Università degli Studi di Roma "Roma Tre"

Facoltà di Economia

Corso di Laurea Economia e Big Data

Metodi di clustering per variabili miste

Relatore

Chiar.mo Prof.

Francesco Dotto

Candidato

Leonardo Suriano

matr. 589966

Anno accademico 2024-2025

Seduta di Laurea di Luglio 2025

Indice

1 Introduzione al clustering **2**

1.1 Apprendimento statistico 2

1.2 Tipologie di apprendimento statistico 2

2 Clustering **5**

2.1 Introduzione 5

2.2 Cos'è un cluster? 6

2.2.1 Misure di similarità 7

2.3 Metodi partizionali 11

2.3.1 Metodo K-means 11

2.3.2 Metodo K-medoids 13

2.4 L'indice di silhouette 16

2.5 Clustering per variabili miste 17

2.5.1 Distanza di Gower 17

3 Caso Pratico **20**

3.1 Data cleaning e implementazione 20

3.2 Risultati e analisi 22

3.2.1 Profilazione dei cluster 24

Conclusioni **27**

Bibliografia **28**

Capitolo 1

Introduzione al clustering

1.1 Apprendimento statistico

L'apprendimento statistico si riferisce a una serie di strumenti per modellizzare e comprendere set di dati complessi. La crescente disponibilità di dati, unita all'aumento della capacità computazionale, hanno reso l'apprendimento statistico un campo fondamentale e di forte interesse in molte aree scientifiche, quali il marketing, finanza e altre discipline aziendali.
Questa materia esplora lo studio e l'implementazione di algoritmi che siano in grado di apprendere da un set di dati forniti e di fare delle predizioni su questi costruendo in modo induttivo un modello. L'induzione, infatti, rappresenta un principio fondamentale della statistica, permettendo di dedurre una regola generale a partire da osservazioni particolari.
Lo scopo di questo breve elaborato sarà quindi esplorare le principali tecniche di apprendimento sotto determinate condizioni e valutarne l'efficacia in diversi contesti applicativi.

1.2 Tipologie di apprendimento statistico

In questo contesto possiamo distinguere differenti tipologie di apprendimento a seconda della disponibilità dei dati a cui possiamo accedere e delle necessità che il

sistema deve soddisfare in funzione del compito richiesto:

- Apprendimento supervisionato: in questo caso i dati a disposizione sono coppie di input/output. Utilizzando un formalismo matematico è possibile definire gli input come variabili indipendenti e gli output come variabili dipendenti. Ne deriva che l'apprendimento supervisionato ha il compito di determinare una funzione f in grado di correlare la variabile dipendente y alla variabile indipendente x nel seguente modo:

  $$y = f(x)$$

  L'obiettivo è quello di adattare un modello che relazioni la variabile dipendente (variabile risposta) ai predittori nella cosiddetta fase di "addestramento", con lo scopo di prevedere con precisione l'output per osservazioni future (previsione).
  I metodi che operano nel dominio dell'apprendimento supervisionato sono numerosi e si possono suddividere principalmente in due categorie:

  - Metodi di regressione: quando la variabile dipendente è di tipo continuo (variabile quantitativa);
  - Metodi di classificazione: quando la variabile dipendente è di tipo discreto (variabile qualitativa o categorica).

- Apprendimento non supervisionato: Si definisce apprendimento non supervisionato l'insieme di tecniche e di strumenti statistici volti a individuare strutture, pattern o relazioni nei dati senza l'ausilio di etichette o output predefiniti. Viene richiamato per il caso in cui si possiede solo un insieme di funzioni $X_1, X_2, ...X_p$ misurate su n osservazioni.
  Non possediamo una variabile di risposta associata Y, pertanto, il fine di questa metodologia è di tipo informativo piuttosto che predittivo. Si parla infatti di analisi esplorativa dei dati, nella quale non è possibile verificare a priori la validità delle ipotesi formulate, poiché non esiste un insieme di

output predefiniti con cui confrontare i risultati ottenuti.

L'assenza di una variabile di risposta impone un approccio basato sulla scoperta di strutture latenti nei dati, piuttosto che sulla predizione di un valore specifico. Per strutture latenti si intende un insieme di schemi o relazioni nascoste che quindi non sono direttamente osservabili, ma che possono essere estratte attraverso l'ausilio di metodi statistici o algoritmi che riescano ad individuare e spiegare queste relazioni per semplificare e interpretare i dati in base agli obiettivi imposti. Questo processo di trasformazione e semplificazione può tradursi nell'implementazione di algoritmi che si basano sulle analogie tra i diversi dati di input per cercare di ricostruire un raggruppamento o per evidenziare particolari caratteristiche presenti nei dati. Possiamo suddividere l'insieme di tecniche per l'apprendimento non supervisionato in due macrocategorie:

- L'analisi delle componenti principali (PCA);
- Metodi di Clustering.

Dal prossimo capitolo, l'attenzione sarà rivolta esclusivamente al Clustering, analizzando le principali tecniche, le metriche di valutazione e le applicazioni pratiche di questa metodologia.

Capitolo 2

Clustering

2.1 Introduzione

Il clustering si riferisce ad un insieme molto ampio di tecniche di analisi multivariata dei dati, volte alla selezione e raggruppamento di elementi omogenei in un insieme di osservazioni. Si immagini di avere un insieme di oggetti da analizzare in cui, a differenza della classificazione, non è nota l'etichetta di classe di ciascun oggetto. Sebbene la classificazione sia un mezzo efficace per distinguere gruppi o classi di oggetti essa richiede, come già anticipato nel capitolo precedente, una costruzione e un'etichettatura del set di addestramento che risultano spesso costose. Solitamente può essere desiderabile procedere in senso inverso, partizionando prima i dati in gruppi sulla base della loro similarità (utilizzando il clustering) e, successivamente, assegnando le etichette al numero di gruppi così ottenuto. I contesti dove si può applicare questa tecnica sono numerosi, passando dalla sociologia per distinguere gruppi di individui con caratteristiche differenti, alla biologia per derivare le tassonomie delle piante e per categorizzare i geni con funzionalità simili, giungendo poi all'economia e marketing, essenziale per effettuare segmentazioni di mercato e distinguere la clientela. Le potenziali applicazioni sono molte, tuttavia, anche in questo caso, ci sono requisiti e vincoli che devono essere presi in considerazione affinché l'approccio risulti efficace:

- Qualità e quantità dei dati disponibili: i dataset limitati o rumorosi possono compromettere l'accuratezza delle analisi.
- Scalabilità: molti algoritmi di clustering lavorano bene su piccoli insiemi di dati, tuttavia, nella realtà, sono necessari strumenti che riescano ad adattarsi bene anche a dataset molto ampi.
- Alta dimensionalità: un database può contenere diverse dimensioni e attributi. Tuttavia, la capacità computazionale di molti algoritmi e la percezione umana, si limitano a gestire dataset con al più tre dimensioni.
- Capacità di trattare diversi tipi di attributi: molti algoritmi sono progettati per lavorare solo su dati numerici. Tuttavia, nei casi pratici e in contesti reali si può richiedere un'analisi di clustering di altri tipi di dati, quali dati binari, categorici e ordinali, o talvolta una combinazione di essi.

