PROGETTO STATISTICA E ANALISI DEI DATI

Analisi del Disturbo dello Spettro dell'Autismo

Autori
Chiara Puglia 0522501984
Luigi Giacchetti 0522502014
Università degli Studi di Salerno
a.a 2024/2025

1 Introduzione

Il disturbo dello spettro autistico è una condizione neurologica nella quale le persone hanno difficoltà a stabilire relazioni sociali normali, usano il linguaggio in modo anomalo o non parlano affatto e presentano comportamenti limitati e ripetitivi. Inoltre possono presentare anche altri comportamenti come:

- Difficoltà di relazione e comunicazione con gli altri
- Le persone affette da un disturbo dello spettro autistico, inoltre, hanno schemi comportamentali, interessi e/o attività limitati e spesso seguono routine rigide.
- La diagnosi si basa sull'osservazione, segnalazioni di genitori e altre persone e test standardizzati specifici per lo screening dell'autismo.
- La maggior parte delle persone risponde al meglio a interventi comportamentali altamente strutturati.

I tempi di attesa per una diagnosi di ASD sono lunghi e le procedure non sono economicamente efficienti. L'impatto economico dell'autismo e l'aumento del numero di casi di ASD in tutto il mondo evidenziano l'urgente necessità di sviluppare metodi di screening facilmente implementabili ed efficaci. Pertanto, uno screening per l'ASD che sia rapido ed accessibile è necessario per aiutare i professionisti della salute e informare le persone se debbano intraprendere una diagnosi clinica formale.

La rapida crescita del numero di casi di ASD a livello globale richiede la disponibilità di dataset relativi ai tratti comportamentali. Tuttavia, tali dataset sono rari, rendendo difficile condurre analisi approfondite per migliorare l'efficienza, la sensibilità, la specificità e l'accuratezza predittiva del processo di screening per l'ASD. Attualmente, sono disponibili dataset molto limitati sull'autismo associati a diagnosi cliniche o screening, e la maggior parte di essi è di natura genetica.

Di conseguenza, è stato proposto un nuovo dataset relativo allo screening dell'autismo negli adulti che include 20 caratteristiche da utilizzare per ulteriori analisi, in particolare per identificare i tratti autistici più influenti e migliorare la classificazione dei casi di ASD. In questo dataset sono state registrate dieci caratteristiche comportamentali (AQ-10-Child) insieme a dieci caratteristiche individuali che si sono dimostrate efficaci nel distinguere i casi di ASD dai controlli nel campo della scienza del comportamento.

2 Prima Visione del Dataset

La seguente tabella fornisce una descrizione dettagliata degli attributi presenti nel nostro dataset e ad essi è associato il numero corrispettivo di questi ultimi.

Numero attributo	Nome attributo
1-10	Punteggio di screening
11	Età
12	Genere
13	Etnia
14	Ittero
15	Autismo
16	Il paese di residenza
17	Utilizzo dell'applicazione in precedenza
18	Risultato di screening
19	Intervallo di età
20	Relazione
21	Classificazione dell'ASD

Table 1: Lista di attributi del Dataset

Il punteggio di screening, indicati nella "Table 2", da S e seguiti da una numerazione da 1 a 10, rappresentano un punteggio binario che indichiamo con i valori 1 e 0. Se il valore è uguale ad 1 vuol dire che ci stiamo avvicinando alla positività, viceversa per il valore pari a 0.

L'età è una variabile intera che, come si evince dall'altra variabile "Intervallo di età", deve rientrare tra 4 e 11 poiché altrimenti sarebbe da considerare un errore.

Il genere è una variabile di tipo carattere che può assumere i valori "m" per indicare se l'individuo è maschio, mentre "f" per indicare se è femmina.

L'etnia è una stringa che rappresenta il gruppo di provenienza dell'individuo ed è possibile notare che, tavolta contiene degli apici al suo interno e viene segnato come "Others" nel caso in cui l'individuo non faccia parte di una determinata etnia.

L'ittero è una variabile binaria che assume valori "yes" e "no" per rappresentare se l'ittero è presente o meno nell'individuo.

L'autismo è anch'essa una variabile binaria che assume valori "yes" e "no" per rappresentare se uno screening precedente ha rilevato la presenza o meno dell'ASD nell'individuo.

Il paese di residenza è rappresentato da una stringa per indicare dove risiede l'individuo e anche in questo caso è possibile notare che, talvolta, contiene degli apici al suo interno.

L'utilizzo dell'applicazione in precedenza, invece, è una variabile binaria che assume i valori "yes" e "no" per rappresentare se, in screening precedenti, l'individuo ha utilizzato la stessa applicazione di screening o è la prima volta in cui la utilizza.

Il risultato è una variabile intera che oscilla in un intervallo tra 0 e 10 ed è dato dalla somma dei 10 punteggi di screening, nel caso in cui il suo valore superi la soglia, che è rappresentata dal valore 7, vi è una classificazione dell'ASD.

L'intervallo di età è una stringa indicante la fascia d'età che viene espresso in anni in cui il rientra l'individuo.

La relazione è una stringa che indica quale relazione ha l'individuo con il soggetto che ha effettuato lo screening e può presentare "?" nel caso in cui non sia presente questa informazione.

La classificazione dell'ASD è una variabile binaria che assume i valori di "yes" e "no" per rappresentare se il risultato dello screening sia risultato positivo o meno.

2.1 Presentazione dataset

Il dataset in seguito prevede 292 instanze con 21 attributi ciascuno descritti nel paragrafo precedente. L'obiettivo di questo dataset è quello di supportare l'identificazione precoce delle caratteristiche di ASD, permettendo interventi tempestivi per migliorare gli esiti per il bambino.

Fra gli attributi troviamo sia dati demografici quali età e genere, ma anche questionari di valutazione comportamentale ed elementi che possono essere correlati con la presenza di tratti di ASD.

S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	age	gender	ethnicity	jundice	autism	country of res	used app before	result	age desc	relation	ClassASD
1	1	0	0	1	1	0	1	0	0	6	m	Others	no	no	Jordan	no	5	'4-11 years'	Parent	NO
1	1	0	0	1	1	0	1	0	0	6	m	'Middle Eastern '	no	no	Jordan	no	5	'4-11 years'	Parent	NO
1	1	0	0	0	1	1	1	0	0	6	m	?	no	no	Jordan	yes	5	'4-11 years'	?	NO
0	1	0	0	1	1	0	0	0	1	5	f	?	yes	no	Jordan	no	4	'4-11 years'	?	NO
1	1	1	1	1	1	1	1	1	1	5	m	Others	yes	no	'United States'	no	10	'4-11 years'	Parent	YES
0	0	1	0	1	1	0	1	0	1	4	m	?	no	yes	Egypt	no	5	'4-11 years'	?	NO
1	0	1	1	1	1	0	1	0	1	5	m	White-European	no	no	'United Kingdom'	no	7	'4-11 years'	Parent	YES
1	1	1	1	1	1	1	1	0	0	5	f	'Middle Eastern '	no	no	Bahrain	no	8	'4-11 years'	Parent	YES
1	1	1	1	1	1	1	0	0	0	11	f	'Middle Eastern '	no	no	Bahrain	no	7	'4-11 years'	Parent	YES
0	0	1	1	1	0	1	1	0	0	11	f	?	no	yes	Austria	no	5	'4-11 years'	?	NO

Table 2: Prime dieci righe del dataset

2.2 Prime osservazioni e modifiche dataset

In seguito ad uno sguardo dettagliato del dataset si è notato che, nella colonna riguardante la variabile dell'età, sono stati riscontrati dei valori "?" che ha portato come conseguenza alla sostituzione di essi con la media di età maschile poiché ad una presenza di outliers corrisponde la presenza del genere maschile.

Sono state rimosse le variabili "age_desc" e "relation", il primo poiché conteneva informazioni ridondanti in quanto lo screening è mirato ad una fascia d'età tra i 4 e gli 11 anni; mentre il secondo riporta informazioni che non sono ritenute utili al fine dello screening dell'ASD.

Per quanto riguarda le variabili "jundice", "autism", "used app before" e "ClassASD" che presentano come valori "yes" e "no", questi ultimi sono stati sostituiti dai valori 0 ed 1 in cui 0 rappresenta i valori "no" mentre 1 i valori "yes".

Per quanto riguarda la variabile "gender" che presenta come valori "m" e "f", questi ultimi sono stati sostituiti dai valori 0 ed 1 in cui 0 rappresenta i valori "m" mentre 1 i valori "f".

Per quanto riguarda le colonne "ethnicity" e "country_of_res" si è notata la presenza di singoli apici e si è ritenuto necessario la rimozione di questi ultimi per garantire una maggiore chiarezza e omogeneità.

Infine, per quanto riguarda la variabile “ethnicity” sono stati riscontrati dati a “?” e si è scelto di renderli omogenei al valore “Others”, dato che non conoscere l’etnia di un individuo è equivalente a generalizzare la sua stessa provenienza.

2.3 Analisi grafiche

Si è notata una soglia di positività per cui ogni individuo con il valore della variabile “results” maggiore o uguale a 7 ha una classificazione del disturbo dello spettro dell’autismo.

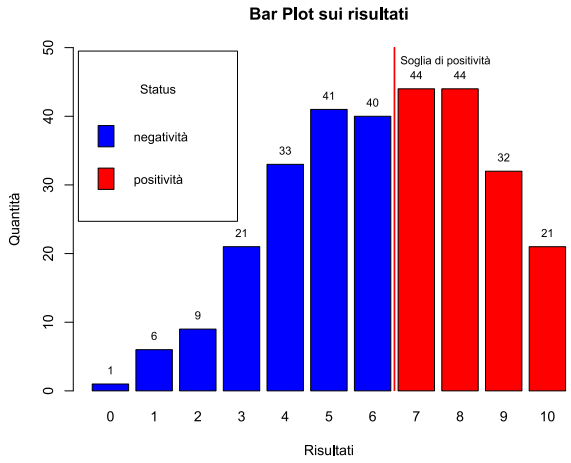
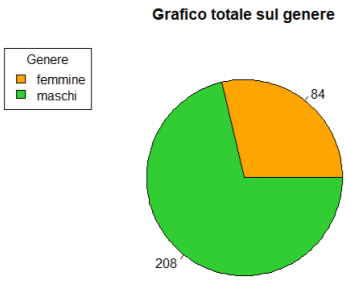
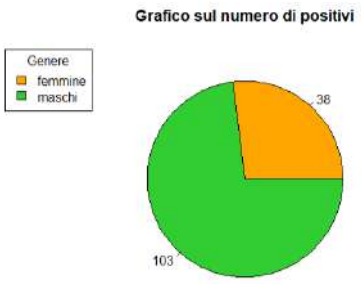


Figure 1: Grafico sulla totalità degli individui rispetto ai punteggi

Tramite i grafici (a) e (b) è possibile dedurre, innanzitutto, che vi è una maggiore casistica di screening per il genere maschile. Inoltre, è possibile notare una leggera maggioranza di classificazioni positivi al ASD per individui di genere maschile.



(a) Grafico a torta sulla totalità degli individui rispetto al genere



(b) Grafico a torta sulla positività al ASD degli individui rispetto al genere

Figure 2: Differenza tra grafico sulla totalità rispetto agli individui positivi al ASD

In relazione a questi due grafici, abbiamo effettuato un ulteriore grafico per mostrare la differenza che vi è tra la percentuale totale, percentuale di positività per l’ASD (Disturbo

dello Spettro dell'Autismo), frequenza relativa totale e frequenza relativa positiva. Di seguito riportiamo il grafico appena descritto:

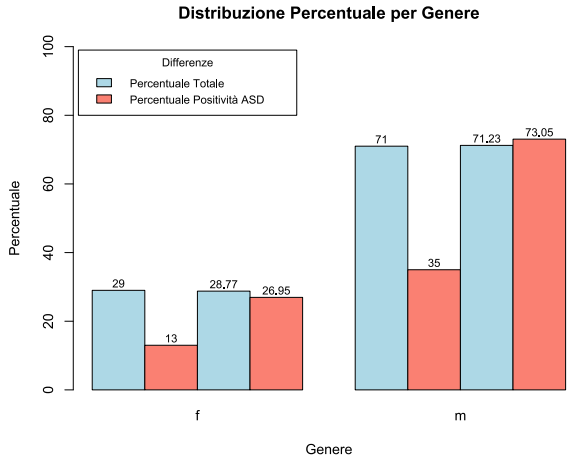


Figure 3: Plot che confronta la percentuale totale, la percentuale di positività e frequenza relativa totale e frequenza relativa positiva

Per quanto riguarda la colonna della variabile età, si è ritenuto opportuno sostituire in quest'ultima i valori uguali a "?" nel caso in cui erano necessarie analisi sulla stessa variabile dell'età anziché su tutte le altre analisi svolte su altre variabili. Siccome la media dell'età maschile risulta essere 6.27451, si è scelto di arrotondare portando così la media a 6 per il grafico dell'età.

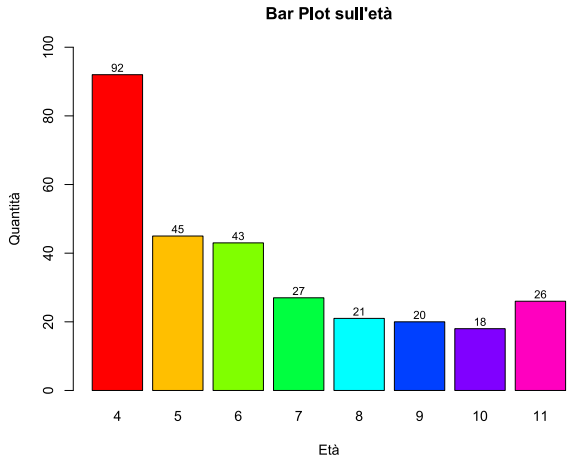


Figure 4: Grafico dell'età effettuato sulla totalità degli individui rispetto ai punteggi

Abbiamo costruito un boxplot in cui mettiamo in correlazione l'età con ClassASD. Notiamo che vi è una maggioranza di screening nell'età infantile poiché vi è assenza del baffo inferiore. Nonostante non vi sia una differenza della mediana tra i negativi e i positivi, è comunque presente un cambio di media minimo (6.181, 6.539) ed inoltre vi è una distanza interquartile maggiore nei positivi.

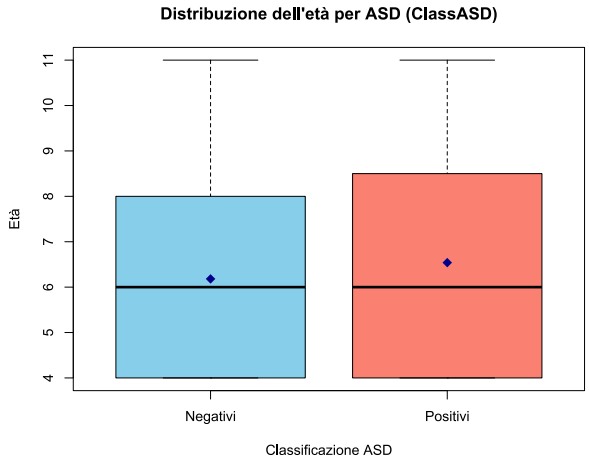


Figure 5: Grafico rappresentante la distribuzione dell'età per individui con ClassASD

Riguardo ai dati sull'etnia si può notare che vi è una maggioranza di bianchi europei seguita da asiatici e medio orientali. Vi è inoltre un'alta presenza di "Others" ovvero di individui la cui etnia non appartiene ai gruppi presenti nel dataset, ed (come si evince dalla fig.6) occupano il secondo posto per quantità.

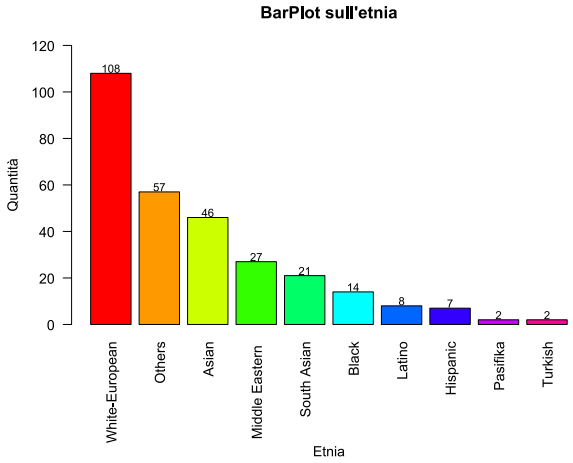


Figure 6: Grafico sull'etnia degli individui

Analizzando le residenze degli individui del dataset si è notato che vi è una maggioranza di individui residenti nel Regno Unito e negli Stati Uniti oltre che in India. In breve quantità anche individui residenti in Australia, Giordania, Nuova Zelanda.

