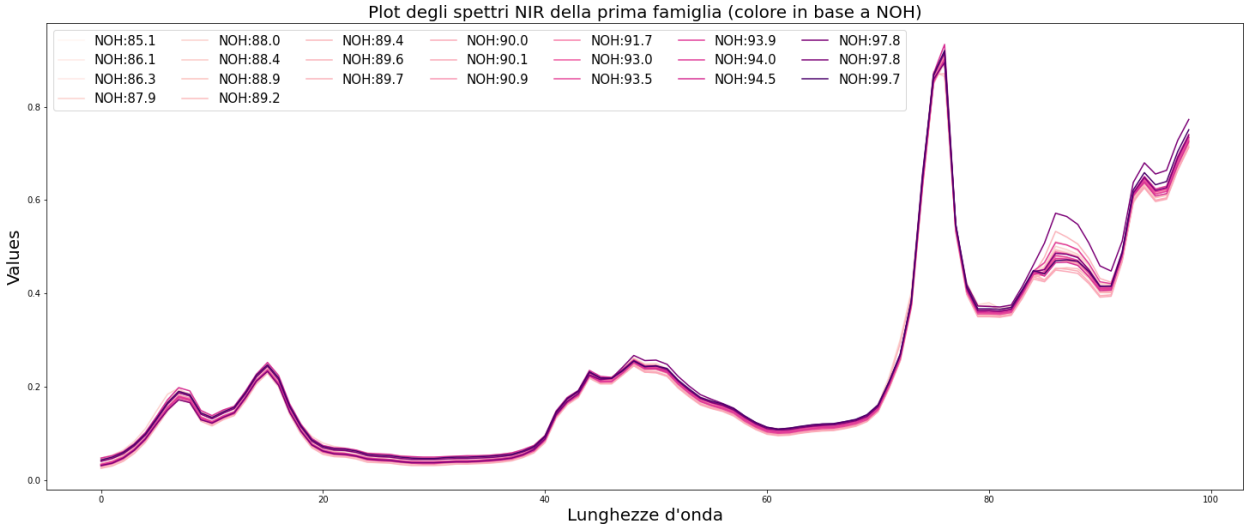
Spettri NIR totali



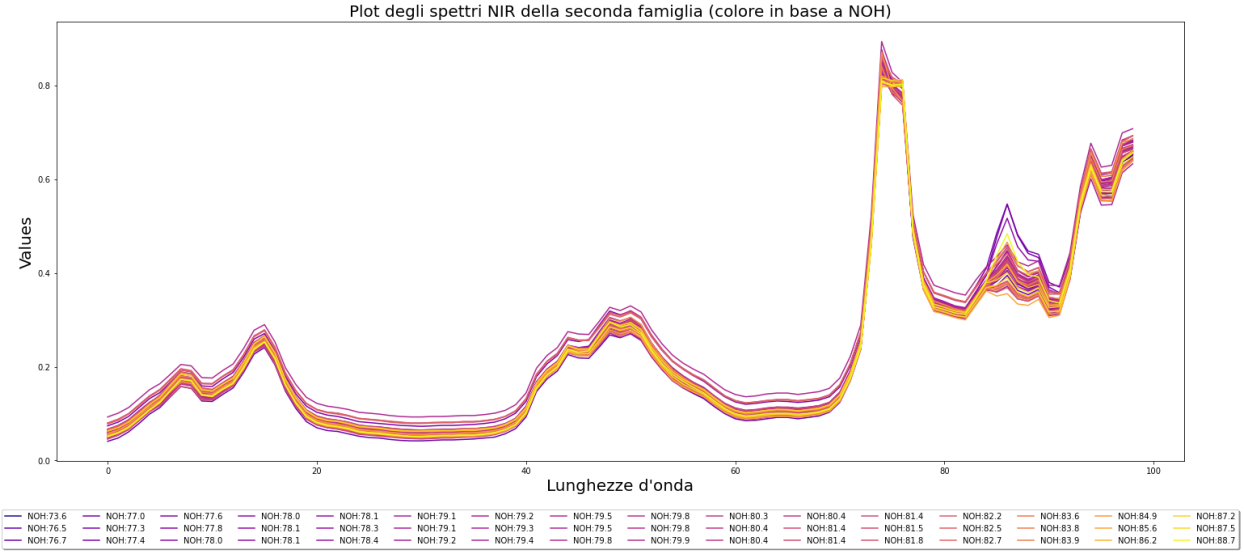
Nei due grafici degli spettri NIR appena visti riscontriamo l'assenza di rumore o variazioni non correlate rispetto alla misura sperimentale. È stato già applicato un processo di smoothing e una riduzione delle variabili. Non sono presenti regioni significative in cui lo spettro risulta completamente piatto, e non vi è un minimo comune per tutti gli spettri analizzati.

Di seguito si riportano gli spettri NIR suddivisi per famiglie con una visualizzazione in base al valore assunto dalla variabile NOH

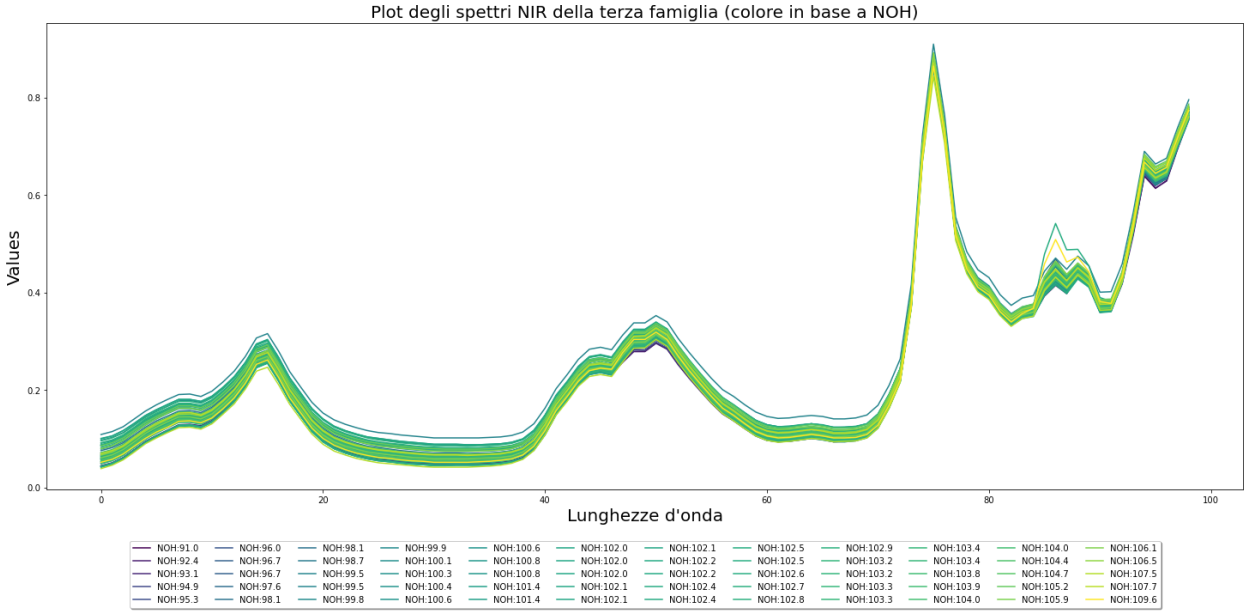
Spettri NIR della prima famiglia di campioni



