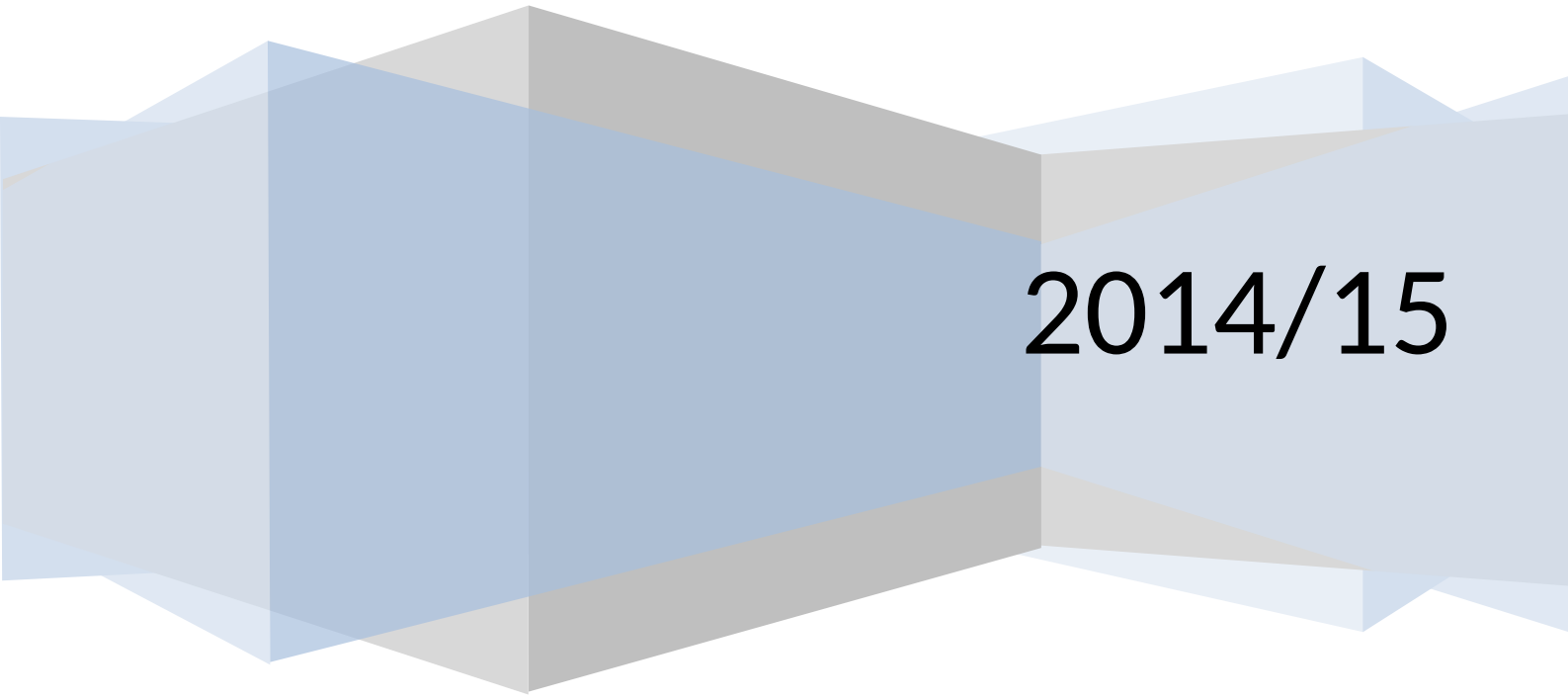
DATA MINING II – Prof.Nanni M., Prof.ssa Monreale A.

PROFILING PURCHASES



2014/15

Sommario

Capitolo 1

Introduzione e obiettivi

Lo scopo di questo progetto è quello di effettuare alcune analisi su un dataset di transazioni reali effettuate da clienti in alcuni supermercati. Nello specifico, andremo a realizzare un *profilo d'acquisto* per ogni cliente, descrivendo: quali sono i **prodotti acquistati sistematicamente** , **dove** (in quale negozio) e la **finestra temporale** in cui è avvenuto l'acquisto.
Successivamente, utilizzando i profili individuati, effettueremo una segmentazione dei clienti, per ogni negozio, al fine di comprendere quali siano i comportamenti d'acquisto tipici.

Capitolo 2

Data Understanding

2.1 Descrizione del dataset

Il dataset analizzato contiene i dettagli di vendita(di un anno di esercizio) di alcuni negozi.
Esso si compone di 6 tabelle:

1. "**articolo**" → contiene la descrizione degli articoli disponibili;
2. "**clienti**" → contiene una sintetica descrizione demografica dei clienti;
3. "**data**" → contiene la descrizione delle date d'acquisto in diverso formato;
4. ~~"**orario**" → contiene la descrizione degli orari d'acquisto in diverso formato;~~
5. "**marketing**" → contiene la descrizione della gerarchia dei prodotti disponibili in negozio;
6. "**venduto**" → è una tabella dei fatti che contiene i singoli prodotti venduti in ogni transazione (scontrino).

Di seguito in Figura 1 vengono rappresentate le relazioni tra le tabelle e i dati effettivamente utilizzati per condurre l'analisi. I metadati relativi alle singole tabelle vengono descritti in ***Appendice A***.

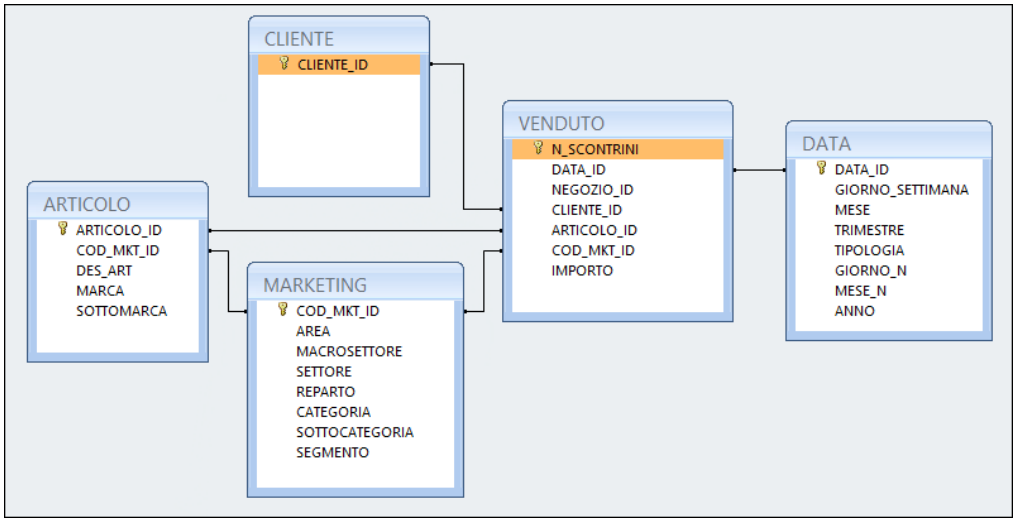


Figura 1 - relazioni tra le tabelle del dataset.

2.2 Qualità dei dati

Per analizzare la qualità dei dati ci siamo serviti di KNIME e precisamente del nodo *Statistics*.
Abbiamo notato la presenza, seppur molto limitata, di missing value nelle tabelle "**articolo**" e "**clienti**".
Nello specifico, nella tabella "**articolo**" i dati mancanti si riferiscono ad attributi quali per esempio *MARCA e SOTTOMARCA*.
I missing value della tabella "**clienti**" si riferiscono agli attributi *FASCIA_ETA*, *SESSO*, *STATO_CIVILE*, *PROFESSIONE*, *TITOLO_STUDIO* e *FASCIA_ANNO_SOCIO*. È facile ipotizzare che i dati raccolti riguardino clienti che hanno sottoscritto una fidelity card e, nel caso di missing value, non abbiano fornito alcune informazioni.
Si è comunque deciso di non eliminare le righe contenenti missing value poiché non influenzano il risultato delle nostre analisi.

2.3 Esplorazione dei dati

Per la creazione dell'*user profile* abbiamo bisogno di tre componenti: il **prodotto**, il **negozio** e la **finestra temporale**; a tal fine analizzeremo la distribuzione dei dati a nostra disposizione, utilizzando il tool KNIME (Figura 2) e Microsoft SQL Server.