L'ultimo punto verrà approfondito nella seconda parte di questo breve elaborato, dove verrà analizzata la capacità di trattare diverse tipologie di attributi nei metodi di clustering. Inoltre verrà presentato un caso pratico in cui questa tecnica sarà applicata per estrarre informazioni significative dai dati misti, fornendo spunti utili per eventuali approfondimenti e applicazioni future.

2.2 Cos'è un cluster?

Per comprendere il concetto di clustering, è fondamentale avere una chiara idea di cosa sia un cluster e di come si formi. Possiamo definire un gruppo (cluster) come una collezione di istanze (osservazioni) tali che:

- le istanze dello stesso cluster sono simili tra loro.
- le istanze di cluster diversi sono dissimili.

Questa prima definizione introduce un aspetto fondamentale della clusterizzazione di dati: il concetto di similarità, inteso come una quantità che riflette la forza della

relazione tra due elementi di dati [1]. Questa forza può assumere domini differenti a seconda della natura delle variabili considerate. Generalmente, un'osservazione viene concepita come un punto definito in uno spazio geometrico bidimensionale nel quale la relazione tra due o più punti viene misurata attraverso il concetto di distanza, che a sua volta può assumere diverse formulazioni a seconda del contesto e del fine che si vuole raggiungere. Ancor prima di esplorare queste misure è necessario però acquisire gli strumenti necessari per comprendere al meglio il processo di clusterizzazione.

Una definizione formale del problema può essere impostata nel seguente modo [2]: Dato un set di osservazioni (oggetti) $X = \{x_1, x_2, \ldots, x_i, \ldots, x_n\}$ dove x_i è un oggetto in uno spazio p-dimensionale, e n è il numero di oggetti in X, allora il clustering è la partizione dell'insieme X in K clusters $\{C_1, C_2, \ldots, C_K\}$ in modo da soddisfare le seguenti condizioni:

- Ogni osservazione deve essere assegnata a un cluster, i.e.

$$\bigcup_{k=1}^{K} C_k = X$$

- Ogni cluster deve contenere almeno un'osservazione, i.e.

$$C_k \neq \emptyset, \quad k = 1, \ldots, K$$

- Ogni osservazione deve appartenere ad uno e un solo cluster, i.e.

$$C_k \cap C_{k'} = \emptyset \quad \text{dove } k \neq k'$$

Queste condizioni stabiliscono la partizione dei cluster secondo una coerenza precisa: ciascun oggetto deve essere incluso esattamente in un singolo cluster, non devono esserci cluster vuoti e non possono esistere oggetti non assegnati. Qualora anche solo una di queste condizioni venisse meno, non si avrebbe più un'assegnazione che rispetti la definizione di partizione in cluster.

2.2.1 Misure di similarità

Passiamo adesso alla descrizione delle misure di similarità per variabili prettamente numeriche, necessarie per introdurre in un secondo momento le principali tecniche

di clustering [3].

Sia X la matrice di dati di ordine $(n \times p)$ contenente i valori di p variabili osservate su n unità da raggruppare:

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix} \tag{2.1}$$

L'elemento generico della matrice è x_{ij}, che rappresenta il valore della variabile j $(j = 1, \ldots, p)$ osservata sull'unità i $(i = 1, \ldots, n)$. Le righe di X corrispondono alle unità. Per ogni unità i, osserviamo un vettore di lunghezza p indicato con

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{ip}).$$

Per confrontare due generiche unità osservate e individuare potenziali raggruppamenti, è necessario definire un criterio di similarità o distanza $d(x_i, x_k)$ tra i vettori $x_i, x_k \in \mathbb{R}^p$, che rispetti però quattro proprietà fondamentali [4]:

- Non negatività: $d(x_i, x_k) \geq 0, \quad \forall x_i, x_j.$
- Separabilità: $d(x_i, x_k) = 0,$ se e solo se $x_i = x_k.$
- Simmetria: $d(x_i, x_k) = d(x_k, x_i).$
- Disuguaglianza triangolare: $d(x_i, x_w) \leq d(x_i, x_k) + d(x_k, x_w).$

La metrica più utilizzata per confrontare variabili continue è la *distanza Euclidea*: Se x_i e x_k indicano i vettori delle variabili osservate per le unità i e k rispettivamente, $i, k = 1, ..., n$, la distanza euclidea è definita come:

$$d(x_i, x_k) = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{kj})^2} = \|x_i - x_k\|_2$$

Questa metrica ha un significato intuitivo di distanza, intesa come lunghezza del segmento avente per estremi i due punti considerati; per una trattazione teorica

viene utilizzata la seconda espressione detta *norma euclidea*. Tuttavia questa metrica non è esente da difetti [5]:

- Lavora bene solo su dataset con cluster "isolati" e "compatti".
- Se due vettori non condividessero neanche un valore di attributo, la loro distanza potrebbe risultare inferiore rispetto ad un'altra coppia di vettori che presenta gli stessi valori di attributo.
- Gli attributi con scale più grandi tendono a dominare la metrica su attributi di scala minore generando così una distorsione (in questo caso si cerca di normalizzare i dati per convertirli sulla medesima scala di valori [6]).

Pertanto, prima di avviare il processo di clusterizzazione, le distanze vengono spesso normalizzate dividendo la distanza di ciascun attributo per l'intervallo (ad esempio, massimo-minimo) di tale attributo, in modo che la distanza appartenga all'intervallo [0,1].

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

Inoltre, per evitare valori anomali (outliers), si tende a standardizzare, quindi sottrarre per ogni valore di ciascun attributo la sua media e a dividere per la deviazione standard anziché per l'intervallo [7].

$$x'_i = \frac{x_i - \mu}{\sigma}$$

Un'ulteriore forma metrica che cerca, anche solo parzialmente, di risolvere il problema degli outliers è la cosiddetta distanza di *Manhattan*:

$$d_1(x_i, x_k) = \sum_{j=1}^{p} |x_{ij} - x_{kj}| = \|x_i - x_k\|_1$$

Questa misura è meno sensibile agli outliers rispetto alla norma euclidea semplicemente perché non eleva al quadrato le differenze. Tuttavia, in questo caso, il significato di distanza è meno intuitivo: Dati due punti in uno spazio n-dimensionale la distanza viene calcolata per vie ortogonali, quindi effettua la somma delle differenze in valore assoluto delle loro coordinate cartesiane [8].