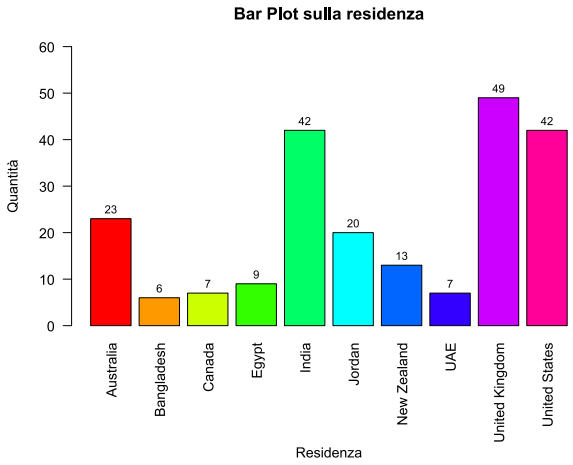


Figure 7: Grafico sulla residenza degli individui

Sono state analizzate le 10 caratteristiche comportamentali presenti nel dataset. Ci si è concentrati, in particolare, su quelle che sono risultate più significative. Tramite un grafico a barre, si è potuto evincere che la terza, la quinta e la decima caratteristica comportamentale sono state quelle che hanno contribuito maggiormente per il campo "results".

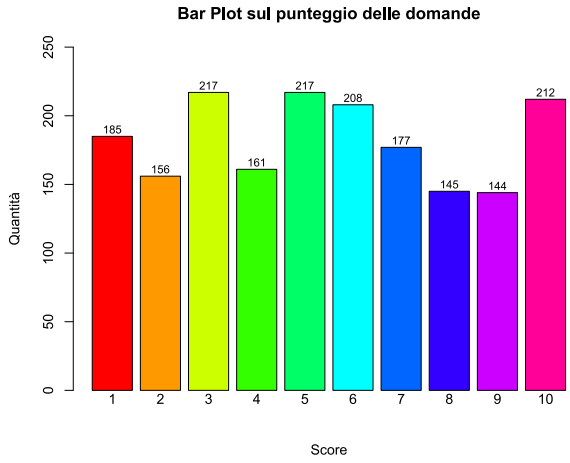


Figure 8: Grafico delle caratteristiche comportamentali che hanno portato ad una classASD

Abbiamo messo in correlazione la colonna "jundice" con la colonna ClassASD. Possiamo notare che gli individui in cui vi è presenza di ittero e che risultano negativi allo spettro dell'autismo, a livello di percentuale essi risultano maggiori rispetto gli individui senza ittero.

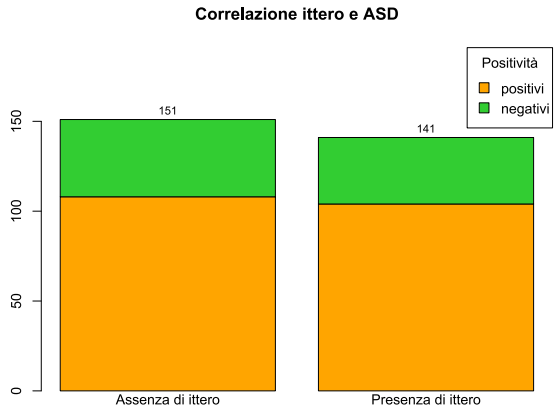


Figure 9: Grafico che mostra la correlazione tra individui con ittero e con ClassASD

3 Domande di ricerca

Sulla base dell'analisi esplorativa effettuata precedentemente, sono state formulate le seguenti domande di ricerca:

RQ1: Quali caratteristiche comportamentali sono maggiormente predittive per una diagnosi accurata della classificazione del disturbo dello spettro dell'autismo (ClassASD)?

RQ2: Come l'analisi della similarità tra cluster, basati su caratteristiche cliniche e comportamentali, consente la distinzione tra pazienti con ClassASD e individui neurotipici?

3.1 RQ1: Osservazioni

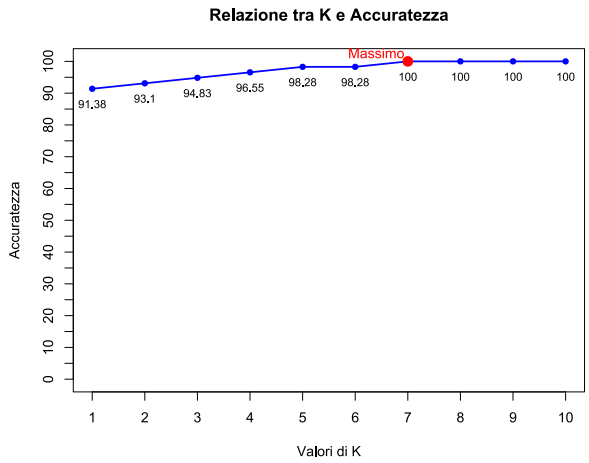


Figure 10: Grafico dell'accuratezza per k rappresentante la grandezza del sottoinsieme

Sulla base del grafico precedente, è stata costruita una tabella che permette di evidenziare la composizione di ciascun sottoinsieme al variare della variabile k che rappresenta la grandezza delle caratteristiche comportamentali (CC nel grafico) all'interno del sottoinsieme. In particolare, è da notare il risultato ottimo globale con k uguale a 7 con CC2, CC3, CC4, CC5, CC8, CC9, CC10 con il picco di accuratezza massima al 100%. Infine, si può notare che la singola caratteristica comportamentale più significativa è CC4 con un'accuratezza massima di 91.37% ed è possibile notare la sua presenza in ogni singolo sottoinsieme.

Osservando il grafico precedente, è possibile notare un plateau al 100% per i valori di k compresi tra 7 e 10.

k	Sottoinsieme ottimale	Accuratezza massima
1	CC4	91.37931
2	CC3, CC4	93.10345
3	CC8, CC9, CC10	94.82759
4	CC1, CC4, CC8, CC10	96.55172
5	CC1, CC4, CC6, CC7, CC10	98.27586
6	CC1, CC3, CC4, CC6, CC7, CC10	98.27586
7	CC2, CC3, CC4, CC5, CC8, CC9, CC10	100
8	CC1, CC2, CC3, CC4, CC5, CC6, CC7, CC8	100
9	CC1, CC3, CC4, CC5, CC6, CC7, CC8, CC9, CC10	100
10	CC1, CC2, CC3, CC4, CC5, CC6, CC7, CC8, CC9, CC10	100

Table 2: Tabella sottoinsiemi per valori di k con relativa accuratezza massima

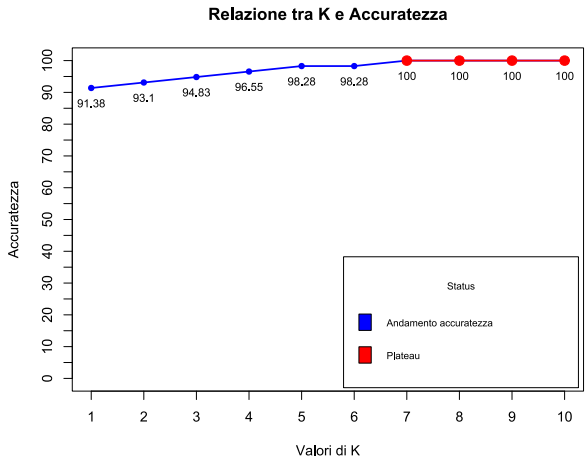


Figure 11: Grafico dell'accuratezza che mostra in rosso l'intervallo del plateau

3.2 Analisi dell'accuratezza: Scelta delle Caratteristiche Comportamentali

Per la selezione delle caratteristiche comportammentali più significative, è stato implementato un processo di selezione delle caratteristiche e di modellazione preditiva in cui è stato utilizzato Random-Forest. Il dataset è stato diviso in 70% training e 30% testing e viene eseguito RFE (Recursive Feature Elimination) per identificare le variabili più importanti tra le 10 caratteristiche comportamentali.

La seguente tabella mostra i risultati ottenuti effettuando la RFE, ovvero le metriche di performance effettuate su ogni singola variabile, in cui RMSE rappresenta l'errore quadratico medio, Rsquared rappresenta il coefficiente di determinazione, MAE rappresenta l'errore assoluto medio e RMSESD, RsquaredSD e MAESD le deviazioni standard delle rispettive metriche.

Variables	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD	Selected
1	1.9149	0.3179	1.5380	0.31121	0.17278	0.24940	
2	1.5727	0.5591	1.2409	0.20976	0.14497	0.19473	
3	1.4445	0.6479	1.1589	0.17418	0.10096	0.16093	
4	1.4162	0.6750	1.1468	0.14580	0.08582	0.12485	
5	1.3274	0.7612	1.0860	0.15085	0.08841	0.13752	
6	1.1221	0.7844	0.8933	0.14070	0.09522	0.11306	
7	0.9938	0.8376	0.7997	0.12379	0.08070	0.11974	
8	0.8820	0.8809	0.7113	0.09411	0.05172	0.08885	
9	0.7407	0.9127	0.6040	0.11060	0.04258	0.09244	
10	0.5970	0.9582	0.4645	0.11337	0.01890	0.07979	

Table 3: Risultati delle metriche di performance

La tabella mostra un miglioramento progressivo delle performance con l'aggiunta di più variabili. Possiamo notare che con 1 variabile, RMSE=1.91, R^2=0.32 mentre con tutte le 10 caratteristiche comportamentali, l'RMSE=0.60, R^2=0.96. L'R^2 di 0.96 con 10 variabili indica un ottimo fit del modello, con un errore quadratico medio (RMSE) basso di 0.60. Le colonne RMSESD, RsquaredSD e MAESD mostrano la deviazione standard delle metriche nella validazione incrociata, indicando la stabilità delle performance. Per quanto riguarda le 5 caratteristiche comportamentali più significative, risultano essere: A3-Score, A4-Score, A6-Score, A8-Score, A9-Score

3.2.1 Lime-Shapley e Random-Forest

Sulla base di quanto descritto nel paragrafo precedente, abbiamo effettuato l'analisi dell'accuratezza con lo scopo di migliorare le prestazioni del modello. Abbiamo utilizzato gli algoritmi lime-shapley e random-forest. Lime-Shapley è stato utilizzato per aiutare a comprendere come una variabile contribuisce alla previsione del modello. Random-forest, invece, è stato utilizzato per la classificazione, ovvero per capire quali variabili hanno maggiore impatto su quest'ultima e per comprendere se l'individuo rientra in questo target (riferito a result).

Sono state utilizzate, a tal proposito, le librerie caret per il machine learning e lime per spiegare la predizione dei modelli complessi tramite modelli interpretabili con lo scopo di facilitare il training e la valutazione dei modelli.

Il dataset viene suddiviso in 70% per il training e 30% per il testing. Di seguito, sono state riportate le immagini che mostrano i risultati ottenuti con accuratezza massima e la relativa matrice di confusione.

Il grafico seguente prende in input, come variabili, gli score A1, A2, A6, A7, A8 con l'aggiunta di Age il cui scopo è di predire il valore di result e di raggiungere l'accuratezza migliore prendendo in considerazione il minor numero delle variabili.

Non sono stati presi in considerazione gli score A3, A4 ed A9 che precedentemente risultavano essere le caratteristiche comportamentali più significative insieme ad A6, A7 ed A8, poichè eseguendo Random-Forest, l'accuratezza del training e del test-set risultava inferiore, non rispecchiando il nostro obittivo.

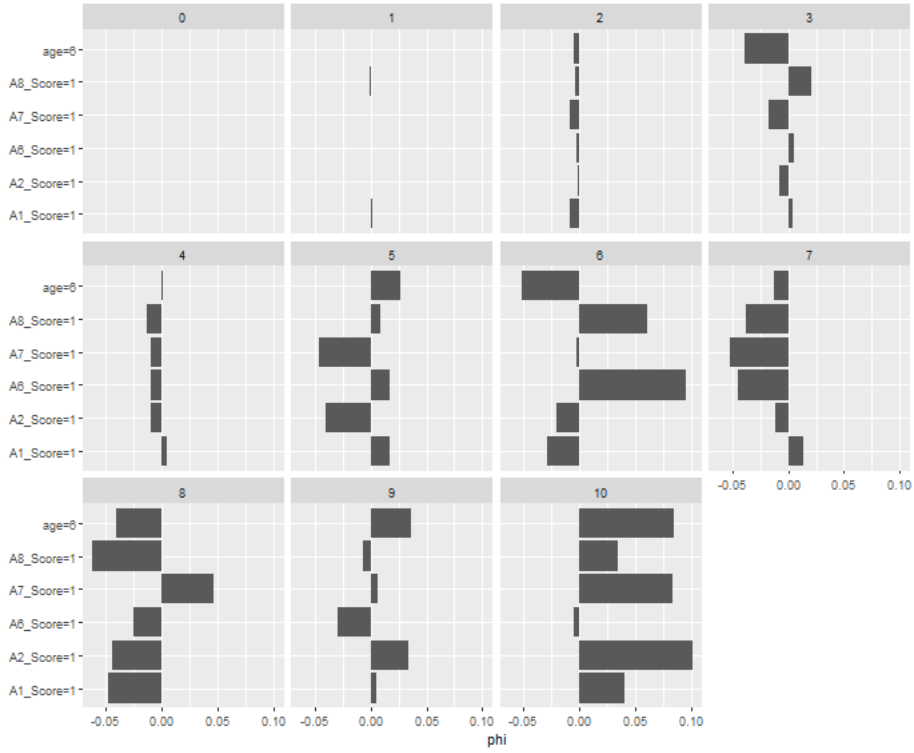


Figure 12: Grafico con gli score A1, A2, A6, A7, A8 e aggiunta di Age

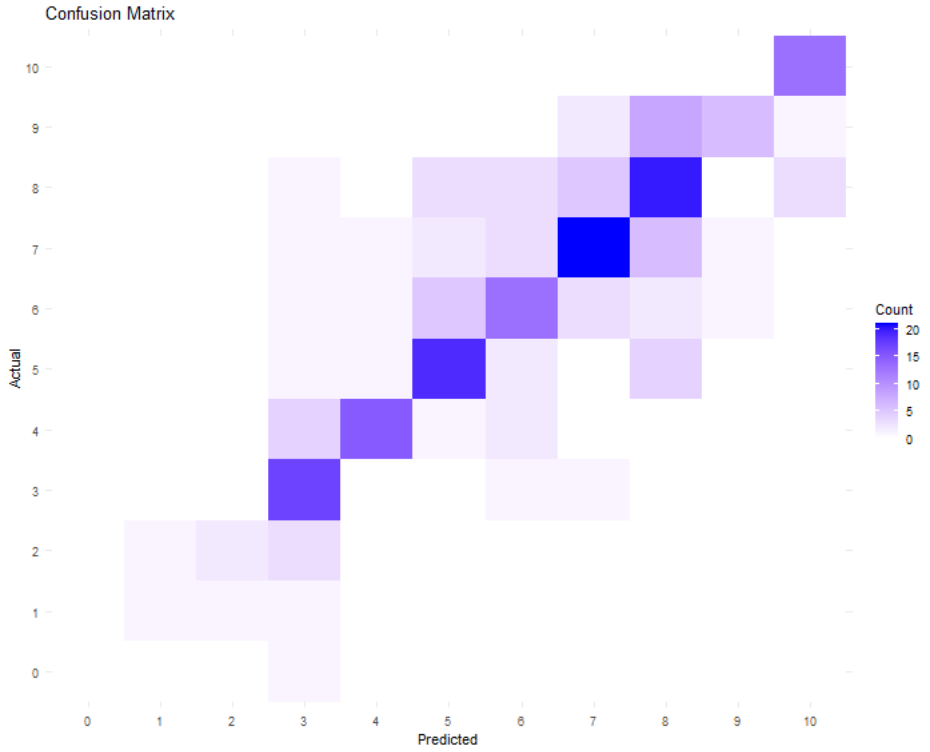


Figure 13: Matrice di Confusione risultante dal training con accuratezza del 63%

Da questo grafico possiamo osservare che è stata ottenuta un'accuratezza del training pari al 63% .