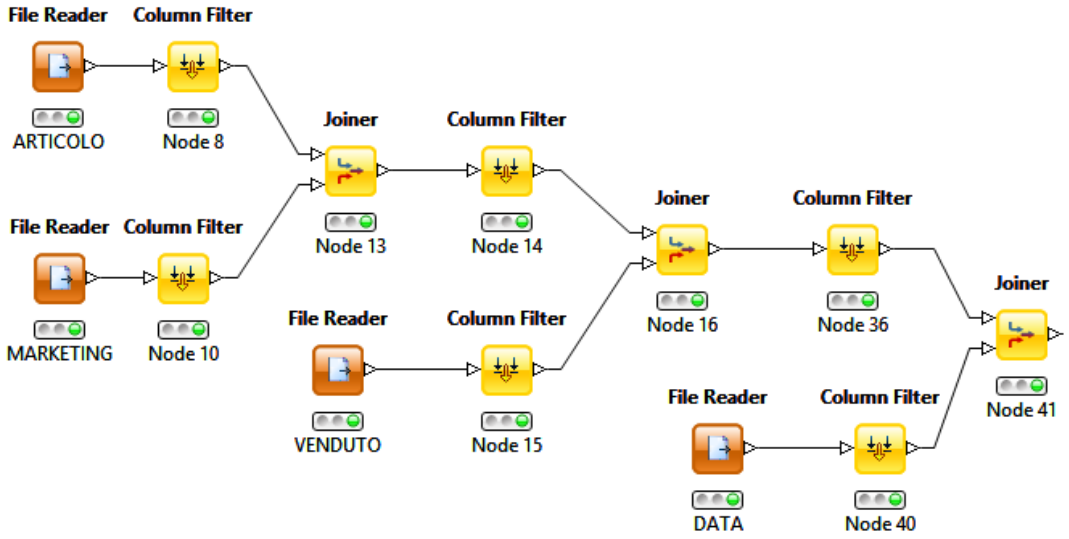


Figura 2 - knime workflow per l'analisi sulla tabella articoli e tabella marketing

ANALISI COMPONENTE PRODOTTO:

Le tabelle "**marketing**" e "**articolo**" sono strettamente correlate in quanto per la nostra analisi la prima contiene la rappresentazione gerarchica dei singoli prodotti presenti in "**articolo"**.

La tabella "**marketing"** è una risorsa fondamentale per il nostro studio, in quanto è uno degli strumenti che ci consente di modulare la granularità del dettaglio dei prodotti in modo che l'analisi non risulti banale (generale) o troppo frammentata (dispersiva).
Abbiamo rappresentato graficamente la distribuzione dei prodotti presenti nella tabella "**articolo**" considerando i diversi livelli di marketing. In Figura 3 sono mostrate le distribuzioni relative ad AREA, MACROSETTORE e SETTORE (non vengono mostrati i livelli con granularità più fine, da reparto a segmento, per questioni di leggibilità).

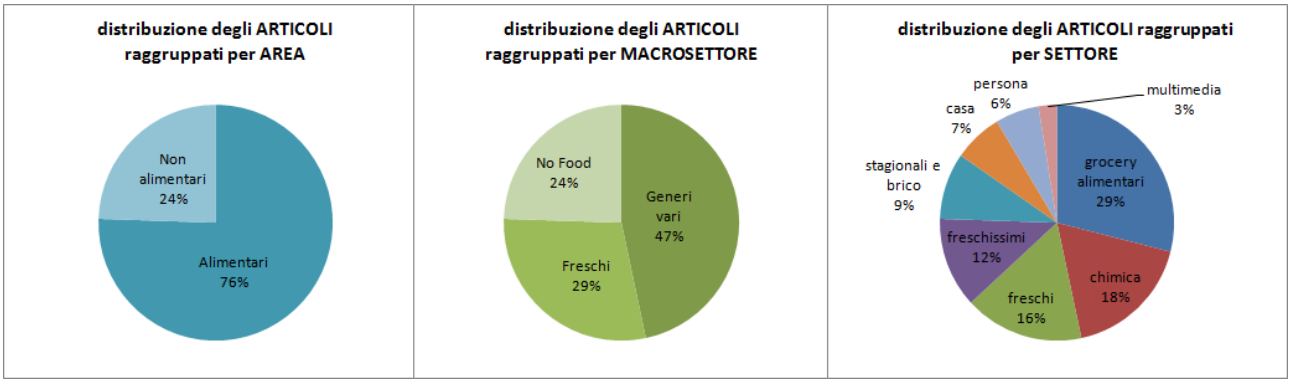


Figura 3 - Distribuzione degli Articoli disponibili raggruppati per diversi livelli di Marketing

Per avere un idea di come sono distribuite le categorie degli "**articoli**" sul "**venduto"** riportiamo graficamente(Figura 4) la loro distribuzione su diversi livelli di "**marketing**".

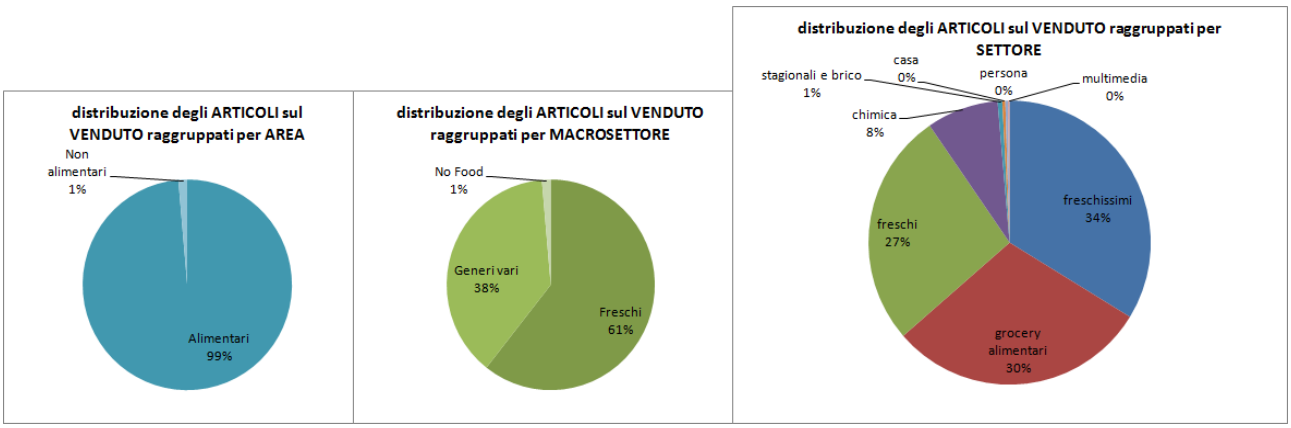


Figura 4 - Distribuzione degli Articoli sul Venduto, raggruppati per diversi livelli di Marketing

E' interessante notare come la distribuzione sia profondamente diversa tra gli articoli effettivamente venduti(Figura 4) rispetto alla distribuzione degli articoli disponibili(Figura 3). Questa informazione potrebbe essere molto utile per ottimizzare la politica di approvvigionamento.

ANALISI COMPONENTE TEMPORALE:

La gerarchia temporale (espressa dalla tabella "**data**") è la seconda leva su cui agiremo per modulare la granularità dell'analisi(Figura 5 e Figura 6).

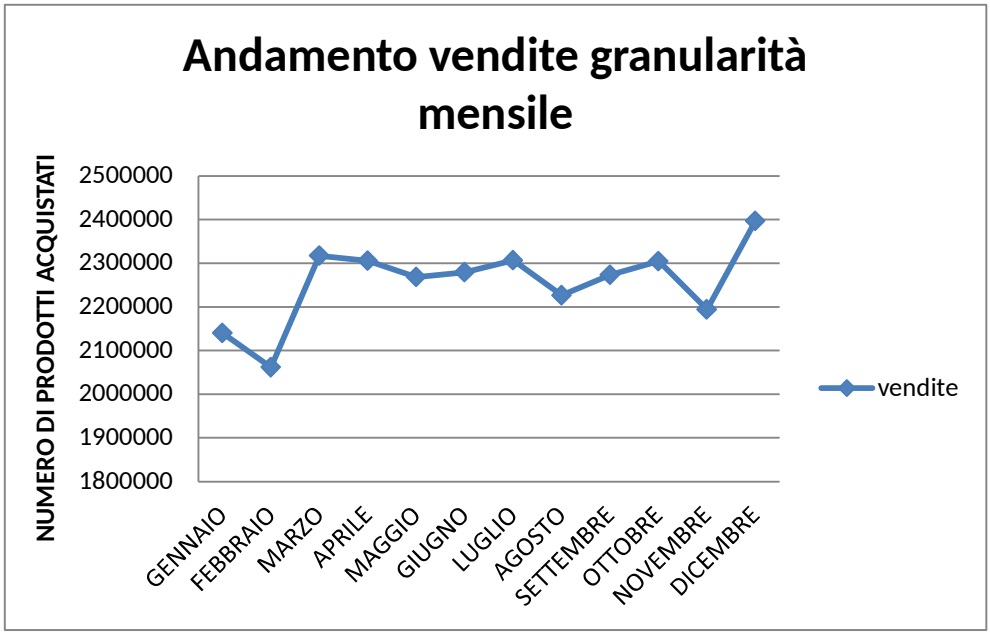


Figura 5 – distribuzione del venduto nel tempo(granularità mensile)

Dalla Figura 5 è possibile osservare un incremento del venduto nel mese di dicembre; questo fenomeno è facilmente spiegabile dal fatto che i clienti tendenzialmente spendono di più nel periodo natalizio. La flessione nei mesi di gennaio e febbraio è ipotizzabile che sia dovuta alla contrazione dei consumi che abitualmente avviene dopo le festività.