Dopo aver presentato le metriche necessarie per comprendere ed individuare la similarità tra due punti, adesso è fondamentale introdurre uno strumento

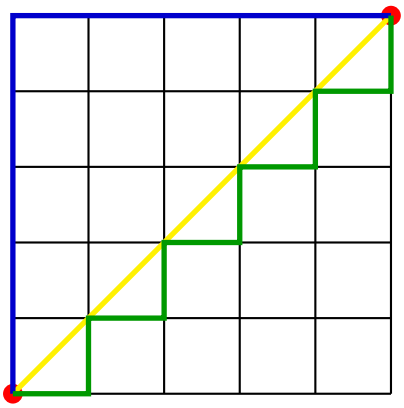


Figura 2.1: Differenza tra distanza Euclidea (gialla) e Manhattan (blu e verde).

che raccolga e rappresenti in forma tabellare le nostre distanze tra coppie di osservazioni x_i e x_k: la matrice delle distanze D.

$$D = \begin{bmatrix} 0 & d(x_1, x_2) & d(x_1, x_3) & \dots & d(x_1, x_n) \\ d(x_2, x_1) & 0 & d(x_2, x_3) & \dots & d(x_2, x_n) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d(x_n, x_1) & d(x_n, x_2) & d(x_n, x_3) & \dots & 0 \end{bmatrix}$$

Come possiamo notare la matrice delle distanze è una matrice simmetrica, inoltre la diagonale principale presenta esclusivamente valori nulli dal momento che la distanza tra un punto e sé stesso è zero. Per la proprietà di simmetria la matrice D è uguale alla sua trasposta ($D = D^T$).
Possiamo distinguere approssimativamente i metodi di clustering in due classi principali:

- Metodi partizionali: dato il numero K di partizioni, si utilizza una riallocazione ricorsiva che sposta gli elementi da un cluster a un altro, minimizzando una funzione di costo.

- Metodi gerarchici: creano una decomposizione gerarchica dei dati con metodi divisivi o agglomerativi.

2.3 Metodi partizionali

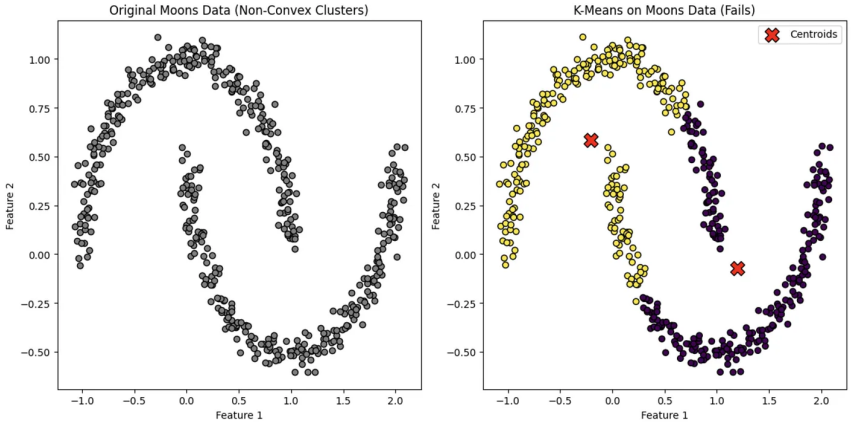
Consideriamo di avere un dataset con N osservazioni, i metodi a partizione identificano un numero definito di punti appartenenti al dataset che siano in qualche modo rappresentativi di ciascun cluster. Questi punti, noti come *centroidi*, vengono utilizzati per suddividere le osservazioni in gruppi, minimizzando una misura di dissimilarità interna. L'assegnazione dei punti ai cluster avviene iterativamente, aggiornando i centroidi fino a ottenere una configurazione stabile (ovvero quando la funzione costo scende al di sotto di un livello prestabilito). A differenza dei metodi gerarchici, bisogna imporre in anticipo il numero desiderato di cluster k, sebbene questo apra a una serie di innumerevoli problematiche che approfondiremo più avanti. Adesso andiamo a studiare i due principali metodi di partizione conosciuti: *K-means* e *K-medoids*.

2.3.1 Metodo K-means

Il metodo K-means è il metodo partizionale più utilizzato dalla comunità scientifica per la sua semplicità ed efficienza computazionale. Infatti a differenza dei metodi gerarchici che presentano una complessità temporale e spaziale pari a $O(n^2)$ [4], il K-means risulta avere una complessità temporale pari a $O(n^2)$ e una complessità spaziale pari a $O((n + k)d)$, dove n è il numero dei dati di input, k il numero dei cluster e d la dimensione dello spazio delle caratteristiche (feature), ovvero il numero di variabili o attributi che descrivono ciascun punto nel dataset [11].
L'idea di base dietro questo algoritmo di clustering consiste nel ricevere in input un parametro k e successivamente partizionare iterativamente un insieme di n osservazioni in modo tale che la variazione totale intra-cluster sia minimizzata. Questa variazione viene espressa attraverso una somma delle distanze euclidee quadratiche di tutte le coppie di osservazioni nel k-esimo cluster, divisa per il numero totale di osservazioni nel k-esimo cluster:

$$\min_{C_1,\dots,C_K} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

L'algoritmo si adatta perfettamente a cluster di forma sferica di dimensioni medio/piccole, ma non è in grado di confrontarsi con dataset che presentano forme non convesse o irregolari [12][13], poiché la minimizzazione della varianza intra-cluster, basata sulla distanza euclidea, tende a produrre cluster convessi per qualsiasi dataset di osservazioni considerate. Di seguito viene riportato un esempio reale di dataset con oggetti disposti a forma di mezzaluna [14]:



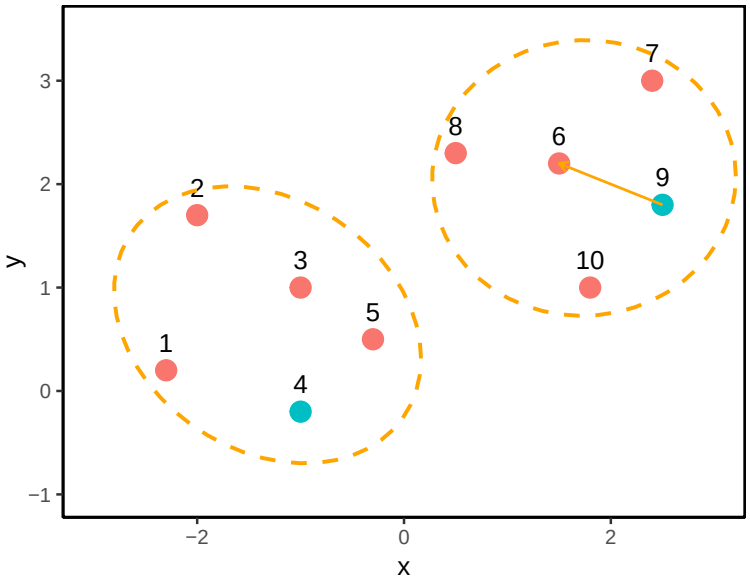
Come possiamo notare l'algoritmo, seguendo una direzione convessa, non è in grado di distinguere le due forme irregolari in due cluster, nonostante la mente umana possa facilmente separare i dati nelle forme più intuitive.
Un altro problema che contraddistingue questa metodologia è sicuramente la sensibilità verso valori estremi (outliers), che tendono ad "attrarre" nella propria direzione il centroide durante di calcolo delle medie delle singole osservazioni, modificando in questo modo la posizione del centroide nelle fasi finali dell'algoritmo. Per attenuare questa problematica, è stato ideato un altro algoritmo dal metodo K-medoids che prende il nome di *PAM* (Partitioning Arround Metods).

2.3.2 Metodo K-medoids

Questa metodologia di clustering a partizione, alternativa al K-means, prende come punto di riferimento del cluster il cosiddetto *medoide*, per definizione il punto più centrale di un cluster. Contrariamente al K-means, nel quale il punto più rappresentativo di un cluster veniva generato artificialmente come baricentro, il K-medoids assume sempre, come punto di riferimento, un punto appartenente alle osservazioni del dataset. Tuttavia, i due metodi condividono numerosi aspetti:

- Bisogna dichiarare in anticipo il numero dei cluster K e selezionare casualmente i medoidi dagli oggetti del dataset.
- Bisogna minimizzare la somma delle dissimilarità tra ciascuna osservazione e il suo punto di riferimento corrispondente (raccolte nella matrice delle distanze).

Come già accennato, l'algoritmo più rappresentativo di questa metodologia è il PAM. Una caratteristica fondamentale di questo algoritmo è la possibilità di utilizzare una qualsiasi metrica di distanza per calcolare la misura di dissimilarità tra due oggetti, questa componente ci consente una maggiore flessibilità quando le variabili di un dataset sono di natura differente. Per variabili numeriche, solitamente, si utilizza la distanza di Manhattan che, grazie alle sue proprietà, aumenta la robustezza dell'algoritmo, limitando la distorsione generata dagli outliers. Possiamo quindi intuire come gli step dei due algoritmi si assomiglino, tuttavia esiste una distinzione importante:
-Per ogni oggetto rappresentativo del cluster, l'algoritmo PAM seleziona un oggetto non rappresentativo e calcola il costo di scambiare i due elementi all'interno dello stesso cluster in relazione ai medoidi dei cluster rimanenti (swapping cost).
Immaginiamo di avere un dataset banale composto da dieci osservazioni e assegniamo casualmente, con K=2, i due medoidi:



La funzione di errore E sarà la seguente:

$$E = \sum_{k=1}^{K} \sum_{p \in C_j} |p - o_j|$$

Dove p è la generica osservazione del dataset e o_j è il medoide. Lo swipping cost S viene calcolato nel momento in cui decidiamo di voler modificare la posizione del medoide di un cluster (nel nostro caso il punto 9), assegnandolo ad un generico punto appartenente al medesimo cluster (per esempio il punto 6). La funzione swapping sarà quindi rappresentata come la differenza tra il tentativo di cambiare i medoidi ($E_(o_4, o_6)$) e la mia configurazione attuale ($E_(o_4, o_9)$):

$$S = E_(o_4, o_6) - E_(o_4, o_9)$$

Se il costo di swapping risulta negativo, allora conviene scambiare la posizione del medoide, nel nostro esempio, dal punto 9 al punto 6. Se il costo di swapping risulta positivo, il medoide corrente è considerato accettabile e non viene sostituito.
Questo processo viene iterato per tutti i punti selezionati casualmente nei K cluster.

Definiamo costo totale di swapping la somma di tutte le differenze nelle funzioni di costo che si ottengono in seguito alla riassegnazione degli oggetti non medoidi. Nel momento in cui questo valore smette di cambiare, allora l'algoritmo ha raggiunto la convergenza e termina automaticamente.

In sintesi, l'algoritmo PAM risulta essere un ottimo sostituto al K-means in termini di robustezza, tuttavia bisogna tener conto di un aspetto fondamentale: i costi computazionali. Infatti, la complessità dell'algoritmo per ogni iterazione è $O(k(n-k)^2)$. Quando il numero delle osservazioni n e il numero di cluster k aumentano, tale calcolo può risultare troppo dispendioso.

2.4 L'indice di silhouette

Dopo aver selezionato l'algoritmo più efficiente in funzione del compito richiesto, è indispensabile individuare il numero ottimale di cluster da generare per raccogliere al meglio i nostri dati. Questa scelta influisce direttamente sulla qualità dei risultati. Se il numero di cluster k risulta essere troppo basso, significa che il metodo non è stato in grado di separare le osservazioni in base alle loro reali differenze, perciò, si rischia di ottenere gruppi eterogenei. D'altra parte, se il numero risulta essere troppo elevato, si va in contro ad una suddivisione eccessiva dei dati, rendendo i cluster poco significativi per le nostre analisi.
L'indice che aiuta nella selezione del numero di gruppi e a valutare la qualità della classificazione è l'indice di Silhouette:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}, \quad s(i) \in [-1, 1]$$

Dove $s(i)$ è la distanza media dell'osservazione i dagli altri elementi del suo cluster, mentre $b(i)$ è la distanza media dell'osservazione i dagli altri elementi del cluster più vicino. Quando $s(i) \approx 1$ significa che l'unità è stata classificata bene; mentre con $s(i) \approx 0$ l'unità risulta essere un *bridge point*, ovvero un'unità nel mezzo di due cluster; infine se $s(i) \approx -1$ significa che l'unità è stata classificata male.
L'indice $s(i)$ tuttavia esprime la classificazione di una singola osservazione, quindi per valutare le prestazioni del nostro modello è necessario effettuare una media degli indici di tutte le osservazioni del dataset. Questa media prende il nome di Average silhouette width:

$$ASW = \frac{1}{n} \sum_{i=1}^{n} s(i)$$

Successivamente, si confrontano i valori dell'ASW per ciascun numero di cluster k con $(k = 1, 2, ..., N)$, dove N è il numero limite di cluster che noi imponiamo per l'analisi.

