

FIRST CLASSWORK – DATA MINING INTRODUCTION

Santo Lo Bianco 1000061604

Il dataset analizzato è **AnonymizedFidelity.csv**, contenente informazioni sugli acquisti effettuati dai clienti di un supermercato. Ogni osservazione del dataset descrive un singolo prodotto acquistato all'interno di uno scontrino e include, in particolare, le seguenti informazioni principali:

- identificativo scontrino;
- descrizione del prodotto acquistato in vari livelli;
- prezzo;
- data e ora;
- tessere fedeltà;

L'obiettivo dell'elaborato è quello di applicare diverse tecniche di data mining al fine di analizzare i comportamenti di acquisto dei clienti. In particolare, le attività svolte sono:

- analisi delle frequenze;
- analisi stratificata per osservare le variazioni temporali;
- estrarre le regole di associazione(mediane algoritmi apriori e fpgrowth);
- clustering con PCA, per individuare gruppi di clienti con comportamenti di acquisto simili.

Prima di procedere con le analisi, il dataset è stato sottoposto a una fase di preprocessing, migliorando la qualità dei dati e rendendoli adeguati al contesto dell'analisi.

Sono state rimosse alcune feature non rilevanti(non coinvolte nell'analisi) al fine di ridurre la complessità computazionale(dato che il file pesa 850 MB+). Successivamente come richiesto dalla traccia sono stati esclusi dal dataset gli articoli aventi il valore “**shopper**” nella variabile `descr_liv4`.

Le feature contenenti informazioni temporali, ossia data e ora, sono state convertite da **object(stringhe)** a **datetime**. Questa conversione è necessaria al fine di estrarre il mese e ore/minuti.

Infine, si è effettuata un'operazione di feature engineering, creando le feature derivate:

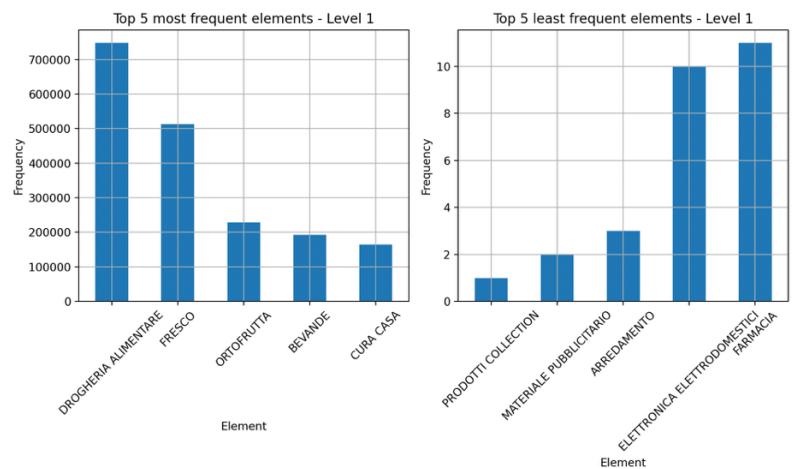
- **fascia_mese**, che suddivide gli acquisti in tre intervalli quali:
 - **RANGE 1**: da gennaio a metà maggio;
 - **RANGE 2**: da metà maggio a settembre;
 - **RANGE 3**: da ottobre a dicembre.
- **fascia_oraria**, che suddivide la giornata in tre fasce temporali:
 - **SLOT 1**: 08:30 – 12:30;
 - **SLOT 2**: 12:30 – 16:30;
 - **SLOT 3**: 16:30 – 20:30.

Lavorando su una versione più pulita e compatta permette di ottenere risultati non ambigui in minor tempo.

