Dipartimento di Impresa e Management
Corso di Laurea Magistrale in Marketing, Major in Analisi e Misure di Marketing
Cattedra di Customer Intelligence e Logiche di Analisi di Big Data

Developing Marketing Personas with Machine Learning in Luxury Fashion Industry

RELATRICE
Chiar.ma Prof.ssa Marina Paolanti

CANDIDATO
Piero Battistel
Matr. 748951

CORRELATORE
Chiar.mo Prof. Luca Romeo

Anno Accademico 2022/2023

Alla mia famiglia

Indice

Indice Figure

Indice Tabelle

Introduzione

L'utilizzo dell'intelligenza artificiale nel *marketing* è solo uno dei possibili campi di applicazione di questo insieme di tecnologie per i diversi domini. Data l'esperienza pregressa, è ormai ben fondata l'idea che l'utilizzo di metodologie basate su *machine learning* e *deep learning* abbiano migliorato diverse realtà aziendali.

Velocizzare i processi, migliorarne la precisione nei risultati e automatizzarli il più possibile sono da sempre obiettivi delle imprese.

Quando si parla di *machine learning* o apprendimento automatico ci si riferisce a un ramo dell'*artificial intelligence* che riguarda la realizzazione di modelli predittivi volti ad analizzare e stimare partendo da un numero di dati molto variabile.

Nel dominio del *marketing*, considerandolo nell'industria del *fashion*, l'apprendimento automatico è stato implementato per diverse finalità finora. Il caso più ricorrente è la segmentazione dei clienti, infatti, soprattutto in grandi realtà dove ci si deve interfacciare con delle moli di dati quotidianamente, è fondamentale avere degli strumenti e dei processi adeguati a gestire ogni singola informazione riducendo la probabilità di generare errori. Inoltre, a causa della globalizzazione e della creazione di *business* sempre più aperti e pronti ad essere attivi in mercati intercontinentali, diventa necessario avere dei processi in grado di gestire questo genere di complessità.

Avere la possibilità di segmentare i clienti in modo accurato e in grado di poter considerare molteplici pattern tra le loro informazioni ne migliora di conseguenza i risultati delle campagne di *marketing* e, più in generale, i risultati di *business* semplificando anche le scelte strategiche.

Per le stesse ragioni, avere un sistema di raccomandazione di prodotti è diventato fondamentale in questo periodo storico; in alcuni contesti si può dire che sia diventato essenziale soprattutto per poter competere. La realizzazione di questi complessi algoritmi destinati a suggerire per ogni singolo cliente i prodotti migliori sulla base di tante variabili di diversa natura come quelle sulla cronologia di ricerche e acquisti passati quanto altre su aspetti demografici, avviene mediante il *machine learning*. Le aziende nell'industria del *luxury fashion*, investendo sugli innovativi metodi di acquisto quali gli *e-commerce*, necessitano di tali tecnologie per migliorare l'esperienza del consumatore.

Infine, altri obiettivi del *machine learning* applicato al *marketing* sono tutte le analisi di *forecasting* su diversi aspetti della clientela. I più evidenti sono sicuramente la previsione

del *customer life time value* ma anche il *churn rate* in quanto una metrica importante soprattutto se conosciuta in anticipo e con meno errore possibile (Anisin, 2022).
Tutt'oggi è quindi possibile assistere o basare le scelte manageriali con delle metodologie fortemente se non esclusivamente basate su dati senza particolari vincoli. La tecnologia odierna è già in grado di prendere in considerazione enormi quantità di dati e restituirne degli output che abbiano un errore davvero irrisorio molto spesso.
Questo cambiamento dovuto all'avanzamento della componente computazionale determina delle situazioni favorevoli siccome, molto spesso, le scelte strategiche sono affiancate da pochi dati oppure dati di bassa rilevanza e, pertanto, realizzate solo in base all'esperienza manageriale. Si evince quindi che avere un modello in grado di poter analizzare svariati milioni o miliardi di dati contemporaneamente utilizzando complessi e specifici calcoli matematici determina un miglioramento oggettivo soprattutto in ottica decisionale. La letteratura attuale non risulta essere così vasta circa l'applicazione di *machine learning* e *deep learning* nell'industria del *fashion* (Mameli, 2021).
Proprio per questo lo studio che qui si presenta vuole essere un contributo per dimostrare ulteriormente come porre in essere complessi algoritmi di intelligenza artificiale in chiave di *business*. Nello specifico, l'elaborato riguarderà la possibilità di sviluppare delle *marketing personas* mediante l'apprendimento automatico. Queste sono delle elaborate descrizioni di clienti di riferimento che aiutano i dipartimenti di *marketing* nel prendere decisioni strategiche ed operative considerando il *target* di riferimento. Il tentativo di innovazione qui proposto è dovuto all'intenzione di voler sviluppare degli strumenti di *marketing* con degli algoritmi di intelligenza artificiale per la creazione di strumenti di *strategic marketing* quando, questi, sono sempre stati posti in essere in base alle esperienze e percezioni dei *marketing* manager o team di *marketing*.
L'analisi qui riportata riguarderà l'applicazione di modelli di *clustering* su di un *dataset* di un importante azienda nel panorama della moda di lusso. Nello specifico, i dati riguarderanno clienti internazionali e molteplici caratteristiche di diversa natura relativi a questi ultimi.
La finalità ultima non è meramente quella di ottenere di profili di acquirenti ma quella di stressare e validare la possibilità di applicazione del *machine learning* in un processo volto a migliorare l'esperienza del cliente finale come obiettivo ultimo (Siegel, 2023).
Partendo da una prima parte rivolta a chiarire gli aspetti principali dell'industria del *luxury fashion* in cui sono stati riportati dati e considerazioni sull'andamento e sulle tendenze

principali, il lavoro si sposta sul *marketing* strategico specificando i concetti di segmentazione, *target* e *marketing personas*.

La trattazione si evolve successivamente con un'analisi dello stato dell'arte volto a definire un divario di ricerca e un'analisi esplorativa del *dataset* oggetto di *clustering* nonché un'accurata descrizione di tutte le tecniche statistiche applicate in fase di *pre-processing* così come degli algoritmi utilizzati.

In conclusione, in base alla scelta del modello con le performance migliori verrà realizzata un'interpretazione dei risultati riportando le implicazioni manageriali degli stessi e del lavoro in generale.

1 Luxury Fashion Industry

L'industria del *luxury fashion* è uno dei settori in più rapida espansione e con le migliori performance, con aziende leader che hanno registrato una crescita rilevante negli ultimi anni.

Il mercato mondiale dei prodotti di lusso è dominato dalla moda, che si prevede crescerà a un ritmo sostenuto nel prossimo lustro (Figura 1).

Il *luxury fashion* è solo uno delle classificazioni di prodotti nel più ampio mercato dei beni di lusso composti anche da orologi e gioielli, prodotti in pelle, cosmetici e fragranze e occhialeria.

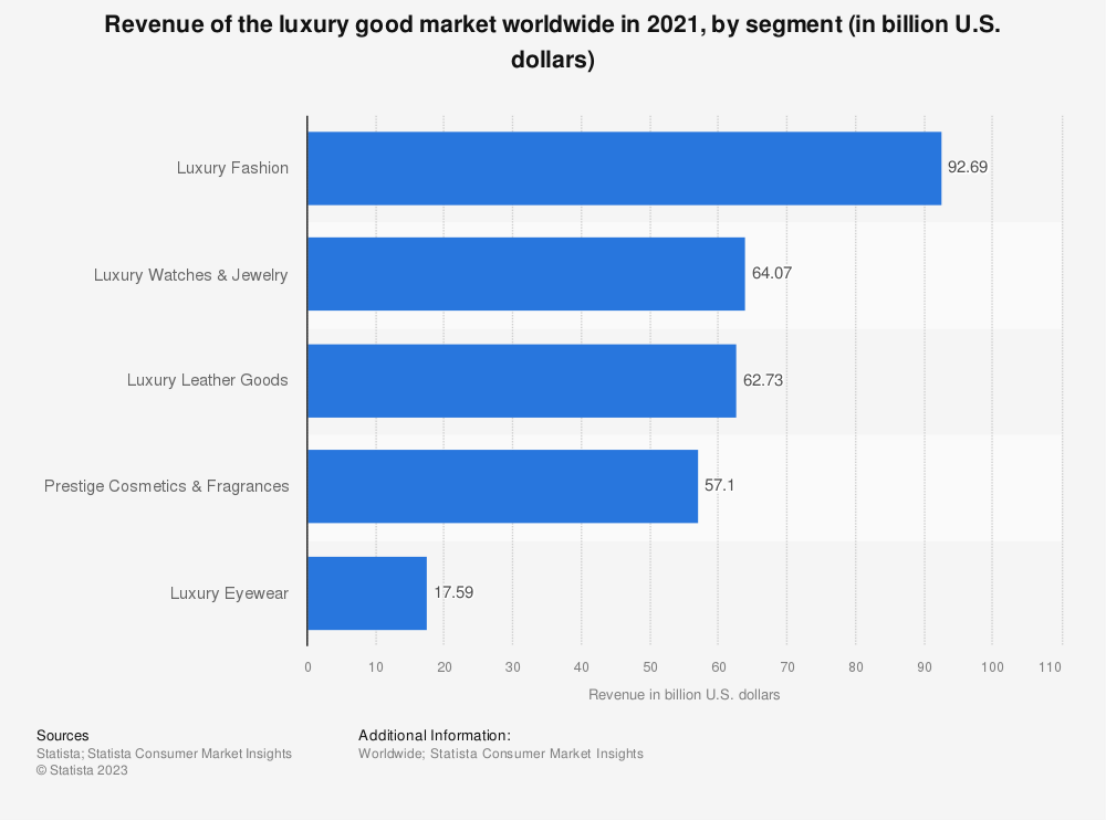

Figura 1 Ricavi del mercato globale dei beni di lusso nel 2021, per segmento (in milioni di dollari U.S.)

In economia, un bene può essere definito di lusso nel momento in cui la domanda aumenta più che proporzionalmente all'aumentare del reddito (Varian, 1992).

La parola lusso affonda le sue radici nel termine latino "*luxus*", che si riferisce alla sovrabbondanza, all'eccesso e all'ostentazione di ricchezza volta a soddisfare i desideri al di là dei bisogni reali. Il concetto di lusso è strettamente legato ai bisogni umani e la sua definizione concreta varia a seconda dell'epoca e della società in esame. Di conseguenza,

il lusso assume forme e significati diversi nel contesto dello spazio e del tempo, nonché a seconda della prospettiva di studio.

Il termine lusso può anche avere un'accezione negativa, associata a immagini di esagerazione e smodatezza. La radice "*lux*" dà anche origine alla parola "*luxuria*", che si riferisce a "esuberanza, profusione, lusso" e "vita lasciva e voluttuosa". Inoltre, il termine richiama il lemma greco "*loxos*", che significa crescita in modo obliquo, che può essere anche interpretato come variazione della normalità.

Il lusso è associato a esperienze, oggetti e servizi con un alto valore simbolico, per i quali i consumatori sono disposti a spendere cifre esorbitanti, ben al di sopra del prezzo medio della categoria di prodotto. Un alto valore figurativo contraddistingue questa sfera e suggerisce una deviazione e una distorsione in cui si mescolano eccesso, unicità e distinzione. Questo descrive un atteggiamento che comporta l'entusiasmo per tutto ciò che è al di fuori dello standard e un notevole allontanamento dal modo abituale di soddisfare i bisogni.

In questo contesto, il concetto di lusso ha un carattere flessibile e polivalente che è difficile ricondurre a un'idea unica e definitiva. Pertanto, una definizione concreta di lusso è soggetta a cambiamenti e interpretazioni nel tempo e nelle diverse società. Spetta alle prospettive e con i contesti individuali definire il significato di lusso. Essendo intrinsecamente ambiguo, oscilla anche tra due polarità opposte. Da un lato, rappresenta la ricchezza e il successo, meritando riconoscimento e ammirazione per le capacità di un individuo. Dall'altro, significa gusto pessimo e assenza di misura, caratterizzato da un attaccamento esclusivo con i beni materiali. Emergono così due diverse interpretazioni del concetto di lusso: il consumo ostentativo, che motiva la necessità di acquisire beni per esibire status e ricchezza, e il consumo edonistico, che favorisce la ricerca di gratificazione e soddisfazione personale.