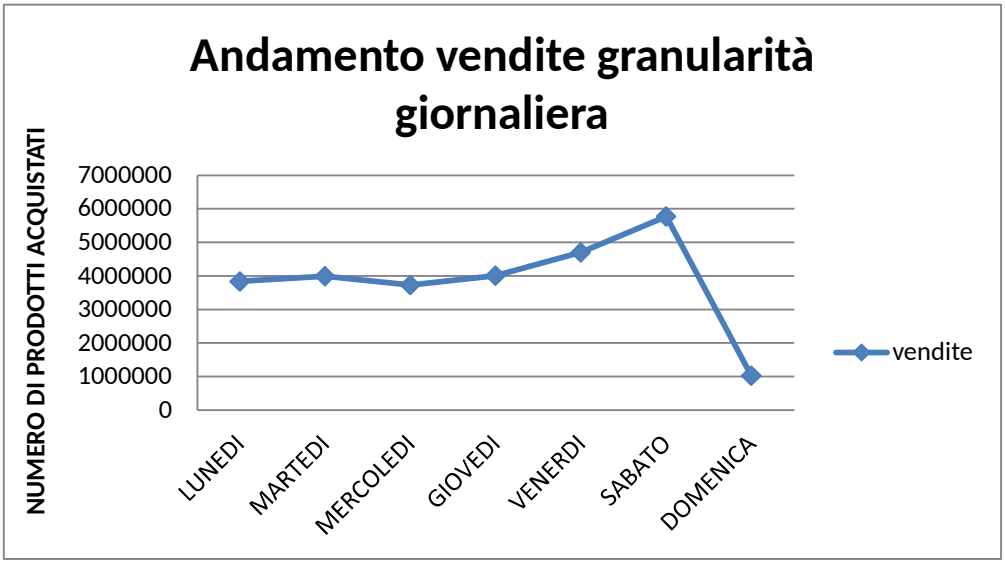


Figura 6 - distribuzione del venduto nel tempo(granularità giornaliera)

Dalla Figura 6 si osserva un comportamento comune di una famiglia italiana che generalmente approfitta del sabato per fare la spesa per l'intera settimana mentre la domenica preferisce dedicarsi ad altre attività.

ANALISI COMPONENTE NEGOZIO:

L'informazione riguardante il negozio è disponibile all'interno della tabella "**venduto**", precisamente grazie al campo NEGOZIO_ID.
Con l'utilizzo di MS SQL Server abbiamo potuto osservare che il dataset si riferisce alle transazioni di 4 negozi identificati dai codici: 70, 101, 102, 103.
Successivamente abbiamo effettuato un'analisi temporale sui negozi individuati.
E' stato interessante notare la similarità della forma delle serie temporali, enfatizzate dai grafici a pila(Figura 7 e Figura 8); ciò indica che l'andamento della frequenza d'acquisto dei clienti è comune in tutti i negozi.

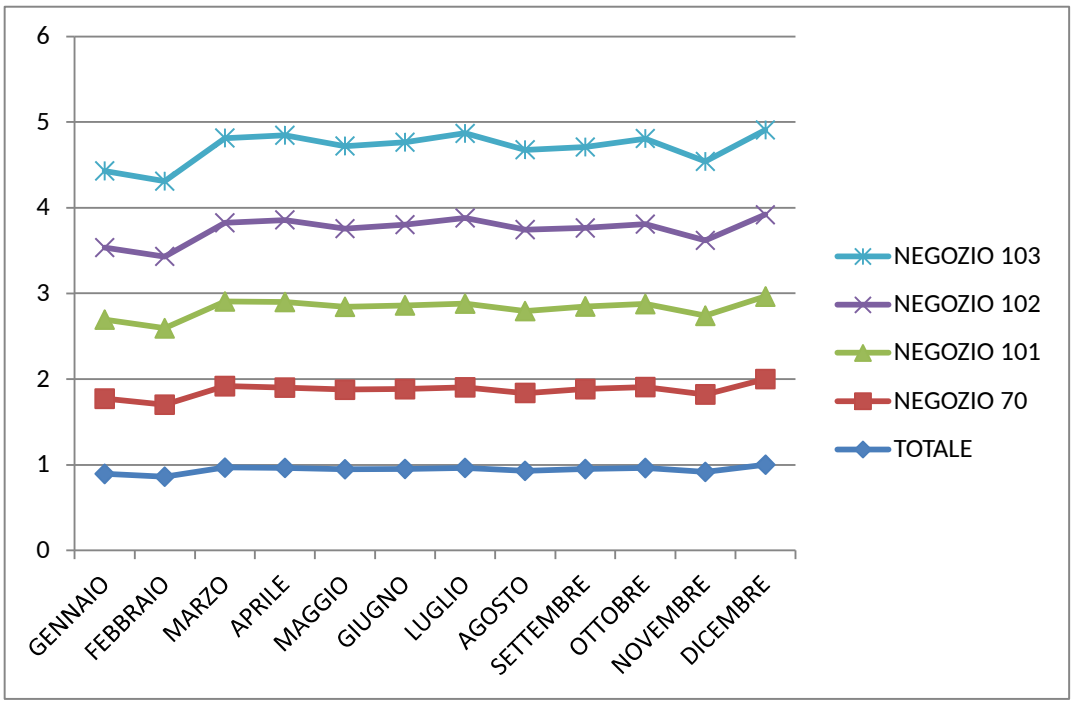


Figura 7 – GRAFICO A PILA, shape delle time series relative ai negozi(su dati normalizzati mensili)

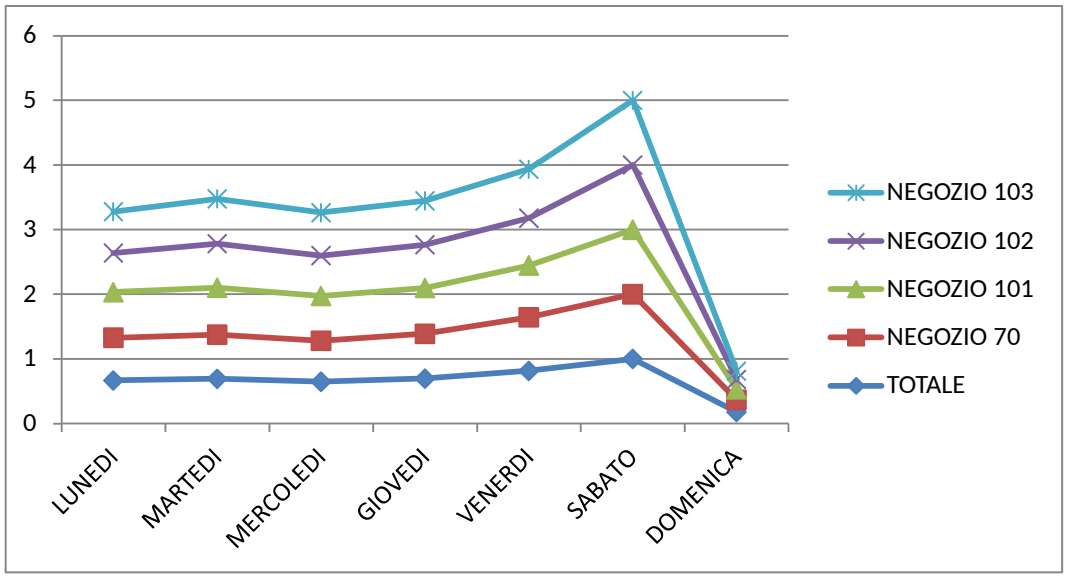


Figura 8 - GRAFICO A PILA, shape delle time series relative ai negozi(su dati normalizzati giornalieri)

Un'ulteriore analisi riguarda la percentuale di prodotti acquistati presso ognuno dei 4 negozi, rispetto al totale delle vendite (Figura 9).
La maggior parte delle vendite si riferisce al negozio 70 (12713953 prodotti venduti), questo potrebbe risultare un problema nelle successive analisi (maledizione della dimensionalità), perché la dispersione dei dati potrebbe influire sui risultati.

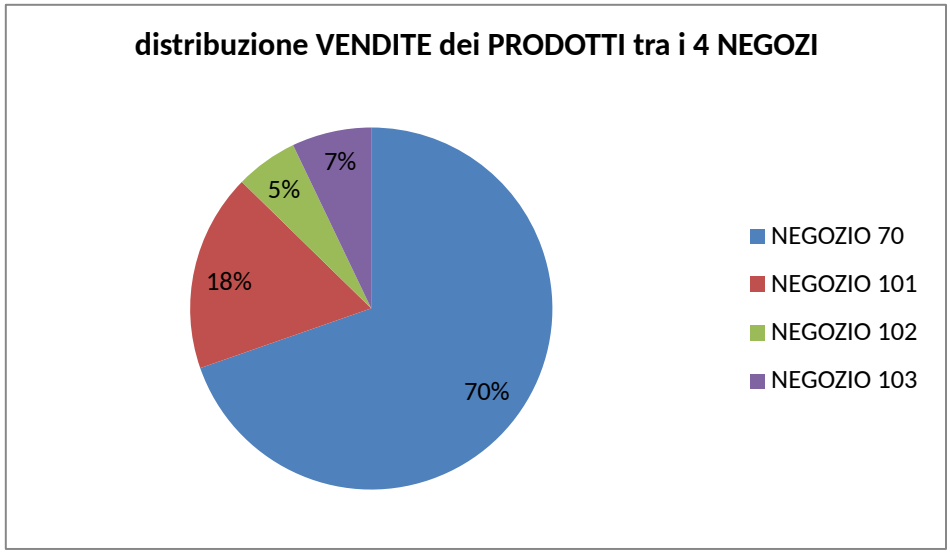


Figura 9 - Percentuale di distribuzione delle vendite del totale dei prodotti rispetto al negozio

Altro aspetto interessante è che la maggior parte degli utenti fa acquisti sempre nel medesimo negozio (Figura 10).
Infatti effettuando una query con MS SQL Server abbiamo riscontrato che su 48658 clienti, con almeno un acquisto, soltanto il 32% fa spese in più negozi (e soltanto l'1% ha effettuato almeno un acquisto in tutti e 4 i supermarket).

