Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in

Statistica per le tecnologie e le scienze

UN CONFRONTO TRA PCR E PPR APPLICATO AD UN DATASET REALE

Relatore: Prof. Erlis Ruli
Dipartimento di Scienze Statistiche

Laureando: Sebastiano Sumiti
Matricola n. 2005384

Anno Accademico 2022/2023

A mia sorella Alessandra.

Indice

Introduzione

La tesi in questione tratterà due diversi metodi nell'ambito delle analisi multivariate e dell'apprendimento automatico che permettono di ridurre la dimensionalità delle variabili, in particolare l'Analisi delle Componenti Principali (da qui in avanti si userà il termine *PCA*) e la Projection Pursuit (denominata con *PP*).
La riduzione della dimensionalità delle variabili applicata ad un dataset è uno degli strumenti più utilizzati in quanto permette di semplificare il problema da analizzare, mantenendo tuttavia la maggior parte delle informazioni iniziali. Inoltre, partendo dalle conclusioni ottenute dai metodi sopracitati, è possibile giungere a modelli più parsimoniosi che possono essere utilizzati in seguito per l'analisi di regressione o di classificazione. La PCA è una tecnica ampiamente utilizzata per esplorare e scomporre dati complessi in componenti principali, al fine di giungere ad interpretazioni più semplici. La PP, invece, si concentra sulla proiezione dei dati in uno spazio a dimensionalità ridotta, in modo da capire le informazioni e i pattern nascosti tra di essi.
La regressione e la definizione del modello statistico con questi due metodi permettono di comprendere la relazione lineare o non lineare tra le variabili indipendenti e dipendenti, risultato che una semplice regressione lineare non può catturare.
L'obiettivo della tesi è, quindi, di applicare i due metodi ad un dataset e capire quale tra essi giunge a delle previsioni migliori riguardo la variabile risposta, che verranno valutate attraverso l'errore quadratico medio. Il primo capitolo presenta la descrizione e le analisi esplorative del dataset in questione; in esso si discuterà sia della variabile risposta relativa alla qualità dei vini, sia delle variabili indipendenti che riguardano le caratteristiche fisiche e chimiche delle unità statistiche. Il secondo e il terzo capitolo trattano della parte teorica: verranno presentati nello specifico e a livello matematico i concetti di componenti principali, analisi delle componenti principali, proiezione di dati in uno spazio d-dimensionale, Projection Pursuit ed infine, modelli di regressione con le componenti principali e regressione con la Projecton Pursuit. Verranno inoltre accennati indici e metodi per valutare una previsione così da giudicare quale metodo può prevedere meglio la variabile dipendente. Il capitolo conclusivo si concentra sull'applicazione pratica di queste tecniche e sulle conclusioni emerse dallo studio.

Capitolo 1

Analisi del Vinho Verde

1.1 Informazioni generali

Il dataset in questione è stato creato dall'unione di due diversi dataset contenenti le informazioni delle varianti di vino rosso e di vino bianco del "Vinho Verde" (Vinho Verde, 2023) . Il Vinho Verde è un vino portoghese proveniente dalla provincia nel nord del paese Minho. Questo vino non indica una precisa varietà d'uva ma piuttosto è preferibile tradurlo, come vino verde, ovvero vino giovane, in opposizione a quello più maturo. La particolarità risiede nella sua necessità di essere consumato tempestivamente ed infatti è consigliato consumarlo entro un anno dall'imbottigliamento. Può essere di tre diverse tipologie: rosso, bianco e rosé; nel caso in questione si tratterranno solamente le prime due. Il dataset relativo al vino bianco contiene le informazioni di 4898 bottiglie di Vinho Verde, mentre il dataset relativo al vino rosso contiene le caratteristiche di un campione di 1599. Al fine di applicare due diversi algoritmi di riduzione della dimensionalità, si lavorerà con un unico dataset nato dall'unione dei due precedenti, quindi contenente 6497 unità statistiche.

1.2 Analisi Esplorativa

Il dataset contiene 11 variabili indipendenti, relative agli aspetti fisico-chimici e sensoriali del vino, e una variabile risposta. Nello specifico si afferma che tutte, ad eccezione della risposta, sono numeriche continue, ovvero possono assumere infiniti valori all'interno di un intervallo finito/infinito. La risposta, *quality*, è una variabile qualitativa ordinale con valori interi appartenente all'insieme [3,9].

Nome della variabile	Tipologia	Valori
fixed acidity	quantitativa continua	7 6.3 8.1 7.2 7.2 ...
volatile acidity	quantitativa continua	0.27 0.3 0.28 0.23 ...
citric acid	quantitativa continua	0.36 0.34 0.4 0.32 0.32 ...
residual sugar	quantitativa continua	20.7 1.6 6.9 8.5 ...
chlorides	quantitativa continua	0.045 0.049 0.05 0.058 ...
free sulfur dioxide	quantitativa continua	45 14 30 47 47 30 30 ...
total sulfur dioxide	quantitativa continua	170 132 97 186 ...
density	quantitativa continua	1.001 0.994 0.995 0.996 ...
ph	quantitativa continua	3 3.3 3.26 3.19 3.19 3.26 3.18 ...
sulphates	quantitativa continua	0.45 0.49 0.44 0.4 ...
alcohol	quantitativa continua	8.8 9.5 10.1 9.9 9.9 ...
quality	qualitativa ordinale	6 6 6 6 6 6 6 ...

Tabella 1.1: Tabella riassuntiva delle variabili con la relativa tipologia

Fixed Acidity

La prima variabile relativa agli aspetti fisico-chimici del vino è *Fixed Acidiy*, ovvero acidità fissa. Gli acidi fissi, ovvero acidi che non si volatilizzano facilmente, contribuiscono alla struttura e alla stabilità del vino, influenzando il suo gusto e la sua capacità di invecchiamento. Nel contesto del Vinho Verde, l'acidità fissa, essenzialmente derivante dagli acidi tartarico, malico e citrico, dona al vino una piacevole sensazione di freschezza e contribuisce ad una struttura ben bilanciata. Tra questi acidi, il tartarico emerge come elemento preponderante, esercitando un'influenza positiva più marcata rispetto agli altri e plasmando in modo significativo il carattere finale del prodotto.
Relativamente al dataset, l'acidità fissa è una variabile quantitativa continua che assume valori da un minimo di 3.8 ad un massimo di 15.9. La media e la mediana presentano valori molto vicini tra di loro, rispettivamente 7.215 e 7.000.

Min.	1st. Quartile	Mediana	Media	3rd. Quartile	Max.	std. deviation
3.800	6.400	7.000	7.215	7.700	15.900	1.296

Tabella 1.2: Tabella riassuntiva della variabile Fixed Acidity con i relativi indici di sintesi

La vicinanza tra i valori di media e mediana suggerisce la possibilità di una distribuzione normale della variabile in esame ma, tuttavia, tale supposizione viene contraddetta dal test di Shapiro-Wilk che respinge l'ipotesi nulla di normalità con un p-value prossimo allo zero.
Dal grafico sottostante è possibile vedere come i vini rossi, seppur siano in numero minore, assumono valori più alti rispetto ai vini bianchi per quanto riguarda l'acidità fissa. Infatti, il primo gruppo presenta primo e terzo quartile pari a 7.10 e 9.20, con uno scarto interquartilico, IQR, di 2.10; il secondo gruppo, invece, presenta rispettivamente i valori di 6.3,7.3 e l'IQR pari a 1.

Figura 1.1: Diagramma di dispersione delle unità statistiche (asse x) rispetto alla variabile fixed acidity (asse y), divise rispetto alla tipologia del vino.

Volatile acidity

La seconda variabile è la *volatile acidity*, in italiano acidità volatile. Questa caratteristica fisico-chimico è responsabile del classico odore di aceto presente nel vino. In quantità moderate, l'acidità volatile può contribuire alla complessità aromatica del vino, ma a livelli eccessivi si può giungere ad un sapore e ad un odore sgradevoli. Relativamente al Vinho Verde, l'acidità volatile permette di garantire un profumo molto fresco e leggero, garantendo in questo modo la sua classica vivacità e vitalità. Nel dataset in questione, si può affermare che la *volatile acidity* è una variabile quantitativa continua, che assume valori nel range 0.080 e 1.580, con una standard deviation pari a 0.165. Seppur media e mediana assumano valori vicini, rispettivamente 0.340 e 0.290, il test di Shapiro-Wilk rifiuta l'ipotesi nulla che la variabile in questione segua una distribuzione normale.

Min.	1st. Quartile	Mediana	Media	3rd. Quartile	Max.	std. deviation
0.080	0.230	0.290	0.340	0.400	1.580	0.165

Tabella 1.3: Tabella riassuntiva della variabile Volatile Acidity con i relativi indici si sintesi

Dalla tabella (1.3), emerge che il 75% delle osservazioni manifesta un valore inferiore a 0.4 ed inoltre, analizzando la figura (1.2), si osserva che la maggioranza di questi corrisponde a punti verdi, indicando chiaramente la prevalenza di vini bianchi in questa fascia di valori.

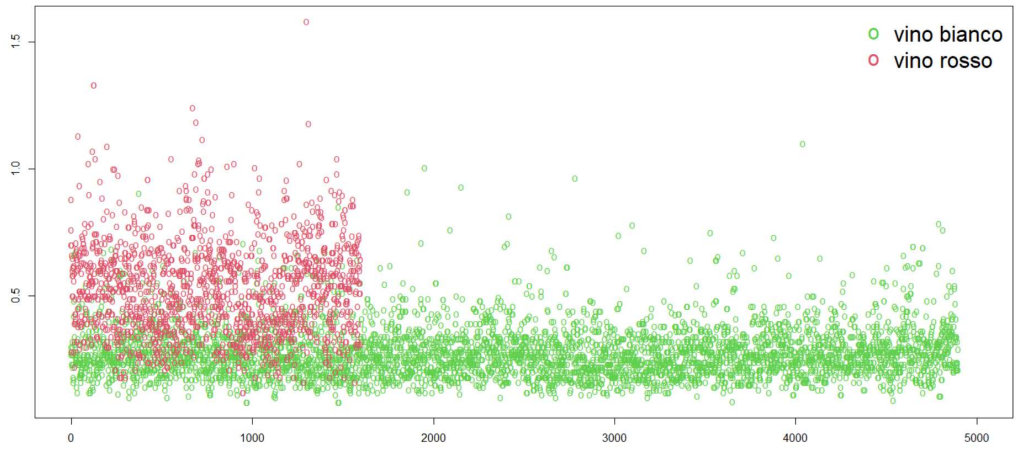

Figura 1.2: Diagramma di dispersione delle unità statistiche (asse x) rispetto alla variabile volatile acidity (asse y), divise in base alla tipologia del vino.

Citric Acid

La variabile *Citric Acid*, in italiano acido citrico, si riferisce alla presenza o meno nel vino dell'omonimo acido, uno degli acidi presenti naturalmente nell'uva. Esso caratterizza il sapore del vino, contribuendo a renderlo più fresco e vivace. Riguardo al vino in questione, l'acido citrico alleggerisce e rinfresca molto il gusto della bottiglia, permettendo una maggiore vitalità e bilanciando l'acidità con altri elementi del vino come il fruttato.
Relativamente al dataset, il valore medio e mediano coincidono assumendo il valore 0.31. Il valore massimo che assume la variabile è 1.66 mentre il valore minimo è quello nullo. Il restante delle informazioni sono riassunte nella tabella (1.4).

Min.	1st. Quartile	Mediana	Media	3rd. Quartile	Max.	std. deviation
0.000	0.250	0.310	0.310	0.390	1.660	0.145

Tabella 1.4: Tabella riassuntiva della variabile Citric Acid con i relativi indici di sintesi

A differenza delle variabili viste in precedenza, la variabile *citric acid* assume valori simili sia per i vini rossi che per i vini bianchi, affermando che la presenza di tale acido nel vino non è influenzata dalla tipologia ma dalle caratteristiche dell'uva e del luogo.
Il test di Shapiro-Wilk rileva una significativa deviazione dall'ipotesi nulla di una distribuzione gaussiana, con un livello di significatività praticamente trascurabile.

Figura 1.3: Diagramma di dispersione delle unità statistiche (asse x) rispetto alla variabile citric acid (asse y) divise rispetto alla tipologia del vino.

Residual Sugar

Il concetto di *residual sugar*, o zucchero residuo, in un vino si riferisce alla frazione di zucchero non convertita durante il processo di fermentazione e persistente nel prodotto finito. Durante la fermentazione, il lievito catalizza la trasformazione degli zuccheri presenti nell'uva in alcol e anidride carbonica. L'insieme degli zuccheri non metabolizzati costituisce l'elemento noto come zucchero residuo, il quale conferisce e definisce la dolcezza intrinseca del vino.
Nel contesto del Vinho Verde, generalmente i vini sono asciutti o leggermente frizzanti con bassi livelli di zucchero. Tuttavia, è importante precisare che la quantità di zucchero residuo, e quindi della dolcezza del vino, varia da produttore a produttore in base alle pratiche enologiche ed è per questo motivo che i valori della variabile variano in un range molto ampio, passando da un minimo di 0.600 ad un massimo di 65.800. Il p-value prossimo a zero del test di Shapiro-Wilk e la distanza significativa tra i valori di media e mediana, rispettivamente di 5.443 e di 3.000, portano a rifiutare l'ipotesi di normalità della variabile.

Min.	I Quartile	Mediana	Media	III Quartile	Max.	std. deviation
0.600	1.800	3.000	5.443	8.100	65.800	4.758

Tabella 1.5: Tabella riassuntiva della variabile Residual Sugar con i relativi indici di sintesi

Dalla figura (1.4) si può notare come i vini bianchi del Vinho Verde sono più dolci rispetto ai vini rossi, con una differenza delle medie di 3.852 in più per i primi rispetto che ai secondi.

Figura 1.4: Diagramma di dispersione delle unità statistiche (asse x) rispetto alla variabile Residual Sugar (asse y) divise rispetto alla tipologia del vino.

Chlorides

La variabile *Chlorides*, in italiano cloruri, si riferisce agli ioni di cloro dovuti alle caratteristiche del suolo, alle pratiche agricole e alle tecniche di produzione e di conservazione del vino. Essi fanno parte della categoria dei sali minerali disciolti e contribuiscono alla struttura del vino e alla relativa percezione del sapore. Tuttavia, un elevato valore della variabile può danneggiare molto il vino, eccedendo ad un sapore salato e perdendo il bilanciamento ed è per questo motivo che *Chlorides* assume valori in un range ridotto, passando da un minimo di 0.009 ad un massimo di 0.611 con una deviazione standard di 0.035.

Min.	1st. Quartile	Mediana	Media	3rd. Quartile	Max.	std. deviation
0.009	0.038	0.047	0.056	0.065	0.611	0.035

Tabella 1.6: Tabella riassuntiva della variabile Chlorides con i relativi indici di sintesi

Figura 1.5: Diagramma di dispersione delle unità statistiche (asse x) rispetto alla variabile Chlorides (asse y) divise rispetto alla tipologia del vino.

Dal grafico (1.5) si può notare come i pallini rossi, che rappresentano i vini rossi, giacciono sopra i pallini verdi, che rappresentano invece i vini bianchi. Le medie dei due gruppi sono significativamente diverse tra di loro, in quanto i vini rossi presentano una media che è il doppio rispetto a quella dei vini bianchi (0.087 per il primo gruppo e 0.046 per il secondo).

Free Sulfur Dioxide

Il *free Sulfur Dioxide*, in italiano anidride solforosa libera, è un composto chimico contenente zolfo e ossigeno, noto anche con la formula SO_2. Viene utilizzato all'interno dei vini per diverse motivazioni, tra le quali quella di avere la capacità di agire come antiossidante e antimicrobico, ovvero di prevenire l'ossidazione e di proteggere il vino dalla proliferazione di batteri. Nel Vinho Verde, la scelta dell'uso di SO_2 e della relativa quantità variano da produttore a produttore durante le fasi di vendemmia e di imbottigliamento. In particolare, i vini bianchi di questa particolare zona presentano in media una maggiore quantità di SO_2 rispetto ai comuni vini bianchi. Questa scelta è motivata dalla richiesta di una maggior attenzione alla conservazione e alla maturazione del vino, dovuto alle proprie caratteristiche come la frizzantezza e le condizioni climatiche del nord del Portogallo. Infatti, in queste zone, la vicinanza all'oceano Atlantico e l'elevata umidità portano ad una frequente creazione di muffa che viene combattuta proprio da questo composto chimico. Il grafico (1.6) mostra esattamente come i vini bianchi stiano sopra ai vini rossi, indicando come la quantità di SO_2 sia molto più presente nel primo gruppo rispetto che al secondo. Le medie per i due gruppi sono rispettivamente 35.308 per i vini bianchi e 15.875 per i vini rossi.