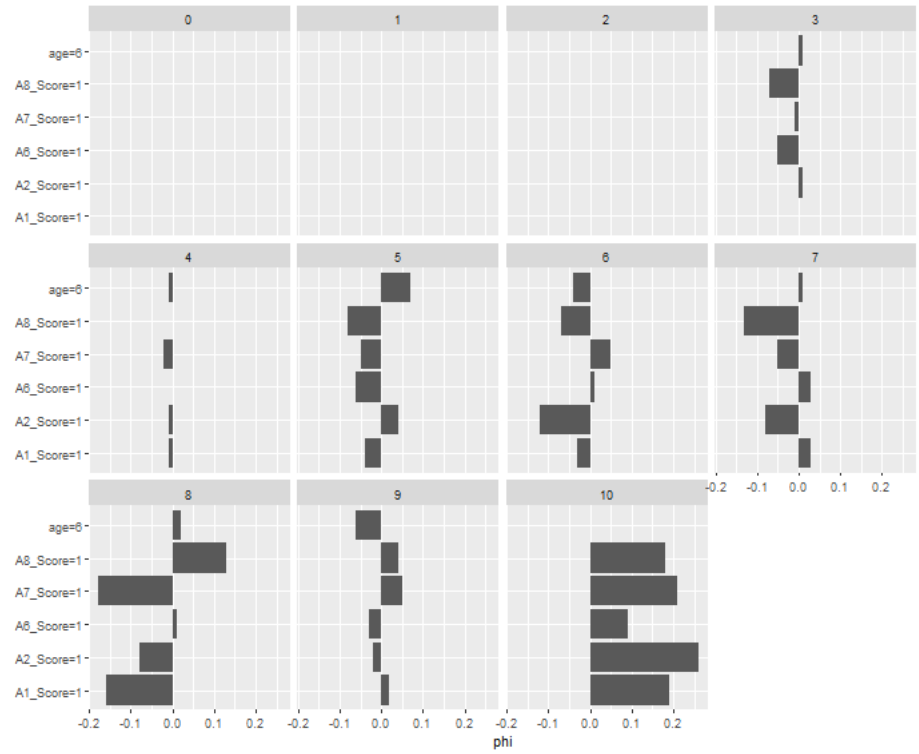


Figure 14: Grafico con gli score A1, A2, A6, A7, A8 e aggiunta di Age

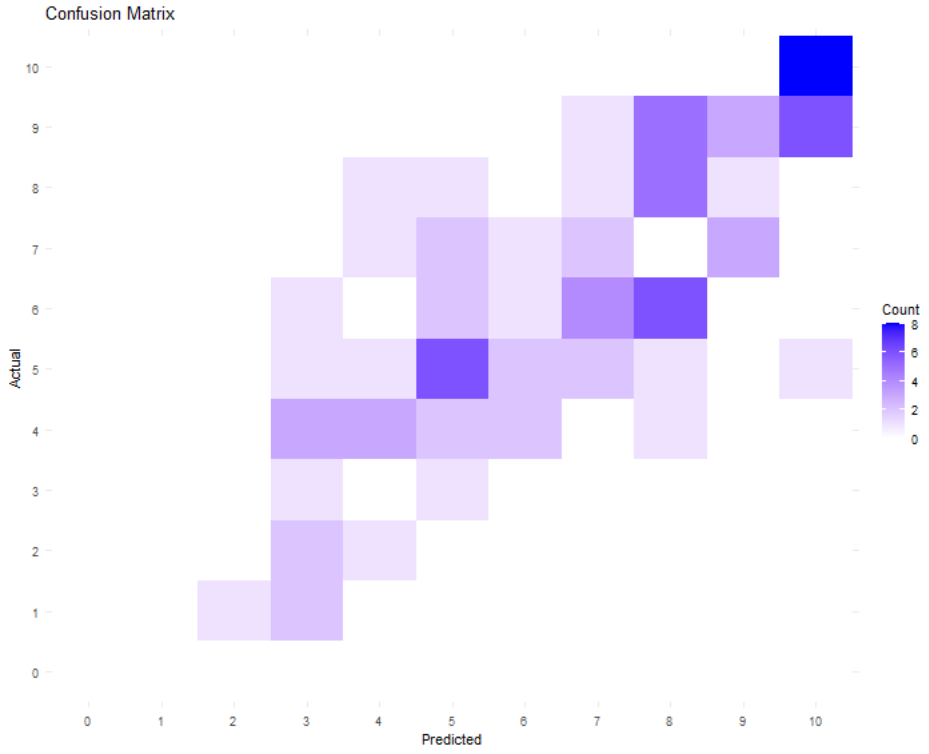


Figure 15: Matrice di confusione risultante dal testing con accuratezza del 34%

Da questo grafico possiamo osservare che è stata ottenuta un'accuratezza per il testing pari al 34%.

L'obiettivo di quest'analisi consiste nel validare la robustezza del modello e vedere quali features sono decisive per rendere il modello robusto. Dai risultati ottenuti possiamo concludere che, il comportamento ottenuto dalla matrice di confusione non rispecchia il comportamento ipotizzato nelle analisi precedenti, poichè risulta che, ad esempio, se il valore attuale è pari 6 (rappresentante gli individui neurotipici), il valore predetto è pari a 7 e 8 che rappresentano gli individui con ClassASD. Di conseguenza Random-Forest non è abbastanza preciso nella predizione dei risultati.

3.2.2 KNN e WKNN

In seguito ai risultati ottenuti dall'analisi precedente, abbiamo utilizzato KNN per fare previsioni sul dataset e per calcolare le metriche di performance come la matrice di confusione, precision, recall, F1-score e accuratezza. Riguardo la scelta delle variabili, le caratteristiche comportamentali che sono risultate più significative per quest'analisi sono: A2, A4, A7, A8 ed A9. La loro importanza è stata determinata dall'utilizzo di due metriche importanti:

- MeanDecreaseAccuracy: misura la riduzione dell'accuratezza del modello quando una variabile viene rimossa. Più alta è questa misura, più la variabile è importante per le predizioni.
- MeanDecreaseGini: indica quanto una variabile contribuisce alla purezza degli split su Random-Forest. Un valore più alto suggerisce una maggiore capacità discriminativa.

In seguito, è stata riportata la tabella rappresentante per ogni caratteristica significativa, i valori rispettivamente di MeanDecreaseAccuracy e MeanDecreaseGini:

Variables	MeanDecreaseAccuracy	MeanDecreaseGini	Selected
A2	0.1080	13.675	
A4	0.0719	13.847	
A7	0.0729	12.687	
A8	0.0835	12.616	
A9	0.0505	13.026	

Table 4: Risultati di MeanDecreaseAccuracy e MeanDecreaseGini

L'algoritmo determina la classe di osservazione in base alla distanza che vi è tra i vicini, che nel nostro caso rappresenta la distanza che intercorre tra ogni singola variabile presa in considerazione. Per ogni punto nei dati di test, il KNN cerca i 7 vicini più prossimi (k = 7) nei dati di addestramento. Il dataset, inoltre, viene normalizzato per farsì che non ci siano variabili che dominino su altre.

Per ogni osservazione nel test set, il modello cerca nei dati di training i 7 punti con la distanza più bassa. L'algoritmo quindi prende le classi di questi 7 punti e assegna la classe

più frequente al punto di test.

Il modello, inoltre, viene addestrato sui dati di addestramento (X-train e y-train) e fa previsioni su questi ultimi, restituendo la classe predetta per ogni punto nel training set. La matrice di confusione che ne risulta ci aiuta a capire quanto bene il modello ha classificato correttamente i dati di addestramento.

Successivamente, KNN fa previsioni sui dati di test (X-test) utilizzando il dataset di addestramento per calcolare la vicinanza. La matrice di confusione risultante ci permette di capire quanto bene il modello ha generalizzato su nuovi dati che non ha visto prima.

Metrica	Precision					
Accuracy	0.0000000	0.0000000	0.0000000	0.5000000	0.3636364	0.2857143
	0.0000000	0.2222222	0.4444444	0.5333333	1.0000000	

Table 5: Risultati dell'Accuracy per la metrica Precision per ogni livello

Metrica	Recall					
Accuracy	0.0000000	0.0000000	0.0000000	0.2000000	0.3333333	0.5000000
	0.0000000	0.1538462	0.3333333	0.4705882	0.5000000	

Table 6: Risultati dell'Accuracy per la metrica Recall per ogni livello

Metrica	F1-Score					
Accuracy	0.0000000	0.0000000	0.0000000	0.2857143	0.3478261	0.3636364
	0.0000000	0.1818182	0.3809524	0.5000000	0.6666667	

Table 7: Risultati dell'Accuracy per la metrica F1-Score per ogni livello

Le tabelle precedenti mostrano i risultati ottenuti dell'accuratezza per ogni livello di Precision, F1-Score e Recall. Per quanto riguarda l'accuratezza ottenuta per il training risulta pari al 44% mentre per il test-set, risulta pari al 32%

A differenza del KNN, l'algoritmo WKNN pondera i vicini in base alla loro distanza rispetto al punto da classificare. In questo caso, per il valore k=6, utilizza i 6 vicini più prossimi per classificare un punto. Come in KNN, viene calcolata la distanza tra i vicini e viene scelto lo schema dei pesi migliore. Successivamente viene effettuata la normalizzazione dei dati per far sì che tutte le variabili siano ridotte nella stessa scala, difatti le features vengono normalizzate tra 0 ed 1. Infine si procede con l'addestramento del modello durante la quale vengono calcolate accuratezza, precisione, recall e F1-Score.

Le seguenti tabelle mostrano l'accuratezza raggiunta per ogni livello di Precisione, Recall e F1-Score, mentre per quanto riguarda l'accuratezza ottenuta per il training risulta pari al 48% mentre per il test-set, risulta pari al 31%

Metrica	Precision					
Accuracy	0.0000000	0.0000000	0.33333333	0.50000000	0.45454545	0.35714286
	0.07142857	0.11111111	0.11111111	0.46666667	0.75000000	

Table 8: Risultati dell'Accuracy per la metrica Precision per ogni livello

Metrica	Recall					
Accuracy	0.0000000	0.0000000	0.3333333	0.3333333	0.6250000	0.3846154
	0.1250000	0.1000000	0.1000000	0.3888889	0.4615385	

Table 9: Risultati dell'Accuracy per la metrica Recall per ogni livello

Metrica	F1-Score					
Accuracy	0.0000000	0.0000000	0.33333333	0.40000000	0.52631579	0.37037037
	0.09090909	0.10526316	0.10526316	0.42424242	0.57142857	

Table 10: Risultati dell'Accuracy per la metrica F1-Score per ogni livello

In conclusione, WKNN migliora KNN dando più peso ai vicini più vicini ma contemporaneamente, prendendo in considerazione l'accuratezza raggiunta sia per il training che per il test set, i risultati ottenuti non rispecchiano il nostro obiettivo. Ciò ha portato a prendere in considerazione un metodo ulteriore.

3.2.3 Artificial Neural Network (ANN)

ANN è una rete neurale utilizzata per la classificazione di un dataset. A differenza del modello WKNN precedente, qui si usa una rete neurale con 3 layer nascosti per effettuare la classificazione. L'obiettivo è quello di predire la classe "result" in base ad un insieme di variabili selezionate (in questo caso sono state selezionate alcune variabili, ovvero A2, A4, A7, A8, A9). Il dataset è stato diviso in 70% training e 30% testing.

Sono stati utilizzati Keras, Caret/Dplyr per la gestione dei dati e TensorFlow come backend. Keras è un formato che si occupa della conversione dei dati in matrici. Viene effettuata la normalizzazione delle feature con scale(), fondamentale per le reti neurali poichè migliora la stabilità numerica ed evita che variabili con scale diverse dominino il modello. Viene effettuata la conversione delle etichette in interi da 0 a N-1 ed applica il one-hot encoding, necessario per la classificazione multi-classe con softmax.

Il modello è una rete feedforward (MLP - Multilayer Perceptron) con tre hidden layers. Ogni strato ha una quantità definita di neuroni (64, 32, 16) e usa la funzione di attivazione ReLU. Lo strato di output utilizza la funzione Softmax, restituendo una distribuzione di probabilità sulle classi. Il modello viene addestrato per 100 epoche con un batch size di 16. Il 20% del training set viene usato come validation set.

Infine, viene calcolata l'accuratezza sul test set, vengono generate le matrici di confusione per il training e il test set per analizzare le prestazioni classe per classe. L'accuratezza ottenuta per il test-set è pari al 33%.

Successivamente, abbiamo eseguito lo stesso codice per 20 volte per verificare se l'accuratezza ottenuta rientrava in un certo intervallo di valori. Per ogni iterazione del codice, i risultati ottenuti sono stati salvati all'interno del file ANN-result.txt. Sulla base di questa precisazione, possiamo concludere che l'accuratezza per il test- set assume valori compresi tra il 32% e il 41%.

I risultati ottenuti sono stati riportati nella seguente tabella:

Metrica	Accuracy					
	0.375	0.3863636	0.4090909	0.3522727	0.375	0.3636364
	0.3409091	0.3636364	0.3636364	0.3636364	0.3295455	0.3409091
	0.3409091	0.3636364	0.3863636	0.3863636	0.375	0.375
	0.375	0.3863636				

Table 11: Accuracy ottenuta utilizzando la rete neurale ANN

Successivamente, è stata effettuata un'analisi in cui sono state prese in considerazione tutte le 10 caratteristiche comportamentali, prendendo sempre in considerazione Artificial Neural Network. Anche in questo caso, Il dataset è stato suddiviso in training e test set, e la rete viene addestrata con un'architettura a un solo strato nascosto contenente 11 neuroni. Successivamente, viene eseguita una validazione incrociata (cross-validation) per determinare l'accuratezza del modello.

Infine, vengono calcolati i valori di Shapley, che permettono di analizzare l'importanza delle variabili predittive nel modello. Tali valori, sono stati stampati e visualizzati all'interno del grafico sottostante:

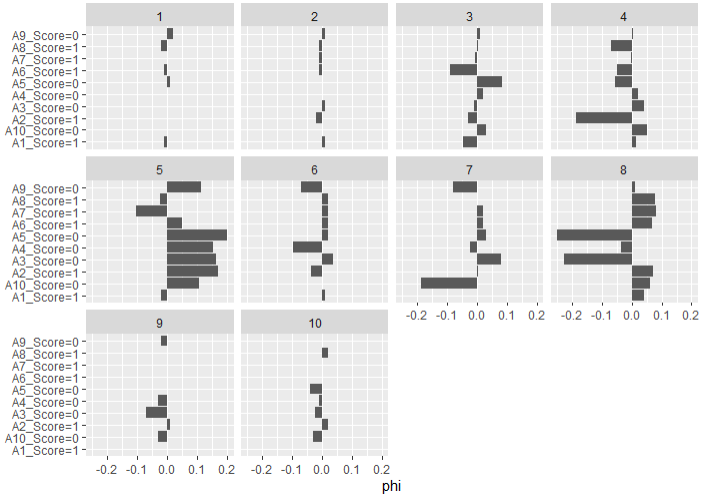


Figure 16: Grafico rappresentante le 5 variabili più significative

Dal grafico precedente, è possibile notare che le caratteristiche comportamentali più significative per quest'analisi sono: A3, A5, A6, A7, A8. Esse hanno portato ad un'accuratezza pari all'84%.

3.2.4 Accuratezza Training per il modello ANN

Il seguente grafico, rappresenta i valori di Shapley per il modello ANN il cui scopo è quello di interpretare l'importanza e l'influenza delle variabili predittive sulle decisioni del modello. Sull'Asse Y vengono rappresentate le variabili di input utilizzate nel modello, denominate A1, A2, ..., A10 e le analizza per capire quale di queste ha avuto un impatto maggiore sulla predizione. Sull'Asse X, invece, vengono rappresentati i valori SHAP, i quali forniscono un contributo ulteriore per ciascuna variabile alla predizione finale del modello.

I Valori positivi suggeriscono che la variabile aumenta la probabilità di una determinata classe mentre i valori negativi indicano che la variabile riduce la probabilità di quella classe.

In seguito a queste considerazioni, è possibile osservare che le variabili A7, A5, A4, A3, A1, A6, A2, A10 sembrano avere un impatto significativo sulle predizioni, poiché hanno

barre più lunghe rispetto alle altre. Le variabili A9 e A8, invece, hanno valori SHAP più piccoli, suggerendo che sono meno influenti nelle decisioni del modello. Inoltre, le variabili A4 e A7 sembrano avere un ruolo chiave con contributi più marcati rispetto alle altre variabili.

