Università di Pisa

Anno Accademico 2014 – 2015

Corso di Laurea Magistrale in
Business Informatics

Data Mining

Advanced topics on Data Mining and case studies

Profiling Purchases

Indice

Introduzione

L'obiettivo di questo elaborato è di illustrare alcune analisi svolte su un insieme di transazioni riguardanti i clienti di una catena di supermercati.
L'analisi si concentra su quattro negozi (identificati come appartenenti alla catena di supermercati COOP) e su un intervallo temporale compreso tra il 2 gennaio 2010 e il 31 dicembre 2010. Gli scopi di questo progetto sono:

1. Definire un **profilo utente** per ogni cliente appartenente all'insieme dei dati;
2. Analizzare i due negozi più importanti con lo scopo di comprendere quali siano i **profili di acquisto utente tipici** che si svolgono in questi.

Al fine di rispettare il primo obiettivo, la strada seguita è stata quella di individuare le abitudini di acquisto *{segmento, giorno_settimana, negozio_id}* più rilevanti per ciascun cliente, sia nei confronti della frequenza di acquisto, sia per quanto riguarda la percentuale d'incidenza sull'importo speso. Una volta determinati i risultati di questa prima analisi, è stata eseguita la segmentazione dei clienti rispetto ai loro profili e verificato quali di questi fossero i profili maggiormente diffusi in ogni negozio selezionato.

1. Esplorazione dei dati

1.1 Descrizione

Nella fase di esplorazione dei dati sono state analizzate 6 tabelle: **Articolo.csv**, **Clienti.csv**, **Orario.csv**, **Marketing.csv**, **Data.csv**, **Venduto.csv**. Di seguito è proposta la descrizione di ciascuna tabella e degli attributi che la compongono. Per ogni attributo è riportata una spiegazione sintetica e, nel caso di dati mancanti, il numero di osservazioni non contenenti un valore per l'attributo analizzato.

Articolo.csv contiene 21.026 osservazioni, ognuna delle quali descrive un articolo venduto. Sono presenti 22 attributi:

- ARTICOLO_ID, ARTICOLO_ID_1: rappresentano entrambi l'identificativo dell'articolo.
- COD_MKT_ID: è un riferimento alla tabella Marketing.csv e rappresenta l'identificativo del mercato cui appartiene l'articolo.
- COD_ART: è un attributo univoco che identifica l'articolo.
- DES_ART: è una stringa che contiene la descrizione per un determinato articolo.
- MARCA_INIZIALE, MARCA: sono due attributi di testo che rappresentano rispettivamente l'iniziale della marca e il nome completo della marca dell'articolo. Per 5.289 articoli, tali attributi non sono specificati.
- SOTTOMARCA: attributo testuale che riporta la sottomarca dell'articolo. Per 8.801 articoli la sottomarca non è disponibile.

- UDM, QTA: il primo è un attributo di testo che rappresenta l'unità di misura dell'articolo, il secondo è un attributo numerico ed indica l'ammontare per una determinata unità di misura.
- FL_COOP, FL_BIO, FL_CELIACI, FL_DOP, FL_IGP, FL_TIPICO, FL_SOLIDALE, FL_DOCG, FL_CARRELLO: sono tutti attributi *flag* che indicano se quella particolare caratteristica è presente o meno per l'articolo.
- PRES_MKT: campo testuale che descrive una particolare caratteristica dell'articolo. Per 11.027 articoli tale attributo assume valore "Descrizione non disponibile".
- RILEVANZA: assume diversi attributi testuali che descrivono la rilevanza dell'articolo. Abbiamo riscontrato 9.213 articoli con rilevanza uguale a "Descrizione non disponibile".
- CAPOSTIPITE_ID: campo numerico intero che riporta l'identificativo del capostipite dell'articolo.

Clienti.csv è una tabella formata da 48.712 clienti (osservazioni) per ognuno dei quali sono presenti 8 differenti colonne:

- CLIENTE_ID: è il codice identificativo del cliente.
- FASCIA_ETA: attributo categorico che identifica la fascia di età del cliente. Per 2.400 clienti tale attributo non è specificato. Inoltre ci sono 11 clienti con fascia di età *110-120* e 228 con fascia di età *100-110* che potrebbero rappresentare valori anomali, tuttavia, non essendo di intralcio alla nostra analisi, abbiamo preferito mantenerli come tali.
- SESSO: attributo testuale che può assumere valore "Uomo" o "Donna" e per 2.400 clienti assume valore "ND".
- STATO_CIVILE: attributo di testo che descrive lo stato civile del cliente. Per 2.440 clienti tale attributo non è disponibile.
- PROFESSIONE, TITOLO_STUDIO: descrivono rispettivamente la professione e il titolo di studio del cliente. In entrambi i casi, per 2.400 clienti assumono valore "Non disponibile".
- FASCIA_ANNO_SOCIO: campo di testo che indica il decennio nel quale il cliente è diventato socio del negozio.
- FL_INVIO_RIVISTA: campo *flag* che identifica se al cliente viene inviata o meno la rivista del negozio.

Orario.csv contiene 795 righe e 8 colonne:

- ORARIO_ID, ORARIO_ID_1: codici univoci che identificano un orario.
- ORARIO, ORA, MINUTO: formano una gerarchia in cui ORARIO descrive l'orario completo (*es:* 18:36) mentre ORA e MINUTO riportano rispettivamente i valori interi rispettivi dell'ora e del minuto (*es:* ora = 18, minuto = 36).

- FASCIA_ORARIA: è un campo testuale che riporta la fascia oraria in cui è avvenuto l'acquisto.
- FASCIA_ORARIA_N: è un valore intero (da 1 a 5) che identifica una determinata FASCIA ORARIA.

Marketing.csv contiene 2.974 osservazioni per ognuna delle quali sono stati utilizzati 17 attributi differenti:

- COD_MKT_ID, COD_MKT_ID_1: codici univoci della tabella Marketing.csv.
- COD_MKT: ulteriore codice univoco per la tabella Marketing.csv.
- AREA → MACROSETTORE → SETTORE → REPARTO → CATEGORIA → SOTTOCATEGORIA → SEGMENTO: descrivono la gerarchia di un prodotto in ordine decrescente, dal più generale al più specifico.
- COD_AREA, COD_MACROSET, COD_SETT, COD_REP, COD_CATEG, COD_SUBCATEG, COD_SEGMENTO: attributi numerici che forniscono un identificatore numerico per l'attributo cui si riferiscono.

Data.csv è una tabella formata da 341 osservazioni e 16 colonne di attributi:

- DATA_ID, DATA_ID_1: codici univoci che identificano una data.
- GIORNO: attributo di testo che descrive un giorno nel formato "GG MM" (*es:* 9 MARZO).
- GIORNO_N: attributo numerico che identifica il giorno del mese (*es:* 9)
- MESE, MESE_N: il primo è un attributo testuale che identifica il mese (*es:* MARZO), mentre il secondo è il numero associato a tale mese (*es:* 3).
- ANNO: attributo intero che assume sempre valore 2010.
- GIORNO_SETTIMANA, GIORNO_SETTIMANA_N: il primo è un campo testuale che rappresenta il giorno della settimana (*es*: MARTEDI), il secondo è un intero che contiene il numero del giorno della settimana (*es*: 2).
- TRIMESTRE, TRIMESTRE_N: il primo è un attributo categorico che identifica un trimestre (*es*: GENNAIO-MARZO), mentre il secondo è il valore numerico associato a quel trimestre (*es*: 1).
- DATA: attributo che descrive la data completa "GG-MM-AA" (*es*: 09-MAR-10).
- SETTIMANA_ANNO, SETTIMANA_COMMERCIALE: il primo è un attributo intero che identifica la settimana dell'anno, il secondo è un attributo numerico intero che identifica la settimana commerciale.
- PERIODO_ID: attributo numerico intero che identifica un periodo dell'anno.
- TIPOLOGIA: attributo di testo che assume solo due valori ovvero "Feriale" o "Festivo".

Venduto.csv rappresenta la tabella dei fatti, per questo motivo è quella più numerosa tra tutte, ed è formata da 18.134.427 osservazioni, ognuna delle quali è descritta da 11 colonne:

- DATA_ID, ORARIO_ID, CLIENTE_ID, ARTICOLO_ID, COD_MKT_ID: rappresentano riferimenti agli attributi che identificano ciascuna delle 5 tabelle, rispettivamente, Data.csv, Orario.csv, Clienti.csv, Articolo.csv e Marketing.csv.
- NEGOZIO_ID: identifica il negozio in cui è avvenuto un acquisto. Assume quattro differenti valori (70, 101, 102, 103).
- IMPORTO: ammontare di denaro speso per l'acquisto di un certo articolo in relazione alla quantità di pezzi o di peso acquistata.
- QTA_PEZZI, QTA_PESO, VOLUME: descrivono rispettivamente il numero di pezzi acquistati di un certo articolo, il peso e il volume relativo a quell'articolo.
- N_SCONTRINI: è un attributo numerico che identifica la transazione, ovvero lo scontrino in cui si è verificato l'acquisto di un determinato articolo.