Min.	1st. Quartile	Mediana	Media	3rd. Quartile	Max.	std. deviation
1.000	17.000	29.000	30.530	41.000	289.000	17.749

Tabella 1.7: Tabella riassuntiva della variabile Free Sulfur Dioxide con i relativi indici di sintesi

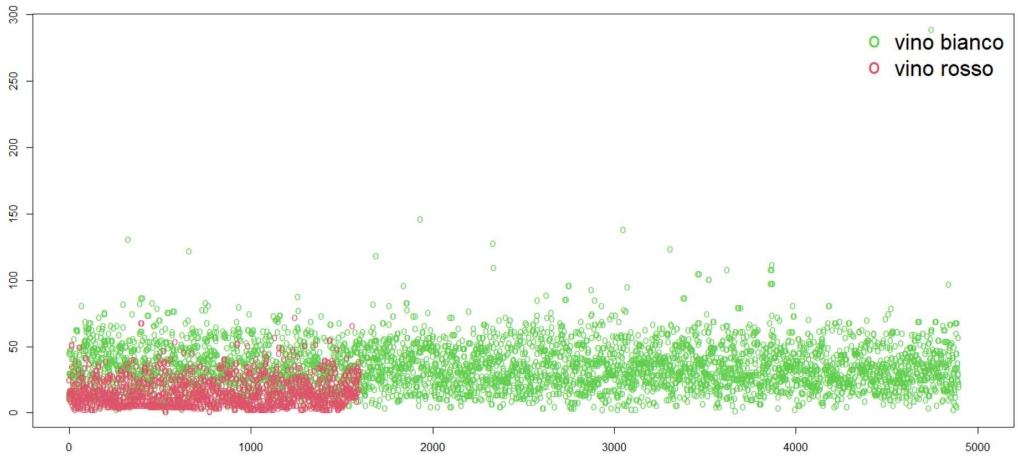

Figura 1.6: Diagramma di dispersione delle unità statistiche rispetto alla variabile Free Sulfur Dioxide divise rispetto alla tipologia del vino.

Total Sulfur Dioxide

La successiva variabile relativa agli aspetti chimico-fisici è *Total Sulfur Dioxide*, ovvero anidride solforosa totale. Essa è la somma degli equivalenti di anidride solforosa libera, variabile precedente analizzata, e di quella legata ad altre molecole. Oltre ai motivi descritti nel paragrafo precedente, l'anidride solforosa totale previene l'ossidazione, preserva la freschezza e mantiene la chiarezza e la purezza del vino, evitando in questo modo la formazione di sedimenti o di altre particelle indesiderate. Analogamente all'esempio precedente, si osserva una vasta distribuzione di valori, e il 75% delle osservazioni si colloca al di sotto della soglia di 156.000.

Min.	1st. Quartile	Mediana	Media	3rd. Quartile	Max.	std. deviation
6.000	77.000	118.000	115.700	156.000	444.000	56.522

Tabella 1.8: Tabella riassuntiva della variabile total Sulfur Dioxide con i relativi indici di sintesi

Dal grafico 1.7 è ancora più evidente come i vini bianchi del Vinho Verde presentino un valore più alto del composto chimico rispetto ai vini rossi del medesimo posto ed infatti la media dei primi, pari a 138.361, è circa 3 volte più grande rispetto alla media dei secondi pari a 46.467.

Figura 1.7: Diagramma di dispersione delle unità statistiche rispetto alla variabile Total Sulfur Dioxide divise rispetto alla tipologia del vino.

Density

La densità è una proprietà fisica che indica quanta materia è contenuta in una determinata quantità di volume di vino. La formula matematica, $\frac{massa}{volume}$, è spesso misurata in chilogrammi per litro ($\frac{kg}{L}$) ed è influenzata da vari fattori, tra cui il contenuto alcolico, gli zuccheri residui e la temperatura. Tale variabile fornisce diverse informazioni agli enologi, in particolare permette di valutare la maturità delle uve, la fermentazione e la qualità del prodotto finale. Tuttavia, bisogna precisare che l'informazione fornita dalla sola densità non garantisce un quadro completo del vino. Altre variabili, che verranno valutate in seguito come il pH e l'alcol, sono altrettanto cruciali per una completa descrizione.

La deviazione standard pari a 0.003 e lo scarto interquartilico pari a 0.005 affermano che i valori assunti dalla variabile sono molto centrati intorno a 0.995, ovvero la mediana. Non è presente una significativa differenza tra i vini bianchi e rossi, indicando che i vini del Vinho Verde presentano valori simili a prescindere dalla loro natura.

Min.	1st. Quartile	Mediana	Media	3rd. Quartile	Max.	std. deviation
0.987	0.992	0.995	0.995	0.997	1.040	0.003

Tabella 1.9: Tabella riassuntiva della variabile Density con i relativi indici di sintesi

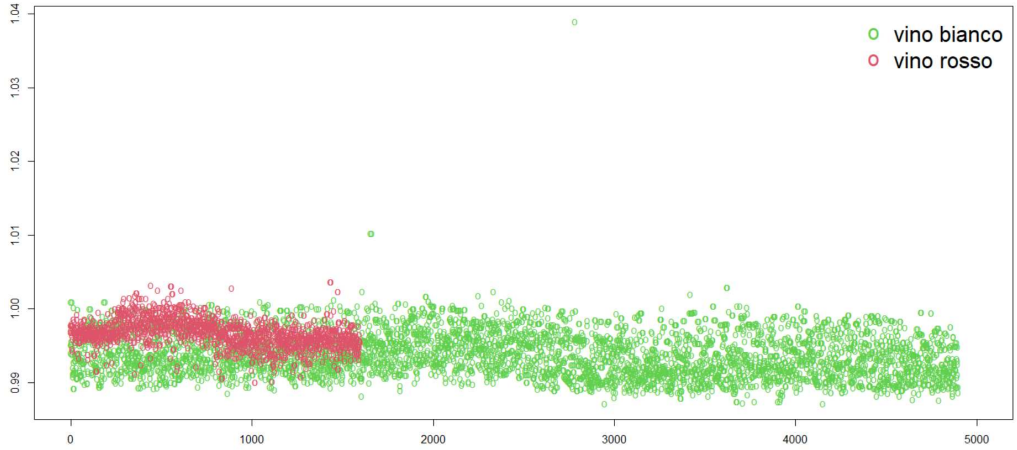

Figura 1.8: Diagramma di dispersione delle unità statistiche rispetto alla variabile Density divise rispetto alla tipologia del vino.

pH

Il *pH* è una misura dell'acidità o della basicità di una soluzione e varia da 0 fino a 14. Il valore 7 indica un *pH* neutro, un valore superiore indica che la soluzione è basica, mentre un valore inferiore indica che la soluzione è acida. Il *pH* è un fattore importante nel vino, influendo nella stabilità chimica del prodotto e nel suo percorso di invecchiamento. L'acidità, e quindi una maggior vitalità, e la basicità, ovvero una maggior morbidezza nel gusto, vengono gestiti dai diversi produttori attraverso la fase di pigiatura e quella di fermentazione.
La gamma di *pH* ideale si trova in genere tra 3 e 4 ed in particolare, il Vinho Verde presenta un valore tra 3 e 3.4. Relativamente al dataset in questione, si nota che l'80% delle unità sono comprese in questo range ed inoltre il 50% di esse è compreso in un range ancora più ridotto, che va dal 3.110 a 3.320. La deviazione standard è pari a 0.161 e quindi, come nel caso precedente, i valori sono centrati attorno alla mediana di 3.210. I vini rossi dalla figura 1.9 sembrano assumere un valore più alto per la variabile *pH* rispetto ai vini bianchi, ma tale differenza per i motivi spiegati in precedenza non sembra essere significativa.

Min.	1st. Quartile	Mediana	Media	3rd. Quartile	Max.	std. deviation
2.720	3.110	3.210	3.219	3.320	4.010	0.161

Tabella 1.10: Tabella riassuntiva della variabile pH con i relativi indici di sintesi

Figura 1.9: Diagramma di dispersione delle unità statistiche rispetto alla variabile pH divise rispetto alla tipologia del vino.

Sulphates

La prossima variabile analizzata è *Sulphates*, tradotto in italiano come solfati oppure solfiti. Questi sono dei composti chimici contenenti zolfo che vengono comunemente utilizzati come additivi nella produzione del vino, grazie alle loro proprietà antimicrobiche, ovvero di eliminare microorganismi indesiderati, e antiossidanti, ovvero di prevenire l'ossidazione nel vino. I solfiti possono essere aggiunti dai diversi produttori nelle quantità desiderate nelle fasi di pigiatura dell'uva, di fermentazione e di imbottigliamento. La scelta della quantità da utilizzare modifica diverse caratteristiche ma deve essere specificata sulla bottiglia, in modo che il consumatore sia consapevole della presenza di solfiti, specialmente per coloro che ne possono essere sensibili o allergici.

Min.	1st. Quartile	Mediana	Media	3rd. Quartile	Max.	std. deviation
0.220	0.430	0.510	0.531	0.600	2.000	0.149

Tabella 1.11: Tabella riassuntiva della variabile Sulphates con i relativi indici di sintesi

Il range di valori che tale variabile assume va da un minimo di 0.220 $\frac{mg}{L}$ ad un massimo di 2.000 $\frac{mg}{L}$, con la relativa standard deviation di 0.149. Le ulteriori informazioni sulla distribuzione della variabile sono riportate nella tabella 1.11.
Dal grafico, appare che la quantità di solfiti presente nei vini rossi sia maggiore rispetto a quella nei vini bianchi ed, infatti, il test di Wilcoxon afferma che la differenza in media dei due gruppi risulta essere significativa.

Figura 1.10: Diagramma di dispersione delle unità statistiche rispetto alla variabile Sulphates divise rispetto alla tipologia del vino.

Alcohol

La variabile *Alcohol*, o alcol, in un vino si riferisce alla percentuale di alcol etilico per volume presente nel prodotto. É un parametro fondamentale che influenza sia le caratteristiche organiche del vino che le sensazioni ed i gusti percepiti in bocca. La gestione della fermentazione e il controllo della maturazione dell'uva sono aspetti importanti per determinare quanto zucchero viene convertito in alcol. In genere, i vini del Vinho Verde presentano una percentuale moderata di alcol, garantendo al prodotto freschezza e leggerezza. Questo bilanciamento e questa elevata bevibilità vengono giustificati anche dalla tabella 1.12 dove appare, infatti, che il 75% delle unità statistiche si trovano al di sotto dell'11%.

Min.	1st. Quartile	Mediana	Media	3rd. Quartile	Max.	std. deviation
8.000	9.500	10.300	10.490	11.300	14.900	1.193

Tabella 1.12: Tabella riassuntiva della variabile Alcohol con i relativi indici di sintesi

Figura 1.11: Diagramma di dispersione delle unità statistiche rispetto alla variabile Alcohol divise rispetto alla tipologia del vino.

Quality

La variabile risposta è *Quality*, o qualità, variabile qualitativa ordinale, che ha la funzione di assegnare un giudizio ad un vino conferendo un voto da 3 a 9. In particolare, 3 e 4 indicano una qualità inferiore, 5 e 6 indicano invece una qualità media, 7 e 8 una buona qualità, e per ultimo 9 viene assegnato ai vini che vengono giudicati eccellenti. Dai boxplot, figura 1.12, si afferma che sia i vini bianchi che i vini rossi presentano principalmente valori tra 4 e 6 ma il primo gruppo è giudicato leggermente meglio rispetto al secondo come si può notare dal valore della media, indicato mediante il punto nero.

Min.	1st. Quartile	Mediana	Media	3rd. Quartile	Max.	std. deviation
3.000	5.000	6.000	5.818	6.000	9.000	0.873

Tabella 1.13: Tabella riassuntiva della variabile Quality con i relativi indici di sintesi

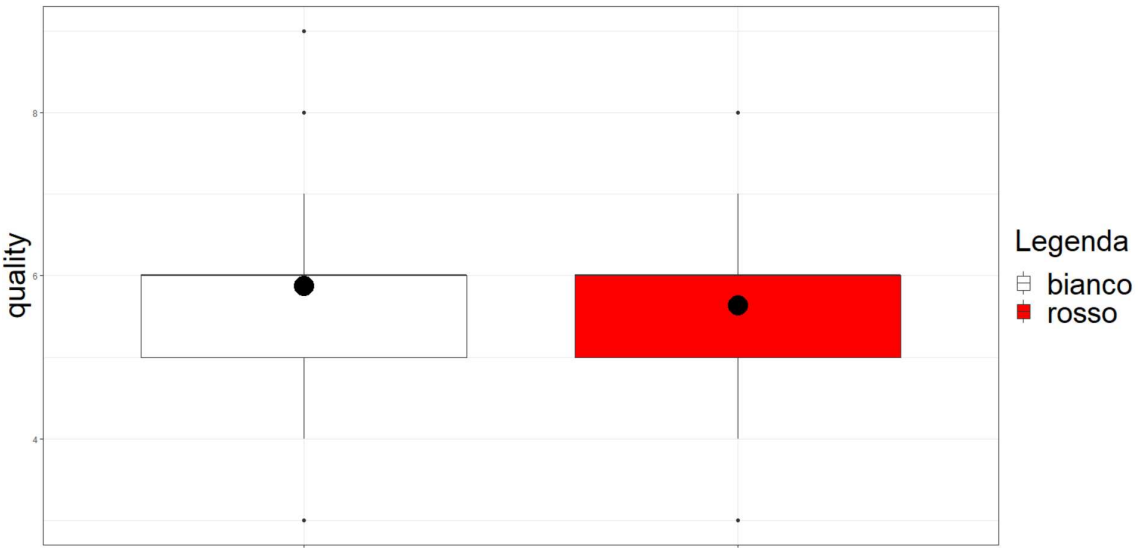

Figura 1.12: Boxplot della variabile Quality rispettivamente per i vini bianchi e per i vini rossi.