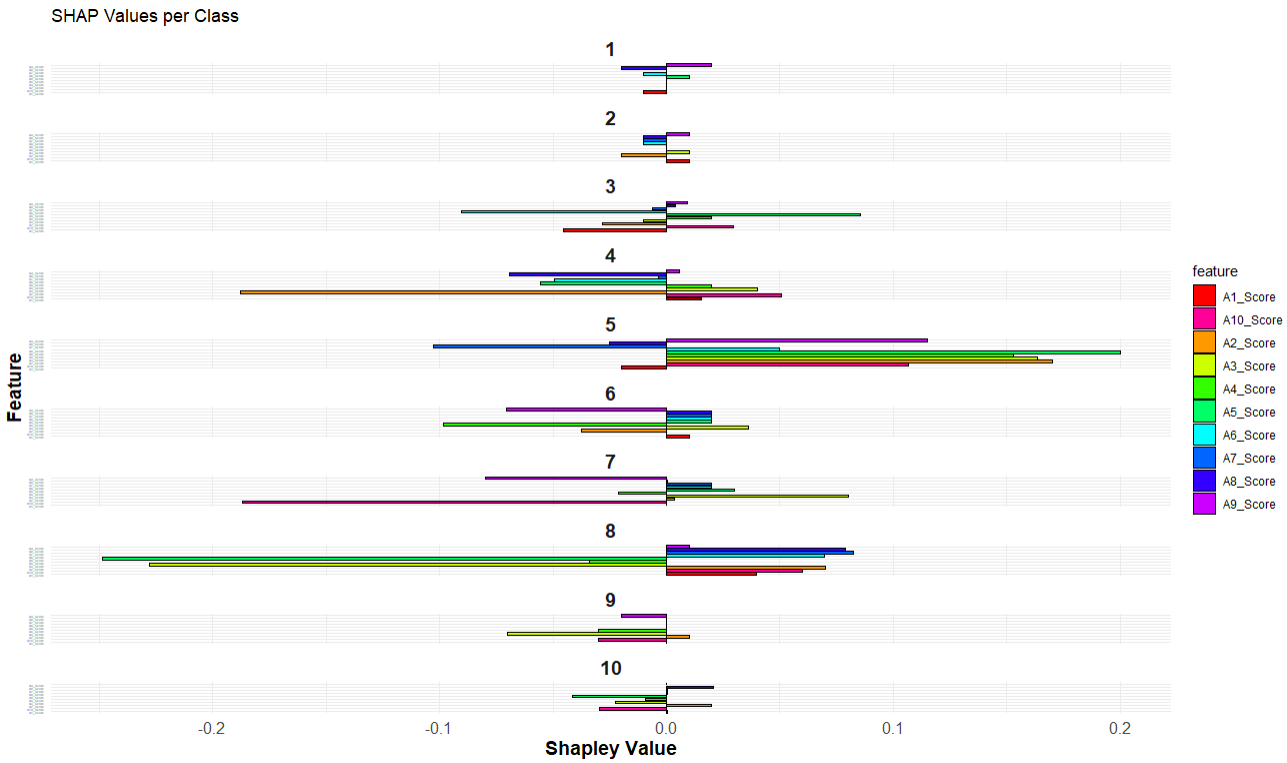


Figure 17: Grafico del training set risultante utilizzando Shapley per ANN

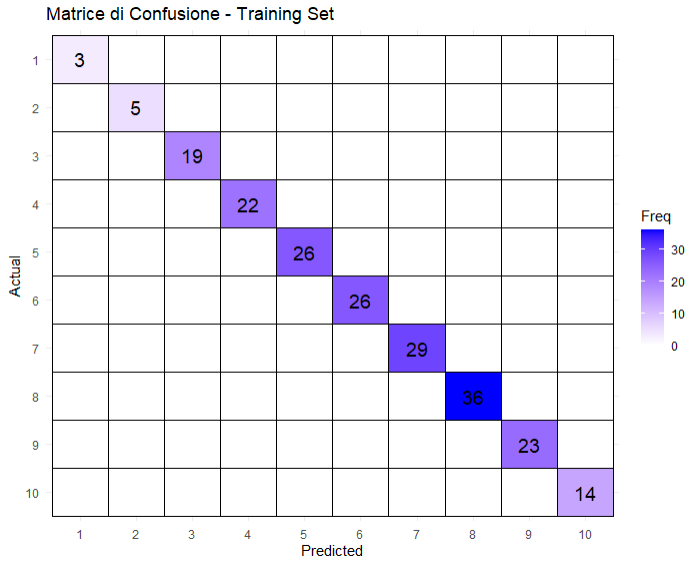


Figure 18: Matrice di confusione del training

Il grafico precedente, mostra la matrice di confusione che visualizza il confronto tra le classi reali (True Class) e le classi predette (Predicted Class). Le righe rappresentano le classi reali mentre le colonne rappresentano le classi predette. I valori nelle celle indicano la percentuale di campioni che appartengono alla classe reale (riga) e che sono stati classificati come una determinata classe (colonna). Per quanto riguarda le celle colorate sulla diagonale rappresentano le previsioni corrette, con valori vicini al 100%, mentre le eventuali celle colorate al di fuori della diagonale (che in questo caso non sono presenti) rappresentano gli errori di classificazione, ovvero le percentuali di campioni classificati in una classe sbagliata.

A tal proposito, sono presenti diverse scale di colori per rappresentare ciò che abbiamo appena descritto. Le celle più chiare indicano valori più bassi mentre le celle più scure indicano valori più alti. Il blu intenso nella classe 8 (con 36 occorrenze) mostra il valore più alto della matrice, indicando che il modello ha classificato correttamente più istanze appartenenti a questa classe rispetto alle altre.

Esaminando la matrice, si può osservare che il modello ha classificato correttamente tutte le istanze perché tutti i valori si trovano sulla diagonale principale. Questo suggerisce che il modello si adatta quasi perfettamente ai dati di training, senza errori evidenti. La classe 8 è quella con il maggior numero di esempi correttamente classificati (36). La classe 1 ha il minor numero di esempi classificati correttamente (3), il che potrebbe indicare una minore rappresentazione di questa classe nei dati di training. Tutte le altre classi hanno una distribuzione bilanciata di esempi correttamente classificati, con valori che variano tra 5 e 29.

In conclusione, le performance del modello risultano abbastanza alte, anche se in alcune classi è presente un numero di occorrenze basse.

3.2.5 Accuratezza test-set per il modello ANN

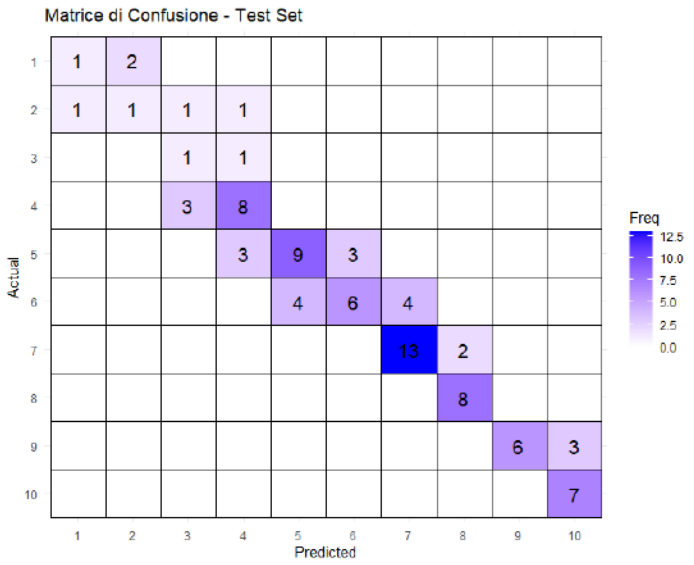


Figure 19: Matrice di confusione risultante dal test set

Il grafico precedente, rappresenta la matrice di confusione del Modello ANN, applicata su un set di test. Questa tabella viene utilizzata per valutare le prestazioni del modello di classificazione su dati mai visti prima. Le righe rappresentano le classi reali (True Class), mentre le colonne rappresentano le classi predette (Predicted Class). I valori nelle celle mostrano la percentuale di campioni appartenenti a una classe reale e assegnati dal modello a una classe predetta.

I colori delle celle rappresentano l'accuratezza. Le celle più scure sulla diagonale rappresentano classi correttamente classificate, mentre le celle colorate presenti al di fuori della diagonale, rappresentano frequenze basse, indicando che poche osservazioni sono state classificate per quella categoria.

Il modello mostra errori di classificazione, segnalando che ha più difficoltà a generalizzare su nuovi dati. Ad esempio, nelle classi 1, 2 e 3 sono presenti pochi esempi e presentano errori evidenti, con molte istanze classificate erroneamente in altre classi. Nella classe 4 ci sono 8 esempi correttamente classificati, ma 3 sono stati predetti come classe 5. La classe 5, invece, ha un numero simile di predizioni corrette ed errate (9 corrette, 3 errate in classe 6 e 3 in classe 4). La classe 7 è quella con il maggior numero di esempi correttamente classificati (13), ma alcune osservazioni sono state erroneamente assegnate alla classe 8. Infine, le classi 9 e 10 mostrano alcune predizioni errate (es. la classe 9 ha 6 esempi corretti, ma 3 erroneamente assegnati alla classe 10).

In conclusione, il modello presenta prestazioni buone nonostante ci siano valori al di fuori della diagonale principale che determinano un calo dell'accuratezza rispetto al training set.

3.2.6 SVM: Support Vector Machines

Dato che i modelli precedenti non hanno prodotto risultati migliori per il calcolo dell'accuratezza, è stato utilizzato SVM, un algoritmo di machine learning per la classificazione e per la regressione. Vengono selezionate una percentuale di dati da utilizzare per l'addestramento mentre la restante percentuale viene utilizzata per il test. Il modello viene addestrato su 10 variabili predittive (da A1-score ad A10-Score) per classificare i dati in base al valore di result. Il training utilizza una tecnica di validazione incrociata a 20 fold, che serve per ottenere una stima più affidabile delle prestazioni del modello. I dati vengono anche centrati e scalati per migliorare la visibilità del modello.

In seguito all'addestramento, viene utilizzata una funzione per la previsione dei nuovi dati utilizzando il modello appena addestrato. Viene estratta l'accuratezza massima ottenuta durante la validazione incrociata, fornendo un'indicazione di quanto bene il modello sta performando.

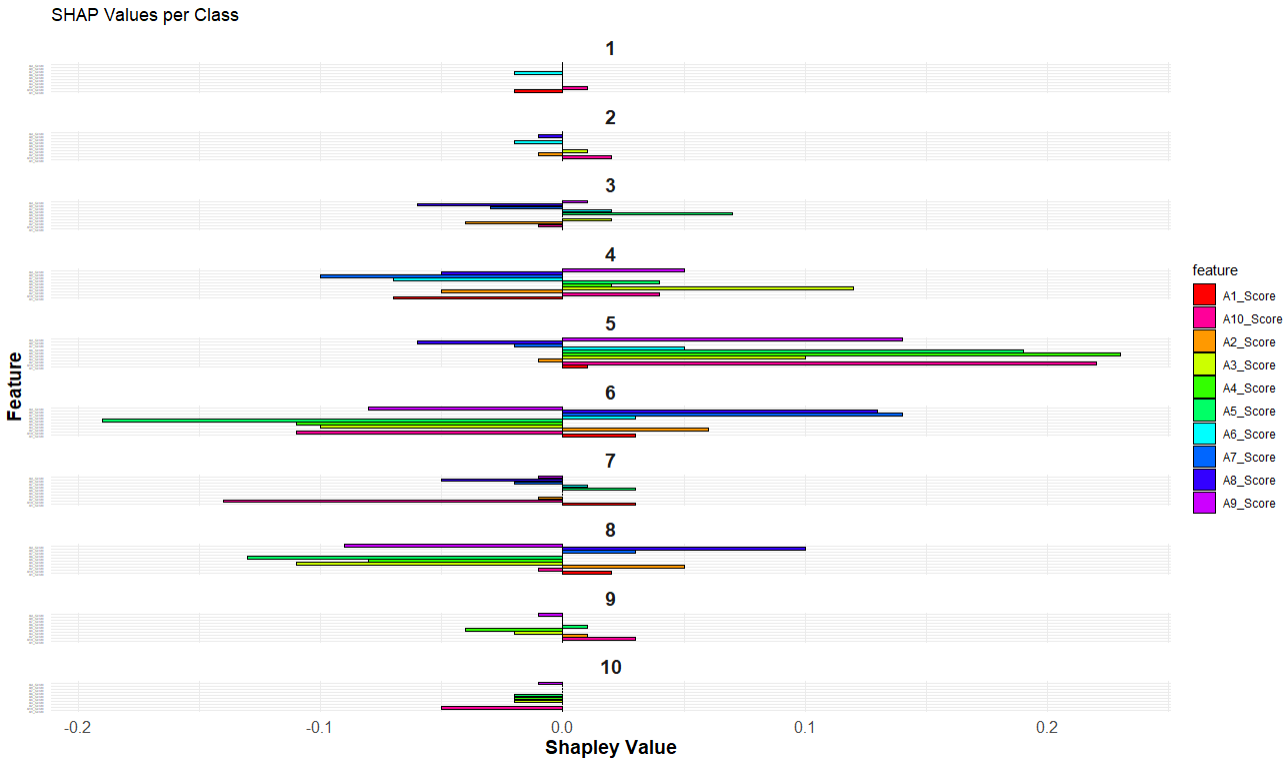


Figure 20: Grafico dei valori di Shapley per SVM

Il grafico precedente mostra i valori di Shapley per diverse variabili predittive per il modello SVM (elencate sull'asse delle ordinate). Sull'asse delle ascisse invece, viene mostrato il valore di Shapley che indica l'importanza e l'impatto che questo valore (positivo o negativo) può avere nel risultato del modello. In questo caso, A5, A3 e A6 sembrano avere un impatto più forte, mentre A10 è quella che presenta un impatto minore. Tuttavia, A1, A6 e A8 sembrano avere maggiore impatto vicino la soglia di positività per una migliore classificazione del disturbo dello spettro dell'autismo.

3.2.7 Accuratezza Training set per il modello SVM

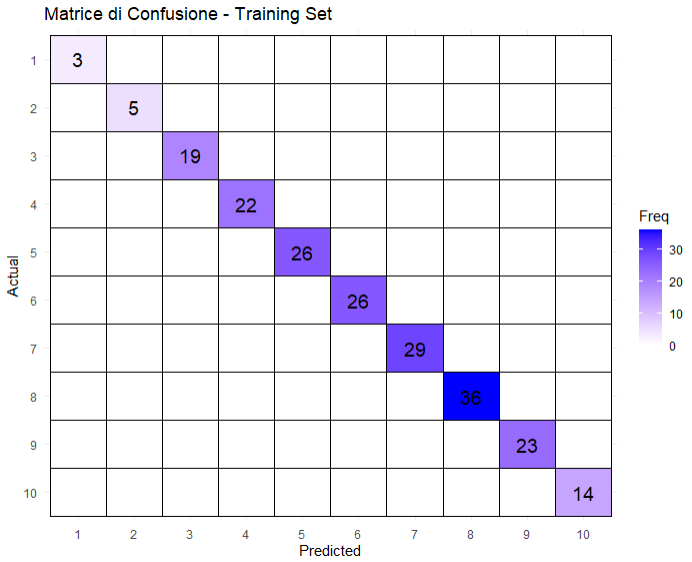


Figure 21: Grafico della Matrice di confusione del training SVM

La figura precedente, rappresenta la matrice di confusione che consente di valutare le prestazioni del modello. Le righe rappresentano le classi reali (effettive) delle osservazioni mentre le colonne rappresentano le classi predette dal modello. La diagonale principale contiene i valori delle predizioni corrette, ovvero i casi in cui il modello ha classificato correttamente un'osservazione. I valori fuori dalla diagonale rappresentano gli errori di classificazione, ovvero i casi in cui il modello ha previsto una classe diversa da quella effettiva.

Inoltre, sono presenti diverse sfumature di colore che rappresentano il grado di classificazione. Le celle più chiare indicano una bassa frequenza di osservazioni classificate in quella categoria, mentre le celle più scure indicano una frequenza più alta, ossia il modello ha classificato molte osservazioni in quella categoria. In questo caso, La cella più scura è nella classe 8 (36 osservazioni classificate correttamente), che rappresenta il valore più alto nella matrice.

Questa matrice suggerisce che il modello sta performando molto bene sul training set, con quasi tutte le osservazioni classificate correttamente. La maggior parte dei valori si trovano sulla diagonale principale, indicando un'alta accuratezza sul training set. Non sono presenti errori di classificazione significativi, dato che non sono visibili numeri al di fuori della diagonale. Alcune classi hanno meno campioni rispetto ad altre (es. classe 1 ha solo 3 esempi, classe 2 ha 5 esempi).

3.2.8 Accuratezza Test-set per il modello SVM

Il grafico seguente, rappresenta la matrice di confusione per il modello SVM che consente di analizzare le prestazioni del modello sui dati di test, ovvero su dati che il modello non ha mai visto prima.

Anche in questo caso, Le righe indicano le classi effettive (verità a terra) mentre le colonne indicano le classi predette dal modello. I numeri all'interno delle celle rappresentano la frequenza delle predizioni per ogni combinazione di classe reale e classe predetta. La diagonale principale rappresenta le predizioni corrette, ovvero i casi in cui la classe prevista corrisponde alla classe effettiva. Le celle fuori dalla diagonale evidenziano errori di classificazione.

Per quanto riguarda la classificazione degli errori, vengono utilizzate diverse sfumature di colore. Colori più chiari indicano un numero basso di osservazioni classificate in quella categoria mentre colori più scuri indicano un numero maggiore di osservazioni classificate in quella categoria. Possiamo ossevare in questo caso che la cella più scura si trova nella classe 5 (15 osservazioni classificate correttamente), il che suggerisce che questa classe ha avuto il miglior riconoscimento da parte del modello.

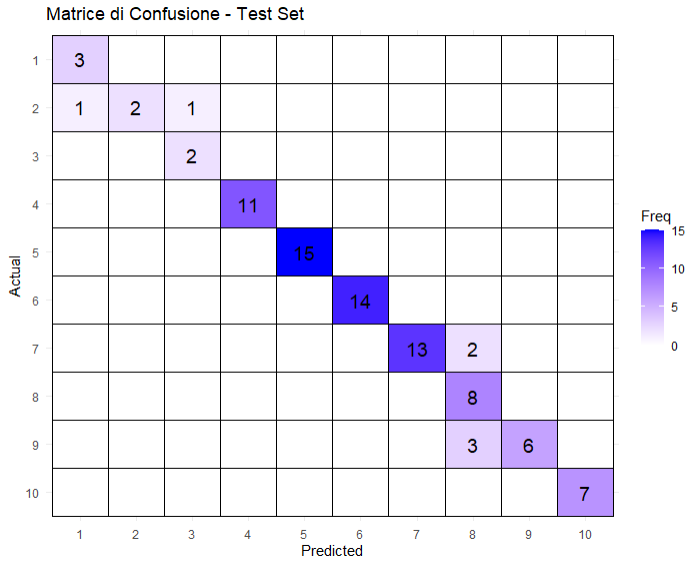


Figure 22: Grafico della Matrice di confusione del test-set per SVM

Il modello SVM ha ottenuto un discreto risultato, poichè la maggior parte delle osservazioni si trovano sulla diagonale principale, indicando che il modello ha fatto diverse classificazioni corrette. Le classi che hanno un buon numero di predizioni corrette sono: 4, 5, 6, 7, 8 e 10, ma la classe 5 ha la performance migliore (15 classificazioni corrette), seguita dalla classe 6 (14 classificazioni corrette) e dalla classe 7 (13 corrette).