1.2 Selezione dei dati

In fase iniziale di analisi è stato scelto di non considerare le tabelle **Articolo.csv**, **Clienti.csv** e **Orario.csv** poiché contenenti informazioni non rilevanti al nostro scopo. L'attenzione principale è stata, infatti, rivolta alla tabella **Venduto.csv** contenente le transazioni dei clienti e nella quale la distribuzione di ciascun attributo risultavano variegate e con un alto *range* di possibili valori, ma in linea con la natura dei dati e non affette, quindi, né da valori anomali, né da rumore. Dalla tabella abbiamo eliminato i seguenti quattro attributi poiché non rilevanti al fine dell'analisi:

- ORARIO_ID: eliminato perché troppo dettagliato nei confronti dell'analisi condotta;
- QTA_PESO e VOLUME: eliminati poiché non rilevanti al fine dei risultati. E' stato notato come sia l'attributo QTA_PEZZI, sia l'attributo IMPORTO, fornissero congiuntamente il dettaglio desiderato, senza che vi fosse la necessità di comprendere anche i due attributi qui considerati;
- ARTICOLO_ID: come dettaglio di analisi rispetto al prodotto venduto, è stato preferito l'attributo SEGMENTO (proveniente dalla tabella **Marketing.csv**)

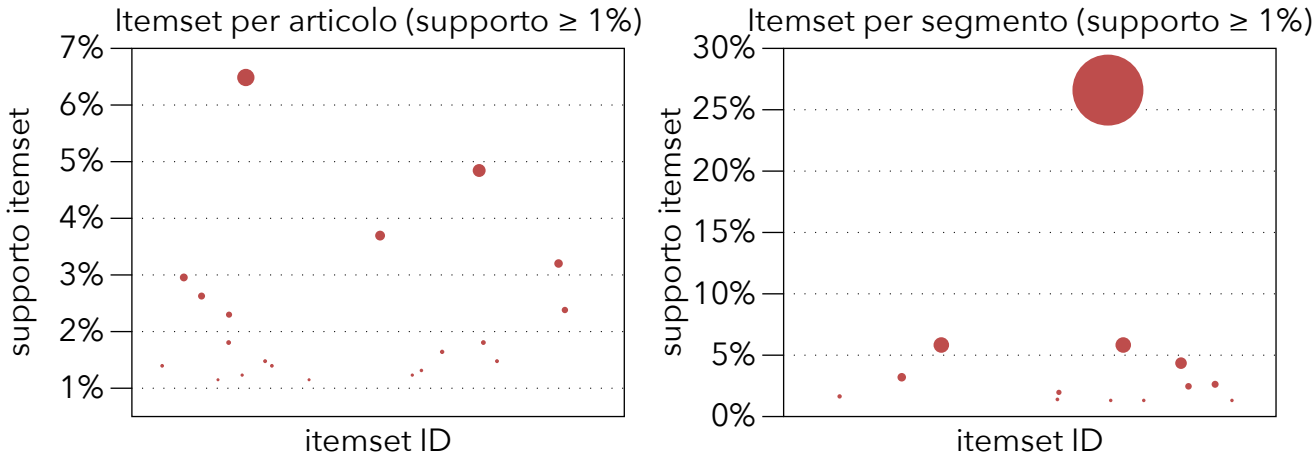

Figura 1 - Itemset generati per articolo_id e per segmento (cliente selezionato n.448736). La grandezza dei punti nei grafici è impostata secondo il supporto di ciascun itemset (raggio 1pt per il valore minimo).

poiché rappresentante il giusto livello di dettaglio nella gerarchia del prodotto. Considerare il singolo articolo avrebbe, infatti, frammentato i profili utente tra innumerevoli insiemi {*articolo*, giorno settimana, negozio}, ciascuno con bassa significatività. A tal proposito si può osservare la Figura 1 rappresentante gli itemset con supporto superiore all'1% nei due casi appena descritti, ovvero, creati secondo l'articolo o secondo il segmento. Come si può facilmente notare, nel primo caso (itemset per articolo) il supporto si disperde in diversi itemset, abbassando così la significatività dei risultati stessi, mentre, nel secondo caso (itemset per segmento), si concentra in alcuni itemset rendendo così i risultati maggiormente indicativi.

Rispetto alle altre tabelle, invece, per quanto riguarda la tabella **Data.csv,** è stato selezionato unicamente l'attributo GIORNO_SETTIMANA. È stato, infatti, ritenuto più opportuno incentrare l'analisi sul singolo giorno della settimana piuttosto che rispetto a intervalli temporali più ampi quali mese o trimestre. Questo poiché un possibile impiego per l'analisi svolta potrebbe essere, ad esempio, quello di ottimizzare settimanalmente la presenza dei prodotti sugli scaffali. Impieghi simili ma condotti su scadenze mensili o trimestrali, sono stati scartati poiché troppo limitati e privi di fondamento per dati appartenenti a un solo anno di esercizio.
Infine, dalla tabella **Marketing.csv** è stato selezionato solo l'attributo SEGMENTO per quanto spiegato in precedenza nell'eliminazione dell'attributo ARTICOLO_ID dalla tabella **Venduto.csv**.

1.3 Pulizia dei dati

Al fine dell'analisi condotta, è stato opportuno eseguire una selezione sui clienti atta a individuare i più importanti ed escludere i restanti. L'importanza di un cliente è stata individuata attraverso quattro parametri:

1. **Intervallo in giorni trascorsi tra la prima e l'ultima spesa effettuata**: sono stati scartati i clienti per cui tra la data della prima spesa e la data corrispondente all'ultima spesa effettuata, passassero tra 0 e 31 giorni (31 incluso). Tali clienti, infatti, avendo eseguito tutti i propri acquisti concentrandosi in un breve periodo, non sono stati considerati interessanti al fine dell'analisi condotta, in particolare poiché incentrata su un solo anno d'esercizio (2010). Clienti simili sarebbero stati rilevanti se lo studio si fosse svolto su più anni, permettendo così un'analisi verso le abitudini di acquisto anche per questi clienti. Dalla Figura 2 si può osservare, inoltre, come gli acquisti dei clienti qui considerati si concentrino maggiormente nei periodi di festività o vacanza e siano quindi affetti da possibile stagionalità. L'eliminazione di tali consumatori ha quindi permesso la pulizia efficace di una considerevole parte dei dati, scartando 9.877 clienti, ovvero il 20,3% del totale.

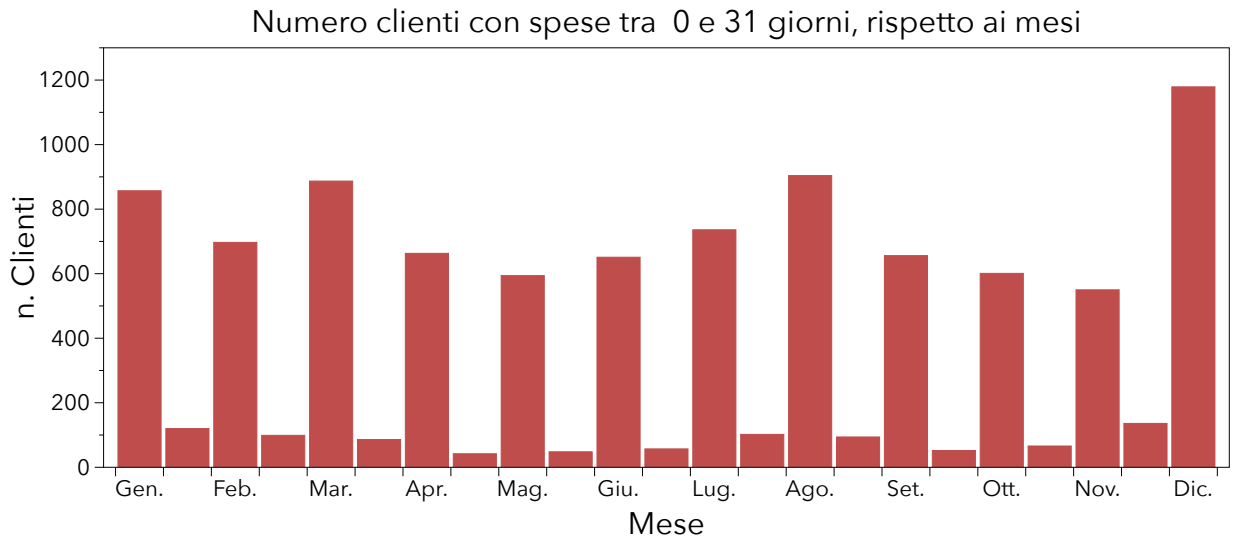

Figura 2 - Distribuzione, secondo i mesi, dei clienti eliminati per acquisti compiuti tra 0 e 31 giorni.

2. **Importo totale**: la distribuzione dell'importo totale, visibile in Figura 3, presentava una forte densità nella parte sinistra dell'istogramma, sintomo che buona parte dei clienti avesse effettuato acquisti per un importo complessivo poco rilevante (compreso tra 2,32 e 155,81 Euro). È stato quindi individuato come valore soglia per l'importo, quello riferito al primo quartile della distribuzione, ovvero corrispondente a 155,81 Euro. In seguito l'analisi è proceduta verso lo studio della quantità totale dei pezzi acquistati.
3. **Quantità totale**: una situazione molto simile a quella verificatasi nel caso dell'importo totale, è stata riscontrata anche nella distribuzione della quantità totale, permettendo in questo caso di stabilire un valore soglia pari a 82, ovvero corrispondente al primo quartile della distribuzione, visibile in Figura 4.