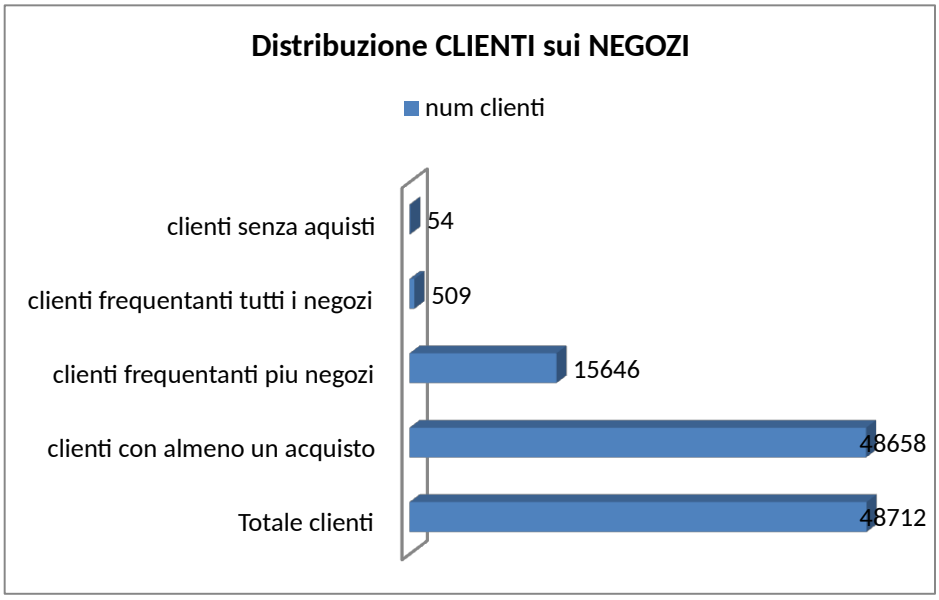


Figura 10 - Distribuzione dei clienti sui negozi del dataset

Capitolo 3

User Purchase Profiles

In questa fase modelleremo un *profilo d'acquisto* per ogni cliente che descriva: un set di prodotti acquistati sistematicamente, dove vengono acquistati e quando.
Il profilo avrà la forma di un insieme di terne del tipo (***cosa***, ***dove***, ***quando***).

3.1 Pre-Processing

La fase preliminare per la definizione dei *profili d'acquisto* consiste nell'identificazione dei livelli di granularità sia del prodotto sia delle finestre temporali.
Per quanto riguarda il prodotto si è scelto di utilizzare la CATEGORIA poiché risulta essere non troppo generico, e non troppo specifico, permettendoci di effettuare analisi non banali.
Ad esempio, se avessimo scelto una granularità più grossa, come l'attributo AREA, nel 99% dei casi avremmo ottenuto fra gli itemset frequenti il valore "*alimentari*" (Figura 4). Viceversa, se avessimo scelto una granularità più fine, ad esempio l'attributo DES_ART (costituito da 21026 valori) non saremmo riusciti a trovare itemset con supporto accettabile.

Stesso discorso per la scelta della granularità della finestra temporale. Abbiamo scelto l'attributo GIORNO_SETTIMANA poiché ci consente di osservare la sistematicità degli acquisti dei singoli clienti in un lasso di tempo congruo affinché l'analisi non risulti banale.
Se avessimo scelto una granularità più grossa, come il MESE, avremmo ottenuto regole troppo generali; gli item set trovati avrebbero contenuto prodotti di consumo quotidiano, come latte e pane, con la conseguente omogeneità dei profili. D'altra parte, scegliendo una granularità più fine, come la FASCIA ORARIA(attributo della tabella "**orario**"), avremmo ottenuto dei dati poco affidabili, in quanto nonostante i clienti siamo abitudinari, l'orario d'acquisto può essere influenzato da molteplici fattori ed imprevisti.

Dopo aver scelto le granularità, utilizzando il nodo *column aggregator* di KNIME abbiamo composto la tripla (prodotto, negozio, tempo) per ogni riga della tabella "**venduto**".

Altro aspetto che abbiamo considerato è che il 37% dei clienti del dataset ha meno di 6 scontrini. Questa informazione non influisce sul risultato di questa fase di analisi ma sarà un aspetto da tenere in considerazione nella successiva fase di clustering (Capitolo 4).

3.2 Individuazione Profilo d'Acquisto

Attraverso il tool KNIME abbiamo costruito un workflow che, a partire dalla join tra le tabelle del dataset (come mostrato in Figura 2), esegue un loop per creare un profilo per ogni cliente (Figura 11).

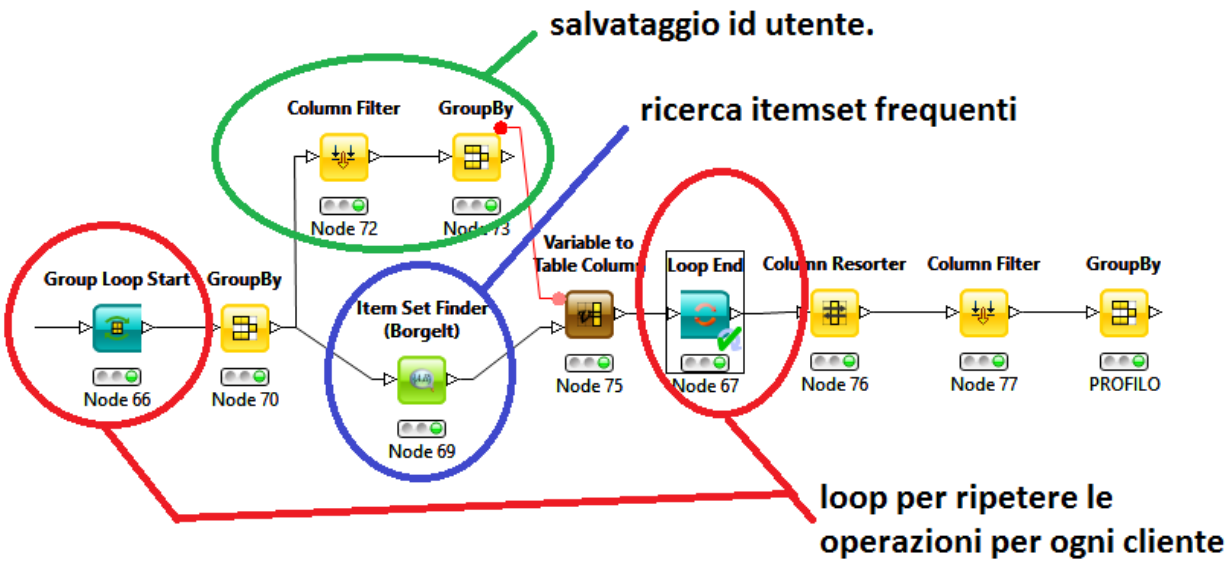


Figura 11 - KNIME workflow per il Profiling dei clienti

Come si può vedere dall'immagine il workflow è composto da 3 componenti principali:

1. *Group Loop Start* e *Group Loop End* delimitano l'inizio e la fine della ricorsione. *Group Loop Start* è stato impostato affinché selezionasse l'insieme delle transazioni che si riferiscono ad un utente;
2. *Item Set Finder (Borgelt)* dato in input un insieme di oggetti (strutturati come [prodotto, negozio, tempo]) ricerca item set frequenti. Abbiamo impostato il nodo per lavorare con l'algoritmo Apriori con item set size 3 e con supporto minimo pari al 2.5%. La scelta di una soglia di support così bassa è stata necessaria a causa delle dimensioni del dataset;
3. *Column Filter* e *GroupBy* sono stati necessari per tener traccia dell'CLIENTE_ID.

3.3 Risultati e valutazioni

Un sample dei *profili d'acquisto* ottenuti è mostrato in Figura 12.

