POLITECNICO DI TORINO

Corso di Laurea Magistrale in Ingegneria Gestionale
AA. 2022/2023

Sessione di Laurea 29 marzo 2023

Progettazione e sviluppo di metodologie data-driven e dashboard informative per il supporto decisionale. Caso di studio: dipartimento di merchandising di un'azienda di moda.

Relatore:
Tania Cerquitelli

Candidato:
Asia Sofia Rosso

A mia nonna, il mio esempio di vita

Sommario

Indice delle figure

Indice delle tabelle

1. Introduzione

L'oggetto della tesi è frutto di una collaborazione con un'azienda di consulenza informatica che sta emergendo sempre più sul territorio nazionale. L'esperienza formativa si è basata sulla possibilità di offrire supporto decisionale a una nota azienda di moda, cliente dell'azienda di consulenza, in ambito merchandising management, team che si occupa di definire stagionalmente il pool di prodotti da proporre sul mercato, stabilirne la fascia di prezzo in base alla clientela e suggerire le eventuali campagne promozionali.

L'obiettivo del lavoro è quello di aiutare il cliente a selezionare i prodotti da riproporre e quelli da scartare durante la pianificazione stagionale della collezione. Più precisamente il supporto al cliente viene fornito da due punti di vista: estraendo conoscenza dai dati storici per evidenziare le informazioni più importanti e generando una dashboard intuitiva in modo da rendere i KPI facilmente accessibili ai vari attori aziendali.

L'approccio che si utilizza in questo operato segue la prospettiva data-driven, poiché si ritiene molto importante assumere decisioni dai dati raccolti. I dati sono stati analizzati sfruttando due tecniche differenti per valutarne l'efficacia e la possibile applicabilità ad un contesto aziendale: analisi ABC e KDD (Knowledge Discovery in Databases). L'analisi ABC è un metodo più semplice e più facilmente applicabile in un contesto aziendale ma in grado di estrarre meno informazioni rispetto al più completo KDD. I due metodi sono poi confrontati per definire quale possa essere considerato migliore in base al contesto specifico.

2. Merchandising Management

In primo luogo è importante definire brevemente il ruolo del merchandising management, funzione aziendale che si desidera supportare al fine di assumere decisioni più consapevoli e mirate.

Il merchandising management è di cruciale importanza per le aziende basate su prodotti perché si occupa dell'ottimizzazione dei profitti aziendali cercando di massimizzare le vendite [1]. Più nello specifico, chi gestisce il merchandising deve possedere una serie di competenze propedeutiche al raggiungimento dei target aziendali tra cui: profonda conoscenza del mercato e del consumatore, saper pianificare il pool di prodotti da inserire nella collezione di ogni mercato e competenze analitiche per carpire le informazioni dai dati storici. Saper osservare il settore di riferimento è molto utile per prendere consapevolezza delle opportunità e dei fattori di rischio presenti ed anche per elaborare strategie di marketing vincenti. Invece, essere consapevoli delle abitudini dei vari tipi di clientela è fondamentale per differenziare correttamente l'offerta in termini di prezzo, stile, materiale e colore. Inoltre, il merchandiser deve pianificare stagionalemente la collezione da lanciare sul mercato e in questo contesto un compito molto delicato è costituito dalla capacità di definire il giusto rapporto tra prodotti nuovi, riproposti (carry-over) e continuativi (still-valid). Oltre a ciò, è importante che sappia analizzare correttamente i dati del venduto, sia di sell-in (prodotti venduti dai produttori ai rivenditori) che di sell-out (vendita al consumatore finale) in modo da comprendere successi, fallimenti e le rispettive cause [1], [2]. Infine, alle attività precedenti bisogna aggiungere anche la capacità di progettare la disposizione dei prodotti nei negozi, la strategia di determinazione dei prezzi e delle campagne promozionali in collaborazione con il marketing department [3]

La figura del merchandiser ha assunto sempre più importanza strategica nel canale retail. È una figura professionale dinamica e curiosa con una forte sensibilità verso il mercato esterno e con una grande capacità di comunicazione in quanto interagisce costantemente con tutti i dipartimenti aziendali [4].

Un riassunto delle principali attività che svolge il merchandiser è visualizzabile in figura 2-1.

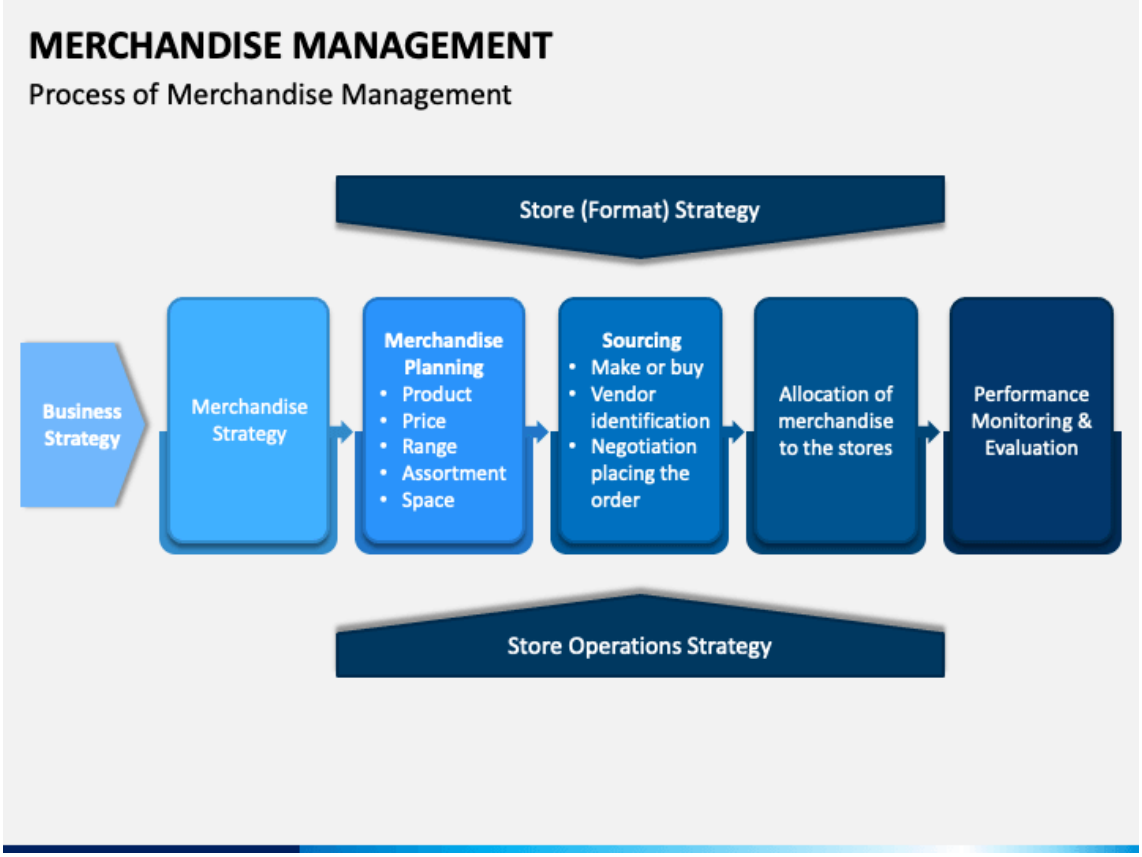


Figura 2-1. Merchandise Management. Fonte: SketchBubble

In quest'ottica si è deciso di provare a mettersi nei panni del merchandiser aziendale e svolgere una breve analisi di mercato (vedi capitolo 5) per comprendere meglio il contesto interno ed esterno all'azienda in modo da poter essere più efficaci durante il processo di estrazione della conoscenza.

3. Analisi ABC

Per quanto riguarda l'approccio metodologico che si è voluto seguire, si è stabilito di approfondire l'analisi ABC poiché deriva dal principio di Pareto, un metodo già largamente utilizzato nell'azienda. L'obbiettivo è quello di suddividere gli elementi (es: prodotti, clienti, ecc.) nelle tre categorie A, B e C, con lo scopo di valutare il loro impatto sulla metrica target (es: fatturato) al fine di stabilire quali siano gli oggetti più rilevanti e quali, invece, quelli più critici. Per le aziende categorizzare le diverse parti di un insieme può essere utile per definire gli elementi di priorità in quanto sono le più fruttuose [5].

Prima di descrivere l'analisi ABC bisogna considerare l'enunciato del teorema di Pareto, il quale deve il suo nome a un economista italiano, Vilfredo Pareto. Tale principio afferma che, analizzando grandi quantità di dati, la maggior parte degli effetti che si possono verificare è dovuto a un numero di cause ristretto o, in altre parole, circa l'80% degli effetti è imputabile al 20% di cause [6]. Bisogna però fare attenzione a non trascurare la parte che contribuisce meno al fatturato ma, anzi, potrebbe essere utile intraprendere azioni commerciali per ottenere una crescita economica sul periodo medio-lungo. È importante sottolineare che la divisione 80/20 è un valore di riferimento, sebbene nella pratica la proporzione possa variare. Ciò che non varia è il fatto che il massimo valore di efficienza può essere ottenuto con un numero limitato di risorse [7].

L'analisi ABC si può svolgere ordinando in un foglio di calcolo tutti gli articoli in modo decrescente in base alla variabile di riferimento (es. fatturato). Le soglie delle tre categorie possono essere stabilite a piacere, ad esempio:

- Categoria A: fatturato minore del 80%;
- Categoria B: fatturato dall'80% al 95%;
- Categoria C: fatturato maggiore del 95%.

Un esempio della suddivisione degli articoli in base al fatturato utilizzando l'analisi ABC è in figura 3-1.

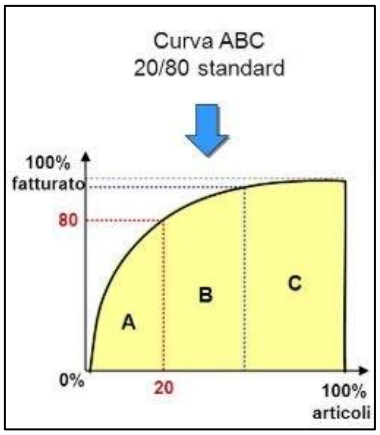


Figura 3-1. Analisi ABC. Fonte: ilprogressonline

Se i dati che si possiedono coprono uno storico abbastanza ampio le quantità medie di ogni categoria possono essere statisticamente significative [8]. L'obiettivo dell'analisi ABC consiste nel cercare di capire come ottenere il miglior risultato possibile impiegando il minor sforzo possibile o il minor numero di risorse e quindi aumentare la produttività [9]. L'analisi ABC è un modello univariato che presenta dei limiti, uno tra questi è l'instabilità. Supponendo di voler categorizzare i clienti in base alla quantità di fatturato che generano per l'azienda, è possibile che i clienti cambino categoria nel corso del tempo generando problemi sulle azioni correttive che si applicherebbero alle categorie sbagliate. Un'altra criticità consiste nel fatto che nell'ambito della vendita al dettaglio bisogna fare attenzione ad alcuni parametri, tra cui la stagionalità oppure il lancio di nuovi prodotti che, essendo appena introdotti, hanno un basso volume di vendita. L'analisi ABC attribuisce l'importanza di un prodotto secondo la frequenza delle vendite o il fatturato che genera ma bisogna fare attenzione perché la frequenza non equivale sempre all'importanza economica. Nel mondo della vendita al dettaglio, ad esempio, subentra l'effetto merceologico ovvero il fatto che alcuni prodotti vengono più pubblicizzati o messi in vetrina rispetto ad altri per attrarre la clientela e quindi hanno maggiori possibilità di essere venduti [8].

L'analisi ABC verrà utilizzata, in ogni area di mercato, per suddividere i prodotti venduti al cliente in tre categorie in base al fatturato che generano. Successivamente ogni categoria verrà analizzata cercando di far emergere le caratteristiche che la definiscono, come il prezzo medio, il costo medio, lo sconto medio o anche variabili qualitative come il colore, il materiale e lo stile più frequente.

I risultati ottenuti dall'analisi ABC sono poi confrontati con quelli ottenuti dal KDD per valutare pregi, limiti di entrambi e definire quale possa essere la tecnica migliore da applicare in azienda per analizzare il venduto, osservare le performance dei prodotti a fine stagione e assumere scelte consapevoli.

4. Business intelligence

In tutti i settori dirigenziali sta crescendo sempre di più l'importanza di assumere decisioni basate sui dati. La possibilità di accedere costantemente ai dati ha reso i big data parte integrante degli studi di management. L'obiettivo dei big data analytics è quello di estrarre conoscenza utile dal volume e dalla varietà dei dati per supportare la fase decisionale, ovvero decision making, in modo da assumere decisioni informate. La rapida e costante crescita dei dati ha generato l'esigenza di creare diversi strumenti per interrogare i database aziendali. È possibile svolgere tre diversi tipi di analisi utilizzando i big data:

- Analisi descrittive come report, dashboard e visualizzazioni;
- Analisi di scoperta che catturano i segnali precoci attraverso riassunti, estrazione di caratteristiche da immagini, ecc.;
- Analisi predittive guidate da modelli econometrici fino a complessi modelli di auto-apprendimento [10]

Con il termine Business Intelligence si intende l'insieme di metodi matematici per analizzare i dati grezzi di business, con l'obiettivo di estrarre informazioni e conoscenze quantitative dai dati storici collezionati solitamente all'interno di un data warehouse e, successivamente, di presentare i risultati nel modo più semplice e intuitivo possibile [11]–[13]. La business intelligence può essere scomposta in quattro passaggi fondamentali:

1. Raccolta e trasformazione dei dati (metodo ETL - Extract, Transform, and Load) per aggregare dati da più fonti;
2. Individuazione delle tendenze, di modelli o valori anomali tramite algoritmi di data mining, navigazione analitica sui dati storici, analisi dei KPI;
3. Visualizzazione dei dati sintetici di business, misurati tramite KPI, per semplificare la condivisione delle informazioni;
4. Processo decisionale in base alle informazioni ottenute in tempo reale per eliminare inefficienze o adattarsi al mercato.

In particolare, per l'obiettivo che questa tesi vuole conseguire si cercherà di migliorare il lavoro svolto dalla funzione merchandising analizzando le performance di vendita e le tendenze di acquisto dei consumatori [12]. Per questo motivo è possibile adattare il

termine ‘Business Intelligence’ al contesto e trasformarlo in Demand Intelligence. Con questa espressione si fa riferimento all’applicazione di modelli matematici di Data mining al contesto del Demand Planning [11].

5. Knowledge discovery in databases (KDD)

L'introduzione di data warehouse aziendali e l'utilizzo delle metodologie di Business Intelligence costituiscono validi strumenti a supporto del knowledge discovery in databases (KDD), perché permettono di:

- identificare i dati rilevanti per i processi decisionali reperendo le informazioni nei sistemi OLTP transazionali e convogliare verso i sistemi multidimensionali OLAP;
- rappresentare e visualizzare i dati multidimensionali secondo differenti prospettive di analisi;
- estrarre conoscenza attraverso l'identificazione di pattern significativi e ricorrenti nei dati storici (algoritmi di Data Mining).

La gerarchia di metodologie di data warehousing e successivamente di business intelligence rappresenta perciò il percorso logico sequenziale di acquisizione, trasformazione, consolidamento ed esplorazione dei dati di business, per sintetizzare le informazioni (sotto forma di report per il management aziendale) e per estrarre conoscenza ad elevato valore aggiunto [14].

Il KDD è un processo che si pone l'obiettivo di analizzare grandi quantità di dati per ottenere modelli e informazioni utili, ovvero, cerca di estrarre conoscenza di alto livello da dati di basso livello. Alcuni esperti considerano il KDD sinonimo di Data Mining mentre altri considerano il data Mining come una fase del KDD. In questo lavoro, si utilizzerà la seconda filosofia di pensiero.

Il processo KDD, come mostrato in figura 5-1, può essere quindi descritto come una sequenza iterativa dei seguenti passaggi:

1. *Data preprocessing*: preparazione dei dati per il data mining. Può essere costiutito dalle seguenti fasi:
    a. Data integration: integrazione di più sorgenti di dati;
    b. Data cleaning: rimozione del rumore e dei dati inconsistenti;
    c. Data selection: selezione dei dati rilevanti ai fini delle analisi dal data warehouse;

d. Data transformation: in questa fase i dati vengono trasformati o consolidati in modo che il processo di estrazione della conoscenza risultante possa essere più efficiente e i modelli trovati possano essere più facili da comprendere. Alcune strategie di trasformazione dei dati sono la normalizzazione, l'aggregazione e la discretizzazione.

2. *Data mining*: processo essenziale in cui si applicano i metodi per estrarre conoscenza dai dati. Esiste una serie di metodi di data mining tra cui l'estrazione di itemset frequenti, le regole di associazione, l'analisi di clustering, l'analisi degli outlier, la classificazione;
3. *Pattern evaluation*: valutazione dei modelli per identificare quelli veramente interessanti che rappresentano la conoscenza sulla base di misure di interesse;
4. *Data visualization*: utilizzo di tecniche di visualizzazione e rappresentazione della conoscenza per mostrare i risultati e le nozioni estratte agli utenti [15].

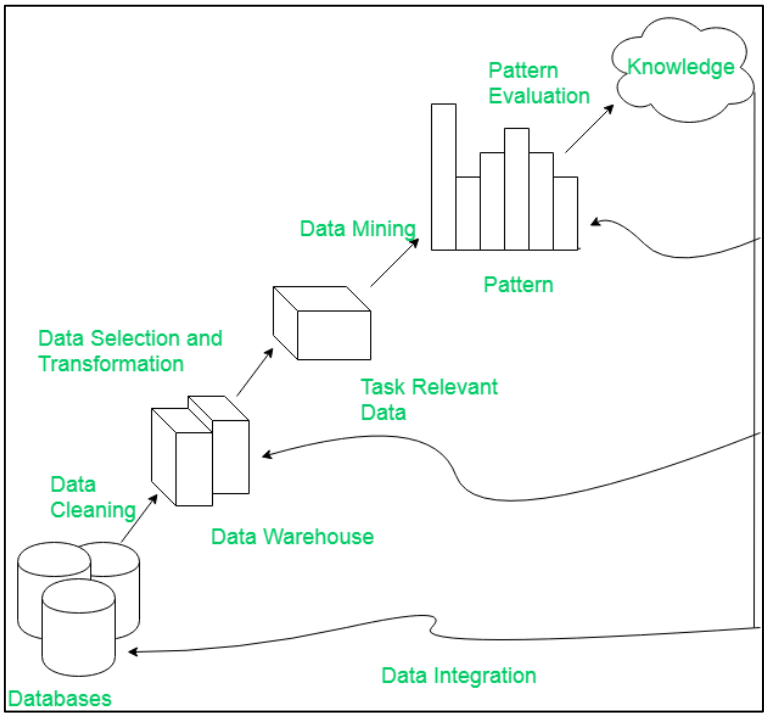


Figura 5-1. KDD Process in Data Mining. Fonte: GeeksforGeeks

L'applicazione del KDD a questo caso studio è un'idea innovativa e ricca di potenzialità. Sarà interessante confrontare le informazioni estrapolate con questo metodo con quelle ottenute dalle categorie generate dall'analisi ABC.

Quando si opera con un nuovo strumento è bene tenere a mente sia i pregi che i difetti. I limiti del modello possono essere riassunti in [16]:

- Problemi di privacy: l'applicazione del processo comporta la raccolta e gestione di dati di grandi dimensioni che possono comprendere informazioni sensibili su persone;
- Qualità dei dati: il processo dipende molto dalla qualità dei dati in input. È bene quindi trascorrere sufficiente tempo nello step di preparazione dei dati affinché si cerchi di ridurre al minimo le incoerenze e si massimizzi l'accuratezza.
- Costo elevato: il processo KDD può essere costoso in quanto richiede notevoli investimenti in strumenti come software o hardware e in tecnici esperti o risorse umane.

5.1 Data preprocessing

A causa delle grandi dimensioni e della provenienza dei dati da fonti eterogenee, i database del mondo reale spesso contengono dati rumorosi o non coerenti o mancanti. I dati di bassa qualità portano a estrarre informazioni di eguale qualità. È quindi molto importante svolgere una fase di preprocessing, pre elaborazione, dei dati per migliorarne la qualità e quindi i risultati ottenibili tramite gli algoritmi di data mining. È noto infatti che l'applicazione delle tecniche di preprocessing prima degli algoritimi di data mining migliora sostanzialmente la qualità complessiva dei modelli che si estraggono e il tempo per estrarli [11], [15].

Lo step di data preprocessing può essere ulteriormente suddiviso in quattro passaggi, come si può osservare in figura 5-2 [15]:

- *Data cleaning*. Il processo di pulizia dei dati tenta di gestire i missing values, identificare gli outlier e correggere le incoerenze nei dati. Il rumore è un errore casuale o una varianza in una variabile misurata e può essere identificato sfruttando il clustering. In presenza di dati mancanti è possibile adottare diverse strategie per gestirli:
    - Ignorare la tupla eliminandola;
    - Riempire a mano il valore mancante ma ciò richiede solitamente molto tempo ed è difficilmente applicabile in presenza di dataset di grandi dimensioni;

    - Inserire una costante globale per riempire tutti i valori mancanti;
    - Utilizzare il valore medio o la mediana;
    - Utilizzare il valore più probabile sfruttando la conoscenza ottenibile con tecniche statistiche come la regressione o l'albero decisionale.
- *Data integration*. Il processo di integrazione spesso è richiesto per unire i dati provenienti da più fonti aziendali. Una questione importante nell'integrazione dei dati è costituita dalla ridondanza delle informazioni. Un attributo può essere considerato ridondante se si deriva da un altro attributo o da un insieme di attributi. Alcune ridondanze possono emergere grazie all'analisi di correlazione.
- *Data reduction*. Le tecniche di riduzione dei dati sono applicate per ridurre il dataset in termini di volume, pur mantenendo l'integrità dei dati originali cosicchè gli algoritmi di data mining possano essere più efficienti. Per ridurre la dimensione, ad esempio, si possono eliminare gli attributi correlati. Questo step può essere definito come *data and feature selection* e può essere sviluppato grazie all'ausilio della *correlation analysis*. Per far ciò, in presenza di attributi numerici, è possibile utilizzare diverse funzioni tra cui il coefficiente di correlazione lineare di Pearson che presenta un valore compreso tra -1 e 1:
    - 1: Correlazione direttamente proporzionale perfetta;
    - -1: Correlazione inversamente proporzionale perfetta;
    - 0: Variabili indipendenti.
- *Data transformation*. Durante questa fase i dati vengono trasformati e consolidati nella forma appropriata per il data mining. Alcune strategie sono:
    - *Costruzione di attributi*: nuovi attributi vengono generati per aiutare il processo di mining;
    - *Aggregazione:* alcuni dati vengono aggregati lungo una dimensione, ad esempio, le vendite giornaliere possono essere aggregate per mostrare le vendite mensili o annuali;
    - *Normalizzazione*: i dati vengono modellati su una scala comune senza distorsioni negli intervalli di valori o perdita di informazioni. Cerca di dare a tutti gli attributi lo stesso peso. È particolarmente utile per gli algoritmi che coinvolgono misure di distanza come il clustering. Esistono diversi metodi di normalizzazione.

- *Discretizzazione*: il valore di un attributo numerico viene sostituito con un intervallo.

Figura 5-2. Step data preprocessing

5.2 Data mining

Dopo la fase di data preprocessing si passa allo step di data mining, il cui scopo è quello di trovare pattern interessanti all'interno del dataset tramite l'estrazione di informazioni non evidenti grazie all'utilizzo di algoritmi sofisticati [17].

Un pattern è un modello ricorrente o eccezionale presente nei dati, in genere, è una rappresentazione sintetica e ricca di semantica del dataset. Affinché un pattern sia rilevante deve essere validato sui dati con un certo valore di confidenza, deve essere facilmente comprensibile dall'utente, non deve essere conosciuto in precedenza e potenzialmente applicabile [18]. In generale, un'attività di data mining può essere classificata in due categorie:

- Descrittiva: trovare pattern interpretabili che descrivano i dati;

- Predittiva: utilizzare alcune variabili per prevedere il valore di altre variabili.

Per ottenere i pattern dai dati e in base alle esigenze si possono utilizzare diversi algoritmi di data mining [15], [17], [19] riassunti in tabella 5-1.

Tabella 5-1. Algoritmi di data mining

METODO	***TIPOLOGIA***	***DESCRIZIONE***
CARATTERIZZAZIONE E DISCRIMINAZIONE DEI DATI	descrittiva	Fase preliminare alla classificazione per esplorare i dati. Si confronta la distribuzione dei valori degli attributi per i record appartenenti ad una medesima classe (caratterizzazione) e si confronta l'andamento dei valori di un attributo nelle diverse classi di appartenenza (discriminazione).
RICERCA DI PATTERN FREQUENTI, REGOLE DI ASSOCIAZIONE	descrittiva	È possibile ottenere pattern frequenti e quindi identificare associazioni interessanti tra gruppi di record di un dataset.
CLASSIFICAZIONE	predittiva	L'algoritmo di classificazione utilizza le informazioni disponibili riferite al passato per identificare un modello matematico che consenta di prevedere a quale classe apparterranno le informazioni future.
REGRESSIONE	predittiva	La regressione è ampiamente usata nell'ambito delle reti neurali. La regressione permette di predire il valore della variabile dipendente sulla base di valori di altre variabili, dette regressori, assumendo un modello di dipendenza lineare/non lineare.
CLUSTERING	descrittiva	Gli algoritmi di clustering mirano a segmentare una popolazione eterogenea in un certo numero di sottogruppi che contengono caratteristiche simili tra loro.

La ricerca sul data mining presenta molte sfide, tra cui: la metodologia di estrazione, l'interazione con l'utente, l'efficienza, la scalabilità e la gestione di diversi tipi di dati. La

ricerca sul data mining ha avuto un grande impatto sulla società e continuerà a farlo [15].
Le aziende grazie alle tecniche di data mining possono cercare di estrarre informazioni utili per il proprio business come conoscere meglio i propri clienti, sviluppare strategie di marketing migliori o aumentare le vendite [17].

Nel contesto della pianificazione della domanda è stato coniato il termine *Demand Intelligence* che si riferisce alle analisi quantitative della domanda storica mediante algoritmi di Data Mining per poter assumere scelte più consapevoli. Più nel dettaglio la demand intelligence ricerca relazioni statisticamente significative e frequenti in un insieme eterogeneo di dati prodotto – mercato. Dai risultati ottenuti dalle analisi si possono generare conoscenza di business che rendono l'azienda più competitiva.
I modelli di Data Mining utili per il Demand Intelligence possono essere diversi [11]:

- *Clustering*: raggruppamento di segmenti prodotto-mercato i cui attributi hanno elevata similarità, in cui è possibile dedurre regole generali di comportamento comuni a tutti gli elementi appartenenti al cluster che potrebbero essere applicati anche ai nuovi prodotti che verranno aggiunti in futuro;
- *Regole di associazione*: studiare le possibili correlazioni tra prodotti diversi venduti nello stesso periodo per lo stesso cliente al fine di individuare regolarità fra gruppi di prodotti acquistati spesso in modo congiunto;
- *Classificazione*: determinare se alcune azioni di marketing possano avere successo su certi clienti o meno.

5.2.1 Clustering

Il clustering è un campo di ricerca moderno nell'ambito del data mining. È correlato all'apprendimento non supervisionato (metodo che raggruppa e interpreta i dati solo in base ai dati in input) nel contesto del machine learning. Il clustering è il processo di partizione di un insieme di dati in sottoinsiemi. Ogni sottoinsieme rappresenta un cluster in modo tale che gli oggetti contenuti in un cluster siano simili tra loro mentre gli oggetti che appartengono a cluster diversi siano dissimili (vedi figura 5-3). In altre parole, è molto importante che:

- gli elementi contenuti in uno stesso gruppo possiedano valori simili degli attributi utilizzati per generare i cluster;
- gli elementi che appartengono a cluster differenti possiedano caratteristiche significativamente diverse nei valori degli attributi usati per clusterizzare.

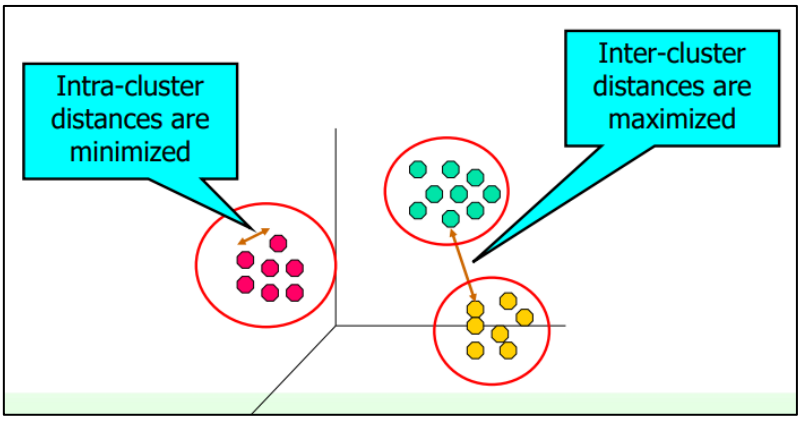


Figura 5-3. Introduction to Data Mining, McGraw Hill 2006

Questo metodo è utile in quanto può portare alla scoperta di gruppi di dati precedentemente sconosciuti e affinché le analisi siano interessanti deve fornire risultati non banali e interpretabili a posteriori [11], [15]. Il clustering può essere applicato per comprendere meglio i dati o per riassumere e ridurre la dimensione del dataset [20].