3.2.9 Considerazioni finali

In seguito alle analisi effettuate, possiamo concludere che il modello SVM e la rete neurale ANN hanno prodotto risultati migliori, sia nella predizione dei risultati con la matrice di confusione, che per il calcolo dell'accuratezza, poichè per SVM abbiamo ottenuto un'accuratezza pari al 96.4% mentre per ANN abbiamo ottenuto un'accuratezza pari all'84.14%.

3.3 RQ2: Osservazioni sul Clustering

In questa sezione, andremo ad esaminare come i pazienti con ClassASD si distinguano dagli individui neurotipici. Abbiamo utilizzato come algoritmo il k-means con $k = 2$, il cui scopo è quello di identificare due gruppi di individui (rappresentati da punti) con caratteristiche simili e analizzare le differenze tra di essi.

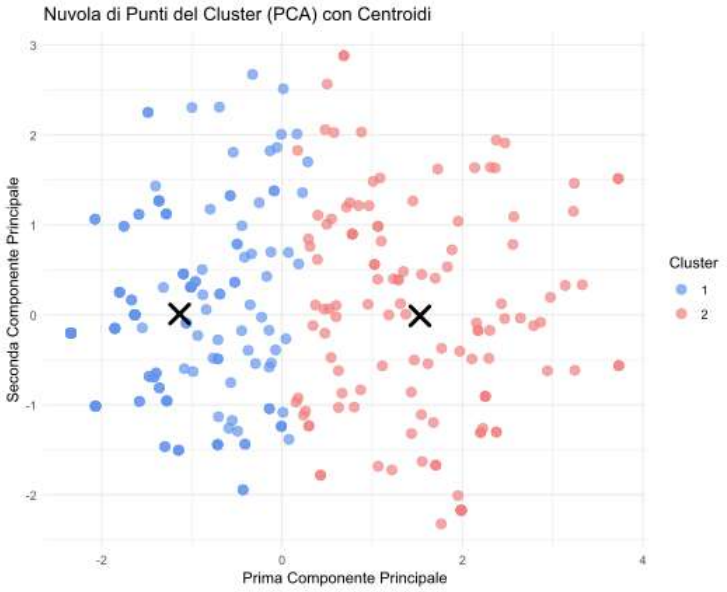


Figure 23: Grafico rappresentante la nuvola dei punti di Cluster 1 e Cluster 2

Il grafico precedente mostra una rappresentazione bidimensionale del clustering ottenuta tramite K-means con i dati proiettati sulle prime due componenti principali di una PCA (Principal Component Analysis). L'asse orizzontale rappresenta la Prima Componente Principale, mentre l'asse verticale indica la Seconda Componente Principale.

I punti nel grafico sono suddivisi in due cluster distinti, colorati in blu e rosso, identificati nella legenda sulla destra. Al centro di ciascun cluster è presente un centroide, rappresentato dalla "X" nera. Questo punto rappresenta il centro medio dei dati appartenenti al cluster corrispondente. I cluster sono chiaramente separati lungo la prima componente principale, suggerendo che questa dimensione è la più rilevante per la differenziazione tra i gruppi.

In conclusione, il metodo K-means ha suddiviso i dati in due gruppi ben distinti. Dato che i centroidi risultano ben separati, suggerisce che i cluster sono ben definiti nella trasformazione PCA.

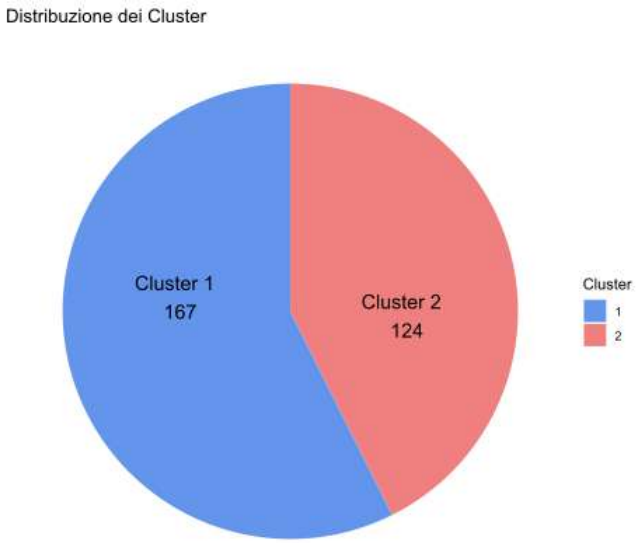


Figure 24: Grafico rappresentante gli individui (somma dei punti) con ClassASD e individui Neutotipici

Il seguente grafico rappresenta una tabella di contingenza che mostra la distribuzione della variabile Result in relazione a due cluster distinti. La colorazione della heatmap riflette le frequenze dei dati e la suddivisione per ClassASD. Sull'Asse Y (Cluster) vengono rappresentati due gruppi distinti, indicati come Cluster 1 e Cluster 2, mentre sull'Asse X (Result) viene rappresentato il valore della variabile "Result", che varia da 1 a 10.

Per la rappresentazione della frequenza dei dati, sono state fornite diverse sfumature di colore. Il colore blu associa la classe 0 di ClassASD mentre il rosso associa la classe 1 di ClassASD. Le sfumature più scure indicano una maggiore frequenza di dati in quella cella, quelle più chiare invece indicano una minore frequenza di dati.

All'interno di ogni cella della tabella sono riportati i valori numerici, che indicano la quantità di elementi in quel determinato incrocio di Cluster e Result. Sul lato destro della heatmap, ci sono due scale di riferimento. La frequenza, che è rappresentata da diverse tonalità di grigio, che indicano il numero di dati presenti in ogni cella e ClassASD rappresentata da due colori (blu e rosso), che indicano le due classi della variabile "ClassASD" (0 e 1).

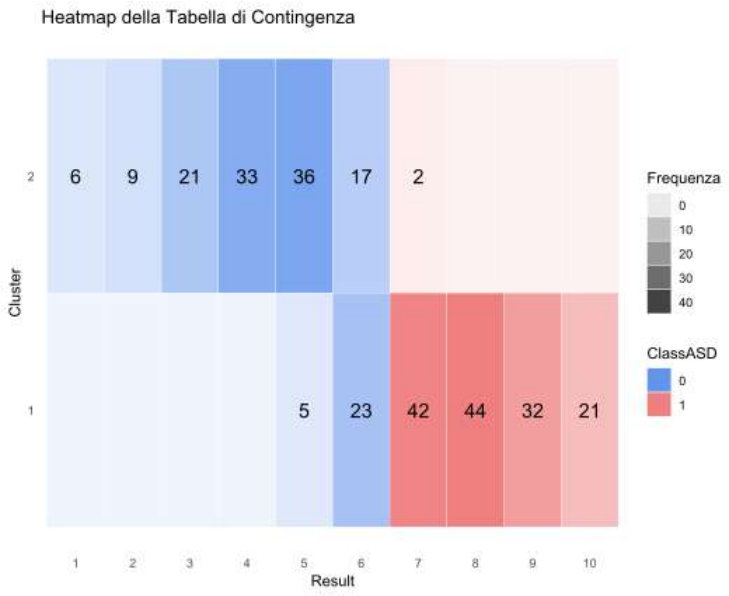


Figure 25: Confronto tra ClassASD e i due Cluster

In conclusione, la tabella di contingenza fornisce una visione chiara della distribuzione dei dati tra i due cluster e della relazione con la variabile ClassASD. Il fatto che il Cluster 1 sia prevalentemente associato alla classe 1 e il Cluster 2 alla classe 0 indica una separazione piuttosto netta tra i due gruppi.

Successivamente a questa tabella, è stato effettuato il calcolo del Silhouette Score, ovvero della distanza tra centroidi utilizzata per valutare la qualità della separazione dei cluster. Tale valore è stato calcolato sia per il Cluster 1 che per il Cluster 2. Per il Cluster 1, abbiamo un Silhouette Score pari a 0.41 mentre per il Cluster 2 il valore del Silhouette score è pari a 0.43. Effettuando la media tra i due, abbiamo ottenuto un valore pari a 0.42. Questo valore dimostra che il Cluster 1 e Cluster 2 sono abbastanza distinti, anche se sono presenti alcune zone in cui i due Cluster si sovrappongono.

3.4 Clustering Gerarchico

In base a quanto detto nel paragrafo precedente, viene effettuata un'analisi di clustering gerarchico per identificare pattern nei dati comportamentali. Partendo da 10 variabili (da A1 ad A10-Score), standardizza i dati e calcola le distanze tra le osservazioni. Tali osservazioni vengono raggruppate in 2 cluster principali, visualizzando la struttura attraverso dendrogrammi.

Per ogni cluster, viene creata una visualizzazione dettagliata che mostra come le osservazioni si raggruppano, con distanze rappresentate sull'asse verticale. Sono presenti anche funzionalità per contare le osservazioni e organizzare le etichette in modo ordinato. L'obiettivo è quello di scoprire strutture naturali nei dati comportamentali e visualizzare come le diverse osservazioni si relazionano tra loro in termini di similarità.

Nella seguente figura, viene mostrato il dendrogramma per i due cluster generati:

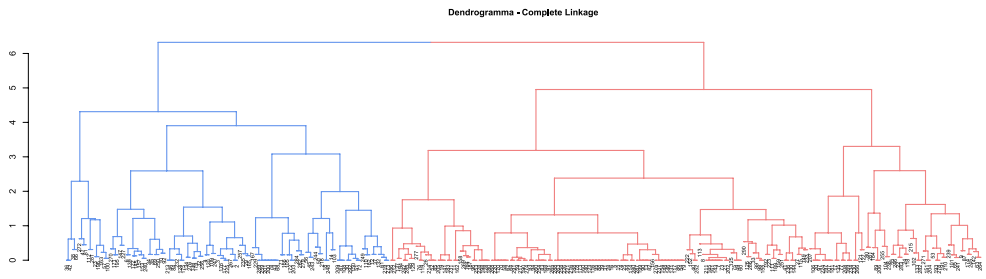


Figure 26: Dendrogramma K-means Cluster1 e Cluster2 ClassASD e neurotipici

3.4.1 Clustering con PCA

Il seguente grafico, fornisce il risultato di un'analisi di clustering combinata con la PCA. L'obiettivo è quello di rappresentare la suddivisione dei dati in due cluster distinti, evidenziando le loro distribuzioni nello spazio bidimensionale definito dalle due principali componenti principali, PC1 (sull'asse X) e PC2 (sull'asse Y).

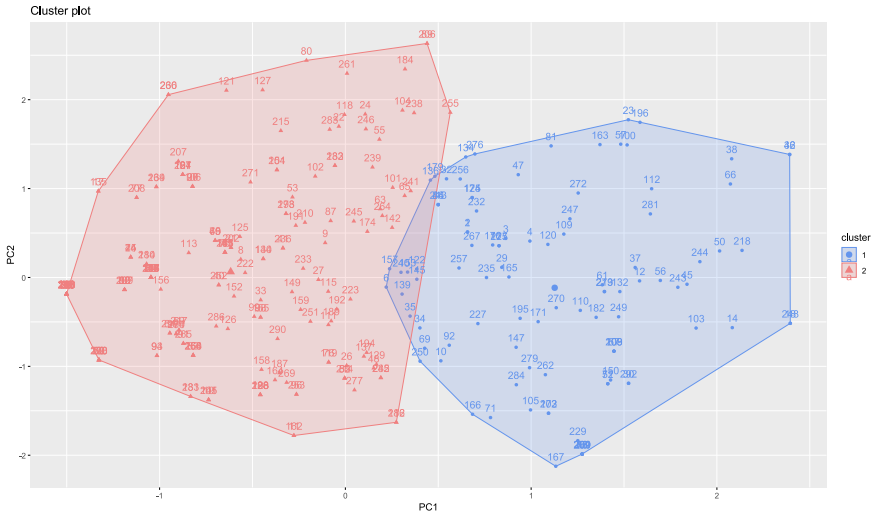


Figure 27: Clustering sia con centroidi che con PCA

Ogni punto nel grafico rappresenta un'osservazione del dataset e viene etichettato con un numero identificativo. Intorno ai punti appartenenti ai due cluster, sono state disegnate delle aree poligonali che delimitano lo spazio occupato da ciascun cluster, aiutando così a

visualizzare meglio la separazione tra i due gruppi.

Dal punto di vista della distribuzione, il cluster rosso occupa principalmente la parte sinistra del grafico, mentre il cluster blu si trova maggiormente a destra. Tuttavia, si nota una zona centrale in cui i due cluster si avvicinano, indicando una possibile sovrapposizione tra i due gruppi.

Il clustering è stato effettuato utilizzando il metodo K-Means basato sui centroidi e successivamente visualizzato nello spazio delle componenti principali, derivato tramite PCA. Questo approccio è utile per comprendere meglio la separazione tra i cluster e identificare eventuali pattern nei dati.

Nel complesso, il grafico mostra come i dati sono stati suddivisi in due cluster distinti, evidenziando una chiara separazione tra di essi nello spazio bidimensionale delle componenti principali. L'uso di PCA ha permesso di ridurre la dimensionalità dei dati, facilitando così la loro rappresentazione visiva.

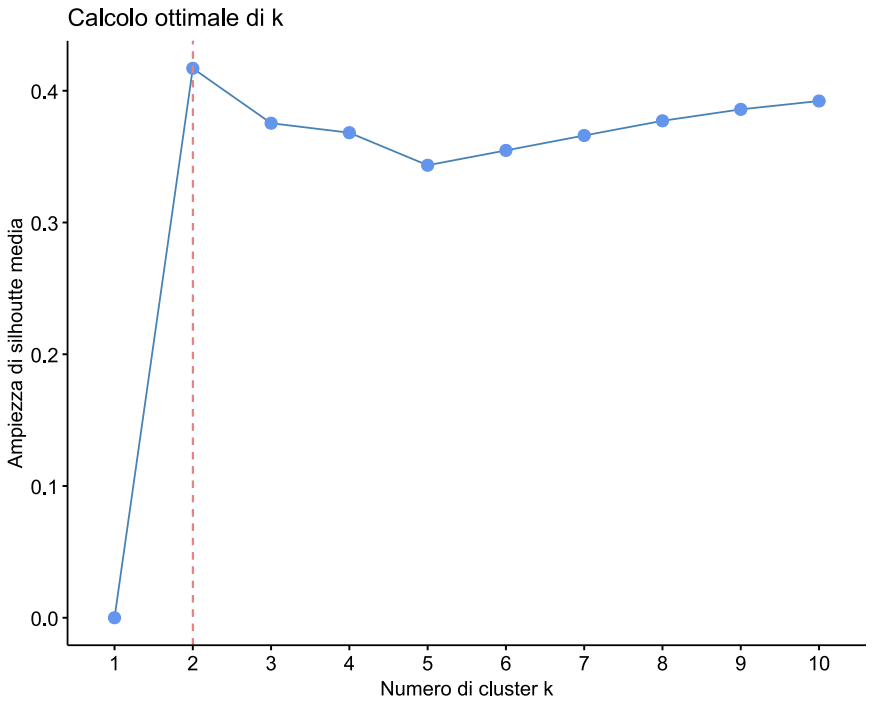


Figure 28: Grafico per il calcolo del valore ottimale di k

Il grafico precedente, mostra il calcolo ottimale del numero di cluster k basato sull'ampiezza della silhouette media. L'asse orizzontale rappresenta il numero di cluster k, mentre l'asse verticale indica il valore medio della silhouette, una misura di coesione e separabilità dei

cluster. Il grafico mostra una curva che cresce rapidamente fino a un massimo intorno a $k = 2$, per poi decrescere leggermente e stabilizzarsi per valori maggiori di k.

La presenza della linea tratteggiata verticale su $k = 2$ suggerisce che questo sia il valore ottimale per il numero di cluster, poiché rappresenta il punto in cui l'ampiezza della silhouette è massima. Questo rafforza la validità del clustering visto nell'immagine precedente, dove i dati sono stati separati in due regioni distinte dopo l'applicazione della PCA.

Il seguente grafico, rappresenta la distribuzione o ampiezza del Silhouette Score per ciascun punto all'interno dei cluster, fornendo un'indicazione di quanto bene ogni elemento è assegnato al proprio cluster rispetto agli altri. Il suo scopo è quello di valutare la qualità del clustering.

Il grafico è diviso in due sezioni, ciascuna colorata in modo diverso (blu e rosso), corrispondenti ai due cluster identificati. L'asse verticale rappresenta il valore del silhouette score, che varia tra -1 e 1: valori vicini a 1 indicano che il punto è ben assegnato al proprio cluster, valori vicini a 0 suggeriscono che il punto si trova al confine tra due cluster, mentre valori negativi indicano che probabilmente il punto è stato assegnato in modo errato.