Row ID	CLIENT...	Sorted list(ItemSet)
Row0	802	[[.SABATO, 70, ALTRI ORTAGGI],[.SABATO, 70, BOVINO]]
Row1	1980	[[.VENERDI, 103, BISCOTTI],[.VENERDI, 103, INSALATE],[.VENERDI, 103, LATTE],[.VENERDI, 103, PASTA DI SEMOLA],[....
Row2	1984	[[.MERCOLEDI, 103, BOVINO]]
Row3	3414	[[.SABATO, 70, BIBITE],[.SABATO, 70, CAFFE'],[.SABATO, 70, CONSERVE DI PESCE],[.SABATO, 70, CONSERVE DI VERDU...
Row4	4022	[[.SABATO, 70, BOVINO],[.SABATO, 70, PANE PROD. ESTERNA]]
Row5	4126	[[.GIOVEDI, 70, BIBITE],[.MARTEDI, 70, ASCIUGATUTTO],[.MARTEDI, 70, DETERGENTI SUPERFICI],[.MARTEDI, 70, VERD...
Row6	4611	[[.SABATO, 70, BOVINO],[.SABATO, 70, PASTICCERIA PRODUZIONE ESTERNA]]
Row7	5220	[[.VENERDI, 70, ACQUE],[.VENERDI, 70, AGRUMI],[.VENERDI, 70, ALIMENTI INFANZIA],[.VENERDI, 70, ALIMENTI PREPA...
Row8	5918	[[.GIOVEDI, 70, COMPLETAMENTO IGIENE PERSONA],[.GIOVEDI, 70, PANE PROD. ESTERNA],[.GIOVEDI, 70, SPECIALITA' ...
Row9	6356	[[.LUNEDI, 101, LATTE]]
Row10	8009	[[.MARTEDI, 101, APERITIVI],[.MARTEDI, 101, BIBITE],[.MARTEDI, 101, RICORRENZA PASQUA]]
Row11	8012	[[.SABATO, 102, LATTE],[.SABATO, 102, UOVA],[.VENERDI, 102, CAFFE'],[.VENERDI, 102, CARTA IGIENICA],[.VENERDI, ...
Row12	8041	[[.GIOVEDI, 70, ALTRE VERDURE],[.GIOVEDI, 70, ALTRI ORTAGGI],[.GIOVEDI, 70, BOVINO],[.GIOVEDI, 70, CONDIMENTI...
Row13	8775	[[.MARTEDI, 70, BASE FARINA],[.MARTEDI, 70, CONIGLIO],[.MARTEDI, 70, ROSTICCERIA],[.MARTEDI, 70, VERDURE PR...
Row14	8875	[[.GIOVEDI, 101, AVVOLGENTI CUCINA],[.GIOVEDI, 101, PASTA FRESCA]]
Row15	8887	[[.GIOVEDI, 103, BIBITE],[.GIOVEDI, 103, CONSERVE DI FRUTTA],[.GIOVEDI, 103, FETTE BISCOTTATE],[.GIOVEDI, 103, ...
Row16	9549	[[.GIOVEDI, 70, ALTRI ORTAGGI],[.GIOVEDI, 70, CIPOLLE E AGLIO],[.GIOVEDI, 70, POMODORI]]
Row17	9937	[[.GIOVEDI, 103, ALIMENTI PER GATTI],[.GIOVEDI, 103, DESSERT],[.GIOVEDI, 103, SUCCHI DI FRUTTA],[.MARTEDI, 103...
Row18	9953	[[.VENERDI, 101, AGRUMI],[.VENERDI, 101, ALTRI ORTAGGI],[.VENERDI, 101, BANANE],[.VENERDI, 101, COCOMERI/ME...
Row19	9957	[[.MARTEDI, 70, CONSERVE POMODORO],[.MARTEDI, 70, POLLAME]]
Row20	9961	[[.LUNEDI, 70, LATTE],[.LUNEDI, 70, PASTA DI SEMOLA],[.MARTEDI, 70, ALIMENTI PER GATTI],[.MARTEDI, 70, LATTE],[....

Figura 12 - Tabella profilo Utenti (sample 20 row)

Gli utenti profilati costituiscono circa il 65% dei clienti che hanno effettuato almeno un acquisto nell'anno; il restante 35% non presenta acquisti sistematici.
Dopo aver individuato tutti i profili abbiamo effettuato qualche analisi statistica al fine di osservare se i dati profilati fossero rappresentativi e in linea con le distribuzioni del totale del venduto (Figura 13, Figura 14, Figura 15).

	CATEGORIA	VENDUTO
1	LATTE	996098
2	PANE PROD. ESTERNA	837267
3	FORMAGGI FRESCHI	824026
4	BOVINO	819079
5	SALUMI A SERVIZIO ASSISTITO	771136
6	PASTA DI SEMOLA	719297
7	ACQUE	664741
8	VERDURE PREPARATE PRODUZIONE ESTERNA	639171
9	YOGURT	596914
10	POMODORI	537220
11	ALTRI ORTAGGI	469997
12	BIBITE	447275
13	AGRUMI	437506
14	SPECIALITA' BASE PANE PROD. INTERNA	423781
15	ALTRE VERDURE	422413
16	BISCOTTI	410948
17	CONSERVE POMODORO	397391
18	SALUMI LIBERO SERVIZIO	384145
19	BANANE	359114

categoria piu' venduta prima della profilazione

CAT_PRODOTTO	
LATTE	13860
PANE PROD. ESTERNA	13372
FORMAGGI FRESCHI	9172
ACQUE	8979
BOVINO	7387
VERDURE PREPARATE PRODUZIONE ESTERNA	7341
YOGURT	6435
PASTA DI SEMOLA	6144
SPECIALITA' BASE PANE PROD. INTERNA	6106
BIBITE	6079
SALUMI A SERVIZIO ASSISTITO	5670
BISCOTTI	5246
POMODORI	4927
ALTRI ORTAGGI	4492
AGRUMI	4428
SALUMI LIBERO SERVIZIO	4240
PASTICCERIA PRODUZIONE ESTERNA	4148
UOVA	4104

categoria piu venduta dopo la profilazione

Figura 13 - CATEGORIA su dati totali vs CATEGORIA su dati profilati

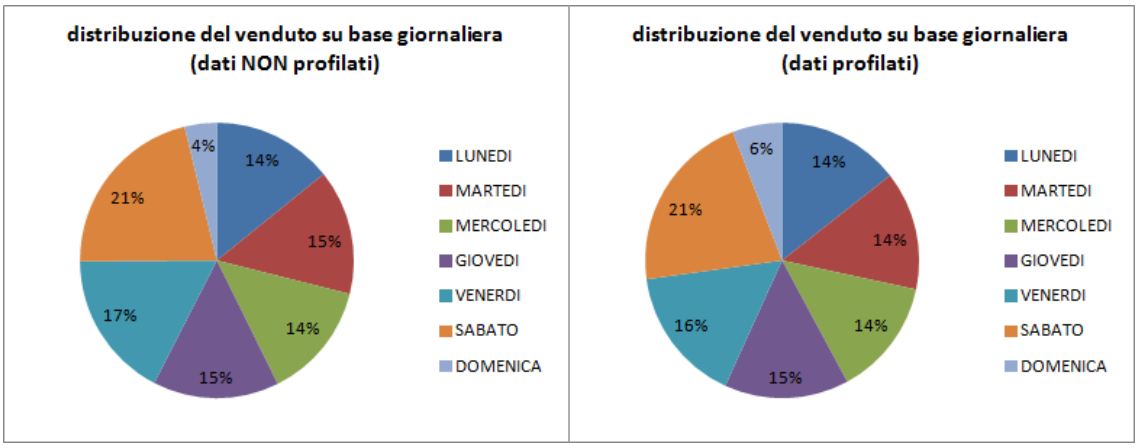


Figura 14 - distribuzione del venduto su base giornaliera su dati non profilati vs distribuzione del venduto su base giornaliera su dati profilati

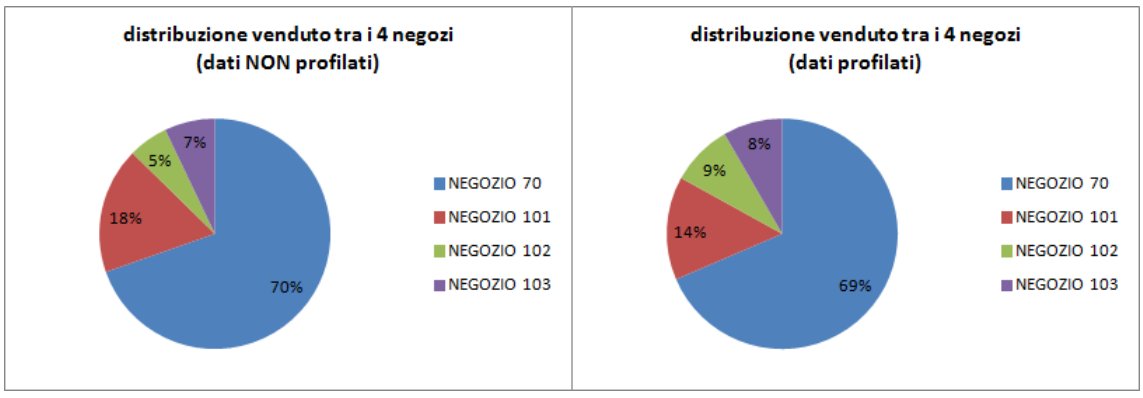


Figura 15 - distribuzione del venduto tra i 4 negozi sui dati non profilati vs distribuzione del venduto sui 4 negozi sui dati profilati

Osservando i grafici, risulta evidente la similarità tra le distribuzioni dei dati profilati e quelli originari.

Capitolo 4

Store Analysis

In quest'ultima fase andremo a segmentare i clienti(partendo dai *profili* estratti nella fase precedente) al fine di individuare i profili tipici d'acquisto per ogni negozio.