Prima di applicare tecniche di clustering è bene tenere a mente importanti informazioni riguardo alcuni requisiti minimi [15]:

- *Scalabilità*: alcuni algoritmi di clustering funzionano bene con dataset non grandi ma al giorno d'oggi i database potrebbero contenere milioni di oggetti;
- *Capacità di gestire diversi tipi di attributi*: alcuni algoritmi funzionano meglio se gestiscono alcuni tipi di attributo semplici ad esempio numerici. Recentemente, si stanno sviluppando nuovi algoritmi in grado di gestire tipi di dati più complessi;
- *Scoperta di cluster con forma arbitraria*: certi algoritmi si basano sulla misura di distanza Euclidea o di Manhattan e generalmente trovano cluster di forma sferica con dimensioni simili. Un cluster però potrebbe avere qualsiasi forma per questo è importante applicare anche algoritmi che sappiano identificare raggruppamenti di forma insolita;

- *Requisiti di conoscenza del dominio per determinare i parametri di input*: per definire i parametri di input al clustering è importante conoscere bene il contesto per questo talvolta chi svolge analisi sui dati deve lavorare a stretto contatto con gli esperti di dominio. Il clustering può essere molto sensibile alla scelta dei parametri.
- *Capacità di gestire dati rumorosi:* i dati reali spesso contengono dai rumorosi o mancanti o sbagliati e questo potrebbe generare cluster di bassa qualità.

In letteratura sono stati sviluppati molti algoritmi di clustering che possono essere suddivisi in base a diversi aspetti: criteri di partizione degli oggetti, misure di similarità, spazio di clusterizzazione, ecc. È importante sottolineare che diversi algoritmi presenti in letteratura anche se applicati a uno stesso set di dati possono fornire raggruppamenti differenti.

Prendendo spunto dal libro 'Data Mining' [15] è possibile fare una prima suddivisione degli algoritmi di clustering riassunti in tabella 5-2.

Tabella 5-2. Algoritmi di clustering

Metodo	Caratteristiche	Algoritmi
Clustering partizionale	Crea un insieme iniziale di k partizioni, con k parametro inserito in input. L'algoritmo cerca di migliorare iterativamente il partizionamento riallocando gli oggetti e spostandoli da un gruppo a un altro.	K-MEANS, KMEDOIDS, CLARANS
Clustering gerarchico	Il set di dati viene decomposto gerarchicamente. In base al tipo di decomposizione si può parlare di algoritmo agglomerativo (bottom-up) o divisivo (top-down).	AGGLOMERATIVO, DIVISIVO
Clustering density - based	Gli oggetti vengono raggruppati analizzando la densità di punti in un intorno di raggio prestabilito.	DBSCAN, OPTICS
Clustering grid-based	Gli oggetti vengono inseriti in un numero finito di celle che formano una struttura a griglia e successivamente si esegue il clustering sulla struttura a griglia.	STING, CLIQUE

5.2.2 K-Means Clustering

Il k-means è uno degli algoritmi di clustering più utilizzati e performanti. Fa parte dei metodi partizionali e si basa sul concetto di centroide (punto centrale, solitamente la media dei punti nel cluster). Questo algoritmo prevede che ogni punto del dataset venga assegnato al cluster con il centroide più vicino. La vicinanza a un cluster può essere misurata in base a diversi tipi di distanza (Euclidea, Minkowski, Manhattan, Cosine Similarity, …). Se il centroide è un valore fittizio che corrisponde al baricentro del cluster allora è giusto parlare di k-means, se il centroide corrisponde a un dato reale allora l'algoritmo è definito k-medoids [15], [17], [20].

I passaggi del k-means sono:

1. Scelta di *k, numero di cluster;*
2. Scelta dei centroidi iniziali. Se nulla viene specificato la prima scelta è random;
3. L'algoritmo assegna ogni dato al cluster con il centroide più vicino;
4. Per ogni cluster si calcola nuovamente il centroide come baricentro del gruppo;
5. Si torna al punto 3.
6. Il processo termina quando la posizione dei centroidi non varia.

È importante sottolineare che il metodo sia iterativo e che la scelta dei centroidi iniziali possa avere un impatto sulle performance dell'algoritmo. Un esempio di applicazione dell'algoritmo è mostrato in figura

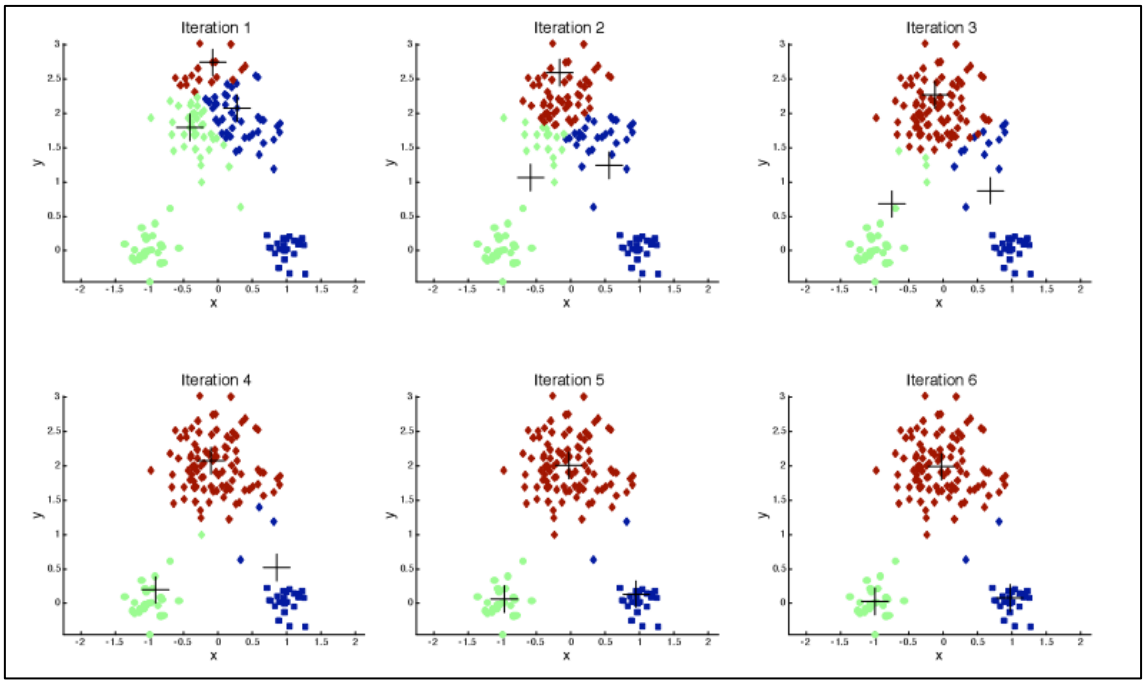


Figura 5-4. K-means. Fonte: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Per valutare la bontà del k-means la metrica più utilizzata è l'SSE, Sum of Squared Error, che aumenta al crescere della distanza tra i punti di un cluster e il relativo centroide.

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} dist^2(m_i, x)$$

Dove:

- x è un preciso valore nel cluster C_i
- k = numero di cluster.
- m_i = centroide del cluster C_i.
- $dist$ = distanza calcolata con un metodo tra Euclidean, Manhattan, Cosine Similarity, …

Il primo step dell'algoritmo consiste nel definire il valore k, numero dei cluster, parametro di input. A questo proposito esiste un metodo detto *Knee Approach* che permette tramite l'elbow graph di rappresentare graficamente il valore dell'SSE per una serie di valori k. Si sceglie il valore di k ottimale in corrispondenza del gomito che si nota graficamente. Scegliendo così il valore di k il guadagno che si otterrebbe aggiungendo un centroide sarebbe trascurabile e, inoltre, la riduzione della misura di

qualità non sarebbe più interessante [20]. Un esempio dell'elbow graph è disponibile in figura 5-5.

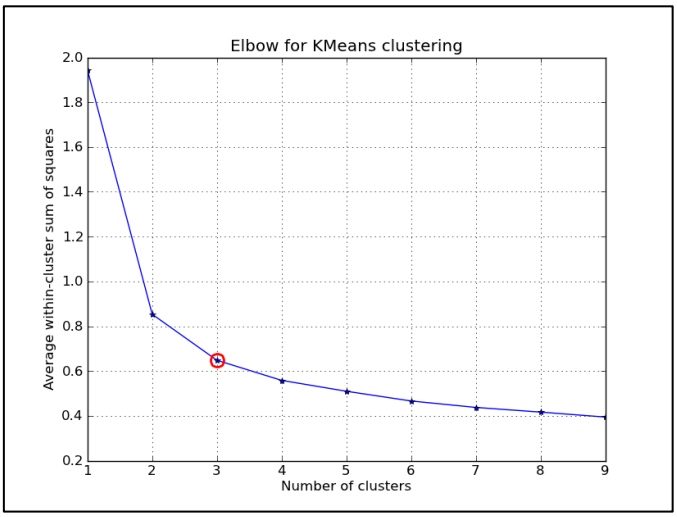


Figura 5-5. Elbow Method. Fonte: Machine Learning Interviews

Generalmente, se k aumenta allora l'SSE diminuisce. Con l'elbow method si cerca quel valore di k per cui la diminuzione dell'SSE può essere considerata trascurabile [15], [20].

Affinché il k-means performi bene è necessario svolgere una fase di preprocessing accurata in cui si normalizzano i dati e si eliminano gli outliers. Inoltre, per ottimizzare i risultati si possono svolgere una serie di fasi di post processing come:

- Eliminare cluster di piccole dimensioni che potrebbero rappresentare valori anomali;
- Unire cluster vicini che presentano un valore di SSE basso;

Un passaggio cruciale nell'applicazione del k-means è la *scelta dei centroidi inziali* dato che l'algoritmo mira a convergere su un insieme ottimale di centroidi e l'appartenenza di un punto a un cluster si basa sulla distanza di questo dal centroide tramite iterazioni successive. Quindi, migliore è la scelta dei centroidi iniziali minore sarà il numero di iterazioni necessarie per giungere alla configurazione ottimale. Esistono diversi metodi per inizializzare i centroidi tra cui ad esempio lasciare al caso la definizione dei primi centroidi ma come abbiamo visto questo potrebbe portare a inefficienze. Una soluzione migliore potrebbe consistere nell'applicazione del clustering gerarchico per trovare i centroidi iniziali che possono essere successivamente passati al k-means per raggruppare i dati [20], [21].

Il K-means presenta alcuni limiti quando i cluster sono di dimensioni diverse o hanno differenti densità. Inoltre, bisogna prestare particolare attenzione al fatto che i dati possano contenere valori anomali [20].

5.2.3 Agglomerative Clustering

Gli algoritmi gerarchici rappresentano un altro possibile metodo di partizionamento in cluster dei dati. Possono essere suddivisi in due tipologie [15], [20], [22]:

- *Agglomerativo*: inizialmente ad ogni dato è associate un cluster diverso. Quindi ci sono tanti cluster quanti sono i dati. In seguito, si procede aggiungendo di volta in volta un legame tra due cluster per unirli utilizzando metriche opportune. Il processo si itera fino a che si otterrà un unico cluster contenente tutti i dati.
- *Dissociativo*: il processo è speculare rispetto al clustering agglomerativo. I dati inizialmente appartengono tutti allo stesso cluster e, ad ogni iterazione, vengono suddivisi creando un gruppo in più, creando cluster più piccoli.

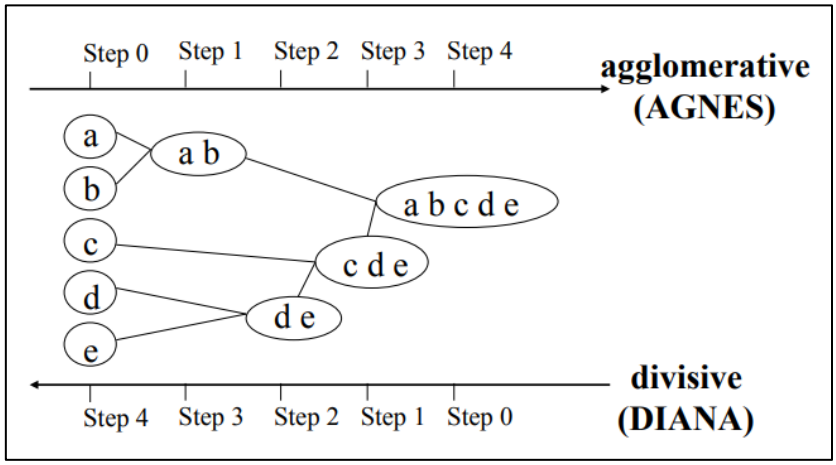


Figura 5-6. Hierarchical clustering. Fonte: Exploreing K-Means with Internal Validity Indexes for Data Clustering in Traffic Management System

Per rappresentare le iterazioni del clustering gerarchico si presta bene un grafo ad albero detto *dendrogramma.* Questo strumento mostra come gli oggetti vengano raggruppati o divisi step by step dall'algoritmo. Un esempio di dendrogramma può essere visualizzato in figura 5-7.

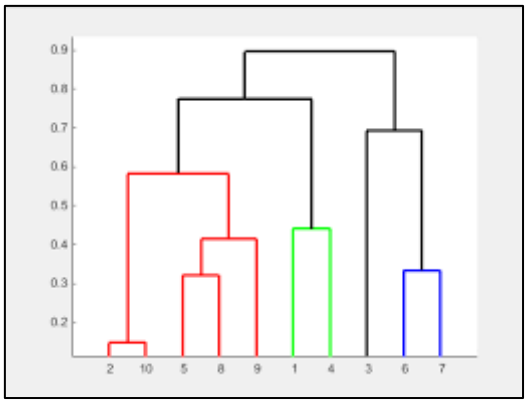


Figura 5-7. Dendrogramma,
Fonte: Clustering gerarchico. MaggioliDEVELOPERS

Per valutare la distanza tra i cluster è possibile utilizzare diverse metriche in base alla necessità. Le più frequenti sono [22], [23]:

- Min o Single Link = distanza minima tra due punti appartenenti a cluster diversi. È in grado di gestire forme non ellittiche ma è sensibile al rumore e agli outlier;
- Max o Complete Link = distanza massima tra due punti appartenenti a cluster diversi. È meno sensibile al rumore e agli outlier rispetto al Single Link, tende a trovare cluster di forma globulare;
- Group Average = distanza media tra i punti appartenenti a due cluster diversi;
- Distanza tra i centroidi = si considera solo la distanza tra i centroidi;
- Ward's Method = la distanza è la somma dei quadrati delle distanze tra i punti appartenenti a due cluster diversi. Può essere utilizzato per inizializzare il k-means.

Una peculiarità degli algoritmi gerarchici è che una volta presa la decisione di fondere due cluster, approccio agglomerativo, o separare due cluster, approccio divisivo, questa non è reversibile e quindi non può essere annullata. Inoltre, alcuni tra i problemi più frequenti di questo approccio sono la difficoltà di gestire cluster di dimensioni diverse e la sensibilità al rumore e agli outlier. Tra i punti di forza dell'algoritmo gerarchico si può citare il fatto che non sia necessario definire a priori un numero di cluster come è necessario per il k-means. Inoltre, il metodo permette di generare tutte le combinazioni possibili, quindi, è sufficiente variare la soglia di clusterizzazione per modificare il numero di raggruppamenti senza dover eseguire nuovamente il processo. Rispetto al k-

means il costo computazionale è maggiore dovendo generare tutte le combinazioni possibili [20], [22].

5.2.4 DBSCAN

Il DBSCAN, *Density-Based Spatial Clustering of Applications with Noise*, è uno tra i possibili metodi di clustering ed è stato sviluppato da *Martin Ester, Hans-Peter Kriegel, Jörg Sander e Xiaowei Xu* [24]. È un algoritmo density-based che, perciò, si basa sulla nozione di densità e definisce un cluster come l'insieme di punti contenuti entro una certa densità [25].

Affinché l'algoritmo funzioni correttamente è necessario determinare Eps e MinPts. Il primo definisce la distanza massima tra due punti mentre il secondo parametro corrisponde al minimo numero di punti richiesti per formare un cluster.

I dati possono essere suddivisi in tre categorie [20], [23], [24]:

- Core point: è un valore che ha più di *MinPts* dati adiacenti con distanza minore di *Eps*;
- Border point: è un punto di confine che ha meno di *MinPts* all'interno del raggio *Eps* ma si trova nelle vicinanze di un core points;
- Noise point: sono tutti i dati che non sono core point né border point.

È possibile visualizzare le tre tipologie di punti graficamente in figura 5-8 e in figura 5-9.

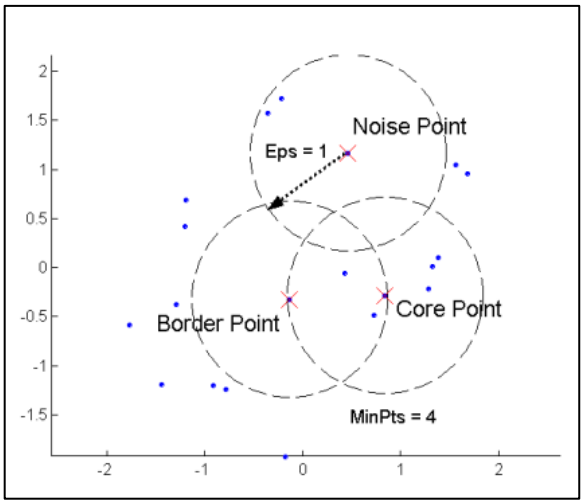


Figura 5-8. Core, Border e Noise Points. Fonte: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

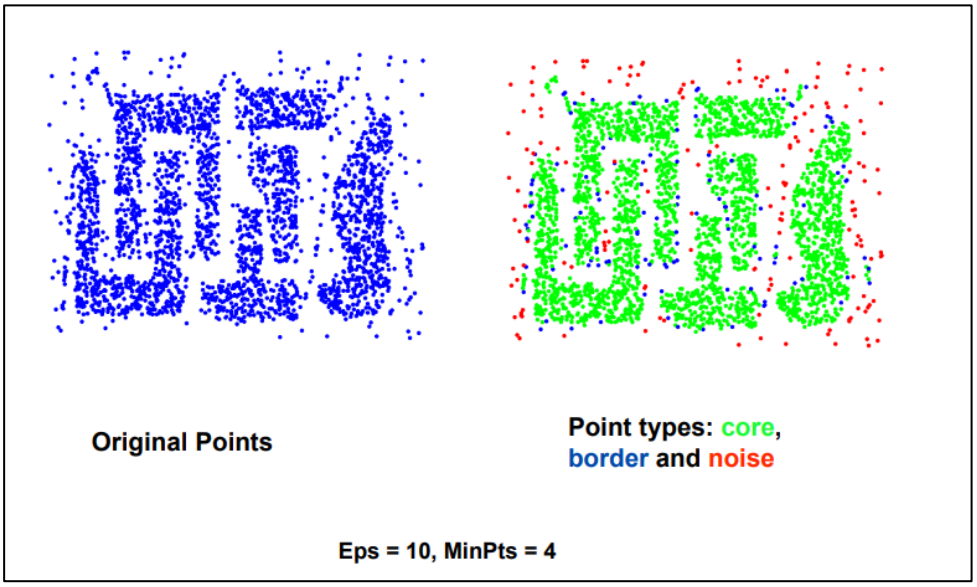


Figura 5-9. Core, Border e Noise Points. Fonte: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Uno dei metodi per determinare Eps e MinPts, i due parametri di input dell'algoritmo, consiste nell'idea che per i punti di un cluster il k-esimo vicino più prossimo si trova più o meno alla stessa distanza, a meno che, la densità abbia elevata variabilità. I punti rumorosi invece hanno il k-esimo vicino più distante. Il valore k è definito come il numero di dimensioni del dataset più uno. Per determinare Eps si si utilizza il grafico *k-nearest neighbor* che presenta sull'asse delle ascisse i punti ordinati per distanza crescente e sull'asse delle ordinate, Eps, le distanze ordinate di ogni punto dal suo k-esimo. Un esempio di kNN graph è disponibile in figura 5-10.

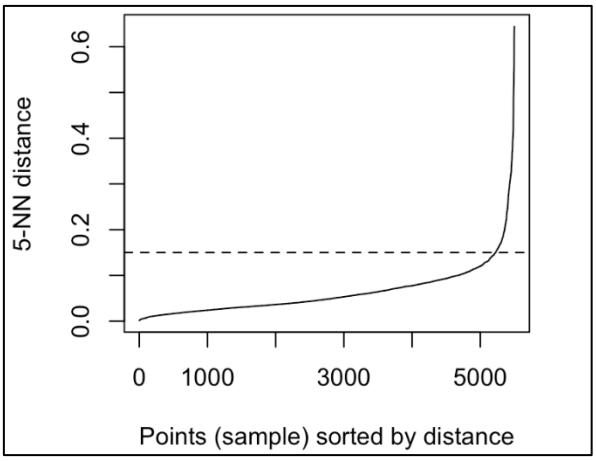


Figura 5-10. K-neares neighbor. Fonte: DataNovia - ADVANCED CLUSTERING

Il DBSCAN non performa bene in presenza di dati con densità variabile e in presenza di dataset di grandi dimensioni.

5.3 Pattern evaluation

È molto importante valutare i pattern ottenuti dai modelli di data mining in quanto solo alcuni sono rilevanti e di interesse. Un pattern o modello può essere significativo se coerente con l'obiettivo che ci si è posti, oppure talvolta, la rilevanza di un pattern può dipendere dal giudizio dell'utente che lo analizza [15]. Giunti a questo punto, quindi, occorre interpretare e valutare i modelli ottenuti e se necessario si può considerare di ricominciare dal primo step modificando il preprocessing dei dati. Il processo di KDD, infatti, è iterativo in quanto i passaggi possono essere ripetuti ed interattivo perché richiede un frequente confronto con l'analista di dominio in modo da fare scelte mirate nella configurazione del procedimento [26]. È importante tenere a mente che la qualità dei risultati ottenuti dipende fortemente dalla qualità dei dati e, anzi, è importante spendere circa il 60% degli sforzi nel processo di KDD per la preparazione dei dati [15], [26].

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage" è una citazione di Jain e Dubes nel libro Algorithms for Clustering Data che sottolinea l'importanza della validazione dei cluster [27]. La valutazione dei pattern, il processo di valutazione della qualità dei pattern scoperti, è importante per determinare

se i pattern sono utili e se ci si può fidare di loro. Esistono metriche diverse per valutare le analisi che possono essere categorizzate in tre macrogruppi [20]:

- *Indici interni* = misurare la bontà del clustering senza aver bisogno di informazioni esterne. Servono per valutare il grado di coesione e separazione dei gruppi. Es. *SSE*, *cluster cohesion*, *cluster separation*. L'obiettivo sarebbe quello di massimizzare la coesione e la separazione.

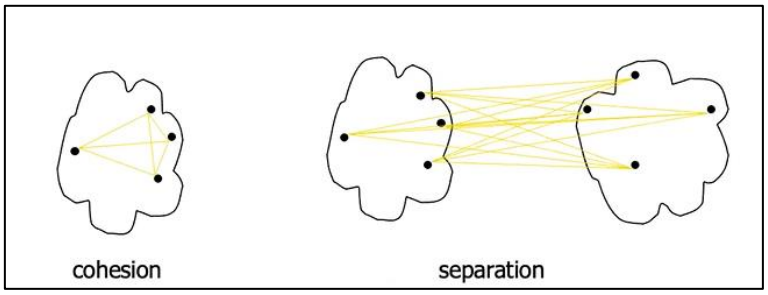


Figura 5-11. Indici interni. Fonte: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

La *Silhouette* è una metrica che comprende in un unico valore sia la coesione che la separazione.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \in [-1, 1]$$

- a(i): la distanza media del dato *i* rispetto agli altri oggetti all'interno dello stesso cluster (minore è il valore, meglio è);
- b(i): è il minimo tra la distanza media di *i* rispetto a qualsiasi altro cluster di cui i non è membro.

Più il valore di silhouette è prossimo ad uno, migliore è la performance dell'algoritmo utilizzato. Per ogni oggetto della base dati si può calcolare la silhouette. Successivamente la media dei valori di silhouette per tutti i dati di un cluster rappresenta quanto coesi siano i valori nel raggruppamento. In seguito, la media dei valori di silhouette dei cluster può rappresentare la bontà del clustering applicato al dataset. In particolare, cluster con un valore di Silhouette prossimo a 1 sono molto ben rappresentati dall'algoritmo, cluster con valori bassi di Silhouette (circa 0) descrivono una situazione in cui le osservazioni si posizionano tra due cluster. Infine, un valore di Silhouette negativo può descrivere un caso in cui è possibile che i dati siano posizionati nel cluster sbagliato [28].

- *Indici esterni* = necessitano di una informazione esterna, a priori come l'etichetta di cluster, per valutare la bontà del clustering. Es. entropy, purity.
- *Indici relativi* = confrontano partizionamenti diversi ottenuti sugli stessi dati. Misura quanto sono simili i due partizionamenti. Ad esempio il Rand index.

In assenza dell'etichetta di classe si possono utilizzare solo gli indici interni e relativi e quindi: SSE, coesione, separazione, silhouette o Rand Index.

L'ispezione visiva è un altro metodo di validazione dei pattern nel data mining in cui i dati vengono ispezionati visivamente alla ricerca di pattern. Ciò può essere fatto osservando un grafico o un diagramma dei dati, oppure osservando i dati grezzi stessi. Questo metodo viene spesso utilizzato per individuare i valori anomali o i modelli insoliti ed è forse il metodo più comune che il data miner possa utilizzare [29].

5.4 Data visualization

L'ultimo step del KDD consiste nella rappresentazione della conoscenza o data visualization, cioè una tecnica che utilizza strumenti di visualizzazione per mostrare i risultati del data mining. Alcuni strumenti molto utilizzati oggi sono: report, tabelle, regole di caratterizzazione, dashboard. La domanda che ci si potrebbe porre in questa fase è come poter trasmettere i dati in modo efficace agli utenti finali. È anche possibile sfruttare le tecniche di visualizzazione per scoprire relazioni tra i dati che non si riuscirebbero a notare osservando i dati grezzi [15]. In ambito big data la visualizzazione dei dati è indispensabile per analizzare enormi quantità di informazioni e prendere decisioni basate sui dati. È vantaggioso rappresentare i dati in modo efficace sfruttando colori e ricorrenze in quanto la cultura umana è di tipo visivo.

In primis è importante determinare il tipo di report che si vuole rappresentare. In generale è utile associare il tipo di report al gruppo di destinatari poiché le esigenze sono diverse. Ogni tipo di report, infatti, deve definire in modo chiaro ed esauriente i requisiti per l'interfaccia utente. Il report può essere [30]:

- *Dashboard*: è destinata al gruppo esecutivo dell'azienda, quindi, deve sintetizzare in modo esauriente le metriche di alto livello che valutano l'andamento aziendale nel complesso. È importante che gli oggetti del report siano auto esplicativi e contestualizzati in modo da non creare confusione nel lettore. L'obiettivo è di identificare la situazione il più rapidamente possibile.

- *Analitico*: è il report più comune con l'obiettivo di supportare gli utenti a individuare la motivazione di una serie di domande tramite l'interpretazione del report. Un esempio è il report di analisi delle vendite.
- *Operativo*: i report operativi sono progettati per essere visualizzati in tempo reale in modo da monitorare costantemente la situazione e prendere decisioni il più tempestivamente possibile.

6. Analisi di mercato

6.1 Obiettivo

Come è stato anticipato in precedenza, l'analisi di mercato è uno strumento essenziale per il merchandising management al fine di conoscere meglio il settore di riferimento ed il comportamento della clientela target. Invero, permette di prendere consapevolezza su opportunità e rischi presenti nel mercato e di elaborare strategie di marketing vincenti [31]. L'obiettivo di questo capitolo è rispondere a interrogativi come quale sia la situazione dell'azienda di moda, quale sia la grandezza del mercato, se sia in crescita o decrescita e chi siano i potenziali clienti e quanto siano propensi a spendere per acquistare i prodotti dell'azienda.

6.2 L'azienda

L'impresa di riferimento opera nel mercato dei beni di lusso e il suo core business è costituito dalla produzione e vendita di articoli in pelle. L'azienda intende sviluppare la propria leadership nel mercato globale della pelletteria attraverso la diffusione di un prodotto di qualità ma con un valido rapporto qualità-prezzo. Il Gruppo è presente in cento paesi con oltre quattrocento negozi nelle città più prestigiose e di recente ha iniziato a sfruttare le opportunità fornite dall'e-commerce.