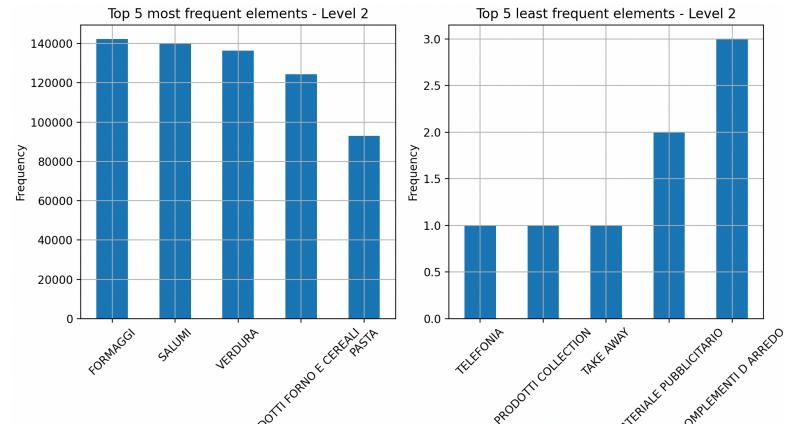
L'analisi delle frequenze di acquisto dei prodotti ha permesso di identificare categorie dei prodotti maggiormente acquistate e quelle meno rappresentate, fornendo una prima visione d'insieme delle preferenze dei consumatori.

I risultati ottenuti sono i seguenti:

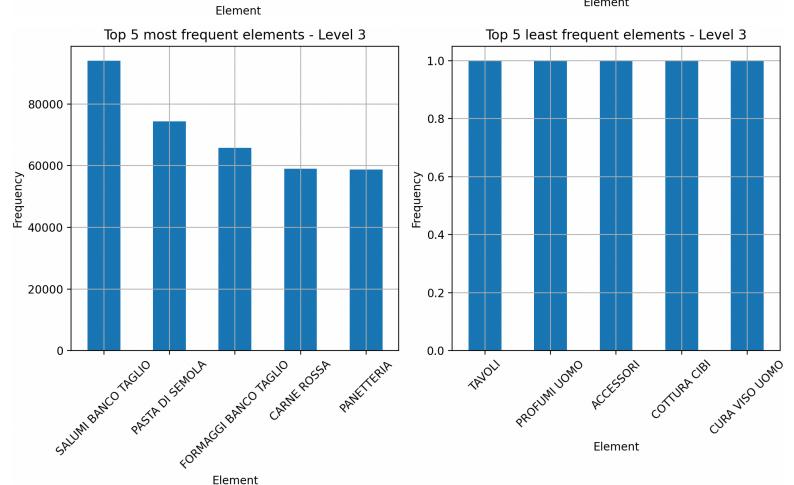
Il livello 1 descrive il prodotto in modo generale e, di conseguenza, sotto la stessa categoria possono essere raggruppati molti prodotti differenti. Dall'analisi delle frequenze emerge che le categorie tipicamente vendute nei supermercati risultano essere le più presenti nel dataset, come ad esempio la drogheria alimentare (circa 750.000 occorrenze). Al contrario, categorie solitamente associate ad altre tipologie di esercizi commerciali, come la farmacia, mostrano una frequenza nettamente inferiore.



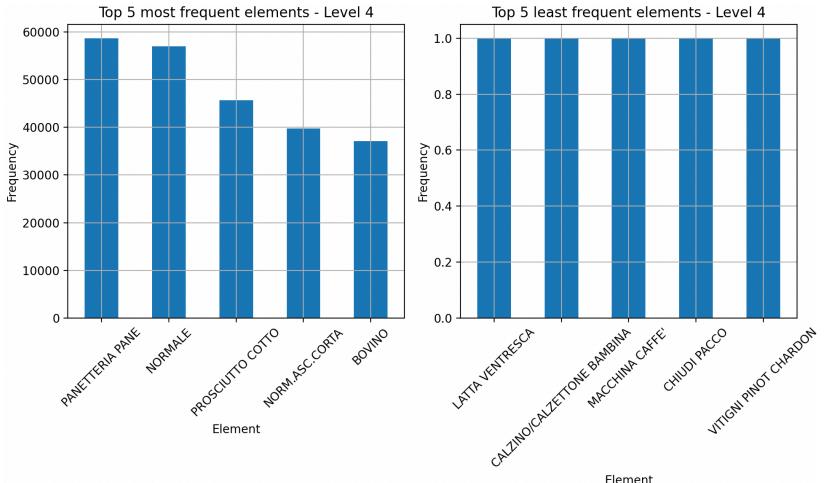
Il livello 2 offre un grado di dettaglio maggiore rispetto al precedente, permettendo di distinguere meglio la tipologia del prodotto. Dall'analisi emerge che i prodotti freschi e prima necessità dominano la classifica (superando le 120.000 occorrenze). Al contrario, le categorie meno frequenti riguardano articoli marginali non pertinenti con ciò che vende in genere un supermercato(telefonia).