Il lusso ha subito una metamorfosi, cambiando forma e valore in diversi periodi storici. Secondo Lipovetsky e Roux in "*Le luxe éternel*" (Lipovetsky, 2003), il lusso non è iniziato con la semplice produzione di oggetti costosi, l'ostentazione e la sontuosità. Prima di essere un simbolo di civiltà materiale, il lusso "paleolitico" era un fenomeno culturale, un atteggiamento mentale che vedeva l'affermazione dell'uomo come essere sociale e non animale. Anche la religione ha avuto un ruolo nell'emergere del lusso, dove questo ha assunto una forma sacra di simbolismo e di vicinanza al divino. Con lo sviluppo delle grandi civiltà del mondo antico, i beni di lusso si riferivano sempre più alla ricchezza, al

privilegio e al potere, soddisfacendo desideri che andavano oltre i bisogni primari. Tuttavia, in alcune culture, come l'antica Grecia, si sono osservate connotazioni negative in quanto è stata osservata la potenziale minaccia determinata dal lusso per la società.

Come già ribadito, oggi il significato del lusso dipende da molteplici e diversi fattori. In alcune culture, le spese per questa categoria di beni sono considerate un sottoprodotto della debolezza dell'anima umana, mentre in altre sono considerate una prova tangibile del successo. Il lusso rimane quindi un concetto complesso e sfaccettato, che racchiude in sé diverse interpretazioni e significati. (Cabigiosu, 2020)

LVMH (che ospita marchi come *Louis Vuitton*, *Fendi* o *Loro Piana*), *Kering* (con *Gucci*, *Balenciaga*, *Saint Laurent* e altri), *Ralph Lauren*, *PVH* (proprietario di *Calvin Klein* e *Tommy Hilfiger*), *Canali*, *Armani* e tante altre sono le aziende più note nella *luxury fashion industry*.

Negli ultimi anni il settore ha registrato un consolidamento significativo, con grandi gruppi come *LVMH* e *Kering* che hanno ampliato il loro portafoglio attraverso acquisizioni di marchi di lusso più piccoli. Si prevede che questa tendenza continui, poiché il settore cerca di massimizzare le economie di scala e di competere in un mercato sempre più affollato.

Nonostante l'impatto della pandemia sul settore produttivo, la moda di lusso ha continuato a registrare buoni risultati, trainata dalla forte domanda in Asia e Nord America. Nel 2020, secondo Statista, il mercato globale dei beni di lusso è stato valutato 256 miliardi di euro, con l'Asia che ha rappresentato la quota di mercato maggiore, pari al 38%.

Il mercato del *luxury fashion* genererà globalmente circa 108,70 miliardi di euro nel 2023. Valore in crescita siccome, globalmente, si prevede che arrivi a 125 miliardi di euro entro

il 2028; il *CAGR* (2023-2028) si attesta intorno al 3% (Figura 2).

Figura 2 Ricavi dell'industria globale della moda di lusso 2018-28 (in milioni di euro)

Nello specifico, quest'anno così come per i precedenti, *USA*, Cina, Giappone e *UK* sono i paesi con maggiore fatturato generato dal mercato (Figura 3).

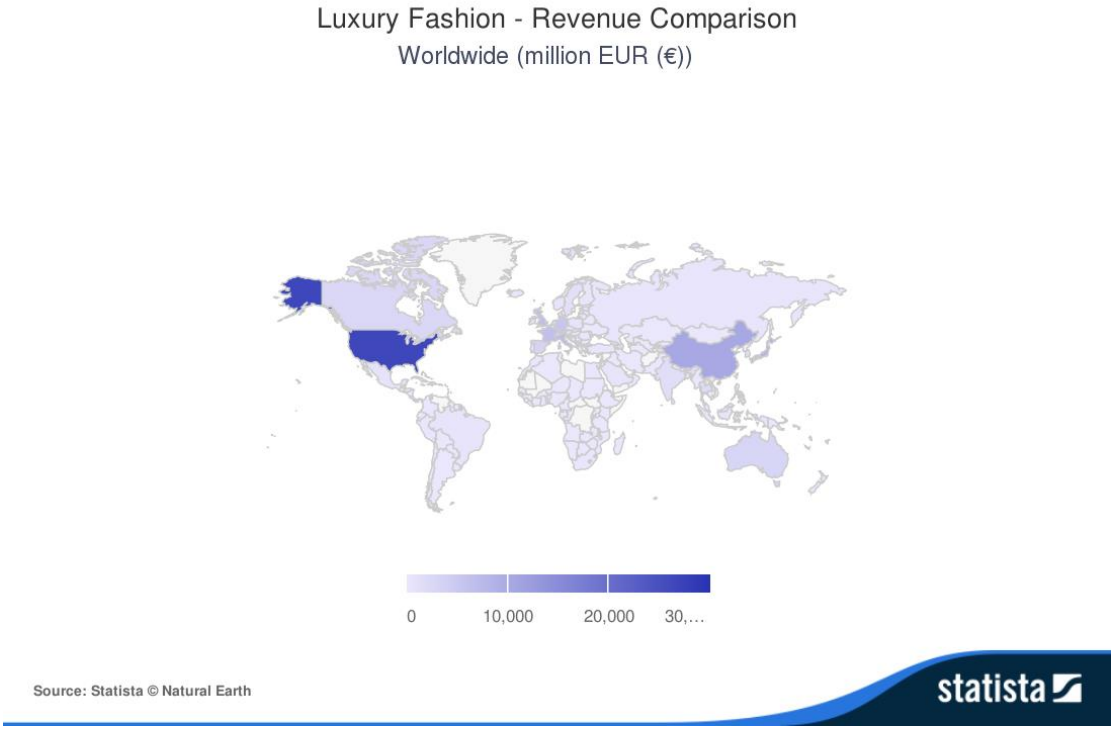

Figura 3 Confronto dei ricavi 2023 per le nazioni nel luxury fashion (in milioni di euro)

In aggiunta è previsto che i canali di vendita digitali siano al 22% del totale ma il dato è in aumento di circa due punti percentuali per anno. Durante il 2025 si prevede che ancora il 74% delle vendite avverrà offline (Figura 4).

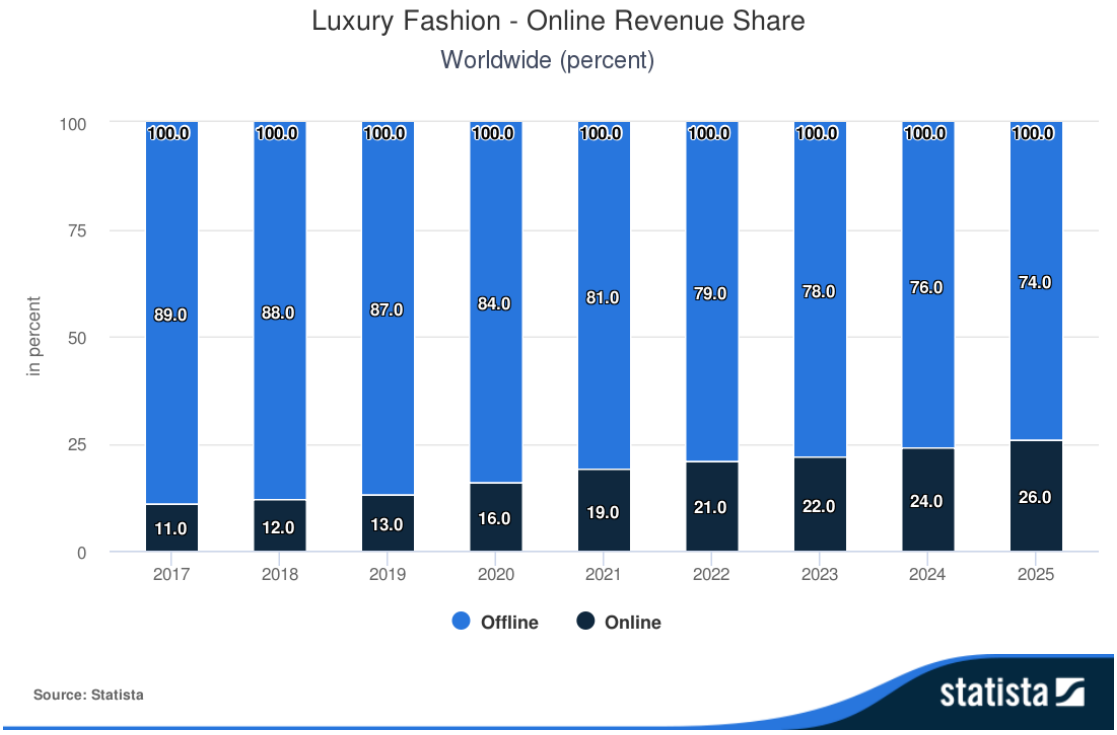

Figura 4 Quote di ricavi online e offline nel mercato globale della moda di lusso

Un altro aspetto inerente ai canali di vendita è la distinzione tra gli acquisti mediante cellulare o via computer. Per il 2023, *Statista* ha previsto una distinzione ancora molto vicina al 50% per ciascuno di questi. Più in particolare, la percentuale di *revenue* generata via cellulare è leggermente superiore con un valore pari al 57%. Anch'esso in aumento di un punto percentuale ogni anno. In aggiunta, la **Error! Reference source not found.** riporta la distribuzione delle quote di mercato globali per canale di vendita rendendo evidente che gli *store monobrand* e l'*online* siano stati i canali più importanti nel 2022 raggiungendo il 55% del totale (Statista, 2023).

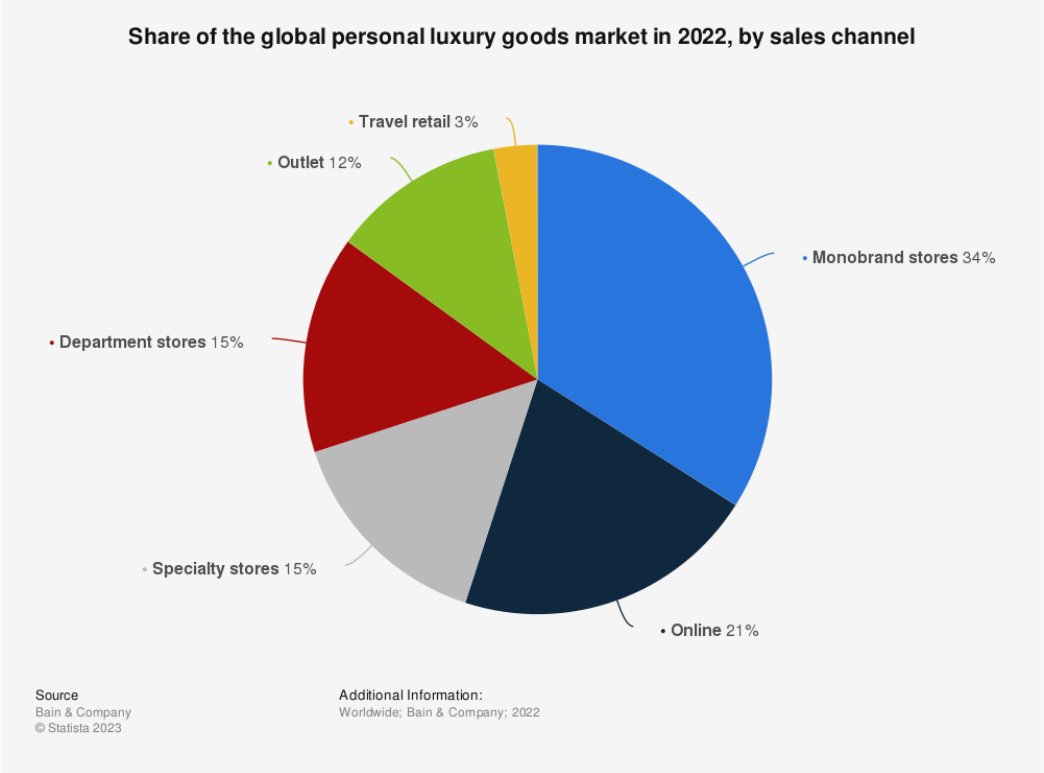

Figura 5 Quote di mercato globali nel settore dei beni di lusso personali per canale di vendita (2022)

Considerando la filiera produttiva a monte invece, il mercato globale della produzione di abbigliamento ha registrato una crescita moderata per tutto il periodo storico pre-pandemia, con un'accelerazione verso il 2018. Tuttavia, il mercato ha subito un calo di valore nel corso del 2019 e del 2020 a causa della pandemia globale da coronavirus, la quale ha interrotto le catene di approvvigionamento e la produzione nei Paesi in via di sviluppo, provocando un calo della produzione di indumenti e tessuti. Inoltre, a causa degli isolamenti necessari per gestire la crisi sanitaria, si è registrata una diminuzione delle spese non essenziali, tra cui proprio l'abbigliamento.

