Chapter 1

Clustering

In questo capitolo mostreremo il comportamento degli algoritmi di clustering KMeans, DBSCAN e Hierarchical applicati al nostro insieme di dati.

Il dataset utile per questa fase è stato ottenuto eliminando gli attributi categorici `education, sex` e `status` e l'attributo `credit_default` poichè il clustering rientra tra gli addestramenti di tipo non supervisionato.

1.1 KMeans

Il miglior parametro k con cui eseguire KMeans è stato stimato calcolando la *SSE* variando k su un range da 2 a 20, con metrica di distanza euclidea. In figura 1.1 sono riportati i risultati ottenuti. Si nota che il cambio di pendenza della curva si trova quando k vale 7. A partire da ciò abbiamo eseguito l'algoritmo per $k \in [3, 10]$, reiterando per ogni passo 50 volte KMeans per evitare che la scelta casuale dei centroidi influenzasse i risultati e calcolando in questo caso anche l'indice di Silhouette per i cluster trovati. In seguito a queste analisi abbiamo constatato come i cluster ottenuti per $k = 4$ siano i più significativi, la nostra considerazione è stata rafforzata dal fatto che il clustering ottenuto per $k = 4$ presenta il valore dell'indice di Silhouette più alto trovato.

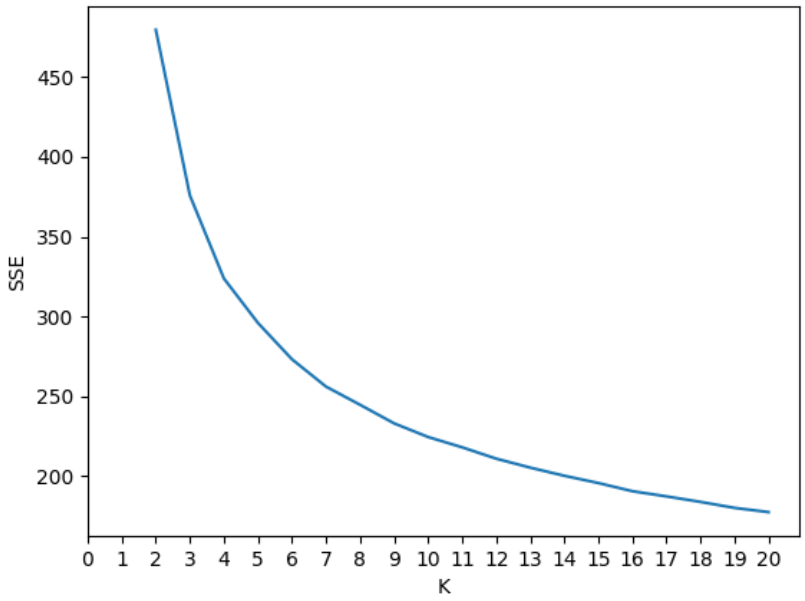

Figure 1.1: SSE

Di seguito analizziamo i cluster trovati, assegnando loro un nome che contraddistingua le caratteristiche di ogni cluster.

Senza rischio Cluster costituito da 3032 persone. Pagano in modo puntuale le loro spese ogni mese senza registrare ritardi e compiono solitamente spese di bassa entità. Non costituiscono alcuna minaccia per la banca.

Piccoli pagatori Gruppo di 4832 persone. Sono soliti fare un uso abbastanza intensivo del revolving credit per pagare le loro spese. Anch'essi registrano spese di bassa entità e non commettono gravi ritardi nel pagamento, per questo motivo rientrano tra clienti credibili per la banca.

Grandi pagatori Cluster di 1083 persone. Come per i *piccoli pagatori*, anch'essi sono soliti usare la modalità di pagamento rateizzata anche se le loro spese sono generalmente molto elevate. Non si registrano comunque grossi ritardi ed é per questo che rientrano comunque tra il gruppo di clienti credibili.

Ritardatari Gruppo formato da 1053 persone. Registrano spese di media entità ma al contrario degli altri tre gruppi si verificano gravi ritardi nei pagamenti. Sono il cluster di persone che, infine, finisce in credit default e non sono clienti credibili per la banca.

In figura 1.2 si sono plottate una selezione delle coordinate dei centroidi, mostrando chiaramente dove i quattro cluster trovati differiscono maggiormente.