Capitolo 2

Analisi delle componenti principali e Regressione con le componenti principali

2.1 Analisi delle componenti principali

2.1.1 Introduzione

La Principal Component Analysis (PCA), in italiano Analisi delle Componenti Principali, è una tecnica estremamente nota nell'ambito della statistica e dell'analisi dei dati. Venne introdotta da Karl Pearson nel 1901 e sviluppata da Harold Hotelling nel 1933. L'obiettivo di questo metodo è la riduzione della dimensionalità dei dati, ovvero semplificare le unità statistiche e trasformare i dati da uno spazio p-dimensionale ad uno spazio d-dimensionale con d<p. Questo ha lo scopo di facilitare l'utilizzo del dataset, mantenendo al contempo la massima quantità di informazione possibile. La PCA entra all'interno degli approcci non supervisionati, cercando di ricreare, per quanto possibile, la struttura originaria dei dati attraverso una trasformazione lineare delle variabili iniziali.
Sia $\mathcal{X}$ un vettore casuale con $\mu = \mathbb{E}(X)$, vettore delle medie, e Σ matrice di covarianza. Siano $\lambda_1, \gamma_1, \lambda_2, \gamma_2, \ldots, \lambda_p, \gamma_p$ le coppie di autovalori e autovettori di Σ, con $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq 0$. Dati $a_1, a_2, \ldots, a_p$ vettori reali con $a_i = (a_{i1}, a_{i2}, ..., a_{ip})$ si consideri le combinazioni lineari:

$$\begin{aligned} Y_1 &= \mathbf{X}a_1 = a_{1,1}X_1 + a_{1,2}X_2 + \ldots + a_{1,p}X_p \\ Y_2 &= \mathbf{X}a_2 = a_{2,1}X_1 + a_{2,2}X_2 + \ldots + a_{2,p}X_p \\ &\vdots \\ Y_p &= \mathbf{X}a_p = a_{p,1}X_1 + a_{p,2}X_2 + \ldots + a_{p,p}X_p \end{aligned} \tag{2.1}$$

I nuovi vettori casuali presenteranno quindi:

$$\begin{aligned} &\mathbb{E}\left(Y_i\right) = a_i^{\mathrm{T}}\mu, \\ &\mathbb{V}\left(Y_i\right) = a_i^{\mathrm{T}}\Sigma a_i, \\ &\mathbb{C}\left(Y_i, Y_j\right) = a_i^{\mathrm{T}}\Sigma a_j \end{aligned} \tag{2.2}$$

2.1.2 Componenti Principali

Definiamo ora le componenti principali come le nuove variabili casuali $Y_1, ..., Y_p$ che presentano varianza massima e sono ortogonali tra loro. In particolare, queste variabili sono ordinate in modo che la prima componente, nata dalla combinazione lineare $Y_1 = a_1^t \mathcal{X}$, massimizzi la varianza $V(a_1^t \mathcal{X}) = a_1^t \Sigma a_1$ sotto il vincolo di determinabilità $a_1^t a_1 = 1$. La i-esima componente principale con $i = 2, \ldots, p$ è data dalla trasformazione $Y_i = a_i^t \mathcal{X}$ e massimizza la varianza $\mathrm{V}(a_i^t \mathcal{X})$ sotto i vincoli $a_i^t a_i = 1$ e quello di incorrelazione con tutte le componenti precedenti $C(Y_i, Y_j) = C(a_i^t X, a_j^t X) = 0$ *con* $j < i$. Questi due punti precedenti sono definiti dal teorema principale delle componenti principali e dal teorema di massimizzazione vincolata di forme quadratiche.
Le componenti principali godono di diverse proprietà che saranno essenziali nelle successive sezioni. La prima proprietà permette di affermare che la somma delle varianze delle cp è pari alla varianza totale. La proporzione della varianza totale spiegata dalla i-esima componente è, quindi, data da:

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + ... + \lambda_p} = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j} \tag{2.3}$$

La seconda e la terza proprietà affermano rispettivamente che: il segno della correlazione tra Y_i e Y_j con $i \neq j$ è data dall'autovettore corrispondente, e che il metodo della PCA non è invariante rispetto all'operazione di standardizzazione. Nella pratica, e quindi come nel caso del dataset in questione, μ e Σ sono ignoti e vanno stimati estrapolando le informazioni dai dati a disposizione. Questi parametri verranno quindi sostituiti dalle relative stime campionarie, rispettivamente:

$$\begin{aligned} S &= [s_{ij}] = \mathcal{X}(\mathcal{I}_n - (\tfrac{1}{n}) 1_1 1_n^t) \mathcal{X}^t = \mathcal{X} H \mathcal{X}^t \\ \bar{x} &= (\tfrac{1}{n}) \mathcal{X}^t 1_n \end{aligned} \tag{2.4}$$

2.1.3 Obiettivi della PCA

L'applicazione di questo metodo ad un dataset reale e non, viene effettuata per diverse ragioni:

- riduzione della dimensionalità: la PCA permette di ridurre il numero di variabili da considerare riducendo lo spazio da una dimensione p a una dimensione d con $d < p$, semplificando quindi l'analisi senza perdere le informazioni significative;

- identificazione di pattern nascosti: successivamente alla riduzione della dimensionalità è possibile visualizzare i dati in un nuovo spazio dato dalle componenti principali, permettendo di vedere pattern nascosti e relazioni tra i dati. Ad esempio, è possibile tramite la PCA visualizzare cluster di punti e tendenze che non erano visibili nella rappresentazione originale. Per questo motivo l'analisi delle componenti principali è utilizzata anche nella parte esplorativa;

- riduzione del rumore: essenziale all'interno del metodo della PCA è la selezione del numero appropriato di componenti da considerare. Generalmente, se

si mantenessero le prime N componenti principali si otterrebbe una rappresentazione circa uguale dei dati originali con il rumore parzialmente ridotto. Questo è spiegato dal fatto che le componenti principali consecutive tendono a catturare meno varianza e quindi meno rumore. La scelta corretta del numero di componenti porta ad una riduzione significativa del rumore ma senza mai raggiungere la rimozione completa. É importante una corretta ricerca di equilibrio tra la riduzione del rumore e il mantenimento delle informazioni importanti;

- algoritmi di regressione e di classificazione: le componenti principali offrono inoltre la possibilità di utilizzare modelli di regressione e di classificazione. L'utilizzo di queste componenti al posto delle variabili, oltre a semplificare il modello, permette anche di gestire la multicollinearità in maniera efficace, dovuto alla ortogonalità delle cp.

I passi che bisogna compiere per effettuare questa analisi sono diversi ma tutti si basano su principi matematici e statistici solidi. Il primo punto consiste nell'operazione di standardizzazione dei dati. La standardizzazione dei dati è il processo matematico che permette di trasformare le variabili aleatorie in modo da avere media zero e deviazione standard uno. Essendo la PCA una tecnica che trova il fondamento nella selezione delle componenti principali in base alla varianza dei dati, la standardizzazione assicura che tutte le variabili presentino la stessa scala e quindi vengano considerate in maniera equa durante il calcolo. Inoltre, permette di rendere i dati adimensionali, privandoli quindi dell'unità di misura, garantendo una maggior facilità di utilizzo per gli algoritmi di ottimizzazione utilizzati nella regressione e nella classificazione. La facilità dell'interpretazione è l'ultimo guadagno da questa prima operazione sui dati. I coefficienti delle componenti riflettono quanto ciascuna variabile contribuisce in termini di deviazione standard. Matematicamente, l'operazione di standardizzazione viene effettuata tramite la formula:

$$\mathcal{Z} = [z_{ij}] = \mathcal{H}\mathcal{X}\mathcal{D}^{-\frac{1}{2}} \tag{2.5}$$

dove $\mathcal{D} = diag(\mathcal{S}) = diag(s_{11}, s_{22}, ..., s_{pp}) \quad e \quad \mathcal{H} = \mathcal{I}_n - (\frac{1}{n})11^t$.

Successivamente, viene calcolata la matrice di covarianza, arrivando ad avere $S_z = R$. Il terzo passo prevede la decomposizione della matrice di covarianza attraverso la decomposizione spettrale oppure la decomposizione a valori singolari. Queste operazioni sono tecniche matematiche che permettono rispettivamente di diagonalizzare una matrice X attraverso i suoi autovettori e autovalori o di, nel caso della svd, scomporre una matrice X in tre componenti fondamentali: le matrici unitarie di sinistra e di destra e una matrice diagonale contenente i valori singolari.
Il quarto passo è il più importante, dato che riguarda la selezione del numero di componenti principali. È bene precisare immediatamente che è obbligatorio selezionare le componenti principali consecutive tra di loro poiché l'ordine di tali componenti principali è determinato dalla loro importanza nella spiegazione della varianza nei dati. Per illustrare meglio, è sbagliato scegliere solamente la quarta,la sesta e

l'ottava componente per ridurre la dimensionalità di un dataset, poiché ciò comprometterebbe l'efficacia dell'intero metodo.
La prima componente principale è quella che cattura la massima varianza totale dei dati, la seconda componente è ortogonale alla prima e cattura la massima varianza rimanente, e così via. L'ordine imposto è quindi decrescente in termini di importanza e quindi di varianza spiegata. Al fine di individuare il numero corretto di cp si possono considerare diversi aspetti:

- Varianza totale spiegata: questo criterio si basa sul numero di componenti principali per ottenere una data proporzione della varianza spiegata. Nella pratica, è possibile imporre questo numero a priori, sulla base del numero di variabili p iniziale. Se il numero di variabili nel dataset iniziale è inferiore alla decina, allora una buona percentuale di varianza spiegata dalle componenti è l'85 %. Se invece la dimensione p delle variabili è superiori a 10, la soglia scelta può variare a seconda del caso. Spesso tale decisione viene semplificata attraverso il grafico dello screeplot, dove nell'asse delle ascisse compare il numero di componenti principali e nell'asse delle ordinate, invece, la proporzione di varianza spiegata da ciascuna componente. La linea che collega i diversi punti presenta nella maggior parte dei casi un "punto di gomito" in corrispondenza del numero di componenti corretto da usare. In altre parole, aiuta a capire rapidamente ed efficacemente quante componenti principali contribuiscono significativamente alla varianza totale;

- Entità degli autovalori: un ottimo strumento per determinare il numero di componenti principali da utilizzare in un'analisi delle componenti principali è esaminare l'entità degli autovalori. Questo è connesso al fatto che più un autovalore è grande, maggiore è la quantità di varianza spiegata dalla componente principale corrispondente. È utile ricordare che la proporzione di varianza spiegata da una cp è data dalla formula (2.3).
  Lo scatterplot può anche considerare nelle ordinate l'entità degli autovalori ordinati in senso decrescente. Il criterio che viene utilizzato è quello di conservare tutte le componenti principali con valore maggiore o uguale a 1, oppure le componenti principali fino al "punto di gomito" che si verrà a formare nel grafico.

- Nelle analisi dei dataset, spesso si predetermina il numero di componenti principali. Ciò è motivato dalla necessità di agevolare la rappresentazione grafica, limitandosi generalmente al massimo a tre componenti. Questa scelta predefinita facilita l'interpretazione visiva dei risultati e contribuisce a una comprensione più chiara dei modelli sottostanti.

Una volta individuato il numero corretto k di componenti principali, il passo successivo è quello di trasformare i dati, ovvero di proiettare i dati originali sul nuovo spazio di dimensione ridotta. La trasformazione dei dati riguarda una moltiplicazione tra $\mathcal{X}$, matrice dei dati originali, e $\mathcal{L}$, matrice dei loading. Queste loadings costituiscono le colonne della matrice di trasformazione. Alla fine dell'intero processo si otterrà $\mathcal{X}_{trasformata} = \mathcal{X}\mathcal{L}^t$, che rappresenta i dati proiettati nello spazio d-dimensionale, catturando però la massima varianza spiegata dei dati originali.

2.1.4 Applicazioni della PCA

La riduzione della dimensionalità attraverso la PCA è un metodo che è possibile applicare in moltissimi ambiti e argomenti. Questi riportati qui sono solamente alcuni esempi.

- Applicazione in ambito medico e biomedico: l'analisi delle componenti principali permette di semplificare il dataset ed in parallelo identificare possibili pattern, tendenze e cluster nei dati clinici. Identificando correttamente le caratteristiche comuni di diverse condizioni patologiche dalle immagini, è possibile giungere ad una classificazione automatica e alla diagnosi precoce di malattie. Questo procedimento è sempre accompagnato da un filtraggio del rumore presente nelle immagini, che migliora la qualità visiva e consente una valutazione più precisa delle condizioni cliniche.
  Come ultimo esempio, si nomina la PCA nel contesto della segmentazione del genoma umano. Il genoma umano è costituito da un gran numero di varianti genetiche che differenziano ciascun individuo e gruppi di individui. L'uso delle cp riduce la dimensione di questi dati, semplificando le possibili analisi ed i modelli che si vogliono effettuare, senza tuttavia perdere la variazione genetica significativa.

- Analisi delle immagini e visione artificiale: l'applicazione della PCA in questo ambito è estremamente significativa per svariati motivi. Le immagini, specialmente quelle ad alta risoluzione, contengono una quantità enorme di informazioni. Le componenti principali permettono di ridurre la dimensionalità di queste immagini, mantenendo allo stesso tempo le caratteristiche più rilevanti. Ciò è particolarmente utile per l'archiviazione e la trasmissione di foto e nell'ambito dei sistemi di riconoscimento facciale. In quest'ultimo è necessario identificare e dare importanza alle sole caratteristiche chiave, che permettono di distinguere un soggetto da un altro.

- Uso di metodi informatici e statistici per la finanza: uno dei metodi utilizzati nell'ambito dei portafogli finanziari e di investimento per analizzare la struttura delle correlazioni tra diversi asset finanziari è proprio l'analisi delle componenti principali. Questo permette di comprendere come le diverse attività si muovono in relazione l'una con le altre, andando a identificare eventuali cambiamenti nelle dinamiche del mercato. Inoltre, questo metodo permette di individuare pattern nei dati storici e di ridurre il rumore. Ciò può migliorare la capacità dei modelli di previsione, cogliendo tendenze sottostanti e generando stime più accurate.

- Sentiment Analysis: l'applicazione della PCA permette di gestire e interpretare grandi quantità di dati testuali. In primo luogo, l'identificare un numero corretto di cp permette di ridurre la dimensione dei vettori di parole e di comprendere quali di questi stilemi, o relative combinazioni tra essi, esprimono meglio un determinato stato d'animo. Incorporare quindi l'analisi delle componenti principali nella sentiment analysis è un ottimo strumento di pre-elaborazione dei dati testuali, aiutando a migliorare la qualità delle rappresentazioni dei testi.

2.2 Regressione con le componenti principali

2.2.1 Introduzione e stima dei coefficienti

La regressione con le componenti principali, PCR, è un approccio che sfrutta l'analisi delle componenti principali per trattare in maniera efficiente la dimensionalità e la multicollinearità tra le variabili indipendenti (Kuhn, Johnson; 2013). Dopo aver applicato la tecnica dell'analisi delle componenti principali sul dataset iniziale, si sceglie un insieme di combinazioni lineari delle $\mathcal{X}_j$ che sono, per definizione, combinazioni non correlate delle variabili originali. Successivamente, le prime k componenti vengono selezionate al fine di adattare un modello di regressione per la variabile risposta. Le variabili esplicative utilizzate sono derivate dalla combinazione lineare, generando, quindi, delle nuove variabili, diverse rispetto a quelle iniziali, che sono incorrelate e ordinate per varianza decrescente. Matematicamente, si afferma che, data una $\mathcal{X}$, matrice dei dati contenente le variabili indipendenti, la decomposizione spettrale della relativa Σ, matrice di varianze e covarianze, porta a:

$$\Sigma\gamma = \lambda^2\gamma \tag{2.6}$$

Indichiamo con γ_j le diverse componenti principali ordinate rispetto all'ordine introdotto dagli autovalori λ_j. Ricordiamo che la più grande componente principale è la direzione che massimizza la varianza dei dati proiettati e l'ultima componente minimizza questa varianza. La regressione con le componenti calcola, quindi, le nuove variabili esplicative derivate $z_j = x\gamma_j$ ed effettua successivamente la regressione di y su $z_1, z_2, ..., z_J$ con $J \leq p$. Si definisce con $\mathcal{Z}_{nxJ}$ la matrice delle componenti principali selezionate costituita dal nuovo set di predittori. Poiché le z_j sono ortogonali tra di loro, la regressione viene calcolata come la somma delle regressioni univariate:

$$y^{pcr} = \bar{y} + \sum_{j=1}^{J} \gamma_j z_j + \epsilon$$

oppure in termini matriciali:

$$\mathcal{Y} = \mathcal{Z}b + \varepsilon$$

dove b è il vettore dei coefficienti di regressione γ_j univarati stimati ed ε è il vettore degli errori residui. La stima dei coefficienti b può essere ottenuta, per esempio, utilizzando il metodo dei minimi quadrati che cerca quel vettore b che minimizza la somma dei quadrati degli errori residui, ovvero:

$$\min_b \|Y - Zb\|^2 \tag{2.7}$$

dove $\|\cdot\|^2$ rappresenta la norma euclidea al quadrato. La soluzione analitica del problema è:

$$\hat{b} = (\mathcal{Z}^t\mathcal{Z})^{-1}\mathcal{Z}^t\mathcal{Y}$$

Questa formula fornisce la stima del vettore dei coefficienti che minimizza l'espressione (2.7). Una volta che si è stimato il vettore dei coefficienti, è possibile utilizzare

il modello per fare previsioni $\hat{y}^{pcr}$ sui nuovi dati. Esplicitiamo ora le applicazioni della PCA nella regressione. In primo luogo, è possibile affermare che la PCA, essendo una tecnica che viene utilizzata anche nella fase esplorativa, rende più chiare le relazioni tra variabili e permette dunque migliori interpretazioni che sarebbero impossibili sennò da ricavare. La riduzione della dimensionalità è un concetto chiave, in quanto permette di arrivare a modelli più parsimoniosi partendo da modelli complessi. Questo comporta il fatto che, non considerando tutte le variabili originali, la PCR cerca di identificare le componenti principali più rilevanti trovando un compromesso tra varianza spiegata e informazioni. Riducendo il numero di variabili, il modello si semplifica, e quindi si può giungere ad una maggiore generalizzazione e interpretabilità. Si può quindi dire che si giunge ad una maggior stabilità e a delle migliori prestazioni del modello di regressione.
Un ulteriore problema che la PCR risolve in maniera efficiente è la multicollinearità delle variabili. La multicollinearità si verifica quando due o più variabili indipendenti nello stesso modello statistico sono fortemente correlate tra di loro e questo in genere comporta problemi nella stima dei coefficienti nel modello e nelle errate comprensioni. Le componenti principali nel modello, essendo per definizione ortogonali tra di loro e non correlate, arrivano a stime migliori e quindi a letture più corrette.