4.1 Pre-Processing

La fase preliminare consiste nell'estrazione di alcune features (Figura 16) che rappresentino i profili d'acquisto.
Nel nostro caso abbiamo utilizzato:

a. Count(Itemset) → indica il numero di item (ogni item è una terna di valori [categoria, negozio, tempo]) contenuti nel profilo;
b. CAT_PRODOTTO(supp%) → indica la categoria del prodotto con supporto (relativo percentuale) maggiore, ovvero la categoria con maggior frequenza tra gli acquisti sistematici (inizialmente avevamo scelto la moda come feature sulla categoria, ma ci siamo resi conto che non era molto solida dato che i valori non erano pesati);
c. Count(CAT_PRODOTTO) → indica il numero delle diverse categorie acquistate sistematicamente;
d. GIORNO(supp%) → indica il giorno con supporto maggiore;
e. Count(GIORNO) → indica il numero dei diversi giorni in cui il cliente effettua acquisti sistematici.

Come accennato alla fine del paragrafo 3.1 circa il 37% dei clienti del dataset ha meno di 6 scontrini. Ricordiamo che l'obiettivo di questa analisi è definire dei profili d'acquisti tipici, per ogni negozio, ed in questo senso profili di utenti con pochissimi scontrini andrebbero a inficiare la bontà dei risultati del clustering. Per questa ragione filtriamo i clienti mantenendo solo quelli con almeno 6 scontrini (è stata scelta una soglia "bassa" per non ridurre ulteriormente la dimensione del dataset).

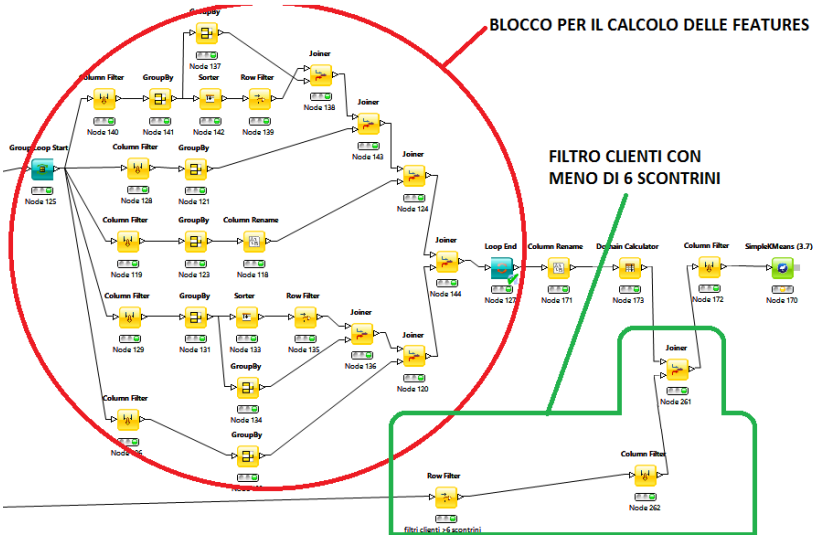
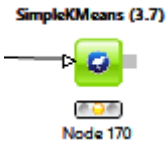


Figura 16 - KNIME workflow per la fase di store analysis dei clienti

4.2 Custumer segmentation



Per la fase di clustering abbiamo utilizzato il nodo di KNIME *SimpleKMeans(3.7)* della libreria di WEKA.
Si è scelto questo nodo per i seguenti motivi:

- Gestisce sia attributi numerici che categorici;
- Fornisce dei centroide, che considereremo profilo d'acquisto tipico del relativo negozio;
- Esegue una normalizzazione automatica dei dati.

4.2.1 Negozio 70

Tramite lo studio della curva dell'SSE [1] al crescere del numero di cluster, abbiamo trovato il valore di k che utilizzeremo per effettuare il clustering(Figura 17).

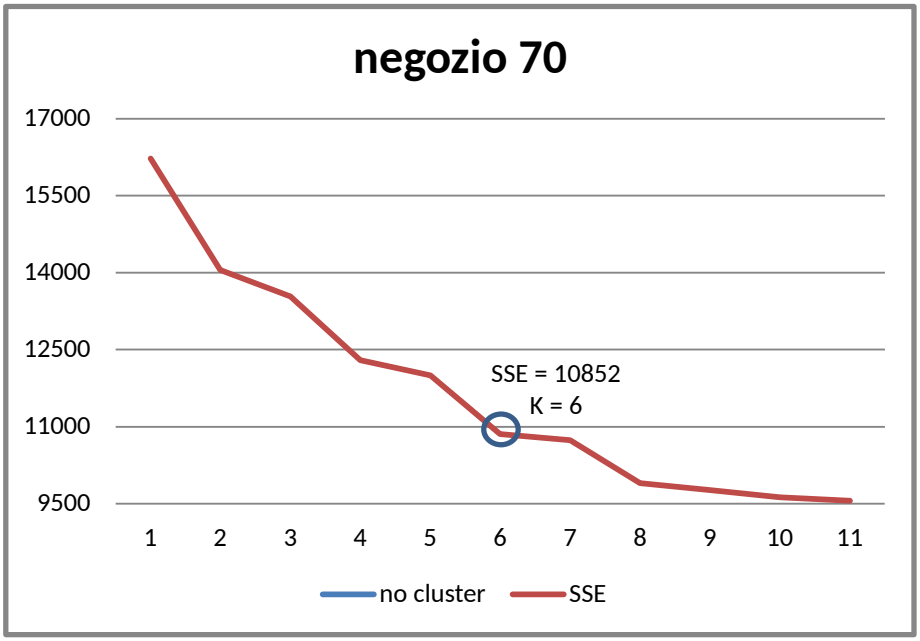


Figura 17 - curva SSE su negozio 70

[1] L'SSE è la somma delle differenze dei quadrati fra ciascuna osservazione e la media del suo gruppo.

In Figura 18 mostriamo i cluster ottenuti applicando il K-means con k = 6 sui profili dei clienti del negozio 70.

```
kMeans
======

Number of iterations: 21
Within cluster sum of squared errors: 10852.425663836333
Missing values globally replaced with mean/mode

Cluster centroids:
                                 Cluster#
Attribute          Full Data           0           1           2           3           4           5
                     (9355)      (1545)      (2306)      (1215)       (775)      (2009)      (1505)
=============================================================================================================
CAT_PRODOTTOsupp       LATTE PANEPRODESTERNA       LATTE PANEPRODESTERNA PANEPRODESTERNA       LATTE       LATTE
CountCAT_PRODOTTO     5.6599      5.5735      5.8812      3.9374      15.129      2.8303       5.701
CountItemset          6.8095      7.2984      6.6002      4.5852     19.5445      3.0483       6.887
GIORNOsupp            SABATO      GIOVEDI      SABATO     DOMENICA      VENERDI      VENERDI      LUNEDI
CountGIORNO           1.9425      2.5126      1.5997       1.758      3.7639      1.2449      2.0246


Clustered Instances

0       1545 ( 17%)
1       2306 ( 25%)
2       1215 ( 13%)
3        775 (  8%)
4       2009 ( 21%)
5       1505 ( 16%)
```

Figura 18 - K-means su negozio 70

I profili tipici d'acquisto che emergono riguardano utenti che effettuano acquisti regolari due giorni a settimana, e tra i prodotti più frequenti troviamo LATTE e PANE_PROD_ESTERNA.
I clienti risultano distribuiti omogeneamente tra i cluster, fatta eccezione per il numero 3 costituito solo dall'8% dei clienti. Questa situazione è giustificabile dal fatto che il suo centroide sia caratterizzato da circa 20 itemset frequenti e da 15 diverse categorie di prodotti acquistate con sistematicità, in 4 diversi giorni.
L'unica anomalia riscontrata è la mancanza di cluster il cui centroide sia caratterizzato da valore di *GIORNOsupp* pari a MARTEDI o MERCOLEDI, nonostante (come mostrato in Figura 14) abbiano una distribuzione molto maggiore rispetto alla DOMENICA.
Sommariamente non spiccano profili d'acquisto particolarmente differenziati.

4.2.2 Negozio 101

Tramite lo studio della curva dell'SSE al crescere del numero di cluster, abbiamo trovato il valore di k che utilizzeremo per effettuare il clustering (Figura 19).