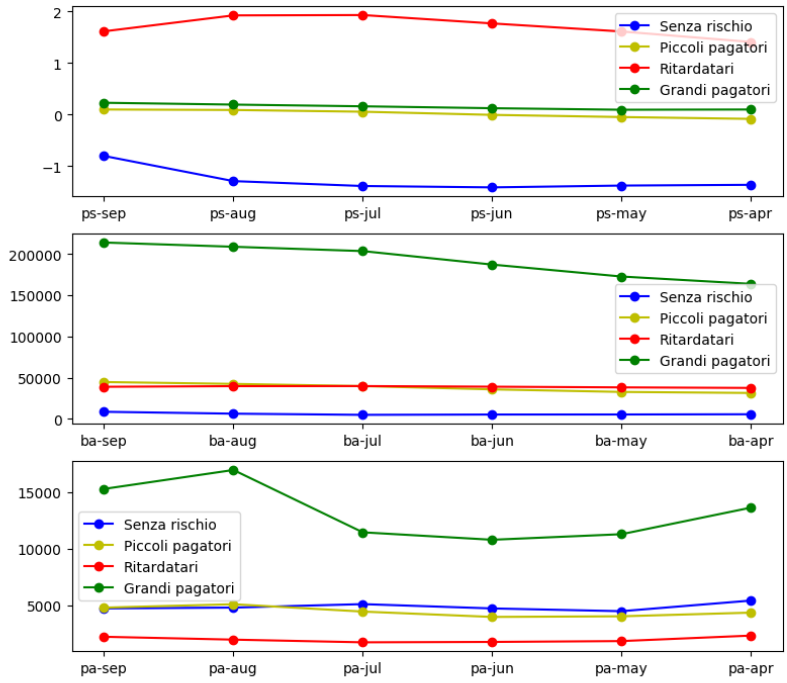

Figure 1.2: Caratteristiche dei cluster

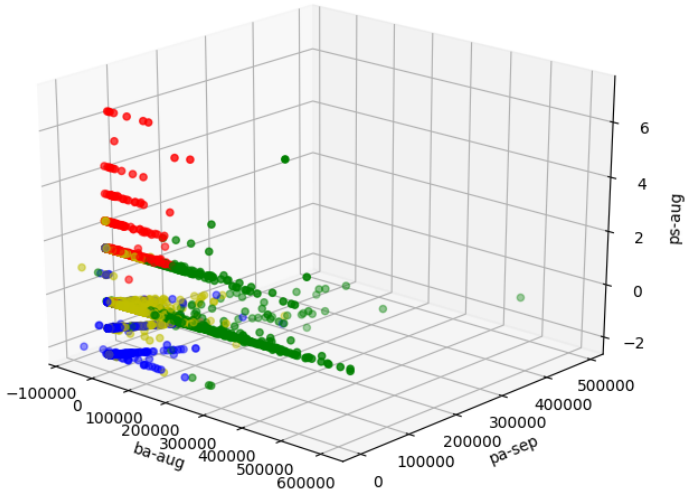

Figure 1.3: Distribuzione dei cluster su un pagamento mensile

Infine si è plottato su un un grafico 3D (figura 1.3 la distribuzione dei cluster su un pagamento mensile (in questo caso si e' preso il pagamento del mese di agosto) ed abbiamo notato che la distribuzione è rispettata per ogni terna di attributi validi costruibili sull'insieme dei mesi disponibili. In particolare il cluster dei *ritardari* si posiziona sempre a valori molto elevati di `payment status`, mentre vale esattamente il contrario per il cluster dei *senza rischio*. Le tre dimensioni scelte per l'esempio sono `billing amount august`, `payment status august` e `payment amount september`.

Infine riportiamo una tabella contenente la media e la deviazione standard di ogni cluster individuato. Per ragioni di spazio per gli attributi `payment status`, `payment amount` e `billing amount` mostriamo i valori degli ultimi due mesi.

Cluster	ps-sep	ps-aug	pa-sep	pa-aug	
Senza rischio	$-0.8(\pm 1.0)$	$-1.3(\pm 0.7)$	$4.7k(\pm 12.0k)$	$4.8k(\pm 14.3)$	
Piccoli pagatori	$0.09(\pm 0.71)$	$0.08(\pm 0.71)$	$4.8k(\pm 10k)$	$5k(\pm 13.3k)$	
Grandi pagatori	$0.22(\pm 0.75)$	$0.19(\pm 0.71)$	$15k(\pm 35k)$	$16k(\pm 56k)$	
Ritardatari	$1.60(\pm 1.18)$	$1.90(\pm 1.03)$	$2k(\pm 3.4k)$	$1.8k(\pm 3.0k)$	
	ba-sep	**ba-aug**	**limit**	**age**	
Senza rischio	$8.7k(\pm 21k)$	$6.4k(\pm 16k)$	$215k(\pm 126k)$	$36(\pm 8)$	
Piccoli pagatori	$44k(\pm 39k)$	$42k(\pm 35k)$	$130k(\pm 113k)$	$35(\pm 9)$	
Grandi pagatori	$213k(\pm 93k)$	$208k(\pm 88k)$	$281k(\pm 115k)$	$37(\pm 8)$	
Ritardatari	$38k(\pm 35k)$	$39k(\pm 36k)$	$79k(\pm 68k)$	$34(\pm 8)$	
	sex	**status**	**education**	**default**	
Senza rischio	F	Single	University	17%	
Piccoli pagatori	F	Single	University	18%	
Grandi pagatori	F	Single	University	19%	
Ritardatari	F	Single	University	61%	

Si noti come il gruppo dei ritardatari ha un limite imposto dalla banca molto più basso rispetto agli altri cluster. Segno che la banca ha valutato in modo abbastanza buono i profili nella scelta di concessione del credito.