Il livello 3 scende ulteriormente nel dettaglio merceologico. Dall'analisi delle frequenze emerge una chiara prevalenza dei prodotti da banco assistito e dei beni di largo consumo, evidenziando l'importanza del servizio al banco. Al contrario, articoli molto specifici, non del settore, sono ignorati dai clienti, indicando che saranno prodotti residuali.



Ultimo è il **livello 4** che indica direttamente il nome del prodotto(non prendendo una categoria più ampia). Dalle analisi emerge che gli alimenti di consumo quotidiano dominano la classifica. Al contrario oggetti estremamente particolari registrano un solo acquisto.



In generale si è visto come generi alimentari di consumo quotidiano prevalgono durante gli acquisti e come articoli di nicchia o che solitamente sono acquistati in altre attività vengano ignorati dai clienti.

Ma durante l'anno o durante il giorno, gli acquisti sono gli stessi? O a seconda del periodo(sia breve che lungo) la tipologia del prodotto acquistato cambia?

Analizziamo come cambia il Liv1 e Liv4 durante l'anno e verifichiamo se ci sono variazioni.



L'analisi stratificata temporale evidenzia che le abitudini sono sostanzialmente stabili con drogheria alimentare e fresco predominante in generale.

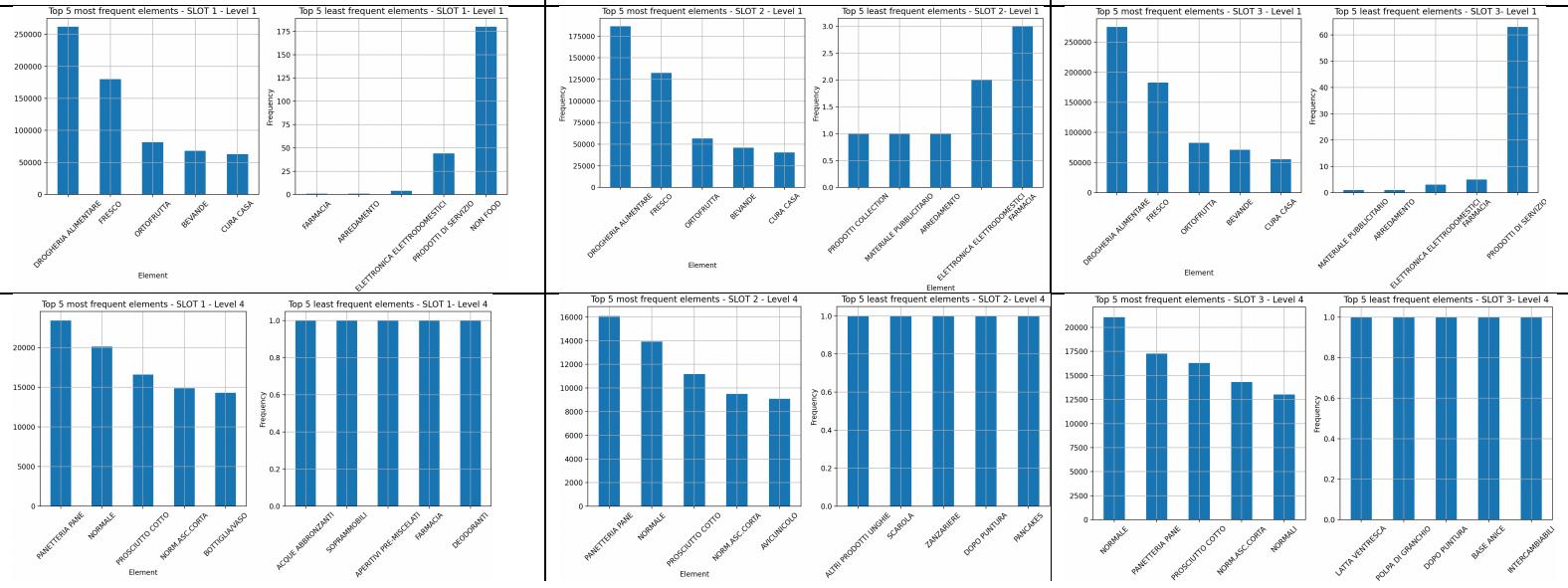
Se si analizza il livello 4, quello che descrive nel dettaglio i prodotti, possiamo vedere che i prodotti di uso quotidiano(pane, prosciutto,...) sono quelli più acquistati indipendentemente dalla stagione.