   A questo punto si è proceduto eseguendo una selezione congiunta, considerando sia l'importo sia la quantità totale e selezionando, quindi, unicamente i clienti che superassero contemporaneamente entrambe le soglie imposte in precedenza per tali valori. In questo modo sono stati eliminati altri

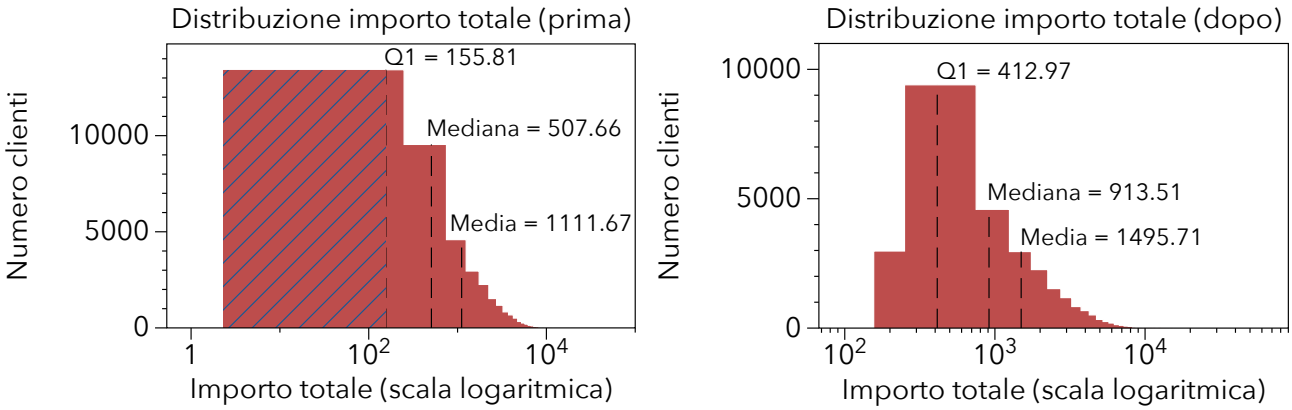

Figura 3 - Distribuzione dell'importo rispetto al numero dei clienti. Il primo grafico rappresenta la distribuzione prima della pulizia effettuata per "Importo totale", dove è stata evidenziata la parte dei dati cancellata dalla pulizia, stessa, mentre il secondo illustra la distribuzione risultante dall'eliminazione dei clienti con un importo ≤ 155,81.

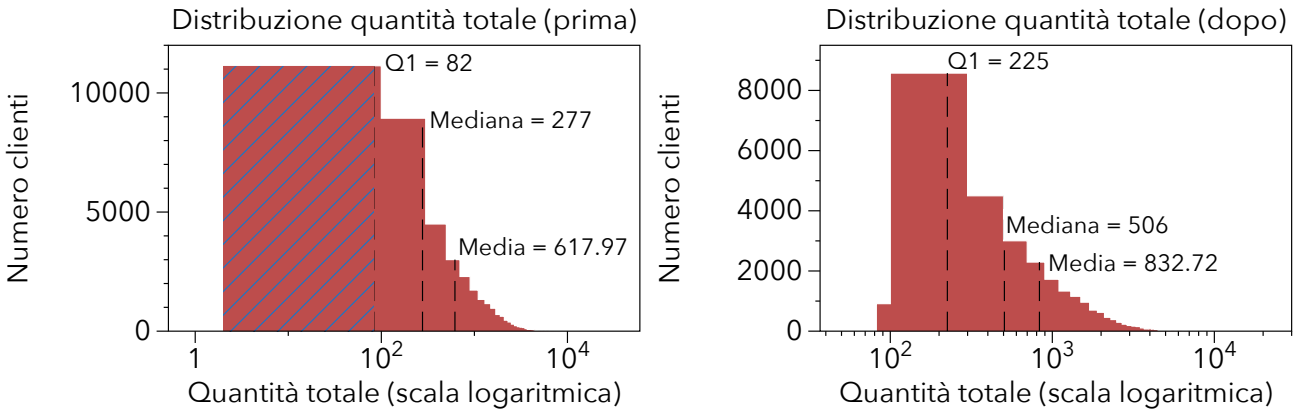

Figura 4 - Distribuzione della quantità rispetto al numero dei clienti. Il primo grafico rappresenta la distribuzione prima della pulizia effettuata per "Quantità totale"; è stata evidenziata la zona contenente i dati eliminati dalla pulizia stessa. Il secondo, invece, illustra la distribuzione risultante dall'eliminazione dei clienti con una quantità ≤ 81.

10591 clienti che, aggiunti ai 9877 eliminati in precedenza, hanno portato la percentuale dei clienti esclusi dall'analisi al 42% dei clienti iniziali.

4. **Numero spese**: per quest'ultimo indicatore non è stato tenuto conto della distribuzione dei dati, bensì è stata posta l'ipotesi che i clienti rilevanti al fine dell'analisi fossero quelli ad aver eseguito più di dieci spese nell'arco di tutto l'anno 2010. In tale modo sono stati esclusi altri 3987 clienti.

In conclusione, i quattro punti utilizzati per la pulizia dei dati hanno portato all'esclusione di 24.455 clienti, corrispondenti al 50,2% di quelli iniziali. Percentuale che può sembrare elevata, poiché superiore alla metà dei dati originari, ma necessaria per indirizzare l'analisi verso i soli clienti più rilevanti e offrire quindi risultati maggiormente indicativi.

1.4 Preparazione dei dati

Con l'estromissione del 50,2% dei clienti, sono state cancellate, dalla tabella **Venduto.csv**, tutte le transazioni appartenenti a questi, portando l'insieme totale delle transazioni da 18.134.427 righe a 17.170.926 osservazioni (riduzione del 5%).
In seguito sono state unificate le tabelle **Venduto.csv**, **Marketing.csv** e **Data.csv**, con i soli attributi selezionati in precedenza.

Come si evince dalla selezione dei dati, e per quanto spiegato in precedenza, è stato scelto di incentrare l'analisi a livello di segmento, per quanto riguarda gli articoli acquistati, e a livello di giorno della settimana, per quanto riguarda la data di acquisto. Dalla tabella risultante da tali operazioni, infine, sono stati eliminati gli attributi DATA_ID, COD_MKT_ID e N_SCONTRINI al fine di agevolare l'analisi da compiere in seguito.

La tabella sulla quale è stata eseguita l'analisi atta all'individuazione dei profili utente presentava quindi i seguenti attributi:

- CLIENTE_ID;
- NEGOZIO_ID;
- SEGMENTO;
- GIORNO_SETTIMANA;
- IMPORTO;
- QTA_PEZZI.

2. Users Profiling

2.1 Modellazione

Per individuare i profili utente è stato creato un modello iterativo che analizzasse un solo cliente alla volta, creando per ciascuna transazione un itemset costituito da {NEGOZIO_ID, SEGMENTO, GIORNO_SETTIMANA}. In seguito sono stati determinati gli itemset più rilevanti per ciascun cliente secondo due criteri:

- **Itemset più frequenti**: la frequenza di ciascun itemset è stata calcolata attraverso l'algoritmo Apriori impostando, come grandezza minima dell'itemset una costante pari a 3, e come supporto minimo un valore di 0,01.
- **Incidenza percentuale sull'importo totale**: gli itemset tra loro uguali sono stati aggregati sommando gli importo corrispondenti a ciascuno di essi. Per ogni aggregazione così costruita è stata quindi calcolata la percentuale di incidenza sull'importo totale speso dal cliente come *importo corrispondente all'aggregazione/importo totale*.

Per entrambi i criteri, i dati sono stati ordinati in senso decrescente, nel primo caso per il supporto, nel secondo per l'incidenza percentuale. A questo punto, per estrarre gli itemset più indicativi per ciascun cliente, è stato creato un codice in linguaggio *java* che ha permesso di selezionare automaticamente gli itemset fino alla soglia più opportuna per ciascun caso. In questo modo non è stato necessario impostare un numero costante di itemset da selezionare per ogni cliente, bensì, per ciascuno di questi è stato creato un profilo rappresentato dalla numerosità di itemset più opportuna.
Dalla Figura 3, si può notare come avvenga la selezione degli itemset in modo da selezionare solo quelli più significativi: nel momento in cui gli itemset tendono ad assumere il medesimo supporto o la stessa incidenza percentuale sull'importo totale, allora la scelta degli itemset rilevanti si arresta.

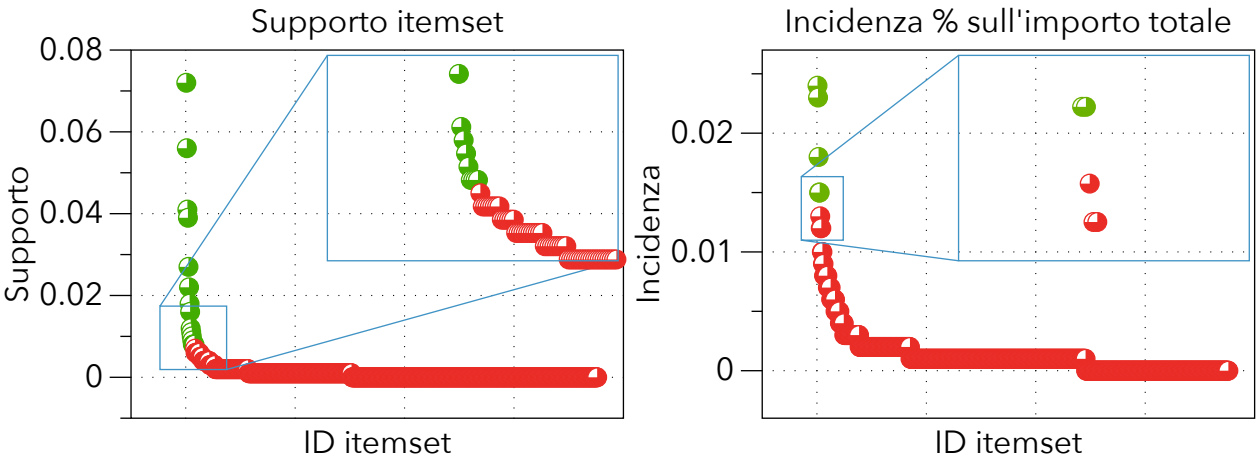

Figura 5 - Itemset (cliente 328104) disposti per supporto e per incidenza percentuale sull'importo totale. I punti verdi indicano che l'itemset è stato selezionato come rilevante, mentre quelli rossi indicano che l'itemset è stato scartato dalla selezione.