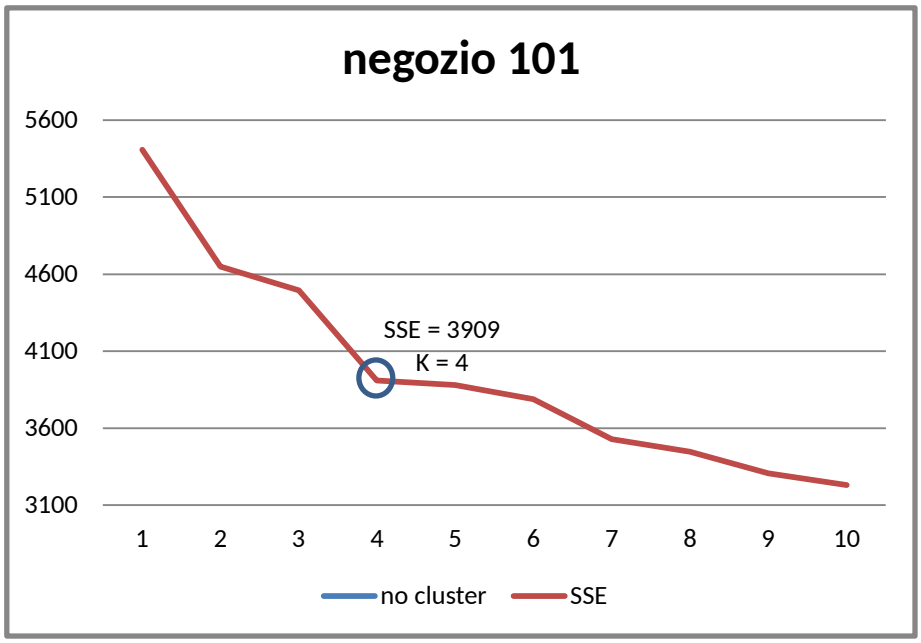


Figura 19 - curva SSE su negozio 101

In Figura 20 mostriamo i cluster ottenuti applicando il K-means con k = 4 sui profili dei clienti del negozio 101.

```
kMeans
======

Number of iterations: 8
Within cluster sum of squared errors: 3909.484398530888
Missing values globally replaced with mean/mode

Cluster centroids:
                                        Cluster#
Attribute               Full Data             0             1             2             3
                           (3201)        (1227)         (657)         (530)         (787)
=================================================================================================
CAT_PRODOTTOsupp           LATTE PANEPRODESTERNA PANEPRODESTERNA         ACQUE         LATTE
CountCAT_PRODOTTO         4.7173        3.5086        7.0533        5.7906        3.9288
CountItemset              5.9378         4.088        9.4916        6.9774         5.155
GIORNOsupp                SABATO        SABATO        LUNEDI       VENERDI       MARTEDI
CountGIORNO               2.0672        1.5265        2.9437        2.0906        2.1626


Clustered Instances

0       1227 ( 38%)
1        657 ( 21%)
2        530 ( 17%)
3        787 ( 25%)
```

Figura 20 - K-means su negozio 101

I profili tipici d'acquisto che emergono riguardano utenti che acquistano regolarmente LATTE, ACQUA e PANE_PROD_ESTERNA, che effettuano spese regolari in due-tre giorni diversi.
I cluster ottenuti risultano equamente distribuiti quindi risulta molto difficile fare una distinzione netta di profili d'acquisto.

4.2.3 Negozio 102

Tramite lo studio della curva dell'SSE al crescere del numero di cluster, abbiamo trovato il valore di k che utilizzeremo per effettuare il clustering (Figura 21).

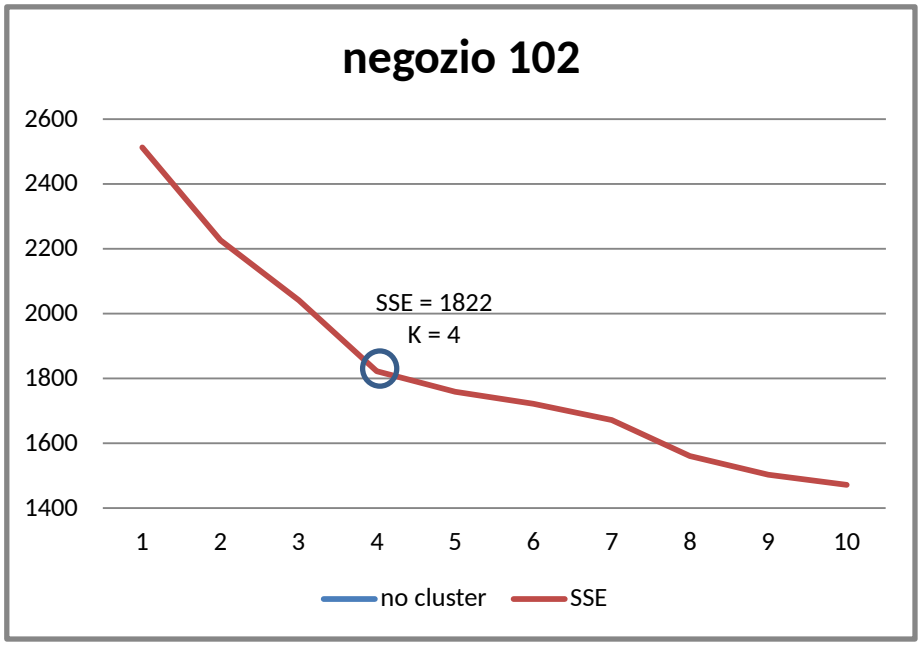


Figura 21 - curva SSE su negozio 102

In Figura 22 mostriamo i cluster ottenuti applicando il K-means con k = 4 sui profili dei clienti del negozio 102.

```
kMeans
======

Number of iterations: 7
Within cluster sum of squared errors: 1854.6382929404112
Missing values globally replaced with mean/mode

Cluster centroids:
                                        Cluster#
Attribute               Full Data             0             1             2             3
                           (1481)        (254)         (674)         (339)         (214)
=============================================================================================
CAT_PRODOTTOsupp           LATTE         ACQUE         LATTE PANEPRODESTERNA         BIBITE
CountCAT_PRODOTTO         5.2566        5.2756         3.822        7.1209        6.7991
CountItemset              6.8251        7.1654        4.7315        9.9233        8.1075
GIORNOsupp                SABATO     MERCOLEDI        SABATO       GIOVEDI       VENERDI
CountGIORNO               2.2525        2.6024        1.7077        3.1917        2.0654


Clustered Instances

0       254 ( 17%)
1       674 ( 46%)
2       339 ( 23%)
3       214 ( 14%)
```

Figura 22 - K-means su negozio 102

Il cluster del negozio 102 presenza una distribuzione di clienti “abbastanza” bilanciata , con una maggior densità nel cluster 1. Il suo profilo tipico è l’utente che effettua molte spese il Sabato, con pochi acquisti frequenti concentrati in 1,2 giorni. Negli altri cluster il cliente tipo acquista prettamente ACQUA, PANE_PROD_ESTERNA e BIBITE con regolarità.

4.2.4 Negozio 103

Tramite lo studio della curva dell'SSE al crescere del numero di cluster, abbiamo trovato il valore di k che utilizzeremo per effettuare il clustering (Figura 23).

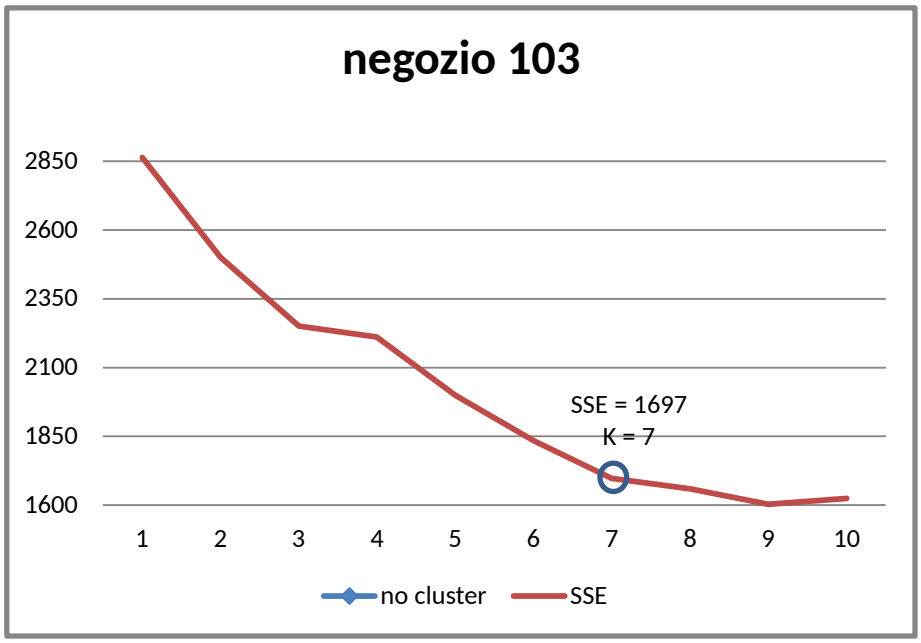


Figura 23 - curva SSE su negozio 103

In Figura 24 mostriamo i cluster ottenuti applicando il K-means con k = 7 sui profili dei clienti del negozio 103.

```
kMeans
======

Number of iterations: 10
Within cluster sum of squared errors: 1697.1042342399455
Missing values globally replaced with mean/mode

Cluster centroids:
                                                Cluster#
Attribute               Full Data                     0          1          2          3          4          5          6
                          (1719)                  (204)       (89)      (532)      (175)      (264)      (211)      (244)
=============================================================================================================================
CAT_PRODOTTOsupp           LATTE  SALUMIASERVIZIOASSISTITO     ACQUE      LATTE      ACQUE      LATTE FORMAGGIFRESCHI PANEPRODESTERNA
CountCAT_PRODOTTO         5.2414                    6.951     4.4157       3.75     6.0971     6.4318     5.9005     4.8934
CountItemset              6.5503                   9.0637     5.0562      4.438        7.4     8.4924     7.1422      6.377
GIORNOsupp                SABATO                   LUNEDI   DOMENICA     SABATO  MERCOLEDI   VENERDI     GIOVEDI    MARTEDI
CountGIORNO               2.0942                   2.6324     1.6742     1.6034     2.1771     2.5871     2.0427     2.3197


Clustered Instances

0       204 ( 12%)
1        89 (  5%)
2       532 ( 31%)
3       175 ( 10%)
4       264 ( 15%)
5       211 ( 12%)
6       244 ( 14%)
```

Figura 24 - K-means su negozio 103

I cluster ottenuti risultano abbastanza equamente distribuiti. I profili tipici d'acquisto che emergono riguardano utenti che acquistano regolarmente, LATTE(prevalentemente il sabato), ACQUA(prevalentemente mercoledì e domenica) , SALUMIASERVIZIOASSISTITO(prevalentemente lunedì), FORMAGGIFRESCHI(prevalentemente giovedì) e PANE_PROD_ESTERNA(prevalentemente il martedì).
Gli altri attributi sono piuttosto omogenei tra cluster e identificano persone che fanno acquisti in 2 giorni distinti acquistando circa 5 categorie diverse di prodotti.