La crisi *COVID-19* ha avuto un impatto diffuso sui mercati e sui settori a livello mondiale, con quelli più globalizzati più esposti alle interruzioni della catena di approvvigionamento. Il settore dell'abbigliamento è stato tra i più colpiti dalla pandemia. Le misure di quarantena, la riduzione degli eventi sociali e le politiche di lavoro da casa hanno ridotto la domanda di abbigliamento formale e da lavoro. Tuttavia, la crisi sanitaria ha aumentato la domanda di prodotti *athleisure* e sportivi e di stili casual, poiché i consumatori desiderano sempre più condurre uno stile di vita sano.

Nel 2019, abbigliamento e calzature hanno rappresentato il 4,6% della spesa totale delle famiglie nell'UE. Tuttavia, come già ribadito, la pandemia *COVID-19* ha avuto un impatto

significativo sulla spesa dei consumatori, determinando una diminuzione della spesa delle famiglie per l'abbigliamento.

L'*e-commerce* ha contribuito a compensare la chiusura dei negozi, in quanto i consumatori hanno spostato le loro abitudini di spesa online durante la pandemia. Nel 2020, il settore della vendita al dettaglio di abbigliamento ha rappresentato il 14,9% del totale della vendita al dettaglio online globale, raggiungendo un valore di 359,6 miliardi di dollari.

Mentre l'industria manifatturiera di abbigliamento in Asia è cresciuta con un *CAGR* del 4% (2016-2020), raggiungendo un valore di 314,7 miliardi di dollari nel 2020, l'industria statunitense è diminuita con un tasso di variazione annuale composto (*CARC*) del -2,0% (2016-2020), raggiungendo un valore di 50,7 miliardi di dollari nel 2020.

Diversi Paesi della regione Asia-Pacifico hanno attraversato negli ultimi anni un periodo di rapida urbanizzazione, che ha portato alla crescita della classe media e all'aumento dei redditi, diventando un motore più dominante della crescita economica e spingendo la domanda di prodotti. La Cina è uno dei mercati dell'abbigliamento in più rapida crescita, il più grande produttore mondiale ed è spesso chiamata la "fabbrica del mondo". Il mercato è stato recentemente stimolato dal basso costo del lavoro, dalla manodopera qualificata e dalle infrastrutture, che ne fanno una destinazione attraente per i produttori. Inoltre, le aziende utilizzano sempre più spesso la Cina come hub produttivo per servire il crescente mercato dei consumatori e per le esportazioni. Le problematiche causate dalla pandemia hanno spinto molte aziende di abbigliamento a trasferire le loro basi produttive più vicino al Paese di origine per ridurre i costi di spedizione e minimizzare il rischio di interruzione della catena di approvvigionamento. Ad esempio, l'italiana *Benetton* ha potenziato la produzione in Serbia, Croazia, Turchia, Tunisia ed Egitto, per dimezzare la produzione in Asia entro la fine del 2022.

Il segmento femminile è stato il più remunerativo dell'industria dell'abbigliamento nel 2020, con un fatturato totale di 179,2 miliardi di dollari, pari al 36% del valore complessivo del settore.

Il segmento maschile ha contribuito con ricavi per 124,4 miliardi di dollari nel 2020, pari al 25,0% del valore complessivo del settore. Storicamente, il segmento femminile è stato più remunerativo di quello maschile, in quanto le donne spendono solitamente una percentuale maggiore del loro reddito in abbigliamento. Tuttavia, negli ultimi anni, gli uomini hanno aumentato la spesa per questo genere di prodotti così come quelli per la

cura del corpo, grazie alla crescente consapevolezza della moda e al miglioramento del tenore di vita, unitamente all'elevata capacità di spesa dei consumatori a medio reddito, in particolare nei Paesi asiatici.

I marchi sportivi come Nike e Adidas dominano il mercato globale dell'abbigliamento maschile, con quote di mercato rispettivamente del 2,7% e del 2,4%. Tuttavia, marchi premium e di lusso come Burberry e Ralph Lauren stanno espandendo le loro collezioni di abbigliamento maschile, indicando che il segmento in questione potrebbe diventare il più redditizio in futuro.

Era stato previsto che l'industria manifatturiera dell'abbigliamento si fosse ripresa nel 2021, con una previsione di crescita dei ricavi del 7,1% su base annua, grazie all'introduzione dei vaccini da parte delle principali economie e all'allentamento delle restrizioni di isolamento. Per i consumatori che hanno risparmiato durante la pandemia è stata registrata una maggiore propensione al consumo, con conseguente aumento della domanda di abbigliamento e accessori. Non a caso, il Dipartimento del Commercio degli Stati Uniti ha riscontrato una crescita delle vendite al dettaglio del secondo trimestre del 2021 pari al 28% rispetto all'anno precedente.

Tuttavia, la crescente tendenza verso un abbigliamento più sostenibile potrebbe inibire la crescita del mercato nel futuro. Molti consumatori sono sostenitori delle diverse forme di riciclo per gli abiti o dell'investimento di un ammontare di denaro superiore per ridurre il numero di capi di abbigliamento acquistando articoli di qualità superiore quindi, riducendo la domanda di abbigliamento di nuova produzione.

La domanda di abiti di seconda mano è aumentata nel corso del biennio 2020-2021 a causa della carenza nella catena di approvvigionamento di abiti nuovi (Figura 6). Di conseguenza, molti rivenditori di moda hanno lanciato la loro gamma di prodotti usati, tra cui *Levi Strauss & Co* e *Urban Outfitters*, che gestiscono entrambi negozi dell'usato online.

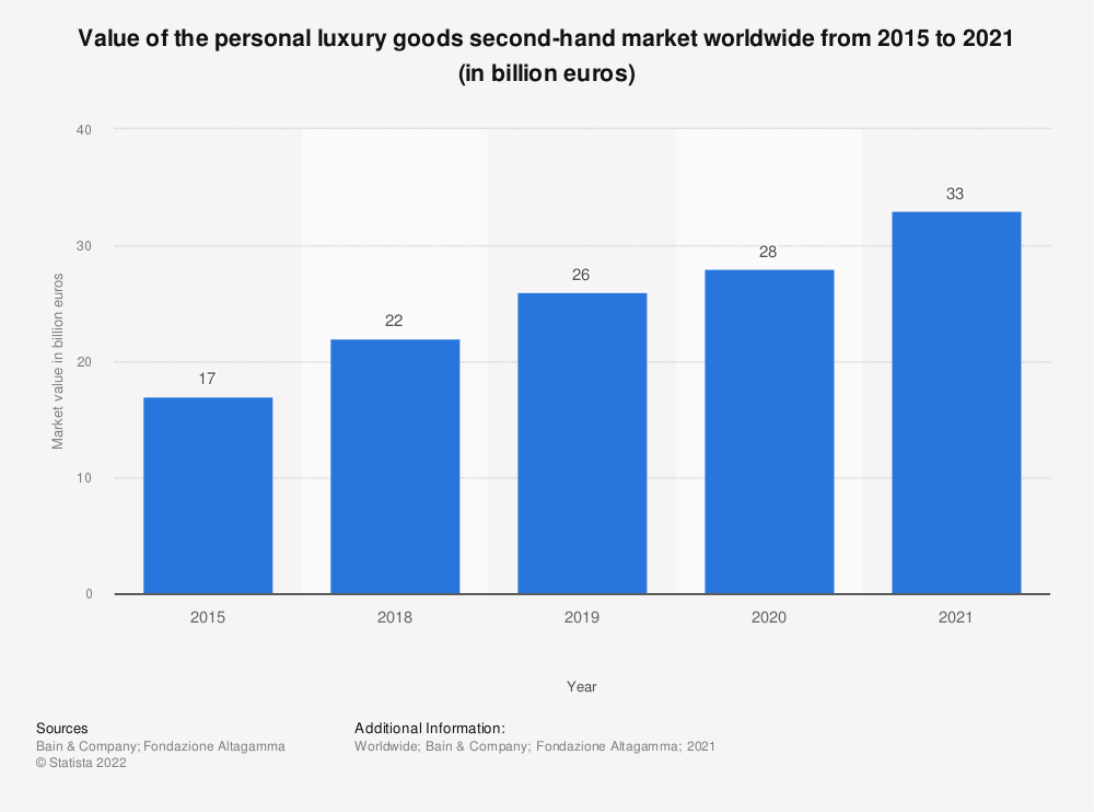

Figura 6 Valore del mercato globale dei prodotti di lusso di seconda mano (2015-2021) in milioni di euro

Aziende come *ThredUp* hanno capitalizzato la crescente domanda di articoli di seconda mano, mentre altre, come *Poshmark*, sfruttano il mercato della rivendita, in continua crescita, per aiutare i privati a vendere i propri abiti ad altri. Secondo *GlobalData*, il settore della rivendita dovrebbe crescere 11 volte più velocemente rispetto al settore dell'abbigliamento al dettaglio nel periodo fino al 2025, grazie alla crescente consapevolezza del peso ambientale del *fast fashion* (MarketLine, 2021).

Il report annuale di *McKinsey and Company* "*State of Fashion 2022*" elabora l'analisi dando maggiore peso alla componente manageriale delle aziende.

I dirigenti dell'industria della moda globale sono stati cautamente ottimisti per il 2022 sebbene le non poche difficoltà relative alla catena di approvvigionamento e alla domanda irregolare dei consumatori.

Le vendite di articoli di moda hanno ripreso slancio grazie al fatto che i consumatori, sempre più fiduciosi, dando sfogo al loro potere d'acquisto represso, hanno rinfrescato i loro guardaroba con la ripresa della vita sociale in tutto il mondo.

Mentre il settore del lusso dovrebbe essersi ripreso completamente con la fine del 2021, l'industria della moda in generale è tornata ai livelli di performance precedenti alla pandemia solo dal secondo trimestre 2022.

La *sentiment analysis* sui consumatori ha reso evidente una traiettoria positiva, soprattutto nei mercati con alti tassi di vaccinazione e di risparmio.

Nel segmento del lusso in Cina, la domanda di consumatori in attesa si era manifestata con il cosiddetto "*revenge shopping*", riscontrato anche nel più ampio mercato della moda negli Stati Uniti all'inizio del 2022.

In Europa, la fiducia dei consumatori nella ripresa economica è più cauta, con circa un quarto degli intervistati in un sondaggio di settembre 2021, ottimisti sul fatto che l'economia sarebbe tornata ai livelli pre-pandemici entro la fine del 2021, mentre oltre la metà prevedeva una ripresa solo nel 2022 o più tardi. Nel complesso, la moda è pronta a trarre vantaggio dai fattori macroeconomici fondamentali. I dirigenti si concentrano sulla crescita in un panorama di mercato modificato, anche se l'incertezza sulla ripresa della crisi e l'incoerenza sono persistite nel 2022.