L'analisi della dimensione economico-finanziaria partendo dal bilancio consolidato dell'esercizio degli ultimi 5 anni, consultato grazie alle banche dati Bureau van Dijk e più in particolare AIDA (Analisi informatizzata delle aziende), è molto importante per interpretare la situazione del gruppo. Per motivi di segretezza non si citeranno i valori puntali degli indici, delle voci di stato patrimoniale e di conto economico ma osservando i grafici e analizzandone l'andamento si cercherà di descrivere la situazione aziendale nel suo complesso. Inoltre, i dati sono stati manipolati e modificati per un fattore moltiplicativo in modo da non potersi ricondurre al cliente dell'azienda di consulenza.

6.2.1 **Analisi di redditività**

Gli indici di redditività misurano la capacità dell'azienda di ottenere valore dal capitale che è stato investito.

Osservando la figura 6-1 si può notare che nel 2019 ad una leggera flessione del fatturato (-2 %) corrisponda un incremento del 55% dell'EBITDA suggerendo, quindi, che il gruppo stia generando maggiore ricchezza rispetto all'anno precedente nonostante si sia verificato un freno alla crescita. Nel 2020, invece, a causa del duro impatto della pandemia COVID-19, le vendite sono drasticamente diminuite (-42%) generando ingenti perdite rispetto al 2019 (utile netto -290%).

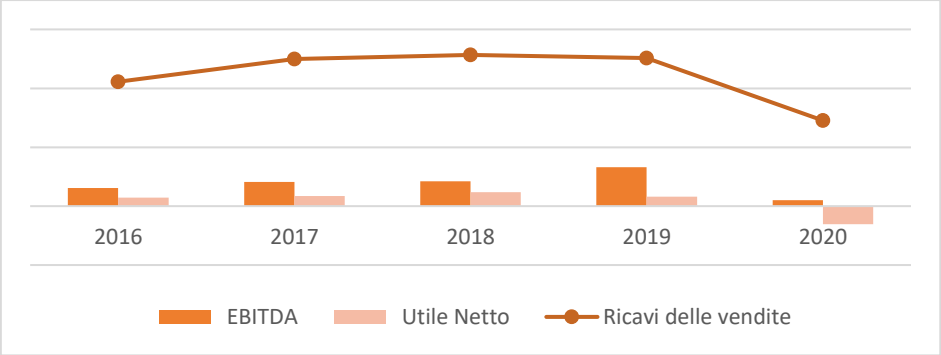


Figura 6-1. Indici di redditività

AIDA non contiene, al momento della consultazione, informazioni sul bilancio consolidato del gruppo per l'anno 2021 ma osservando il bilancio non consolidato emerge una piccola crescita dei ricavi pari al 5% in più rispetto al 2020 (figura 6-2).

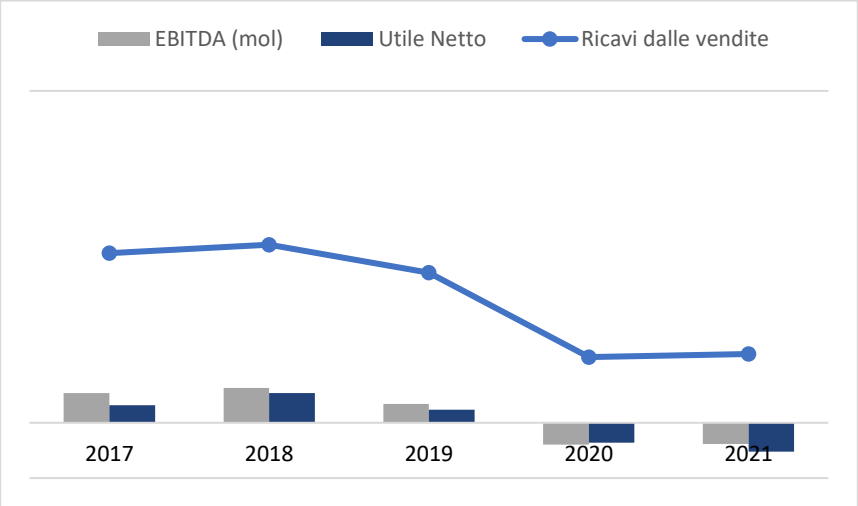


Figura 6-2. Ebitda, utile netto e ricavi dalle vendite.

Il ROS [32], Return on Sales, è un indice di bilancio che esprime la capacità dell'impresa di praticare prezzi di vendita remunerativi rispetto ai costi sostenuti .

$$ROS = \frac{RO}{PL}$$

- RO = reddito operativo;
- PL = produzione lorda (somma di fatturato e variazioni di magazzino).

In generale, si tratta di un indice che assume valori elevati se le aziende sono in grado di offrire i propri prodotti a prezzi alti e con costi operativi più contenuti possibile, altrimenti sarebbe necessario migliorare l'efficienza aziendale. In genere le aziende appartenenti al settore della moda e dei servizi riportano un ROS elevato. Dalla figura 6-3 emerge che l'azienda ha goduto di una buona efficienza operativa fino al 2019, invece, dal 2020 i ricavi sono insufficienti per coprire i costi di gestione e gli oneri e dunque non permettono di remunerare gli azionisti.

Il ROE, Redditività dell'Equity, consente di valutare la remunerazione per i proprietari della società.

$$ROE = \frac{RE}{KN}$$

- RE = risultato d'esercizio (dal conto economico)
- KN = patrimonio netto

È un indicatore di confronto rispetto a investimenti alternativi. Come si può osservare dal grafico, il ROE ha toccato il suo valore massimo nel 2018. Invece, nel 2020 il ROE ha assunto un valore negativo e questo implica che lo squilibrio è talmente grave da erodere il capitale proprio.

Il ROI, Redditività del Capitale Investito, misura quanto rende l'attività operativa dell'azienda rispetto ai mezzi finanziari impiegati.

$$ROI = \frac{RON}{KON}$$

- RON = reddito operativo al netto di ammortamenti e accantonamenti;
- KON = capitale operativo netto.

Perché sia soddisfacente deve essere superiore al costo medio del denaro in prestito e al tasso di remunerazione atteso dall'azionista. Come mostrato in figura, il valore del ROI è gradualmente diminuito dal 2017 al 2019, nonostante riportasse un valore soddisfacente. La situazione si è capovolta con l'avvento del COVID-19, momento nel quale i costi operativi sono rimasti circa costanti mentre il fatturato è diminuito fortemente. In figura il ROS è rappresentato in blu, il ROI in grigio e il ROE in azzurro.

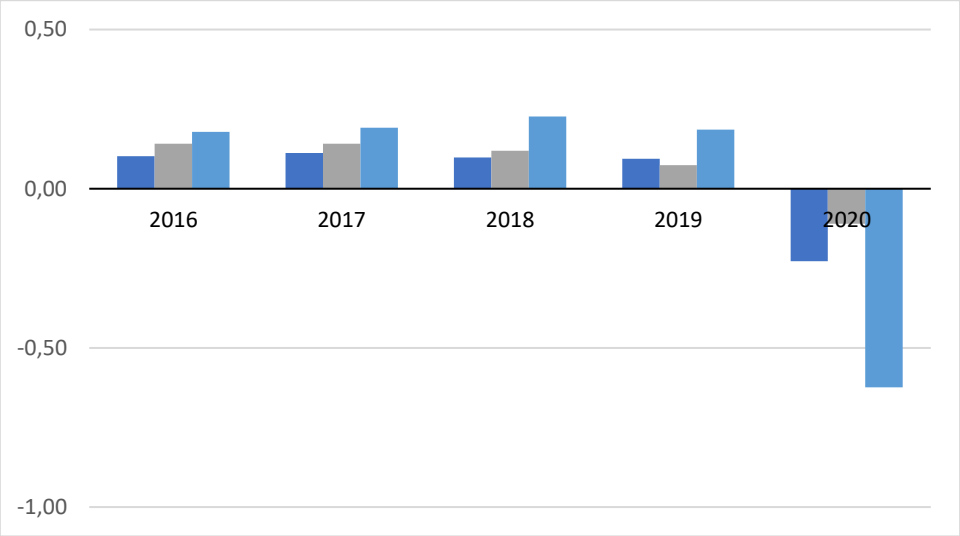


Figura 6-3. Indici di redditività

Analisi di liquidità

Gli indici di liquidità forniscono informazioni sulla situazione finanziaria dell'azienda e descrivono la sua solvibilità, cioè la capacità di mantenere gli impegni presi con i creditori.

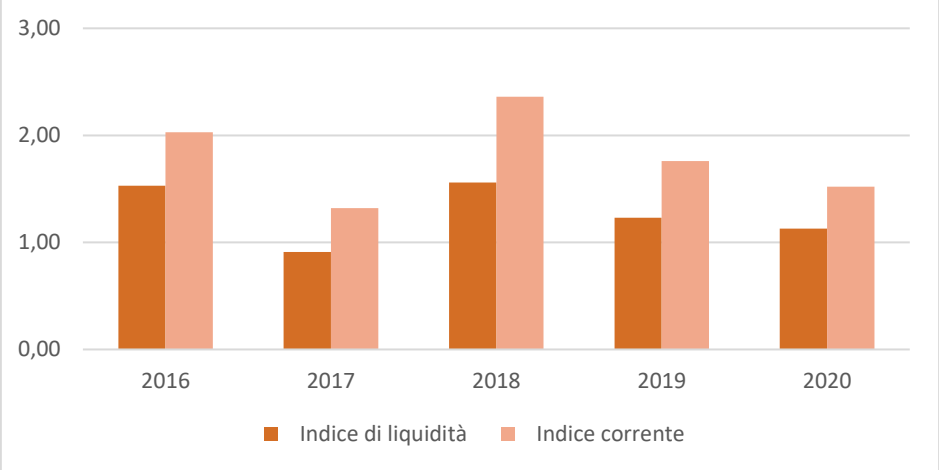


Figura 6-4. indici di liquidità

L’indice corrente si calcola dividendo le attività correnti per le passività correnti:

$$Indice\ corrente = \frac{attività\ a\ breve}{passività\ a\ breve}$$

Questo indicatore rappresenta la possibilità che le attività correnti siano convertite in denaro nei successivi 12 mesi ripagando così le passività a breve termine. L’azienda è considerata in una situazione di tranquillità finanziaria se l’indice assume un valore superiore a 1,5.

L’indice di liquidità o quick ratio, invece, evidenzia la capacità dell'azienda di far fronte agli impegni che scadono nel breve periodo utilizzando solo le risorse disponibili in forma liquida nello stesso periodo di tempo. Caso di ottimo equilibrio finanziario se l’indicatore assume un valore maggiore di 1.

$$Indice\ di\ liquidità = \frac{attività\ a\ breve - rimanenze}{passività\ a\ breve}$$

Analizzando entrambi gli indicatori per l’esercizio 2020, in figura 6-4, notiamo che, nonostante la pandemia, l'azienda si trovi in una situazione di tranquillità finanziaria in quanto le liquidità immediate (cassa e banca) e differite (crediti verso clienti) riescono a coprire le passività correnti.

Analisi di solidità

Un’impresa è solida se è riesce a mantenersi in equilibrio finanziario nel medio-lungo termine, superando eventuali incertezze di breve periodo. I punti centrali da analizzare sono, quindi le modalità di finanziamento delle immobilizzazioni e il grado di autonomia finanziaria. In generale, più la struttura finanziaria è costituita da mezzi propri, rispetto ai mezzi di terzi, meno l’azienda dipende da obblighi contrattuali ed inoltre, in caso di una contrazione significativa delle vendite, un’azienda molto indebitata farebbe più difficoltà a rimborsare i propri debiti.

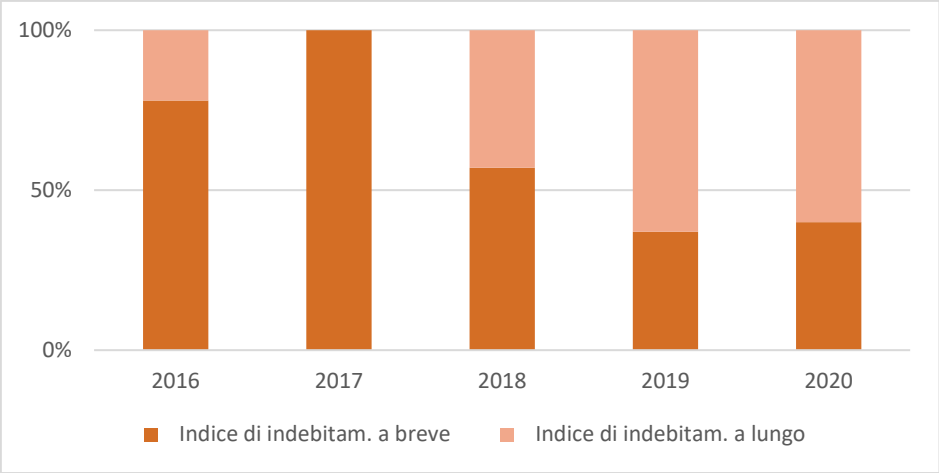


Figura 6-5. Indici di indebitamento

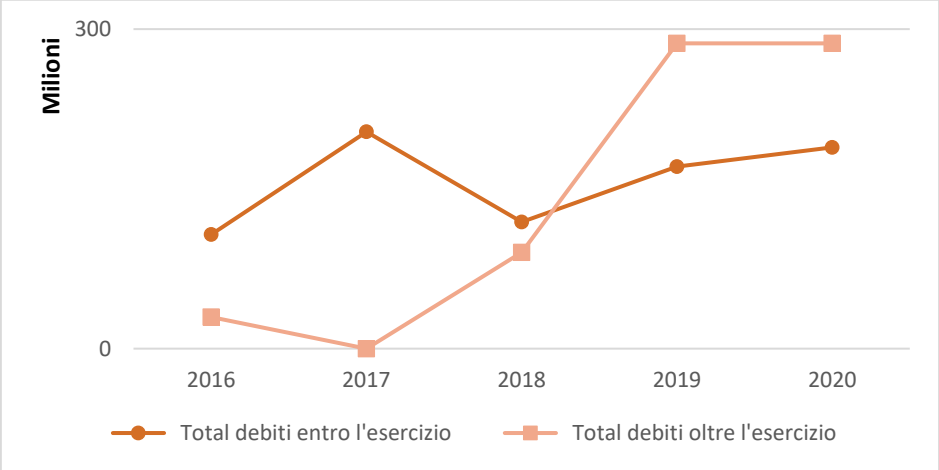


Figura 6-6. Debiti

L'indice di indebitamento è un valore che definisce il rapporto tra il totale delle passività a breve e a lungo termine.

$$Indice\ di\ indebitamento\ a\ breve = \frac{Totale\ debiti\ entro\ l'esercizio}{Totale\ debiti}$$

$$Indice\ di\ indebitamento\ a\ lungo = \frac{Totale\ debiti\ oltre\ l'esercizio}{Totale\ debiti}$$

Dai grafici in figura 6-6 si nota nel 2019 l'azienda abbia contratto un debito nel lungo periodo, modificando la struttura finanziaria.

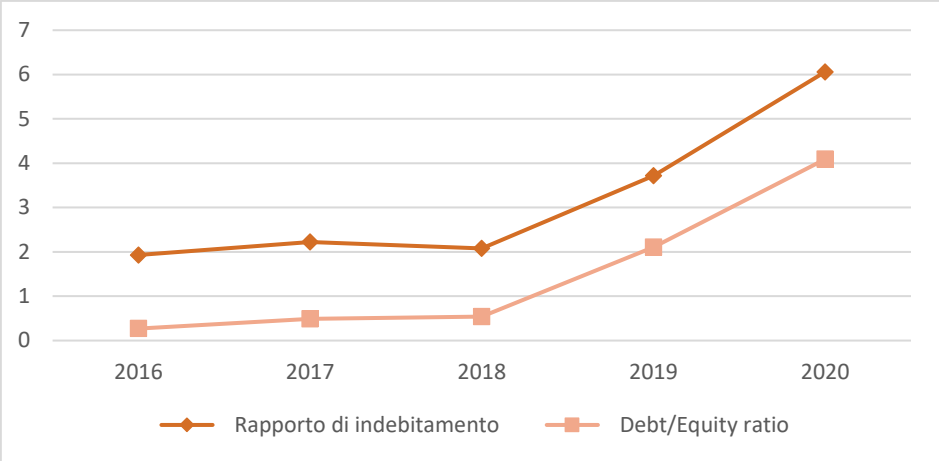


Figura 6-7. Rapporto di indebitamento

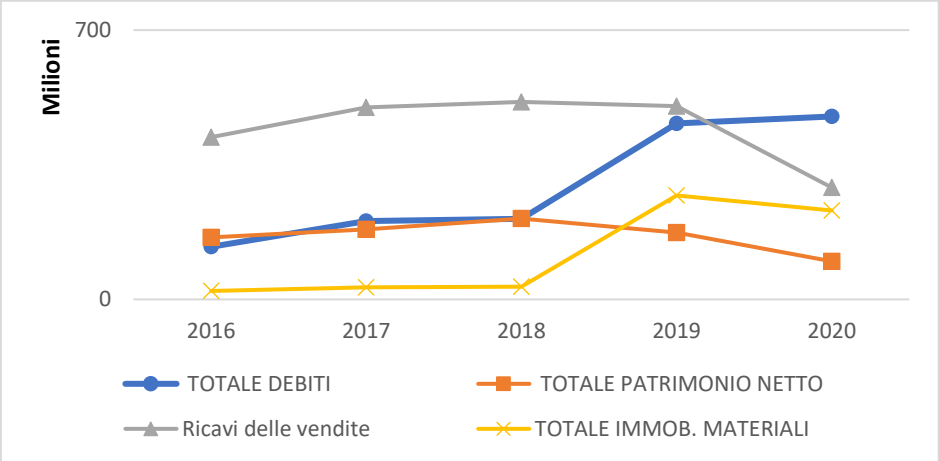


Figura 6-8. Passività, ricavi dalle vendite e immobilizzazioni immateriali.

Il rapporto di indebitamento o leva finanziaria è il rapporto tra il totale delle passività e il patrimonio netto [33].

$$Rapporto\ di\ indebitamento = \frac{totale\ passivo}{patrimonio\ netto}$$

Il debt/equity ratio descrive invece il rapporto tra quantità di debiti contratti dall'azienda e i mezzi propri. Più il rapporto risulta elevato, minore sarà la probabilità che le banche forniscano ulteriori prestiti, in quanto l'indebitamento è già alto.

Osservando le figure 6-7 e 6-8 si può notare che dal 2016 al 2020 la struttura finanziaria dell'azienda si sia capovolta. La quota crescente di debito implica un costo per l'impresa sotto forma di interessi che erodono il reddito d'esercizio e potrebbero tradursi in una remunerazione minore dei soci. Generalmente, i debiti vengono contratti per sostenere piani di investimento, e quindi per stimolare la crescita dell'impresa e la possibilità di produrre utili futuri. Come si evince dalla figura 6-8 il 2019 l'azienda ha

puntato sulla crescita aumentando le proprie immobilizzazioni materiali, il che giustifica la ristrutturazione finanziaria. Purtroppo, però nel 2020 con lo scoppio della pandemia le vendite si sono contratte rendendo più difficile per l'azienda rimborsare il debito in breve tempo e di conseguenza rallentandone la crescita.

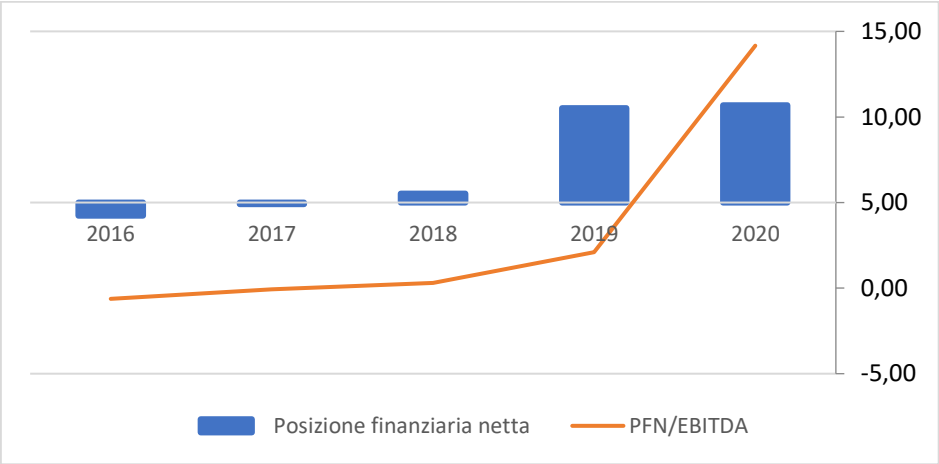


Figura 6-9. Posizione finanziaria netta.

La posizione finanziaria netta [34] si calcola come:

$$pfn = totale\ debiti - attività\ finanziarie\ a\ breve$$

- se negativo comporta una eccedenza della liquidità e delle disponibilità finanziarie rispetto all'indebitamento;
- se positivo l'impresa non riesce a coprire l'indebitamento finanziario con le proprie risorse.

La capacità di rimborsare il debito può essere calcolata come il seguente rapporto [35]:

$$\frac{posizione\ finanziaria\ netta}{Ebitda}.$$

Notiamo che il valore della posizione finanziaria netta sia cresciuto esponenzialmente il 2019 a causa dell'aumento dei debiti finanziari. Inoltre, dal 2019 si riduce la capacità di rimborsare i debiti finanziari (il valore dell'indice aumenta).

Indici di produttività

L'interpretazione degli indicatori di produttività deve essere confrontata con il settore dove opera l'azienda, sia nel tempo. Un aumento dell'indice dei ricavi pro capite o del valore aggiunto pro-capite descrive una più grande efficienza dell'azienda e perciò una condizione di buona salute dell'azienda stessa. Gli indici di produttività sono fortemente

influenzati dal numero dei dipendenti, in quanto un calo di produzione in momenti di crisi come quelli che si stanno attraversando, non sempre sono accompagnati da una diminuzione del numero delle risorse umane.

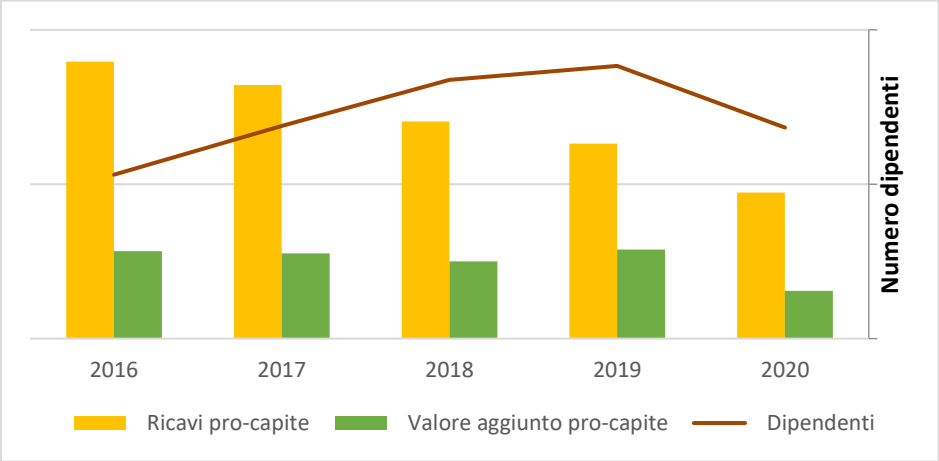


Figura 6-10. La produttività

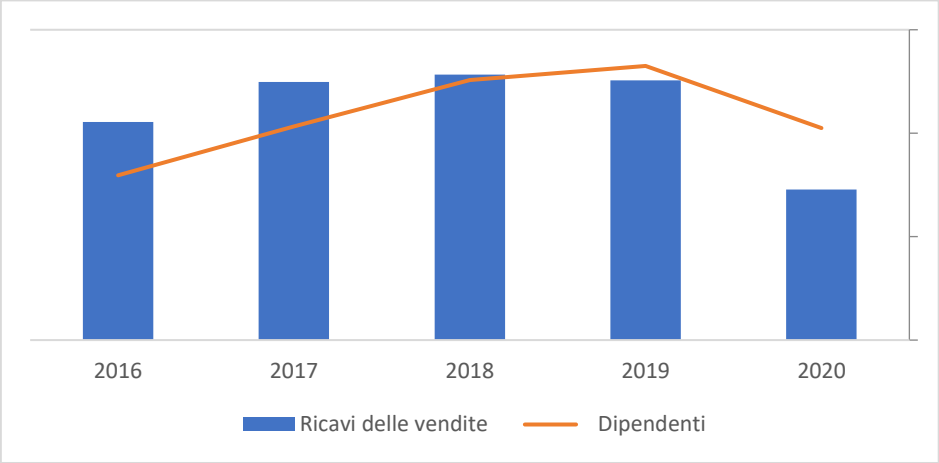


Figura 6-11. Ricavi dalle vendite e numero di dipendenti.

Come emerge dalla figura 6-10 i ricavi pro-capite dei dipendenti diminuiscono costantemente dal 2016 al 2020. Il valore aggiunto pro-capite è circa costante fino al 2019 ma nel 2020 si contrae del 45% circa rispetto all'anno prima. La forza lavoro ha raggiunto la sua grandezza massima nel 2019 composta da quasi 3000 dipendenti. Nel 2020 con l'arrivo della pandemia COVID19 l'azienda è stata costretta a ridimensionare il suo organico riducendolo di circa il 25% rispetto all'anno precedente.

La figura 6-11 mostra invece la stretta correlazione tra ricavi e numero di dipendenti. Si può notare un periodo di forte crescita dell'azienda in cui assume nuovi dipendenti per aumentare il suo fatturato. A causa della pandemia però nel 2020, la produzione si è

fermata e l'impresa è stata costretta a licenziare parte delle risorse umane per tamponare le perdite.

In conclusione, come emerge dagli indici di liquidità, sia pre che post pandemia, l'azienda presenta un elevato grado di solvibilità nel breve periodo non evidenziando difficoltà a rimborsare i debiti con scadenza nell'anno mediante le attività destinate a essere realizzate nel breve termine. Nel corso degli anni emergono due eventi rilevanti che hanno cambiato la storia dell'azienda: investimento con contrazione di un debito nel 2019 e pandemia COVID19 nel 2020. Dal 2019 in poi, la posizione finanziaria netta dell'impresa è peggiorata così come la capacità dell'azienda di rimborsare il debito. Prima della pandemia l'impresa godeva di una buona capacità di generare valore dal capitale investito e il numero di dipendenti cresceva in modo costante. Nel 2020 a causa delle chiusure forzate l'impresa ha generato perdite in quanto i ricavi si sono contratti del 40% circa e i costi operativi hanno superato il valore della produzione.

6.3 Il mercato di riferimento

L'azienda fa parte del mercato della pelletteria di lusso che comprende sia la preparazione della pelle pregiata che la produzione degli articoli finiti: borse, cinture, guanti, ecc. Questo mercato rappresenta il cuore del Made in Italy. L'Italia è rinomata per gli investimenti in ricerca e sviluppo dei prodotti, per la cura nello sviluppo dei prototipi e per l'attenzione al dettaglio. Per questo motivo molte aziende del lusso mondiale si affidano alla competenza italiana per migliorare le loro collezioni [36].

Il mercato globale della pelletteria di lusso ha generato un fatturato di 63 miliardi nel 2022, con una previsione di crescita del 17% nel 2023 e ipotizzando di raggiungere quasi 85 miliardi nel 2027 come mostrato in figura 6-12 [37].

