Relazione Progetto IDM-2025: Analisi dei Pattern di Acquisto Retail

Michela Maria Tasca - 1000084036

19 dicembre 2025

Indice

1 Introduzione

L'obiettivo del progetto è l'analisi di un dataset transazionale composto da circa 2.3 milioni di record. L'analisi copre l'esplorazione dei dati, il mining di regole di associazione e la segmentazione della clientela tramite tecniche di riduzione dimensionale e clustering.

2 Istruzioni per la Riproducibilità

2.1 Ambiente di Esecuzione e Requisiti

Per garantire il corretto funzionamento degli algoritmi, è necessaria l'installazione delle seguenti librerie:

- `pandas`, `numpy`: per la manipolazione e il calcolo numerico;
- `matplotlib`: per la generazione degli output grafici;
- `scikit-learn`: per l'implementazione della SVD e degli algoritmi di preprocessing;
- `mlxtend`: per il mining delle regole di associazione (Apriori e FP-Growth);
- `hdbscan`: per l'analisi di clustering basata sulla densità.

2.2 Configurazione e Avvio

Il codice segue rigorosamente il paradigma *Object-Oriented* ed è organizzato secondo la struttura di directory richiesta. Per riprodurre i risultati:

1. Posizionare il dataset originale nominato `AnonymizedFidelity.csv` nella cartella radice *first_classwork/*.
2. Accedere alla directory dei sorgenti: `cd first_classwork/src/`.
3. Eseguire lo script principale: `python main.py`.

Il sistema gestisce automaticamente la creazione della directory *results/*, all'interno della quale verranno salvati tutti i grafici in formato `.png` e i risultati tabellari in formato `.csv`.

2.3 Note sulle Prestazioni e Tempi di Esecuzione

Data la mole del dataset, l'esecuzione dell'algoritmo *Apriori* sarà computazionalmente costosa.

Inoltre, l'intera pipeline di analisi richiede tempi di elaborazione significativi. In particolare:

- **Tempo stimato**: Circa 10 minuti per il completamento di tutte le Task.

- **Configurazione**: Per ottimizzare il rapporto tra tempi di calcolo e qualità dei risultati, il parametro `min_support` è stato impostato a 0.15 per le regole di associazione. Qualora l'ambiente di esecuzione dovesse riscontrare errori di memoria (*MemoryError*), si consiglia di elevare tale soglia a 0.05 nella chiamata al metodo `assoc_analyzer.run_task3_4`.
- **Comportamento atteso**: È normale che il sistema sembri inattivo durante le fasi di *Frequent Itemset Mining* e di costruzione della matrice Cliente-Prodotto. Il programma termina correttamente generando tutti gli output previsti nella cartella `results/`.

3 Preprocessing e Analisi Esplorativa (Task 1 & 2)

Il dataset è stato preventivamente pulito rimuovendo i record relativi a resi (quantità negative) ed escludendo le voci "shopper", non rilevanti ai fini dell'analisi.

3.1 Stratificazione Temporale

Sono state eseguite analisi di frequenza stratificate per:

- **Mesi**: Suddivisione in tre range (Gen-Mag, Mag-Set, Ott-Dic).
- **Fasce Orarie**: Mattina (08:30-12:30), Pranzo (12:30-16:30) e Sera (16:30-20:30).