Negli Stati Uniti, si stima che i risparmi nel primo trimestre del 2021 siano stati 3,1 volte superiori a quelli del primo trimestre del 2019. Dunque, si prevede che le generazioni più giovani e i consumatori a medio reddito dimostreranno la più forte propensione alla spesa per il tempo libero, con la moda tra le prime tre categorie.

In Cina, l'aumento dei redditi previsto contribuirà a un incremento atteso di 10.000 miliardi di dollari in termini di crescita dei consumi tra il 2021 e il 2030.

Per quanto riguarda la ripresa del settore, i leader sono più fiduciosi: il 75% dei dirigenti del lusso, il 61% dei dirigenti della fascia media del mercato e il 50% dei dirigenti della fascia alta prevedevano condizioni commerciali migliori nel 2022 rispetto al 2021.

Si prevede che l'industria della moda globale crescerà in modo disomogeneo tra le varie aree geografiche a causa degli *shock* economici e sanitari indotti dalle pandemie. Il mercato della moda cinese si è ripreso dalla pandemia ed è tornato alle vendite pre-Covid. Nonostante la previsione di un ritorno più lento ai livelli di vendita pre-pandemia in Europa, i dirigenti di questa regione erano i più ottimisti per l'anno 2022, probabilmente a causa di fattori quali la presenza relativamente forte dei marchi di lusso europei nei mercati globali. Il 67% dei dirigenti europei prevedeva condizioni commerciali migliori rispetto al 2021. Questo dato si confronta con il 57% dei dirigenti in Nord America e il

52% in Asia, dove la maggior parte dei mercati chiave è già tornata alle vendite pre-pandemia (McKinsey, 2021).

Considerando un altro studio del 2021 realizzato da *MarketLine* specificatamente sull'industria dei beni di lusso si evincono ulteriori elementi.

La pandemia da *COVID-19* ha portato a un calo significativo del 18% del mercato mondiale dei beni di lusso nel 2020.

Tra i segmenti, si prevede che il settore della gioielleria e dell'orologeria dovrà affrontare sfide più significative, soprattutto a causa della sua dipendenza dai canali di vendita al dettaglio fisici e dagli articoli con un prezzo superiore alla media. I mercati europei e statunitensi hanno contribuito in modo significativo al calo, mentre il mercato cinese e il resto del mondo hanno registrato una crescita notevole nonostante la crisi pandemia. I consumatori cinesi di beni di lusso costituiscono una porzione in rapida espansione del mercato globale, grazie all'ascesa di una classe media particolarmente interessata alla tipologia di prodotti.

Il mercato tedesco dei beni di lusso dipende in larga misura dagli acquisti dei turisti cinesi, che sono stati significativamente colpiti dalle restrizioni ai viaggi causate dalla pandemia. I mercati occidentali devono affrontare la sfida delle giovani generazioni che mostrano un minore interesse per i marchi di lusso tradizionali, di conseguenza questi, hanno bisogno di aiuto per adattarsi a questo cambiamento. Pertanto, sebbene esista una certa domanda locale, una parte significativa della crescita del mercato deriva dai consumatori di altri paesi. In Cina, le restrizioni ai viaggi hanno avuto un impatto positivo sul mercato, in quanto i consumatori cinesi hanno riorientato la loro spesa all'interno del Paese anziché all'estero. È notevole che i principali marchi di fascia alta come *LVMH*, *Kering* ed *Estee Lauder* abbiano registrato una crescita delle vendite nel secondo trimestre come risposta immediata alla pandemia.

A sostenere il mercato statunitense è il numero consistente di miliardari presenti nella nazione. Secondo Forbes, attualmente negli Stati Uniti si conta un numero record di 724 miliardari, tra cui 13 persone tra le prime 20 più ricche del mondo. Tuttavia, il mercato ha dovuto affrontare delle sfide nella sua espansione a causa del cambiamento dei comportamenti dei consumatori. I consumatori *millennial*, che costituiscono la parte più significativa della base di consumatori attivi, sono più inclini a spendere in servizi di lusso esperienziali, come cene di lusso e vacanze di alto livello, piuttosto che in beni di lusso tangibili. In questa generazione, un importante gruppo di consumatori noto come *HENRY*

(*High Earners - Not Rich Yet*) è significativo. Questo segmento, che si colloca tra il 75° e il 95° percentile della distribuzione del reddito (equivalente a guadagni annuali compresi tra 100.000 e 250.000 dollari), dimostra di preferire le vacanze di lusso, lo shopping online e il noleggio di articoli di lusso rispetto all'acquisto dei prodotti. Riconoscendo l'importanza di questo gruppo demografico come futuro consumatore benestante, i marchi del lusso hanno iniziato ad allineare le loro offerte di prodotti, le strategie di *marketing* e i canali di distribuzione per soddisfare le loro preferenze e stimolare la spesa presente e futura.

L'industria globale dei beni di lusso presenta un'intensa concorrenza guidata da importanti operatori, limitate opportunità di crescita e un'elevata concentrazione di mercato.

Sebbene i singoli consumatori costituiscano gli acquirenti di beni di lusso, la loro influenza è relativamente limitata. Tuttavia, la loro ricchezza, insita nella natura dei prodotti di lusso, conferisce loro una maggiore flessibilità di scelta, in quanto il prezzo di acquisto ha un'importanza minore rispetto ai tipici mercati al dettaglio. In generale, il potere degli acquirenti è considerato moderato.

I fornitori in questo mercato possiedono un certo livello di controllo sugli operatori, data la criticità di materie prime come pelle, oro o diamanti, che spesso non hanno sostituti diretti. Di conseguenza, il potere dei fornitori è valutato come moderato.

Il mercato dei beni di lusso presenta un'accessibilità favorevole per i nuovi operatori. L'assenza di elevati costi di cambiamento tra operatori storici e nuovi operatori e più in generale, la mancanza di elevate barriere all'ingresso consente ai nuovi agenti di trovare un mercato accomodante riuscendo a stabilire una nicchia specifica per i loro prodotti. Pertanto, la probabilità di nuovi ingressi risulta essere elevata.

Inoltre, il mercato offre numerosi potenziali alternative ai beni di lusso stessi. All'interno di ogni segmento specifico, come borse, orologi e gioielli, esistono molteplici opzioni non di lusso, ognuna con i suoi vantaggi rispetto alle controparti di lusso. Pertanto, la minaccia rappresentata dai sostituti è rilevante (MarketLine, Global - *Luxury* Goods, 2021).

2 Strategic Marketing

Il *marketing* strategico affonda le sue radici in un continuo lavoro di definizione avvenuto nel corso degli anni grazie a diversi contributi.

Tra i più importanti vi sono sicuramente Peter Drucker, Micheal Porter, Philip Kotler, Theodore Levitt, Al Ries e Jack Trout.

Peter Drucker definì nel 1973 il *marketing* strategico come:

"*Strategic marketing as seen as a process consisting of analyzing environmental, market competitive and business factors affecting the corporation and its business units, identifying market opportunities and threats and forecasting future trends in business areas of interest for the enterprise, and participating in setting objectives and formulating corporate and business unit strategies. Selecting market target strategies for the product-markets in each business unit, establishing marketing objectives as well as developing, implementing and managing the marketing program positioning strategies in order to meet market target needs*".

Si evince che questo concetto riguardi una specifica fase del processo di *marketing* nel suo complesso (Mongay, 2006).

Nello specifico, un processo di *marketing* può essere scomposto in sei componenti interconnesse. La formazione della strategia di *marketing* è esattamente il primo *step* necessario per poter iniziare.

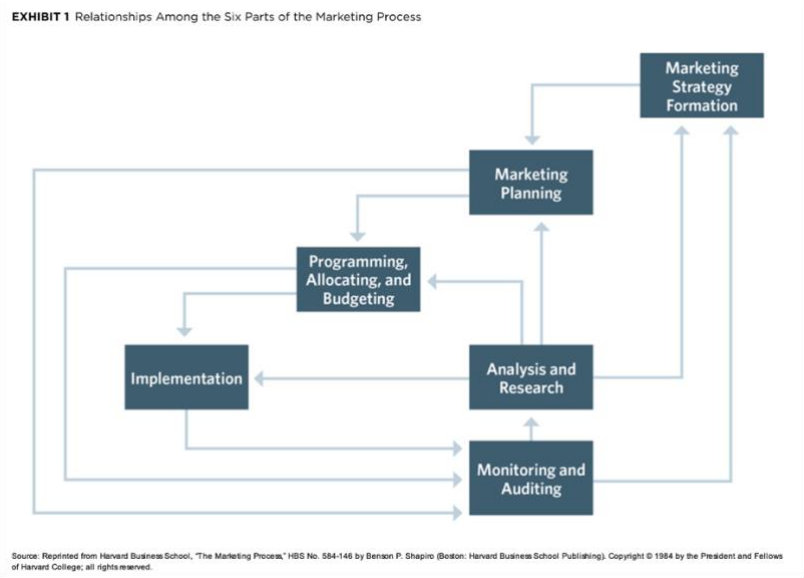

Figura 7 Relazione tra le sei parti del processo di marketing

Il *framework* di cui sopra (Figura 7), proposto per la prima volta da Benson Shapiro in "*The Marketing Process*", pone in evidenza soprattutto come le diverse componenti siano connesse tra loro e come effettivamente venga data attuazione all'intero processo di *marketing*.

Definendo la strategia di *marketing* si stabiliscono degli obiettivi di lungo periodo da cui ne derivano anche le modalità con cui l'azienda intende approcciarsi al mercato. Sulla base di ciò, considerando anche la precedente definizione di *strategic marketing*, si può scomporre ulteriormente il concetto in segmentazione della domanda, scelta dei segmenti *target* e posizionamento.

Sebbene queste siano le fasi da sempre ricondotte a questo concetto, nella trattazione più moderna è possibile sintetizzare il processo di formazione della strategia di *marketing* in tre grandi step ossia l'analisi, le decisioni e i risultati.

Figura 8 Schema del processo di formazione della Strategia di Marketing

Nello schema riportato nella Figura 8 si nota come tutto parta da diverse analisi: clienti, azienda, concorrenza, collaborazioni e contesto. Tra queste è stata data maggiore importanza all'analisi del cliente in linea con quello che è l'orientamento più attuale e utilizzato dalla maggior parte delle imprese negli ultimi lustri.

Soprattutto in questa prima fase si evince come avere un elevato numero di dati qualitativi e quantitativi sia un fattore chiave su cui poter basare tutto ciò che segue.

Sebbene la parte di decisioni sia scissa in *aspiration* e *action plan*, ciò che maggiormente risulta essere parte fondante della strategia di *marketing* è la prima in cui si realizza la segmentazione del mercato, si decide quali saranno i segmenti *target* nonché quale posizionamento si vuole occupare nella mente dei consumatori.

In questa specifica fase, le *personas*, declinate in ottica di *marketing*, sono utilizzate come strumenti per facilitare ulteriormente le scelte strategiche conoscendo meglio il cliente *target*.

Allo stesso tempo, risultano anche le scelte operative di *marketing mix* riguardanti prodotto, comunicazione, distribuzione e prezzo. Le prime tre sono definite come delle scelte di creazione del valore mentre l'ultima è di cattura del valore.

La scelta del *marketing mix* dipende e deriva soprattutto dal *target* e dal posizionamento; esattamente per questo, nella fase di *decisions* è stata apposta una numerazione.

Di seguito, ne risulteranno diversi output; per quanto concerne gli output dovuti alle scelte di *aspiration*, questi saranno tutti strettamente legati ai clienti. Nello specifico, in base al *target* di riferimento e al posizionamento che si vorrà raggiungere si potranno stimare diversi dati inerenti all'*acquisition*, alla *retention* e al tasso di acquisto (Dolan, 2014).