Nel caso in cui un itemset fosse scelto sia per un alto supporto, sia per un'alta incidenza sull'importo totale, è stato selezionato un'unica volta riportando entrambi gli indicatori. Nel caso in cui un itemset fosse selezionato solo rispetto a uno dei due indicatori, e mancasse il secondo, è stato inserito un valore nullo per il valore mancante.

2.2 Descrizione dei risultati

Avendo creato un profilo per ciascuno dei 24.455 clienti e non potendo quindi riportare la descrizione di ognuno, si illustrano di seguito quattro diversi profili utente per lunghezza e caratteristiche. Ciascun profilo utente è riportato in ordine decrescente per il supporto e, a parità di questo, in ordine decrescente per l'incidenza percentuale sull'importo totale:

- **Cliente 112.456**: il profilo utente qui considerato ha una lunghezza pari a 15 itemset. Come si può osservare nella Tabella 1, le prime 7 righe totalizzano un supporto pari a circa il 94% del totale e una incidenza sull'importo totale di poco superiore al 70%. Inoltre, si distribuiscono lungo tutta la settimana ma restando concentrate sul solo segmento Affettato, escluso il Mercoledì nel quale il cliente acquista anche nel segmento Naturali. Si noti, anche, come gli acquisti che compongono il profilo avvengano sia nel negozio 70, sia nel negozio 102. Da ciò si può quindi osservare come il cliente in esame abbia un profilo ben distribuito per quanto riguarda i giorni della settimana ma altamente concentrato per quanto riguarda il segmento di acquisto.

ID	Itemset	Supporto	% Importo
1	{Affettato, Lunedì, 102}	0,181	0,154
2	{Affettato, Venerdì, 102}	0,156	0,133
3	{Naturali, Mercoledì, 102}	0,150	0,029
4	{Affettato, Sabato, 102}	0,144	0,128
5	{Affettato, Mercoledì, 102}	0,125	0,111
6	{Affettato, Giovedì, 102}	0,103	0,088
7	{Affettato, Martedì, 102}	0,078	0,064
8	{Bambini, Martedì, 70}	0,006	0,112
9	{Affettato, Giovedì, 70}	0,006	0,005
13	{Extralarge, Venerdì, 70}	-	0,045
10	{Junior, Giovedì, 102}	-	0,031
14	{Normali, Venerdì, 70}	-	0,019
11	{Altri tipici, Lunedì, 102}	-	0,009
12	{Bresaola, Lunedì, 102}	-	0,009
15	{S. Daniele, Venerdì, 70}	-	0,009

Tabella 1 - Profilo utente, cliente 112.456.

Da questo punto di vista si può quindi affermare che un profilo utente così definito assicura un'alta percentuale di precisione rispetto al segmento preferito dal cliente e una valida indicazione sul fatto che l'utente non si concentri su alcuni giorni bensì sia abituato a effettuare acquisti ben distribuiti lungo tutto l'arco della settimana.

- **Cliente 807.935**: in questo caso sono stati individuati nove itemset come parte del profilo utente. Dalla Tabella 2, è possibile osservare come già solo le prime tre righe descrivano le abitudini del consumatore poiché, sommando i relativi supporti, si ottiene un supporto totale dell'80%. Allo stesso modo si può notare come i tre primi itemset racchiudano complessivamente anche la parte più consistente dell'importo speso dal cliente, totalizzano, infatti, il 58% dell'incidenza sull'importo totale speso. Nel complesso gli itemset descrivono una alta variabilità nei segmenti preferiti dall'utente ma anche una valida abitudine dell'utente nell'effettuate le proprie spese tutte nello stesso negozio e concentrate principalmente nel Martedì.

ID	Itemset	Supporto	% Importo
1	{Vetro superiore ml.330, Giovedì, 70}	0,388	0,279
2	{Vetro superiore ml.330, Venerdì, 70}	0,290	0,208
3	{Vetro superiore ml.330, Martedì, 70}	0,125	0,090
4	{Naturali, Martedì, 70}	0,018	0,005
5	{Normale, Martedì, 70}	0,015	0,013
6	{Frizzanti, Giovedì, 70}	0,008	0,014
7	{<500 ml, Martedì, 70}	0,008	0,011
8	{Amari, Martedì, 70}	-	0,019
9	{Amari, Venerdì, 70}	-	0,019

Tabella 2 - Profilo utente, cliente 807.935.

- **Cliente 388.952**: in questo caso il profilo utente, osservabile nella Tabella 3, nonostante sia abbastanza corposo, è ampiamente spiegato dalla prima riga. Di fatto, l'acquisto nel segmento *monoc. <110gr* il *venerdì* nel negozio *70*, racchiude in se il 57% del supporto totale e quasi il 40% della spesa. Confrontato con gli altri itemset mostra come solo tale abitudine descriva con buona probabilità una forte tendenza d'acquisto del cliente. In aggiunta, si può notare come il profilo sia distribuito prevalentemente sul Venerdì (undici itemset su sedici) e unicamente nel negozio 70.

ID	Itemset	Supporto	% Importo
1	{Monoc. <110gr, Venerdì, 70}	0,570	0,363
2	{Pluric. 110-390gr, Venerdì, 70}	0,120	0,070
3	{Lavorazioni interne, Venerdì, 70}	0,050	0,094
4	{Pasto completo, Venerdì, 70}	0,047	0,085
5	{Monoc. <110gr, Giovedì, 70}	0,043	0,026
6	{Alta qualità, Venerdì, 70}	0,010	0,007
7	{Normale, Venerdì, 70}	0,008	0,013
8	{Lavorazioni interne, Martedì, 70}	0,007	0,011
9	{ Monoc. <110gr, Mercoledì, 70}	0,006	0,002
10	{Lavorazioni interne, Giovedì, 70}	0,005	0,008
11	{Pluric. 110-390gr, Giovedì, 70}	0,005	0,003
12	{Antirughe/antietà/ristrutturanti, Venerdì, 70}	0,005	0,069
13	{Altri senza risciacquo, Venerdì, 70}	-	0,007
14	{Blended, Venerdì, 70}	-	0,007
15	{Cartoons, Venerdì, 70}	-	0,007
16	{Fissatori dentiere, Venerdì, 70}	-	0,007

Tabella 3 - Profilo utente, cliente 388.952.

- **Cliente 714.091**: in questo caso è qui presentato un profilo utente molto sintetico e mirato verso due soli segmenti, un singolo giorno e un unico negozio. I due itemset selezionati descrivono un supporto di oltre il 65% e un'incidenza percentuale sull'importo totale del 30% complessivo. Sintomo, dunque, che le abitudini di acquisto del cliente siano rilevanti quanto a segmenti, giorno e negozio rispetto alla quantità acquistata ma relativamente variegate, e quindi non del tutto indicative, riguardo all'importo speso; tale aspetto deriva, infatti, dal fatto che le regole qui individuate non spiegano il 70% dell'importo totale speso.

ID	Itemset	Supporto	% Importo
1	{Cannellini, Sabato, 70}	0,329	0,150
2	{Ceci, Sabato, 70}	0,329	0,150

Tabella 4 - Profilo utente, cliente 714.091.

2.3 Valutazione dei risultati

Analizzando l'insieme dei profili utente, è emerso come tutti descrivessero adeguatamente le abitudini di acquisto dei diversi clienti. Nel complesso, è stato riscontrato che i differenti profili fossero simili, per numerosità d'itemset, supporto e percentuale d'incidenza sull'importo totale, ai profili dei quattro clienti illustrati in precedenza. Inoltre, è emerso che, in media, per ciascun cliente siano stati selezionati tra i cinque e i sei itemset per descrivere il rispettivo profilo utente (la media si attesta a 5,48). Riteniamo dunque i risultati ottenuti ampiamente soddisfacenti.

3. Analisi dei negozi

3.1 Preparazione dei dati

Per affrontare l'analisi riguardante i negozi, sono stati individuati i due negozi più rilevanti. Tale ricerca è stata eseguita sia riguardo all'importo incassato dai singoli negozi, sia al numero di spese avvenute in ciascun negozio. Per entrambi gli indicatori, i due negozi più importanti, sui quali è stata incentrata l'analisi, sono stati quelli identificati dai codici 70 e 101.