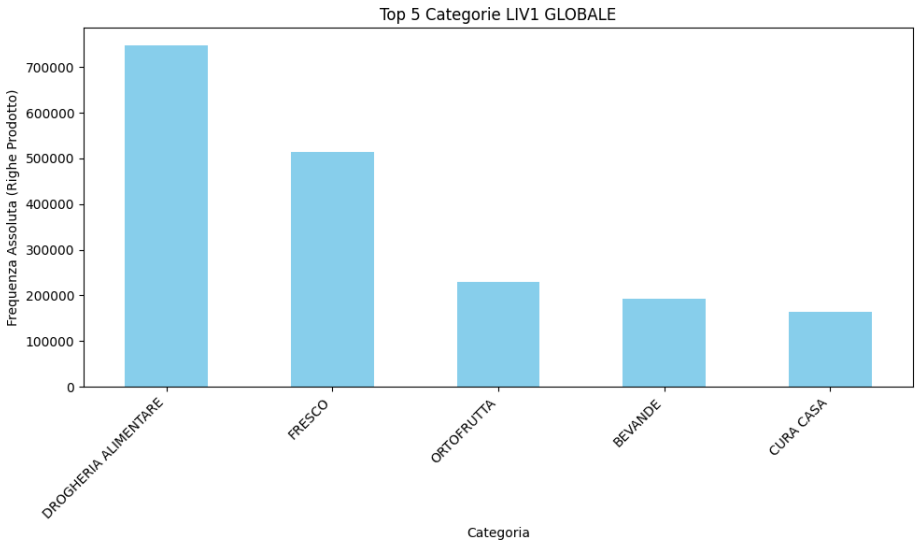


Figura 1: Visualizzazione delle prime 5 categorie (livello 1)

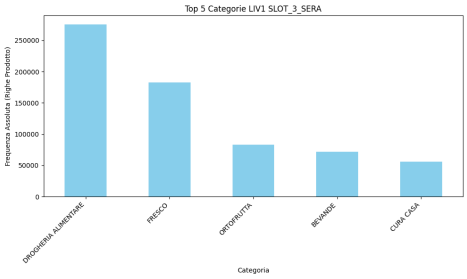
Dall'analisi globale del primo livello di merchandising (Fig. 1), emerge chiaramente la natura alimentare del dataset, con le categorie *Drogheria Alimentare* e *Fresco* che rappresentano oltre la metà delle transazioni complessive.

Passando alla stratificazione oraria, l'analisi comparativa dei 12 grafici relativi alle frequenze assolute di vendita, suddivise per slot temporali (Mattina, Pranzo, Sera) e livelli gerarchici (LIV1-LIV4), permette di trarre le seguenti conclusioni:

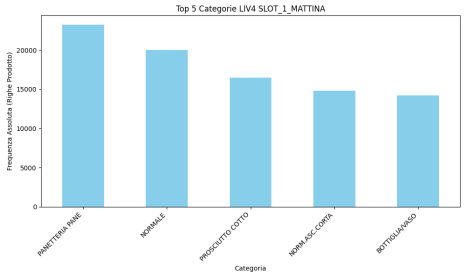
- **Stabilità del paniere macro (LIV1 e LIV2):** Si osserva una forte resilienza delle categorie principali rispetto all'orario. *Drogheria Alimentare* e *Fresco* dominano costantemente il LIV1. Analogamente, al LIV2, le categorie *Formaggi*, *Salumi* e *Verdura* mantengono le prime posizioni in tutti gli slot.
- **Variazioni volumetriche:** Lo slot **Sera** registra i volumi di righe prodotto più elevati in assoluto (picco di ≈ 275.000 per la Drogheria), indicando che la maggior parte delle transazioni si concentra nella fascia finale della giornata. Lo slot **Pranzo** risulta essere quello con la frequenza minore.
- **Comportamenti specifici per slot (LIV3 e LIV4):**

  - **Mattina:** È l'unico slot in cui la categoria *Panetteria Pane* (LIV4) appare come leader indiscussa, a conferma di un acquisto mirato al prodotto fresco di giornata.
  - **Sera:** Al LIV3 compare la categoria *Latte UHT*, assente nella top 5 degli altri slot. Questo suggerisce una tendenza all'acquisto di scorte per la colazione del giorno successivo durante la spesa serale.
  - **Pranzo:** Si osserva una rilevanza maggiore della categoria *Avicolo* al LIV4, probabilmente legata alla preparazione dei pasti pomeridiani.

- **Prodotti Anchor:** La *Pasta di Semola* (LIV3) e il *Prosciutto Cotto* (LIV4) mostrano una presenza costante nelle top 5 di ogni fascia oraria, qualificandosi come prodotti ad alta rotazione indipendentemente dalle abitudini temporali del consumatore.