2.2.2 Limiti e problemi della PCR

La PCR, tuttavia, è criticata per diverse ragioni. Il fatto che la prima componente principale spieghi la massima varianza rispetto a tutte le componenti non implica che essa sia anche la più utile per prevedere la variabile risposta y. Si afferma quindi che uno dei limiti più grandi della regressione con le componenti principali è l'assenza di un ordinamento in termini predittivi e quindi la scelta corretta del numero di componenti per prevedere un vettore y è più complicata rispetto alla scelta fatta in precedenza nella PCA.
Un secondo limite riguarda il problema dell'overfitting quando il dataset viene diviso in più parti. Una tecnica che viene utilizzata per giudicare correttamente la previsione di una variabile risposta da parte di un metodo o di un algoritmo è la divisione in insieme di stima, o *training set*, e insieme di verifica, o *test set*. Il *training set*, che in genere contiene il 75% delle osservazioni iniziali, ha l'obiettivo di adattare i diversi modelli candidati. Il *test set*, invece, contiene il restante 25% delle unità statistiche che vengono utilizzate come fossero informazioni reperibili solo in futuro. Ha, quindi, lo scopo di valutare le prestazioni dei modelli disponibili e di confrontarli per scegliere il più accurato. Se si effettua l'analisi delle componenti principali sull'intero dataset si sta commettendo un errore di overfitting, dato che si utilizzano sia i dati di oggi (insieme di stima) che le informazioni reperibili solo domani (insieme di verifica). Se invece si effettua la PCA unicamente sul *training set*, i loadings ottenuti devono poi essere riportati e trasformati in modo che si adattino al meglio al *test set*. Questo problema genera successivamente errori nella valutazione e nell'interpretazione dei coefficienti quando si effettua la regressione con le componenti principali (Azzalini, 2012).

Capitolo 3

Projection Pursuit e PPR

3.1 Projection Pursuit

3.1.1 Principi e fondamenti

La Projection Pursuit è una tecnica statistica che viene utilizzata per l'analisi esplorativa di un dataset. Ideata da J.B. Kruskal nel 1969, e successivamente perfezionata dalla collaborazione di Jerome H. Friedman e John Tukey, mira a trovare una trasformazione dei dati che permetta di massimizzare la varianza della variabile dipendente. Questo approccio è particolarmente utile quando si ipotizza che le relazioni tra le variabili esplicative e la variabile risposta sono complesse e non possono essere descritte da una semplice regressione lineare. È bene precisare all'inizio del capitolo che questo metodo incoraggia lo sviluppo di metodologie capaci di cogliere strutture significative nei dati e di visualizzare successivamente le informazioni trovate in una, due oppure tre dimensioni (Jones,1987).
La Projection Pursuit si fonda su concetti chiave specifici che delineano le diverse metodologie. A differenza dell'analisi delle componenti principali, la PP non impone che la ricerca si basi unicamente sulla variabilità dei dati. Essa, infatti, da maggiore importanza nella ricerca della funzione di proiezione. Questa funzione determina quali aspetti o pattern vogliamo analizzare attraverso la proiezione dei dati da uno spazio di dimensione p ad uno di dimensione d con $d < p$. Gli indici di ricerca sono gli oggetti che vengono utilizzati per selezionare la corretta proiezione ottimale. Si esploreranno dettagliamene questi concetti nel corso del capitolo.
Nella PCA, l'obiettivo primario è identificare le direzioni di massima varianza nei dati. Tuttavia, questa massimizzazione della varianza non permette di garantire le rilevazioni di strutture significative nei dati non lineari. La Projection Pursuit generalizza e amplia la ricerca delle proiezioni ottimali. Quest'ultime, infatti, sono indirizzate verso la massimizzazione della varianza associata a strutture significative, o in altre parole mirano a trovare quegli assi o quelle direzioni in cui i dati si estendono il più possibile. Questo concetto chiave nella projection pursuit garantisce una maggior flessibilità al metodo e permette sia di trovare informazioni più rilevanti e più significative, sia di adattare modelli di dati più complessi, in quanto la trasformazione delle variabili non si limita al solo caso lineare ma permette anche trasformazioni non lineari, come ad esempio curve o superfici. Per concludere, la PP regola la massimizzazione della varianza a seconda degli obiettivi specifici dell'ana-

lisi per arrivare a risultati di maggior interesse rispetto alle tecniche che utilizzano approcci più lineari.

3.1.2 Indici di proiezione

Sia $\mathcal{X} \sim (0, \mathcal{I})$ un vettore aleatorio di dimensione p, con $\vec{0}$ vettore delle medie e matrice d'identità di ordine d come matrice di varianza e covarianza, e sia $a \in \mathcal{R}^d$ un vettore di direzione. Un indice di proiezione $\mathcal{Q}$ che misura la deviazione dalla normalità è una funzione che assegna un numero reale alla coppia $(\mathcal{X}, a)$ e soddisfa le condizioni per cui:

- $\mathcal{Q} \geq 0 \quad \forall(\mathcal{X}, a)$
- $\mathcal{Q}(\mathcal{X}, a) = 0 \iff \mathcal{X}_a = a^t \mathcal{X} gaussiano$

Gli indici di ricerca sono gli oggetti che vengono utilizzati per selezionare la corretta proiezione ottimale. Questi sono definiti in base agli obiettivi specifici dell'analisi e dal contesto dell'applicazione e sono quindi scelti a priori. Per esempio, se il fine dell'analisi fosse la rilevazione di cluster, la funzione di proiezione potrebbe quindi essere basata sulla varianza interclasse; se invece volessimo ricercare le code pesanti in una distribuzione si potrebbe utilizzare una funzione basata sulla curtosi. Qui di seguito verranno analizzate delle diverse tipologie di indici di proiezioni, e per ciascuno verrà fornita una breve descrizione.

Indice di massimizzazione della varianza

Il primo indice di proiezione analizzato è l'indice di massimizzazione della varianza, conosciuto anche come Variance Index.

$$\mathcal{Q}_{VI} = \frac{\sum_{i=1}^{n} \left(w_i \cdot (x_i - \bar{x})\right)^2}{\sum_{i=1}^{n} w_i^2} \tag{3.1}$$

dove:

- w_i sono i pesi associati a ciascuna osservazione nella proiezione
- x_i sono le osservazioni
- $\bar{x}$ è la media delle osservazioni

Tale formula permette di calcolare la varianza in maniera ponderata considerando il contributo relativo di ciascuna osservazione nella proiezione. Ciò permette di enfatizzare i particolari pattern nascosti, in linea quindi con gli obiettivi specifici dell'analisi della Projection Pursuit. La normalizzazione $\sum_{i=1}^{n} w_i^2$ garantisce che la varianza sia valutata in modo coerente rispetto ai pesi assegnati, evitando che il risultato sia sproporzionato per pesi particolarmente alti o bassi. La scelta dei pesi è un processo che richiede una combinazione di intuizione e sperimentazione, in quanto la scelta più corretta dipende dalle specifiche caratteristiche dei dati. Si possono

selezionare pesi maggiori o per importanza delle osservazioni, se la volontà è quella di enfatizzare alcune unità statistiche rispetto ad altre, oppure per distribuzione delle variabili, se si vuole bilanciare l'impatto delle variabili.
L'indice di massimizzazione della varianza può essere utilizzato per:

- Rilevare strutture complesse
- Analisi di cluster
- Scoperta di ouliers
- Esplorazioni di relazioni non gaussiane.

Indice di ricerca di cluster

Questo indice, conosciuto come Cluster Search Index, è stato progettato per individuare le proiezioni che mettono in evidenza i diversi cluster nei dati multivariati. La formula matematica che lo caratterizza è:

$$\mathcal{Q}_{CSI} = \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n} d_{ij}}{\sum_{i=1}^{n} w_i} \tag{3.2}$$

dove:

- w_i sono i pesi associati a ciascuna osservazione
- d_{ij} sono le distanze tra le osservazioni i e j

Al numeratore è presente la somma delle distanze tra tutte le coppie di osservazioni. Questa somma valuta la dispersione delle unità statistiche nella proiezione, rilevando quanto le diverse regioni dello spazio delle osservazioni siano separate. La normalizzazione di $\sum w_i$ permette, come nel caso precedente, di valutare in maniera coerente il calcolo. Questo indice permette di scoprire cluster nascosti, non evidenti nelle proiezioni lineari, identificando proiezioni che massimizzano la distanza tra gruppi di osservazioni. Esso permette inoltre di esplorare la similarità e la dissimilarità tra i cluster.
É quindi possibile affermare che una somma delle distanze più bassa indica una maggiore coesione all'interno dei cluster, mentre una somma più elevata suggerisce invece una separazione tra essi.

Indice di Linearità/Non Linearità

Un altro indice da considerare nell'ambito della Projection Pursuit è l'Indice di Linearità/Non-Linearità, noto anche come Linearity Index. Questo indicatore svolge un ruolo chiave nell'analisi e nella valutazione del livello di linearità presente nelle proiezioni. La sua inclusione nella ricerca fornisce un approfondimento essenziale per comprendere la complessità delle relazioni tra le variabili e migliora la capacità di delineare pattern non lineari all'interno dei dati proiettati. Questa metrica costituisce un elemento fondamentale per una valutazione completa e approfondita delle caratteristiche delle proiezioni ottenute mediante Projection Pursuit, contribuendo

così a una comprensione più accurata della struttura sottostante dei dati. La formula matematica che viene utilizzata è:

$$\mathcal{Q}_{LI} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{n}\sum_{j=1}^{n} w_i \cdot w_j \cdot \operatorname{corr}(x_i, x_j)\right| \tag{3.3}$$

La parte fondamentale che lo differenzia dagli altri è la presenza della correlazione ponderata tra le osservazioni e questo permette di valutare se le proiezioni siano linearmente correlate. La formula calcola successivamente una media ponderata delle correlazioni tra tutte le coppie di osservazioni, offrendo un indicatore sulla linearità generale della proiezione. L'utilizzo di questo indice guida ad una scelta corretta sulla trasformazione della variabile da applicare.

Indici di entropia

L'entropia viene definita come la quantità media di sorpresa associata a un evento in un sistema di informazione. Maggiore è l'entropia, maggiore è l'incertezza o il disordine nel sistema, indicando quindi una maggiore imprevedibilità degli eventi (Shannon,1948). Gli indici di entropia nella PP consentono di misurare la complessità informativa delle proiezioni, rilevando la quantità di informazione e di disordine presenti intrinseca nei dati. Questo permette dunque al metodo di selezionare le proiezioni che massimizzano la varietà informativa. Le formule matematiche degli indici di entropia che la Projection Pursuit comunemente utilizza sono:

$$Q_E(X,a) = \int f_a \log f_a, \qquad \text{e} \qquad Q_{E'}(X,a) = Q_E(X,a) + \log\left[(2\pi e)^{\frac{1}{2}}\right] \tag{3.4}$$

Questi due indici differiscono unicamente per una costante pari a $log[(2\pi \exp)^{\frac{1}{2}}] \approx 0.62$

3.1.3 Ottimizzazione

L'obiettivo successivo richiede quindi di trovare quel vettore a che permetta di massimizzare l'indice di proiezione Q(v) attraverso

$$v = \operatorname{argmax}_{v \in R^d}\left(Q\left(v\right)\right) \tag{3.5}$$

La ricerca del vettore di proiezione ottimale v può essere effettuata utilizzando diversi metodi di ricerca e di ottimizzazione numerica che esplorano lo spazio delle possibili proiezioni. Questi appartengono ad un'area di studio che è in continua evoluzione con i miglioramenti delle tecniche di machine learning e la crescente complessità dei dati da analizzare. Questi metodi sono ad esempio:

- Discesa del gradiente: nella Projection Pursuit la discesa del gradiente misura la bontà di una particolare proiezione nell'ambito del clustering, della massimizzazione della varianza o di altri criteri d'interesse. Si suppone, quindi, di avere $\mathcal{Q}(\theta)$ come funzione obiettivo, dove θ rappresenta il parametro della proiezione che vogliamo ottimizzare. Il passo iniziale usa come partenza una

proiezione (θ_0) che può essere scelta anche casualmente. Il gradiente per una funzione f (o Δf) è un vettore composto dalle derivate parziali rispetto a ciascuna funzione e, oltre ad indicare la pendenza della funzione in ogni punto, permette di determinare la direzione in cui muoversi al fine di raggiungere il minimo. Successivamente i parametri della proiezione vengono aggiornati nella direzione opposta al gradiente e moltiplicati per un tasso di apprendimento α che assume un ruolo fondamentale in tale algoritmo. Questo parametro critico non assume un valore specifico ma varia da caso a caso: un valore troppo alto può portare a mancata convergenza, mentre un valore troppo basso potrebbe ingannare e far convergere l'algoritmo a massimi/minimi locali. Questo passaggio viene ripetuto iterativamente:

$$\theta_{n+1} = \theta_n - \alpha \cdot \nabla \mathcal{Q}(\theta_n) \tag{3.6}$$

- Newton-Raphson: un altro algoritmo di ottimizzazione che permette di ottenere una particolare proiezione è il metodo di Newton-Raphson. Presa una funzione obiettivo da massimizzare $\mathcal{Q}(\theta)$ con θ parametri della proiezione, l'aggiornamento è dato da:

$$\theta_{n+1} = \theta_n - \mathcal{H}^{-1} \Delta Q(\theta_n) \tag{3.7}$$

  L'utilizzo della matrice hessiana rende questo metodo più potente rispetto a metodi che si basano solo sul gradiente perché, in questo modo, viene data importanza anche alla curvatura che assume la funzione obiettivo in un punto dello spazio. Tuttavia, l'utilizzo della matrice H richiede uno sforzo maggiore dal punto di vista di calcolo e limita l'utilizzo di questo metodo alle sole funzioni che sono derivabili almeno due volte e che presentano la matrice hessiana non singolare.

Una volta selezionato il criterio più corretto per l'analisi da effettuare, si itera il metodo fino a convergenza, ovvero fino a quando i cambiamenti successivi nei parametri in (3.6) e (3.7) diventano inferiori ad una determinata soglia, oppure fino al raggiungimento del numero massimo di iterazioni scelto a priori. Viene quindi scelto il vettore $v = (v_1, v_2, .., v_k)$ che permette di massimizzare l'indice di proiezione più adeguato al dataset da analizzare.

3.2 Regressione con la Projection Pursuit

La projection pursuit regression costruisce un modello a partire dalla direzione trovata. Le variabili indipendenti del dataset originale vengono proiettate su queste direzioni per creare delle nuove variabili. Supponiamo per esempio di aver trovato K direzioni di proiezione ottimali $v_1, v_2, .., v_k$ e successivamente di calcolare le proiezioni dei dati originali su queste direzioni, ottenendo $z_1, z_2, \ldots, z_k$. Si ottiene quindi il modello:

$$\mathcal{Y} = \beta f(XW) + \epsilon = \beta_0 + \sum_{j=1}^{p} \beta_j (f(\mathcal{X}_j w_j)) + \epsilon \tag{3.8}$$

- Y è la variabile dipendente
- p è il numero di variabili indipendenti
- X è la matrice delle variabili indipendenti
- W è la matrice di proiezione ottimale
- $f(\cdot)$ è una funzione di trasformazione
- β_j sono i coefficienti univariati associati alle variabili trasformate
- ϵ è il termine di errore

La formula esplicita,invece, per calcolare le previsioni è data da:

$$\hat{Y} = \beta_0 + \sum_{j=1}^{p} \beta_j z_j + \epsilon \tag{3.9}$$

dove z_j sono i nuovi dati trasformati.