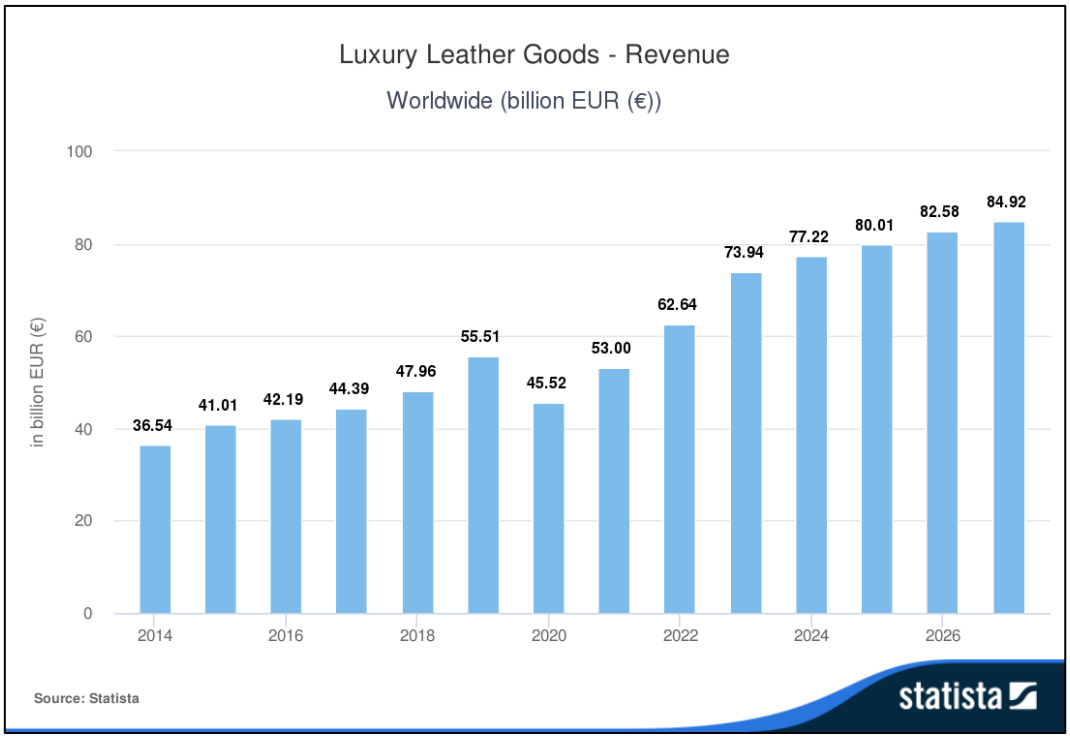


Figura 6-12. Ricavi beni di lusso in pelle. Fonte: Statista

I prodotti in pelle rappresentano una delle aree più fortemente in crescita nel mercato dei beni di lusso con potenzialità di sviluppo superiori rispetto alla media dei tassi di crescita del mercato come emerge in figura 6-13 [36], [37].

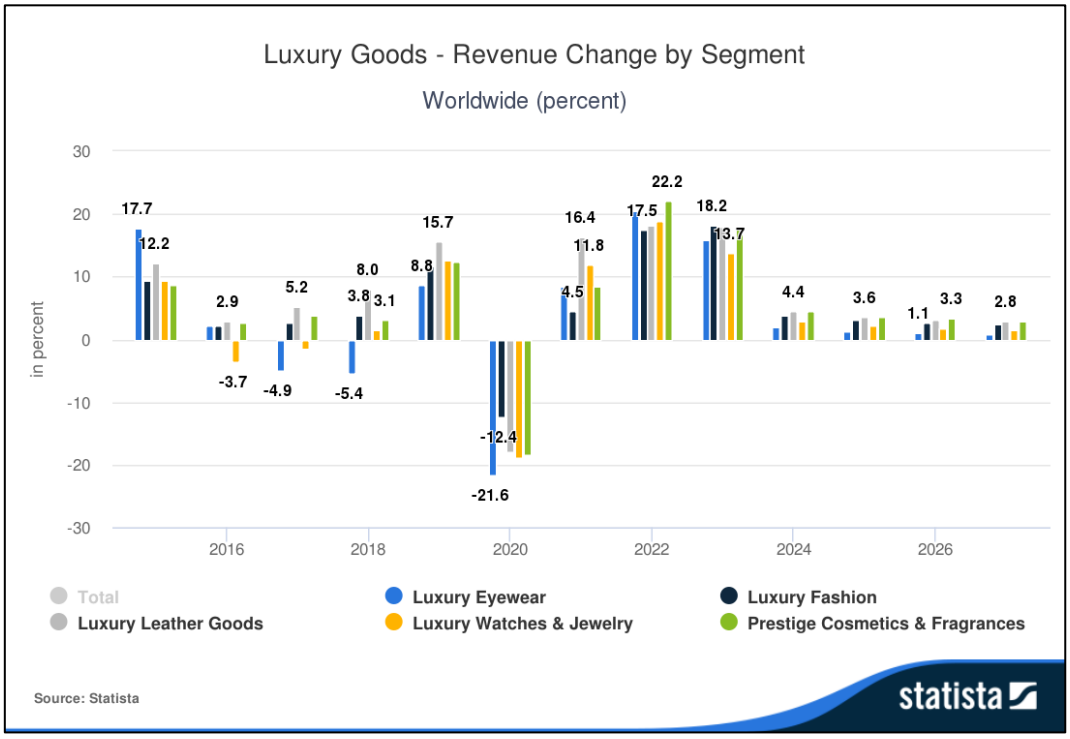


Figura 6-13. Tasso di crescita dei ricavi per segmento di mercato. Fonte: Statista

Come si vede dalla figura 6-13 il Covid ha avuto un grande impatto sul settore della pelletteria di lusso che ha registrato un forte stop alla crescita nel corso del 2020 [36]. La situazione si è poi ripresa gradualmente nel 2021 e riuscendo a ottenere livelli di fatturato nel 2022 superiori rispetto al periodo pre-pandemico [37].

Analizzando la situazione globalmente, il continente asiatico è il più profittevole, seguito dalle Americhe e dall'Europa (vedi figura 6-14) mentre gli Stati Uniti generano la maggior parte dei ricavi del settore seguiti dalla Cina (vedi figura 6-15) [37], [38]. Una ricerca di *BoF Insights* evince che anche se nel tempo i gusti e i comportamenti di acquisto cambiano, i consumatori di Stati Uniti e Cina rimangono fidelizzati al marchio che fornisce loro funzionalità, stile e uno status symbol [38]. Le stime positive di Altagamma Consensus stimano una crescita solida per il 2023 suddivisa su tutti i mercati con Asia e Medio Oriente in testa davanti a Europa e Usa [39].

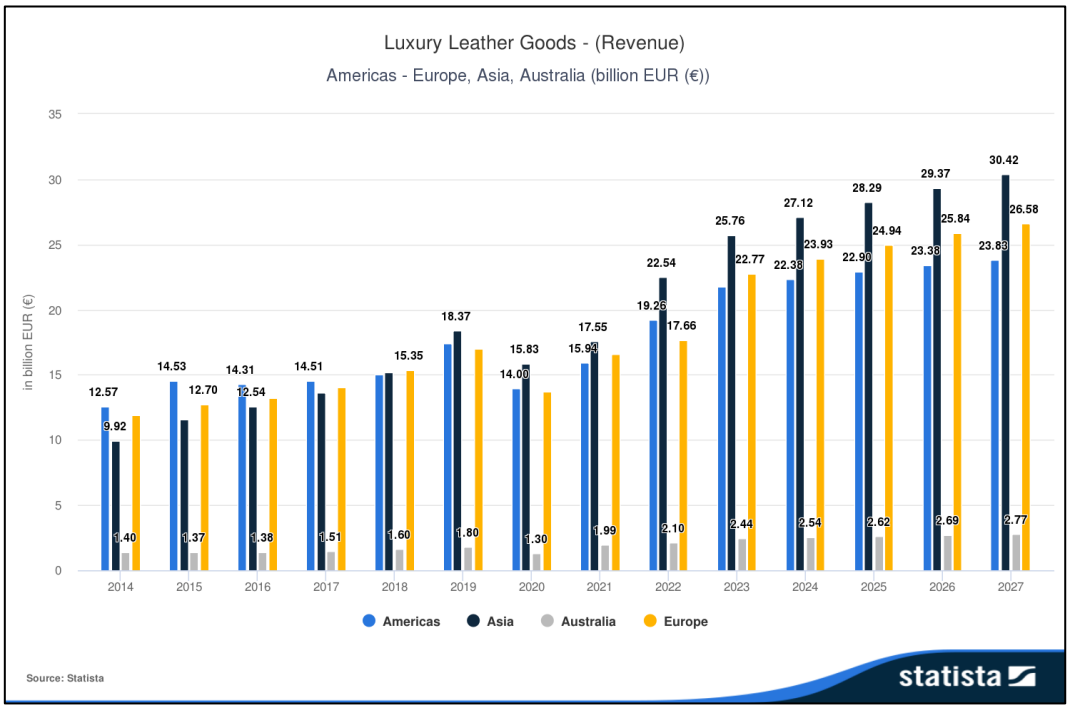


Figura 6-14. Ricavi beni di lusso in pelle. Confronto tra America, Europa, Asia, Australia. Fonte: Statista

10

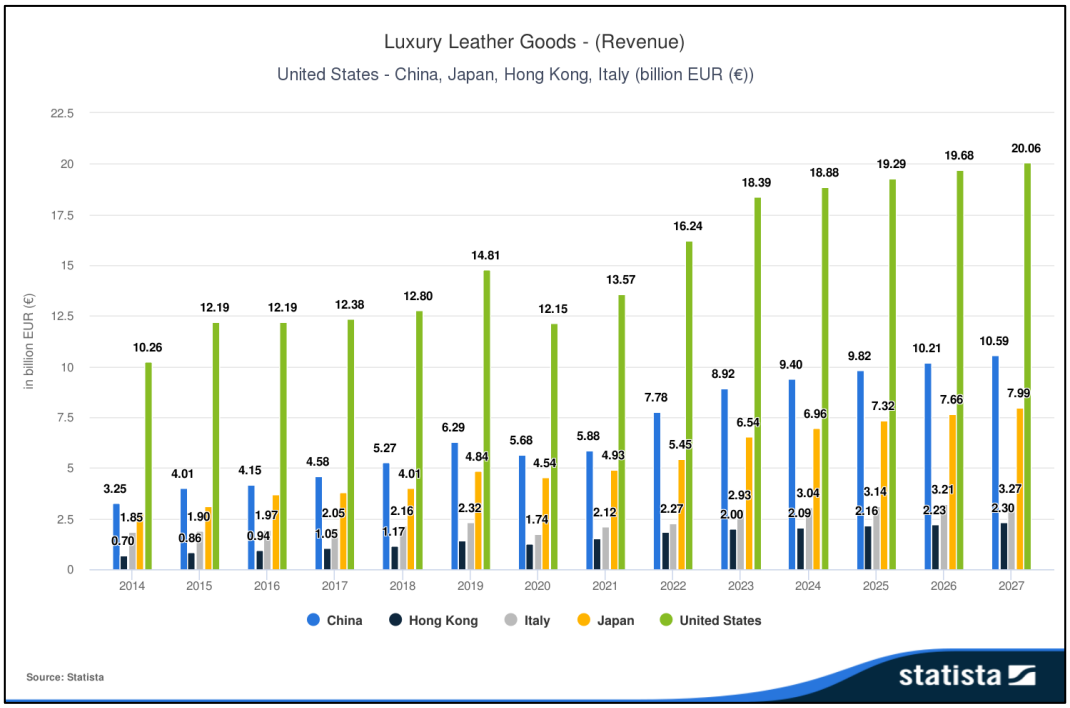


Figura 6-15. Ricavi beni di lusso USA, Cina, Hong Kong e Italia. Fonte: Statista

La clientela del lusso è costituita prevalentemente da individui con un patrimonio netto elevato: HNWI – High net worth individual [38]. Secondo AltaGamma entro il 2025 i Millenial e la Generazione Z costituiranno oltre il 60% del mercato rispetto al 40% nel 2019 (figura 6-16) [40], [41]. In generale il lusso conta oggi circa 400 milioni di consumatori nel mondo [39].

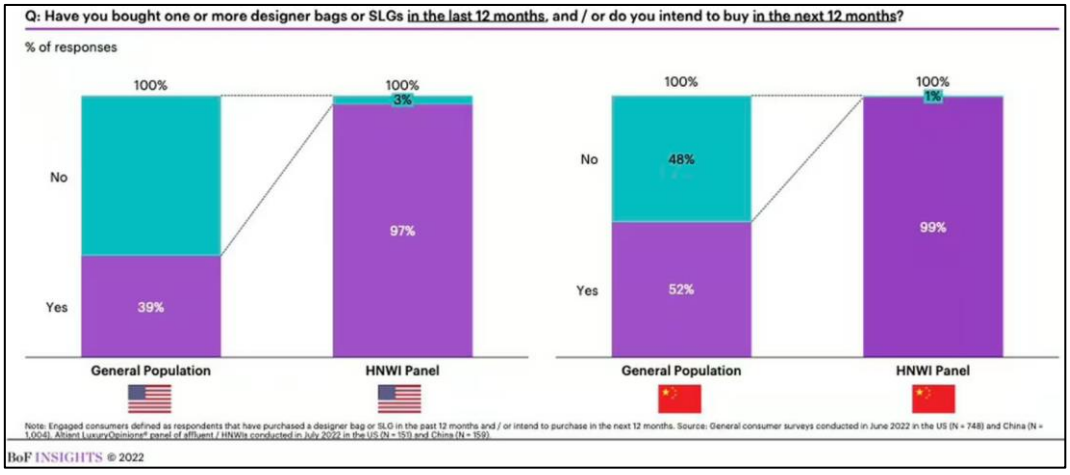


Figura 6-16. Intervista BoF INSIGHTS 2022.

Il mercato dei beni di lusso è costituito da un numero sempre più elevato di competitor. Per rimanere significativi i marchi cercano di possedere uno stile specifico costituito da forma, materiali e varianti di colore in modo da sottolineare la loro unicità [38]. Le

peculiarità dei marchi del mercato di lusso infatti sono: l'unicità, il pregio, la qualità eccellente e il prezzo elevato. Per questo motivo le aziende che operano nel settore dei beni di lusso devono riuscire a innovarsi costantemente, percepire le tendenze del mercato e adottare strategie per soddisfare la domanda dei consumatori [42].

Concludendo, emerge che le potenzialità del settore dei beni di lusso in pelle sono notevoli offrendo un quadro positivo per le aziende del settore e in particolare le azioni strategiche dovrebbero puntare sulle nuove generazioni [37], [40].

7. Caso studio

L'azienda di consulenza gestisce da parecchi anni il data warehouse del cliente, grazie al cui consenso, è stato possibile accedere allo storico delle vendite ed estrarre i dati necessari per le analisi. Il settore della vendita al dettaglio è un'area di applicazione adatta al data mining, poiché raccoglie enormi quantità di dati sulle vendite. La quantità di dati raccolti continua ad espandersi rapidamente, soprattutto a causa della crescente popolarità delle attività commerciali condotte sul Web. L'analisi dei dati dopo esser stati raccolti può essere uno strumento utile per ottimizzare il raggiungimento degli obiettivi della funzione merchandising. Inoltre, grazie all'esperienza del tutor aziendale è stato possibile conoscere in modo approfondito le funzioni svolte dal merchandising e scegliere gli attributi di interesse per le analisi. In particolare, l'obbiettivo che ci si è posti è quello di aiutare il merchandising a definire gli articoli che stiano performando meglio e peggio nel corso della stagione ed analizzare a posteriori per ogni area geografica in ambito full price quali siano le caratteristiche dei prodotti più e meno performanti.

7.1 Dataset

Generando un prompt su Microstrategy, software utilizzato dal cliente, si estraggono i dati relativi alle ultime due stagioni fall – winter (FW2021, FW2022) che coprono un periodo temporale da maggio 2021 a novembre 2022. Il dataset contiene circa 275mila record e più di 35 colonne. In particolare, gli attributi selezionati dal dwh sono:

1. Pl. Region e Pl. Region Desc: il primo attributo è una stringa di caratteri che descrive il codice dell'area geografica di competenza. Può essumere 9 diversi valori (AU, CA, CN, EMEA, HK, JP, RU, SG, US) la cui descrizione è riportata nel secondo attributo, 'Planning Region Desc', il quale contiene per esteso l'area geografica: AUSTRALIA, CANADA, CHINA, EMEA, HKMC, JAPAN, RUSSIA, SINGAPORE, USA.
2. Pl. Channel: stringa di caratteri che riporta il canale di vendita della merce e può assumere 10 diversi valori (tra parentesi si porta la variabile 'planning region desc'):
    - AL (others);

- CO (corner);
- DP (department store);
- EC (e-commerce);
- EM (e-commerce market place);
- ER (e-commerce retail);
- FR (franchising);
- OU (outlet);
- PR (property full price);
- SS (showroom sales).

3. Pl. Channel Desc: contiene per esteso la descrizione del canale di vendita. È suddiviso in:
    - CORNER = area dedicata a un singolo brand (monomarca) all'interno di un punto vendita multimarca come ad esempio un grande magazzino;
    - DEPARTMENT STORE = tipologia di vendita nata a Parigi nella metà dell'Ottocento in cui l'edificio ha una sua autonomia architettonica e internamente è diviso tra reparti specializzati;
    - E-COMMERCE = vendita di prodotti ai rivenditori (B2B).
    - E-COMMERCE MARKET PLACE = piattaforma digitale in cui più rivenditori mettono a disposizione i propri prodotti. È come se fosse un grande centro commerciale;
    - E-COMMERCE RETAIL = vendita di prodotti online al cliente finale, rappresenta il negozio online;
    - FRANCHISING = è una formula di collaborazione tra imprenditori indicata per chi vuole avviare una nuova impresa ma non parte da zero e affilia la propria attività ad un marchio già affermato;
    - OTHERS;
    - OUTLET = punto vendita fisico che offre al pubblico i prodotti di fine serie di grandi marche a pprezzi scontati;
    - PROPERTY FULL PRICE = vendita al dettaglio dei prodotti a prezzo pieno.
    - SHOWROOM SALES = Spazi illustrare e vendere a clienti e distributori il prodotto dove viene ideato e realizzato.

4. Merchandise Class: sigla che descrive la categoria merceologica;
5. Merchandise Class Desc: descrizione della categoria merceologica;
6. Statistical Class: sigla alfanumerica di tre caratteri che descrive la linea del prodotto;
7. Statistical Class Desc: stringa alfanumerica che specifica il significato del valore espresso in 'statistical class'. Ad esempio, *statistical class = MB2* e *statistical class desc = 'MAIN BAG 2'* (nome fittizio).
8. Product: codice prodotto costituito da 6 valori numerici. Sono presenti 3352 valori distinti;
9. Product Desc: concatenazione del valore presente nell'attributo 'style desc' e del valore presente nella variabile 'material desc';
10. Style: codice alfanumerico di sette caratteri che definisce il modello del prodotto;
11. Style Desc: descrizione del modello del prodotto. Sono presenti 1354 valori distinti.
12. Material: riporta il materiale con cui è stato fatto il prodotto. È un codice alfanumerico di 6 caratteri. È costitutito da 1100 elementi differenti.
13. Material Desc: è costituito da 1097 valori. Descrive i materiali con cui sono stati fabbricati i prodotti.
14. Color: codice alfanumerico di 5 cifre che corrisponde al colore del prodotto. È costituito da 1335 valori.
15. Color Desc: riporta la descrizione dei colori che compongono i prodotti concatenati con un '+'. Sono presenti 1332 valori differenti;
16. Last Coded Season: Riporta il codice dell'ultima stagione di vendita del prodotto.
17. Last Coded Season Desc: Riporta la descrizioned dell'ultima stagione di vedntia;
18. Variant: le varianti possibili per ogni prodotto sono 72. È un codice alfanumerico di 3 caratteri.
19. Variant Desc: descrive la tipologia delle varianti.
20. Variant Type: per ogni prodotto si riporta anche il tipo della variante. L'attributo è costituito da 57 diversi valori.

21. Variant Type Desc: descrive il tipo di variante.
22. First Coded Season: riporta il codice della prima stagione in cui il prodotto è stato messo in vendita;
23. Consolidated Product Type: definisce una categoria di prodotto stabilita dall'azienda. Può essere: C, N, O, S
24. Consolidated Product Type Desc: riporta la descrizione della categoria di prodotto C = CARRY OVER, N = NEWNESS, O = OLD PRODUCT, S = STILL VALID.
25. Commercial Macroseason: codice alfanumerico che riporta il tipo di stagione commerciale costituito da: (SS oppure FW) + anno di riferimento + (FP oppure OP);
26. Commercial Macroseason Desc: codice alfanumerico che riporta il tipo di stagione commerciale costituito da: (Spring Summer oppure Fall Winter) + anno di riferimento + (Full Price oppure Off Price);
27. Consolidated Is Markdown: definisce il fatto che il prodotto sia stato scontato. Può assumere tre valori: N, O, Y descritti nell'attributo 'consolidated is markdown;
28. Consolidated Is Markdown Desc: No, Old product, Yes;
29. Consolidated Is Markdown LY: riporta se il prodotto sia stato venduto in sconto l'anno precedente. Valori possibili: N, O, Y, ND (Not Defined);
30. Consolidated Is Markdown LY Desc: descrizione estesa di 'consolidated is markdown LY';
31. Sales Commercial Week: periodo temporale corrispondente alle vendite riportate nel set di dati. I dati estratti si riferiscono a 108 settimane commerciali.

Le metriche estratte dal data warehouse sono:

32. Net Sales €: fatturato netto;
33. Discount €: sconto applicato al prodotto;
34. Net Discount €: sconto del prodotto al netto dell'iva;
35. Sales TB €: costo dei prodotti venduti;
36. Sales U: numero delle unità vendute.

I dati nel data warehouse possiedono già un buon livello di qualità ma per migliorare le performance degli algoritmi che verranno applicati in seguito è necessario eseguire ulteriori passaggi di data preprocessing.

7.2 Data preprocessing

La fase di preparazione dei dati è stata svolta sfruttando il software *RapidMiner* [43], un tool di data mining open source che possiede un'interfaccia drag-and-drop facilmente personalizzabile in base al caso di studio. Sfruttando quindi RapidMiner si genera la pipeline per svolgere la data preprocessing. In particolare, questo step è utile per preparare i dati per la successiva fase di data mining.

L'obiettivo è supportare il team merchandising a decidere quali tra i prodotti nuovi venduti a prezzo pieno (mercato Full Price) di una stagione riproporre la stagione successiva e quali escludere. A questo proposito si suddivide il dataset nelle 2 stagioni di interesse: FW2021_FP e FW2022_FP. Inizialmente il focus si sposta verso la stagione FW2021_FP da cui si cercherà di estrarre informazioni utili sulle performance dei prodotti *nuovi* in ogni area di mercato per poi confrontare i risultati ottenuti con le performance dei prodotti riproposti la stagione successiva, FW2022_FP.

I dati di partenza della fase di preprocessing corrispondono alla stagione autunno-inverno 2021 (FW2021_FP). Sfruttando l'operatore Read Excel si importano i dati precedentemente estratti dal database. Di seguito, dalla figura 7-1 alla figura 7-8 è possibile osservare l'intera pipline generata su RapidMiner.

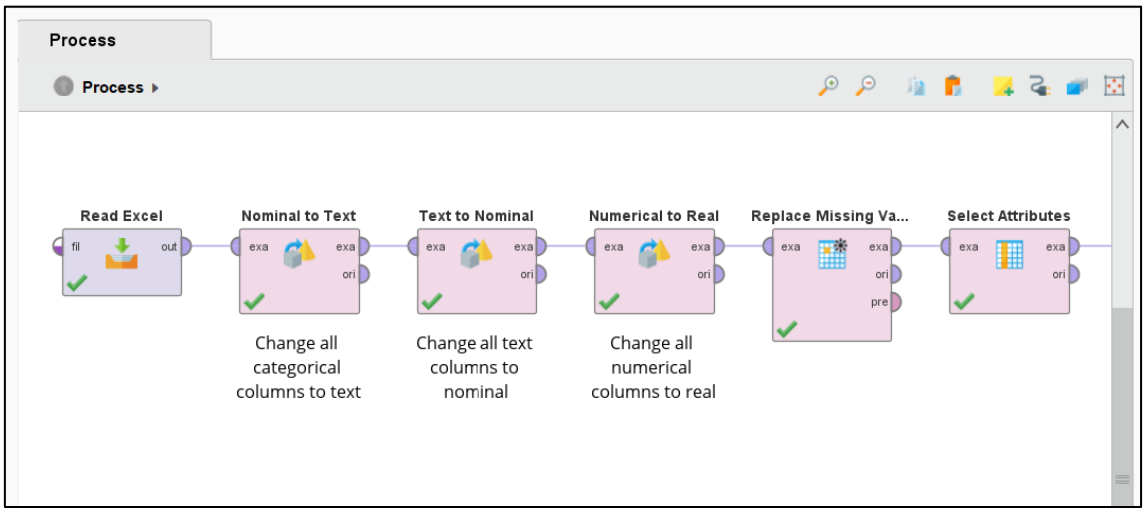


Figura 7-1. Pipline data preprocessing 1

In primis si esegue lo step di *Data Cleaning* ovvero di pulizia dei dati in cui si gestiscono i valori mancanti e si correggono eventuali incongruenze tra i dati. Più precisamente si aggiorna il tipo di dato e si gestiscono i valori mancanti tramite l'operatore *Replace Missing Values*. L'unica colonna che presenta valori mancanti è quella di Net Discount, per questo si decide di riempire le celle vuote con una costante globale pari a 0, in modo da riportare uno sconto pari a zero nel caso in cui questo non sia stato applicato.

Il secondo step che si affronta è quello di Data Reduction, per eliminare attributi superflui e ridurre il dataset in termini di volume cosicché gli algoritmi di data mining possano essere più efficienti. Analizzando il dataset dopo l'estrazione si nota che alcune colonne sono prive di valore in quanto non possiedono informazioni utili per l'obiettivo. Sono stati rimossi quindi due attributi: *Consolidated Is Markdown LY*, *Consolidated Is Markdown LY Desc*. Inoltre, poiché lo studio si sofferma sul product core business dell'azienda la classe merceologica presenta un unico valore e può essere rimossa. In precedenza, è stato detto che si analizzerà una stagione alla volta e quindi l'attributo *Commercial Macroseason* assumerà un solo valore allora viene eliminato con l'attributo contenente la sua descrizione, ovvero *Commercial Macroseason Desc*. Inoltre, si decide di eliminare lo sconto a lordo dell'iva poiché ai fini delle analisi è più utile lo sconto netto (*Net Discount*). Tutte queste operazioni sono state eseguite sfruttando l'operatore *Select Attributes*.