2.5 Clustering per variabili miste

Tradizionalmente, le tecniche di clustering sono state sviluppate per dati numerici. In questo breve excursus, abbiamo dunque esaminato i principali metodi di questa disciplina, proponendo soluzioni che ci permettono di affrontare problemi di natura numerica. Tuttavia, quando ci confrontiamo, come nel nostro caso, con un contesto economico, la realtà risulta spesso più complessa. È quindi necessario introdurre, in questa nuova sezione, metodi di clustering che siano in grado di operare su qualsiasi tipologia di variabile, sia essa numerica, categorica o una combinazione delle due. Un aspetto centrale sarà la scelta della metrica di distanza, poiché nei dati misti non è possibile applicare direttamente le distanze che abbiamo approfondito nello scorso capitolo, ma sarà indispensabile individuare una metrica alternativa.

2.5.1 Distanza di Gower

Un approccio diffuso per trattare i dati misti è la distanza di Gower come metrica di dissimilarità [15][16].
La distanza di Gower s_{ijk} tra due osservazioni i e j è definita come la media normalizzata delle differenze tra le variabili corrispondenti, in modo da poter confrontare variabili di natura diversa. A seconda del tipo di variabile, s_{ijk} viene calcolata come segue:

- Per le variabili numeriche si utilizza, come abbiamo già anticipato, il metodo di normalizzazione, in modo tale da modificare il dominio di appartenenza:

$$s_{ijk} = \frac{|x_{ik} - x_{jk}|}{R_k}$$

  dove R_k è il range (intervallo) della variabile k. La distanza sarà un valore compreso tra 0 e 1.

- Per le variabili categoriali, la distanza è definita come:

$$s_{ijk} = \begin{cases} 1, & \text{se } x_{ik} = x_{jk} \\ 0, & \text{se } x_{ik} \neq x_{jk} \end{cases}$$

- Per le variabili dicotomiche, a differenza di quelle categoriali, è importante considerare il valore informativo tra due istanze. In altre parole, le variabili dicotomiche comprendono caratteristiche che possono assumere valori "True" o "False", ma che vengono impiegate in contesti in cui solo l'intersezione della loro presenza ("True") fornisce un'informazione rilevante, mentre l'assenza non ha un significato informativo:

$$s_{ijk} = \begin{cases} 1 & \text{se } x_{ik} = \text{True AND } x_{jk} = \text{True} \\ 0 & \text{altrimenti.} \end{cases}$$

Il punteggio finale che determina la distanza tra due osservazioni eterogenee è semplicemente la media di tutti i punteggi per le caratteristiche disponibili:

$$S_{ij} = \frac{\sum_{k=1}^{N}(s_{ijk} \cdot \delta_{ijk})}{\sum_{k=1}^{N} \delta_{ijk}}.$$

Dove δ_{ijk} è un indicatore che serve a determinare se una determinata caratteristica k può essere confrontata tra le due osservazioni i e j ed assume i seguenti valori:

- $\delta_{ijk} = 1$ Se la caratteristica k è disponibile per entrambe le osservazioni i e j.

- $\delta_{ijk} = 0$ Se almeno una delle due osservazioni i o j non possiede la caratteristica k, quindi il confronto non può essere effettuato.

Il numeratore somma tutti i punteggi s_{ijk}, ma solo per le caratteristiche che possono essere confrontate ($\delta_{ijk} = 1$). Il denominatore, invece, conta quante caratteristiche sono effettivamente confrontabili tra i e j.
Dopo aver ottenuto il punteggio finale, è necessario convertirlo in una metrica

di distanza che rispetti una certa coerenza. Infatti il punteggio S_{ij} può risultare controintuitivo, dal momento che assume un valore pari a 1 quando la similarità è massima, mentre un valore pari a 0 quando la similarità è minima. La distanza tra due osservazioni i e j verrà, d'ora in avanti, definita ne seguente modo:

$$D_{ij} = \sqrt{1 - S_{ij}}.$$

La radice quadrata è stata volutamente inserita per garantire una proprietà fondamentale delle metriche di distanza che abbiamo già trattato, ovvero la disuguaglianza triangolare. Consideriamo tre osservazioni a, b e c:

$$\sqrt{1 - S_{ab}} + \sqrt{1 - S_{bc}} \geq \sqrt{1 - S_{ac}}$$

In alternativa si potrebbe usare anche:

$$D_{ij} = 1 - S_{ij}.$$

Nota come *dissimilarità di Jaccard*, la quale viene impiegata maggiormente per le variabili dicotomiche.

Capitolo 3

Caso Pratico

Nei capitoli precedenti è stata affrontata una panoramica teorica sul clustering e su tutti gli strumenti necessari per questa metodologia. Nella seguente sezione verrà presentato un caso pratico che consentirà di applicare concretamente le tecniche illustrate, utilizzando un dataset composto da variabili miste relativo ai comportamenti di acquisto di un gruppo di consumatori [17].

3.1 Data cleaning e implementazione

Per prima cosa è stato necessario effettuare una pulizia dei dati attraverso una serie di operazioni standard che garantissero l'affidabilità e la coerenza del dataset, per evitare potenziali distorsioni nell'analisi successiva. Le azioni effettuate sono state le seguenti:

- Rimozione dei dati mancanti: l'algoritmo PAM scelto per questa analisi non è in grado di gestire dati incompleti e l'imputazione dei valori avrebbe potuto introdurre bias nei risultati.
- Eliminazione delle variabili non rilevanti: alcune colonne nel dataset considerato rappresentavano dati che non avrebbero in alcun modo contribuito all'analisi di clusterizzazione della clientela, come ad esempio: "Customer

ID" come variabile univoca e "Satisfaction score", considerato non rilevante per l'analisi comportamentale in questo contesto.

- Gestione delle variabili categoriche (convertite in fattori per una memorizzazione più efficiente) e normalizzazione delle variabili numeriche.

- Riduzione del dataset a 40.000 osservazioni: il calcolo e l'archiviazione delle distanze tra tutte le coppie di punti richiedono una matrice quadrata di dimensione $N \times N$, che nel nostro caso, con $N > 40000$, avrebbe superato i limiti di memoria disponibili in R.