Capitolo 4

Applicazione al dataset

4.1 Introduzione

In questo capitolo si applicheranno gli algoritmi precedentemente analizzati al dataset Vinho Verde, attraverso R e Rstudio. Le librerie e le funzioni principali utilizzate sono:

- funzione *prcomp*: è il comando che si utilizza in questo studio per applicare la PCA al dataset.
- libreria *pls*: per l'adattamento del modello con l'uso delle componenti principali;
- libreria *ggplot2* e *factoextra*: le librerie in questione permettono di rappresentare i risultati ottenuti dalla PCA e dalla PP per valutare il processo;
- libreria *corrplot*: tale strumento è stato esaminato per comprendere le correlazioni presenti tra le componenti principali e le variabili originali;
- libreria *Pursuit*: tale libreria permette di utilizzare diverse funzioni per l'analisi della Projection Pursuit tra cui GrandTour, funzione per trovare l'indice e per ottimizzare;
- funzione *ppr*: è il comando che viene utilizzato per adattare un modello con la Projection Pursuit;
- libreria *tidymodels*: permette di dividere il dataset in un insieme di stima e un insieme di verifica.

4.2 PCA

Sia $\mathcal{X}$ la matrice dei dati relativi al caso di studio, contenente le 11 variabili dipendenti. Al fine di applicare correttamente l'analisi delle componenti principali, si applica l'operazione di standardizzazione alla matrice, in modo da evitare i possibili problemi di scala e di privare delle diverse unità di misura.
Successivamente, si applica la PCA alla matrice dei dati standardizzati e si ottiene l'output contenente:

- *rotation*: la matrice di rotazione rappresenta gli autovettori che definiscono le combinazioni lineari delle variabili originali per formare le componenti principali, ovvero fornisce una guida su come ciascuna variabile contribuisce alla formazione delle cp. Per spiegare meglio questo concetto chiave all'interno della PCA, viene riportata la colonna relativa alla combinazione lineare per la prima componente principale.

Variabile	**Contributo su PC1**
fixed acidity	0.239
volatile acidity	0.381
citric acid	-0.152
residual sugar	-0.346
chlorides	0.290
free sulfur dioxide	-0.431
total sulfur dioxide	-0.487
density	0.045
pH	0.219
sulphates	0.294
alcohol	0.106

Tabella 4.1: Contributo delle variabili relativo alla prima componente principale.

Si può notare che una maggiore presenza di *volatile acidity* e *chlorides* è associata ad un aumento positivo della prima componente principale, mentre una maggior presenza di *citric acid* e di *free sulfur dioxide* ne comporta una riduzione. La visione dei carichi di ciascuna variabile fornisce ulteriori interpretazioni per l'analisi delle componenti principali;

- *sdev*: le deviazioni standard delle componenti principali, corrispondenti agli autovalori, indicano quanto ciascuna cp contribuisce alla varianza totale dei dati;

- *center*: la media delle variabili originali, utilizzate per l'analisi delle componenti principali. Essendo stata effettuata l'operazione di standardizzazione, tale vettore risulterà nullo;

- *scale*: i fattori di scala riflettono la deviazione standard originale di una specifica variabile durante la PCA. La tabella sottostante indica i fattori di scala associati alle diverse variabili di input.

Variabile	**Fattore di Scala**
Fixed Acidity	1.300
Volatile Acidity	0.160
Citric Acid	0.150
Residual Sugar	4.760
Chlorides	0.040
Free Sulfur Dioxide	17.750
Total Sulfur Dioxide	56.520
Density	0.003
pH	0.160
Sulphates	0.150
Alcohol	1.190

Tabella 4.2: tabella con le variabili e i relativi fattori di scala

Si può osservare che *free sulfur dioxide* e *total sulfur dioxide* hanno associati i più alti valori dei fattori di scala e quindi rappresentano le variabili con maggior variabilità. *Density*, invece, presenta la più piccola variabilità;

- x: la matrice rappresenta le osservazioni nei nuovi spazi delle componenti principali, ossia i dati proiettati nelle dimensioni definite dalle cp estratte durante l'analisi. Ogni riga della matrice corrisponde a un'osservazione dei dati originali, mentre le colonne rappresentano le diverse componenti principali;

- *summary*: l'output di questo comando fornisce un riassunto delle informazioni più utili ottenibili dall'analisi delle componenti principali. Questo oggetto contiene le deviazioni standard di ciascuna componente principale, la proporzione di varianza spiegata e la proporzione cumulativa.

CP	Std.deviation	Prop. of Variance	Cumulative Prop.
PC1	1.741	0.275	0.275
PC2	1.580	0.227	0.502
PC3	1.248	0.142	0.644
PC4	0.985	0.088	0.732
PC5	0.848	0.065	0.797
PC6	0.780	0.055	0.852
PC7	0.723	0.048	0.900
PC8	0.708	0.045	0.945
PC9	0.581	0.031	0.976
PC10	0.477	0.021	0.997
PC11	0.181	0.003	1.000

Numero di componenti

La selezione del numero di componenti principali è una fase fondamentale che incide significativamente sull'interpretazione e l'efficacia dell'analisi stessa. La determinazione del numero ottimale di componenti principali da includere nell'analisi è stata

basata su un approccio ponderato, combinando diverse considerazioni.
Innanzitutto, è stata valutata la varianza spiegata dalle singole componenti principali. Gli autovalori associati a ciascuna componente forniscono una misura della quantità di varianza che la componente riesce a spiegare. La scelta è stata orientata verso le prime 3 componenti principali le cui entità presentano autovalori significativamente superiori ad altri. Tale criterio permette di mantenere le componenti principali che catturano la maggior quantità di varianza nei dati, garantendo al contempo una riduzione significativa della dimensionalità.

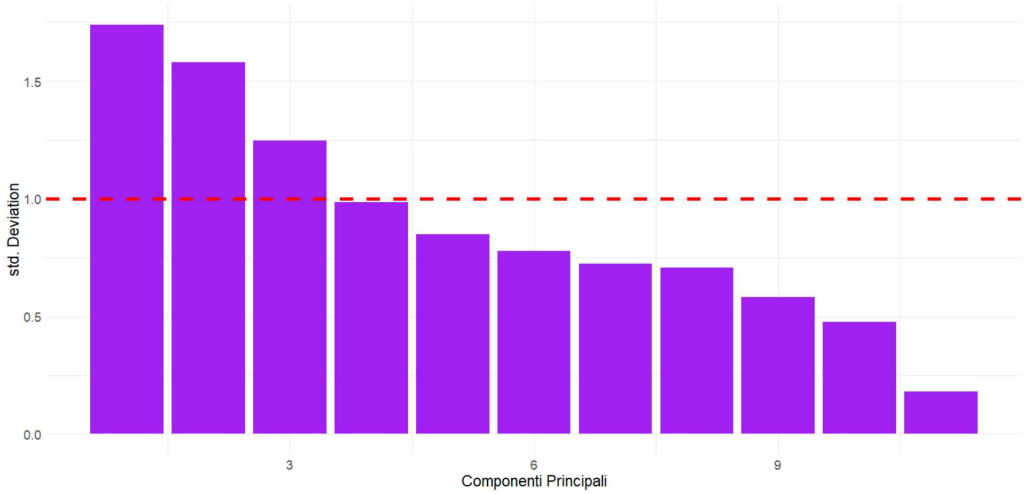

Figura 4.1: Istogramma dei valori degli autovalori (asse y) rispetto alle componenti (asse x).

Un secondo criterio preso in considerazione è la deviazione standard delle componenti principali: componenti con deviazioni standard superiori a 1 indicano una maggiore capacità di spiegare la varianza nei dati, influenzando positivamente la rappresentazione globale del modello.
Infine, la scelta è stata validata considerando la proporzione cumulativa di varianza spiegata. Si è considerato che il 64% della varianza spiegata potesse essere una ottima percentuale per il caso in questione dato che il numero di variabili presenti è superiore alla decina.

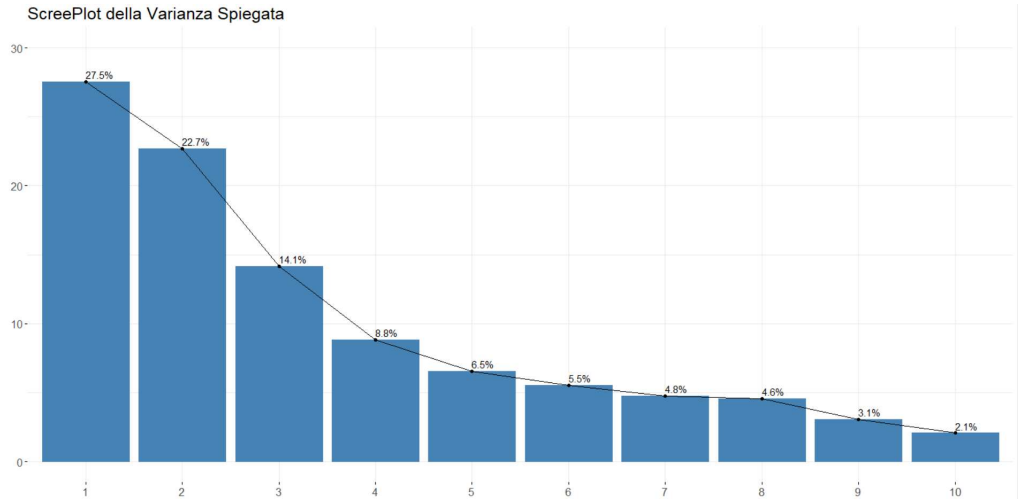

Figura 4.2: Istogramma delle percentuali di varianza spiegata per ciascuna componente principale.

Complessivamente, la scelta di includere 3 componenti principali è stata quindi guidata da una combinazione di criteri che mirano a massimizzare la varianza spiegata, a garantire la significatività delle componenti selezionate e a ridurre al minimo la perdita di informazione durante la riduzione della dimensionalità.

Conclusioni

La decisione di utilizzare tre componenti principali consente di condensare le informazioni provenienti da 11 variabili in un grafico tridimensionale. Tuttavia, per una rappresentazione visuale, verranno esaminati i confronti tra le tre componenti in due dimensioni.

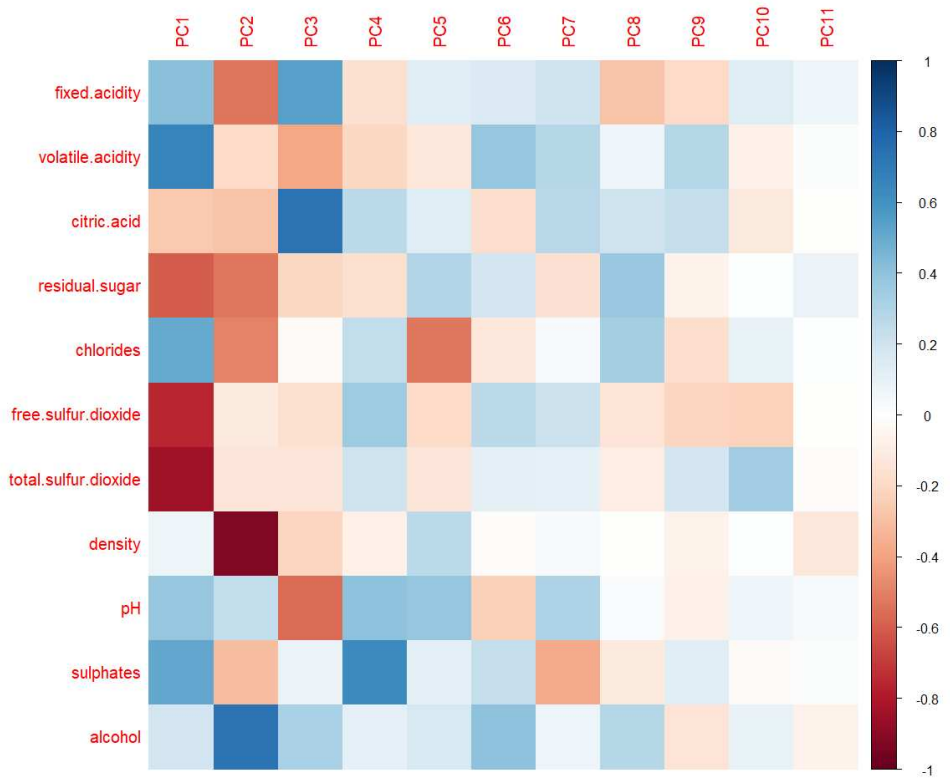

Figura 4.3: Grafico a doppia entrata relativo alle correlazione tra le variabili originarie, asse y, e le cp, asse x.

Dall'analisi del grafico (4.3) e dagli autovalori, emerge che la prima componente principale è fortemente influenzata dalle variabili originarie con correlazioni estreme. In particolare, *free sulfur dioxide* e *total sulfur dioxide* mostrano una correlazione prossima a -1, mentre *volatile acidity* presenta una correlazione positiva. Ciò implica che un aumento di *volatile acidity* è associato a una diminuzione delle altre due variabili legate all'anidride solforosa.
La seconda componente principale è caratterizzata dalle variabili *density*, con correlazione negativa, e *alcohol*, con correlazione positiva. In analogia al caso precedente, queste due variabili mostrano una relazione inversa, indicando che un aumento di una variabile si traduce in una diminuzione dell'altra.
La terza componente principale, simile alla prima, coinvolge tre variabili. Tuttavia, in questo caso, *fixed acidity* e *citric acidity* presentano correlazioni positive, mentre *pH* mostra una correlazione opposta rispetto alle prime due variabili.
Queste informazioni emergono anche dal grafico del biplot, ovvero la rappresentazione che permette di visualizzare contemporaneamente sia le osservsazioni che le variabili in uno spazio bidimensionale, facilitando la comprensione delle relazioni e delle tendenze presenti nei dati multivariati.

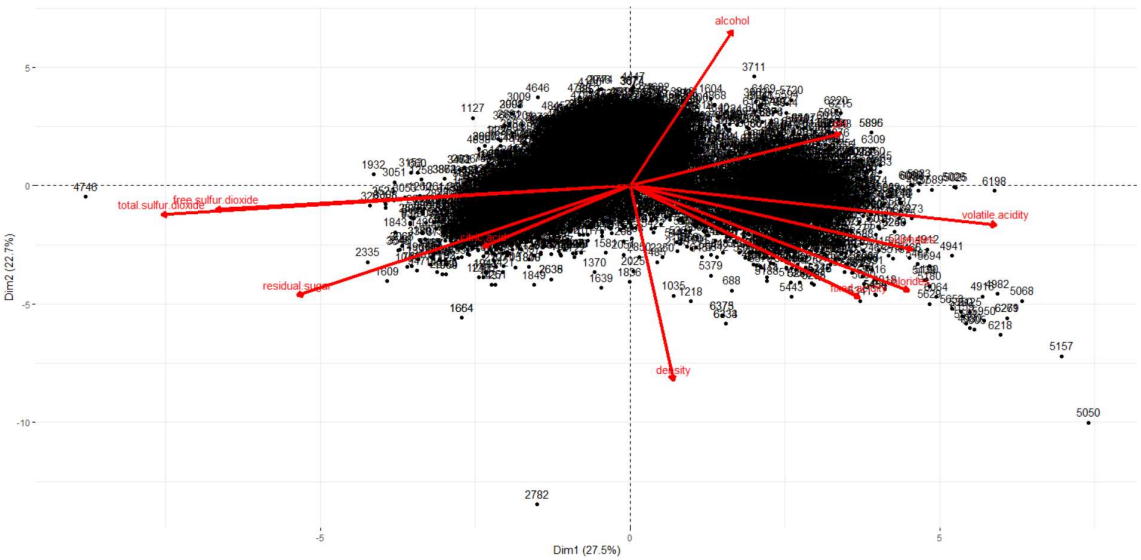

Figura 4.4: Biplot delle prime due componenti principali.