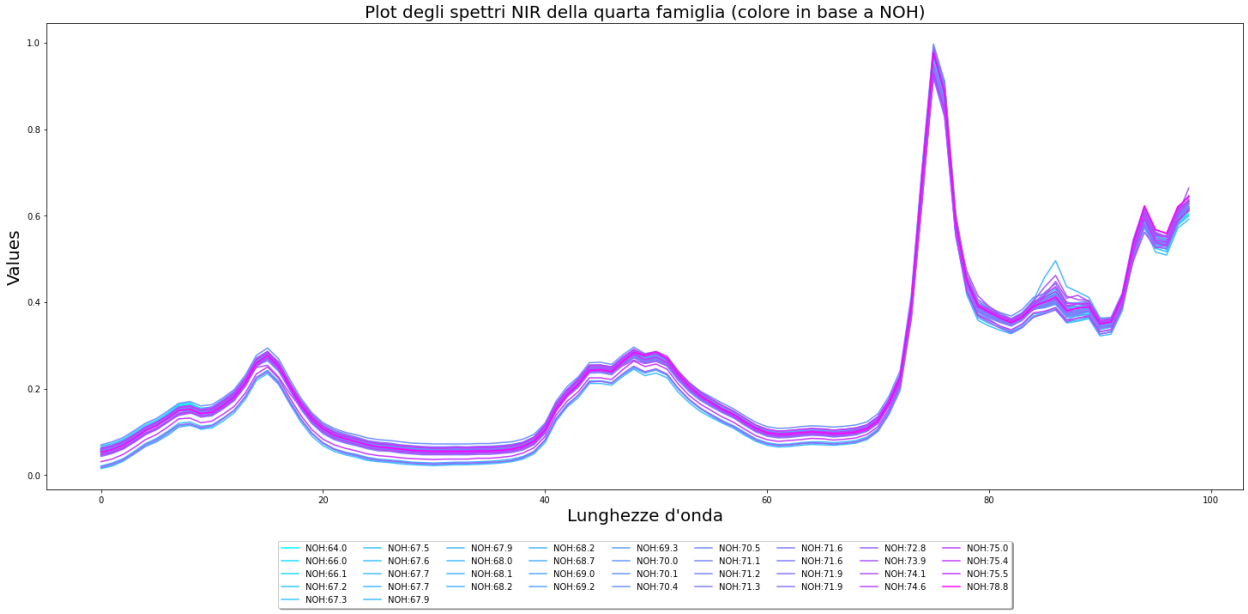
Spettri NIR della seconda famiglia di campioni