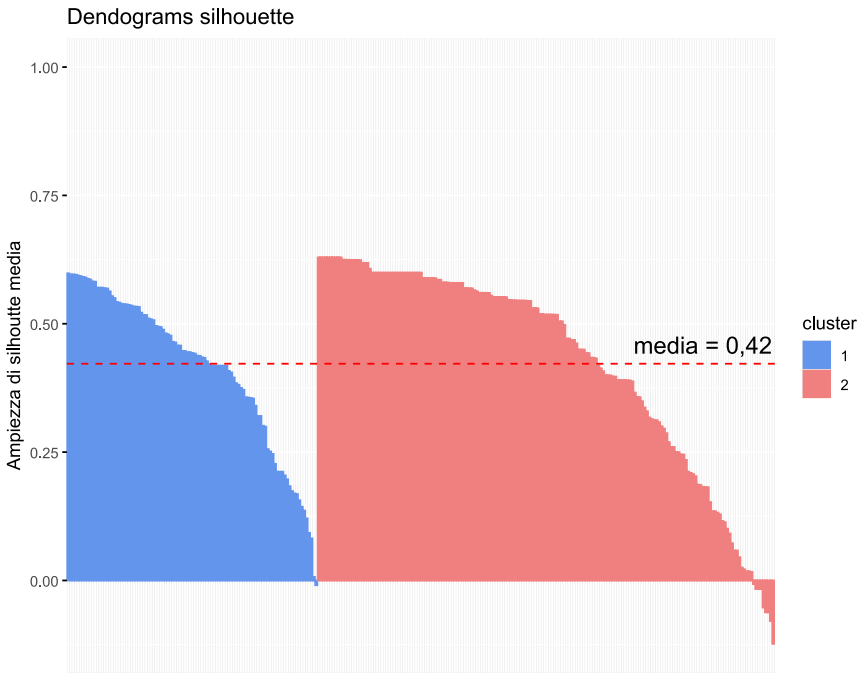


Figure 29: Media del silhouette score tra i due Cluster

La linea tratteggiata orizzontale evidenzia il valore medio della silhouette, pari a 0,42, un valore discreto che indica una buona coesione all'interno dei cluster, anche se con una certa variabilità. Il cluster colorato in blu mostra una distribuzione più compatta, con valori di silhouette generalmente elevati, mentre il cluster in rosso presenta maggiore dispersione e alcuni punti con valori di silhouette bassi o negativi.

Possiamo concludere che il valore medio di silhouette score conferma che la scelta di $k = 2$ cluster è appropriato, poiché il valore di 0,42 è abbastanza alto da giustificare la separazione in due gruppi.

Sulla base del calcolo del valore ottimale di k, è stata realizzata una heatmap con lo scopo di mostrare la distribuzione della frequenza dei dati in funzione dei due cluster ottenuti dal clustering gerarchico e della classificazione binaria ClassASD (0 e 1).

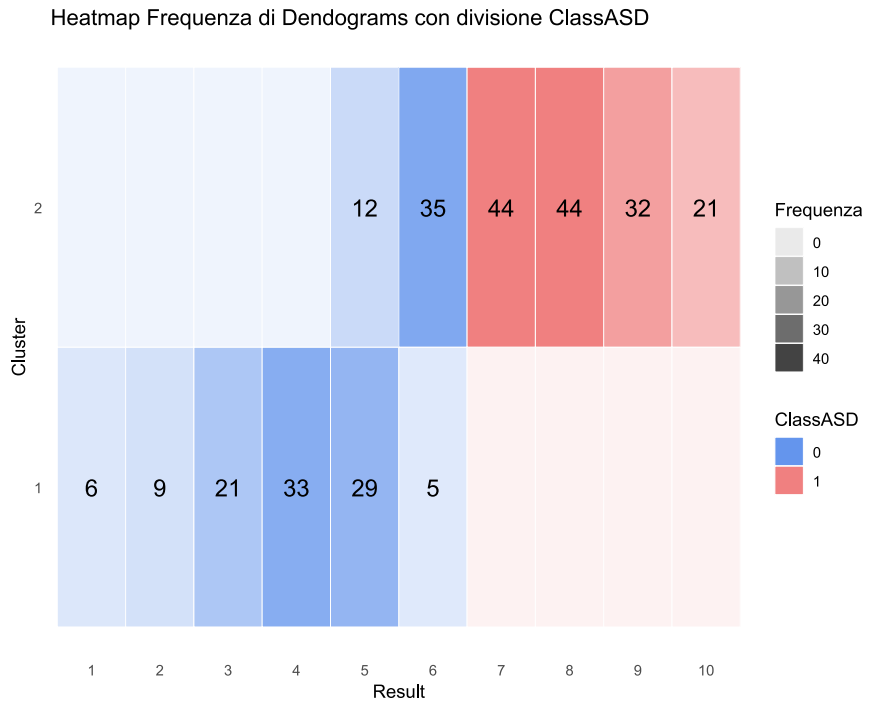


Figure 30: Tabella di contingenza Clustering Gerarchico tra ClassASD e i due Cluster

Da questa tabella possiamo effettuare diverse osservazioni. Il Cluster 1 contiene un numero significativo di osservazioni associate alla classe ClassASD = 0 (blu), mentre il Cluster 2 è dominato dalla classe ClassASD = 1 (rosso).

Inoltre, possiamo notare come si verifica un cambiamento graduale della distribuzione dei dati man mano che ci si sposta verso risultati più alti. Fino alla colonna 5-6, i valori sono prevalentemente associati alla classe ClassASD = 0, mentre nelle ultime colonne predomina ClassASD = 1 e questo è abbastanza coerente con il grafico del silhooette score, in cui il Cluster 2 mostrava maggiore variabilità.

3.5 DBSCAN Clustering

Il seguente grafico, rappresenta la distribuzione dei cluster ottenuti con DBSCAN e li confronta con la variabile ClassASD. L'asse orizzontale, etichettato come "Somma dei punteggi

(result)", mostra un valore numerico che rappresenta un punteggio complessivo ottenuto da un test o da una valutazione, mentre l'asse verticale indica i cluster assegnati dall'algoritmo, con due distinti gruppi numerati 0 e 1.

Ogni punto nel grafico rappresenta un'osservazione, il cui colore varia in base alla variabile ClassASD: i punti di colore blu (cornflowerblue) corrispondono alla categoria ClassASD = 0, mentre quelli di colore rosso (lightcoral) rappresentano ClassASD = 1. Per evitare sovrapposizioni e rendere la distribuzione più visibile, è stato applicato un leggero jitter, che aggiunge una piccola variazione ai punti, consentendo di distinguere meglio i dati anche quando sono molto vicini tra loro.

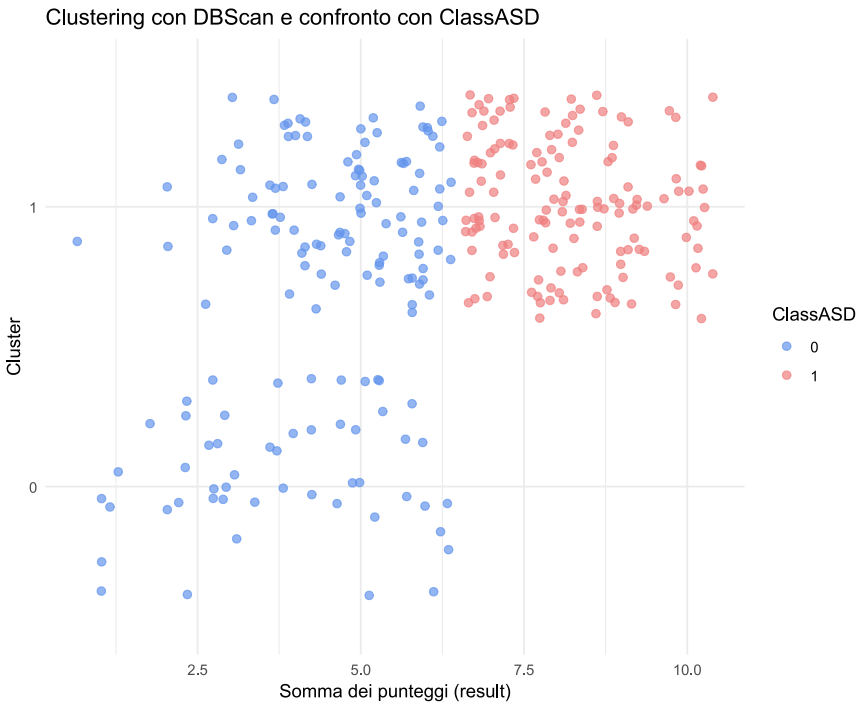


Figure 31: Grafico sulla distribuzione dei due Cluster e confronto con ClassASD per DBSCAN

Dall'osservazione del grafico, emerge un chiaro schema nella distribuzione dei punti. I dati con un valore basso di "Somma dei punteggi" tendono a concentrarsi nel Cluster 0 e sono prevalentemente di colore blu, suggerendo che la maggior parte degli individui in questa fascia appartiene alla categoria ClassASD = 0. Al contrario, man mano che il valore della Somma dei punteggi aumenta, i punti tendono a raggrupparsi nel Cluster 1 e diventano prevalentemente rossi, indicando che una quota maggiore di individui in questa regione appartiene alla categoria ClassASD = 1.

Nel complesso, il grafico mostra una relazione tra il punteggio complessivo e la classificazione effettuata da DBSCAN, suggerendo che l'algoritmo ha identificato correttamente due gruppi distinti con una buona corrispondenza con la variabile ClassASD.

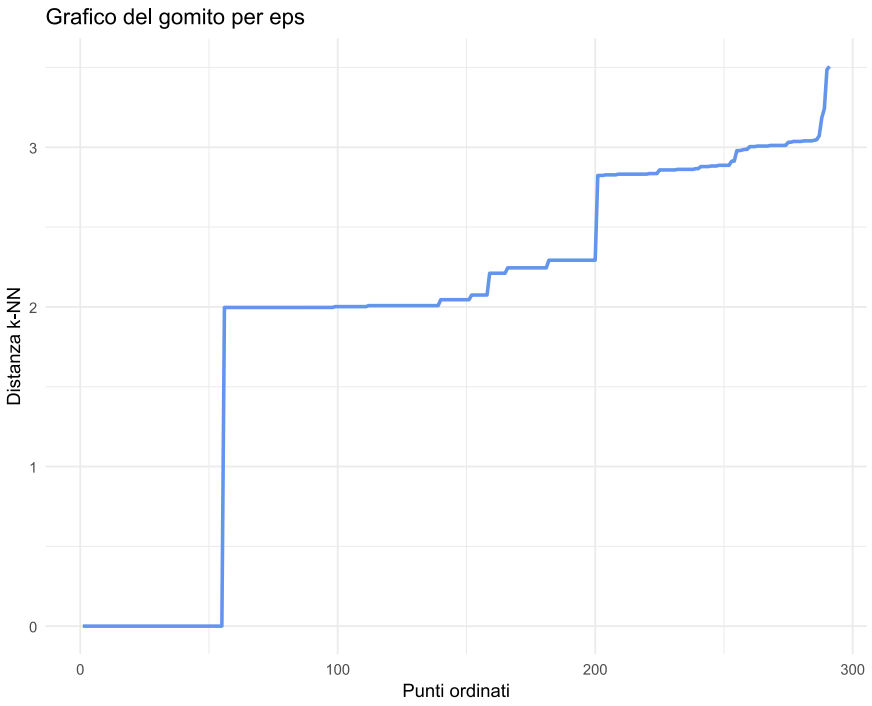


Figure 32: Grafico del gomito per eps

Il grafico precedente, mostra il gomito per la selezione del parametro eps nell'algoritmo di clustering DBSCAN. Questo tipo di grafico aiuta a determinare un valore ottimale per eps osservando il punto in cui la curva presenta un cambiamento significativo nella sua pendenza, noto come "gomito". L'asse X rappresenta i punti ordinati, mentre l'asse Y mostra la distanza k-NN (tipicamente la distanza rispetto ai k vicini più prossimi, con k scelto in base a minPts). La curva è una linea spezzata crescente che parte da valori vicini a zero e mostra un andamento inizialmente piatto, seguito da un graduale incremento e infine un forte aumento nella parte finale.

All'inizio, le distanze k-NN sono molto piccole, indicando che i punti sono relativamente vicini tra loro. Ad un certo punto (circa tra 100 e 200 punti ordinati), la curva inizia a crescere con maggiore pendenza, suggerendo un cambiamento nella densità dei punti. Il punto di gomito si trova dove il tasso di crescita della distanza cambia bruscamente, il che aiuta a selezionare un valore appropriato per eps.

Dopo aver analizzato l'andamento della curva, è stato scelto come valore eps 3.0 e

minPts=11 per eseguire il clustering con DBSCAN, poichè scegliere un valore troppo basso, si giungeva ad una frammentazione dei dati in troppi piccoli cluster con la conseguente classificazione di troppi punti come rumore. Allo stesso tempo, scegliere un valore troppo alto, si finirebbe per aggregare troppi punti in un unico cluster, perdendo la capacità di distinguere strutture più fini nei dati.

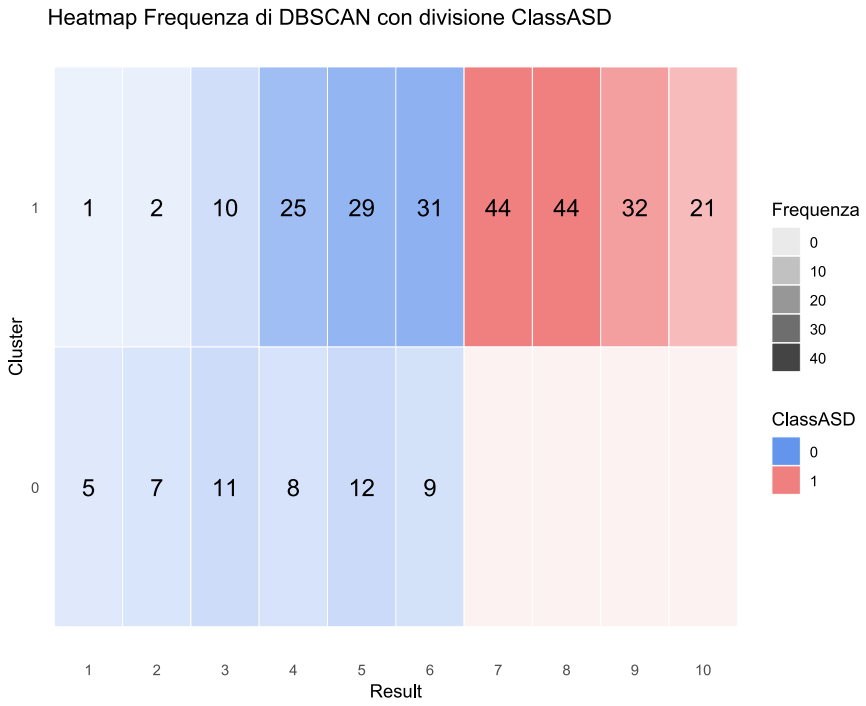


Figure 33: Tabella di Contingenza di Cluster 1 e Cluster 2 con DBSCAN

Il grafico precedente mostra una heatmap che rappresenta la frequenza dei dati raggruppati dall'algoritmo DBSCAN, con lo scopo di visualizzare la distribuzione delle osservazioni nei cluster individuati dall'algoritmo in relazione ai valori della variabile Result e alla classificazione ClassASD.

Da questo grafico, emergono diverse osservazioni: il Cluster 1 è il più popoloso, con un numero significativo di osservazioni soprattutto per valori di Result tra 5 e 10. Per valore di result da 5 a scendere, la maggior parte delle osservazioni in Cluster 1 appartiene a ClassASD = 0 (blu). Invece, per valori di Result da 6 a salire, le osservazioni in Cluster 1 appartengono prevalentemente a ClassASD = 1 (rosso), con le celle più scure attorno a Result 7 e 8 (entrambe con una frequenza di 44). Il Cluster 0 ha generalmente frequenze inferiori e sembra contenere solo osservazioni con ClassASD = 0, con numeri più bassi in tutta la gamma di Result.

Il passaggio dal blu al rosso avviene intorno a Result = 6, suggerendo che questo valore potrebbe essere un punto di separazione tra i due gruppi. Questo grafico evidenzia come DBSCAN abbia identificato due gruppi distinti e come essi siano correlati alla variabile ClassASD. Il clustering segue una struttura coerente con la distribuzione di ClassASD: i dati con basso Result tendono a essere in Cluster 0 e ClassASD = 0, mentre quelli con alto Result si raggruppano in Cluster 1 e ClassASD = 1. Questo suggerisce che il valore di Result potrebbe essere un discriminante chiave nella classificazione delle osservazioni, confermando l'idea che ClassASD sia strettamente legato a questa variabile.

3.6 Generazione dei dati: Large Language Model

Per la generazione del dataset, abbiamo utilizzato Chat GPT. Lo scopo del nuovo dataset ottenuto è quello di confrontarlo con il dataset originario per capire se i dati generati sono conformi con i dati del dataset originario. Il dataset generato è composto da 292 campioni, stesso numero del dataset originario con lo scopo di avere un numero di punteggi bilanciati.