Il grafico (4.4) offre una visualizzazione della proiezione dei dati in uno spazio bidimensionale definito dalle due componenti principali. La prima componente principale (asse x) è distintamente influenzata dalle variabili legate all'anidride solforosa, evidenziato dalla direzione estrema delle frecce lungo l'asse delle ascisse. D'altra parte, la seconda componente principale è graficamente delineata dal dualismo tra *density* e *alcohol*.
È interessante notare che alcune unità, come la 2782-esima, non sono adeguatamente spiegate utilizzando solo due componenti principali. Ciò suggerisce che la varietà e la complessità dei dati potrebbero richiedere una rappresentazione tridimensionale o l'inclusione di ulteriori componenti principali per catturare appieno la varianza nei dati.

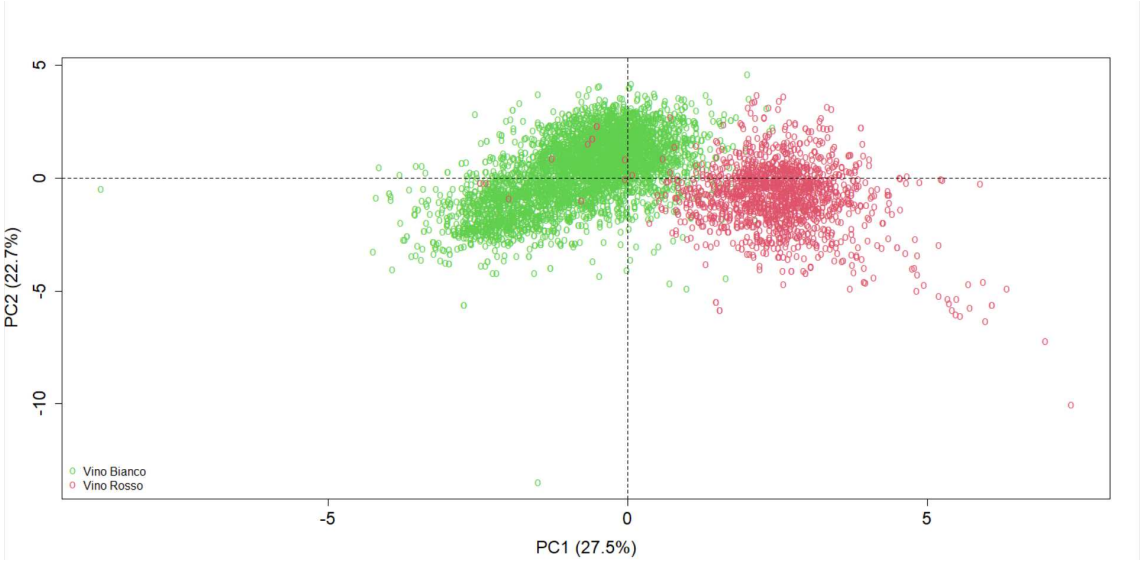

Figura 4.5: Diagramma di dispersione delle unità statistiche rispetto la prima (asse x) e la seconda (asse y) componente principale.

Dall'analisi del grafico (4.5), emerge chiaramente che l'impiego delle prime due componenti principali consente di distinguere i vini bianchi dai rossi. Questo risultato è evidenziato dal fatto che i punti rossi presentano valori positivi sulla prima componente principale, mentre i punti verdi mostrano valori più negativi. Tuttavia, per quanto riguarda la seconda componente principale, la distinzione tra le due tipologie di vino non è altrettanto evidente. Infatti, sia i punti rossi che quelli verdi si distribuiscono intorno al valore nullo su questa componente.
Questa osservazione suggerisce che la prima componente principale spiega una percentuale maggiore di varianza rispetto alla seconda (rispettivamente, 27.5% e 22.7%) e che l'utilizzo di sole due componenti principali non è sufficiente per catturare in modo completo e dettagliato le differenze tra le due tipologie di vini, evidenziando la complessità della struttura dei dati.

4.3 Regressione con le componenti principali

Attraverso l'applicazione della PCA al dataset, si sono ridotte le dimensioni mantenendo le informazioni più rilevanti. Utilizzando queste informazioni, intendiamo implementare un modello di regressione per prevedere la variabile risposta *quality*. Il dataset è stato suddiviso casualmente in un insieme di stima, che contiene il 75% delle osservazioni (4872 unità), e un insieme di verifica, che rappresenta la restante parte. Il primo sarà impiegato per adattare e ottimizzare i diversi modelli candidati, mentre il secondo sarà utilizzato per valutare le prestazioni predittive dei modelli e selezionare quello più accurato. Questo approccio consente di simulare la previsione su dati futuri, contribuendo a garantire la validità e l'efficacia del nostro modello (James, Witten, Hastie, Tibshirani, 2013).
L'utilizzo di 3 componenti principali permette di implementare un modello lineare

del tipo:

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \epsilon$$

Il modello ha un termine di intercetta significativo (p-value prossimo allo zero), indicando che c'è una relazione significativa tra le variabili indipendenti e dipendenti. I valori ottenuti sono:

$$\hat{y} = 5.827 + 0.041 z_1 - 0.183 z_2 - 0.137 z_3 \tag{4.1}$$

I coefficienti associati a PC1, PC2 e PC3 indicano l'effetto stimato di ciascuna componente su *quality*. La prima componente, legata maggiormente alle variabili dell'anidride solforosa e all'*acidità volatile*, presenta un coefficiente positivo e quindi, un aumento di una unità di z_1 comporta un aumento di circa 0.041 unità in *quality*. La seconda, definita prevalentemente dalla densità del vino e dall'alcol presente, assume invece un valore negativo, indicando quindi che l'aumento di una unità porta ad una riduzione del giudizio dato. La terza ed ultima componente, legata all'acido citrico, all'acidità fissa e al ph del vino, assume un valore negativo, e quindi ad un aumento di z_2 ne consegue una riduzione della variabile risposta.
La significatività dei coefficienti è valutata attraverso i valori dei livelli di significatività osservati. In questo caso, tutti i coefficienti sono significativi (pvalue < 0.05), indicando che le tre componenti principali sono statisticamente significative nella predizione della qualità. Tuttavia, bisogna precisare che l'elevata numerosità del campione e l'utilizzo di un numero ridotto di coefficienti abbassa i livelli di significatività osservati, e quindi i risultati ottenuti potrebbero essere distorti.
Il valore assunto dall'R^2 è 0.155, il che suggerisce che il modello spiega solamente il 15.5% della variazione nella variabile dipendente. Il valore assunto da questo indice è basso, mettendo in dubbio l'utilizzo di tre sole componenti principali. La scelta del numero di componenti è stata effettuata trovando un equilibrio tra numero di coefficienti impiegati, R^2 e l'errore quadratico medio. Quest'ultimo è una misura della deviazione tra i valori previsti e quelli osservati in un insieme di dati, ed è utile a valutare la precisione di un modello in termini predittivi. Gioca un ruolo fondamentale per valutare quanto il modello si discosti dai dati reali, e quindi l'obiettivo durante il processo di stima è quello di minimizzare l'EQM per arrivare ad una maggiore precisione nelle previsione.

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

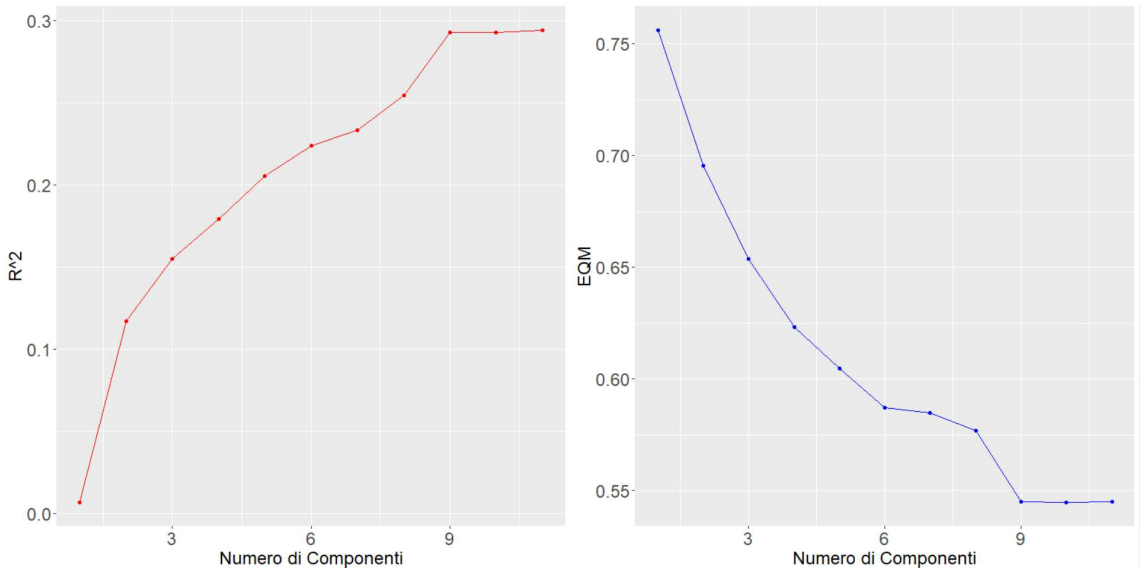

Figura 4.6: Grafici contenenti rispettivamente i valori degli R^2 e degli EQM (asse y) assunti dai diversi modelli rispetto al numero di componenti (asse x) .

L'utilizzo di 3 componenti ha quindi permesso di arrivare a prevedere e confrontare 1625 osservazioni. I risultati ottenuti sono riportati nella tabella sottostante.

Quality	Valori Osservati	Valori Previsti
3	9	1
4	59	0
5	539	279
6	717	1329
7	254	16
8	46	0
9	1	0

Tabella 4.3: tabella di confronto tra valori osservati di *quality* e valori previsti dal modello (4.1)

Dalla Tabella 4.3 emerge chiaramente che il modello produce previsioni insoddisfacenti, classificando approssimativamente l'82% delle osservazioni nell'etichetta 6, e circa il 99% dei vini viene giudicato medio per la qualità.
L'evidente tendenza del modello a valutare erroneamente un'ampia percentuale di osservazioni suggerisce la presenza di un possibile errore nella percezione della relazione lineare tra la variabile di risposta e le variabili indipendenti. Questo fenomeno stimola la considerazione di approcci alternativi, come l'adozione di relazioni non lineari della Projection Pursuit, al fine di catturare in maniera più accurata la complessità e la varietà dei dati sottostanti.

4.4 Projection Pursuit

L'applicazione della Projection Pursuit al dataset standardizzato viene inizializzata attraverso la procedura del *Grand Tour*. Tale studio non si limita semplicemente a rivelare i diversi pattern non lineari del dataset, ma si propone di utilizzare ulteriori mezzi nell'analisi dei dati per visualizzare dinamicamente le diverse proiezioni bidimensionali. Nello specifico, l'applicazione dell'interpolazione consente di passare in modo graduale da una proiezione all'altra anziché farla sbrigativamente, migliorando la comprensione dell'osservatore. L'interpolazione può essere applicata a vari aspetti delle proiezioni, tra cui:

- Rotazione dello spazio;
- Variazioni casuale dato dalle perturbazioni nello spazio dei dati;
- Cambiamenti dei valori dei parametri;

La figura sottostante permette di visualizzare quattro diversi passaggio del *Grand Tour* per comprendere meglio il relativo utilizzo.

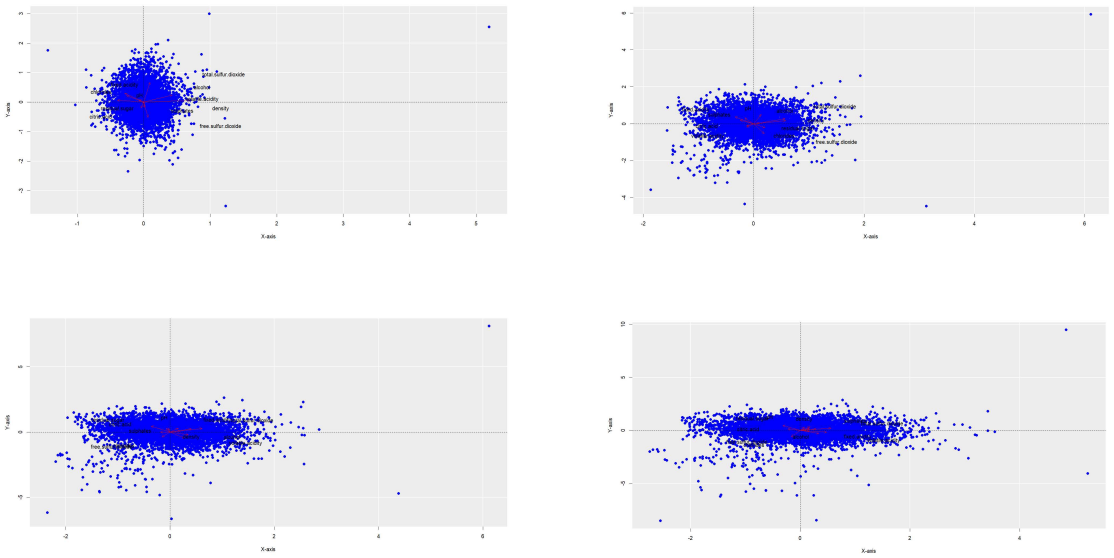

Figura 4.7: Diagramma di dispersione delle osservazioni proiettate in uno spazio bidimensionale attraverso il metodo *Grand Tour*

Come affermato nel capitolo teorico, la Projection Pursuit mira a trovare quella proiezione che permette di visualizzare nel modo migliore i dati proiettati in uno spazio d-dimensionale (in questo caso d=2). Ad esempio, la proiezione illustrata nella prima figura risulta incapace di catturare appieno le informazioni e la diversità presenti nei dati, poiché le unità sono condensate all'interno di un cerchio. La proiezione nell'ultima figura, invece, comprende la variazione lungo l'asse delle ascisse dato che le unità risultano più dilatate.
Successivamente, vengono analizzati le diverse proiezioni effettuate rispettivamente in 1 e in 2 dimensioni:

- Proiezione in 1 dimensione attraverso l'indice di curtosi: Attraverso un'analisi preliminare, è stato scelto l'indice di curtosi come misura chiave per la proiezione unidimensionale. Questo indice valuta la forma della distribuzione dei dati, indicando un'elevata curtosi per code più spesse e picchi più alti, e una bassa curtosi per code più sottili e picchi più bassi rispetto a una distribuzione normale. La ragione di questa selezione è motivata dal fatto che le proiezioni della Projection Pursuit (PP) sono progettate per catturare strutture più nascoste nei dati. In questo contesto, l'applicazione dell'indice di curtosi permette di evidenziare eventuali deviazioni dalla normalità nei dati. La massimizzazione dell'indice di curtosi favorisce la rivelazione di tali deviazioni, mentre la minimizzazione mira a ottenere proiezioni più simili a una distribuzione gaussiana. Questa scelta strategica consente di affinare la sensibilità delle proiezioni della PP alle caratteristiche complesse e non evidenti dei dati, migliorando così la capacità di identificare pattern significativi.