2.1 Customer Segmentation

Le aziende non hanno la possibilità di servire e stringere una relazione con tutto il mercato nel suo complesso. Oggigiorno è diventato necessario definire solo specifici segmenti da soddisfare nel modo più efficiente possibile. La segmentazione dei clienti è quel processo atto a trasformare un contesto eterogeneo e dominato da molteplici differenze in omogeneo e caratterizzato da gruppi tra loro diversi per molteplici aspetti ma simili al loro interno. In altre parole, ogni segmento di clienti sarà formato da persone con requisiti e desideri simili e, al tempo stesso, diverse rispetto agli altri segmenti.

La segmentazione è uno step fondamentale per poter arrivare alla realizzazione delle *marketing personas*. Per poter raggruppare i clienti in dei *cluster* si possono utilizzare due grandi gruppi di caratteristiche.

Una prima riguarda tutte le informazione di natura maggiormente descrittiva come variabili geografiche, demografiche e psicografiche dando importanza anche ai bisogni dei consumatori nonché ad eventuali loro manifestazioni circa il prodotto o il servizio dell'azienda.

D'altro canto, invece, un altro gruppo di dati su cui spesso si basa l'attuazione della segmentazione è di natura comportamentale. Rientrano in questa classificazione tutte le informazioni sui consumatori definibili come delle risposte ai marchi, alle occasioni di utilizzo nonché ai benefici.

TABLE 9.1	Major Segmentation Variables for Consumer Markets
Geographic region	Pacific Mountain, West North Central, West South Central, East North Central, East South Central, South Atlantic, Middle Atlantic, New England
City or metro size	Under 5,000; 5,000–20,000; 20,000–50,000; 50,000–100,000; 100,000–250,000; 250,000–500,000; 500,000–1,000,000; 1,000,000–4,000,000; 4,000,000+
Density	Urban, suburban, rural
Climate	Northern, southern
Demographic age	Under 6, 6–11, 12–17, 18–34, 35–49, 50–64, 64+
Family size	1–2, 3–4, 5+
Family life cycle	Young, single; young, married, no children; young, married, youngest child under 6; young; married, youngest child 6 or older; older, married, with children; older, married, no children under 18; older, single; other
Gender	Male, female
Income	Under $10,000; $10,000–$15,000; $15,000–$20,000; $20,000–$30,000; $30,000–$50,000; $50,000–$100,000; $100,000+
Occupation	Professional and technical; managers, officials, and proprietors; clerical sales; craftspeople; forepersons; operatives; farmers; retired; students; homemakers; unemployed
Education	Grade school or less; some high school; high school graduate; some college; college graduate; post college
Religion	Catholic, Protestant, Jewish, Muslim, Hindu, other
Race	White, Black, Asian, Hispanic, Other
Generation	Silent Generation, Baby Boomers, Gen X, Millennials (Gen Y)
Nationality	North American, Latin American, British, French, German, Italian, Chinese, Indian, Japanese
Social class	Lower lowers, upper lowers, working class, middle class, upper middles, lower uppers, upper uppers
Psychographic lifestyle	Culture-oriented, sports-oriented, outdoor-oriented
Personality	Compulsive, gregarious, authoritarian, ambitious
Behavioral occasions	Regular occasion, special occasion
Benefits	Quality, service, economy, speed
User status	Nonuser, ex-user, potential user, first-time user, regular user
Usage rate	Light user, medium user, heavy user
Loyalty status	None, medium, strong, absolute
Readiness stage	Unaware, aware, informed interested, desirous, intending to buy
Attitude toward product	Enthusiastic, positive, indifferent, negative, hostile

Tabella 1 Sintesi delle variabili di segmentazione (geografiche, demografiche, psicografiche e comportamentali)

La Tabella 1 qui riportata è una sintesi delle principali variabili utilizzate per la segmentazione definita precedentemente. Queste sono solo alcune delle possibili informazioni che si possono utilizzare per questa fase. Con il susseguirsi delle innovazioni nascono ogni anno nuovi fattori che possono essere oggetto di raggruppamento dei clienti per finalità di *marketing*.

La segmentazione basata su variabili geografiche cattura gli aspetti territoriali e spaziali ed è da sempre utilizzata dalle aziende per realizzare campagne e attività di *marketing* in base alle peculiarità dell'area, della zona, del paese o del continente.

La demografia è lo studio e l'analisi delle popolazioni umane considerate sulla base di diversi indici e metriche come ad esempio il sesso, il reddito, la razza, la religione, l'occupazione, l'età, la dimensione della famiglia, la classe sociale, la generazione, la nazionalità, il livello di istruzione e il *family life cycle*. Pertanto, sono considerati aspetti relativi sia alla composizione della popolazione così come al loro sviluppo e ai caratteri generali. Di conseguenza, si evince che questa dimensione della persona ingloba sia variabili qualitative che quantitative contrariamente a quanto visto con la dimensione geografica.

La psicografia è una branca della psicologia che riguarda la descrizione dei fatti della coscienza. Nello specifico, nell'ottica qui presentata, può essere vista come la disciplina che prende in considerazione psicologia e demografia per comprendere in modo più elaborato i consumatori. Gli aspetti presi in considerazione in questo caso sono tratti psicologici e della personalità, i valori delle persone, e lo stile di vita.

Se per le dimensioni demografiche e geografiche, l'osservazione delle informazioni è più immediata, per questa modalità di segmentazione sono stati creati delle metodologie specifiche. Tra le più utilizzate ricorre il VALS (acronimo di *Values, Attitudes and Lifestyles*) *framework* di *Strategic Business Insights*, il quale si fonda su otto classi (Strategic Business Insights, 2009). La classificazione degli individui avviene mediante questionari basati su domande principalmente attitudinali e in parte anche demografiche. Il database su cui si basa questo strumento psicografico è costantemente aggiornato e, ad oggi, si basa su dei dati ottenuti da più di 80000 sondaggi.

Il *framework* (Figura 9) si basa su due dimensioni; quella orizzontale ossia la motivazione della persona e quella verticale ossia le risorse dell'individuo. Di conseguenza ogni classe riporta aspetti sui valori, sulle attitudini e sullo stile di vita delle persone.

In base alla motivazione si distinguono le persone motivate in base agli ideali, le quali sono influenzate dalla conoscenza e dai principi, le persone motivate dalla realizzazione le quali realizzano le loro scelte in base ai risultati dei loro pari e le persone motivate dalla *self-expression* le quali si caratterizzano per essere alla ricerca del rischio, della varietà e delle attività fisiche e sociali.

Per risorse e quindi quanto misurato dall'asse verticale del *framework*, si intendono i tratti della personalità e aspetti demografici. I tratti della personalità considerati sono ad esempio il vigore, la fiducia in sé, la capacità di guida, il desiderio di novità, l'intellettualità, lo spirito innovativo, la vanità, l'impulsività ecc.

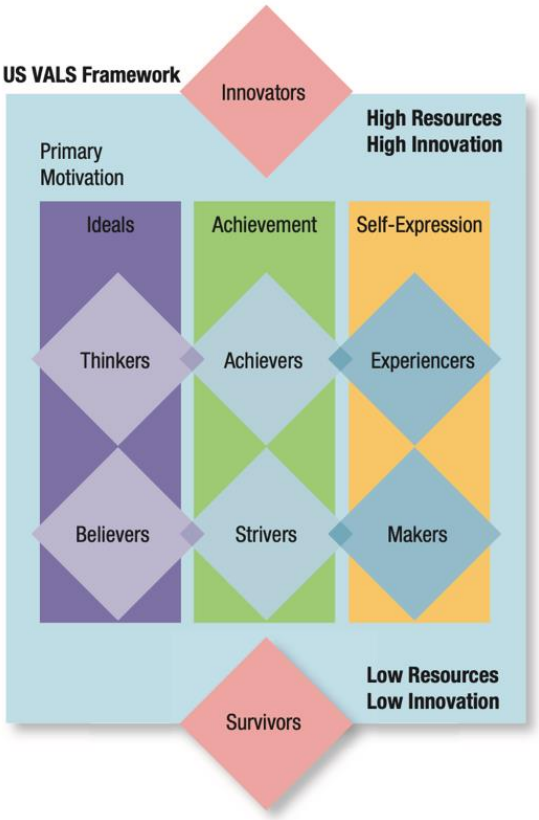

Figura 9 VALS Segmentation System Fonte: www.strategicbusinessinsights.com/vals 2014

Infine, la segmentazione comportamentale concerne tutto ciò che può essere definito come una risposta dei consumatori al prodotto o al servizio. Pertanto, ci si riferisce all'utilizzo che un consumatore fa del prodotto, all'attitudine verso un servizio così come anche la conoscenza di un bene di un'azienda.

Le variabili utilizzate nella segmentazione comportamentale sono quindi i bisogni o i benefici dei consumatori, i ruoli decisionali nonché tutte quelle riconducibile all'utilizzo del prodotto. Così come per la dimensione demografica, anche in questo caso figureranno dati quantitativi e qualitativi.

Affinché la segmentazione sia efficace è necessario che siano rispettati i seguenti principi: misurabilità, sostanzialità, accessibilità, differenziabilità e azionabilità.

Il primo criterio riguarda la possibilità di misurare ogni singola variabile necessaria ai fini della segmentazione. Per sostanzialità, in *Kotler P., 2016* gli autori si riferiscono alle dimensioni in termini di consumatori e di profitti adeguati all'investimento necessario per poter acquisire, gestire e mantenere i clienti.
L'accessibilità e l'azionabilità sono due aspetti simili ma che riguardando due aspetti diversi. Il primo si rifà alla possibilità di raggiungere fisicamente il segmento mentre l'azionabilità riguarda la capacità di poter realizzare prodotti, servizi ed esperienze attrattive per il segmento.
Per ultimo, la differenziabilità ossia la peculiarità del segmento di essere differente dagli altri nonché rispondente in modo diverso in base a prodotto, distribuzione, comunicazione e prezzo diversi. (Kotler P., 2016)

2.2 Targeting

Il *targeting* è quell'attività strategica di *marketing* indirizzata alla scelta di uno o più segmenti come individui per i quali verranno prese decisioni di *marketing mix* nonché a cui verranno indirizzati gli sforzi di *business* per generare ricavi e profitti.
Secondo Michel Porter, un mercato o un segmento può essere considerato attrattivo in base a una valutazione strutturata su cinque forze.
Queste sono la rivalità dei concorrenti nell'industria, la minaccia dei nuovi entranti, il potere contrattuale dei fornitori, il potere contrattuale dei clienti e la minaccia dei prodotti sostitutivi.
Maggiore sarà il numero di concorrenti, la quota di mercato da loro detenuta oppure, ancora, il capitale a disposizione per quel mercato o segmento, minore sarà l'attrattività dello stesso.
Per le medesime ragioni, minori saranno le barriere all'ingresso del mercato, maggiore sarà la repulsività del segmento.
I poteri contrattuali dei fornitori e dei clienti sono connessi con l'appetibilità di uno o più *cluster* di mercato dal momento che, maggiori essi saranno, più elevata sarà la difficoltà di poter prendere decisioni o porre in essere cambiamenti avendo come *target* dei clienti rientranti in questi scenari. Quindi maggiore sarà il peso della dipendenza con gli altri *stakeholder*, maggiore diventerà complesso gestire l'attività di *marketing* nonché di *business*.

Infine, se, per un determinato segmento o mercato dovesse esserci una situazione attuale o potenziale di prodotti o servizi sostituti tale per cui, per i clienti potrebbe essere effettivamente poco oneroso cambiare scelta di acquisto, l'attrattività sarà minore. (Porter, 1980)

L'attrattività del segmento è solo una delle considerazioni che si possono fare nel momento in cui si intende definire il *target*. Altri aspetti da considerare altrettanto fondamentali sono gli obiettivi aziendali e le risorse che si dispongono.

Una matrice utile per poter realizzare la scelta del *target* basandosi sull'attrattività del *target* e sulla fattibilità di poterlo soddisfare è la "*GE-McKinsey nine-box matrix*" realizzata per General Electric. Il *framework* (Figura 10) è atto a prioritizzare gli investimenti dell'azienda.