Terminata questa prima fase, si è passati alla determinazione del numero ottimale di cluster per l'algoritmo, utilizzando l'indice di silhouette come criterio di valutazione. Questo processo è stato condotto in modo iterativo, analizzando i valori dell'indice su un intervallo di possibili configurazioni di cluster (da 2 a 10), al fine di identificare la partizione che massimizza la coesione interna e la separazione tra i gruppi. Il risultato di questa analisi è stato rappresentato graficamente, consentendo di individuare facilmente il numero ottimale di cluster come il valore dell'ascissa corrispondente alla silhouette width più elevata che, come possiamo notare, è pari a 2:

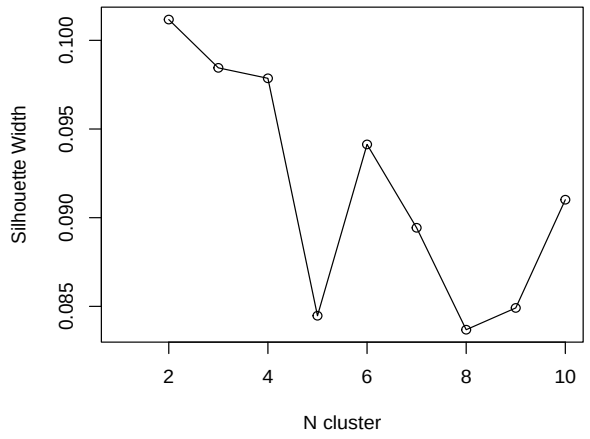


Figura 3.1: Silhouette

3.2 Risultati e analisi

Il passo successivo è stato applicare nuovamente l'algoritmo PAM, utilizzando il numero ottimale di medoidi individuato. Questa operazione ha consentito di ottenere i risultati statistici suddivisi nei due gruppi identificati, permettendo un'analisi più precisa delle caratteristiche distintive dei cluster, sotto forma di output riportato di seguito:

Cluster 1		**Cluster 2**	
age	**gender**	**age**	**gender**
Min: 12	Female: 4046	Min: 12	Female:16061
1st Qu.: 27	Male: 14096	1st Qu.: 27	Male: 5797
Median: 30		Median: 30	
Mean: 30.01		Mean: 29.97	
3rd Qu.: 33		3rd Qu.: 33	
Max: 47		Max: 48	
income	**education**	**income**	**education**
Min: 5006	Bachelor: 8598	Min: 5000	Bachelor: 3546
1st Qu.: 19652	College: 3864	1st Qu.: 14511	College: 12062
Median: 31875	HighSchool: 3771	Median: 23927	HighSchool: 4176
Mean: 30126	Masters: 1909	Mean: 25324	Masters: 2074
3rd Qu.: 40830		3rd Qu.: 35745	
Max: 50000		Max: 49997	
region	**loyalty_status**	**region**	**loyalty_status**
East: 2755	Gold: 1731	East: 9221	Gold: 2135
North: 3614	Regular: 11019	North: 4388	Regular: 13188
South: 3596	Silver: 5392	South: 4364	Silver: 6535
West: 8177		West: 3885	
purchase_amount	**purchase_frequency**	**purchase_amount**	**purchase_frequency**
Min: 1286	frequent: 3625	Min: 1204	frequent: 4407
1st Qu.: 6704	occasional: 5317	1st Qu.: 4947	occasional: 6585
Median: 10951	rare: 9200	Median: 8212	rare: 10866
Mean: 10563		Mean: 8837	
3rd Qu.: 14107		3rd Qu.: 12311	
Max: 26204		Max: 24217	
product_category	**promotion_usage**	**product_category**	**promotion_usage**
Beauty: 1010	Min: 0.0000	Beauty: 1006	Min: 0.0000
Books: 2804	1st Qu.: 0.0000	Books: 3126	1st Qu.: 0.0000
Clothing: 5661	Median: 0.0000	Clothing: 2387	Median: 0.0000
Electronics: 2994	Mean: 0.3026	Electronics: 8938	Mean: 0.3062
Food: 2856	3rd Qu.: 1.0000	Food: 3164	3rd Qu.: 1.0000
Health: 1854	Max: 1.0000	Health: 2153	Max: 1.0000
Home: 963		Home: 1084	

Per una migliore comprensione dei risultati sono stati generati grafici relativi alle variabili analizzate, in modo da facilitare il confronto tra i diversi gruppi identificati.