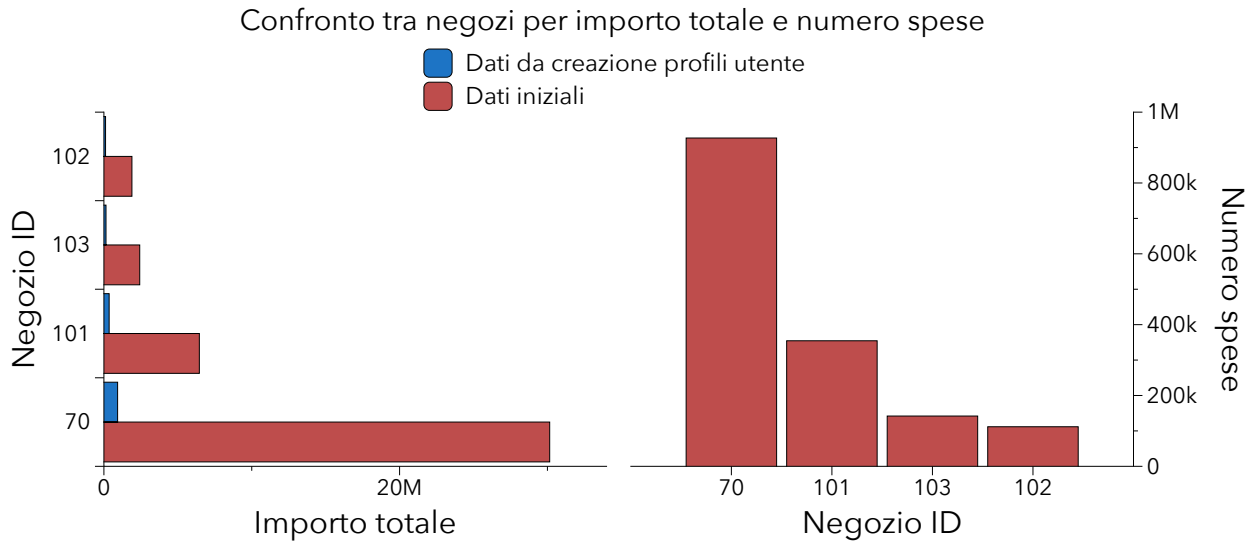

Figura 6 – Confronto tra i negozi. Nel grafico a sinistra, sono riportati sia gli importi calcolati dai dati completi, sia quelli calcolati dai soli profili utente. Si noti come in entrambi i casi, i due negozi maggiormente rilevanti sono il negozio 70 e il negozio 101. Nel grafico a destra, invece, è riportato il numero di spese effettuate nei singoli negozi. Anche in questo caso i negozi con maggior numero di spese sono il numero 70 e il numero 101.

Utilizzando come punto di partenza la tabella dei profili utente ottenuta in precedenza, da questa sono state eliminate due colonne, ovvero sia quella riferita al supporto, sia quella riferita all'incidenza percentuale sull'importo totale di ciascuna abitudine d'acquisto, poiché non necessarie al fine dell'analisi. La tabella risultante sulla quale è stato, quindi, svolto lo studio indirizzato verso i due negozi selezionati, presentava i seguenti quattro attributi:

- CLIENTE_ID;
- SEGMENTO;
- NEGOZIO_ID;
- GIORNO_SETTIMANA.

In seguito, per analizzare i due negozi, sono state separate le righe appartenenti all'uno o all'altro negozio in due tabelle differenti. Dalla tabella riguardante il negozio 70, sono state rimosse le righe inerenti ai quattro segmenti che per frequenza e perché rappresentanti un segmento di mercato tipico delle abitudini alimentari dei consumatori italiani, non avrebbero offerto un risultato di particolare rilevanza

nell'analisi del negozio. I quattro segmenti così selezionati e la percentuale dei clienti che ha acquistato in ciascuno di questi sono stati i seguenti:

- NATURALI (54%);
- LAVORAZIONI INTERNE (52%);
- NORMALE (34%);
- PARZIALMENTE SCREMATO (24%).

Attraverso tale esclusione sono stati eliminati 504 clienti (3% del totale) poiché aventi un profilo esclusivamente legato ai quattro segmenti eliminati. Inoltre, il profilo utente di 361 consumatori (2% del totale) è venuto a coincidere con quello di altri clienti dai quali precedentemente si differenziavano grazie ai segmenti esclusi.
Per le medesime ragioni prima indicate rispetto al negozio 70, dalle righe inerenti al negozio 101, sono state eliminate quelle relative agli stessi segmenti selezionati in precedenza, riportanti ora però percentuali differenti:

- LAVORAZIONI INTERNE (50%);
- NATURALI (41%);
- NORMALE (21%);
- PARZIALMENTE SCREMATO (18%).

L'esclusione dei quattro segmenti ha portato all'eliminazione di 560 clienti (9% del totale), e alla modifica sostanziale del profilo di 245 clienti (4% del totale), portandoli ad avere un profilo uguale a quello di clienti dai quali prima si differenziavano.
Va qui precisata una nota rispetto alla pulizia dei dati appena descritta: date le basse percentuali sia per i clienti così eliminati sia per i clienti ai quali l'eliminazione dei segmenti ha causato una netta modifica nel profilo clienti, si ritiene che tale azione non abbia inciso in modo significativo sulla segmentazione dei profili utenti, ovvero, non abbia alterato considerevolmente la composizione dei cluster, rispetto a quella che sarebbe stata la composizione se non fossero stati omessi dall'analisi i quattro segmenti prima elencati.

3.2 Modellazione

Nella segmentazione dei clienti è stato preferito svolgere l'analisi sulla base di due sole variabili, piuttosto che rispetto alle tre variabili originarie (CLIENTE_ID, SEGMENTO, GIORNO_SETTIMANA). Tuttavia, essendo tutte e tre di fondamentale importanza, sono state individuate tutte le possibili coppie di valori SEGMENTO, GIORNO_SETTIMANA e a ciascuna coppia così individuata è stato associato un valore univoco COPPIA_ID che la identificasse così nell'analisi. In questo modo ogni riga della tabella poteva essere quindi identificata unicamente tramite i due attributi CLIENTE_ID e COPPIA_ID.

Cliente	Segmento	Giorno_settimana	Coppia_ID
802	Bianchi	Sabato	4573
802	Classica	Sabato	4643
1123	Classica	Sabato	4643
2607	Frizzanti	Venerdì	5681
2607	Bianchi	Venerdì	4573

Tabella 5 – Tabella di esempio riportante i valori di COPPIA_ID associati a ciascuna riga.

Cliente	4573	4643	5681
802	1	1	0
1123	0	1	0
2607	1	0	1

Tabella 6 – Matrice Pivot costruita dalla tabella di esempio a lato e secondo la struttura utilizzata.

In seguito è stata creata una matrice Pivot riportante per ogni riga un cliente e per ogni colonna un valore COPPIA_ID. In ciascuna cella è stato inserito il valore 1 nel caso in cui il cliente corrispondente alla riga avesse effettuato un acquisto nel giorno della settimana e nel segmento corrispondente all'ID della coppia in colonna, e 0 altrimenti (si veda a tal proposito l'esempio riportato in Tabella 5 e Tabella 6).
Nella tabella pivot così costruita ogni riga è stata convertita in un valore binario. Tale valore è stato in seguito utilizzato, attraverso la misura *Cosine Distance*, per calcolare le distanze tra una riga (e quindi un cliente) e tutte le possibili altre righe (e quindi tutti i possibili altri clienti), necessarie per l'utilizzo dell'algoritmo K-Means. Si è quindi proceduto all'avvio di un ciclo di calcolo nel quale è stato iterativamente impostato un valore per la variabile *k* dell'algoritmo, da un minimo di 1 a un massimo di 7. Per ciascun ciclo, e dunque per ogni possibile valore assunto dalla variabile k, è stato calcolato l'errore SSE e quindi, una volta terminati tutti i possibili valori, è stata tracciata la curva corrispondente all'andamento dell'errore SSE. Da questa è stato infine scelto il valore k corrispondente al punto di flesso di tale curva e proceduto di conseguenza all'analisi dei cluster calcolati dall'algoritmo K-Means con il valore *k* impostato secondo quanto scelto.

3.2.1 Negozio 70: scelta numero dei cluster

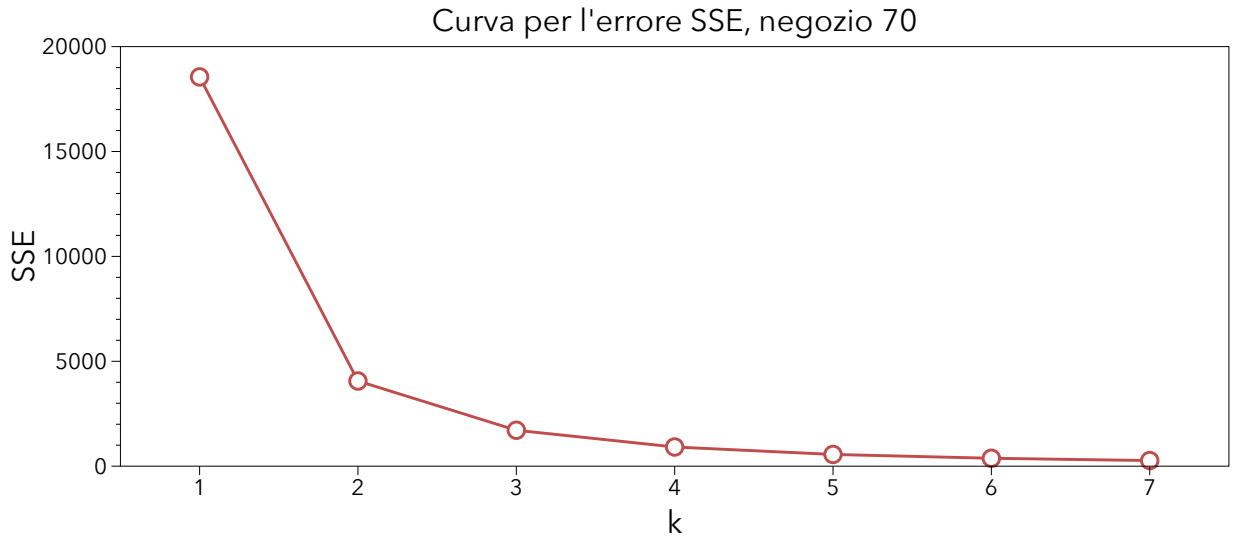

Figura 7 – Curva dell'errore SSE per il negozio 70. Il valore individuato per la costante *k* è stato *k* = 3.