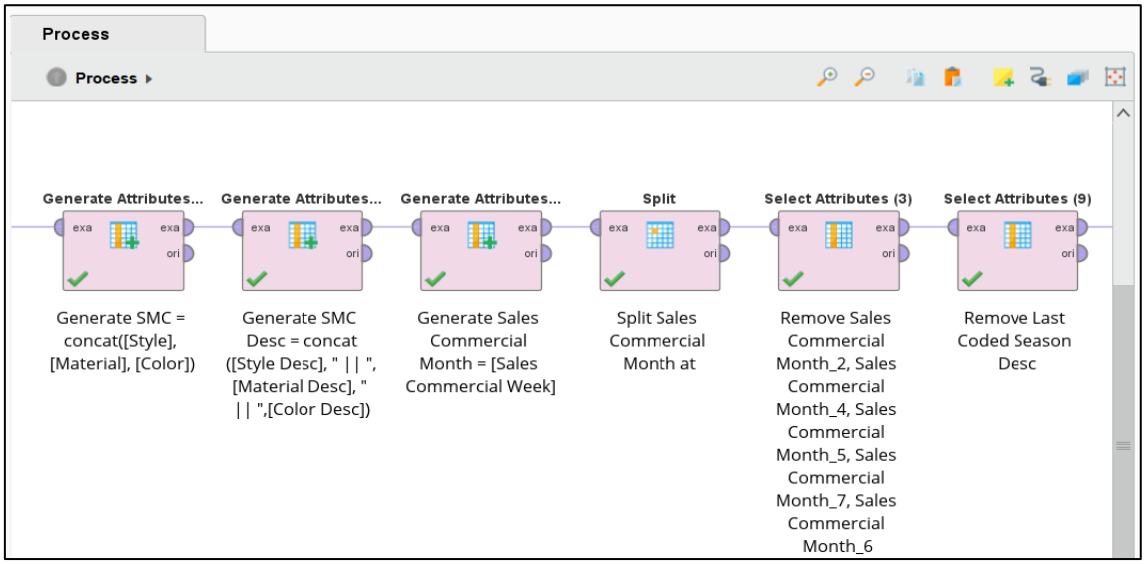


Figura 7-2. Pipline data preprocessing 2

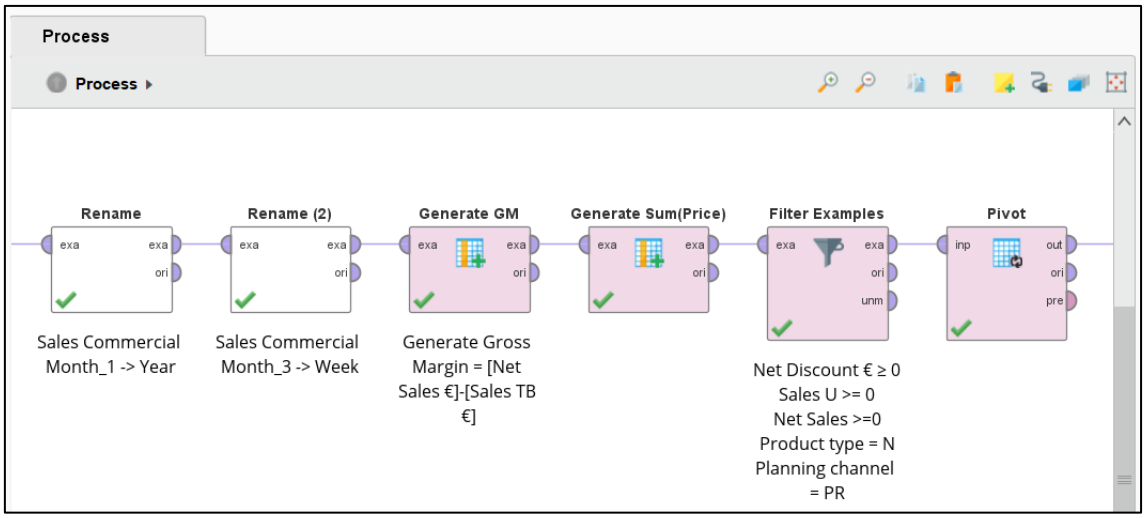


Figura 7-3. Pipline data preprocessing 3

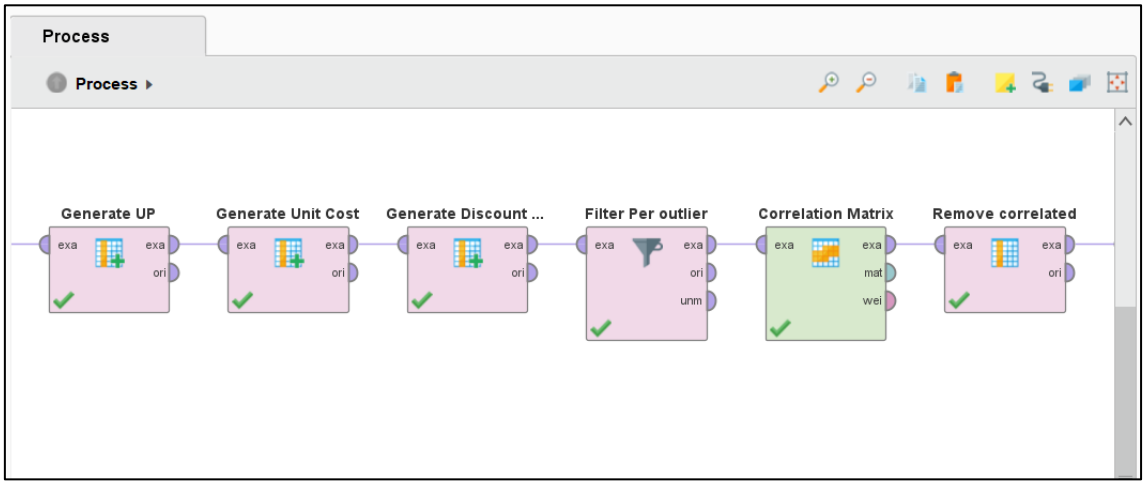


Figura 7-4. Pipline data preprocessing 4

Durante la terza fase (*Data transformation)* si modificano i dati nella forma appropriata per il data mining. In primis, si costruiscono nuovi attributi:

- SMC è la concatenazione di style, material, color in modo da definire una nuova granularità di prodotto. Il codice risultante è un codice alfanumerico;
- SMC Desc è la descrizione delle componenti style, material e color che definiscono il prodotto. Gli attributi sono concatenati con il simbolo '||';
- Year = si estrae l'anno dalla stringa Sales Commercial Week;
- Week = si estrae la settimana dalla stringa Sales Commercial Week;
- Gross Margin=Net Sales-Sales TB, indica il profitto diretto che un'azienda ottiene dalla vendita di un prodotto;

Successivamente si prosegue con la fase di *Data Reduction* filtrando alcune colonne solo per i valori di interesse. Innanzitutto, poiché le analisi sono da svolgere per i prodotti NEWNESS, si filtra l'attributo *Consolidated product type*. Nel dataset i dati sono aggregati settimanalmente e questo comporta un problema nel calcolo delle unità vendute (Sales U) e del fatturato prodotto (Net Sales) in quanto eventuali resi possono essere compensati con le vendite e generare informazioni errate come il fatto che Sales TB sia pari a zero o Net Sales sia negativo. Per questo motivo si decide di selezionare solo i dati legati a *Sales U* e *Net Sales* maggiori di 0. Dopodiché, si filtra *Net Discount* per un valore maggiore o uguale a zero in quanto uno sconto negativo definirebbe uno storno dello sconto e quindi può essere trascurato. Infine, l'attributo *Planning Channel*, come si è visto nel capitolo in cui si descrive il dataset, può assumere nove differenti valori ma ai fini dell'analisi solo uno di questi è rilevante: property full price.

Giunti a questo punto si aggregano i dati di vendita per prodotto e area di mercato sfruttando l'operatore PIVOT che svolge la funzione di GROUP BY in linguaggio SQL. In seguito, si generano altri tre attributi:

- Unit Cost = (Sales TB) / (Sales U), indica il costo di ogni unità di prodotto;
- Unit Price (derived) = (Net Sales+Net Discount) / (Sales U), prezzo unitario del prodotto ottenuto dal fatturato, dallo sconto applicato e dalle unità vendute.
- Discount Percentage = sconto percentuale applicato in media a ogni prodotto nel corso della stagione.

Si passa poi alla fase di *outlier detection* al fine di verificare la presenza di outlier all'interno del dataset, utilizzando una strategia univariata. Analizzando la distribuzione dei dati si notano dati che presentano valori eccessivamente diversi rispetto al contesto quindi si decide di filtrare per:

- $Sum(Net\ Sales\ €) \leq 100K$
- $Sum(Sales\ U) < 250$

come mostrato in figura 7-5 e in figura 7-6.

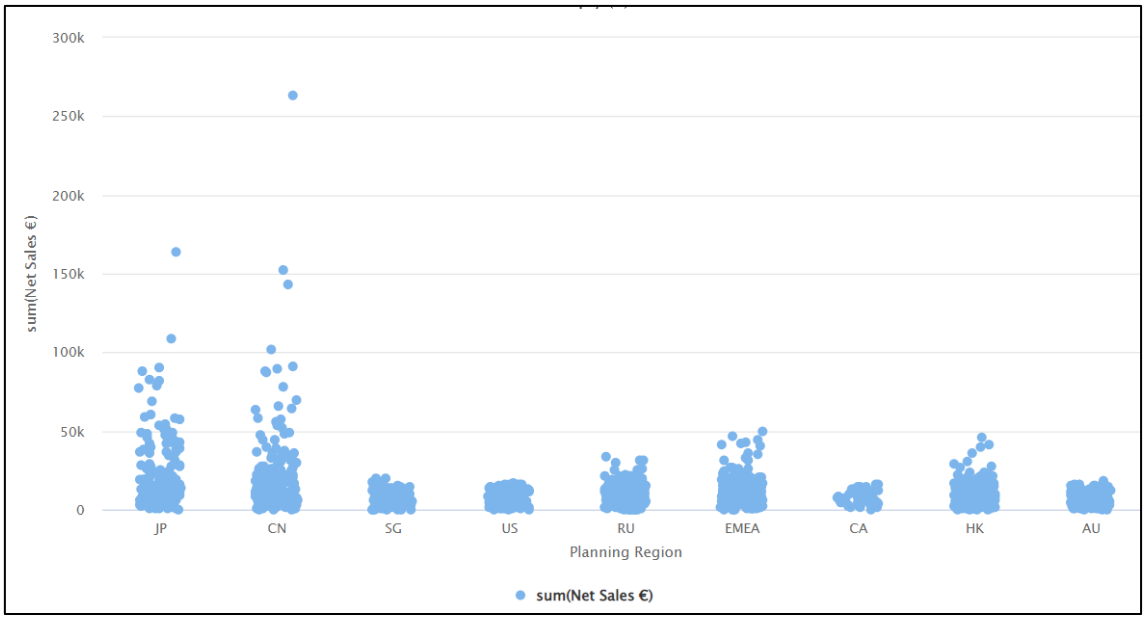


Figura 7-5. Outlier

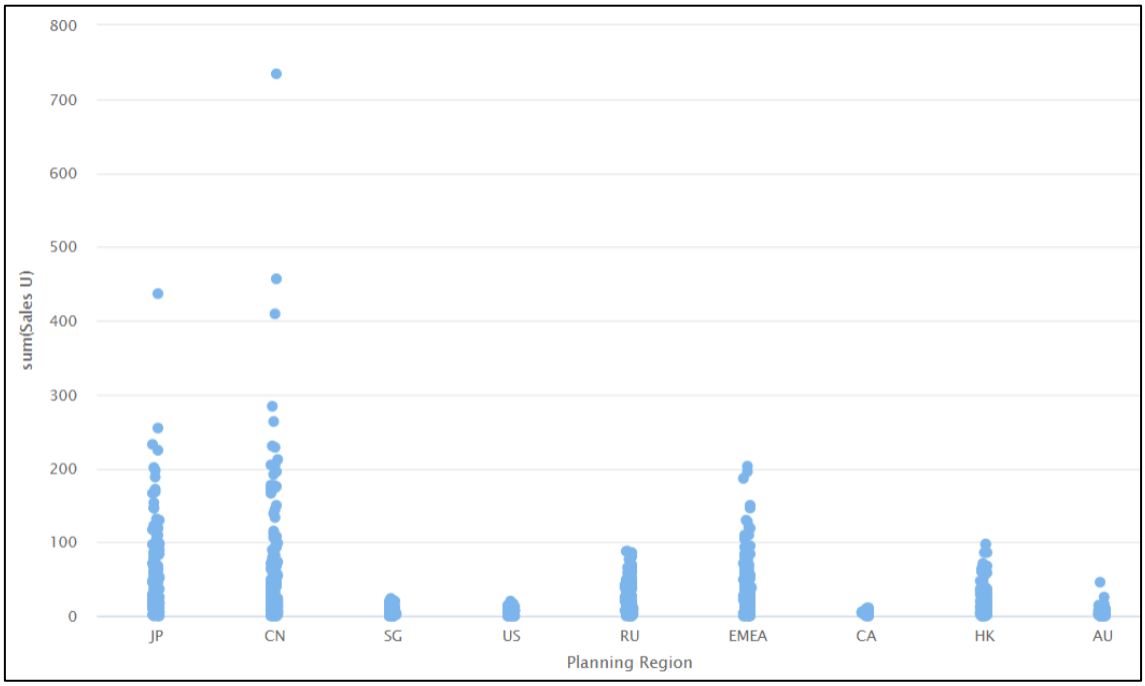


Figura 7-6. Outlier

Dopo aver rimosso gli outlier si prosegue con la fase di *correlation analysis* e tramite la matrice di correlazione si osserva quali siano i legami tra gli attributi e nel caso in cui esista una correlazione maggiore di 0,75, si procede tenendo solo una variabile, quella che si considera la più significativa.

Attributes	sum(Sales U)	sum(Gross Margin)	sum(Net Sales €)	sum(Sales TB €)	Average Price	Average Cost	Average Discount Percentage
sum(Sales U)	1	0.941	0.960	0.973	-0.002	-0.010	-0.040
sum(Gross Margin)	0.941	1	0.996	0.960	0.150	0.077	-0.065
sum(Net Sales €)	0.960	0.996	1	0.982	0.129	0.084	-0.056
sum(Sales TB €)	0.973	0.960	0.982	1	0.082	0.097	-0.035
Average Price	-0.002	0.150	0.129	0.082	1	0.789	0.159
Average Cost	-0.010	0.077	0.084	0.097	0.789	1	-0.015
Average Discount Percentage	-0.040	-0.065	-0.056	-0.035	0.159	-0.015	1

Figura 7-7. Matrice di correlazione

Dalla matrice di correlazione emerge una stretta relazione tra Sum (Sales U), Sum (Net Sales €), Sum(Gross Margin) e Sum(Sales TB €). Confrontandosi con l'esperto di dominio si decide di mantenere solo Sum (Net Sales €) in quanto si ritiene essere il più significativo per il business e si passa alla rimozione degli altri. Si può notare anche che il prezzo medio e il costo medio siano fortemente correlati e si decide di mantenere il costo medio unitario (Average Cost).

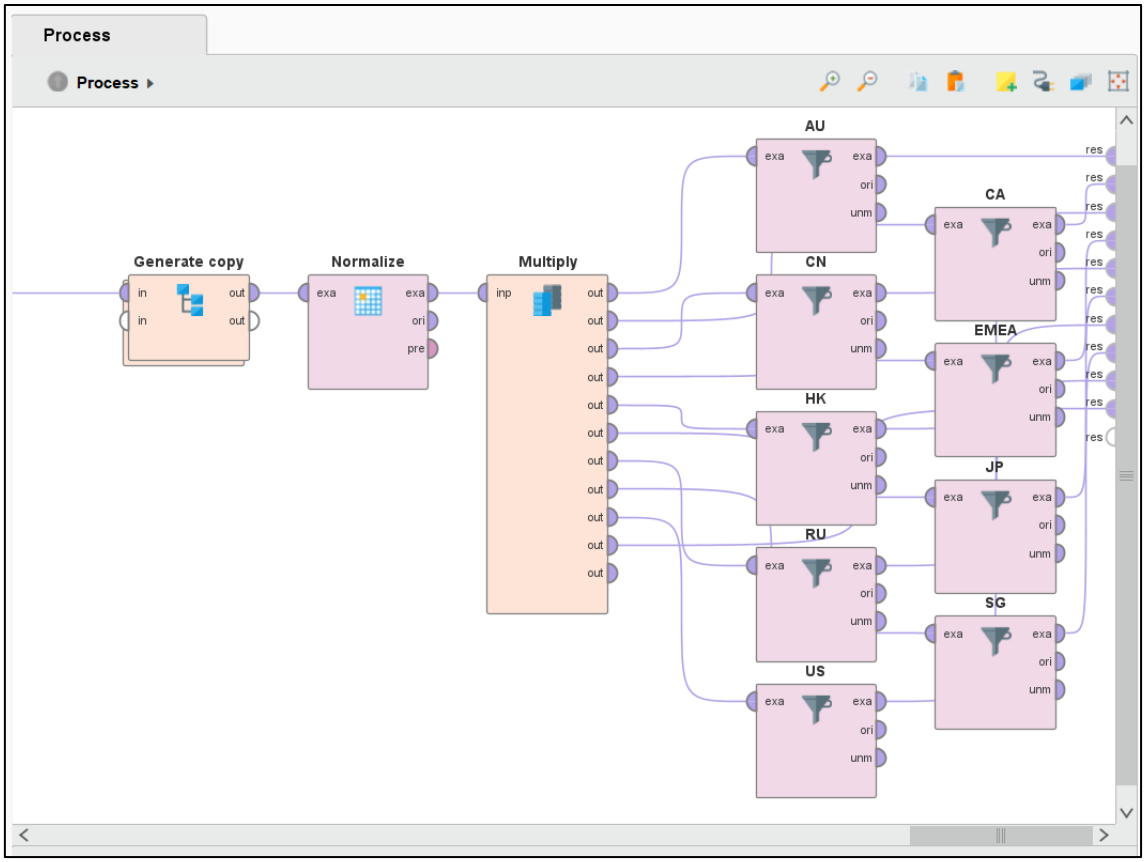


Figura 7-8. Pipline data preprocessing 5

L'ultimo step di data preprocessing è costituito dalla normalizzazione degli attributi che saranno oggetto della successiva fase di data mining. La fase di normalizzazione degli

attributi numerici è molto utile per gli algoritmi di data mining che coinvolgono misure di distanza come il clustering che si utilizzerà in seguito. Si utilizza il metodo *range transformation* che normalizza tutti i valori in un range specificato tra un minimo e un massimo, che in questo caso sarà rispettivamente pari a 0 e 1. Tutti gli altri valori vengono scalati, in modo da rientrare nell'intervallo indicato. Per non perdere l'informazione sul valore reale degli attributi si decide di generare una copia delle metriche rimanenti, a valle della correlation analysis, che potrebbero essere oggetto del clustering creando gli attributi:

- Nrom_AC
- Norm_DP
- Norm_NS

rispettivamente collegati a costo unitario medio, sconto percentuale medio, fatturato totale.

Infine, l'ultimo passaggio consiste nella suddivisione dei prodotti per ogni area di mercato in quanto in accordo con l'esperto di dominio si decide di proseguire le analisi per reagion. Per questo motivo vengono generati nove file Excel che saranno l'input della fase successiva.

7.3 Data mining

La fase di data mining è stata svolta utilizzando il tool Google Colaboratory [44] uno strumento gratuito di Google che permette di programmare in Python scrivendo codice direttamente in Cloud dal proprio browser. Il codice che si utilizza per le analisi è stato frutto di una profonda ricerca online e di una costante rielaborazione.

In questo capitolo il set di dati precedentemente preparato è pronto per essere sottoposto alla fase di data mining. Il risultato ottenuto sarà frutto della comparazione di diversi algoritmi e di diverse variabili in input. Si è scelto di segmentare i prodotti per ogni area di mercato confrontando tre diversi metodi di clustering: K-means, gerarchico agglomerativo e DBSCAN. Inoltre, ogni algoritmo è stato testato con input diversi:

- Norm_NS, Norm_AC, Norm_DP ovvero il valore normalizzato della metrica Net Sales, Average Cost e Discount Percentage;

- Norm_NS, Norm_DP ovvero il valore normalizzato della metrica Net Sales e Discount Percentage
- Norm_NS, Norm_AC ovvero il valore normalizzato della metrica Net Sales e Average Cost;

Infine, i tre scenari verranno successivamente confrontati tramite la *silhouette*, una misura di validazione dei cluster, per scegliere quale sia il metodo più adatto per il caso studio.

7.3.1 K-MEANS

Il primo algoritmo di clustering che si vuole testare è il k-means, uno degli algoritmi più utilizzati e performanti. Il metodo richiede come valore di input k, il numero di cluster da generare. Per questo motivo si genera su Google Colab l'elbow graph in modo da rappresentare graficamente il valore dell'SSE per una serie di valori k. Inoltre, si decide di rappresentare anche il valore di Silhouette per valori di k da 2 a 10. Successivamente si sceglie il valore di k in corrispondenza del gomito che si nota nell'elbow graph a cui corrisponde la configurazione migliore in quanto il guadagno che si otterrebbe aggiungendo un centroide sarebbe trascurabile, poiché la riduzione della misura di qualità non sarebbe più rilevante. Contemporaneamente si analizza il valore di Silhouette cercando di massimizzarla. La scelta del k ottimale viene ripetuta per ogni area di mercato.

Il codice utilizzato per svolgere questo step è riportato di seguito:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import K-means
from sklearn import datasets
from sklearn.metrics import silhouette_samples, silhouette_score

df = pd.read_excel("CN.xlsx")
A = 'Norm_NS'
#B = 'Norm_AC'
B = 'Norm_DP'
X = df[[A,B]]

#elbow method
```

```
sse = []
mapping = {}
K = range(1,11)
for k in K:
   kmeanModel = K-means(n_clusters=k)
   kmeanModel.fit(X)
   sse.append(kmeanModel.inertia_)
   mapping[k] = kmeanModel.inertia_

#Elbow graph
plt.figure(figsize=(16,8))
fig=plt.plot(K, sse, 'bx-')
plt.xticks(range(1, 11))
plt.xlabel('Number of Clusters')
plt.ylabel('SSE')
plt.title('Elbow Method')
plt.savefig('Elbow Method US.jpeg')
plt.show()

# Lista di valori di Silhouette per ogni k da 2 a 10
silhouette_coefficients = []

for k in range(2, 11):
  k-means = K-means(n_clusters=k)
  k-means.fit(X)
  score = silhouette_score(X, k-means.labels_)
  silhouette_coefficients.append(score)

# Silhouette graph
plt.figure(figsize=(16,8))
plt.plot(range(2, 11), silhouette_coefficients, 'bx-')
plt.xticks(range(2, 11))
plt.xlabel("Number of Clusters")
plt.ylabel("Silhouette Coefficient")
plt.title('Silhouette')
plt.savefig('Silhouette US.jpeg')
plt.show()
```

Si riportano di seguito rispettivamente l'elbow graph (figure 7-9, 7-11, 7-13) e il grafico della silhouette (figure 7-10, 7-12, 7-14):

- per la Cina per cui si è scelto un valore di k ottimale pari a 3;

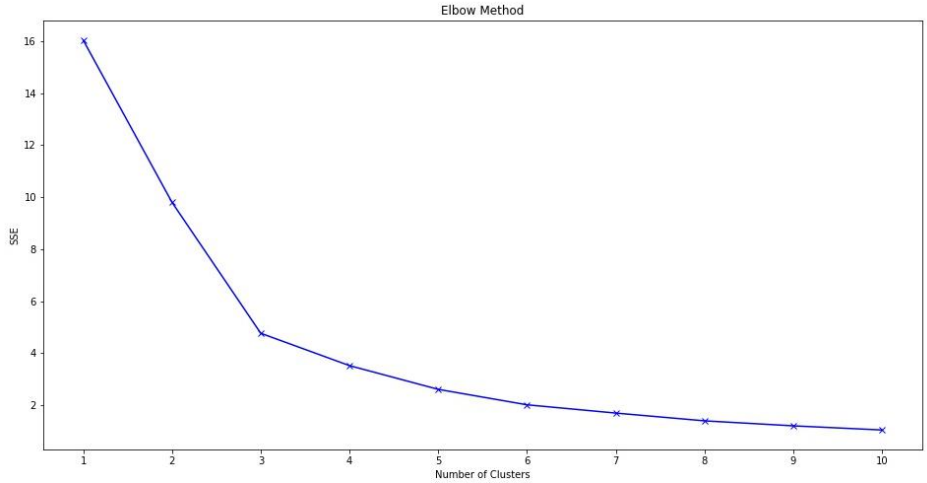


Figura 7-9. Elbow graph Cina

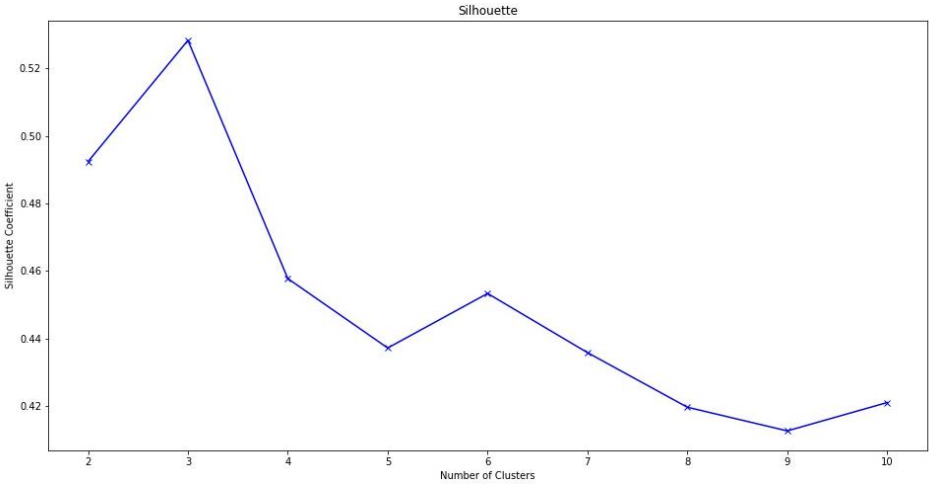


Figura 7-10. Silhouette graph Cina

- per gli Stati Uniti per cui si è optato per un valore di k pari a 4;

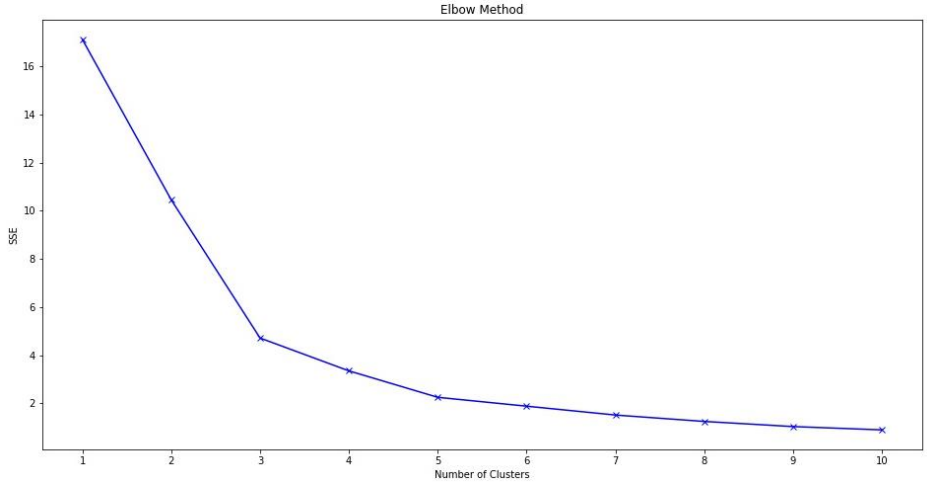


Figura 7-11. Elbow graph USA

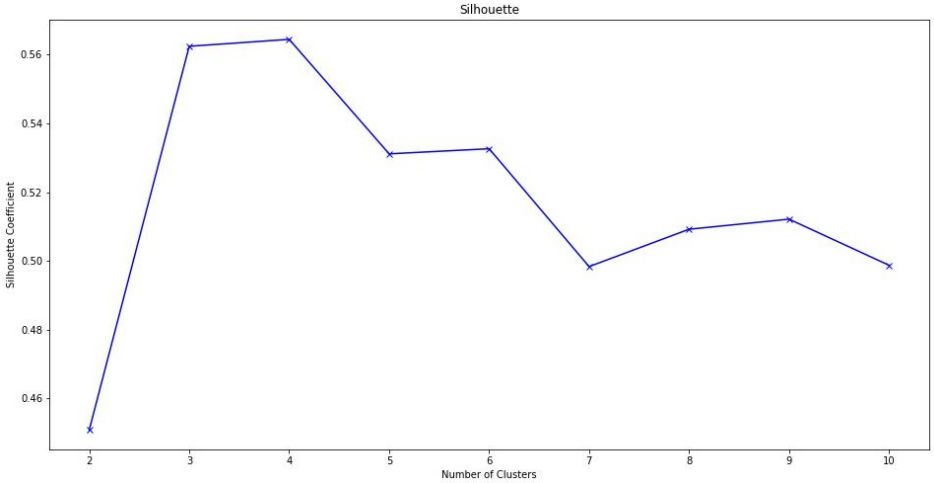


Figura 7-12. Silhouette graph USA

- per EMEA per cui si è scelto un valore di k pari a 5.

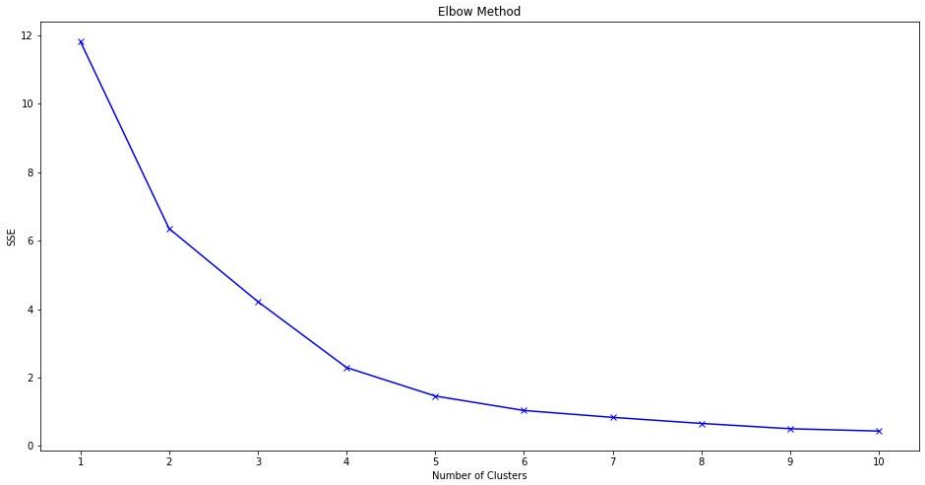


Figura 7-13. Elbow graph EMEA

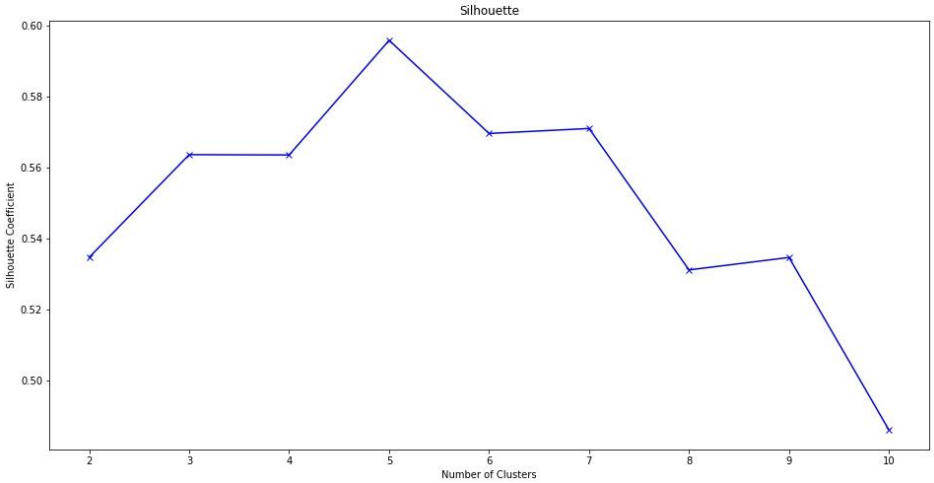


Figura 7-14. Silhouette graph EMEA

I valori di k scelti per ogni area di mercato sono riportati nella tabella 7-1.

Tabella 7-1. Valori k per ogni area di mercato

Area di mercato	k
AU	3
CA	4
CN	3
EMEA	5
HK	4
JP	2
RU	2
SG	2
US	4

Successivamente si implementa con Python l'esecuzione del k-means. Il codice utilizzato a questo scopo è il seguente:

```
# Crea un'istanza di K-means con k cluster pari al risultato ottenuto con il passaggio
precedente
k-means = K-means(init="k-means++",n_clusters=2)

# Si addestra il modello di K-means sul dataset
k-means.fit(X)

# Si assegnano le etichette dei cluster a ogni oggetto del dataset
df['Cluster'] = k-means.predict(X)
df.to_excel('Cluster US=4.xlsx')

# Crea un grafico a dispersione dei dati
plt.scatter(df[A],df[C], c=df['Cluster'], cmap='viridis')
plt.scatter(k-means.cluster_centers_[:, 0], k-
means.cluster_centers_[:, 1], s=200, c='red', marker='x', label = 'Centroids')
plt.title('Clusters k-means')
plt.xlabel(A)
plt.ylabel(C)
#plt.savefig('Cluster k-means US=4')
plt.show()

#per il valore silhouette
labels = k-means.labels_
print('Silhouette:',metrics.silhouette_score(X, labels, metric='euclidean'))
```

Grazie a questo è stato possibile assegnare i prodotti ad ogni cluster, creare un file Excel che sarà successivamente rielaborato e rappresentare graficamente l'output dell'algoritmo.

Si illustrano rispettivamente in figura …, … e …. i cluster di Cina, Stati Uniti ed EMEA.

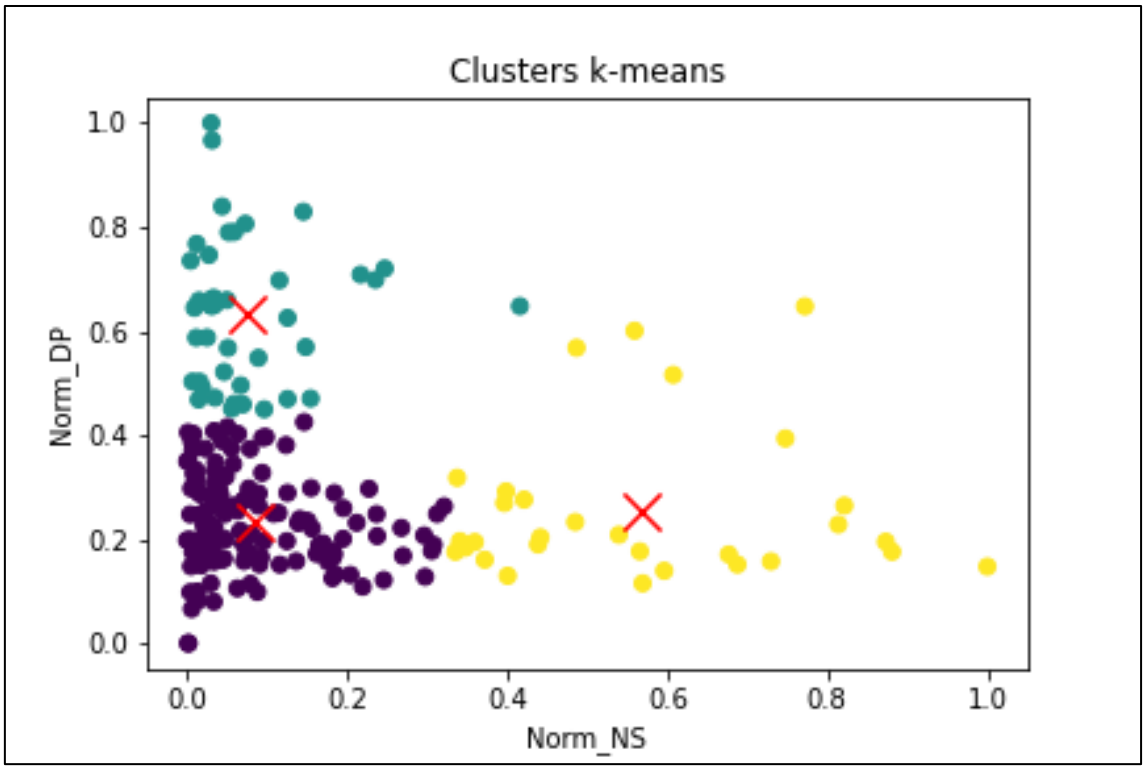


Figura 7-15. Cluster Cina

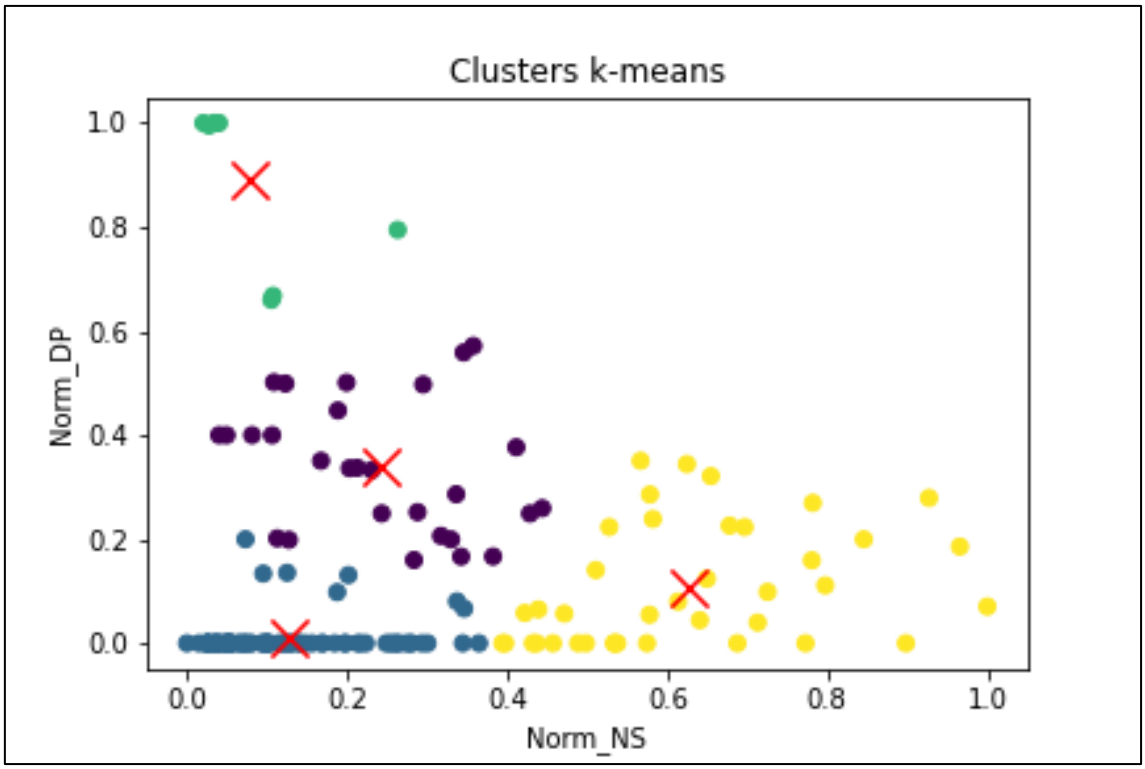


Figura 7-16. Cluster USA

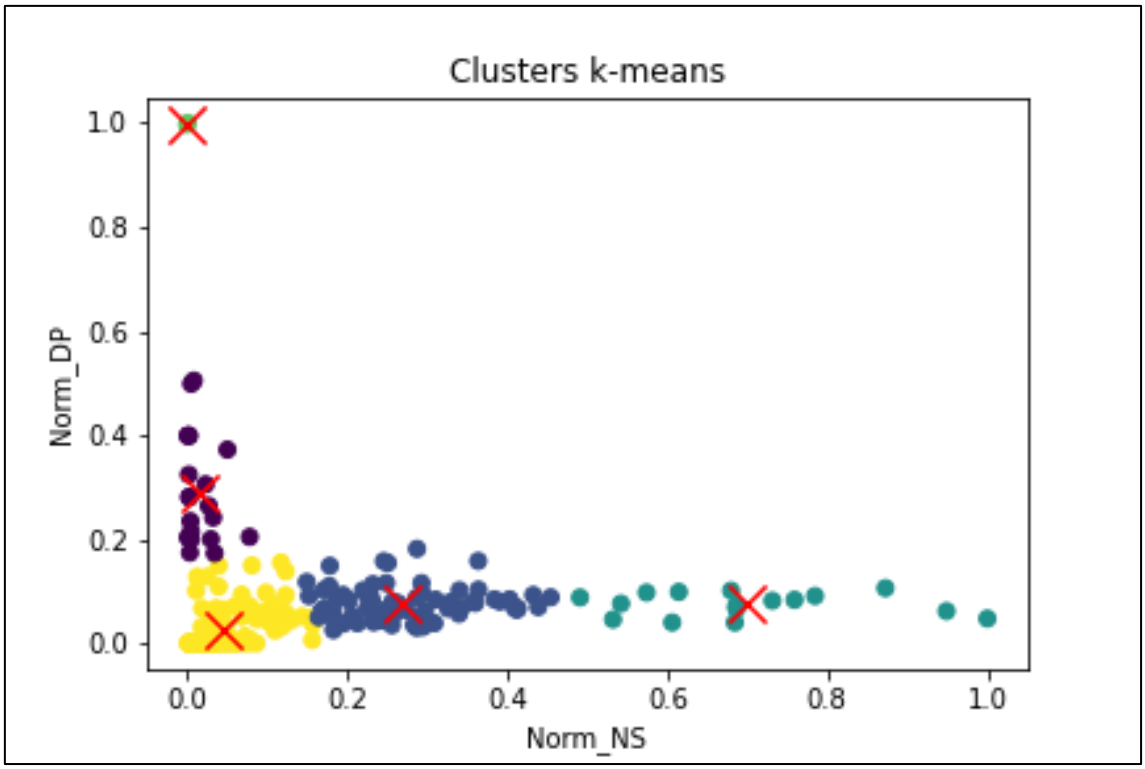


Figura 7-17. Cluster EMEA

Osservando il grafico in EMEA emerge la necessità di una fase di post processing per ottimizzare i risultati. Più nel dettaglio in questo caso si elimina un cluster di piccole

dimensioni, costituito da 3 prodotti, che potrebbe rappresentare valori anomali ma generalmente si potrebbero anche unire cluster vicini che presentano un valore di SSE basso.

Dopo la fase di post processing i valori aggiornati di k sono riassunti nella tabella 7-2.

Tabella 7-2. Valori di k per ogni area di mercato dopo il posto processing

Area di mercato	k
AU	3
CA	4
CN	3
EMEA	5
HK	4
JP	2
RU	2
SG	2
US	4

Un esempio di rappresentazione 3D in presenza di tre variabili considerate come input del clustering è presente in figura 7-18.

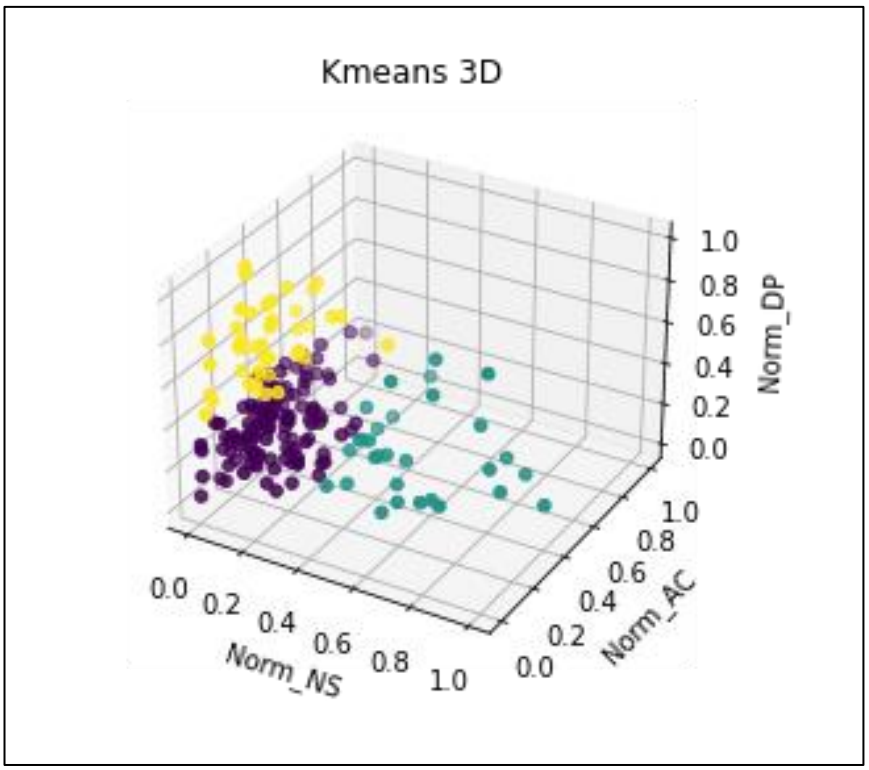


Figura 7-18. K-means 3D

Il codice utilizzato per generare il grafico precedente è:

```
#grafico 3D k-means
from mpl_toolkits import mplot3d
%matplotlib inline
import numpy as np
import matplotlib.pyplot as plt
fig = plt.figure()
ax = plt.axes(projection='3d')
ax.scatter3D(df[A], df[B],df[C], c=df['Cluster'], cmap='viridis')

ax.set_title('K-means 3D')
ax.set_xlabel(A)
ax.set_ylabel(B)
ax.set_zlabel(C)

plt.savefig('K-means 3D')
plt.show()
```

7.3.2 AGGLOMERATIVE CLUSTERING

Il secondo algoritmo di clustering utilizzato è il gerarchico agglomerativo. Inizialmente ci sono tanti cluster quanti sono i prodotti in ogni area di mercato. In seguito, si procede aggiungendo di volta in volta un legame tra due cluster per unirli utilizzando il Ward's Method secondo cui la distanza è la somma dei quadrati delle distanze tra i punti appartenenti a due cluster diversi. Il processo è ripetuto fino a che si otterrà un unico cluster contenente tutti i dati. Per rappresentare le iterazioni del clustering gerarchico è possibile utilizzare il dendrogramma che mostra come gli oggetti vengano raggruppati dopo ogni step. Tra i punti di forza dell'algoritmo gerarchico ricordiamo il fatto che non sia necessario definire a priori un numero di cluster come è stato necessario per il k-means. Inoltre, il metodo permette di generare tutte le combinazioni possibili, quindi, è sufficiente variare la soglia di clusterizzazione per modificare il numero di raggruppamenti senza dover eseguire nuovamente il processo. Rispetto al k-means però il costo computazionale è maggiore dovendo generare tutte le combinazioni possibili.

Il codice in Python utilizzato per applicare il metodo ai dati è:

```
from sklearn.cluster import AgglomerativeClustering
from scipy.cluster.hierarchy import linkage, dendrogram
```

```
from scipy.cluster import hierarchy

# matrice di associazioni (linkage), usando il Ward's linkage
link_matrix = linkage(X, method="ward")
# imposto il nome delle colonne per chiarire i valori che contengono
pd lnk mtx = pd.DataFrame(link matrix, columns=["cluster ID A", "cluster ID B", "dista
nza", "numero dati"])

# visualizzo il dendrogramma
hierarchy.dendrogram(link_matrix, truncate_mode = 'lastp', p=10)
# Plot title
plt.title('Hierarchical Clustering Dendrogram')
# Plot axis labels
plt.xlabel('sample index')
plt.ylabel('distance (Ward)')
# Show the graph
plt.show()

# cluster gerarchico
agglom_clustering = AgglomerativeClustering(n_clusters=3, affinity='euclidean', linkage=
"ward")
# in un unico passaggio addestro il modello ed elaboto la predizione
y_agglomerative = agglom_clustering.fit_predict(X)
# Assegna le etichette dei cluster a ogni oggetto del dataset
df['Cluster'] = agglom_clustering.fit_predict(X)

# visualizzo il risultato
plt.scatter(df[A],df[C], c=y_agglomerative, cmap='viridis')
# Plot title
plt.title('Hierarchical Clustering')
# Plot axis labels
plt.xlabel(A)
plt.ylabel(C)
# Show the graph
plt.show()

#per il valore silhouette
labels = agglom_clustering.labels_
print('Silhouette GERARCHICO:',metrics.silhouette_score(X, labels, metric='euclidean'))
```

In particolare, dal codice si può notare l'esigenza di importare apposite librerie per eseguire l'algoritmo gerarchico agglomerativo. Inoltre, si è deciso di rappresentare il dendrogramma di cui si ha un esempio in figura 7-19 per la Cina ma affinché sia leggibile è stato introdotto uno stop in corrispondenza di massimo dieci cluster differenti.

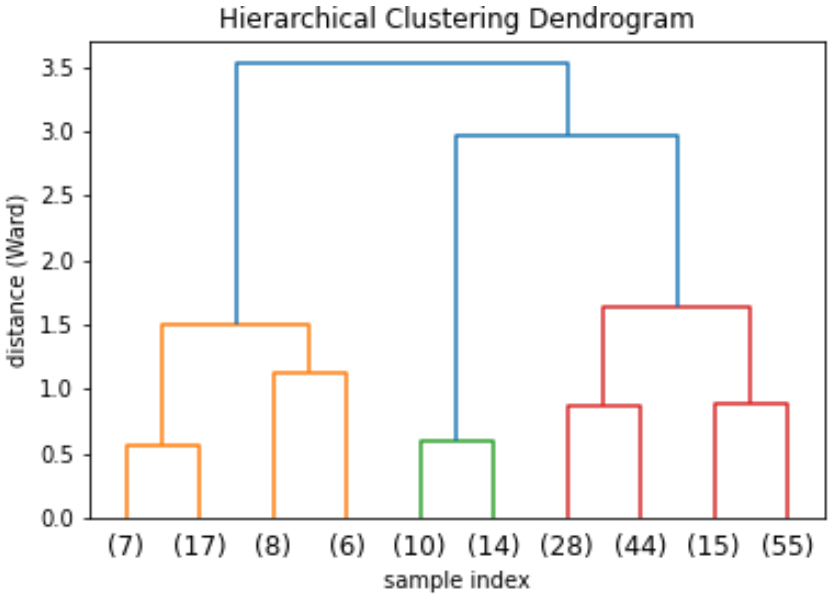


Figura 7-19. Esempio dendrogramma

In modo da rendere l'algoritmo confrontabile con il k-means si è deciso di considerare il numero di cluster pari a quello emerso durante la fase di elbow graph per ogni area di mercato. Successivamente si rappresenta il grafico dei cluster generati, di cui si ha un esempio in figura 7-20, e per ultimo si calcola il valore di silhouette corrispondente.

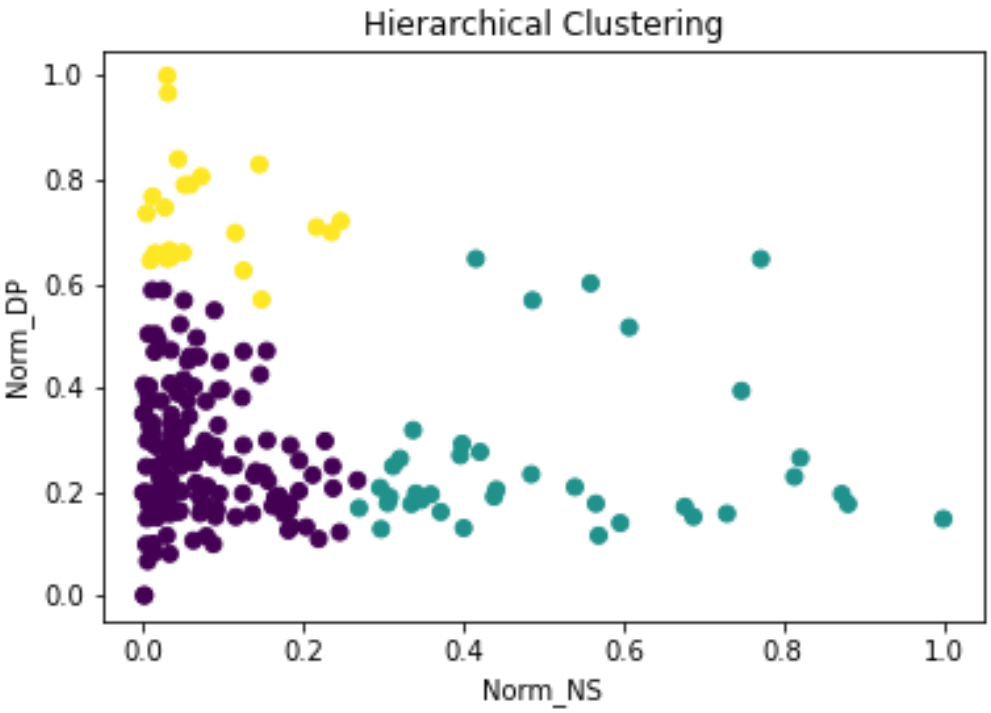


Figura 7-20. Gerarchico agglomerativo

7.3.3 DBSCAN

L'ultimo metodo utilizzato per analizzare i dati è il DBSCAN che definisce un cluster come l'insieme di punti contenuti entro una certa densità.

Affinché l'algoritmo funzioni correttamente è necessario determinare Eps e MinPts, rispettivamente la distanza massima tra due punti e il minimo numero di punti richiesti per formare un cluster. Per questo motivo, si decide di applicare l'algoritmo k – nearest neighbor per cui il valore k è definito come il numero di dimensioni del dataset più uno, in questo caso pari a 3 o 4 a seconda che si utilizzino 2 o 3 variabili rispettivamente. Per determinare Eps si si utilizza il grafico k- nearest neighbor che presenta sull'asse delle x i punti ordinati per distanza crescente e sull'asse delle ordinate, Eps, le distanze ordinate di ogni punto dal suo k°. Un esempio di kNN graph generato su Google Colab è in figura 7-21.

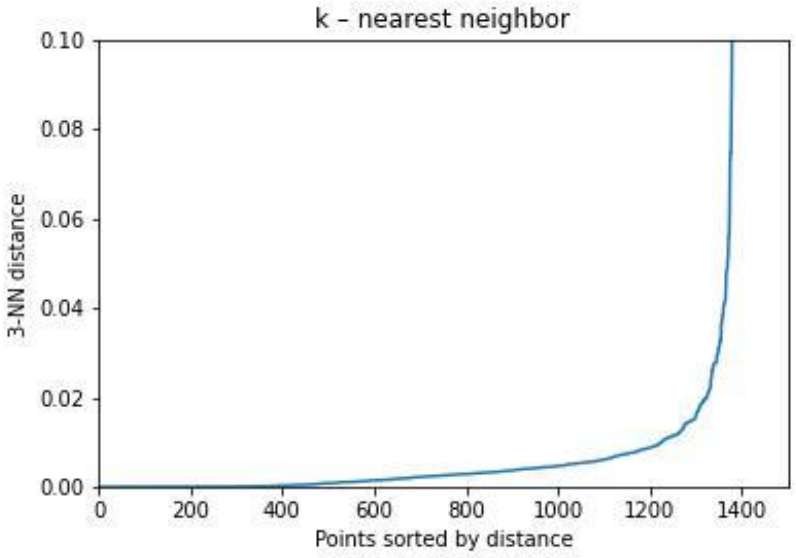


Figura 7-21. kNN

In corrispondenza della massima pendenza del grafico si può stabilire il valore di Eps da utilizzare per applicare il DBSCAN. In questo caso Eps è circa 0,015.

Come si può notare a primo impatto dal grafico in figura 7-22, il DBSCAN non è adatto al caso studio in quanto non genera cluster significativi infatti è noto che il DBSCAN non performi bene in presenza di dati con densità variabile.

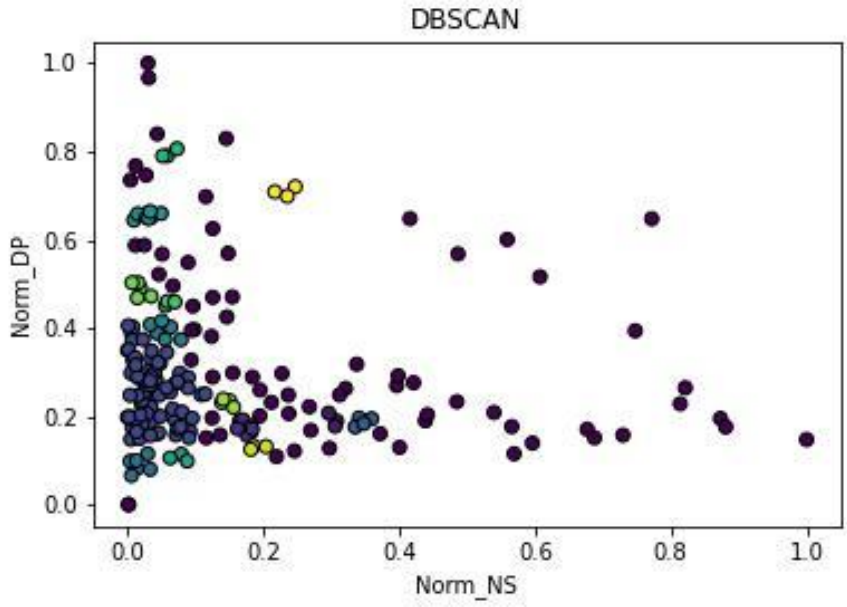


Figura 7-22. Cluster from DBSCAN

Il codice utilizzato per generare il KNN graph è:

```
from sklearn.neighbors import NearestNeighbors

nn = NearestNeighbors(n  neighbors=3).fit(X)
distances, indices = nn.kneighbors(X)
distances = np.sort(distances, axis=0)
distances = distances[:,1]
plt.plot(distances)
plt.axis([0, 220, 0, 0.1])
plt.xlabel("Points sorted by distance")
plt.ylabel("3-NN distance")
plt.title('k – nearest neighbor')
plt.savefig('k – nearest neighbor.jpeg')
plt.show()
```

Si sottolinea che il numero di vicini è stato modificato in base al numero di variabili in input.

Il codice utilizzato per generare i cluster con il DBSCAN, raffigurarli e calcolare il valore di silhouette è:

```
import matplotlib.pyplot as plt
from sklearn.cluster import DBSCAN
from sklearn.datasets import make_moons

# istanzio la classe di clustering DBSCAN
dbscan = DBSCAN(eps=0.015, min  samples=3)
# eseguo fitting e predizione in una volta sola
y_dbscan = dbscan.fit_predict(X)

# visualizzo il risultato
plt.scatter(df[A],df[C], c=y_dbscan, cmap='viridis', edgecolors="black")
plt.xlabel(A)
plt.ylabel(C)
plt.title('DBSCAN')
plt.savefig('DBSCAN CN.jpeg')
plt.show()