4.3 Valutazioni dei risultati

Dai risultati ottenuti possiamo osservare che il clustering sui negozi 70 e 101 non produce informazioni interessanti, infatti i profili d'acquisto tipici estratti sono molto simili.
Con tutta probabilità ciò è dovuto alla dimensionalità dei dati inerenti ai due negozi(a cui si riferiscono gran parte del venduto), infatti dal negozio 102 e 103 otteniamo una segmentazione leggermente maggiore dei clienti seppur con aspetti altrettanto banali.

CONCLUSIONI

La profilazione dei clienti e la loro clusterizzazione per ogni negozio non ha prodotto risultati molto interessanti; probabilmente questo fenomeno ha diverse cause:

E' ragionevole pensare che le features utilizzateper il clustering(specialmente quelle categoriche) non siano appropriate per questo tipo di analisi;

La tecnica di clustering scelta probabilmente non è adeguata al problema (forse un algoritmo Fuzzy avrebbe prodotto risultati più interessanti).

Altra causa può essere legata alla natura stessa dei comportamenti d'acquisto delle famiglie italiane, infatti PANE, LATTE e ACQUA rappresentano la base d'acquisto per tutti i clienti e data la alta cardinalità di clienti, gli acquisti più caratteristici vengono del tutto oscurati da quelli più comuni. Per ovviare al problema in una successiva analisi si potrebbero eliminare le categorie più comuni.

APPENDICE A

Gli attributi barrati sono stati rimossi perché ritenuti non utili ai fini dell'analisi.

ARTICOLO (composta da 21026 righe):

- **ARTICOLO_ID** : (PK[2]) Codice identificativo articolo;
- ~~**ARTICOLO_ID_1**~~ : Codice identificativo articolo(attributo ridondante, verrà rimosso);
- **COD_MKT_ID**: (FK[3] alla tabella *"marketing")*;
- **DES_ART**: Descrizione articolo;
- ~~**MARCA**~~: Marca del articolo;
- ~~**SOTTOMARCA:**~~ Sottomarca dell'articolo(se disponibile);
- ~~**UDM:**~~ Unità di misura (espressa in KG = KILI oppure in NR = PEZZI);
- ~~**QTA:**~~ Descrive la quantità dell'articolo(in KILI oppure in numero di PEZZI);
- ~~**FL_COOP:**~~ Flag che indica se il prodotto è a marchio coop
- ~~**FL_BIO**~~ : Flag che indica se il prodotto è bio;
- ~~**FL_CELIACI**~~: Flag che indica se il prodotto è per celiaci;
- ~~**FL_DOP**~~: Flag che indica se il prodotto è di Denominazione di Origine Protetta;
- ~~**FL_IGP**~~: Flag che indica se il prodotto è Indicazione Geografica Protetta;
- ~~**FL_TIPICO**~~: Flag che indica se il prodotto è tipico;
- ~~**FL_SOLIDALE**~~: Flag che indica se il prodotto è equo-solidale;
- ~~**FL_DOCG:**~~ Flag che indica se il prodotto è di Denominazione di Origine Controllata e Garantita;
- ~~**FL_CARRELLO:**~~ Flag che indica se è stato utilizzato il carrello;
- ~~**PRES_MKT:**~~ Non è chiaro il significato di questo attributo, probabilmente da informazioni sulla strategia di marketing;
- ~~**RILEVANZA:**~~ Non è chiaro il significato di questo attributo.
- ~~**CAPOSTIPITE_ID:**~~ Identificativo per la tabella capostipite(non in nostro possesso)

DATA *(composta da 341 righe)* :

- **DATA_ID:** (PK) Codice identificativo data;
- ~~**DATA_ID_1:**~~ Codice identificativo data(attributo ridondante, verrà rimosso);
- **GIORNO**: Descrizione testuale della data(giorno/Mese);
- ~~**GIORNO_SETTIMANA_N**~~: Descrizione numerica del giorno all'interno della settimana(1..7);
- ~~**GIORNO_N**~~: Descrizione numerica del giorno nel mese;
- ~~**MESE_N:**~~ Descrizione numerica del mese;
- ~~**ANNO:**~~ Descrizione numerica dell'anno;
- **GIORNO_SETTIMANA**: Descrizione testuale del Giorno della settimana(LUNEDI..DOMENICA);
- **TRIMESTRE:** Descrizione testuale del trimestre;
- **DATA**: Data in formato gg-mm-aa;
- **MESE**: Descrizione testuale del mese(GENNAIO..DICEMBRE);
- ~~**TRIMESTRE_N:**~~ Descrizione numerica del trimestre;
- **SETTIMANA_ANNO:** Numero di settimana nell'anno;
- **PERIODO:** Descrizione numerica del periodo;
- **TIPOLOGIA**: Attributo categorico (festivo/feriale);
- ~~**SETTIMANA_COMMERCIALE:**~~ Descrizione numerica della settimana commerciale.

[2] Primary Key
[3] Foreign Key

CLIENTI (composta da 48712 righe):

- **CLIENTE_ID**: (PK) Codice identificativo cliente;
- ~~**FASCIA_ETA:**~~ Range d'età del cliente;
- ~~**SESSO**~~: Genere cliente;
- ~~**STATO_CIVILE:**~~ Stato civile del cliente: spostato, single, divorziato, etc;
- ~~**PROFESSIONE**~~: Professione del cliente;
- ~~**TITOLO_STUDIO:**~~ Grado di istruzione del cliente;
- ~~**FASCIA_ANNO_SOCIO:**~~ Anno (range) di sottoiscrizione al supermarket;
- ~~**FL_INVIO_RIVISTA:**~~ Flag che indica se il cliente al cliente viene inviato il volantino del supermarket.

MARKETING (composta da 2974 righe):

- **COD_MKT_ID:** (PK) Codice identificativo della tabella marketing;
- ~~**COD_MKT_ID_1:**~~ Codice identificativo della tabella marketing (attributo ridondante, verrà rimosso);
- ~~**COD_MKT :**~~ Codice di marketing;
- **AREA → MACROSETTORE → SETTORE → REPARTO → CATEGORIA → SOTTOCATEGORIA → SEGMENTO**: Insieme di attributi che descrivono la gerarchia del prodotto in ordine decrescente (dal generale allo specifico);
- ~~**COD_AREA → COD_MACROSET → COD_SETT → COD_REP → COD_CATEG → COD_SUBCATEG → COD_SEGMENTO:**~~ Insieme di attributi (espressi in maniera numerica) che descrivono la gerarchia del prodotto in ordine decrescente.

VENDUTO (composta da 18134427 righe):

- **N_SCONTRINI:** Codice identificativo della transazione;
- **DATA_ID**: (FK alla tabella "*data"*) *Data di transazione;*
- **ORARIO_ID:** (FK alla tabella "*orario"*) Ora della transazione;
- **NEGOZIO_ID:** ID del negozio alla quale la transazione si riferisce;
- **CLIENTE_ID**: (FK alla tabella *"clienti"*) ID del cliente;
- **ARTICOLO_ID:** (FK alla tabella "*articolo")* ID del prodotto acquistato;
- **COD_MKT_ID:** (FK alla tabella "*marketing");*
- ~~**IMPORTO**~~ : Importo pagato;
- ~~**QTA_PEZZI:**~~ Numeri di elementi, di un certo prodotto, acquistati;
- ~~**QTA_PESO:**~~ Quantità in Kg del prodotto acquistato;
- ~~**VOLUME**~~: Volume del prodotto acquistato.

~~*ORARIO (composta da 795 righe):*~~

- **ORARIO_ID:** (PK) Codice identificativo orario;
- ~~**ORARIO_ID_1:**~~ Codice identificativo orario (attributo ridondante, verrà rimosso);
- ~~**ORARIO:**~~ Descrizione testuale dell'orario (formato hh:mm);
- ~~**MINUTO:**~~ Descrizione numerica dei minuti (1..60);
- ~~**ORA:**~~ Descrizione numerica dell'ora (7..20);
- ~~**FASCIA_FASCIA_ORARIA:**~~ Descrizione della fascia oraria;
- ~~**FASCIA_ORARIA_N:**~~ Descrizione numerica della fascia oraria.