Dalla Figura 7 è possibile osservare come il punto di flesso si trovi in corrispondenza del valore $k = 3$. Da tale specifica, quindi, i cluster risultanti si sono distribuiti secondo quanto segue:

- **Cluster 1**: 10.134 clienti;
- **Cluster 2**: 3858 clienti;
- **Cluster 3**: 4562 clienti;

Dall'esito del risultato, si può notare come il primo cluster sia predominante rispetto agli altri due, poiché raggruppa il 55% dei clienti totali. I due cluster restanti, invece, raccolgono rispettivamente il 21% e il 24% dei clienti.

3.2.2 Negozio 101: scelta numero dei cluster

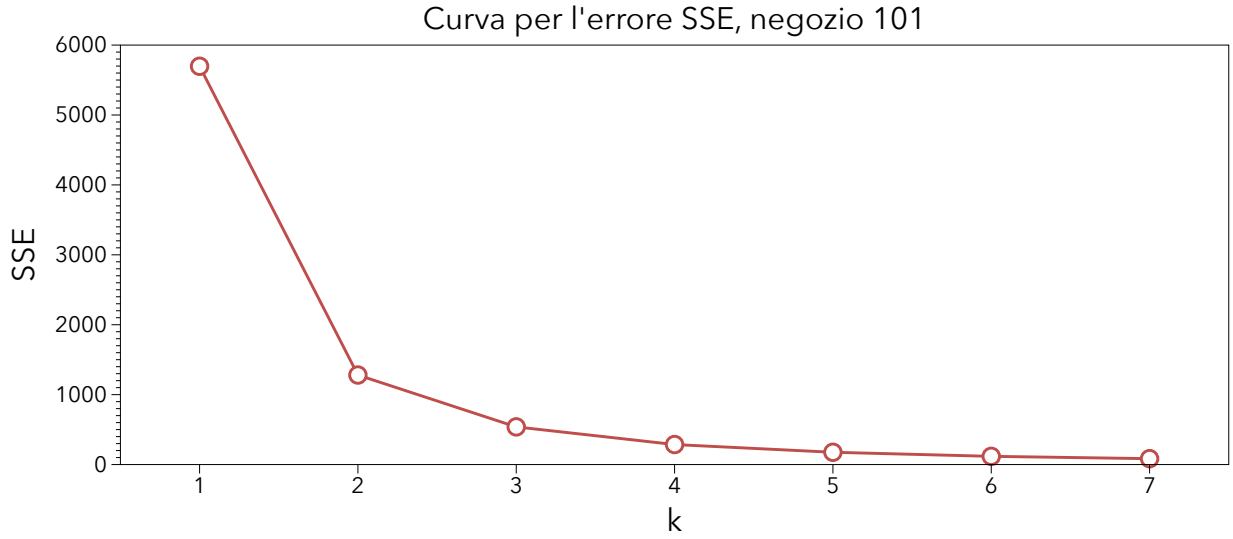

Figura 8 - Curva dell'errore SSE per il negozio 101. Il valore individuato per la costante k è stato $k = 3$.

Dalla Figura 8 è possibile osservare come, anche in questo caso, il punto di flesso si possa individuare in corrispondenza del valore $k = 3$. In questo modo sono stati generati tre cluster, distribuiti nel seguente modo:

- **Cluster 1**: 4103 clienti;
- **Cluster 2**: 654 clienti;
- **Cluster 3**: 873 clienti;

Dal risultato ottenuto, si può notare come il primo cluster sia predominante rispetto ai secondi due, raggruppando da solo il 73% dei clienti totali. Per quanto riguarda i secondi due, invece, questi si distribuiscono quasi equamente raccogliendo rispettivamente il 12% e il 15% dei clienti.

3.3 Valutazione dei risultati

Vengono di seguito esposte le caratteristiche di ciascun cluster individuato precedentemente, separate per i due negozi. Per ogni cluster si riportano le sette abitudini di acquisto più rilevanti, ognuna di queste descritta da cinque attributi:

- **ID**: codice univoco associato a un'abitudine;
- **Segmento**: il segmento corrispondente all'abitudine;
- **Giorno settimana (G. settimana)**: il giorno corrispondente all'abitudine;
- **n. Clienti**: numero di clienti nel cluster che hanno nel proprio profilo utente tale abitudine;
- **%**: rapporto tra *n. Clienti* e il numero totale dei clienti che hanno acquistato nel giorno cui è riferita l'abitudine. Tale indicatore permette di comprendere meglio quale sia l'effettiva importanza del segmento associato all'abitudine rispetto al giorno della settimana cui l'abitudine fa riferimento.

Si noti, infine, come in tutti i cluster, per entrambi i negozi, non compaiono abitudini rilevanti riferite alla domenica. Questo poiché, come riscontrato in precedenza, durante tale giorno i clienti tendono ad effettuare meno acquisti rispetto agli altri giorni della settimana.

3.3.1 Negozio 70

Come si può osservare dalle tre tabelle sottostanti, i tre cluster individuati sono ben distinti l'uno dall'altro.
Il primo cluster evidenzia una particolare abitudine dei clienti nell'effettuare acquisti nel segmento *frizzanti* lungo tutta la settimana, tranne il venerdì in cui gli acquisti si concentrano anche nel segmento *III lavorazione/freschi*.
Per quanto riguarda il secondo cluster, si può notare come le abitudini di acquisto evidenzino una netta preferenza per il segmento *fresco*, lungo tutto l'arco della settimana. In particolare le percentuali associate alle prime cinque abitudini, indicano che oltre il 50% dei clienti acquisti in tale segmento nei giorni cui le abitudini si riferiscono. Al contrario, per quanto riguarda la sesta abitudine, la percentuale non è ugualmente significativa, mentre, rispetto alla settima abitudine, la percentuale indica una ancor più debole rilevanza.
Nel terzo cluster non si evidenzia un segmento prevalente sugli altri, tuttavia si può osservare come gli acquisti siano concentrati prevalentemente il sabato, per questo motivo si può dedurre come queste abitudini possano rappresentare un valido profilo utente per tale giorno.

Cluster 1 - Negozio 70				
ID	Segmento	G. settimana	n. Clienti	%
1	Frizzanti	Sabato	775	36,9 %
2	Frizzanti	Venerdì	703	30,5 %
3	Frizzanti	Mercoledì	602	29,6 %
4	Frizzanti	Giovedì	575	29,1 %
5	Frizzanti	Lunedì	568	25,6 %
6	III Lavorazione / (Freschi)	Venerdì	568	24,6 %
7	Frizzanti	Martedì	542	26,6 %

Cluster 2 - Negozio 70				
ID	Segmento	G. settimana	n. Clienti	%
1	Fresco	Venerdì	1986	64,8 %
2	Fresco	Sabato	1749	60,3 %
3	Fresco	Giovedì	1578	57,5 %
4	Fresco	Martedì	1521	56,5 %
5	Fresco	Mercoledì	1382	55,9 %
6	Fresco	Lunedì	636	30,3 %
7	Classica	Venerdì	257	8,4 %

Cluster 3 - Negozio 70				
ID	Segmento	G. settimana	n. Clienti	%
1	Classica	Sabato	849	14,4 %
2	Rossi	Sabato	706	11,9 %
3	III Lavorazione / (Freschi)	Sabato	669	11,3 %
4	Effervescenti Nat.	Sabato	655	11,1 %
5	Intero	Sabato	645	10,9 %
6	Fresco	Sabato	568	9,6 %
7	Effervescenti Nat.	Venerdì	466	9,4 %

3.3.2 Negozio 101

Come si evince dalle tre tabelle sottostanti, i tre cluster individuati sono ben distinguibili per quanto riguarda le abitudini di acquisto.
Si può notare come nel primo cluster, i clienti acquistino nei segmenti *III lavorazione/(freschi)* e *Frizzanti* lungo tutta la settimana, ad esclusione del lunedì e della domenica. Tuttavia, le percentuali associate a ciascun'abitudine indicano una debole rilevanza delle abitudini stesse, segno che il profilo utente estratto per tale cluster non sia particolarmente attendibile.
Per quando riguarda, invece, il secondo cluster, gli acquisti più frequenti si concentrano nel segmento *altro*. Anche in questo caso si ha una distribuzione delle abitudini di acquisto costante lungo tutto l'arco della settimana. Le percentuali delle prime sei abitudini, molto elevate, indicano un'alta significatività del profilo estratto, aspetto che non può dirsi certo per quanto riguarda la settima abitudine.
Infine, nel terzo cluster, si può osservare come le abitudini individuate siano abbastanza variegate per quanto riguarda il segmento, ma al contempo siano concentrate in tre giorni specifici. Le percentuali, pur non essendo molto elevate, mostrano una non troppo debole rilevanza del profilo estratto.