Di seguito abbiamo riportato il dataset generato:

S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	age	gender	ethnicity	jundice	autism	country of res	used app before	result	age desc	relation	ClassASD
0	1	0	0	1	1	0	0	0	1	9	f	Middle Eastern	yes	yes	New Zealand	yes	4	4-11 years	Health care professional	0
1	1	0	1	1	1	0	0	0	0	4	m	Hispanic	no	no	Isle of Man	yes	5	4-11 years	Health care professional	0
0	0	0	0	0	0	1	0	1	0	11	m	Hispanic	yes	yes	China	no	2	4-11 years	Health care professional	0
0	0	1	1	1	0	0	1	1	1	6	f	Middle Eastern	no	no	New Zealand	no	6	4-11 years	Self	0
0	1	1	0	0	1	1	0	1	0	10	m	Hispanic	yes	yes	Egypt	no	5	4-11 years	Health care professional	0
0	0	0	0	0	0	1	1	0	0	11	m	Hispanic	yes	yes	Oman	no	2	4-11 years	Health care professional	0
0	0	0	1	1	0	0	0	1	0	7	m	Asian	no	no	Japan	yes	3	4-11 years	Health care professional	0
0	0	0	0	0	0	0	0	1	0	10	f	Black	yes	no	Turkey	yes	1	4-11 years	Relative	0
0	1	0	1	1	0	0	1	1	1	11	m	Latino	yes	no	United States	yes	6	4-11 years	Relative	0

Table 2: Prime dieci righe del dataset generato

3.6.1 Clustering Gerarchico per il dataset generato

Nel seguente grafico, viene mostrato il clustering gerarchico applicato al dataset generato, con i dati rappresentati su due componenti principali (PC1 e PC2), derivati da un'analisi PCA. I punti sono suddivisi in tre cluster, indicati con colori distinti: blu (Cluster 1), rosso (Cluster 2) e verde (Cluster 3). Ogni cluster è racchiuso da un contorno convesso che aiuta a visualizzare meglio la distribuzione dei dati nello spazio bidimensionale. I numeri all'interno dei punti rappresentano le etichette degli elementi nel dataset.

Dal punto di vista della distribuzione, è possibile effettuare delle osservazioni. Il Cluster 1 (blu) occupa la parte superiore del grafico, il Cluster 2 (rosso) si trova sulla destra e sembra essere più denso rispetto agli altri ed infine, il Cluster 3 (verde) è posizionato nella parte inferiore sinistra e presenta una distribuzione più diffusa.

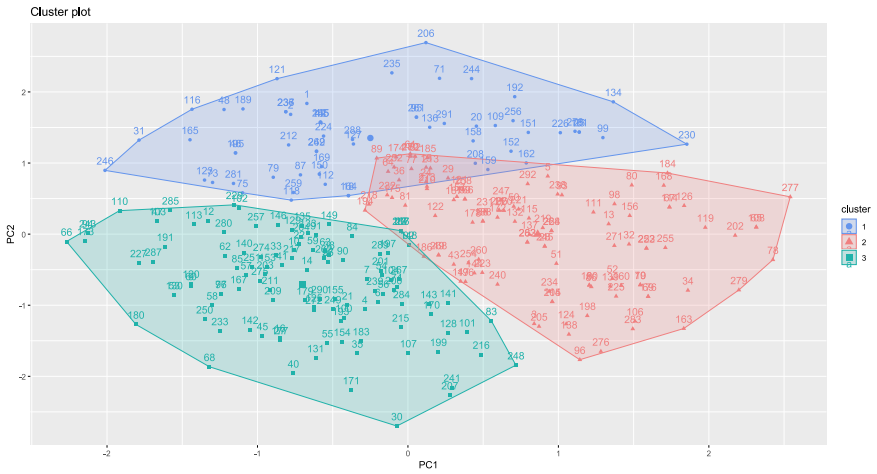


Figure 34: Clustering con il confronto tra centroidi

Sulla base delle informazioni appena fornite, possiamo affermare che questo grafico conferma la struttura del dataset attraverso clustering gerarchico, visualizzando la separazione dei gruppi e la loro distribuzione nello spazio ridotto delle componenti principali.

Per selezionare un numero di cluster necessario per l'analisi del clustering, è stato proposto un grafico, riportato in seguito, che determina il valore ottimale di k, dove k rappresenta il numero di cluster presi in considerazione.

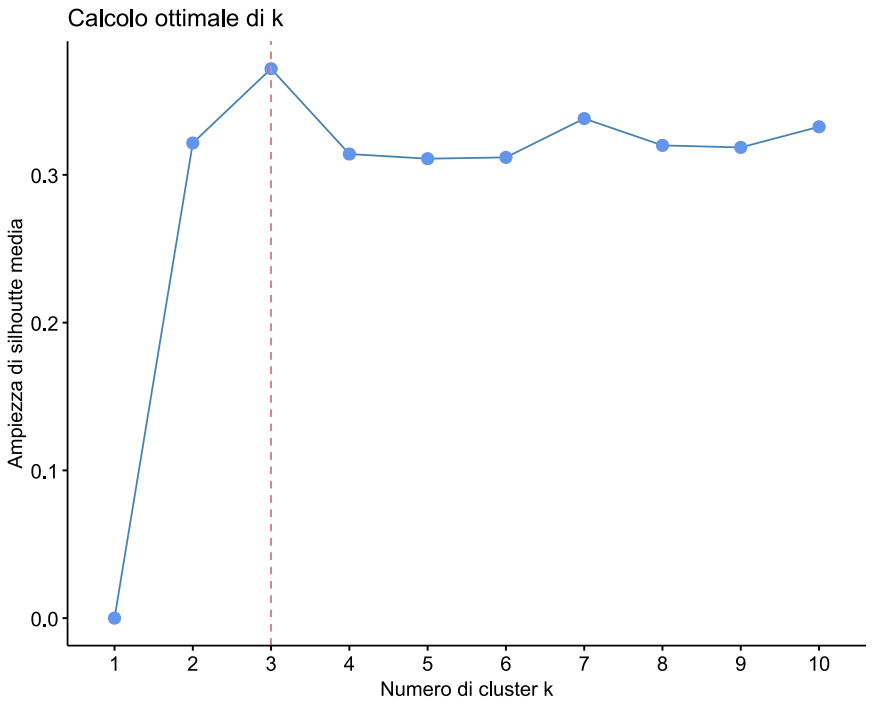


Figure 35: Grafico per il calcolo del valore ottimale di k per dataset generato

Sull'asse X troviamo il numero di cluster k, che varia da 1 a 10, mentre sull'asse Y è riportata l'ampiezza media della silhouette. Il grafico mostra un andamento crescente fino a un picco massimo, dopo il quale l'ampiezza del silhouette score si stabilizza o mostra variazioni meno significative. Il valore ottimale di k, viene evidenziato da una linea tratteggiata verticale, che nel caso specifico corrisponde a $k = 3$. Questo indica che, tra le diverse opzioni considerate, tre cluster rappresentano la scelta migliore per ottenere una buona separazione tra i gruppi mantenendo una coesione interna elevata.

In conclusione, il grafico conferma che il valore ottimale di k è 3 per il clustering dei dati. Questo valore rappresenta il numero di Cluster che sono stati riportati nella tabella di contingenza sottostante.

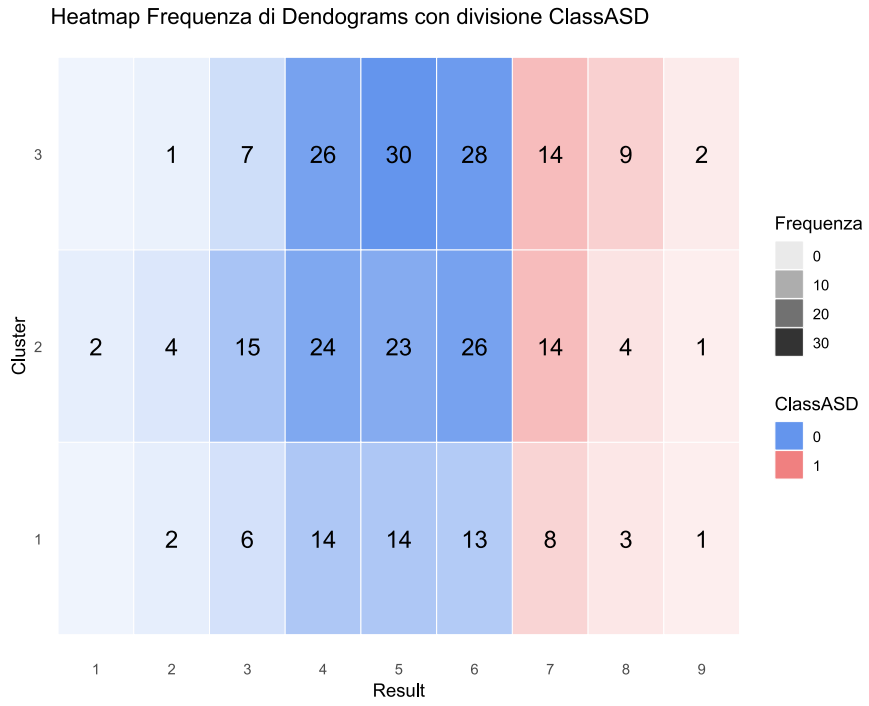


Figure 36: Tabella di contingenza per Cluster 1, Cluster 2 e Cluster 3 del dataset generato

Il grafico precedente, come già anticipato, rappresenta la distribuzione dei dati in tre cluster, denominati Cluster 1, Cluster 2 e Cluster 3, in relazione ai valori di Result e alla suddivisione della variabile ClassASD. Dall'analisi visiva della heatmap, emergono alcune osservazioni. In primo luogo, i valori centrali di "Result" (specialmente 4 e 5) tendono a presentare frequenze più elevate, con molteplici occorrenze nei cluster 2 e 3.

Questo potrebbe suggerire che la maggior parte dei dati si concentra in questi intervalli. Inoltre, la distribuzione tra i cluster mostra una presenza abbastanza bilanciata tra le diverse classi di ASD, anche se con una leggera predominanza di valori più alti nella classe "0" per alcuni cluster.

La heatmap fornisce, in conclusione, un'utile rappresentazione della distribuzione e della frequenza degli elementi nei diversi cluster, aiutando a comprendere meglio la struttura del dataset e le differenze tra i gruppi individuati.

Il grafico seguente, rappresenta la media del silhouette score per tre cluster, che visualizza la qualità della suddivisione dei dati in base al metodo di clustering utilizzato.

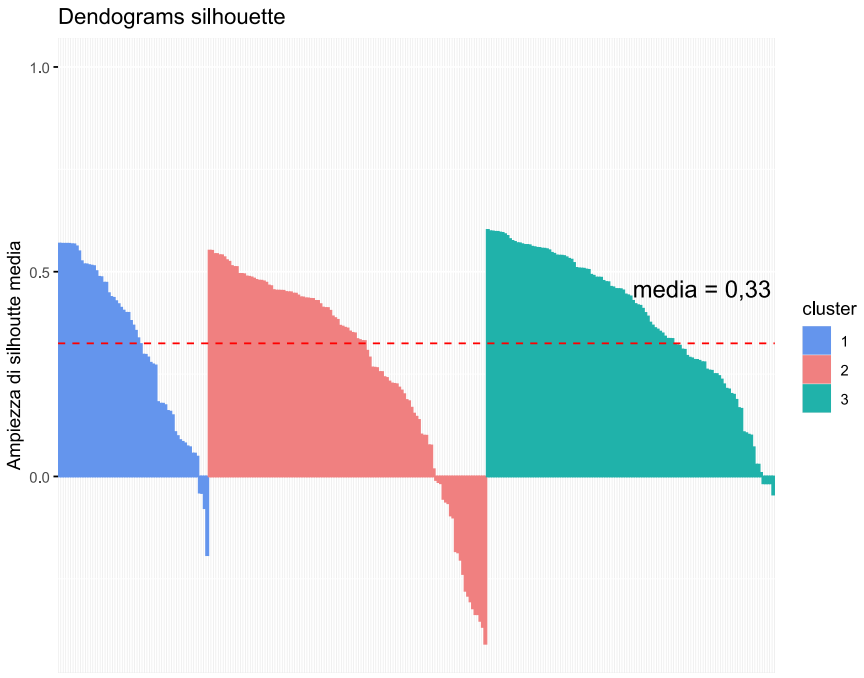


Figure 37: Grafico della media del Silhouette Score per i tre Cluster del dataset generato

Sull'asse X, i punti sono suddivisi nei tre cluster individuati, rappresentati con colori distinti: blu (Cluster 1), rosso (Cluster 2) e verde (Cluster 3). Sull'asse Y è riportata l'ampiezza del silhouette score per ciascun punto del dataset, che misura il grado di appartenenza di ogni osservazione al proprio cluster rispetto agli altri.

Il valore medio del silhouette score, pari a 0,33, è evidenziato da una linea tratteggiata rossa. Questo valore indica che, in generale, la coesione all'interno dei cluster è moderata e la separazione tra i gruppi è relativamente buona, anche se non perfetta. Infatti, si può notare che, in ciascun cluster, esistono alcuni elementi con valori negativi o prossimi allo zero, suggerendo che per questi dati la classificazione potrebbe essere meno chiara.

Nonostante il valore medio non sia molto elevato (idealmente sarebbe sopra 0,5 per una suddivisione molto netta), sembra comunque sufficiente per giustificare la scelta di tre cluster, ovvero di $k = 3$.

3.6.2 Clustering con DBSCAN sul modello generato

La seguente immagine mostra un grafico di dispersione che rappresenta i risultati del clustering ottenuti con DBSCAN, confrontando i cluster individuati con la variabile ClassASD. Da questo grafico possiamo osservare che la maggior parte dei punti si trovano nella fascia

di result compresa tra 2 e 7.5, con una distribuzione abbastanza densa.

Il Cluster 0 contiene prevalentemente punti blu (ClassASD = 0) ed è distribuito su quasi tutta la gamma dei valori di result. Il Cluster 1 contiene quasi esclusivamente punti rossi (ClassASD = 1) e sembra concentrarsi nei valori più bassi di result, ma con una presenza ridotta rispetto al Cluster 0.

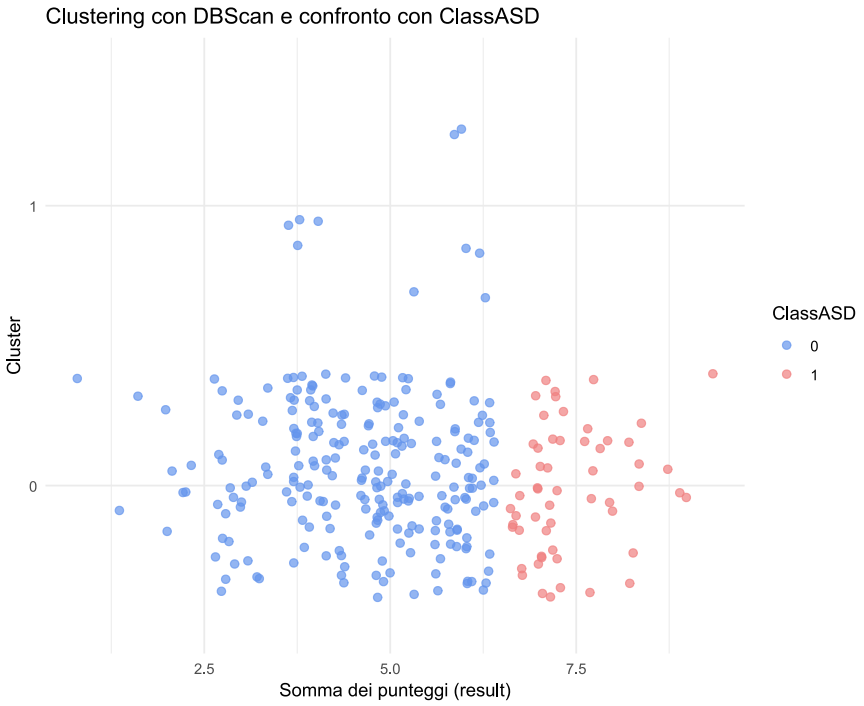


Figure 38: Clustering con DBSCAN per dataset generato e confronto con ClassASD

A partire da result=6.5, si nota una leggera transizione, con alcuni punti rossi che compaiono all'interno del Cluster 0. Non sembra esserci una separazione netta tra le due categorie, ma piuttosto una graduale sovrapposizione.

In conclusione, DBSCAN ha identificato due cluster, ma la separazione rispetto a ClassASD non è netta come negli altri grafici precedenti. Il fatto che il Cluster 0 contenga sia osservazioni di ClassASD = 0 che alcune di ClassASD = 1 potrebbe suggerire che i parametri di eps e minPts non siano perfettamente ottimizzati per distinguere i due gruppi. Il Cluster 1 è più specifico e contiene quasi esclusivamente punti rossi, suggerendo che esiste una certa struttura nei dati, ma DBSCAN non è riuscito a catturare la separazione tra ClassASD = 0 e ClassASD = 1.