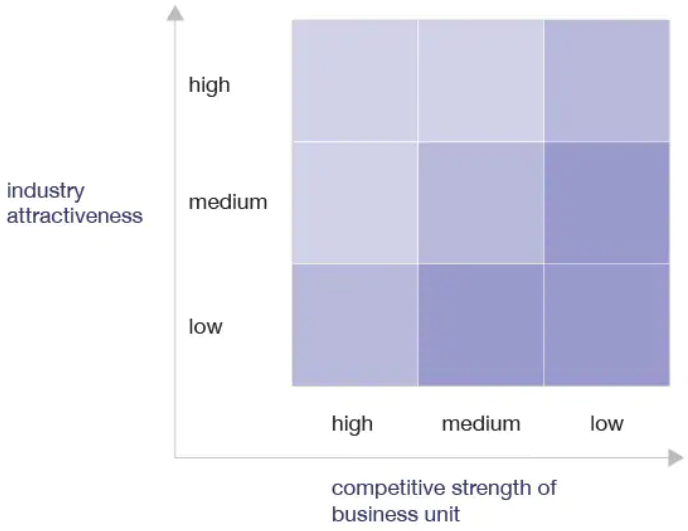

Figura 10 GE-McKinsey nine-box matrix

Sull'asse orizzontale è riportato la forza competitiva delle *business unit* mentre su quello verticale, l'attrattività dell'industria; in corrispondenza di valori alti per entrambe le dimensioni si configura la scelta migliore (McKinsey, Enduring Ideas: The GE–McKinsey nine-box matrix, 2008).

Invece, la figura inserita di seguito (Figura 11), riporta sotto forma di linea continua, i possibili scenari di segmentazione da cui scaturiscono anche le diverse conseguenze che ne derivano per quanto concerne la scelta del *target*. Un'azienda potrebbe decidere di avere come *target* l'intero mercato, più segmenti, un solo segmento oppure singoli individui ognuno considerato come segmento a sé stante. In base allo scenario, ci saranno scelte di *marketing mix* molto differenti (Kotler P., 2016).

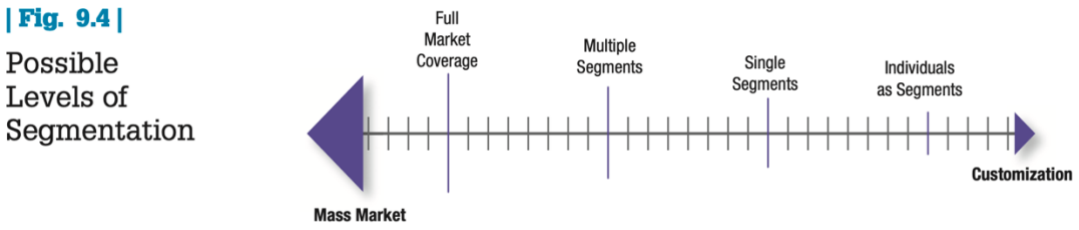

Figura 11 Possibili scelte di target. Marketing Management

2.3 Marketing Personas

Le *personas* sono uno strumento nato per diverse finalità; per quanto concerne lo *strategic marketing*, possono essere definite come uno strumento volto a migliorare l'approfondimento del *target* di riferimento mediante un'accurata descrizione di un archetipo. Da ciò ne deriva un miglioramento della conoscenza delle peculiarità del consumatore finale, ergo la possibilità di prendere decisioni molto più precise sulla base dei bisogni, attitudini, esigenze, problematiche e comportamenti del consumatore finale.
Una credenza comune è quella per cui i dati demografici siano sufficienti per definire in modo omogeneo e rilevante i segmenti. Tuttavia, la correlazione tra le variabili descrittive della dimensione demografica di un individuo e i bisogni dello stesso è molto bassa.
Due individui con le medesime caratteristiche demografiche, possono avere bisogni completamente diversi o addirittura opposti (Kotler P., 2016).
L'idea di creare delle *personas* è nata nell'industria del software con Alan Cooper, il quale ha riconosciuto che le aziende non riuscivano a produrre software di alta qualità perché non prendevano in considerazione le *personas* degli utenti durante il processo di progettazione. Il concetto di creazione di *personas* di Cooper non si limitava allo sviluppo di software, ma poteva essere applicato anche alle vendite e al *marketing*. Questo ha portato all'era in cui i *marketer* hanno iniziato a creare delle *personas* per comprendere e spiegare le esigenze degli acquirenti (Cooper, 1999).
Pur essendoci delle analogie tre *personas* del software e del *marketing* in quanto queste condividono le stesse componenti di base: familiarità, facilità di riconoscimento e tentativo di creare un legame emotivo con il gruppo di utenti che rappresentano, differiscono nel loro approccio. Le *personas* nell'industria del *software* mirano a raccontare la vita degli utenti e il modo in cui il prodotto potrebbe essere utilizzato, mentre le *personas* in ottica di *marketing* considerano il motivo per cui gli utenti hanno bisogno

del prodotto, cosa scatena la realizzazione dei loro bisogni e quali fattori portano alla decisione finale di acquisto.
Riportando la definizione di Adele Ravella fondatrice del *Buyer Persona Institute*, una buyer persona è:
"It's an archetype, a composite picture of the real people who buy, or might buy, products like the ones you sell.
It's an avatar you craft from what you learn in direct interviews with as many buyers as possible. And from behavior observed anywhere else: at industry conferences; in online forums; through social media.
If crafted with skill and insight, the person who emerges in this picture may become as real to you as anyone you can ever remember meeting. In your mind's eye, this person becomes three-dimensional to the point that you can see the world through his or her eyes." (Revella, 2011).

In generale, le *personas* vengono create per fornire una comprensione più olistica ed empatica del pubblico *target*. Sono utilizzate per basare le scelte strategiche sugli attributi dell'individuo (HCD), assicurando che i manager possano comprendere chiaramente il pubblico *target* ed entrare indirettamente in contatto in anticipo con esso attraverso questi strumenti (Koponen, 2017).
In altre parole, le *personas* sono un modo per ridurre la complessità data dall'utilizzo di un elevato numero di variabili per descrivere il *target* di riferimento. Inoltre, rendono dei dati molto spesso di difficile considerazione da un punto di vista operativo, facilmente interpretabili attribuendogli una dimensione umana; in tal modo risulteranno anche maggiormente accessibili alle figure professionali meno analitiche (Jansen, 2020).
Nel corso degli ultimi decenni non si è verificato solo un aumento dell'uso delle *personas* bensì i professionisti hanno anche esplorato vari mezzi per presentarle, come poster, siti web e cartoncini a grandezza naturale (ibidem, 2017).

Figura 12 Esempi di Personas

La creazione delle *personas* è un processo a più fasi che coinvolge dati qualitativi e quantitativi. Inizialmente, i dati quantitativi vengono spesso raccolti attraverso interviste e analisi demografiche del pubblico *target*. Le note e le registrazioni delle interviste vengono riassunte e raggruppate in base a un quadro di riferimento, che può variare a seconda delle esigenze delle *personas.*

Una *personas* è costituita da diversi parametri come si evince dalla Figura 12: un'immagine che permette di visualizzare l'archetipo del consumatore, diversi dati demografici, i bisogni, i *pain point*, i comportamenti del cliente in relazione alla categoria (in che modo interagisce con i prodotti o i servizi), il *decision journey* con tutti i *touchpoint*, gli *stakeholder* di riferimento con particolare rilevanza per quelli fondamentali nel modello decisionale, lo stile di comunicazione e il tono di voce preferito e, infine, le modalità di coinvolgimento preferite (Li, 2020).

Sebbene in principio, il processo di creazione era completamente manuale, ad oggi ci sono stati diversi tentativi di automazione. In aggiunta, con l'aumentare delle moli di dati a disposizione delle aziende, tutt'oggi è oggetto di ricerca la possibilità di realizzare delle *data-driven buyer personas* in modo più efficiente e affidabile. L'intento è quello di limitare le criticità date dai metodi tradizionali di creazione come, ad esempio, i sondaggi e le interviste, le quali presentano dei limiti importanti.

I primi tentativi di automatizzare la generazione di *personas* sono stati effettuati da *Jung et al.* in "*Automatically Conceptualizing Social Media Analytics Data via Personas*" (Jung, 2018) e in "*Persona Generation from Aggregated Social Media Data*" (Jung S. G., 2017) i quali, partendo da dati di social media, hanno realizzato delle *marketing personas* utilizzando un algoritmo di apprendimento non supervisionato chiamato "*non-negative matrix factorization*" (Lee, 1999) il quale mediante una matrice di dati non negativi cerca di approssimarla come il prodotto di più matrici non negative. Questo approccio è servito considerando che il *dataset* di partenza conteneva dei dati aggregati che gli algoritmi di *clustering* non sono in grado di gestire. Pertanto, per raggiugere l'obiettivo hanno disaggregato i dati riducendo difatti la dimensionalità del *dataset* fino a trovare i *p* comportamenti latenti, i quali sono stati considerati come numero di *personas*. Di conseguenza, in base al numero di comportamenti individuati sono stati selezionati gli attributi demografici (Jansen, 2020).

3 Literature Review

Il presente capitolo riguarda una revisione esaustiva dello stato dell'arte al fine di ricercare ed individuare un divario di ricerca per il quale è stato possibile realizzare l'analisi di questo contributo.

Lo studio di Asenio et al., 2022 mira a clusterizzare i contatti ricevuti dalle università in dei gruppi omogenei per definire quale di questi sia più propenso ad iscriversi successivamente ai corsi. Ad essere diverso non è solo il settore in cui è applicato lo studio e per il quale si sviluppano numerose divergenze con quello del *fashion,* ma anche le modalità di raccolta dei dati avvenuta per lo più mediante dei sondaggi. Nonostante ciò, sono stati utilizzati degli algoritmi di *machine learning* quali il *k-medoids* con una distanza *Gower*. Quest'ultima scelta è dipesa dalla tipologia di variabili categoriche e numeriche presenti nel *dataset*, determinando così l'impossibilità di utilizzare la distanza Euclidea e di *Manatthan*. Le variabili utilizzate da Asenio et al. sono state l'età, la media dei voti, la presenza di un'attuale posizione lavorativa, la tipologia di posizione lavorativa, la dimensione dell'azienda, il *budget,* il tempo impiegato per rispondere al sondaggio e il paese di provenienza. Le feature, oltre ad essere specifiche per l'obiettivo del lavoro, sono state raccolte completamente mediante un questionario raggiungendo un totale di 16272 *form* completati da maggio a ottobre 2020. Il numero di *cluster* è stato definito a priori utilizzando la *silhouette width* ossia uno dei metodi più comuni per misurare la somiglianza tra ogni punto in un *cluster* comparandola con il punto più vicino di un altro *cluster*.

Il *K-Medoids* è un algoritmo di *clustering* non gerarchico per il quale ogni *cluster* si struttura attorno a un punto centrale che in questo caso è proprio uno dei dati compresi nel *dataset*. Può essere definito come una variante del *K-Means* sebbene quest'ultimo non sia detto che abbia come centro del *cluster* uno dei dati di *input* dal momento che ogni *cluster* si struttura intorno al punto medio di ogni suo punto. Nonostante ciò, sia il *K-Means* che il *K-Medoids* si basano su una procedura iterativa volta a minimizzare la varianza *within* massimizzando la varianza tra *cluster*. Tuttavia, il *K-Medoids* risolve un problema di interpretabilità in quanto con il *K-Means* potrebbe essere difficile interpretare il centroide non essendo uno dei punti del *dataset*.