In estate si registra un leggero aumento nell'acquisto di bevande, indicando che, pur con una struttura di spesa stabile, le frequenze di alcuni prodotti possono variare stagionalmente.

SLOT 1(08:30 - 12:30)

SLOT 2(12:30 – 16:30)

SLOT 3(16:30 – 20:30)



L'analisi per fascia oraria fornisce un'interessante interpretazione del comportamento dei clienti durante la giornata. Si ha che:

- Mattina(slot1): fascia dominante è quella del pane e il fresco di giornata;
- Primo pomeriggio(slot 2): si registra un calo nelle vendite e si hanno gli stessi prodotti più venduti però con frequenze minori;
- Tardo pomeriggio(slot 3): il volume degli acquisti risale e la spesa è più strutturata, dato che la categoria predominante è “Normale”(descr_liv4). “Normale” identifica la variante standard di prodotti di largo consumo(caffè, burro o carte igienica). Ciò indica una spesa più orientata al rifornimento piuttosto che al consumo diretto.

Dopo aver analizzato la distribuzione delle frequenze e le variazioni temporali, l'attenzione si sposta sull'individuazione di relazioni tra prodotti acquistati insieme. Secondo quanto studiato durante il corso abbiamo visto due algoritmi principali:

- **Apriori;**
- **Fpgrowth;**

Indipendentemente dall'algoritmo utilizzato si è scelto di analizzare le regole di associazione A->B e valutarle attraverso le metriche:

- Support, che indica la frequenza con cui una coppia appare nelle stesse transizioni;
- Confidence, che misura la probabilità di acquistare B dato A;
- Lift, che indica quanto una regola di associazione è realmente utile.

antecedents	consequents	support	confidence	lift
PROSCIUTTO COTTO	PANETTERIA PANE	0.0768446084673961	0.3064069017332055	1.1289223525232612
NORMALI	PROSCIUTTO COTTO	0.059685006627335814	0.3444706941163236	1.3735276973511006
NORMALI	NORMALE	0.06703355251189022	0.38688266396670046	1.5530199276436956
SALAME	PROSCIUTTO COTTO	0.059756477895885854	0.6070226387697182	2.420416080410627
LATTE VACCINO	PROSCIUTTO COTTO	0.07662369727369597	0.4296800991036945	1.713288030592798
LATTE VACCINO	NORMALE	0.06539621072328924	0.36672010493332363	1.4720836157136739
PROSCIUTTO COTTO	PASTE FILATE STAGIONATE	0.07723445175039634	0.30796134614886395	2.2726177053643726
MORTADELLA	PROSCIUTTO COTTO	0.05291472827923175	0.5902732478074944	2.3536302760060064
PROSCIUTTO COTTO	NORMALE	0.08028822413389818	0.3201378274048551	1.2850935745083967
PANETTERIA PANE	NORMALE	0.0628427372196377	0.23153711727670984	0.9294336257745979
PROSCIUTTO COTTO	AVICUNICOLO	0.05081607193908049	0.20262182958107725	1.1887744652599566
AVICUNICOLO	PANETTERIA PANE	0.050822569327130494	0.29817405557885107	1.0985893411062124
PROSCIUTTO COTTO	GRANA E SIMILI	0.06643579281128986	0.26490323583512526	1.7321236817449424
GRANA E SIMILI	NORMALE	0.053337058502482	0.3487552043504121	1.3999691189891557
BOVINO	AVICUNICOLO	0.06234243833978741	0.3673149069749636	2.155022403183117
PANETTERIA PANE	BOVINO	0.06142630662473685	0.22631843535297916	1.333443754241877
BOVINO	PANETTERIA PANE	0.06142630662473685	0.36191715795115226	1.3334437542418773
BOVINO	PROSCIUTTO COTTO	0.05146581074408088	0.30323099303269274	1.2090902789107405

Emergono legami significativi tra prodotti come ‘salumi’, ‘pane’ e ‘formaggi’. Il ‘prosciutto cotto’ risulta essere un prodotto centrale, frequentemente associato ad altri ‘salumi’ e ‘formaggi’.