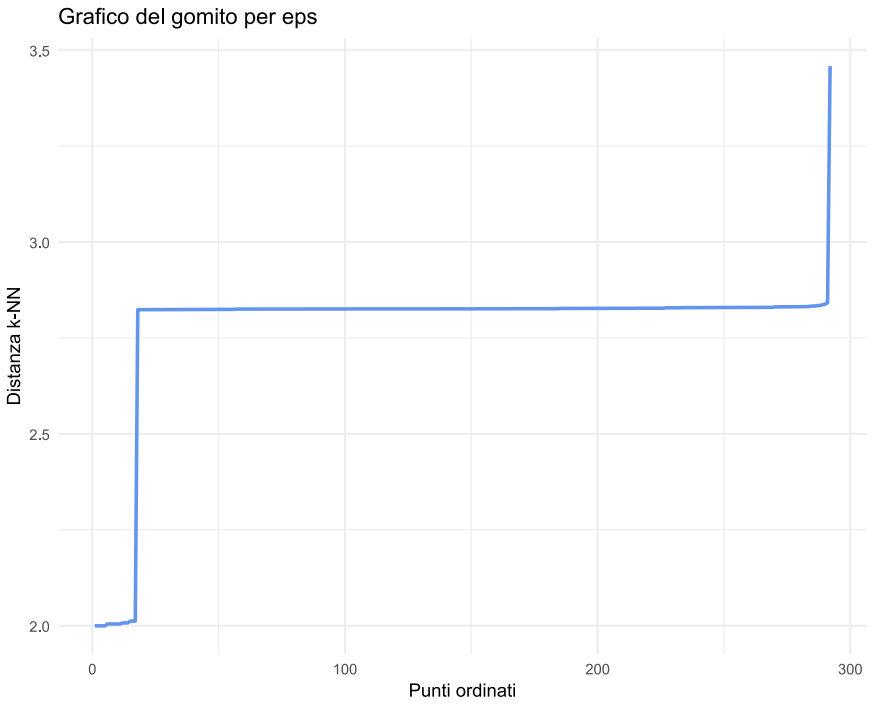


Figure 39: Grafico del gomito per eps Dataset Generato

Anche in questo caso, come per il dataset originale, è stato costruito il grafico del gomito per la scelta del valore eps nell'algoritmo di clustering DBSCAN. Da questo grafico possiamo osservare l'andamento della curva, in cui, all'inizio, sembra quasi piatto, con valori della distanza intorno a 2.0. Questa fase indica che per la maggior parte dei punti la distanza dal loro k-esimo vicino più prossimo è relativamente bassa, segnalando la presenza di una zona ad alta densità di dati.

Ad un certo punto, la curva inizia a salire in modo più marcato, evidenziando un cambiamento nella distribuzione delle distanze. Si forma così il cosiddetto punto di gomito, il momento in cui il tasso di crescita della distanza subisce un brusco cambiamento. In questo caso, il gomito sembra collocarsi intorno a un valore compreso tra 2.5 e 3.0 sulla scala della distanza k-NN.

Dopo aver analizzato l'andamento della curva, è stato scelto come valore eps 2.8 e minPts=10 per eseguire il clustering con DBSCAN, per le stesse motivazioni introdotte precedentemente per il grafico del gomito del dataset originale.

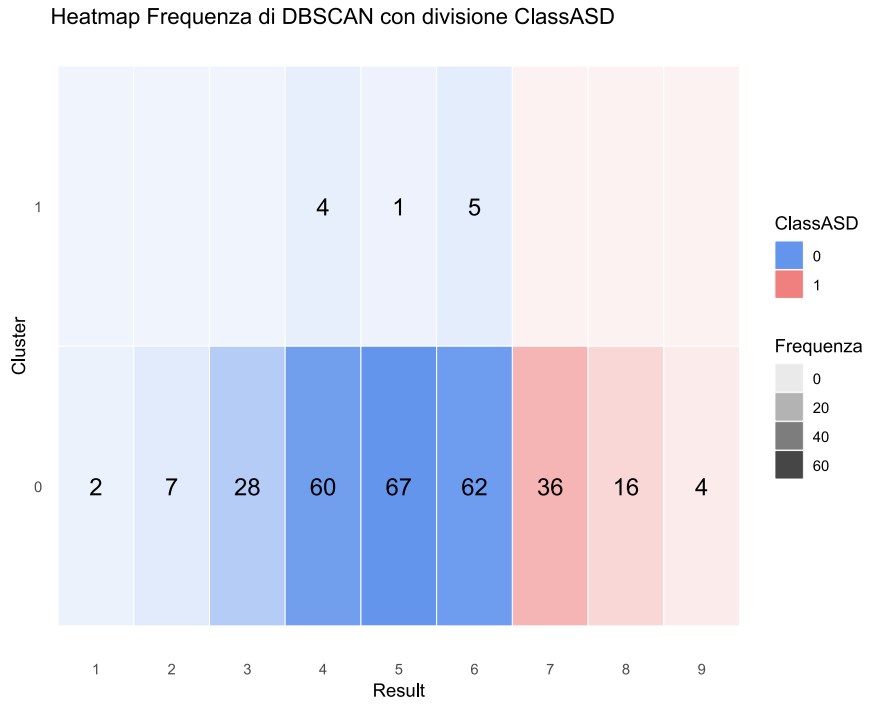


Figure 40: Tabella di contingenza frequenza di DBSCAN con divisione ClassASD

Come per il dataset originale, è stata costruita una heatmap per il dataset generato, che rappresenta una tabella di contingenza tra i cluster ottenuti con DBSCAN e la variabile ClassASD, che indica una classificazione binaria. L'obiettivo è analizzare come si distribuiscono i cluster identificati dall'algoritmo rispetto alla variabile ClassASD.

Osservando il grafico, la maggior parte dei dati è concentrata nelle colonne result 4, 5 e 6, con frequenze particolarmente alte nel Cluster 0, dove si osservano valori massimi di 60, 67 e 62. La classe 0 (blu) è predominante nei punteggi più bassi, mentre la classe 1 (rosso) diventa più rilevante nei punteggi più alti, soprattutto a partire da result 7, dove la frequenza della classe 1 aumenta sensibilmente. Il Cluster 1 invece, presenta frequenze molto più basse e distribuite in modo più uniforme, con valori compresi tra 1 e 5.

In definitiva, il grafico mostra che l'algoritmo DBSCAN ha individuato due cluster principali con una netta distinzione nella composizione di ClassASD a seconda del valore di result. Mentre per punteggi più bassi prevale la classe 0, nei punteggi più alti la classe 1 diventa più frequente.

La heatmap fornisce quindi un'utile visualizzazione della relazione tra il clustering non supervisionato di DBSCAN e la classificazione di ClassASD, permettendo di capire meglio la struttura dei dati e come le due variabili si correlano tra loro.

3.7 Valutazione Comparativa tra Cluster

Il seguente grafico a barre mette a confronto due metriche di valutazione del clustering K-means applicato a due diversi dataset: dataset originale e dataset generato. Lo scopo della è quello di confrontare la qualità del clustering tra i due dataset utilizzando due metriche: Indice di Calinski-Harabasz, rappresentato con barre di colore blu, che serve a valutare la qualità di un clustering, ovvero a misurare quanto i cluster siano compatti e ben separati tra loro e la media della silhouette score, rappresentata con barre di colore rosso.

Dal grafico emerge chiaramente che la media della silhouette score (in rosso) assume valori significativamente più alti rispetto all'indice di Calinski-Harabasz in entrambi i dataset. Questo suggerisce che il clustering è piuttosto coeso e ben separato. Per quanto riguarda l'indice di Calinski-Harabasz (in blu) ha valori molto più bassi rispetto alla silhouette score, il che potrebbe indicare che la separazione tra i cluster è meno marcata secondo questa metrica. In conclusione, i valori delle metriche sono molto simili tra il dataset originale e quello generato, il che suggerisce che la struttura del clustering rimane coerente tra le due versioni del dataset.

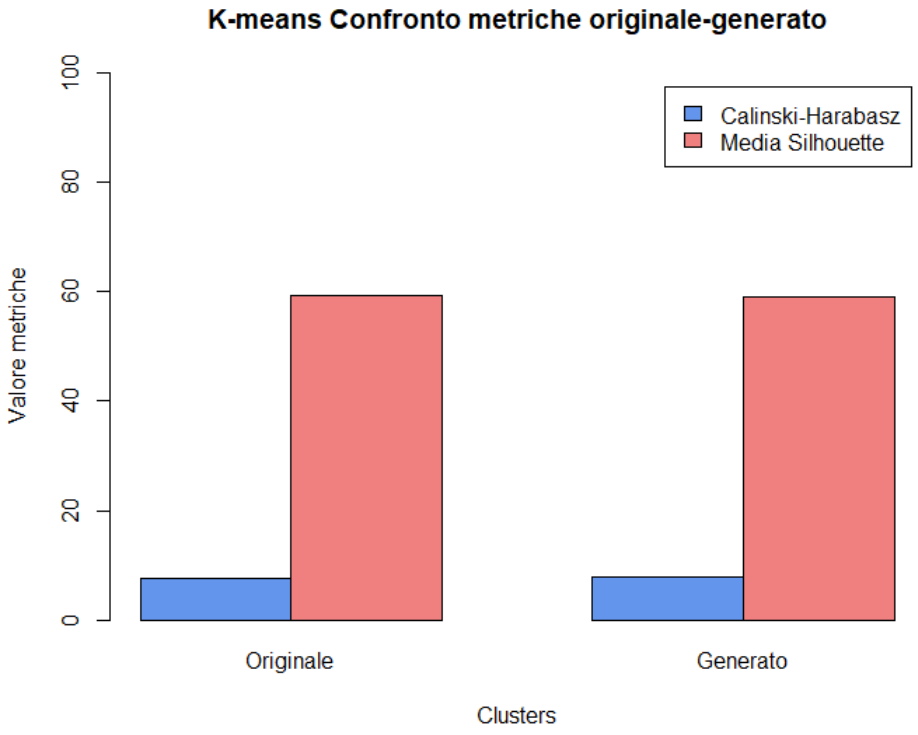


Figure 41: Confronto tra Cluster con K-means

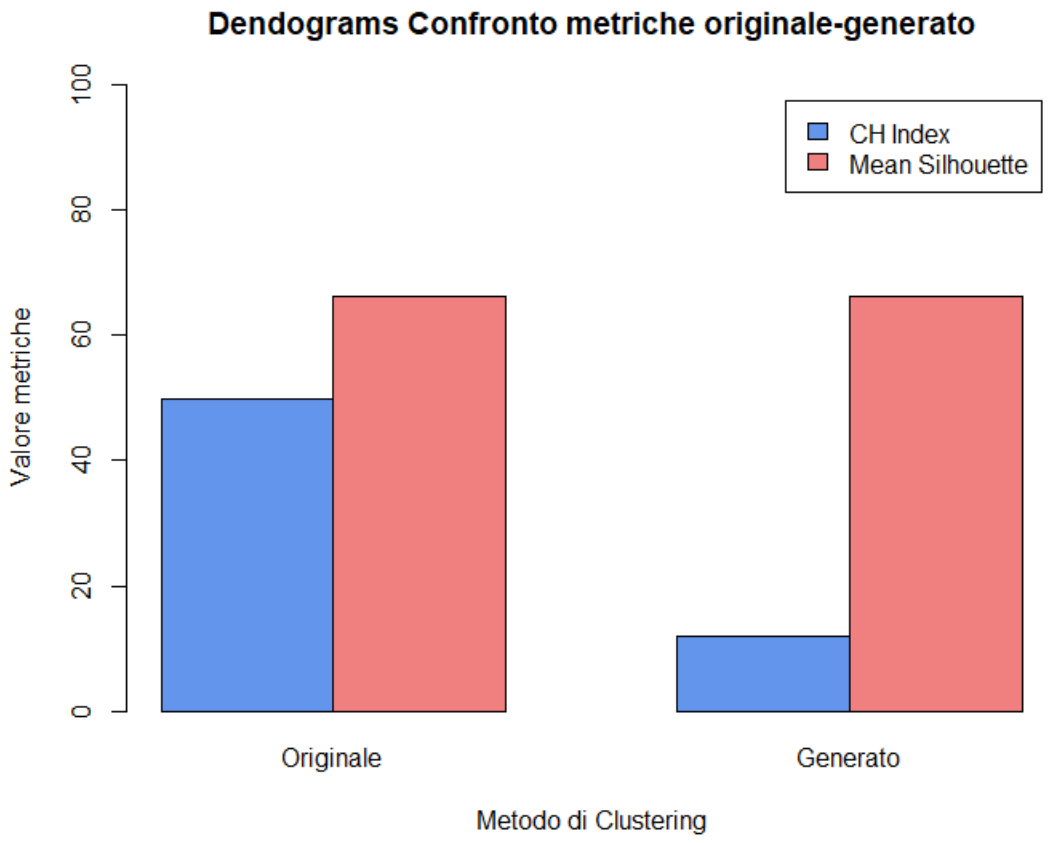


Figure 42: Grafico Valutazione comparativa Clustering gerarchico

Nel grafico precedente, è stata fatta una rappresentazione simile a quella effettuata con K-Means precedentemente, ma in questo caso stiamo prendendo in considerazione la valutazione comparativa tra i due dataset per il Clustering Gerarchico.

Il grafico presenta sull'asse X le due diverse configurazioni del dataset (Originale e Generato), mentre sull'asse Y il valore delle metriche di clustering. Possiamo notare che per quanto riguarda il valore di CH-Index, risulta più basso rispetto alla media del silhouette score in entrambe le configurazioni. Questo suggerisce una separazione non estremamente forte tra i due cluster.

Per quanto riguarda il valore della media del silhouette score, è più alto rispetto al valore di CH-Index. Ciò indica che i cluster presentano una forte coesione. Il confronto tra il dataset originale e quello generato mostra valori simili per entrambe le metriche, indicando che il dataset generato mantiene una struttura di clustering comparabile a quella dell'originale.

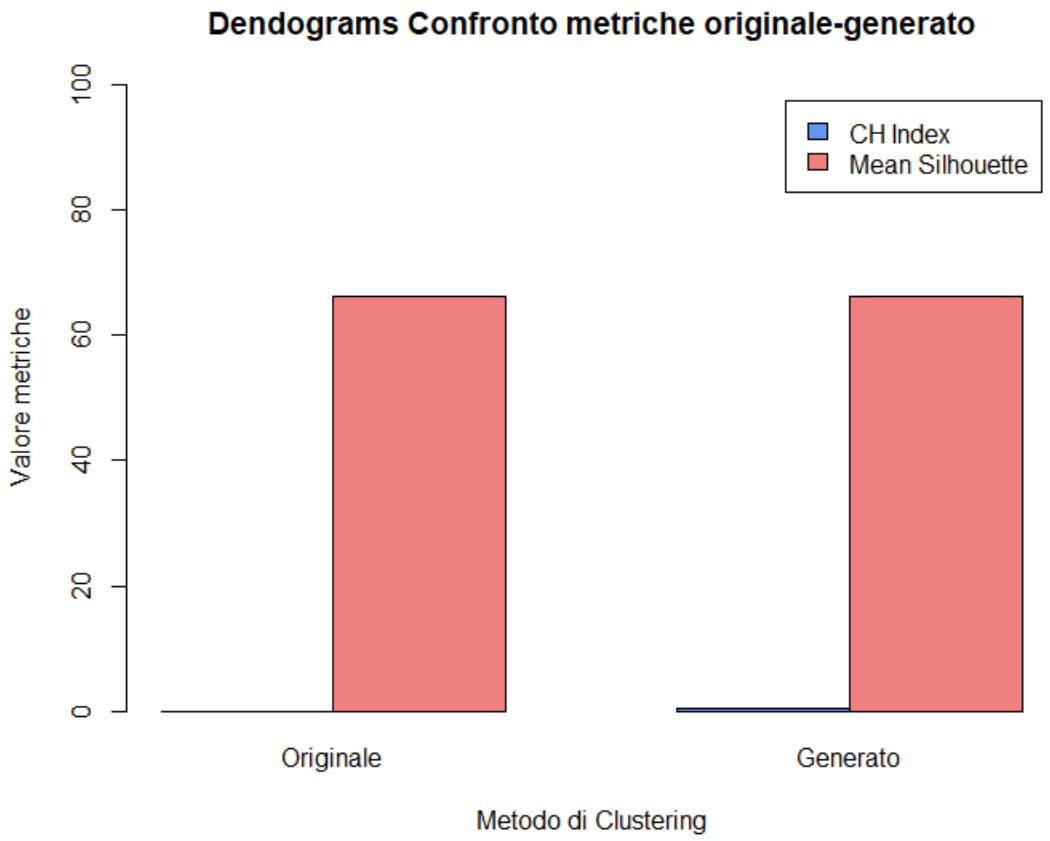


Figure 43: Grafico Valutazione comparativa DBSCAN

il valore del CH Index è quasi nullo per entrambi i metodi, mentre l'Indice di Silhouette è molto più alto e pressoché identico tra Originale e Generato. Questo suggerisce che, secondo il CH Index, i cluster ottenuti non sono particolarmente ben separati o compatti, il che non è sorprendente dato che DBSCAN non si basa su una chiara separazione tra gruppi, ma piuttosto sulla densità dei dati.

Dall'altro lato, l'Indice di Silhouette, che misura quanto bene gli elementi sono assegnati ai loro cluster rispetto agli altri, sembra indicare che entrambi i metodi abbiano una qualità di clustering simile. Questo potrebbe voler dire che la struttura dei cluster rilevati nel dataset generato è coerente con quella dell'originale.

Link Utili

Al seguente link è presente la repository GitHub: **ASD-Child**