Spettri NIR della terza famiglia di campioni

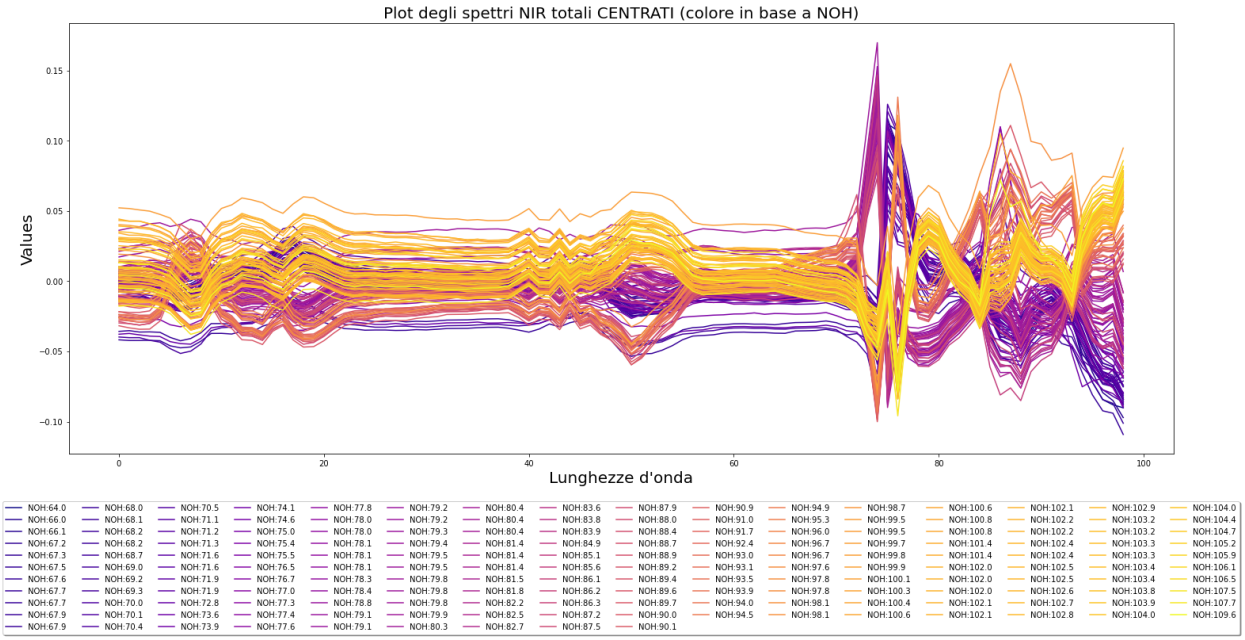


Spettri NIR della quarta famiglia di campioni

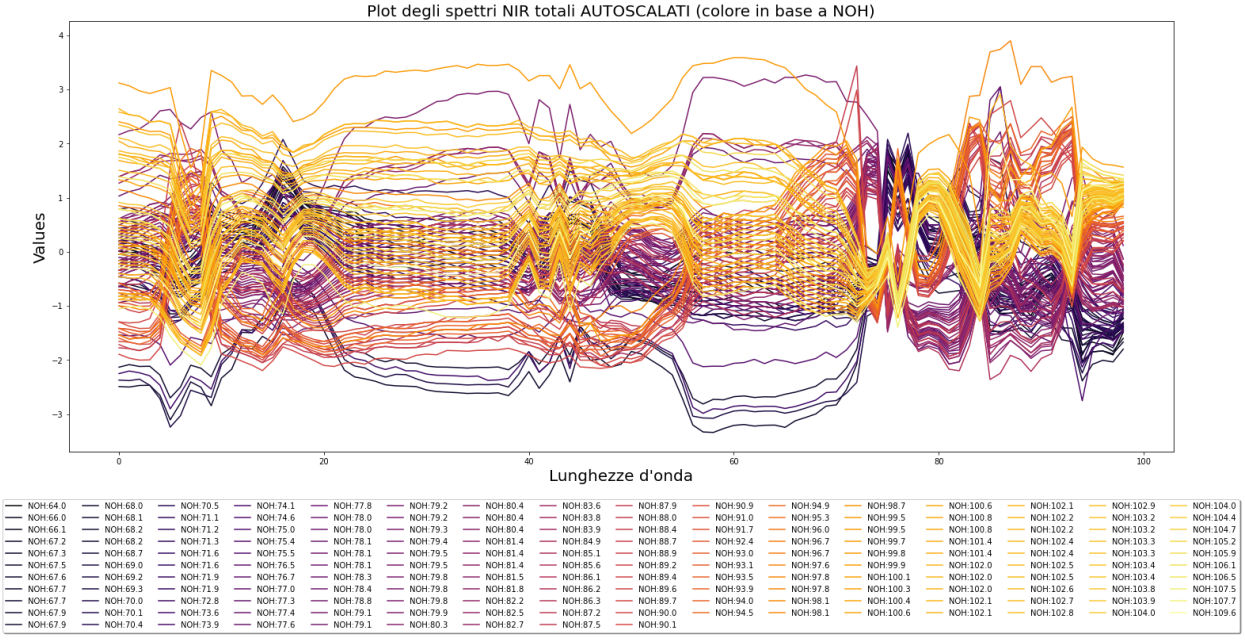


Pre-trattamento

Centratura

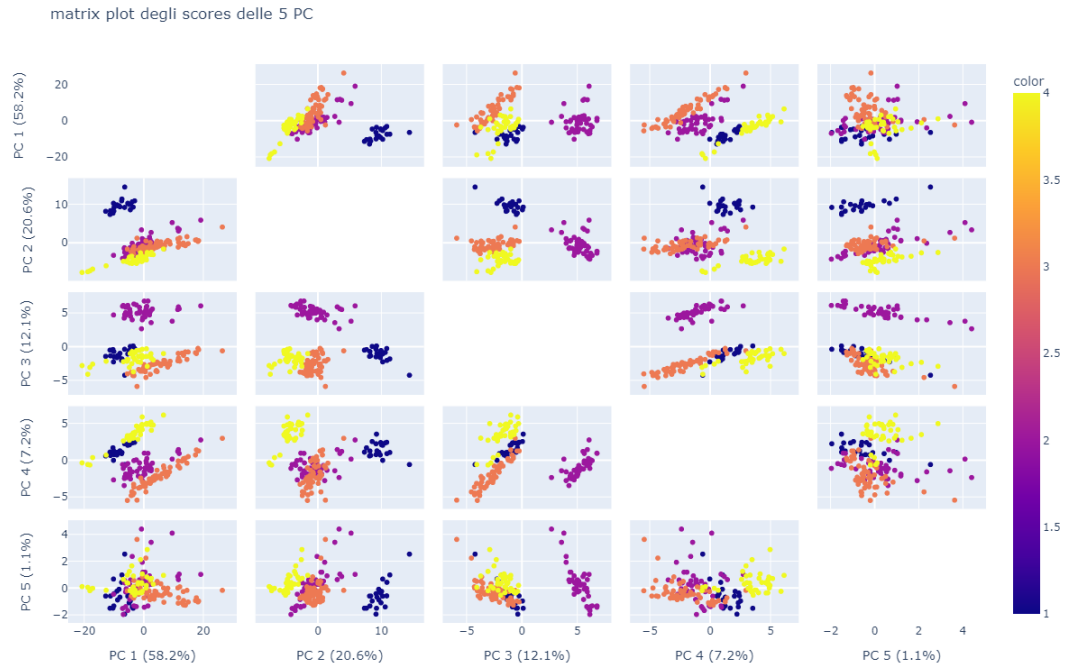


Autoscalatura



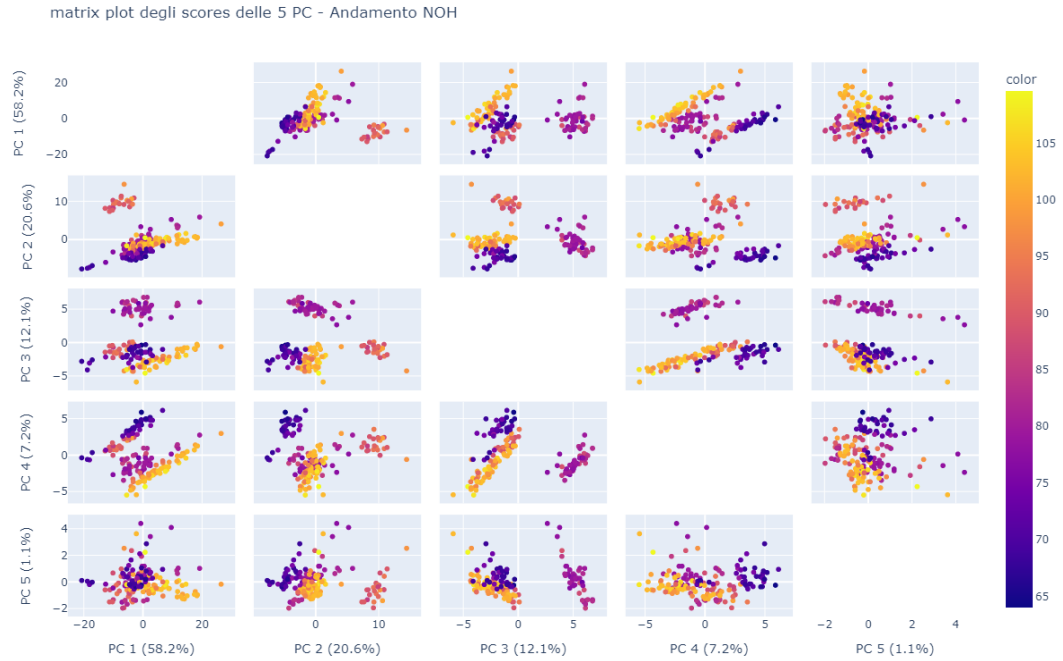
PCA

Scores

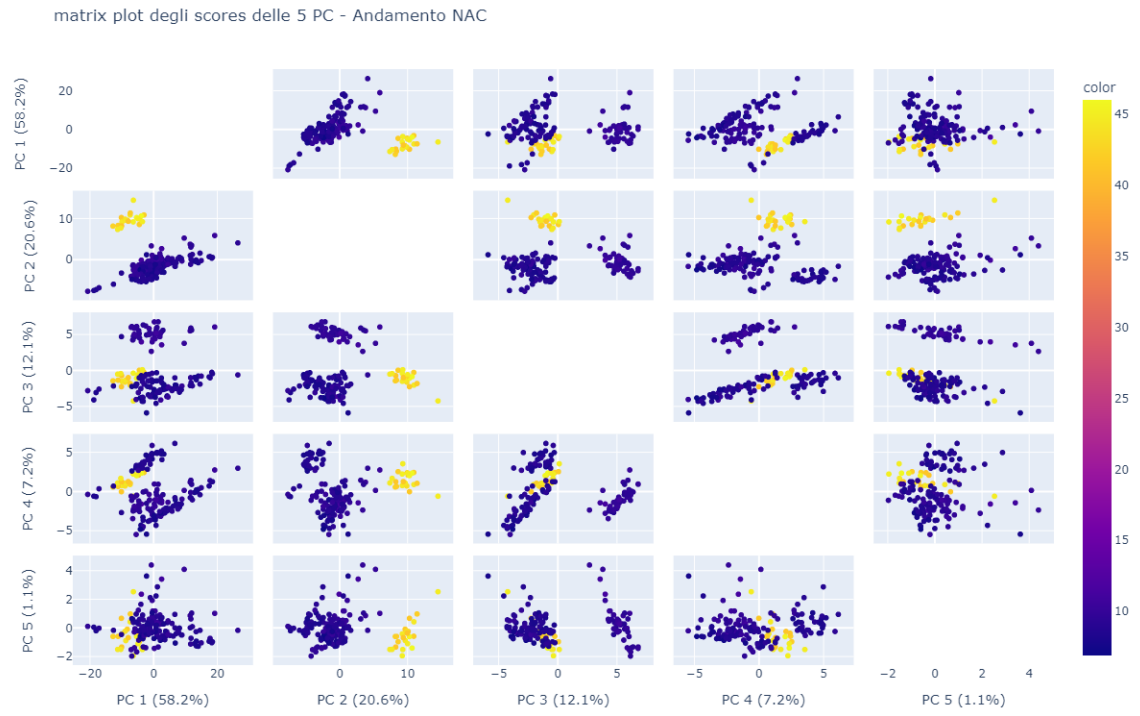


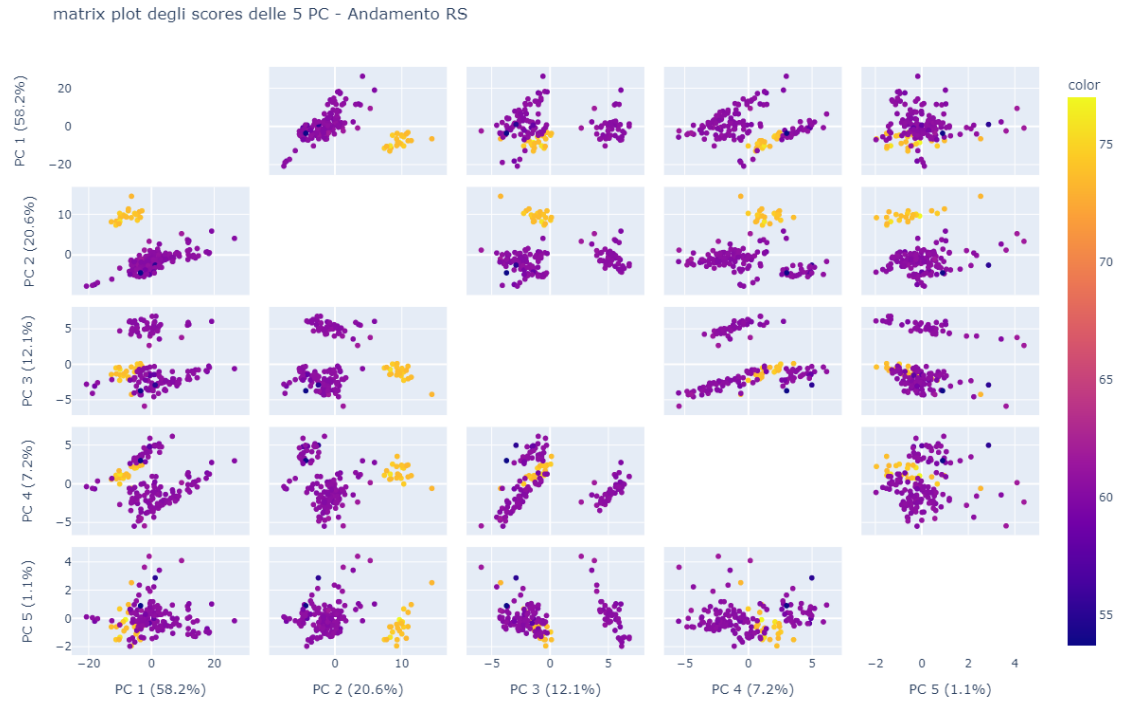
L'analisi della prima componente principale (PC1) rivela una notevole correlazione con lo spettro medio. Questa coincidenza ci fornisce un'importante quantità di informazioni riguardo alle caratteristiche comuni delle famiglie di resine. Tuttavia, per ottenere ulteriori dettagli e approfondimenti sulle diverse famiglie, è necessario osservare e confrontare le componenti principali successive (PC2, PC3, …). Questa analisi comparativa delle componenti principali, come mostrato nel matrix plot soprastante, ci permette di identificare e discriminare in modo più preciso le variazioni tra le famiglie di resine.