(a) Picco volumi (Sera - LIV1)



(b) Focus Panetteria (Mattina - LIV4)

Figura 2: Selezione di grafici rappresentativi delle dinamiche di acquisto.

4 Regole di Associazione (Task 3 & 4)

Per l'estrazione delle regole di associazione sono stati applicati gli algoritmi *Apriori* e *FP-Growth*. Entrambi hanno prodotto un set di 52 regole significative, confermando la coerenza dei risultati. L'analisi si è focalizzata sulle regole con i valori di **Lift** e **Confidence** più elevati.

4.1 Interpretazione delle Metriche

- **Supporto:** indica la rarità o comunanza della combinazione nel dataset.

- **Confidence:** esprime la forza del legame predittivo tra antecedente e conseguente.
- **Lift:** misura quanto la presenza dell'antecedente aumenti la probabilità di acquisto del conseguente rispetto al caso di acquisti indipendenti.

4.2 Risultati

I risultati prodotti dai due algoritmi sono identici: questo è corretto, poiché i due algoritmi differiscono solo per l'efficienza computazionale, ma producono lo stesso set di regole a parità di parametri (supporto minimo 0.015 e confidenza minima 0.5). Le regole estratte evidenziano comportamenti d'acquisto ben definiti:

1. **Segmento Pasta:** La regola {Pasta Lunga, Normale} $\Rightarrow$ {Pasta Corta} presenta il Lift più alto (4.58) e una confidenza del 61%. Questo suggerisce che il consumatore medio tende a fare scorta di diversi formati di pasta nello stesso atto d'acquisto.
2. **Segmento Macelleria:** Si osserva una forte correlazione tra tagli di carne diversi, in particolare {Suino, Normale} $\Rightarrow$ {Bovino} con un Lift di 3.34. Questo pattern indica una spesa programmata per il consumo settimanale.
3. **Salumi e Gastronomia:** Il *Prosciutto Cotto* appare frequentemente come prodotto *pivot* che si associa sia alla pasta che ad altri affettati, confermandosi un prodotto *entry-level* nel carrello della spesa.

In conlusione, è possibile affermare che i valori di Lift superiori a 3 per le top-rule indicano che le associazioni trovate non sono casuali, ma riflettono abitudini consolidate.

5 Clustering e Riduzione Dimensionale (Task 5)

Questa fase ha rappresentato la sfida tecnica maggiore a causa dell'estrema sparsità della matrice Cliente-Prodotto (9375×19422).

5.1 Evoluzione della Metodologia

Inizialmente è stata testata la **PCA** (Principal Component Analysis), che tuttavia ha mostrato limiti evidenti:

- Una varianza spiegata di soli 0.17 con 100 componenti.
- Un clustering K-Means che convergeva verso un unico cluster dominante (99.9% dei clienti), fallendo nella segmentazione.

5.2 Soluzione Finale: Truncated SVD e HDBSCAN

Per gestire meglio la natura sparsa dei dati, si è passati all'utilizzo della **Truncated SVD** (Latent Semantic Analysis) abbinata a una normalizzazione tramite **MaxAbsScaler**.

- **MaxAbsScaler**: Utilizzato per scalare i dati preservando gli zeri della matrice originale.
- **HDBSCAN**: Scelto al posto di K-Means per la sua capacità di gestire il rumore e identificare cluster di forma arbitraria basati sulla densità.
- **Risultato**: L'analisi ha isolato 2 cluster densi, classificando la restante parte dei dati (8747 punti) come rumore statistico, confermando l'eterogeneità dei pattern d'acquisto.

6 Conclusioni

L'integrazione di tecniche diverse ha permesso di evidenziare che, nonostante esistano delle regole di associazione forti tra prodotti, la base clienti globale risulta estremamente varia, rendendo la segmentazione rigida meno efficace rispetto a un'analisi basata sulla densità.