Cluster 1 - Negozio 101				
ID	Segmento	G. settimana	n. Clienti	%
1	III Lavorazione / (Freschi)	Sabato	279	14,1 %
2	Frizzanti	Sabato	258	13,1 %
3	III Lavorazione / (Freschi)	Venerdì	179	10,0%
4	Frizzanti	Venerdì	163	9,1 %
5	Frizzanti	Mercoledì	160	9,9 %
6	Frizzanti	Martedì	154	9,3 %
7	Frizzanti	Giovedì	137	8,3 %

Cluster 2 - Negozio 101				
ID	Segmento	G. settimana	n. Clienti	%
1	Altro	Sabato	285	68,3 %
2	Altro	Venerdì	233	60,2 %
3	Altro	Lunedì	217	59,1 %
4	Altro	Martedì	215	57,2 %
5	Altro	Giovedì	204	57,3 %
6	Altro	Mercoledì	192	56,1 %
7	III Lavorazione / (Freschi)	Sabato	56	13,4 %

Cluster 3 - Negozio 101				
ID	Segmento	G. settimana	n. Clienti	%
1	Intero	Sabato	206	30,1 %
2	Classica	Sabato	198	28,9 %
3	Effervescenti Nat.	Sabato	151	22,1 %
4	Comune	Sabato	135	19,7 %
5	Effervescenti Nat.	Venerdì	123	25,3 %
6	Classica	Venerdì	118	24,3 %
7	Classica	Martedì	96	22,9 %

3.3.3 Confronto tra i negozi

Dal confronto tra i due negozi, e quindi tra i tre cluster individuati per ciascuno di questi, sono emerse diverse analogie. Alcune rilevanti sia per quanto riguarda il giorno della settimana, sia per quando concerne il segmento di acquisto. Altre, indicative unicamente rispetto all'abitudine dei clienti nell'effettuare acquisti lungo tutta la settimana concentrandosi in uno specifico segmento piuttosto che in un altro. Da tale analisi sono emerse quindi quattro maggiori analogie tra i sei cluster analizzati. Va precisato, inoltre, che di seguito sono esposte unicamente le similarità osservate, rimandando la valenza di ogni singolo profilo, e quindi di ciascun confronto, a quanto precisato nei paragrafi precedenti.
Per agevolare il confronto, è stata utilizzata una rappresentazione grafica che ha permesso la veloce ed efficace individuazione di possibili analogie tra i due diversi negozi. Si tenga in considerazione che lo scopo dell'analisi era esclusivamente quello

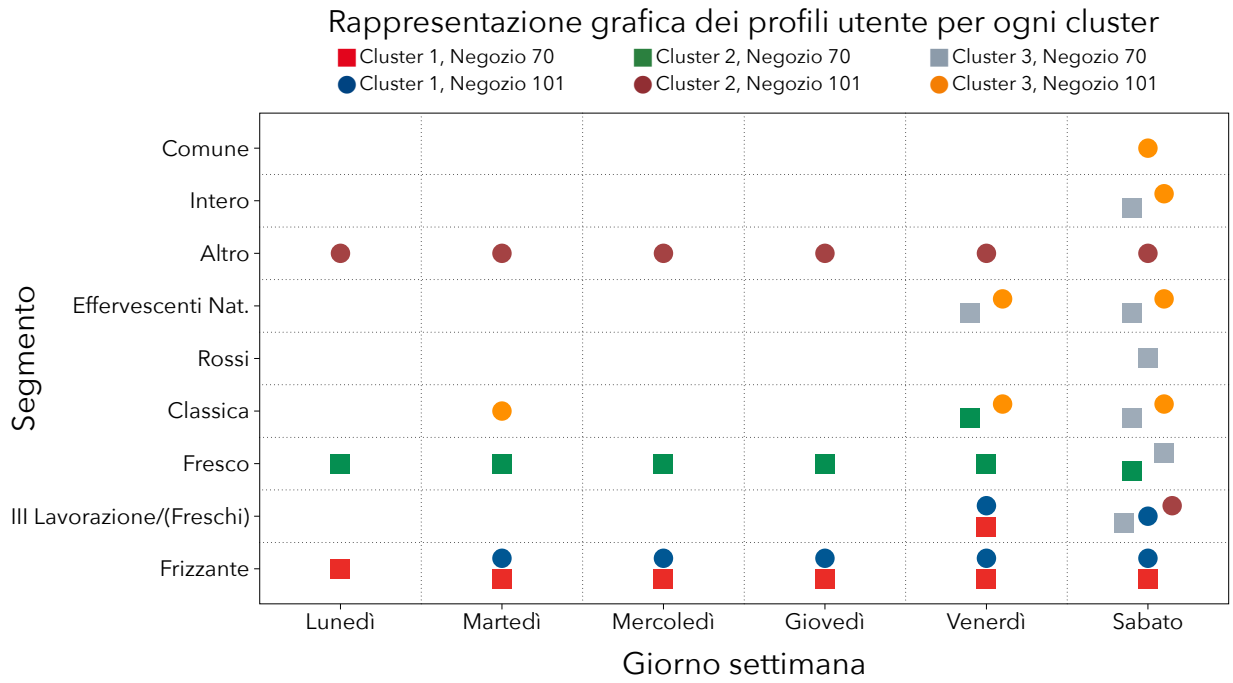

Figura 9 – Rappresentazione grafica utilizzata per individuare facilmente possibili analogie tra i due diversi negozi. A ogni cluster è stato assegnato un diverso colore, mentre, il simbolo a questi assegnato dipende dal negozio di appartenenza: ■ per il negozio 70, ● per il negozio 101.

di individuare abitudini simili appartenenti a cluster di negozi diversi. Non sono state quindi cercate né selezionate analogie tra cluster appartenenti al medesimo negozio. Infine, come ultima osservazione prima di procedere alla descrizione dei confronti individuati, dalla Figura 9 si può facilmente notare come in entrambi i negozi vi sia una concentrazione di acquisti maggiore tra il venerdì e il sabato, piuttosto che lungo il resto della settimana. A differenza degli altri giorni della settimana, infatti, nella giornata del venerdì e in quella del sabato sono presenti, almeno una volta, tutti i sei cluster individuati per i due negozi.
Si procede descrivendo le quattro analogie individuate. Per ciascuna di queste si propone un grafico nel quale sono stati opportunamente omessi giorni della settimana e/o segmenti per i quali non vi era alcuna occorrenza da parte dei due cluster analizzati. Infine la notazione utilizzata per specificare i cluster è la seguente: NEGOZIO_ID.NUMERO_CLUSTER. In questo modo, ad esempio, 70.1 indica il primo cluster del negozio 70.

- **Cluster 70.1 – 101.1**: in questo primo confronto, si può notare come sei delle abitudini dei due cluster siano identiche, mentre solo una sia rivolta sia a un segmento sia a un girono della settimana differente. In particolare, per entrambi i negozi, l'acquisto nel segmento *Frizzante* avviene dal martedì al sabato e solo nel negozio 70, tuttavia, questo avviene anche il lunedì. Inoltre, gli acquisti nel segmento *III Lavorazioni/(Freschi)* avvengono di venerdì in entrambi i negozi e di sabato unicamente nel negozio 101. Le analogie riscontrate per questi due cluster sono le più numerose individuate tra ogni possibile coppia. Va precisato che

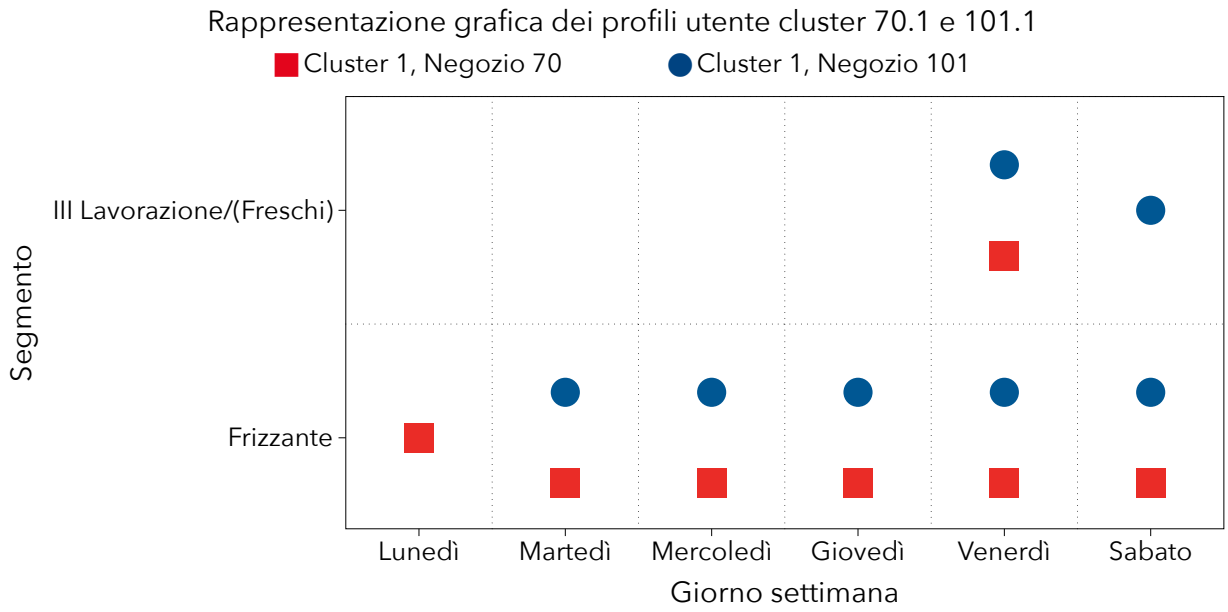

Figura 10 – Illustrazione grafica del confronto tra le abitudini di acquisto del cluster n.1 del negozio 70 e del cluster n.1 del negozio 101.

entrambi rappresentano i due cluster più numerosi per il negozio cui appartengono, sintomo che la parte più numerosa dei clienti abbia abitudini di acquisto molto simili in entrambi i negozi.

- **Cluster 70.3 – 101.3**: il paragone tra questi due cluster mostra come questi abbiano quattro abitudini analoghe: in entrambi i negozi, gli acquisti nella giornata di sabato avvengono nei segmenti *Classica, Effervescenti Nat. e Intero;* inoltre, anche l'acquisto nel segmento *Effervescenti Nat.* di venerdì è comune ad entrambi i cluster.