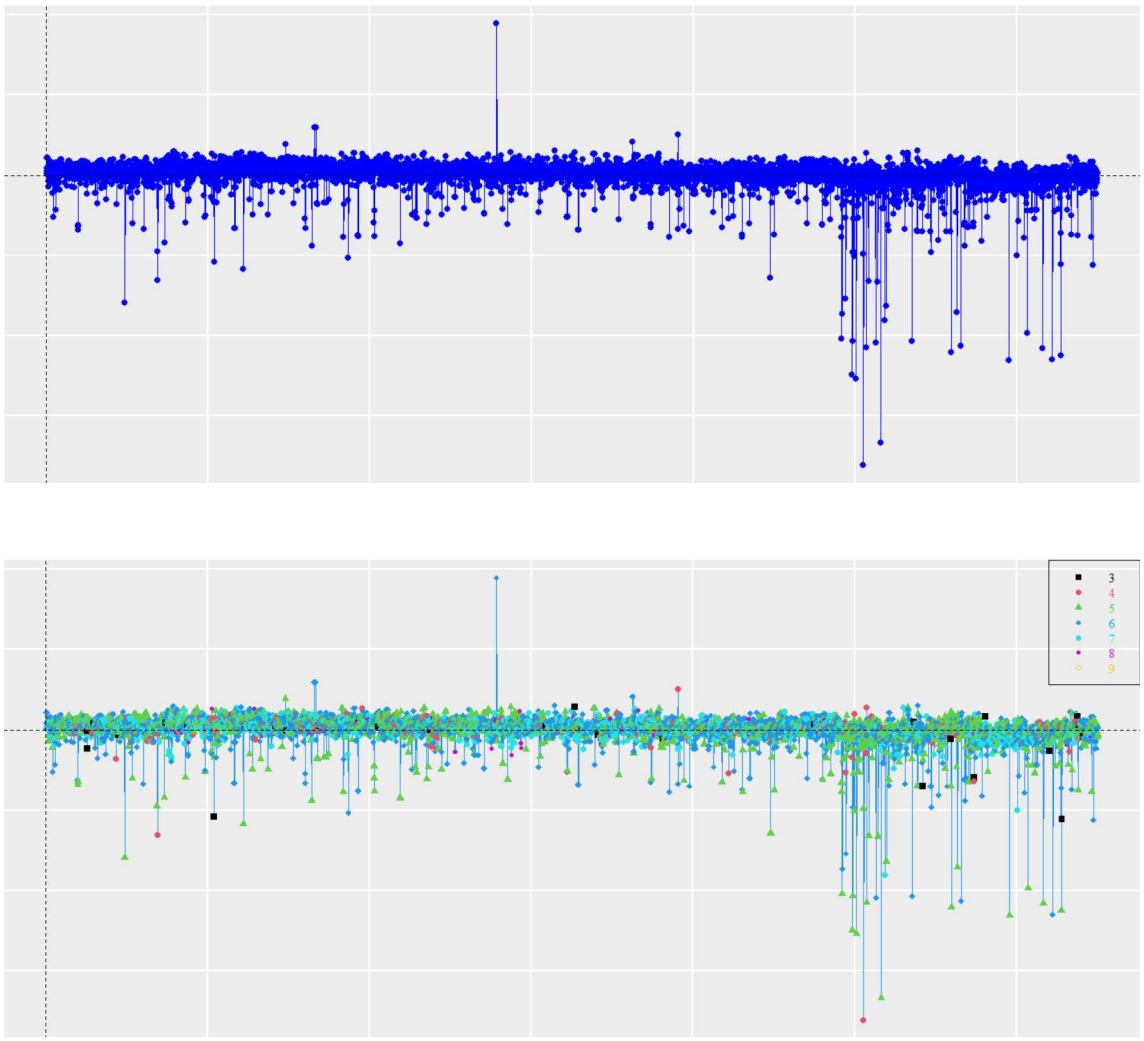

Figura 4.8: Proiezione dei dati in una dimensione attraverso la massimizzazione dell'indice di curtosi con l'assenza e la presenza delle classi.

I vettori di proiezione associati alle variabili che indicano la direzione in cui le variabili contribuiscono maggiormente alla proiezione sono scritti nella tabella sottostante.

Variabile	Asse 1	Asse 2
Fixed Acidity	-0.084	0.173
Volatile Acidity	-0.173	0.230
Acido Citrico	0.242	-0.342
Zucchero Residuo	-0.048	-0.293
Cloruri	0.059	-0.144
Anidride Solforosa Libera	0.074	-0.018
Anidride Solforosa Legata	0.1159	0.507
Densità	0.210	0.267
PH	-0.379	0.308
Solfati	-0.323	0.373
Alcol	0.752	0.300

Tabella 4.4: I vettori di proiezione associati alle variabili per i due assi

Il primo asse è fortemente guidato dall'alcol, indicando che questo contributo ha un impatto significativo sulla direzione principale di variazione nei dati. Ad esempio, attributi proprio come *alcohol* che hanno pesi significativi su entrambi gli assi sono correlati anche alla qualità del vino.
Il valore ottimizzato dell'indice di proiezione risulta essere pari a 74.313 e, nel corso del processo, presenta un andamento che segue la retta stimata $y = 4x - 0.396$.

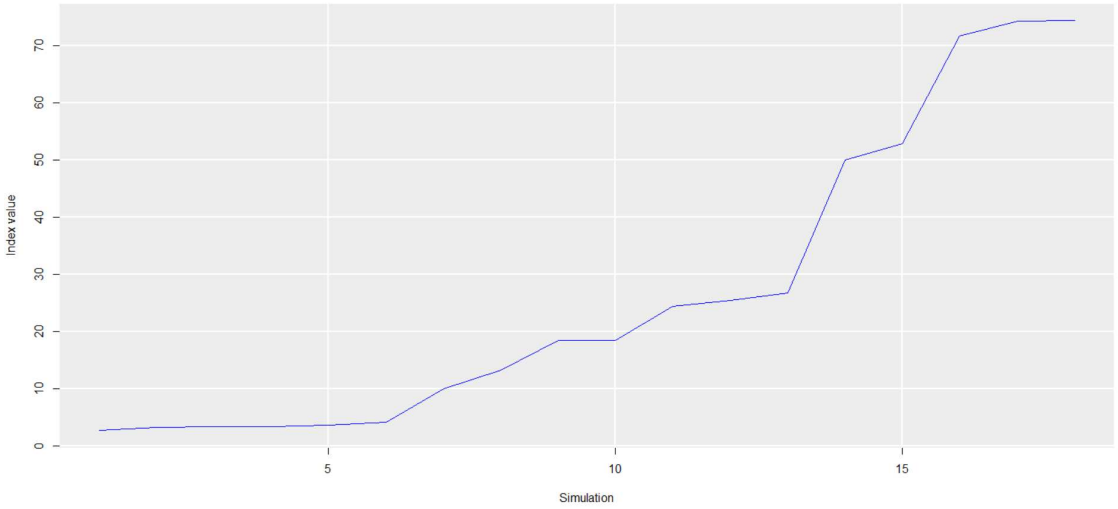

Figura 4.9: Rappresentazione della retta di evoluzione dell'indice di proiezione della curtosi.

- Proiezione in 2 dimensioni attraverso l'indice dell'entropia: Anche nella proiezione in uno spazio a due dimensioni, l'obiettivo è quello di individuare pattern

e strutture nascoste nei dati, per consentire una miglior interpretazione e per giungere a risultati che la PCA non ha evidenziato. L'utilizzo dell'entropia come indice assume un ruolo cruciale nella valutazione della qualità della proiezione ottenuta. L'entropia è una misura di disordine o di incertezza in un sistema, e nello specifico viene impiegata per esaminare quanto bene la proiezione catturi le informazioni rilevanti all'interno dei dati, per esempio la dispersione e la distribuzione dei nuovi punti.
Se tale indice assume un valore basso allora la proiezione ha raggruppato in modo coerente i dati, mentre un'entropia alta può suggerire la presenze della casualità nella distribuzione. La formula matematica scelta è la (3.4), ma si sarebbero potute usare anche altre varianti.

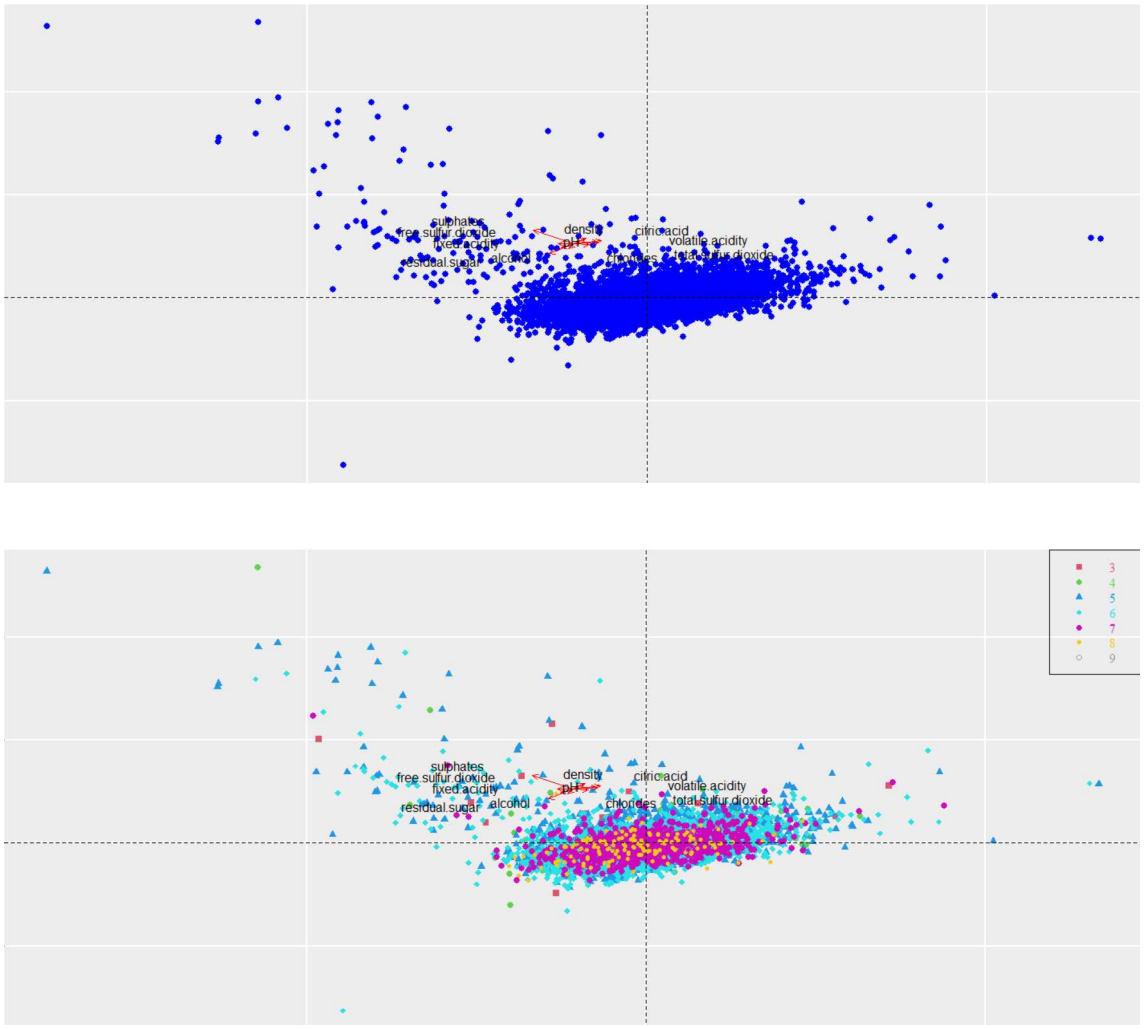

Figura 4.10: Proiezione dei dati in due dimensione attraverso l'indice di entropia con l'assenza e la presenza delle classi.

I vettori di proiezione associati alle variabili che indicano la direzione in cui esse contribuiscono maggiormente alla proiezione sono scritti nella tabella sottostante.

Variabile	Asse 1	Asse 2
Fixed Acidity	-0.237	0.010
Volatile Acidity	-0.411	0.155
Acido Citrico	0.186	0.274
Zucchero Residuo	-0.360	-0.457
Cloruri	-0.190	-0.192
Anidride Solforosa Libera	0.091	0.189
Anidride Solforosa Legata	0.225	-0.031
Densità	0.159	0.289
PH	0.357	0.092
Solfati	-0.598	0.697
Alcol	-0.076	-0.200

Tabella 4.5: I vettori di proiezione associati alle variabili per i due assi.

Le variabili *fixed acidity* e *residual sugar* mostrano contributi negativi sul primo asse, suggerendo che la proiezione è sensibile a valori più bassi di acidità fissa e zucchero residuo. Tuttavia, il contributo del primo è più forte, indicando una maggiore importanza nella direzione opposta. *Residual sugar* e *sulphates* danno un contributo negativo al secondo asse, indicando che la proiezione distingue tra vini con basso contenuto di zucchero residuo e bassa presenza di solfati. L'*alcohol* mostra un contributo negativo sul secondo asse, indicando che vini con un contenuto alcolico più basso possono essere posizionati in quella direzione. *Sulphates*, invece, mostra un forte contributo positivo su quest'asse, suggerendo che la presenza di solfati nel vino può essere un fattore chiave in questa direzione.
Il valore dell'indice dell'entropia parte da 0.289 e giunge al valore ottimo di 0.540. L'andamento della retta viene definita a tratti, in quanto la prima e l'ultima parte presentano una crescita veloce, mentre la parte centrale presenta un aumento moderato.

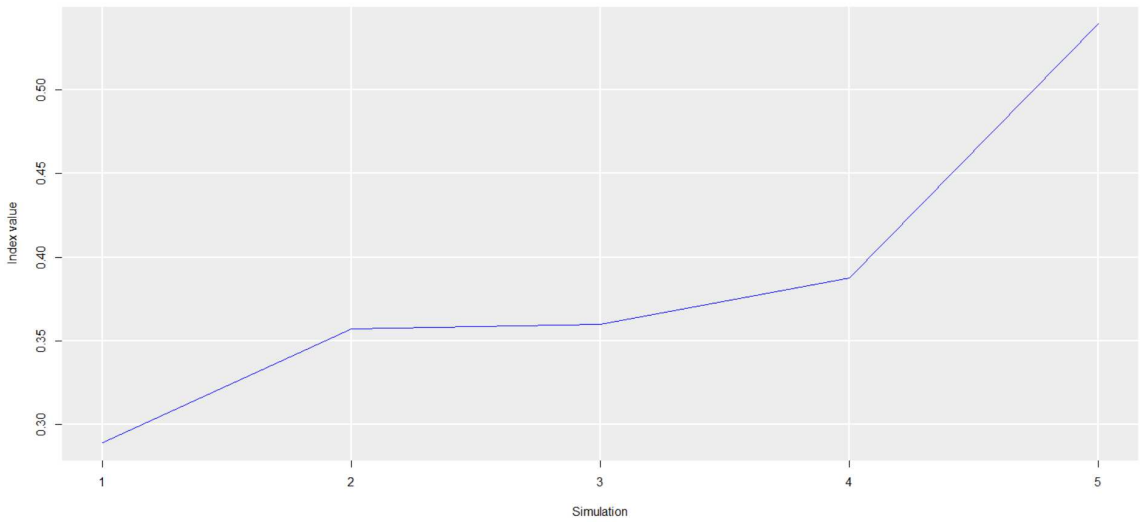

Figura 4.11: Rappresentazione della retta di evoluzione dell'indice di proiezione dell'entropia.

4.5 Regressione con la Projection Pursuit

Nel contesto dell'analisi predittiva, si decide di modellare in modo accurato e flessibile le intricate relazioni tra variabili indipendenti e la variabile risposta. Per affrontare questa sfida, abbiamo scelto di adottare un modello di regressione adatto a strutture non lineari e complesse dato dalle relazioni presenti nei nostri dati. In questa sezione, abbiamo applicato il modello di regressione PPR (Projection Pursuit Regression) al nostro dataset utilizzando la funzione *ppr* in *R*. Questa funzione calcola automaticamente i valori previsti e le misure di bontà di adattamento del modello.