Di seguito si riportano i matrix plots relativi alle separazioni fra le famiglie che si possono ottenere in base alle variabili NOH, RS e NAC



Attraverso l'osservazione dei grafici presenti nella figura, è evidente che le quattro famiglie di resine presentano una separazione in base ai valori del numero di ossidrile (NOH). Questa distinzione visibile ci suggerisce che il NOH è un fattore discriminante significativo per classificare le diverse famiglie di resine. L'analisi del grafico fornisce una chiara evidenza della relazione tra il NOH e la tipologia di resina, consentendoci di identificare facilmente ciascuna famiglia e distinguere le loro caratteristiche specifiche. Questo risultato sottolinea l'importanza del NOH come parametro rilevante nell'analisi e nella classificazione delle resine.

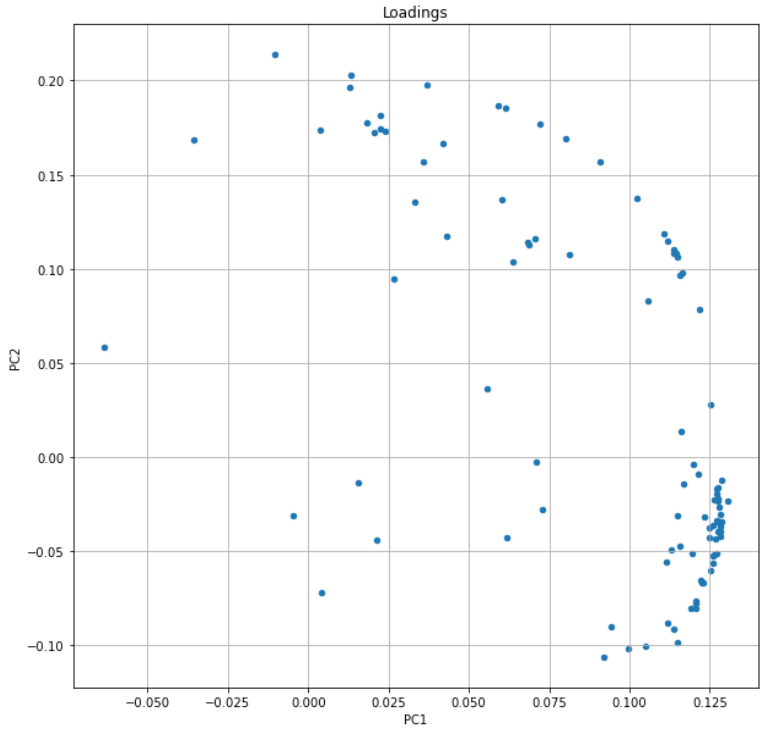




Analizzando i due matrix plots soprastanti, possiamo notare che una particolare famiglia

di resine si distingue dalle altre per il suo range di valori di residuo secco (RS) e numero di acido (NAC). Questa famiglia mostra una distribuzione significativamente diversa rispetto alle altre, caratterizzata da valori di RS e NAC che si discostano notevolmente. Tale discrepanza evidenzia una chiara separazione tra questa famiglia e le altre in termini di composizione chimica. A differenza del caso precedente, in cui si faceva riferimento al NOH come fattore discriminante, in questo caso ci concentriamo sul RS e sul NAC come indicatori chiave per distinguere la famiglia di resine in questione. L'analisi dei grafici ci permette di identificare queste differenze significative nella distribuzione dei valori di RS e NAC, sottolineando l'importanza di queste due variabili nella classificazione e nella caratterizzazione delle famiglie di resine.

Loadings



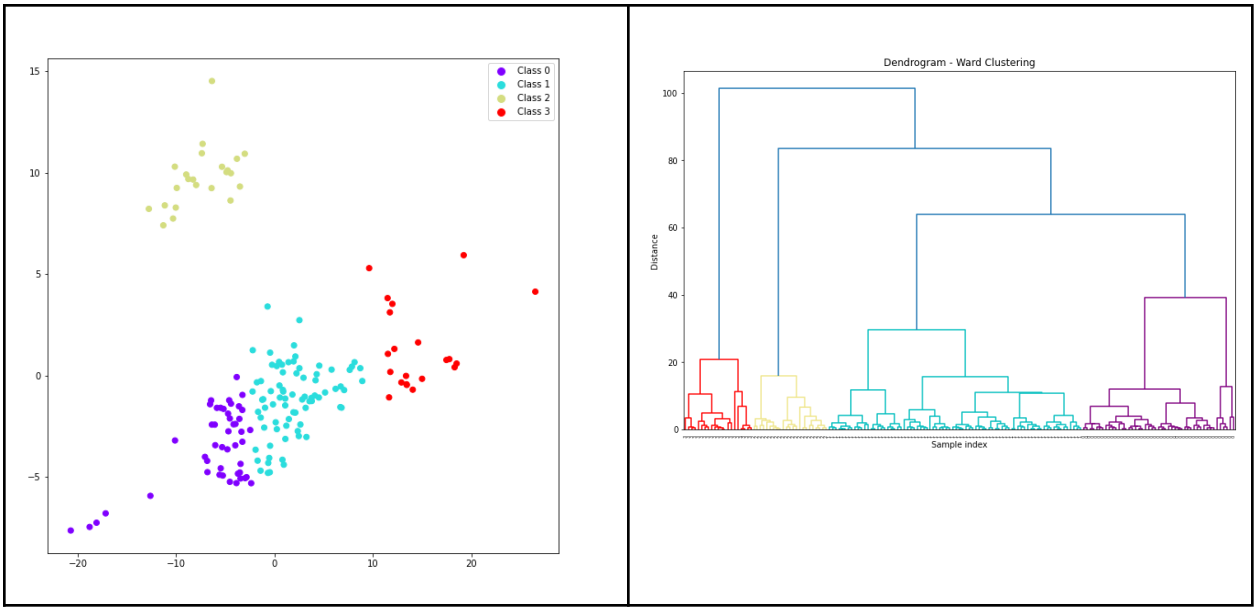
Cluster Analysis

L 'analisi delle componenti principali (PCA) è stata impiegata come una strategia per ridurre la complessità dei dati, consentendo di sintetizzare le informazioni più rilevanti in un numero limitato di variabili continue.