#per il valore silhouette
labels = dbscan.labels_
print ('Silhouette DBSCAN:',metrics.silhouette_score(X, labels, metric='euclidean'))
```

7.4 Pattern evaluation e data visualization

Per scegliere quale algoritmo e quali parametri di input fossero i migliori per ogni area di mercato è stata svolta una fase di validazione dei cluster in cui sono stati confrontati i valori di silhouette per le varie configurazioni analizzate. Le tabelle riportate di seguito, 7-3, 7-4 e 7-5, descrivono i risultati ottenuti per ogni configurazione.

Tabella 7-3. Silhouette Norm_NS, Norm_AC, Norm_DP

	K-MEANS	DBSCAN	GERARCHICO
AU	0,36	-0,11	0,32
CA	0,36	0,23	0,33
CN	0,4	0,17	0,37
EMEA	0,36	0,21	0,26
HK	0,5	0,15	0,42
JP	0,42	0,44	0,44
RU	0,43	0,14	0,39
SG	0,32	0,1	0,34
US	0,45	0,2	0,42
Media	0,41	0,15	0,38

Tabella 7-4. Silhouette Norm_NS, Norm_DP

	K-MEANS	DBSCAN	GERARCHICO
AU	0,62	0,14	0,61
CA	0,54	0,09	0,53
CN	0,53	0,14	0,51
EMEA	0,57	0,16	0,55
HK	0,59	0,18	0,48
JP	0,60	0,31	0,59
RU	0,59	0,08	0,59
SG	0,55	0,18	0,51
US	0,56	0,24	0,51
Media	0,57	0,16	0,55

Tabella 7-5. Silhouette Norm_NS, Norm_AC

	K-MEANS	DBSCAN	GERARCHICO
AU	0,46	0,26	0,42
CA	0,42	0,24	0,37
CN	0,44	0,51	0,42
EMEA	0,4	0,47	0,37
HK	0,42	0,57	0,39
JP	0,45	0,43	0,41
RU	0,49	0,3	0,49
SG	0,46	0,29	0,42
US	0,48	0,39	0,33
Media	0,45	0,36	0,41

A posteriori possiamo dire che per ogni configurazione di variabili in input il k-means è sempre l'algoritmo con valore di silhouette maggiore quindi si proseguono le analisi con questo metodo. In assoluto invece il set di variabili in input migliore per il caso studio è costituito dalle 2 variabili Net Sales e Discount Percentage.

Identificato l'algoritmo migliore e le metriche da mantenere si generano e si estraggono da Python i cluster per ogni area di mercato. Dopo l'applicazione della cluster analysis i prodotti sono stati assegnati a un preciso cluster. Nella tabella 7-6 si riportano il numero di prodotti, minimo, medio e massimo valore delle metriche per ogni cluster.

Tabella 7-6. Soglie clustering. Valori modificati per un valore moltiplicativo.

Area di mercato	B/W	Cluster	Numero prodotti nel cluster	Min di sum(Net Sales €)	Media di sum (Net Sales €)	Max di sum (Net Sales €)	Min di Average Discount Percentage	Media di Average Discount Percentage	Max di Average Discount Percentage
AU									
		0	127	355 €	1.486 €	4.295 €	0%	3%	16%
	WORST	1	27	224 €	1.662 €	4.574 €	25%	45%	79%
	BEST	2	16	4.680 €	7.679 €	19.788 €	0%	8%	18%
CA									
		0	41	398 €	924 €	1.722 €	0%	5%	22%
	WORST	1	5	300 €	456 €	573 €	69%	70%	70%
		2	13	413 €	1.277 €	2.370 €	31%	46%	63%
	BEST	3	16	1.824 €	2.663 €	4.918 €	0%	4%	12%
CN									
		0	207	367 €	12.251 €	44.883 €	0%	18%	33%
	WORST	1	66	998 €	10.791 €	58.016 €	35%	50%	78%
	BEST	2	47	46.837 €	78.921 €	138.964 €	9%	20%	50%
EMEA									
	WORST	0	33	258 €	1.356 €	6.199 €	14%	23%	40%
		1	118	11.693 €	20.975 €	35.191 €	2%	6%	14%
	BEST	2	25	38.021 €	54.098 €	77.251 €	3%	6%	8%
		4	171	302 €	3.703 €	12.292 €	0%	2%	12%
HK									
		0	215	381 €	5.020 €	18.180 €	0%	9%	27%
	WORST	1	55	570 €	4.704 €	12.436 €	28%	47%	78%
	BEST	2	44	20.860 €	34.854 €	58.214 €	3%	8%	24%
JP									
	BEST	0	56	48.543 €	81.037 €	144.635 €	0%	1%	1%
	WORST	1	191	461 €	14.838 €	47.283 €	0%	0%	11%
RU									
	WORST	0	197	347 €	4.617 €	12.690 €	0%	3%	38%
	BEST	1	69	13.518 €	21.877 €	41.192 €	0%	2%	12%
SG									
	BEST	0	154	452 €	2.578 €	16.508 €	0%	7%	31%
	WORST	1	33	282 €	1.339 €	4.888 €	36%	60%	78%
US									
		0	44	495 €	1.925 €	3.364 €	6%	13%	22%
		1	122	209 €	1.120 €	2.803 €	0%	0%	8%
	WORST	2	13	354 €	772 €	2.080 €	26%	35%	39%
	BEST	3	61	3.011 €	4.659 €	7.318 €	0%	4%	14%

Successivamente osservando i cluster graficamente e confrontandoli con i valori tabulari, per ogni area di mercato si è identificato il cluster migliore, BEST, e quello peggiore, WORST. Più precisamente si sono verificati tre casi diversi:

- In Australia, Canada, Cina, EMEA, Hong Kong e USA il cluster migliore è caratterizzato da un elevato valore di fatturato medio e una bassa percentuale di sconto;
- In Giappone e Russia l'analisi si focalizza solamente sul valore del fatturato. Il cluster migliore ha fatturato maggiore;
- A Singapore invece il cluster migliore è costituito dalla minor percentuale di sconto. Il fatturato non è stato determinante per suddividere i cluster ulteriormente.

Uno degli obiettivi della tesi è supportare il merchandising management nelle decisioni aziendali provando a far emergere dai dati informazioni importanti sulla clientela target. La segmentazione del mercato è una strategia che ogni azienda del settore può utilizzare per rafforzare il proprio vantaggio competitivo focalizzandosi sul gruppo di clienti più adatto a cui rivolgere le proprie campagne di marketing. Si decide, quindi, di analizzare i cluster separatamente per tutte le aree di mercato determinando le caratteristiche dei prodotti che li rappresentano come: stile, materiale, colore, prezzo medio, costo medio. Si è deciso di supportare il merchandising tramite report analitici e quindi di rappresentare direttamente alcuni risultati tramite un software di Business Intelligence.

In figura 7-23 è possibile visualizzare le informazioni ottenute per l'area geografica Cina in relazione all'analisi del cluster migliore. Innanzitutto, sono state fornite alcune informazioni legate al contesto come il nome dell'azienda (in questo caso inventato), la stagione di riferimento, l'area geografica, il totale dei ricavi e la percentuale media di sconto. L'obiettivo del report è mostrare facilmente e in modo intuitivo gli elementi più importanti per comprendere cosa sia andato bene nel corso della stagione. Ad esempio, è fornita la classifica dei primi cinque stili, colori e materiali che hanno generato più fatturato nel corso della stagione.

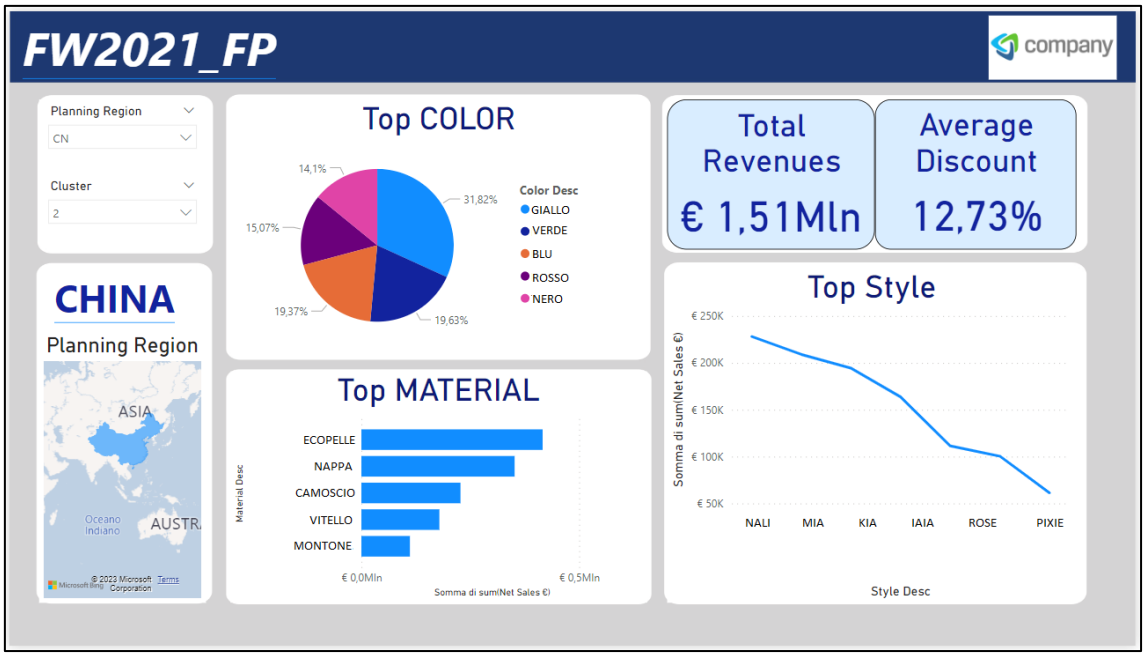


Figura 7-23. Dashboard caratteristiche prodotti più performanti FW2021

Successivamente, come mostrato in figura 7-24, per confronto sono state mostrate le informazioni legate al cluster peggiore inerenti alle stesse caratteristiche discusse per il cluster migliore.

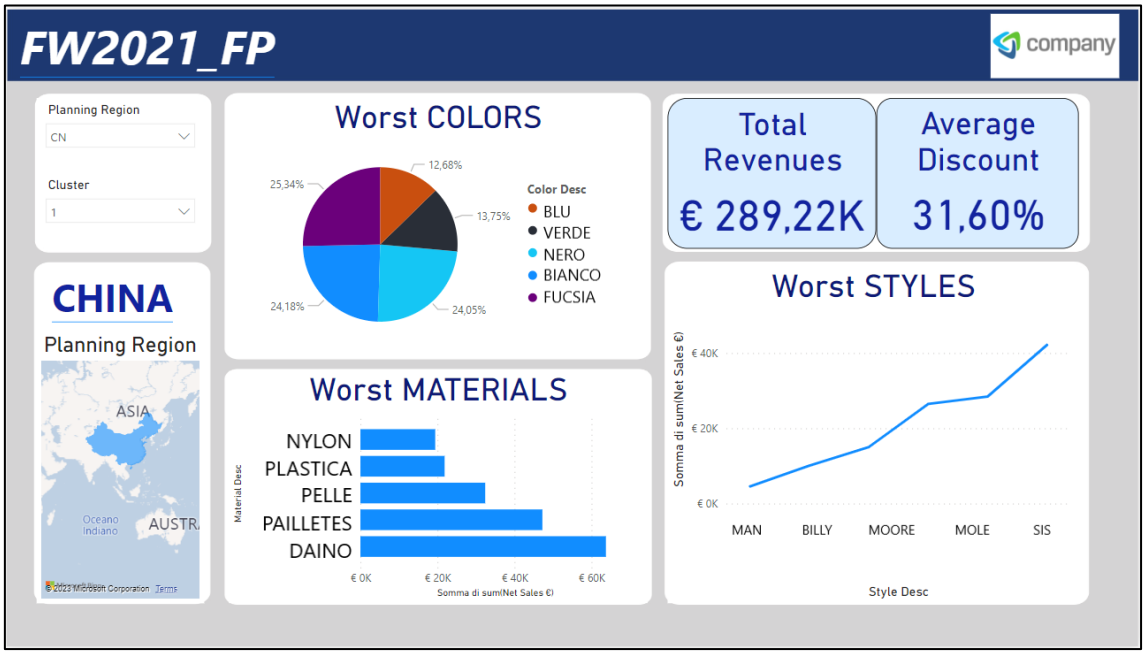


Figura 7-24. Dashboard caratteristiche prodotti meno performanti FW2021

La potenzialità dei report è di aggiornarsi periodicamente in modo automatico per fornire informazioni recenti agli utenti. Inoltre, un grande vantaggio risiede nella

dinamicità delle pagine in quanto uno stesso template può essere analizzato in base alle informazioni selzionate con una serie di filtri impostati come: Region, Cluster, Numero di Top o di Worst elementi da visualizzare.

8. Soglie: K-means vs analisi ABC

Un altro importante obiettivo di questo studio è cercare di rispondere all'esigenza di definire delle soglie sotto cui i prodotti non performano bene e quindi sarebbero da scartare e oltre il quale, invece, rappresentano un'importante risorsa per l'azienda e quindi conviene riproporli. La difficoltà di ciò risiede nel fatto che non sia facile prevedere come i prodotti nuovi possano essere colti dal mercato. Si è deciso di estrarre le potenziali soglie attraverso due metodi diversi: l'analisi ABC e il clustering. Una volta trovate le soglie si valida il modello testandole sui prodotti riproposti nel 2022.

8.1 K-MEANS

A valle del processo KDD il clustering è stato utile per identificare le caratteristiche dei prodotti più performanti e meno performanti. In questo capitolo il metodo viene snaturato e si cerca di estrapolare le soglie da utilizzare stagionalmente per selezionare i prodotti alto-vendenti e basso-vendenti. Si ricorda che i dati presentati sono stati manipolati per privacy.

In particolare, si è deciso di analizzare per ogni area di mercato il cluster migliore e il cluster peggiore da cui si sono estratte le seguenti soglie:

- Max NS W = massimo NetSales del cluster peggiore;
- Min NS B = minimo NetSales del cluster migliore;
- Max AD B = massimo sconto medio percentuale del cluster migliore;
- Min AD W = minimo sconto medio percentuale del cluster peggiore.

In tabella 8-1 sono riassunti e riportati i risultati ottenuti.

Tabella 8-1. Soglie cluster migliore e peggiore

Area di mercato	Max NS W	Min NS B	Max AD B	Min AD W
AU	4.627 €		22%	
CA	573 €	1.824 €	12%	31%
CN	52.427 €		31%	
EMEA	6.199 €	38.021 €	16%	
HK	12.436 €	20.860 €	26%	
JP	47.913 €		20%	
RU	13.104 €		20%	
SG	2.670 €		31%	
US	2.080 €	3.011 €	14%	26%

Osservando la tabella si può notare che alcune soglie siano univoche, calcolate come media delle due precedenti. Questa scelta è stata presa in due casi:

- La distanza tra la soglia di massimo e minimo dei due cluster era molto piccola (differenza percentuale tra le due soglie minore del 30%);
- Le due soglie si sovrapponevano perciò è stato necessario definire un valore medio che le distinguesse.

Inoltre, occorre sottolineare che Giappone e Russia possiedono valori della percentuale media di sconto così basse da essere prossime allo zero, per questo in accordo con l'esperto di dominio è stato deciso di porre la seguente soglia pari al 20%. Questa soglia solitamente in azienda designa i prodotti che la stagione successiva verrebbero categorizzati come 'prodotti vecchi' di cui cessare la produzione.

Per visualizzare meglio la situazione graficamente, in figura 8-1 è possibile osservare le soglie determinate per Hong Kong:

- due soglie distinte per il fatturato medio (asse x);
- una sola soglia per la percentuale di sconto (asse y).

In particolare, il cluster migliore è di colore giallo a cui corrisponde fatturato medio più alto e percentuale di sconto più bassa. Il cluster peggiore è di colore verde acqua ed è caratterizzato da basso fatturato medio e alta percentuale di sconto.

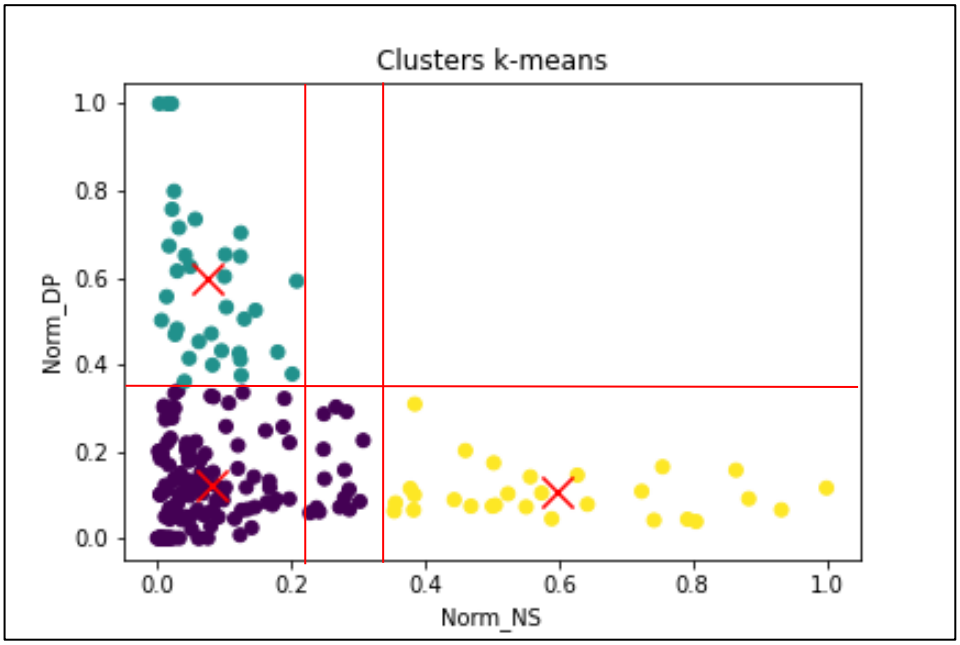


Figura 8-1. Esempio soglie applicate a Hong Kong.

Una volta definite le soglie che determineranno se un prodotto ha performato bene o male si passa a estrapolare quanti e quali prodotti siano effettivamente stati riproposti sul mercato full price durante la stagione FW successiva. Da questo punto di vista è necessario fare un piccolo excursus. In prima battuta, in figura 8-2, sembra che il 72% dei prodotti offerti sul mercato nel 2021 venga riproposto anche nel 2022.

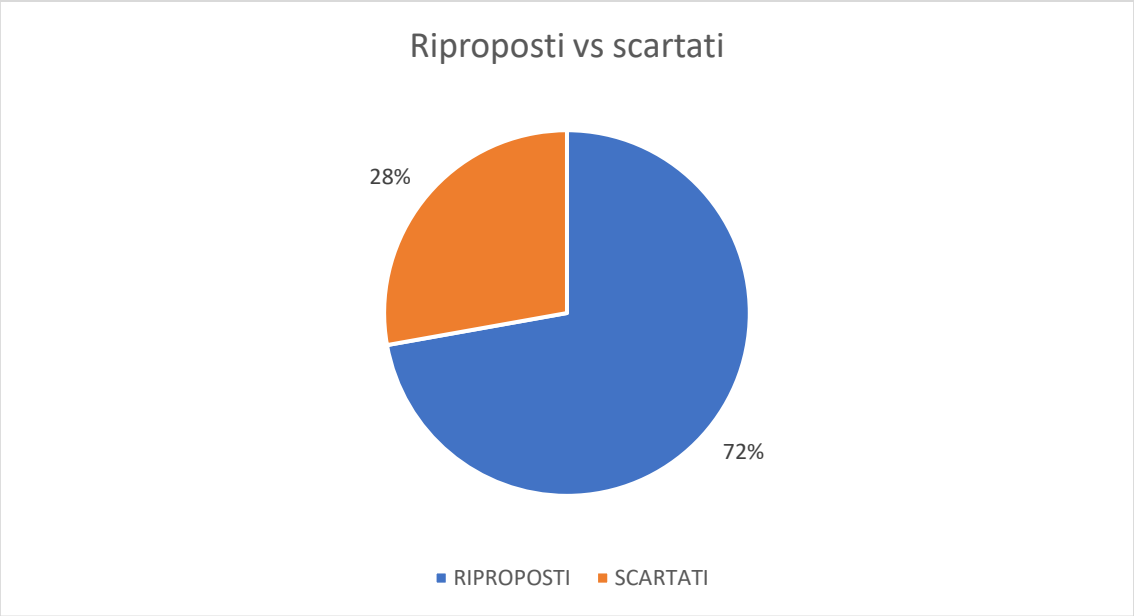
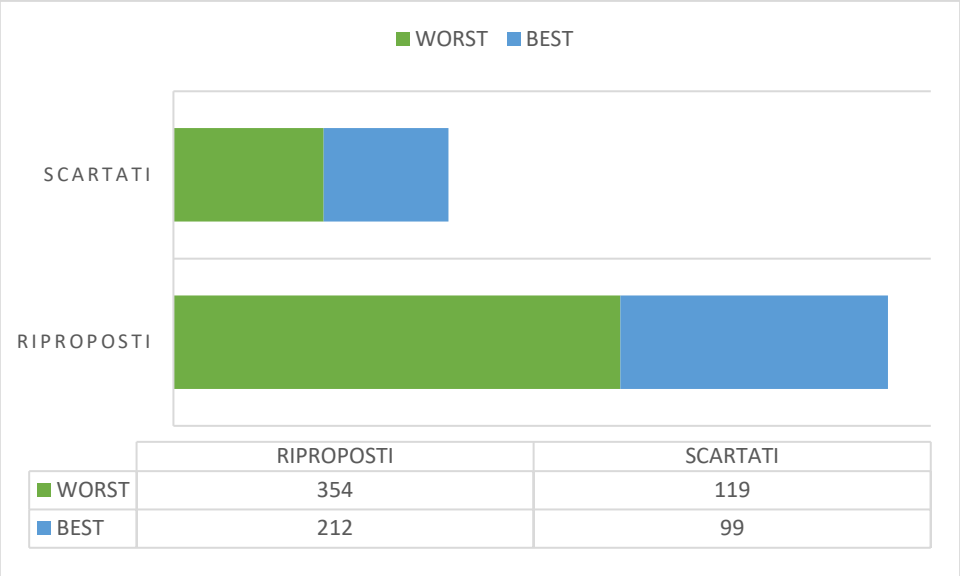


Figura 8-2. Grafico prodotti riproposti vs scartati

Approfondendo l'analisi dei prodotti riproposti e scarti emerge che il 54% dei prodotti scartati faceva parte del cluster peggiore mentre il 46% apparteneva al cluster migliore, contrariamente a quanto si sarebbe potuto immaginare. Non potendo però continuare l'analisi su questo gruppo di prodotti perché assenti nel 2022, si decide di proseguire le analisi solo sui prodotti riproposti. Per questa categoria di articoli si nota che il 63% apparteneva nel 2021 al cluster peggiore mentre il 37% al cluster migliore. Ciò è ben sintetizzato nel grafico in figura 8-3.



	RIPROPOSTI	SCARTATI
WORST	354	119
BEST	212	99

Figura 8-3. Cluster prodotti riproposti vs scartati

Contrariamente a quello che ci sarebbe potuti aspettare la maggior parte dei prodotti riproposti fanno parte della categoria WORST nel 2021. Questo è possibile spiegarlo poiché in realtà l'azienda ha categorizzato gli articoli nel 2022 come segue:

- 79% OLD PRODUCT;
- 21% NOT OLD PRODUCT (comprende CARRY OVER o STILL VALID).

Inoltre, è importante sottolineare che a livello aziendale un prodotto è definito OLD PRODUCT indipendentemente dall'area di mercato ma questo potrebbe influenzare le analisi poiché, come viene spiegato dall'esperto aziendale, la suddetta categoria comprende i prodotti che non sono realmente riproposti alla clientela ma che l'azienda ha deciso di vendere per esaurire le scorte. Per questo motivo si decide di focalizzare le analisi solo sugli articoli realmente ripresentati, cioè quelli che sono compresi nella categoria creata appositamente: NOT OLD PRODUCT. Focalizzandosi su questi emerge che nel 2022 il 67% dei prodotti riproposti apparteneva al cluster migliore e, invece, un terzo dei prodotti era stato associato al cluster peggiore come rappresentato in figura 8-4.

Figura 8-4. Cluster prodotti riproposti

Tornando alle analisi, poiché nel 2022 il fatturato dell'azienda è rimasto circa costante rispetto a quello precedente, si è deciso di non far variare le soglie e di applicare le stesse anche agli articoli della FW22. Come si può notare in figura 8-5 emerge che:

- Il 39% dei prodotti Best nel 2021 continua a posizionarsi nella stessa categoria anche nel 2022 (BEST_BEST);
- Il 57% dei prodotti nel cluster Best nel 2021 subisce un leggero declassamento e finisce nel cluster intermedio (BEST_MIDDLE);
- Il rimanente 4% rappresenta i prodotti che hanno peggiorato le loro performance tanto da entrare a far parte del cluster peggiore (BEST_WORST.)

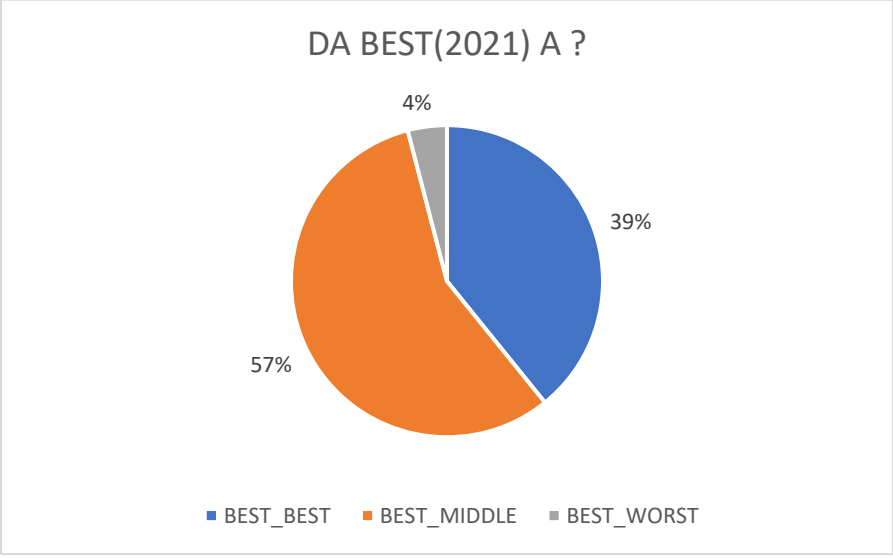


Figura 8-5. Cluster a cui apparterrebbero i prodotti riproposti nel 2022

Del gruppo BEST_WORST fanno parte i prodotti che hanno subito una perdita di fatturato maggiore del 90% ed appartengono alle seguenti aree di mercato: Cina, Stati Uniti e Giappone. Sembra che le soglie siano in grado di determinare i prodotti che hanno ancora elevate potenzialità la stagione successiva.

Per quanto riguarda invece i prodotti che si posizionavano nella categoria Worst nel 2021, la situazione nel 2022 è la seguente (figura 8-6):

- L'86% dei prodotti che facevano parte del cluster peggiore anche nel 2022 continuano a farne parte (WORST_WORST);
- Il 6% dei prodotti si posiziona in una fascia intermedia (WORST_MIDDLE);
- L'8% dei prodotti entra a far parte della categoria più performante (WORST_BEST).

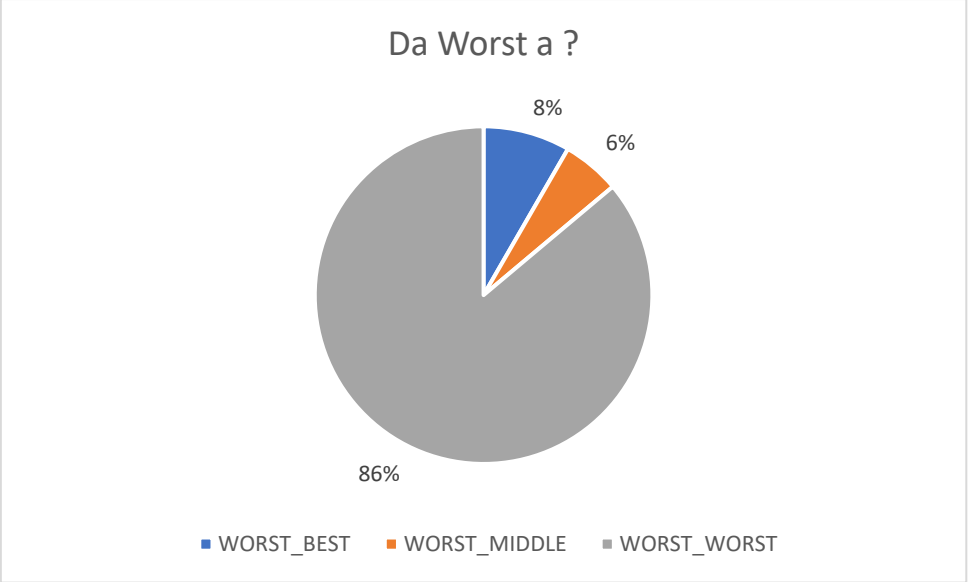


Figura 8-6. Cluster a cui apparterrebbero i prodotti riproposti nel 2022

Solo l'8% dei prodotti stravolge la categoria rispetto alla stagione precedente, ma analizzandoli più attentamente emerge che siano tutti articoli venduti in Russia. È noto, infatti, che i magnati russi apprezzino molto i prodotti italiani, in particolare per quanto riguarda gli articoli di abbigliamento, i mobili e la fabbricazione di altre macchine. Mosca rappresenta circa il 2% delle esportazioni italiane per un equivalente di 7,7 miliardi [45]. Inoltre, il cambiamento degli equilibri geopolitici e l'inizio della guerra potrebbero aver influito su questo evento. Sembra interessante il fatto che addirittura più

dell'85% dei prodotti continua a confermare la sua posizione nel cluster peggiore e quindi sarebbe stato meglio non riproporli.

In conclusione, l'andamento generale per i prodotti che nel 2021 facevano parte della categoria best è rappresentato dal fatto che nel 2022 poco meno della metà continua ad essere nella stessa categoria mentre l'altra metà entra a far parte della categoria intermedia, peggiorando le proprie performance. Raramente i risultati dei prodotti la stagione successiva vengono stravolti rispetto alla precedente entrando a far parte della categoria opposta. Per quanto riguarda invece i prodotti che facevano parte della categoria worst nel 2021 addirittura l'86% di questi continua a posizionarsi nella stessa categoria confermando le ipotesi fatte. Solo una piccola parte migliora il proprio apprezzamento sul mercato ed in particolare nel caso in esame è giustificabile da fenomeni esterni.

Si potrebbe concludere che, in base ai dati analizzati, la definizione delle soglie dal cluster worst come determinante dei prodotti da scartare possa essere un metodo accurato da applicare a supporto decisionale del merchandising team.

8.2 ANALISI ABC

In parallelo al data mining, per ogni area di mercato, si è svolta l'analisi ABC sui prodotti con l'obiettivo di identificare anche con questa metodologia quelli più e meno performanti della FW2021. Inoltre, si vuole indagare se possa essere un buon metodo per determinare le soglie di fatturato che definiscano se un prodotto debba essere scartato o meno la stagione successiva.

Ordinando i prodotti in base al fatturato decrescente si attribuiscono ad una categoria in modo che la cumulata dei prodotti che generano maggior fatturato (categoria A) produca l'80% dei ricavi totali, la successiva categoria (B) produca il successivo 15% dall'80% al 95%, ed infine i prodotti nella categoria C generino il rimanente 5%.

La tabella 8-2 riepiloga i valori di minimo, massimo e media di ogni categoria, per ogni region (valori manipolati per un fattore moltiplicativo).

Tabella 8-2. Soglie analisi ABC

Area geografica	Categoria	MIN	MEDIA	MAX
AUSTRALIA	A	449 €	1.088 €	6.314 €
	B	202 €	332 €	448 €
	C	71 €	150 €	200 €
CANADA	A	306 €	568 €	1.569 €
	B	184 €	246 €	303 €
	C	96 €	145 €	183 €
CINA	A	6.546 €	16.570 €	44.341 €
	B	1.663 €	3.468 €	6.347 €
	C	117 €	900 €	1.662 €
EMEA	A	4.502 €	9.089 €	24.649 €
	B	1.213 €	2.663 €	4.444 €
	C	53 €	461 €	1.192 €
HONG KONG	A	2.407 €	6.697 €	18.575 €
	B	699 €	1.380 €	2.383 €
	C	121 €	398 €	695 €
GIAPPONE	A	8.589 €	19.821 €	46.150 €
	B	3.627 €	5.590 €	8.545 €
	C	147 €	1.253 €	3.530 €
RUSSIA	A	2.652 €	5.656 €	13.144 €
	B	986 €	1.725 €	2.593 €
	C	111 €	398 €	919 €