Possiamo quindi analizzare il lift(se >1 si ha un’influenza positiva) del prodotto che emerge essere in maggior presenza nelle regole di associazione, ossia il ‘prosciutto cotto’. I valori di lift con gli altri prodotti sono:

- **lift(PROSCIUTTO COTTO – SALAME) = ‘2.43’** -> indica una tendenza di acquisto congiunto significativo;
- **lift(PROSCIUTTO COTTO – PASTE SFILATE STAGIONATE) = ‘2.27’** -> tendenza di acquisto congiunto significativo;
- **lift(PROSCIUTTO COTTO - MORTADELLA) = ‘2.35’** -> tendenza di acquisto congiunto significativo;

Altri acquisti congiunti significativi:

- **lift(BOVINO - AVICUNICOLO) = ‘2.15’** -> acquisto combinato di carni;
- **lift(GRANA E SIMILI - NORMALE) = ‘1.39’**-> acquisto congiunto moderato;
- **lift(PANETTERIA E PANE – BOVINO) = ‘1.33’**-> acquisto congiunto rilevante.

Nel complesso, salumi e formaggi rappresentano il nucleo centrale delle associazioni, fungendo da elementi chiave attorno ai quali si costruiscono sia acquisti immediati sia spese più articolate. Si sottolinea infine che, data l’elevata numerosità del dataset, i valori di supporto risultano relativamente bassi.

Dopo che si sono studiate le principali relazioni tra prodotti mediante le regole di associazione, l'analisi si è concentrata sull'individuazione di gruppi di clienti(tesserati) con comportamenti di acquisti simili.

Dato che il dataset è molto grande, si è scelto di applicare una tecnica di riduzione della dimensionalità, ossia la PCA(Principal Component Analysis). Si è creata quindi una matrice cliente – prodotto con:

- righe -> tessera fedeltà(unico modo per ricondurre l'acquisto ad un cliente);
- colonne -> prodotti.

I valori sono le frequenze di acquisto di un prodotto per un dato cliente. La PCA ci permette di ridurre il numero di variabili, mantenendo solo la varianza e di eliminare correlazioni e rumore.

La maledizione della dimensionalità del KMeans è mitigato dalla PCA. Sui dati ridotti è stato poi applicato il KMeans e per la scelta del numero ottimale sono stati utilizzati due criteri:

- metodo del gomito(elbow), basato sull'analisi del numero di cluster e ...;
- silhouette score, che misura la separazione tra i cluster.

La soluzione ottimale trovata è di 3 cluster. La maggior parte dei clienti appartiene al cluster 0(circa l'88%) e il restante alle restanti due classi.

Confrontando ciascun cluster con le frequenze dei prodotti acquistati si è osservato che:

- Cluster 0 – Cliente standard(88,1%): maggioranza dei clienti con acquisti bilanciati;
- Cluster 2 – Cliente con acquisti consistenti(11,5%): minoranza probabilmente famiglie o gestori di attività;
- Cluster 1 – Pet lovers(0,4): quelli in minor parte, che hanno un focus sui prodotti per gli animali.

I gruppi che si sono trovati sono suddivisi principalmente per intensità di acquisto e tipologia.

Le informazioni estratte, ci permettono di avere informazioni su:

- la variabilità delle frequenze dei prodotti acquistati nel tempo;
- le associazioni tra prodotti;
- l'individuazione di gruppi di clienti con comportamenti simili.

Questi risultati offrono spunti concreti per ottimizzare la gestione del punto vendita: è possibile identificare prodotti da promuovere o da ridurre, progettare offerte mirate e pianificare promozioni personalizzate per specifiche categorie di clienti, massimizzando così l'efficacia commerciale.