Una volta completata la PCA, i risultati ottenuti sono stati utilizzati per eseguire un'analisi dei cluster. Questa metodologia ha lo scopo di raggruppare i dati in base alla somiglianza delle loro caratteristiche, permettendoci di individuare gruppi omogenei e differenziati all'interno del dataset. L'analisi dei cluster, basata sui risultati della PCA, ci fornisce un'opportunità unica di scoprire relazioni nascoste tra i dati e di identificare sottoinsiemi distinti all'interno del dataset. Attraverso l'uso di algoritmi di clustering appropriati, possiamo assegnare ogni osservazione a un cluster specifico in base alle sue caratteristiche simili, consentendoci di esplorare la struttura interna del dataset e di rivelare eventuali pattern o tendenze emergenti.

Metodo di Ward

Il clustering gerarchico è stato eseguito utilizzando il criterio di Ward sulle prime due componenti principali.

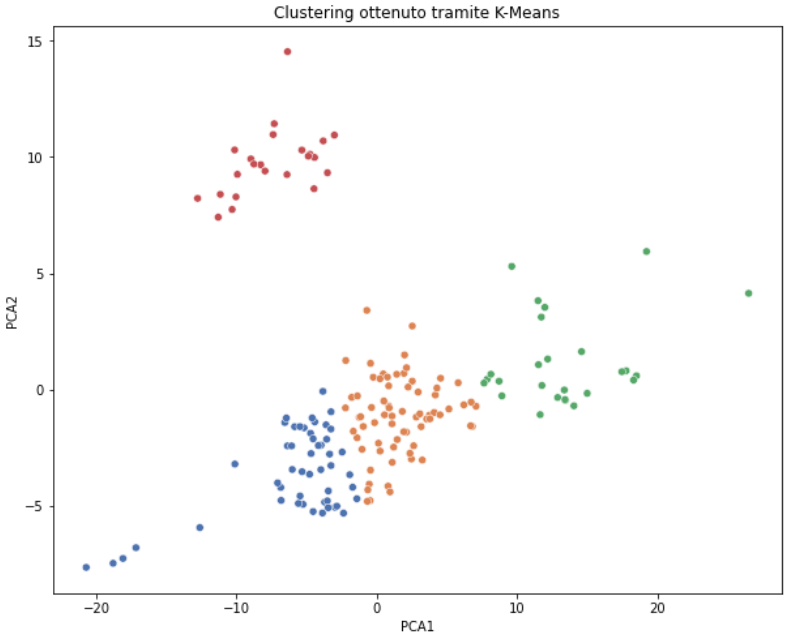


I campioni delle resine sono stati colorati in base al cluster di appartenenza individuato

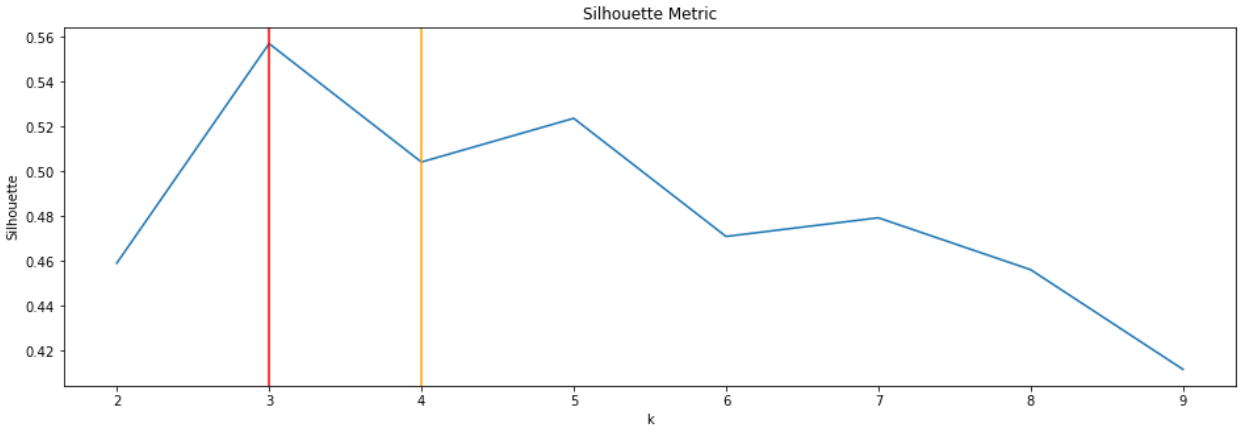
K-Means

Inoltre, si è utilizzato l'algoritmo k-means per effettuare un clustering non gerarchico

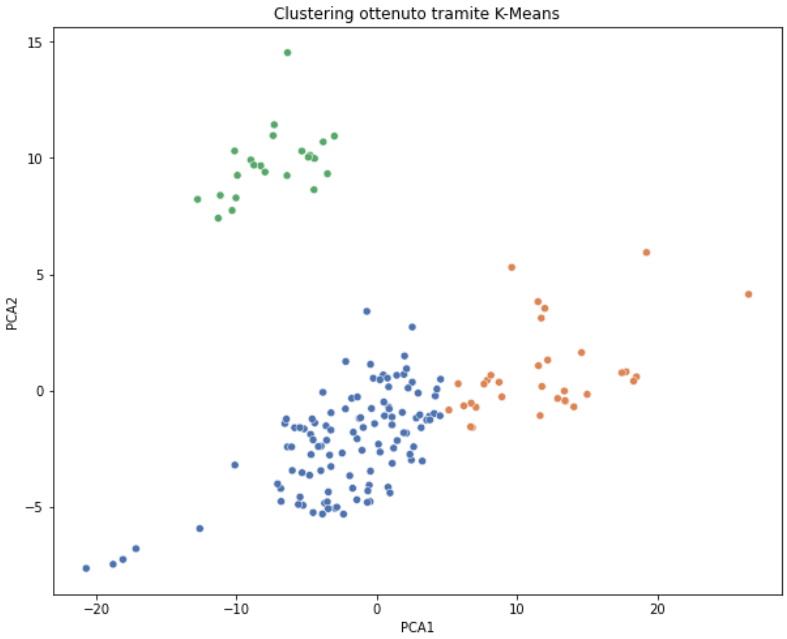
All'inizio si sono cercati sempre 4 cluster, ottenendo il seguente risultato:



Tramite gli indici di silhoutte si è poi valutato il miglior numero di cluster



Si nota che con 3 cluster si ottiene un valore migliore. Di seguito si riporta il clustering ottenuto