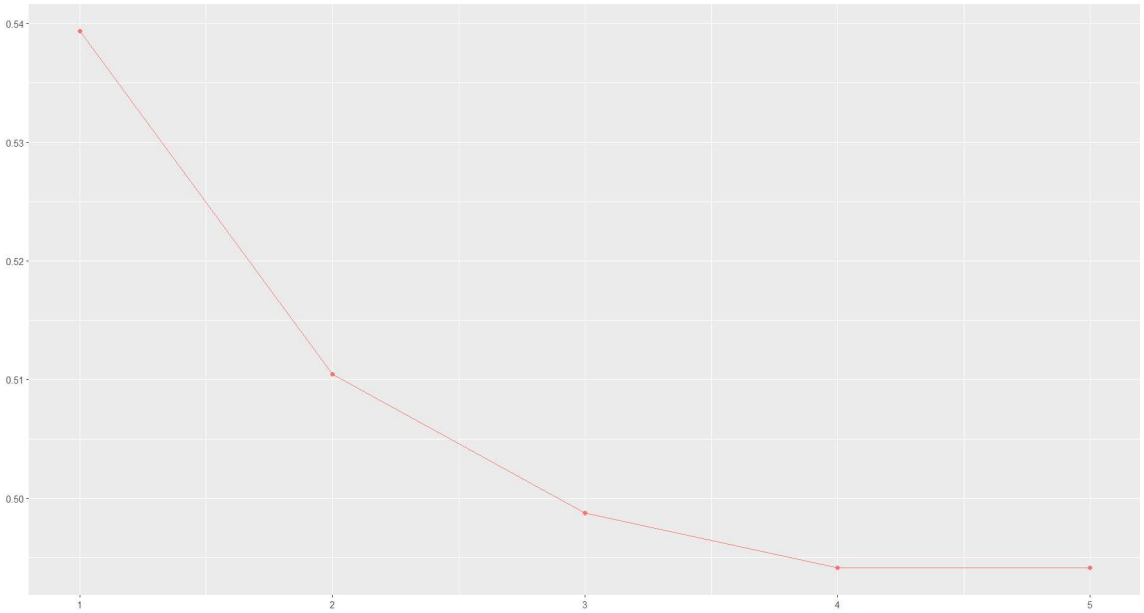

Figura 4.12: Retta del valore assunto dall'EQM rispetto al numero di termini utilizzati nel modello.

La selezione della dimensione d su cui proiettare le unità statistiche è stata effettuata considerando il numero massimo di 5, che rappresenta la metà arrotondata per difetto del numero di variabili iniziali. Il valore dell'errore quadratico medio (EQM) rimane invariato utilizzando 4 o 5 coefficienti. La scelta ponderata tra il principio di parsimonia e il valore di questo indice ci ha condotto a optare per 2 termini. L'utilizzo di tali termini ci consente di formulare il modello nel seguente modo:

$$y = \beta_0 + \beta_1 f_1 + \beta_2 f_2 + \epsilon$$

dove f_1 e f_2 rappresentano i nuovi dati proiettati in uno spazio bidimensionale. Il modello ha un termine di intercetta significativo (p-value prossimo allo zero), indicando anche in questo caso che c'è una relazione significativa tra le variabili indipendenti e la variabile risposta.
I valori dei β ottenuti dall'adattamento del modello sono 0.510 per la prima dimensione e 0.201 per la seconda dimensione, quindi entrambi i valori dei coefficienti hanno un impatto positivo nel contesto, significando che tutte e due stanno contribuendo in modo positivo alla regressione.
Dall'analisi preliminare e dai dati riportati nella tabella (4.5), emerge che la prima

dimensione è principalmente influenzata dalle variabili legate ai solfati, al pH, allo zucchero residuo e all'acidità volatile. L'aumento di queste variabili è associato a un incremento della variabile y nel contesto del modello utilizzato.
Per quanto riguarda il secondo termine, esso sintetizza l'apporto dell'acido citrico, dello zucchero residuo, dei solfati e della densità. Dato che il primo coefficiente è più del doppio del secondo, possiamo affermare che il pH e l'acidità volatile rivestono un ruolo estremamente significativo nella valutazione della qualità del vino, secondo il modello adottato.
L'adattamento del modello può essere effettuato tramite un plot di proiezione, strumento che permette di esplorare visivamente la distribuzione dei dati e la loro rappresentazione nelle prime due dimensioni identificate dal modello. Nello specifico, il plot di proiezione sarà composto da punti, ognuno dei quali rappresenta un'istanza del set di dati. La posizione di ciascun punto nello spazio bidimensionale è determinata dalla proiezione dei dati rispetto ai due termini scelti del modello, con lo scopo di visualizzare cluster, relazioni tra variabili e altre strutture nascoste.

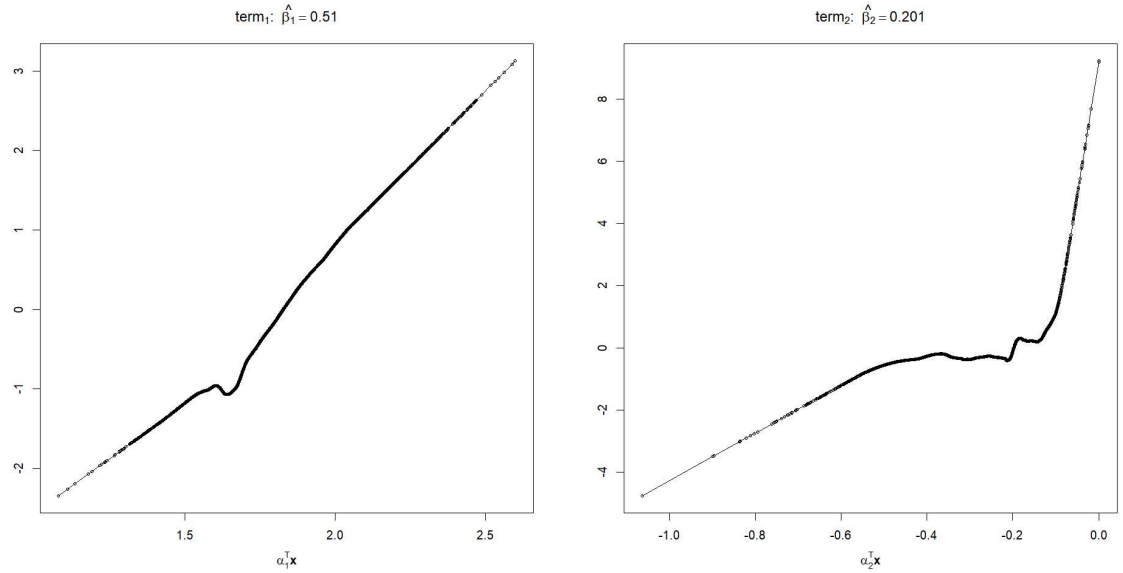

Figura 4.13: Grafico di proiezione del modello scelto con due termini rispetto alla variabile risposta (asse y) e le due dimensioni (asse x).

Nel contesto analitico, osserviamo un trend positivo in entrambe le dimensioni, sebbene i risultati della variabile di risposta siano espressi su scale differenti. La prima dimensione presenta un range di valori compreso tra -2 e +3, mentre la seconda dimensione spazia da un minimo di -4 a un massimo di +8. Per entrambe le variabili, la regione centrale mostra una concavità orientata verso l'alto, con una maggiore accentuazione nella seconda dimensione rispetto che alla prima.

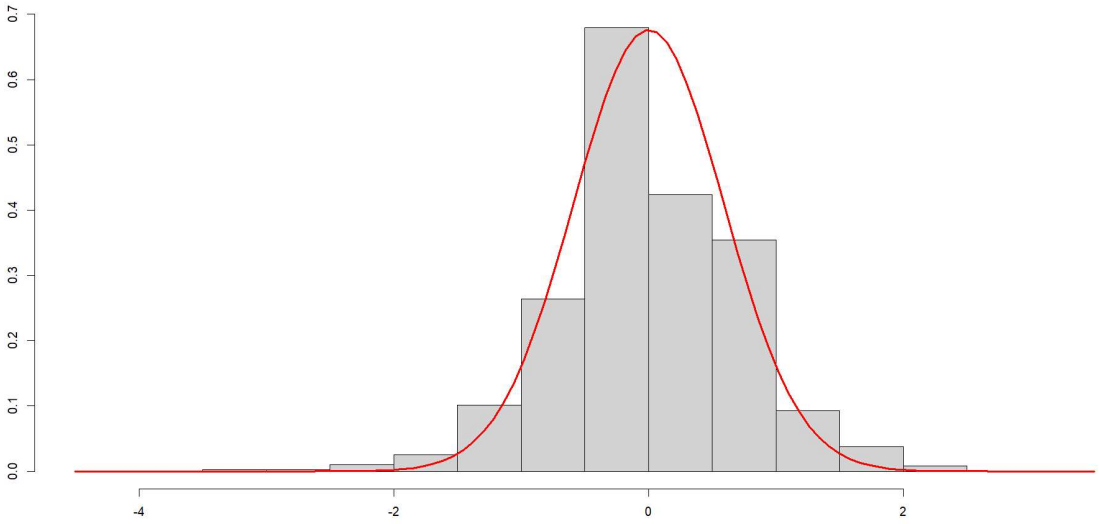

Figura 4.14: Istogramma relativo ai residui confrontato con la densità di una normale (curva rossa).

L'adattamento dei residui a una distribuzione normale è una pratica significativa in quanto permette di valutare quanto bene il nostro modello si adatta ai dati. Nello specifico l'adattamento dei residui, ovvero le differenze tra i valori osservati e quelli predetti dal modello, seguono una distribuzione approssimativamente normale con una media pari a 0 e quindi è possibile affermare che il modello spiega la variabilità in modo bilanciato, senza tendenze simmetriche (Salvan,2020).

Quality	Osservati	PPR
3	30	0
4	216	6
5	2138	2087
6	2836	3771
7	1079	633
8	193	0
9	5	0

Tabella 4.6: Tabella di confronto tra valori osservati di *quality* e valori previsti dal modello

Dalla Tabella 4.6 emerge che il modello produce previsioni insoddisfacenti, assegnando approssimativamente il 58% delle unità al valore 6 e circa il 90% dei vini viene giudicato come medio per quanto riguarda la qualità (*quality* uguale a 5 o a 6). Il modello di PPR prevede male i risultati per diverse ragioni. Innanzitutto, potrebbe essere influenzato da dati non rappresentativi, il che comprometterebbe la capacità del modello di apprendere pattern significativi. Inoltre, una selezione inappropriata del numero di termini potrebbe impedire al modello di catturare correttamente la complessità della variabile *quality* che si sta cercando di modellare e potrebbe contribuire a prestazioni scadenti. Allo stesso tempo, una complessità eccessiva potrebbe portare a fenomeni di overfitting, dove il modello si adatta troppo ai dati e fallisce nella previsione di nuove unità.

Capitolo 5

Conclusioni

Il dataset Vinho Verde contiene le informazioni di 6497 bottiglie di vino, 4898 riguardanti vini bianchi e 1599 rossi. Di queste osservazioni, si sono rilevate diverse caratteristiche relative ad aspetti chimico-fisici, come ad esempio il valore del pH del vino, la percentuale di alcol presente e la quantità di zucchero residuo.
L'obiettivo della tesi è stato di applicare un modello di regressione che riuscisse a prevedere correttamente la risposta *quality*, variabile qualitativa ordinale che assegna un giudizio da 3 a 9 alla bottiglia. Tuttavia, volendo evitare di utilizzare 11 coefficienti nel modello, sono stati applicati due algoritmi che permettono di ridurre la dimensionalità e semplificare il problema, mantenendo tuttavia un'alta percentuale di informazione iniziale. I due algoritmi in questione sono l' Analisi delle componenti principali (PCA) e la Projection Pursuit (PP) e, di conseguenza, i modelli selezionati sono il modello di regressione con le componenti principali (PCR) e quello con la Projection Pursuit (PPR). Il primo metodo semplifica i dati identificando le direzioni lungo le quali la varianza è massima, consentendo una rappresentazione più compatta. D'altra parte, il secondo si concentra su proiezioni complesse dei dati, cercando di rivelare strutture informative al di là delle direzioni di massima varianza, offrendo una visione più dettagliata delle relazioni nei dati.
Alla conclusione dell'applicazione, è emerso che i risultati mostrano delle differenze, poiché l'approccio alla ricerca varia tra i due metodi. La PCA ha mostrato che *free sulfur dioxide* e *total sulfur dioxide* hanno associati i più alti fattori di scala, indicando quindi che queste variabili sono le più importanti per spiegare la variabilità della variabile risposta. Al contrario, la densità del vino risultava la più debole nel captare tale variabilità. Si è deciso di utilizzare un modello di regressione che utilizza 3 componenti principali, arrivando a spiegare una percentuale di *quality* superiore del 60%.
La Projection Pursuit, invece, ha mostrato che le variabili più importanti nella proiezione dei dati in due nuove direzioni sono *fixed acidity*, *residual sugar*, *sulphates* e *pH*. Valori alti o bassi di queste variabili sono significativi nella nuova proiezione delle unità, indicando quindi che sono fattori chiave nell'arrivare a un risultato finale. L'osservazione delle unità in due dimensioni e l'analisi della variabile risposta hanno consentito di sviluppare un modello con due coefficienti.
Dal punto di vista predittivo, entrambi i modelli mostrano risultati insoddisfacenti. La PCR è stata implementata su un set per la stima, con le previsioni calcolate su un set di verifica. Il 75% delle osservazioni viene erroneamente categorizzato in

termini di *quality*, evidenziando che i valori previsti sono principalmente assegnati alle categorie che indicano una qualità mediocre del vino (vale a dire *quality* uguale a 5 e 6).
Il metodo della PPR condivide limitazioni simili al primo modello, ma registra un miglioramento significativo nelle previsioni, assegnando erroneamente solo il 30% delle unità alla categoria sbagliata. Questo progresso è attribuibile all'approccio della PPR nel cercare relazioni non lineari, come curve e superfici, portando a una migliore percentuale di spiegazione della varianza della variabile risposta.
L'esplorazione del dataset attraverso l'utilizzo della PCR e della PPR ha portato alla luce elementi cruciali, fornendo un'analisi approfondita delle interconnessioni tra le variabili considerate. Nonostante questa chiarezza ottenuta, è importante sondare sia le sfide che i punti di forza che si sono manifestati durante l'indagine, al fine di ottenere una comprensione completa della nostra ricerca.
In primo luogo, i limiti dello studio fanno riferimento ad una dimensione ridotta del dataset, in quanto un campione più ampio avrebbe potuto offrire una visione più esaustiva dei modelli. Inoltre, la scelta di utilizzare un numero ridotto di dimensioni nei modelli potrebbe aver limitato la comprensione completa delle dinamiche della variabile risposta. La mancanza di un adeguato controllo sperimentale costituisce un elemento critico di limitazione nello studio dato che viene introdotto un grado di incertezza nei risultati, poiché non si può escludere che le relazioni osservate siano influenzate da variabili estranee non misurate. I futuri sviluppi dovrebbero contemplare l'implementazione di design sperimentali più precisi al fine di confermare le associazioni tra le variabili. Per ultimo, la natura categoriale di *quality* si scontra con l'adattamento di un modello lineare, in quanto i risultati ottenuti nello spazio continuo sono stati approssimati all'intero più vicino.
Tuttavia, lo studio comprende diversi pregi tra cui l'identificazione di pattern significativi non visibili dalla semplice analisi esplorativa e, di conseguenza, ha permesso di approfondire e comprendere delle nuove relazioni complesse presenti nel dataset. L'utilizzo di questi strumenti e la relativa metodologia è facilmente trasferibile ad altri contesti di ricerca, dimostrando che possono contribuire in modo significativo a una vasta gamma di discipline.
In conclusione, nonostante i limiti evidenziati, lo studio ha contribuito significativamente alla comprensione del dataset. Le sfide incontrate forniscono spunti per future ricerche, mentre i pregi sottolineano l'importanza di continuare a esplorare le potenzialità della regressione con le componenti principali e con la Projection Pursuit nell'analisi dei dati.

Bibliografia

Azzalini A.; Scarpa B. (2012). *Data analysis and data mining.* Oxford University Press.

Comissão de Viticultura da Região dos Vinhos Verdes (2023). Vinho verde, like no other wine in the world. *https://www.vinhoverde.pt/en/about-vinho-verde.*

James G.; Witten D.; Hastie T.; Tibshirani R. (2013). *An Introduction to Statistical with Applications in R.* Springer.

Jones, C. M.; Sibson R. (1897). What is projection pursuit? *Journal of the Royal Statistical Society.*

Kuhn M.; Johnson K. (2013). *Applied Predictive Modeling.* Springer, New York.

Salvan A.; Sartori N.; Pace L. (2020). *Modelli Lineari Generalizzati.* Springer.

Shannon, C. E. (1948). *A Mathematical Theory of Communication.* The Bell System Technical Journal.