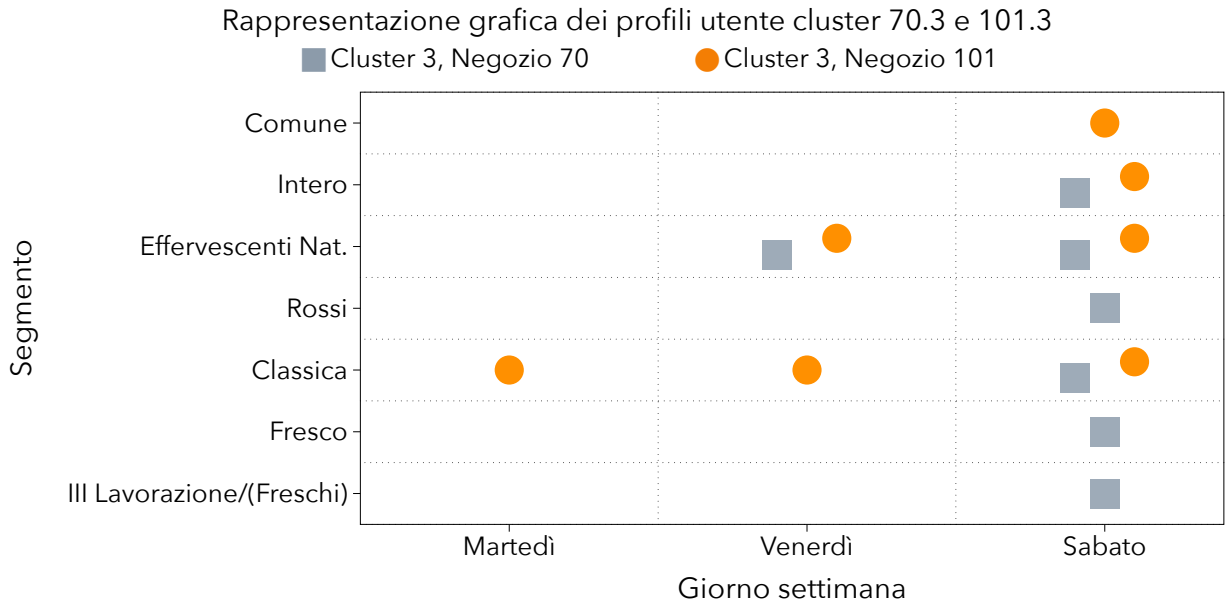

Figura 11 – Illustrazione grafica del confronto tra le abitudini di acquisto del cluster n.3 del negozio 70 e del cluster n.3 del negozio 101.

Le restanti tre abitudini per ciascun cluster sono invece ampiamente diverse. In questo caso entrambi i cluster sono i due che, per numerosità, si collocano a metà tra i tre cluster in ciascun negozio, aspetto che può far riflettere su come la segmentazione abbia prodotto, seppur in modo meno evidente rispetto al caso prima proposto, due cluster in parte simili sia per abitudini sia per posizione rispetto agli altri cluster individuati. Tuttavia si osservi anche che, mentre il cluster 70.3 racchiude il 24% dei clienti totali, il cluster 101.3 ne raggruppa unicamente il 15%.

- **Cluster 70.2 – 101.2**: tra questi due cluster emerge una forte somiglianza delle abitudini di acquisto unicamente riguardo ai giorni in cui tali acquisti avvengono. Mentre i clienti appartenenti al negozio 70 acquistano nel segmento *fresco* in tutti i giorni della settimana, allo stesso modo i clienti del negozio 101 concentrano i propri acquisti nel segmento *altro*. Non vi è quindi una particolare somiglianza come nei due casi precedenti, bensì la sola analogia per la quale le prime sette abitudini d'acquisto di entrambi i cluster prevedano una distribuzione equamente ripartita lungo tutta la settimana degli acquisti effettuati dai clienti nei due negozi.

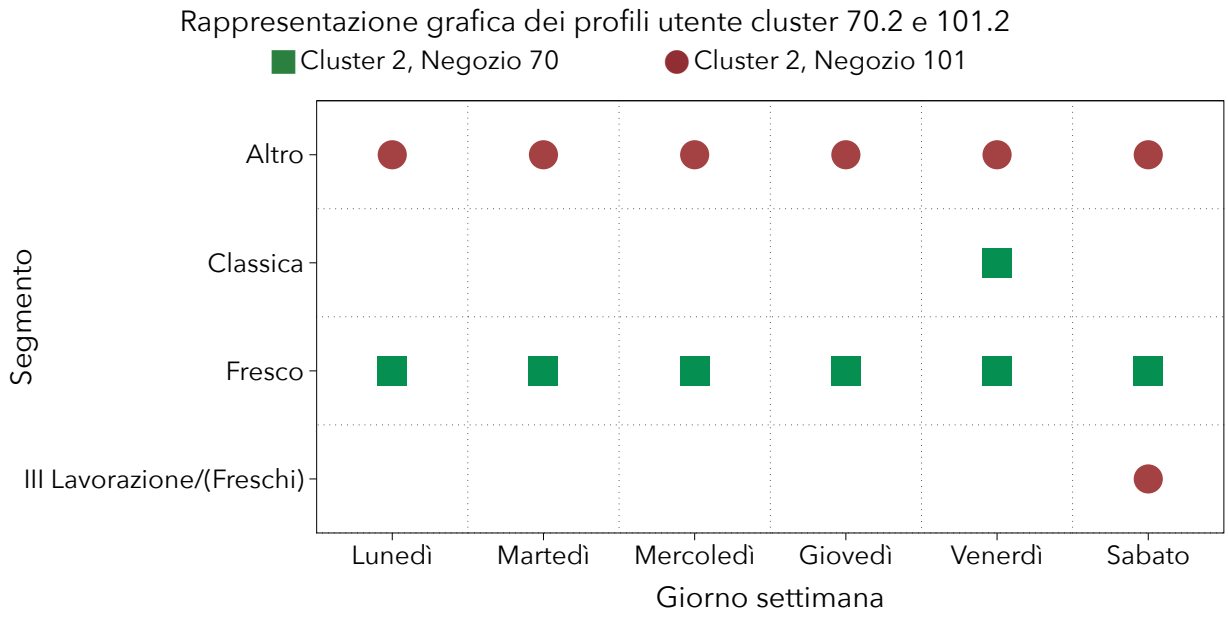

Figura 12 – Illustrazione grafica del confronto tra le abitudini di acquisto del cluster n.2 del negozio 70 e del cluster n.2 del negozio 101.

- **Cluster 70.1 – 101.2**: dalla comparazione dei due cluster si può osservare come la distribuzione settimanale degli acquisti avvenga in un modo quasi identico seppur concentrata in due segmenti differenti. Infatti, i clienti del negozio 70 acquistano prevalentemente nel segmento *Frizzante*, mentre quelli del negozio 101 preferiscono articoli relativi al segmento *Altro*. Un'ultima somiglianza si può notare considerando il segmento *III Lavorazione/(Freschi)*, nel quale i clienti

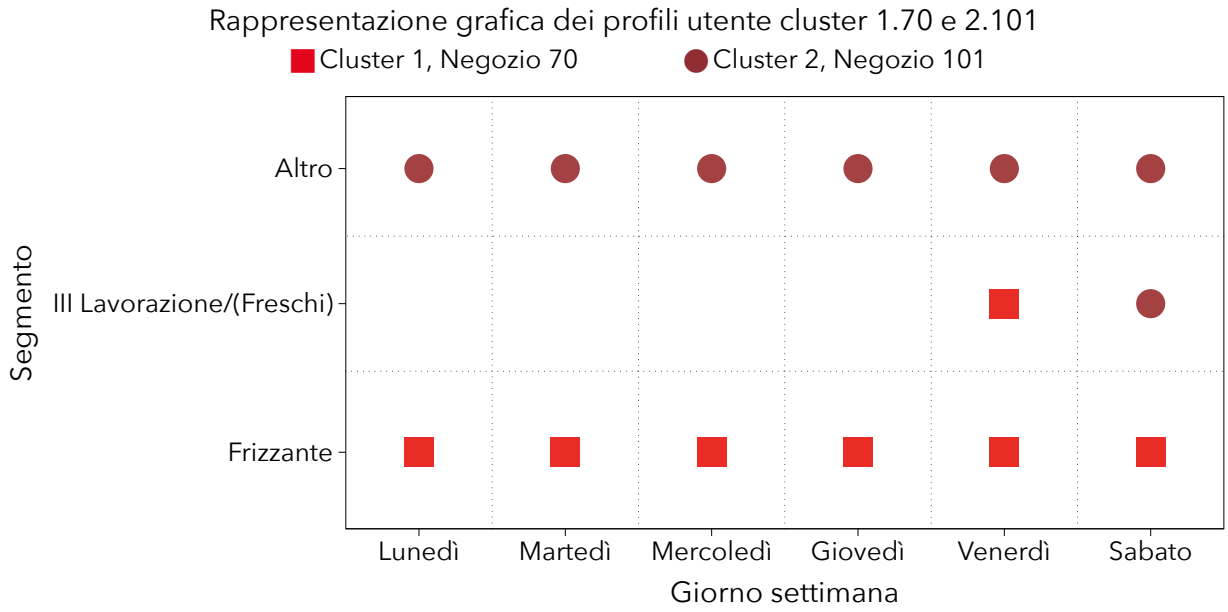

Figura 13 – Illustrazione grafica del confronto tra le abitudini di acquisto del cluster n.1 del negozio 70 e del cluster n.2 del negozio 101.

appartenenti ai due diversi cluster concentrano gli acquisti in due giorni differenti, ovvero Venerdì per il negozio 70 e Sabato per il negozio 101.

In conclusione quindi sono state notate unicamente le precedenti analogie descritte. Dalle quali si può derivare come, per entrambi i negozi, la maggior parte dei clienti si concentri nell'acquisto del segmento *Frizzate* (confronto 70.1 – 101.1). Altre similarità sono state riscontrate seppur caratterizzate da un minor grado di coincidenza.