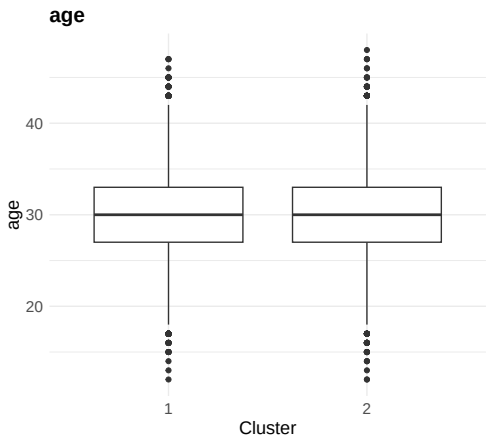


Figura 3.2: Age

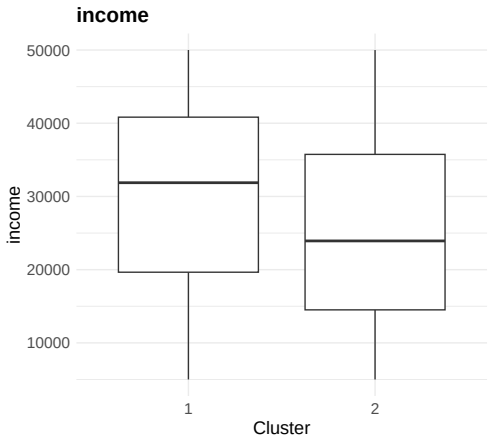


Figura 3.3: Education

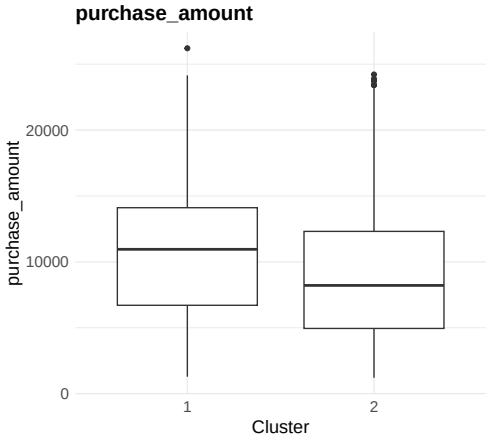


Figura 3.4: Gender

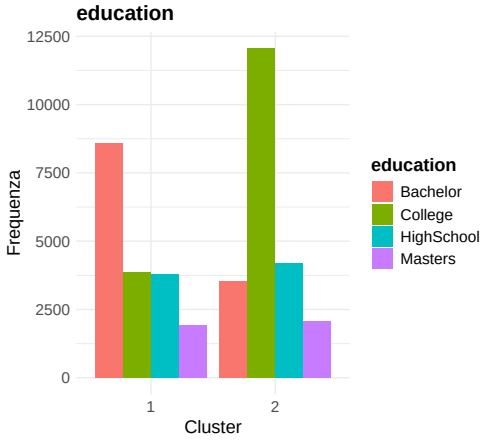


Figura 3.5: Income

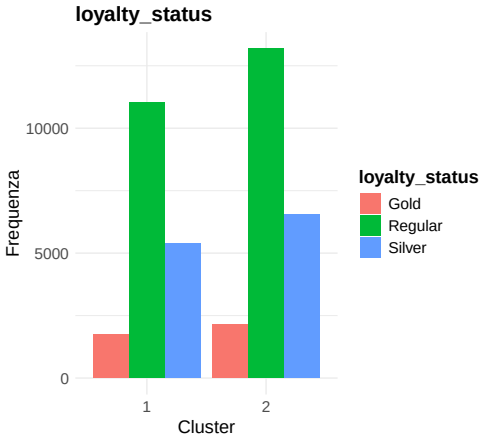


Figura 3.6: Loyalty Status

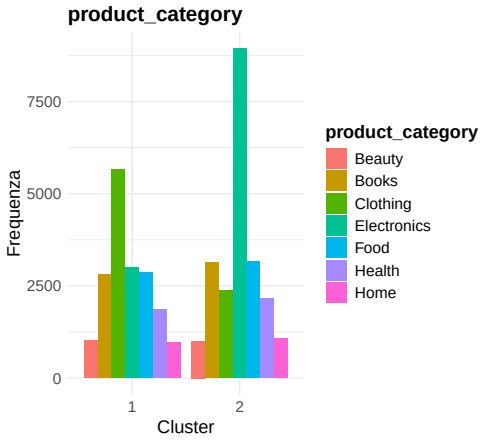


Figura 3.7: Product Category

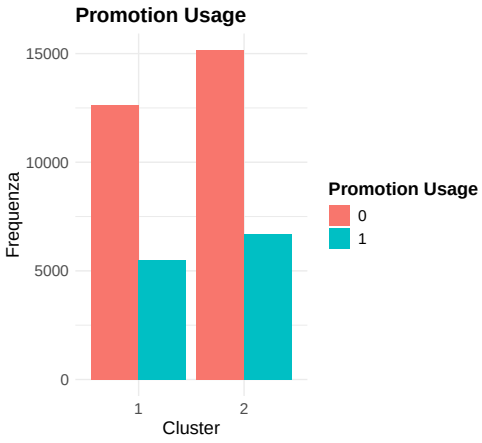


Figura 3.8: Promotion Usage

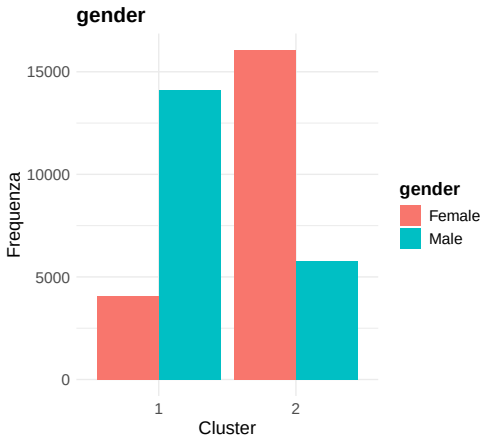


Figura 3.9: Purchase Amount

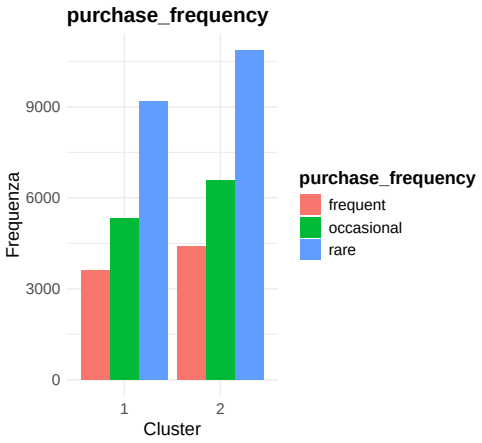


Figura 3.10: Purchase Frequency

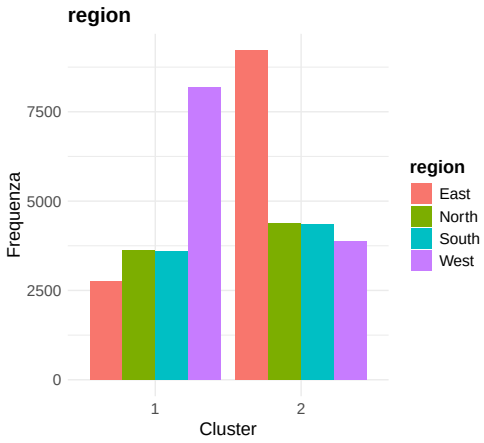


Figura 3.11: Region

3.2.1 Profilazione dei cluster

La partizione del dataset in due cluster ci consente di identificare alcune caratteristiche distintive. Il primo gruppo di individui (composto da 18142 unità) risulta avere maggiori disponibilità economiche, con un reddito annuo mediano nettamente più alto, pari a 31875 dollari, a differenza del secondo gruppo (di 21858 unità) che risulta avere un reddito mediano annuo pari a 23927 dollari. Questa statistica si riflette inevitabilmente sull'importo speso dal cliente ad ogni acquisto: 10951 dollari per il primo gruppo e 8212 dollari per il secondo. Le due variabili risultano dunque correlate positivamente con un indice di correlazione

$R = 0.948$, dove $R \in [-1, 1]$. Viene riportata di seguito una rappresentazione grafica della distribuzione tra le due variabili:

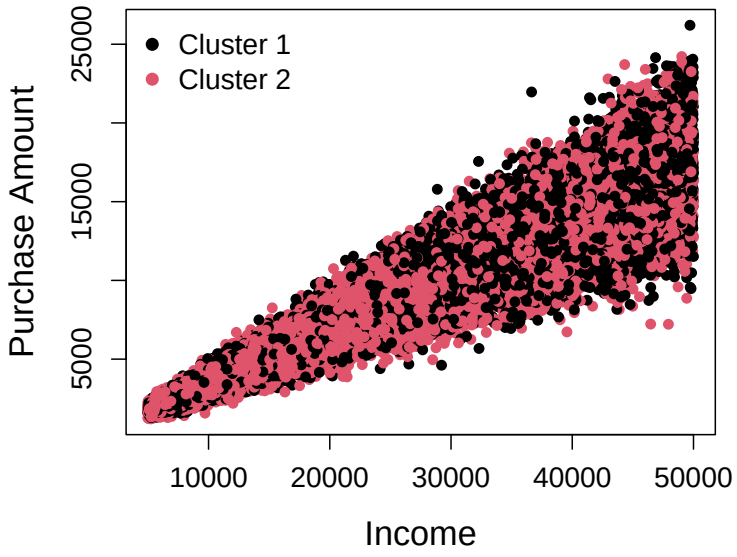


Figura 3.12: Correlazione tra reddito e importo speso per cliente

Le motivazioni alla base di questa differenziazione economica possono essere ricercate nella variabile "education" che rappresenta il livello di istruzione dei clienti. Il primo gruppo è composto in larga parte da individui con titoli di studio più elevati: con una percentuale di laureati pari al $47,4\%$, significativamente superiore rispetto al secondo gruppo dove prevalgono titoli di studio inferiori, come individui diplomati di scuola superiore pari al 55.2%, mentre i laureati si attestano al 16.2%.

Oltre alle differenze economiche e di tipo educativo, emergono anche variazioni significative legate alla residenza geografica e al genere all'interno dei due cluster. Infatti possiamo notare come il cluster 1 sia formato prevalentemente da individui dell'Ovest (45.1%) e da individui maschi (77.7%). Al contrario, il Cluster 2 è caratterizzato da una maggiore concentrazione di individui provenienti dalle regioni Est (42,2%) e da una forte presenza femminile (73,5%).

Le restanti variabili come l'età, l'utilizzo delle promozioni, lo stato di fedeltà e la frequenza di acquisto del cliente non risultano essere statisticamente significative, poiché in percentuale le differenze intercluster non superano l'1%.
Un aspetto invece più interessante riguarda le categorie di prodotti acquistati. A primo impatto, sembrerebbe che il Cluster 1, composto principalmente da uomini con un reddito più elevato, abbia come prodotto principale l'abbigliamento. Al contrario il Cluster 2, composto in gran parte da donne e con un reddito mediano inferiore, sembra preferire l' elettronica. Tuttavia, questa interpretazione basata su dati assoluti può risultare fuorviante. Se si analizzano le percentuali relative di acquisto per genere, emerge un quadro diverso:

Categoria	**Cluster 1**		**Cluster 2**	
	Female	**Male**	**Female**	**Male**
Beauty	4.79%	5.79%	4.84%	3.93%
Books	13.89%	15.91%	15.19%	11.83%
Clothing	**46.42%**	**26.84%**	13.45%	3.92%
Electronics	6.77%	19.30%	**35.93%**	**46.42%**
Food	14.38%	16.13%	15.22%	12.42%
Health	9.07%	10.55%	10.22%	8.81%
Home	4.67%	5.49%	5.14%	4.45%

Tabella 3.1: Distribuzione percentuale dei prodotti acquistati per genere

Nel cluster 1, nonostante gli uomini acquistino complessivamente più abbigliamento in termini assoluti, una quota molto più alta di donne si orienta verso questa categoria: 46,4% delle donne contro 26,8% degli uomini. Allo stesso modo, nel cluster 2, dove l'elettronica sembra dominare in termini assoluti, solo il 35,9% delle donne acquista prodotti elettronici, contro un significativo 54,6% degli uomini. Questa differenziazione si manifesta anche in senso opposto, dove prevale una percentuale maggiore di uomini che acquistano prodotti elettronici nel cluster 2 e di donne che preferiscono abbigliamento nel cluster 1.

Conclusioni

In questo breve elaborato è stato proposto un semplice metodo partizionale di clustering che riuscisse a confrontarsi con variabili di natura mista e ad estrarre insight utili per ipotetiche analisi di marketing orientate verso una segmentazione della clientela.
I risultati ottenuti rappresentano solo un primo passo verso una comprensione più approfondita delle dinamiche comportamentali dei clienti, evidenziando come variabili apparentemente secondarie, come il genere, la residenza e il livello di istruzione, possano influire in modo significativo sulle preferenze di acquisto.
In questa sede, tuttavia, si è voluto dimostrare quanto questo strumento statistico, seppur contraddistinto da rigorose premesse matematiche, possa rivelarsi estremamente versatile nell'analisi di grandi quantità di dati.

Per una visione completa del codice R si rimandi al seguente link:
`https://github.com/LeoSuriano/Personal-Portfolio/blob/main/Script_ClusteringMixed.R`

Bibliografia

1 J Irani, N Pise, M Phatak (2016) *Clustering Techniques and the Similarity Measures used in Clustering: A Survey*

2 Mahamed G.H. Omrana, Andries P. Engelbrechtb and Ayed Salmanc (2007) *An overview of clustering methods*,

3 Paolo Giordani, Maria Brigida Ferraro, Francesca Martella (2020) *An Introduction to Clustering with R*,

4 Rui Xu (2005) *Survey of Clustering Algorithms*

5 Ali Seyed Shirkhorshidi , Saeed Aghabozorgi, Teh Ying Wah (2015) *A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data*

6 Jain AK, Murty MN, Flynn PJ (1999) *Data clustering: a review*

7 Wilson D, Martinez T. (1997) *Improved heterogeneous distance functions*

8 Tedo Vrbanec, Ana Mestrovic (2021) *Relevance of Similarity Measures Usage for Paraphrase Detection*

9 Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2013) *An Introduction to Statistical Learning with Applications in R*

10 (2009) *Distances between Clustering, Hierarchical Clustering 36-350, Data Mining*

11 Mohiuddin Ahmed, Raihan Seraj and Syed Mohammed Shamsul Islam (2020) *The k-means Algorithm: A Comprehensive Survey and Performance Evaluation*

12 Boomija (2008), *Comparison of Partition Based Clustering Algorithms*

13 Gummadi Venkata Nikhil Sai, Robby Aulia Tubagus, Vasala Rohith, Haritha Donavalli (2024) *Comparative Analysis of Kmeans Technique on Non Convex Cluster*

14 Cornellius Yudha Wijaya (2024) *Why K-Means Failed at Non-Convex Shape Data-NBD*

15 *https://crispinagar.github.io/blogs/gower-distance.html*

16 Gower (1971) *A general coefficient of similarity and some of its properties*

17 https://www.kaggle.com/datasets/sanyamgoyal401/customer-purchases-behaviour-dataset/data