1.2 DBSCAN

L'algoritmo DBSCAN è stato eseguito su un dataset modificato rispetto alla esecuzione del KMeans. Questo è stato dovuto per permettere all'algoritmo di funzionare al meglio. Dopo molteplici esperimenti infatti i migliori risultati per DBSCAN sono stati ottenuti su un dataset composto dai soli attributi `payment status` di ciascun mese. Ciò è in linea con la teoria in quanto DBSCAN ha problemi su dati con un numero troppo elevato di dimensioni. Per stimare i parametri ottimali dell'algoritmo si sono utilizzati i grafici del k-dist[2], utilizzando la distanza di Manhatthan. I valori che ci hanno permesso di ottenere un (primo) risultato soddisfacente sono stati $\varepsilon = 0.30$ e $minPoints = 350$. Il primo risultato è stato l'individuazione da parte di DBSCAN di due cluster di dimensioni molto diverse tra loro ma con un significato molto forte.

Senza rischio Composto da 8597 persone, formato da persone che non costituiscono alcun rischio in quanto i loro valori di payment status sono costantemente sotto lo 0.

Ritardatari Composto da sole 454 unità. Ciò che lo contraddistingue è un costante ritardo nei pagamenti dei propri debiti. Queste sono le persone che maggiormente rappresentano un rischio di perdita di credito per la banca.

Data la notevole disparità di dimensioni dei due cluster, abbiamo deciso di eseguire nuovamente DBSCAN sul cluster dei senza rischio al fine di verificare se anch'esso avrebbe trovato la conformazione dei tre cluster individuati da

KMeans. Abbiamo rieseguito l'algoritmo con gli stessi parametri e il risultato è stato l'individuazione di tre cluster con una notevole quantità di rumore.

Rifinanziatori Cluster composto da 3271 persone. Sono contraddistinti da un uso intensivo del revolving credit. Le spese sono di entità media. Per questo motivo sospettiamo sia un merge tra i *Grandi pagatori* e *Piccoli pagatori* trovati da KMeans.

Senza rischio Cluster composto da 635 persone. Sono le persone che pagano sempre in orario, ogni mese rispettano la scadenza e ripagano in pieno il loro debito. Hanno caratteristiche molto simili all'omonimo di KMeans.

No consumption Cluster composto da 703 persone. Questa è la novità rispetto a KMeans. Questo cluster incorpora tutte le persone che non fanno uso del loro credito, evidenziato da un valore di *No consumption* molto ripetuto.

In conclusione, DBSCAN in qualche modo valida i risultati ottenuti da KMeans, in quanto due diversi algoritmi, su due selezioni diverse del dataset hanno trovato dei risultati molto simili, fatta eccezione per il cluster dei *No consumption*.

Di seguito riportiamo una tabella riassuntiva delle caratteristiche dei cluster trovati:

Cluster	**ps-sep**	**ps-aug**	**pa-sep**	**pa-aug**
Senza rischio	$-1.0(\pm 0.0)$	$-1.0(\pm 0.0)$	$6.8k(\pm 14.0k)$	$6.4k(\pm 12k)$
No consumption	$-2.0(\pm 0.71)$	$-2.0(\pm 0.71)$	$4.8k(\pm 14k)$	$5k(\pm 14k)$
Rifinanziatori	$0.0(\pm 0.0)$	$0.0(\pm 0.0)$	$6k(\pm 13k)$	$5.9k(\pm 16k)$
Ritardatari	$1.72(\pm 0.68)$	$2.10(\pm 0.33)$	$2.5k(\pm 3.9k)$	$2.5k(\pm 4.3k)$
	ba-sep	**ba-aug**	**limit**	**age**
Senza rischio	$6.3k(\pm 12k)$	$6.4k(\pm 12k)$	$221k(\pm 123k)$	$36(\pm 8)$
No consumption	$7k(\pm 23k)$	$6k(\pm 19k)$	$248k(\pm 122k)$	$36(\pm 7)$
Rifinanziatori	$93k(\pm 88k)$	$89k(\pm 85k)$	$161k(\pm 128k)$	$35(\pm 9)$
Ritardatari	$52k(\pm 56k)$	$53k(\pm 56k)$	$92k(\pm 68k)$	$35(\pm 9)$
	sex	**status**	**education**	**default**
Senza rischio	F	Single	Graduate school	14%
No consumption	F	Single	Graduate School	11%
Rifinanziatori	F	Single	University	9%
Ritardatari	F	Single	University	68%