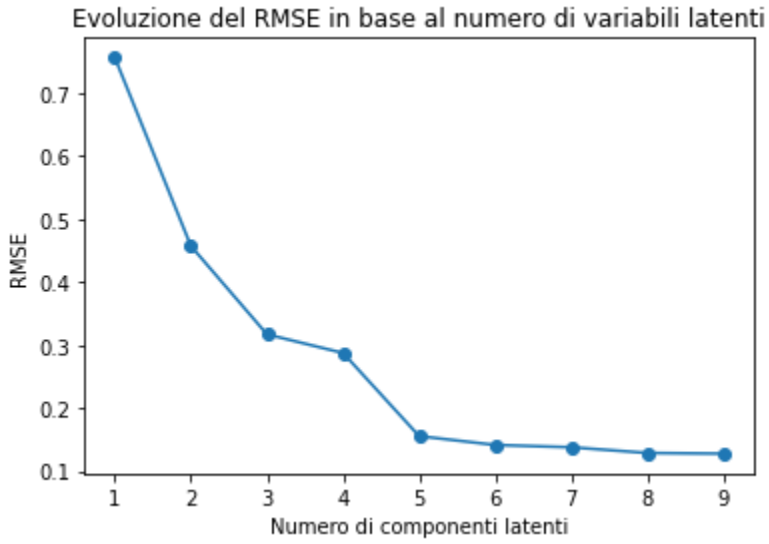

Clustering ottenuto tramite K-Means
PCA2
PCA1
15
10
5
0
−5
−20
−10
0
10
20

PLS

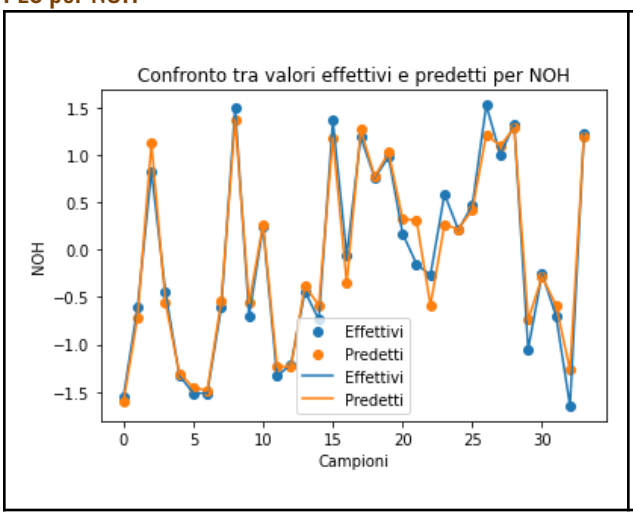
Si è sviluppato un modello PLS (Partial Least Squares) per la previsione dei valori di NOH, RS e NAC utilizzando i dati spettrali come predittori. Per calibrare il modello, si è utilizzata la cross-validation e si è scelto di includere 7 variabili latenti. Questa tecnica ha permesso di valutare le prestazioni predittive del modello in modo accurato e affidabile.

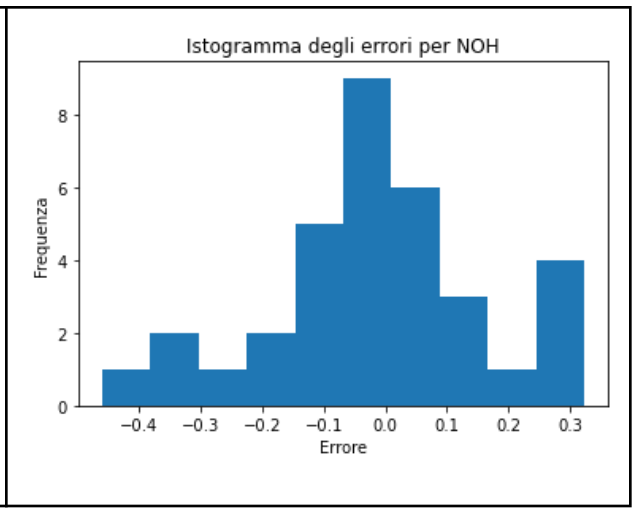
Complessivamente, attraverso l'uso del modello PLS e l'applicazione della cross-validation, si è stati in grado di ottenere previsioni accurate e affidabili per le variabili NOH, RS e NAC basate sui dati spettrali a mia disposizione.

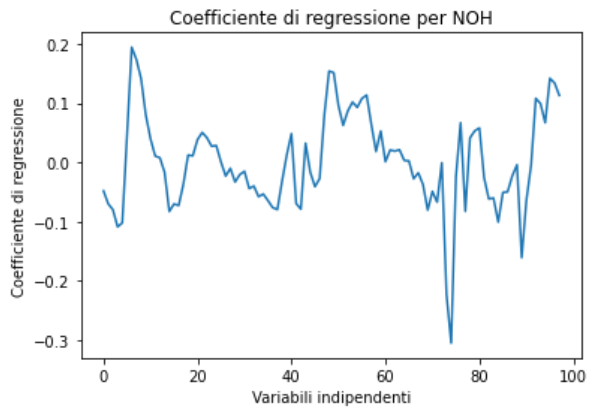
Evoluzione RMSE totale in base al numero di variabili latenti