SINGAPORE	A	530 €	1.284 €	5.267 €
	B	227 €	364 €	529 €
	C	90 €	163 €	226 €
STATI UNITI	A	528 €	1.136 €	2.335 €
	B	195 €	342 €	524 €
	C	67 €	151 €	192 €

Si decide di focalizzare le analisi solo sui prodotti che fanno parte della categoria A, idealmente la migliore, e di quelli che fanno parte della categoria C, la peggiore, così

come è stato fatto per il clustering. Nonostante si ritenga che il gruppo intermedio possa essere ricco di informazioni e possa essere un interessante spunto per sviluppi futuri.

I prodotti che sono presenti sia nel 2021 che nel 2022 si suddividono nel 2021 tra le categorie A, B, C come si nota in figura 8-7.

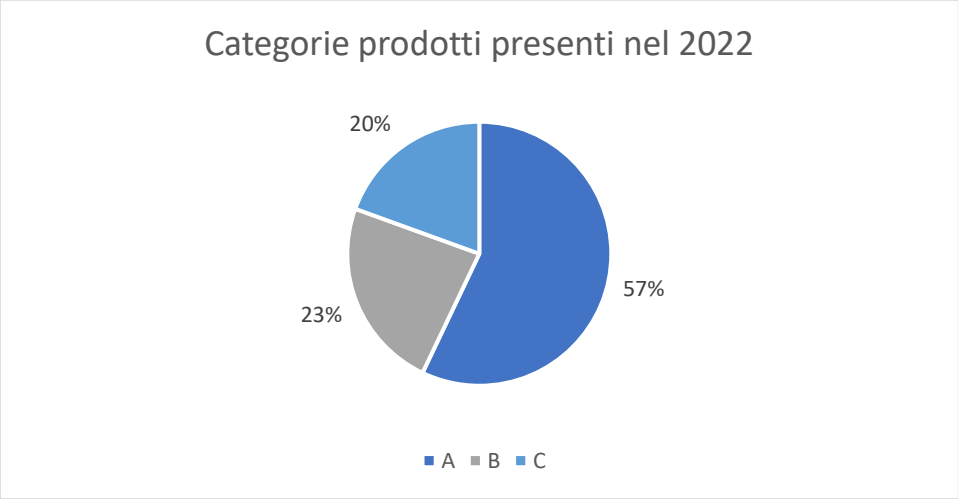


Figura 8-7. Suddivisione prodotti riproposti tra le categorie

Si decide anche in questo contesto di proseguire le analisi solo con i prodotti realmente riproposti, ovvero quelli che appartengono alla categoria NOT OLD PRODUCT, la cui motivazione è stata descritta all'inizio del paragrafo precedente. Ed anche in questo caso le soglie non vengono modificate rispetto al 2021 dato che il fatturato complessivo aziendale è rimasto circa costane. Le soglie per ogni area di mercato sono riepilogate in tabella 8-3.

Tabella 8-3. Soglie categorie

Area di mercato	Max NS C	Min NS A
AU	200 €	449 €
CA	183 €	306 €
CN	1.662 €	6.546 €
EMEA	1.192 €	4.502 €
HK	695 €	2.407 €
JP	3.530 €	8.589 €
RU	919 €	2.652 €
SG	226 €	530 €
US	192 €	528 €

Si inizia analizzando in che categoria rientrano nel 2022 i prodotti che nel 2021 facevano parte della categoria A. In figura 8-8 emerge che:

- Il 63% dei prodotti A nel 2021 continua ad essere nella stessa categoria anche nel 2022 (A_A);
- Il 28% dei prodotti nella categoria A nel 2021 diminuisce leggermente le performance e finisce nella categoria intermedia (A_B);
- Il rimanente 9% rappresenta i prodotti che hanno peggiorato le loro performance tanto da entrare a far parte dell'ultima fascia di prodotti (A_C).

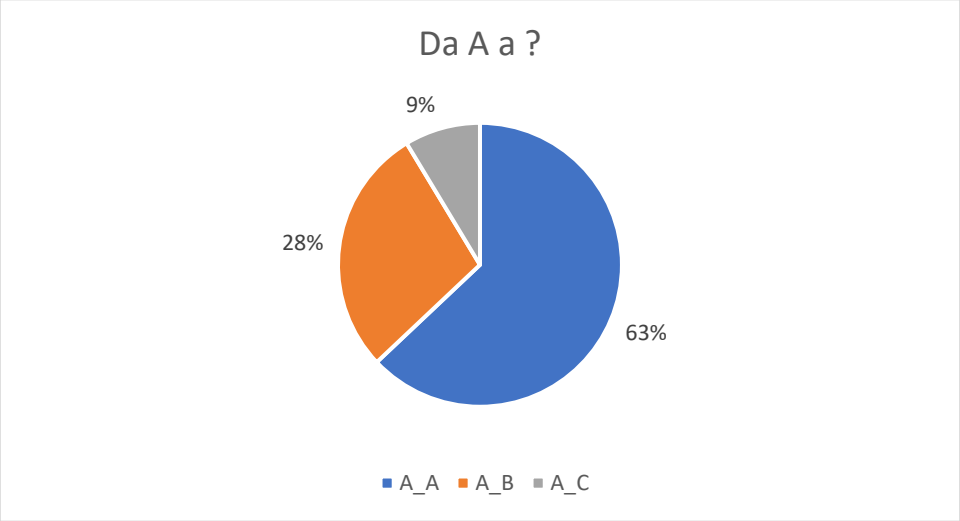


Figura 8-8. Catgoria in cui si posizionerebbero nel 2022 i prodotti riproposti

Per quanto riguarda invece i prodotti che si posizionavano nella categoria C nel 2021, la situazione nel 2022 è la seguente (vedi figura 8-9):

- Il 36% dei prodotti che facevano parte della categoria peggiore anche nel 2022 continuano a farne parte (C_C);
- Il 64% dei prodotti si posiziona in una fascia intermedia (C_B);
- Nessun prodotto entra a far parte della categoria migliore.

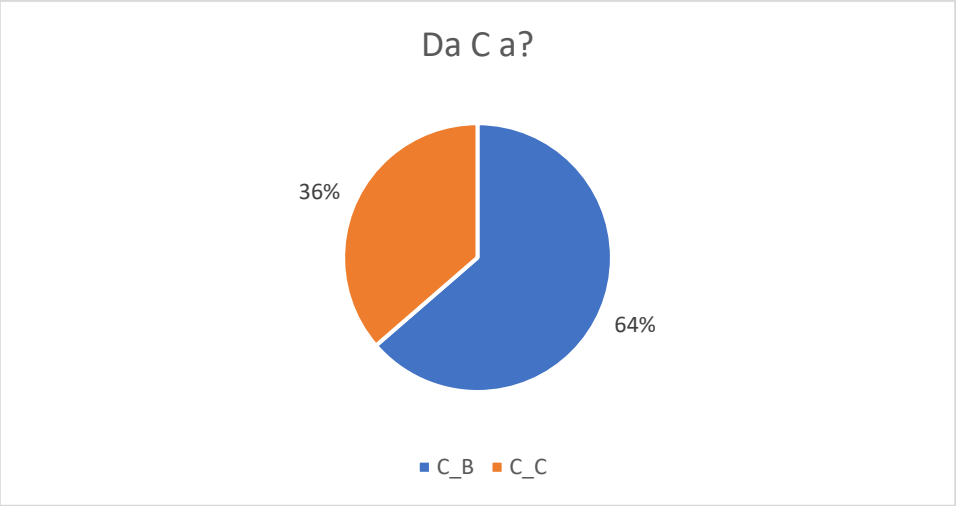


Figura 8-9. Categoria in cui si posizionerebbero nel 2022 i prodotti riproposti

In conclusione, le soglie che delimitano il gruppo di prodotti migliore ottiene un buon riscontro in quanto conferma lo stato dei prodotti per due terzi circa dei casi. Un po' meno di un terzo dei prodotti peggiora leggermente le proprie performance entrando a far parte della categoria intermedia mentre il 10% circa dei prodotti peggiora le proprie performance declassando il proprio livello al più basso.

Per quanto riguarda invece, i prodotti della classe peggiore continuano a far parte della stessa categoria nel 36% dei casi mentre migliorano leggermente la propria categoria nel 64% dei casi.

8.3 Confronto Analisi ABC e K-Means

Focalizzandosi sulle soglie di fatturato dei due metodi si nota che, in generale, la soglia sotto il quale i prodotti vengono considerati da scartare è sempre più bassa per l'analisi ABC rispetto a quella ottenuta con il clustering. Questo determina quindi un intervallo più ampio con le soglie del clustering e quindi giustifica il fatto che si ottenga un risultato più soddisfacente nel confronto tra 2021 e 2022. L'86% dei prodotti categorizzati come worst continua ad essere nella stessa categoria.

Si potrebbe concludere che i due metodi rispecchino due diverse filosofie di pensiero:

- Una meno restrittiva rispetto ai prodotti da scartare (analisi ABC);
- Una più restrittiva che esclude un numero maggiore di prodotti (CLUSTERING).

E viceversa, il range di soglie che determina i prodotti più performanti è maggiore nell'analisi ABC rispetto al clustering. Anche questo confermerebbe due diverse strategie aziendali:

- Strategia più permissiva nel riproporre i prodotti sul mercato (analisi ABC);
- Strategia meno permissiva nel riproporre i prodotti (CLUSTERING).

In questo secondo caso, poiché le soglie sono più stringenti per i prodotti più performanti è più facile che questi peggiorino il proprio cluster da una stagione alla successiva. Questo si traduce in una minor percentuale di prodotti che continua a rimanere nel cluster BEST, solo il 39%. Mentre l'analisi ABC sembra fornire risultati più soddisfacenti in quanto il 63% dei prodotti nella categoria A continua a popolarla la stagione successiva.

Alcuni esempi visivi del posizionamento delle soglie è possibile visualizzarli in figura 8-10 e 8-11 per l'Australia, in figura 8-12 e 8-13 per il Canada e in figura 8-14 e 8-15 per EMEA.

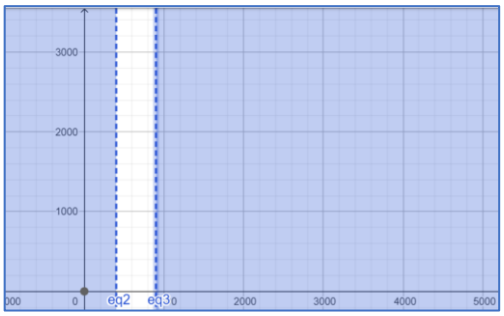


Figura 8-10. Soglie analisi ABC Australia

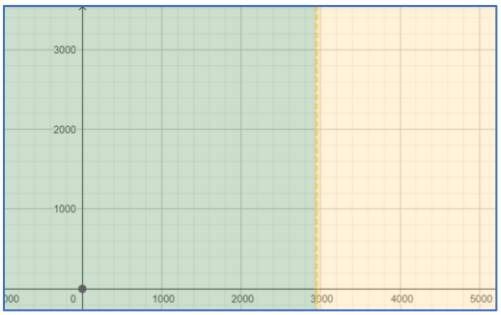


Figura 8-11. Soglie clustering Australia

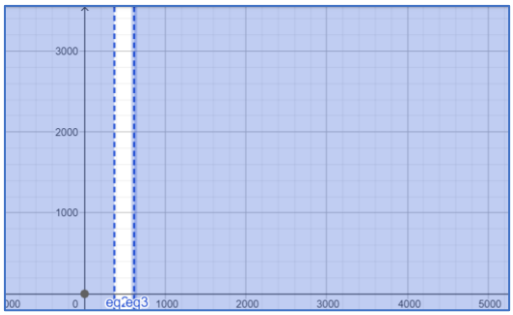


Figura 8-12. Soglie analisi ABC Canada

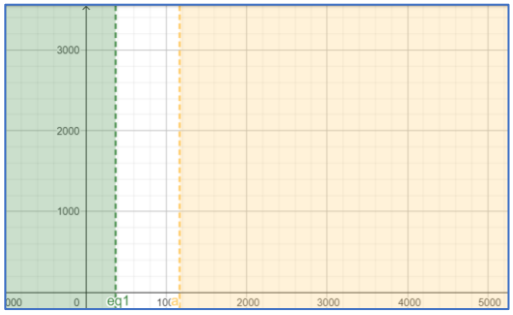


Figura 8-13. Soglie clustering Canada

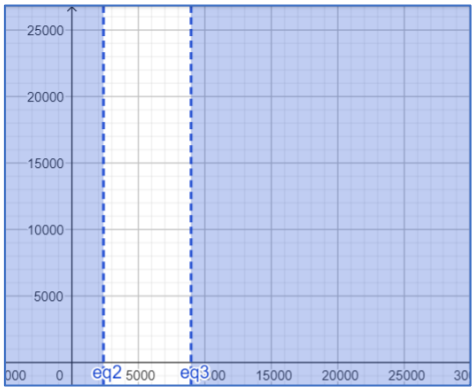


Figura 8-14. Soglie analisi ABC EMEA

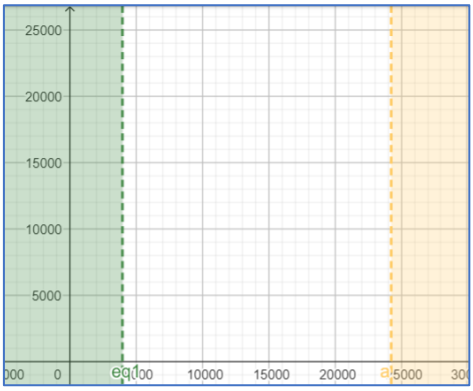


Figura 8-15. Soglie clustering EMEA

Ciò che è emerso in precedenza è confermato dal fatto che analizzando il rapporto tra l'ampiezza del range della categoria più performante e quello della categoria meno performante si nota come per l'analisi ABC sia di molto superiore (vedi tabella).

Tabella 8-4. Rapporto tra l'ampiezza del gruppo migliore e del gruppo peggiore.

Area di mercato	cluster	ABC
AU	3,5	45,7
CA	11,3	14,5
CN	1,6	24,5
EMEA	6,6	17,7
HK	3,1	28,2
JP	2,1	11,1
RU	2,2	13,0
SG	3,5	34,9
US	2,5	14,5

9. Conclusioni

L'obbiettivo della tesi è di offrire supporto decisionale al merchandising di un'azienda di moda operante nel mercato internazionale. Operativamente il supporto è stato fornito per soddisfare due esigenze:

- definizione di soglie utilizzabili per determinare i prodotti da scartare;
- visualizzazione tramite dashboard dei prodotti più e meno apprezzati in ogni area geografica per quanto riguarda stile, colori, materiali, fatturato totale e sconto medio.

L'approccio che è stato utilizzato segue la prospettiva data-driven in quanto tutte le informazioni ottenute sono emerse dai dati raccolti. Il dataset è stato analizzato sfruttando due tecniche differenti: analisi ABC e KDD (Knowledge Discovery in Databases). L'analisi ABC è un metodo univariato, più facile da applicare in un contesto aziendale, che ha permesso di suddividere i prodotti in base al fatturato ed ha fornito risultati interessanti nella fase di validazione delle soglie in quanto più del 60% dei prodotti di categoria A continua a rimanere nella fascia dei prodotti più performanti anche l'anno successivo. È un metodo che per definizione stabilisce soglie meno severe rispetto al KDD e che quindi permetterebbe a più prodotti di essere riproposti da un anno all'altro. Il KDD invece è un procedimento multivariato che permette di considerare contemporaneamente più variabili. In questo studio le variabili considerate sono state due: fatturato e percentuale media di sconto. Il KDD ha portato alla definizione di soglie più rigide rispetto all'analisi ABC che avrebbe permesso ad un minor numero di prodotti di essere riproposti la stagione successiva. Durante la validazione delle soglie con le performance degli stessi prodotti nel 2022 è emerso che il KDD sia molto affidabile per definire i prodotti da scartare, in quanto l'86% dei prodotti che hanno ottenuto bassi risultati nel 2021 non variano la loro condizione nel 2022 e quindi sarebbero potuti esser scartati.

In secondo luogo, sono state create delle dashboard per permettere al merchandising di osservare in modo efficace i risultati del KDD. Ogni schermata è interattiva e permette all'utente di selezionare un'area di mercato e un cluster per mostrare i modelli, i colori e i materiali più apprezzati e meno apprezzati dalla clientela. Ad esempio, è emerso che il

colore nero, tipico della stagione autunno inverno, è il più apprezzato in otto aree di mercato su nove, esclusa la Cina.

È stato deciso di non analizzare i prodotti che si sono posizionati in una fascia intermedia sia per l'analisi ABC che per il clustering al fine di concentrare l'attenzione sulle caratteristiche più rilevanti sia in negativo, prodotti meno apprezzati, che in positivo, prodotti più performanti. Tutti gli articoli che si posizionano in una fascia intermedia possono contenere però informazioni importanti per una migliore comprensione del mercato e costituiscono un interessante sviluppo futuro.

Inoltre, sarebbe curioso provare a ripetere le analisi per altre stagioni al fine di confermare o meno i risultati trovati con i due metodi di analisi.

Un altro possibile sviluppo consisterebbe nel cercare di integrare nel modello del KDD altre variabili come il numero di negozi fisici presenti in ogni area di mercato, il fatto che un prodotto sia stato pubblicizzato o il fatto che un prodotto sia stato esposto in vetrina in modo da rendere più accurate le analisi.

Bibliografia e sitografia

[1] "IL MERCHANDISING MANAGEMENT PER LE AZIENDE DEL SETTORE MODA," 2020. https://www.professionaldatagest.it/news/il-merchandising-management-per-le-aziende-del-settore-moda/ (accessed Sep. 27, 2022).

[2] "Merchandiser: chi è e cosa fa," 2021. https://muwo.it/in-store-out-store-promotion/merchandiser-chi-e-e-cosa-fa/ (accessed Sep. 27, 2022).

[3] "Cos'è il merchandising?" https://www.oracle.com/it/industries/retail/merchandising/what-is-merchandising/ (accessed Sep. 27, 2022).

[4] "Corso di alta formazione per Merchandising Manager." https://www.accademiadellavoro.it/corsi-nel-settore-aeronautico/corso-di-alta-formazione-per-merchandising-manager/#:~:text=Il%20Merchandiser%20Manager%20%C3%A8%20un,vendita%20di%20sua%20competenza%20e (accessed Oct. 10, 2022).

[5] Danea, "Analisi ABC." https://www.danea.it/blog/analisi-abc/ (accessed Oct. 08, 2022).

[6] "La regola di Pareto." https://www.pensierocritico.eu/principio-di-pareto.html (accessed Oct. 08, 2022).

[7] "Principio di Pareto." https://it.wikipedia.org/wiki/Principio_di_Pareto (accessed Oct. 08, 2022).

[8] "Analisi ABC." https://www.lokad.com/it/definizione-analisi-abc-(gestione-di-inventario) (accessed Oct. 08, 2022).

[9] "Il Principio di Pareto: cos'è, vantaggi e svantaggi della Legge 80/20." https://www.businesscoachingitalia.com/il-principio-di-pareto-cose-vantaggi-e-svantaggi-della-legge-80-20/ (accessed Oct. 10, 2022).

[10] A. K. Kushwaha, A. K. Kar, and Y. K. Dwivedi, "Applications of big data in emerging management disciplines: A literature review using text mining,"

International Journal of Information Management Data Insights, vol. 1, no. 2, p. 100017, Nov. 2021, doi: 10.1016/J.JJIMEI.2021.100017.

[11] Damiano Milanato, *Demand Planning. Processi, metodologie e modelli matematici per la gestione della domanda commerciale*. Milano, 2008. doi: https://doi-org.ezproxy.biblio.polito.it/10.1007/978-88-470-0822-9.

[12] "Cos'è la business intelligence?", Accessed: Dec. 14, 2022. [Online]. Available: https://powerbi.microsoft.com/it-it/what-is-business-intelligence/

[13] "Cos'è la business intelligence? La tua guida alla BI e al perché è importante." https://www.tableau.com/it-it/learn/articles/business-intelligence (accessed Dec. 14, 2022).

[14] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Knowledge Discovery and Data Mining: Towards a Unifying Framework", Accessed: Dec. 14, 2022. [Online]. Available: https://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf?utm_campaign=ml4devs-newsletter&utm_medium=email&utm_source=Revue%20newsletter

[15] H. Jiawei, K. Micheline, and P. Jian, *DATA MINING Concepts and Techniques*, III. Morgan Kaufmann, 2011.

[16] "KDD Process in Data Mining", Accessed: Feb. 15, 2023. [Online]. Available: https://www.geeksforgeeks.org/kdd-process-in-data-mining/

[17] "Data mining: cos'è." https://www.bigdata4innovation.it/data-science/data-mining/data-mining-cose-perche-conviene-utilizzarlo-e-quali-sono-le-attivita-tipiche/ (accessed Nov. 10, 2022).

[18] "Cos'è un pattern?" https://datacadamia.com/data_mining/pattern (accessed Nov. 10, 2022).

[19] "Metodologie di analisi: apprendimento supervisionato." https://www.tesionline.it/appunti/economia/laboratorio-informatico-per-le-decisioni-aziendali/metodologie-di-analisi-apprendimento-supervisionato/223/40 (accessed Nov. 10, 2022).

[20] E. Baralis and T. Cerquitelli, “Clustering fundamentals.” Accessed: Dec. 15, 2022. [Online]. Available: https://dbdmg.polito.it/dbdmg_web/wp-content/uploads/2022/03/clustering.pdf

[21] “Centroid Initialization Methods for k-means Clustering.” https://www.kdnuggets.com/2020/06/centroid-initialization-k-means-clustering.html (accessed Dec. 15, 2022).

[22] C. Casadei, “Clustering gerarchico.” https://www.developersmaggioli.it/blog/clustering-gerarchico/ (accessed Dec. 16, 2022).

[23] S. Orlando, “Clustering.” Accessed: Dec. 16, 2022. [Online]. Available: https://www.dsi.unive.it/~dm/Slides/2_Cluster.pdf

[24] “Dbscan.”

[25] A. Ram, S. Jalal, A. S. Jalal, and M. Kumar, “A Density Based Algorithm for Discovering Density Varied Clusters in Large Spatial Databases,” *Int J Comput Appl*, vol. 3, no. 6, pp. 1–4, Jun. 2010, doi: 10.5120/739-1038.

[26] Unife, “Knowledge Discovery in Databases.” Accessed: Oct. 24, 2022. [Online]. Available: https://www.unife.it/ing/lm.infoauto/sistemi-informativi/allegati/25-knowledge_discovery_in_databases.pdf

[27] A. Jain and R. Dubes, *Algorithms for Clustering Data*. Accessed: Dec. 16, 2022. [Online]. Available: https://homepages.inf.ed.ac.uk/rbf/BOOKS/JAIN/Clustering_Jain_Dubes.pdf

[28] “CLUSTER VALIDATION ESSENTIALS.” https://www.datanovia.com/en/lessons/cluster-validation-statistics-must-know-methods/#silhouette-coefficient (accessed Jan. 15, 2023).

[29] “Pattern Evaluation Methods in Data Mining”, Accessed: Nov. 18, 2022. [Online]. Available: https://www.geeksforgeeks.org/pattern-evaluation-methods-in-data-mining/

[30] “Introduzione - Training Microsoft.” https://learn.microsoft.com/it-it/training/browse/?terms=introduzione&products=power-bi (accessed Dec. 10, 2022).

[31] “Analisi di mercato,” 2022. https://www.webmarketingaziendale.it/come-si-fa-analisi-di-mercato/ (accessed Oct. 03, 2022).

[32] “Definizione di ROS.” https://www.insidemarketing.it/glossario/definizione/ros/ (accessed Nov. 11, 2022).

[33] “GLOSSARIO FINANZIARIO - RAPPORTO DI INDEBITAMENTO.” https://www.borsaitaliana.it/borsa/glossario/rapporto-di-indebitamento.html (accessed Nov. 22, 2022).

[34] “Posizione finanziaria netta: cos’è e perché è importante.” https://blog.docfinance.net/posizione-finanziaria-netta-cosa-e-e-perche-tenerne-conto (accessed Nov. 11, 2022).

[35] Valentina Lazzarotti, Rafaela Gjergji, and Federico Visconti, “Scelte strategiche e performance: un modello di analisi per le imprese del settore.” Accessed: Nov. 24, 2022. [Online]. Available: https://www.amaplast.org/archivioFiles/Allegati/Report%20Amaplast%20FINALE.pdf

[36] “IL MERCATO DELLA PELLETTERIA DI LUSSO | ITALIA,” *Businesscoot*. https://www.businesscoot.com/it/studio-di-mercato/il-mercato-della-pelletteria-di-lusso-italia#:~:text=Il%20mercato%20mondiale%20della%20pelle,il%202022%20e%20il%202027. (accessed Oct. 16, 2022).

[37] “Luxury Leather Goods - Worldwide,” *Statista*. https://www.statista.com/outlook/cmo/luxury-goods/luxury-leather-goods/worldwide (accessed Oct. 16, 2022).

[38] “Rapporto sulla pelletteria,” *Assomac*, Jul. 20, 2022. https://assomac.it/it/news/news-dal-mondo/rapporto-sulla-pelletteria/ (accessed Oct. 16, 2022).

[39] “Beni di lusso, un 2022 da super record: ricavi in aumento del 22%,” *ilsole24ore*, Nov. 2022.

[40] BCGxALTAGAMMA, “True-Luxury Global Consumer Insights,” 2021. Accessed: Oct. 16, 2022. [Online]. Available: https://altagamma.it/media/source/PRESS_True%20Luxury%20Global%20Consumer%20Insight.pdf

[41] S. Lazzaroni, “ALTAGAMMA CONSENSUS 2022,” Jun. 2022. Accessed: Oct. 16, 2022. [Online]. Available: https://altagamma.it/media/source/ALTAGAMMA%20CONSENSUS%202022%20UPDATE_1.pdf

[42] F. Natale, “IL SETTORE DEI BENI LUXURY: DINAMICHE COMPETITIVE E STRATEGIE,” LUISS, 2021. Accessed: Feb. 16, 2023. [Online]. Available: http://tesi.luiss.it/28625/1/211311_NATALE_FRANCESCO.pdf

[43] “RapidMiner.” https://rapidminer.com/ (accessed Feb. 25, 2023).

[44] “Google Colab.” https://www.miriade.it/google-colab-il-tool-gratuito-di-google-a-servizio-dei-data-scientist (accessed Dec. 10, 2022).

[45] L. Tremolada, “L’Italia esporta oltre 7,5 miliardi di euro in Russia. La mappa regionale dell’export,” *Il sole 24 ore*, Accessed: Dec. 11, 2022. [Online]. Available: https://www.infodata.ilsole24ore.com/2022/04/03/litalia-esporta-oltre-75-miliardi-euro-russia-la-mappa-regionale-dellexport/