I risultati dello studio sono tre *cluster* tra i quali, uno in particolare, presenta un *conversion rate* significativamente più alto rispetto gli altri. Il *conversion rate* è il tasso di conversione che si registra per una determinata base di utenti o clienti definendo l'azione

che verrà considerata come conversione. Inoltre, per le finalità di *data visualisation* è stata utilizzata la tecnica multivariata *t-Distributed Stochastic Neighbour Embedding,* la quale permette di visualizzate dati multidimensionali su uno *scatterplot* bidimensionale. Sebbene lo studio abbia non poche divergenze rispetto al lavoro qui realizzato, ci sono numerosi spunti ed elementi che possono essere considerati nella realizzazione della *clustering analysis* partendo dall'utilizzo delle tecniche statistiche di *data* visualisation così come le implicazioni basate sui risultati (Asensio, 2022)

Il paper di Cibulková e Sulc riguarda una *clustering analysis* per finalità di segmentazione dei clienti nel settore del turismo. Il *dataset* utilizzato nello studio contiene dati su 5755 viaggi dei clienti e le feature sono molteplici come, ad esempio, il numero di passeggeri che viaggiano insieme, l'età, la nazionalità, il prezzo, il numero di destinazioni/attività prenotate, i dettagli dei percorsi, ecc. In particolare, la maggior parte sono categoriche e binarie nonostante ci sia qualche variabile numerica.

Così come nel lavoro qui realizzato e che verrà nei capitoli successivi presentato, anche l'analisi di Cibulková e Sulc presenta finalità di *marketing*.

La parte di *feature selection* è stata realizzata sulla base di due principali considerazioni. La prima riguarda il dominio di appartenenza dell'analisi tale per cui è stato possibile definire ciò che sarebbe risultato più rilevante per le finalità definite. In secondo luogo, è stata considerata la correlazione che vi è tra le variabili. Essendoci sia variabili categoriche che binarie, è stata utilizzata la tecnica statistica *CATPCA* mediante la quale è stato possibile calcolare il coefficiente di correlazione di *Pearson* anche per le variabili qualitative. La *CAPTCA,* acronimo di *Categorical Principal Component Analysis*, è un metodo di analisi in componenti principali non lineare per le variabili categoriche. Con *feature selection* ci si riferisce all'utilizzo di tecniche statistiche qualitative e/o quantitative volte a determinare quali siano le variabili maggiormente rilevanti ai fini della realizzazione di un modello efficace ed efficiente di *machine learning* o *deep learning*.

In questo contributo sono stati testati tre algoritmi di *clustering* gerarchico aggregativo ossia il metodo del legame completo, il metodo del legame medio e il metodo del legame singolo.

I risultati del processo di *clustering* sono stati comparati e valutati utilizzando il *PSFM* (*Pseudo F Index based on the Mutuability)* e il *WCM* (*Within-Cluster Variability*)

rispettivamente lo pseudo indice F basato sulla mutabilità e la variabilità interna di ogni *cluster* (Rezankova, 2011).

In termini di misure utilizzate per le distanze, figurano la distanza *ES* di Eskin et al. (Eskin, 2002), la distanza *IOF (Inverse Occurence Frequency)* di Sparck-Jones (Sparck Jones, 1972), la distanza di *LIN* di Lin (Lin, 1998), la distanza di *VE* (*Variability Explained*) di Šulc (Šulc, 2016) e il *Simple Matching*.

In base alla combinazioni di metodi e distanze, il numero ottimale di *cluster* varia tra due e quattro. Sulla base delle valutazioni degli autori le soluzioni più promettenti sono tre.

La prima si basa su tre *cluster* con la distanza *ES* e il metodo del legame completo. Successivamente, due o tre *cluster* con la distanza *IOF* o *LIN* e il metodo del legame completo ed infine due *cluster* con la distanza *ES* e il metodo del legame medio (Cibulková, 2018).

Il contributo di Brito et al. raggruppa una *clustering analysis* con delle tecniche di *data mining*. Nello specifico, il lavoro è stato realizzato nel campo del *fashion*, lo stesso per il quale è stata realizzata l'analisi elaborata nei successivi capitoli.

Il *dataset* utilizzato da Brito et al. è di bivolino.com ossia un produttore artigianale di camicie. Le feature contenenti sono state raggruppate nelle seguenti categorie: caratteristiche di prodotto, caratteristiche demografiche e biometriche, geografiche, psicografiche e comportamentali.

Le prime hanno come variabili il materiale, il colore e il colletto. Le caratteristiche demografiche e biometriche sono il sesso, l'età, la grandezza del colletto e l'indice di massa corporea. Il gruppo delle variabili demografiche invece si caratterizza per essere costituito dal paese in cui i clienti vivono. Infine, per le caratteristiche psicografiche vi è la tipologie di lifestyle mentre, per le caratteristiche comportamentali, la sensibilità al prezzo.

L'algoritmo di *clusterizzazione* utilizzato è il *K-Medoids* che a differenza del *K-Means*, riportando quando definito dall'autore, può non essere applicato solo con variabili numeriche.

Gli output dell'analisi dei *cluster* sono stati sei *cluster* di clienti per quanto concerne 29 caratteristiche di prodotto. Nello specifico, la maggior parte dei dati di prodotto riguardavano delle camicie (circa 10775 in totale) descritte da 25 variabili categoriche e 4 numeriche). Inoltre, gli autori precisano che per le variabili numeriche, in fase di descrizione dei *cluster*, è stata utilizzata la media mentre, per le categoriche, la moda.

L'analisi di Brito et al. si espande poi con l'algoritmo *CN2-SD* con il quale sono stati identificati i sottogruppi di clienti considerando le altre caratteristiche (demografiche, psicografiche, geografiche, comportamentali e del *lifestyle*). Il modello è riuscito a trovare sei sottogruppi. Secondo Kloesgen e Wrobel, l'algoritmo *CN2-SD* è una delle tecniche di *data mining* per la scoperta di sottogruppi. Esso può essere definito come un adattamento dell'algoritmo *CN2* il quale si basa su delle regole di classificazione strutturate "*if Cond then* $Class_{value}$". Nella sua forma originaria, il modello si caratterizza per essere predittivo mentre nel suo adattamento (*CN2-SD*) per la scoperta di sottogruppi cambia, diventando appropriato per un task descrittivo (Brito, 2015).

Il lavoro di Yadegaridehkordi et al. del 2021 riguarda l'industria del turismo in Malesia, nello specifico quella alberghiera considerando esclusivamente gli hotel eco-friendly. L'obiettivo è quello di segmentare i loro clienti capendone le esigenze, le preferenze e i comportamenti e aiutando così sia le aziende alberghiere sia il policy maker per la realizzazione di politiche mirate all'espansione di questo genere di attività più attente all'ambiente. L'analisi ha riguardato le recensioni online presenti su TripAdvisor per ogni singolo hotel eco-friendly. Per quanto concerne la parte di *clustering* dello studio, è stato approfondito ed utilizzato solo l'algoritmo non gerarchico *K-Means*. Per la valutare l'algoritmo e definire il numero di *cluster* ideale è stato utilizzato il metodo della Silhouette determinando il numero di *cluster* ideale pari a 5. Le feature utilizzate per l'analisi sono in totale sette ossia "*check-in/front desk*", "*value* (costo-beneficio)", "*location*", "*sleep quality*", "*rooms*", "*cleanliness*" e "*service*". Oltre alla *cluster analysis*, il paper si espande con una parte di *supervised learning* in cui è stato applicato il *CART* (*Classification and Regression Trees*) per predire e classificare le preferenze dei clienti. Per la parte di *data* visualisation, è stata applicata anche una analisi in componenti principali affinché si riuscisse a visualizzare i dati multidimensionali su uno scatterplot bidimensionale. Infine, sono stati menzionati altri approcci come *Latent Semantic Analysis* (LSA) e la *Linguistic Inquiry Word Count* (LIWC) per sviluppi futuri al fine di integrare l'analisi delle recensioni (Yadegaridehkordi, 2021).

Nella ricerca di Purnomo et al. del 2020 si discute dell'applicazione del *machine learning* con tecniche di *clustering* per determinare strategie di *marketing* efficaci. Nello specifico sono stati utilizzati dei dati di un'azienda di arredamento dal 2016 al 2018 la quale ha sia negozi online che offline. Le metriche utilizzate per l'analisi sono quelle costitutive del metodo RFM (*Recency*, *Frequency* e *Monetary Value*). Il *dataset* su cui è stata realizzata

l'indagine è composto da 30240 osservazioni mentre per quanto concerne le variabili queste si dividono in due categorie: i dati dei clienti e dei prodotti nonché i dati delle transazioni. Il numero di clienti registrati è di 192 e le tipologie di prodotti 37. La tabella utilizzata per l'*RFM* contiene le seguenti variabili *transaction_id*, *invoice_id*, *product_code*, *customer_id*, *product_name*, *quantity*, *invoice_date* e *unit_price*. Sulla base di questi parametri sono stati calcolati *Recency*, *Frequency* e *Monetary Value*. *Recency* è stata calcolata sottraendo la data dell'acquisto più recente alla data attuale. La valore *Frequency* invece è stato ottenuto sommando il numero di transazioni per ogni cliente e infine, il *Monetary Value*, sommando il prezzo dei prodotti acquistati da ogni cliente.

L'algoritmo di *clustering* preso in considerazione nel lavoro è il *K-Means* il quale prevede che venga definito il numero di *cluster* a priori. In questo contesto, è stato scelto *l'elbow method* basato sulla somma degli errori quadratici (SSE) il quale ha determinato che il numero di *cluster* migliore possa essere 4. Lo studio in questione non riporta metriche utilizzate per la valutazione del modello di apprendimento automatico. Nonostante ciò, nella parte conclusiva, viene confermato che la *cluster analysis* con il *K-Means* può essere uno strumento utile per affiancare il lavoro dei *marketing manager* nella fase di segmentazione e di definizione della parte strategica (Yoseph, 2018).

La ricerca accademica di Bellini et al. riguarda la realizzazione di un *recommendation system* per un'azienda nell'industria del *fashion* ossia *Tessilform* proprietaria del marchio Patrizia Pepe. Nonostante l'obiettivo ultimo sia un recommendation system, sono stati applicati diversi algoritmi di *clustering*. Gli algoritmi di *clustering* hanno segmentato i clienti affinché i dati fossero organizzati per porre in essere raccomandazioni personalizzate. Il fine ultimo del lavoro è quello di migliorare la costruzione di relazioni di lungo termine con i clienti aumentandone anche la redditività mediante delle raccomandazioni personalizzate e in linea con il concetto di *customer centricity*. In altre parole, l'intento è quello di migliorare l'esperienza dell'utente affinché, aumentandone la soddisfazione, ci sia un ritorno di natura economica per l'azienda. Nel corso della trattazione sono stati citati diversi algoritmi di *machine learning*, utilizzati per la *clustering analysis*; questi sono il *K-Means*, *K-Medoids*, *Clara* e *Self-organizing map* (SOM). L'algoritmo *Clara* è un'estensione del *Partitioning Around Medoids* (*PAM*) adattato a dei *dataset* di dimensione elevata. Invece, il *SOM* è un algoritmo di *clustering* di *deep learning* quindi basato sulle c.d. reti neurali artificiali. Per quanto concerne le

variabili in oggetto, per la componenti di clusterizzazione sono state considerate feature demografiche, la *recency*, la *frequency*, il *monetary value*, il *life time value* (*LTV*) e alcune combinazioni di queste. Per la valutazione delle performance dell'analisi di *clustering* è stato utilizzato il metodo della *silhouette,* il quale è stato considerato anche per la scelta del numero di *cluster* a priori. Il numero di *cluster* ideale in questo lavoro è stato 14 (Bellini, 2021).