PLS per NOH

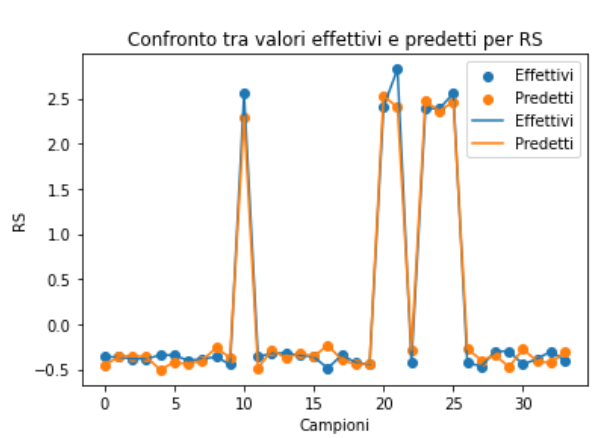


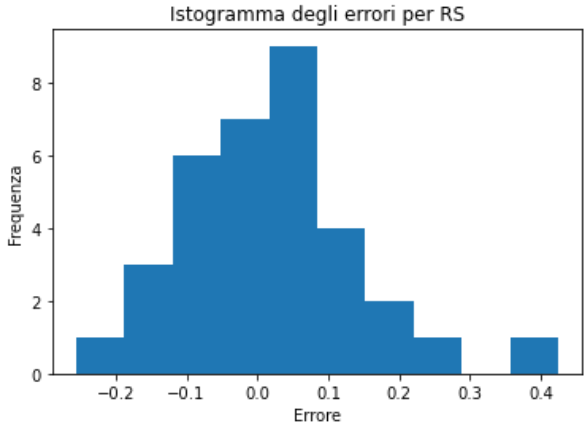


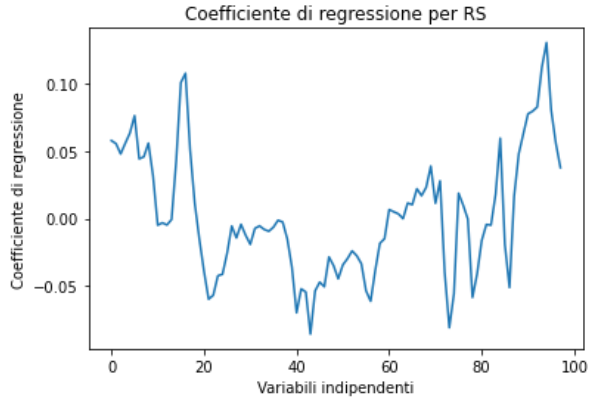


RMSE: 0.0182103368632796

PLS per RS

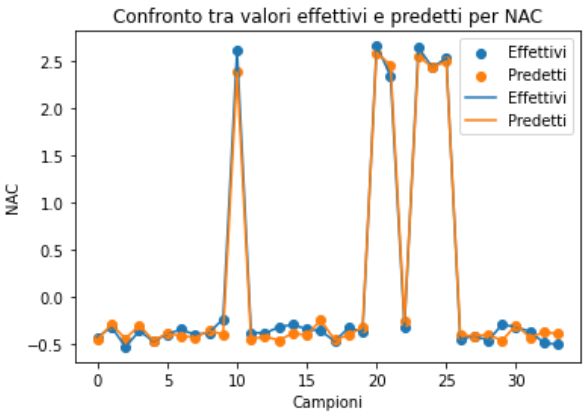


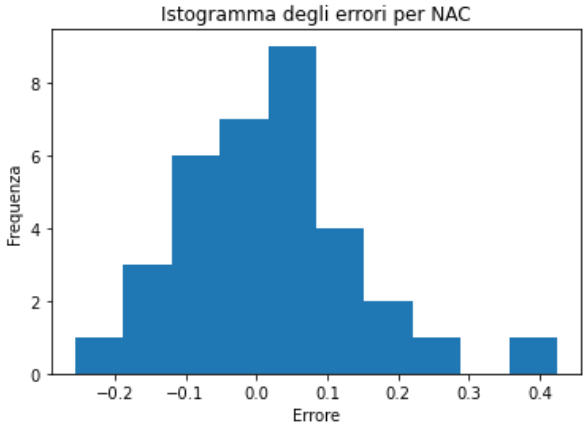


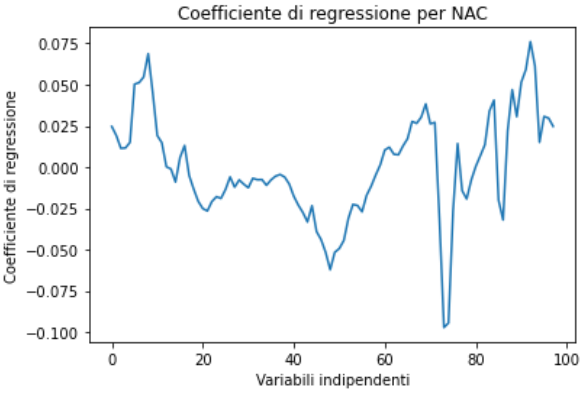


RMSE: 0.012317757600892092

PLS per NAC







RMSE: 0.015060369564383436

Conclusioni

Lo scopo dell'analisi condotta era di esaminare se fossero presenti informazioni rilevanti per le variabili del numero di ossidrile (NOH), residuo secco (RS) e numero di acido (NAC) all'interno dei dati spettrali.

L'obiettivo era quello di stabilire un legame tra le variabili dei dati spettrali e le variabili quantitative y (NOH, RS e NAC), al fine di calibrare e predire le concentrazioni di incognite utilizzando il metodo multivariato di regressione delle Partial Least Squares (PLS).

I metodi di riconoscimento dei pattern utilizzati hanno permesso di esplorare i dati, estrarre informazioni generali, identificare correlazioni e raggruppamenti tra le quattro famiglie di resine preesistenti.

La PCA (Principal Component Analysis) è stata utilizzata solo sui dati spettrali NIR per valutare se contenessero informazioni su NOH, RS e NAC. Attraverso l'analisi delle componenti principali, è stato possibile ridurre la dimensionalità dei dati. I risultati hanno mostrato una forte correlazione tra gli spettri e, come previsto, è stato possibile condensare tutte le informazioni rilevanti in soli due/tre componenti principali. Sulla base di queste componenti principali, è stato possibile descrivere gli oggetti analizzati. La PCA, tuttavia, non tiene conto di tutte le informazioni disponibili.

Successivamente, sono stati eseguiti raggruppamenti (clustering) basati sulle componenti significative per rimuovere il rumore sperimentale presente nei dati. Nell'intervallo di valori coperto da NOH, è stata osservata una distribuzione omogenea, consentendo di separare le quattro famiglie di resine in diversi range di valori. Tuttavia, per quanto riguarda RS e NAC, gli intervalli di valori non erano omogenei e solo una famiglia di resine è stata distinguibile dalle altre.

Al fine di identificare eventuali correlazioni, è stato utilizzato un metodo di regressione per costruire un modello quantitativo. La scelta del metodo PLS è stata determinata dalla sua capacità di trovare la migliore correlazione con la variabile y. Con l'uso di PLS, è stato possibile modellare le tre variabili y (NOH, NAC, RS) in base ai dati spettrali. Questo approccio ha permesso di creare modelli quantitativi che consentono di predire le concentrazioni di queste variabili utilizzando gli spettri analitici disponibili.

Dai valori RMSE calcolati si è notato che gli errori commessi dai modelli di predizione sulle tre variabili sono estremamente bassi:

- NOH: RMSE = 0.0182103368632796
- RS: RMSE = 0.012317757600892092
- NAC: RMSE = 0.015060369564383436

Ciò significa che i predittori riescono a predire efficacemente i valori di NOH, RS e NAC