Il lavoro "*Clustering Consumers Through their Consumption Behavior: Analysis on the Fashion Industry*" di Puiu A. riguarda un'analisi di *clustering* nell'industria della moda; non sono state specificate aziende in particolare. L'obiettivo dello studio è quello di stimare i pattern di consumo degli individui di età compresa tra 18 e 45 anni in Romania mediante una scala a 6 dimensioni dell'inventario degli stili di consumo sviluppata da Sproles e Kendall (1986). Prima di proseguire con l'analisi dei modelli di *machine learning*, il lavoro si è caratterizzato per l'applicazione di un'analisi fattoriale esplorativa e un'analisi fattoriale confermativa per validare le dimensioni della scala. L'algoritmo di *clustering* utilizzato è il *K-Means* definendo il numero ideale di *cluster* attraverso *l'elbow method*, il *silhouette method* e il *gap statistic*. In tal modo, il numero di *cluster* risultante è stato otto mentre la selezione delle feature ha determinato che, nel modello, ci fossero variabili demografiche, variabili relative al coinvolgimento monetario nell'acquisto di articoli di abbigliamento nonché la situazione finanziaria dei consumatori. Per determinare l'associazione tra i *cluster* e misurare la distanza tra le osservazioni è stata usata la distanza Euclidea mentre per misurare le differenze tra i gruppi di consumatori identificati è stata realizzata un'analisi della varianza. In ultima istanza, è stata anche realizzata un'analisi della discriminante per poter definire quale variabile discriminava in misura maggiore i *cluster* di individui trovati (Andreea-Ionela, 2020).

Il lavoro di Wang si basa sull'utilizzo del *deep learning* per segmentare i clienti in ottica di *digital marketing*. L'approccio utilizzato nello studio è quello di *swarm intelligence* modificato e chiamato "*Modified Social Spider Optimization*" mediante il quale sono state selezionate le caratteristiche. Al fine di clusterizzare i clienti, l'algoritmo di *deep learning* è il *Self Organizing Nuerla Network*. I clienti sono stati clusterizzati sulla base dei modelli di acquisto e della quantità spesa. Partendo da sei features, la clusterizzazione basata sulle variabili riconducibili ai *pattern* e comportamenti di acquisto ha determinato sei *cluster*. Successivamente, sulla base dei dati presenti in ogni *cluster*, è stata realizzata una segmentazione mediante *DNN* (*deep neural network*) e risultando in cinque segmenti.

Gli algoritmi sono stati valutati sulla base di metriche e aspetti diversi. L'algoritmo di *clustering SONN* è stato analizzato con il metodo degli errori quantitativi e topologici. Nonostante ciò, il modello DNN è stato valutato con l'accuratezza (Wang, 2022).

"*Market segmentation for profit maximization using machine learning algorithms*" di Janardhanan, S. e Muthalagu, R. si caratterizza per segmentare i clienti sulla base di variabili relativi alle vendite, in altre parole riconducibili al comportamento d'acquisto. L'algoritmo utilizzato per clusterizzare i clienti è il *K-Means*. Oltre ad algoritmi di *clustering*, sono state utilizzate tecniche statistiche quali *ARIMA* (*Auto Regressive Moving Average*) per il forecasting delle vendite, l'analisi della correlazione per capire l'impatto di ogni variabile con le vendite. L'algoritmo ha generato quattro *cluster* partendo da un *dataset* contenente le vendite settimanali, i numeri di prodotti (91 prodotti diversi) e le 45 filiali diverse. La *clustering analysis* in questo caso ha aiutato a creare dei segmenti di clienti per definire il loro comportamento d'acquisto colmando la mancanza di informazioni sui clienti presente nel *dataset* (Janardhanan, 2020).

"*Customer Clustering Using The K-Means Clustering Algorithm in Shopping Mall in Indonesia*" di Jerry Heikal e Iksan Adityo Mulyo e realizza un'analisi sui consumatori di un centro commerciale in Indonesia. Nello specifico è stato utilizzato un algoritmo di *machine learning* per la *clustering analysis* di clienti in base a età, sesso, reddito annuo, *speding score*, stato civile, posizione di residenza, luogo di lavoro, interesse per il prodotto, social media e figli. Il *K-Means* è stato il modello di apprendimento automatico utilizzato, determinando la divisione in tre *cluster*. Il lavoro propone poi delle iniziative e dei cambiamenti in termini di *marketing* operativo per ogni *cluster*. Non sono stati discusse tecniche di valutazione dei modelli tantomeno altre tecniche statistiche utilizzate durante lo studio (Mulyo, 2022).

"*Multi Clustering Recommendation System for Fashion Retail*" di Bellini et al. riguarda il settore della rivendita al dettaglio nell'industria del *fashion*. L'elaborato propone un *recommendation* system per i negozi di moda basato su una *clustering analysis*. Così come il lavoro "*Fashion retail recommendation system by multiple clustering*" del 2022, anche qui, lo studio è stato realizzato con un'azienda nel settore della moda ossia *Tessilform*, società avente il marchio Patrizia Pepe. La validazione dello studio è avvenuta sia online che offline. Lo studio si caratterizza per avere due clusterizzazioni; una prima per quanto concerne le descrizioni degli item e una seconda per i clienti. La *cluster analysis* per le descrizioni degli item è stata realizzata su un *dataset* contenente 50000

item mediante l'algoritmo *K-Medoids* con *Gower distance*. Il metodo per definire il numero di *cluster* ottimale è stato il *Silhouette method* determinando 13 *cluster* ideali. L'analisi dei *cluster* per i clienti è stata realizzata partendo da una parte di *feature engineering* affinché potessero essere ottenute le feature per il metodo *RFM*. Il numero di osservazioni per questa parte del lavoro è di 608447 di cui, una parte, è stata raccolta tra il 2016 e il 2019. Un aspetto interessante del contributo in questione è che alcune *feature* come il sesso, l'età, il family status, il livello di fidelizzazione, il paese e la città di provenienza, non sono state utilizzata per ragioni di incompletezza di informazioni e valori mancanti. Il *Silhouette method* ha determinato cinque *cluster* come valore ottimale. Per clusterizzare i clienti è stato utilizzato un algoritmo *K-Means* con distanza euclidea (Bellini P., 2022).

Il contributo di Namvar et al. riguarda la clusterizzazione di 491 clienti di una banca iraniana considerati in base a 25 feature in parte demografiche e in parte ottenute dal metodo RFM nonché dalla computazione del *Life Time Value*. Il lavoro è partito da due principali *dataset*: uno chiamato "*Customer's transactions*" e un altro "*Customer's Profiles*". Dal primo sono stati ottenute tutte le variabili riconducibili al metodo RFM le quali, insieme alle variabili demografiche sono state oggetto della *clustering analysis* mediante *K-Means*. L'approccio utilizzato si differenzia dai precedenti poiché sono stati applicati due algoritmi *K-Means*. Un primo esclusivamente sulle variabili RFM e un secondo, come continuazione, per le variabili demografiche. In aggiunta al *two-phase clustering* è stata realizzata una classificazione utilizzando un algoritmo di *deep learning* per ottenere la predizione del LTV. Sulla base di questi risultati sono stati definiti dei profili di clienti con le quali sono state definite delle strategie di *marketing*. Con la prima clusterizzazione, i clienti sono stati divisi in tre *cluster*. Invece, con la seconda, per ogni *cluster* risultante dalla prima fase, è stata realizzata la seconda analisi di *clustering* determinando nove *cluster* (Namvar, 2010).

L'elaborato "*Developing Marketing Personas with Machine Learning for Educational Program Finder*" di Koponen M. è l'unico tentativo sebbene solo teorico di dimostrate come gli algoritmi di *machine learning* possano essere di supporto ai *marketing manager* nella realizzazione delle *marketing personas* funzionali alla strutturazione della strategia di *marketing* per l'impresa. Lo studio parte da un'iniziale descrizione dei concetti di dominio come le *personas*, le *marketing personas*, il *machine learning*, i sistemi di raccolta dati per variabili comportamentali necessarie per la realizzazione di una

marketing personas, ecc. Il contesto di riferimento per il contributo in questione è il settore dell'istruzione di livello universitario. Un ulteriore elemento di differenziazione in Koponem è la spiegazione teorica di ogni singolo passaggio del processo per realizzare *marketing personas* con l'apprendimento automatico, partendo dalla raccolta dei dati per le feature per arrivare all'applicazione degli algoritmi di *machine learning* (Koponen, 2017).

Il lavoro di An et al. è un tentativo di applicazione del *machine learning* per la realizzazione di *clustering analysis* su gli utenti dei *social media* di AJ+ ossia un canale *online* di *news* di proprietà di Al Jazeera Media Network. Nello specifico, sono stati uniti dei dati raccolti su tre principali social media: Twitter, YouTube e Facebook. Considerando i seguaci dei diversi canali di comunicazione sono stati collezionati mediante *data* mining le seguenti caratteristiche per ogni unità statistica.

Da Twitter sono stati considerati i dati relativi a interessi e punti di vista, analizzando cos' circa 195000 utenti.

I ricercatori hanno utilizzato YouTube per collezionare altri dati relativi a interessi e demografiche. In questo caso, sono stati considerati 188000 account.

Su Facebook, i ricercatori hanno raccolto centinaia di migliaia di link condivisi da oltre 54.000 follower sul social media, esaminando in particolare i domini condivisi da questi utenti.

L'unione di queste informazioni è stata utilizzata per la realizzazione di un algoritmo che fosse in grado di generare *personas* in tempo reale considerando l'età, il genere, il paese di origine, gli interessi, le quote relative alla pagina giornalistica e i video più visti.

L'algoritmo di *clustering* utilizzato è *il K-Means++* esclusivamente sui dati raccolti su Facebook. Questa variante del *K-Means* ne migliora la scelta del seme iniziale. Per la scelta del numero di *cluster* ideale, è stato applicato l'elbow method con la percentuale di varianza spiegata. Il risultato è stato la generazione di sette *cluster*. Partendo dalla definizione dei *cluster*, sono state poi definitele altre caratteristiche demografiche e di interessi ottenute dagli altri social (An, 2016).

Il contributo di Yan et al. intitolato "*Customer segmentation using real transactional data in e-commerce platform: A case of online fashion bags shop*" riguarda l'industria del *fashion* e ha come obiettivo quello di clusterizzare i clienti in base a un *dataset* costituito da feature relative alle transazioni effettuate da ogni cliente. Esempi di feature utilizzate sono state il discount applicato all'ordine, il valore massimo pagabile dal cliente per

transazione, il valore minimo pagabile dal cliente per transazione, il valore medio pagabile dal cliente per transazione, il ritardo nella consegna, il numero di prodotti in un singolo acquisto, il tipo di prodotto, la città del cliente, la percentuale di resi, l'ammontare speso, ecc. Affinché fossero selezionate solo le variabili maggiormente rilevanti è stata realizzata un'analisi di correlazione determinando la decisione di prendere in analisi solo 13 variabili. L'algoritmo di *clustering* utilizzato è il *Fuzzy C-Means* il quale ha determinato la presenza di tre *cluster* e a cui è seguito un algoritmo di classificazione di *deep learning* per predire il tipo di cliente (Yan, 2021).

Il lavoro di Rodrigues e Ferreira riguarda la realizzazione di un recommendation system ibrido in cui è stata realizzata anche una *clustering analysis*. L'industria per la quale il lavoro trova applicazione è quella dei profumi. Nello specifico, il contributo è stato realizzato, per la parte di *clustering*, con un *dataset* costituito da 3245 clienti di catene di profumerie le cui osservazioni risalgono ad un periodo che intercorre tra il 2012 e il 2013. In questo caso le feature di riferimento sono state ottenute dal metodo RFM e dal calcolo del LTV. Per la valutazione della *cluster analysis* sono stati utilizzati due indici ossia il *total Within Sum of Squares* (*WSS*) e l'indice *Calinski-Harabasz*. L'algoritmo utilizzato è il *K-Means* che ha posto in essere tre *cluster* (Rodrigues, 2016).

Un altro contributo rilevante ai fini di questo studio è il lavoro di Tu et al. "*Using cluster analysis in persona development*". Nello specifico, si tratta di una descrizione elaborata del processo di creazione di una Persona mediante sia tecniche quantitative sia qualitative. Per le prime ci si riferisce alla *clustering analysis*, ad esempio, mentre per le successive si possono citare le osservazioni e le interviste. Sebbene sia un contributo prettamente teorico e poco applicativo in termini di algoritmi, presenta una spiegazione ancora esaustiva e precisa del lavoro nel suo complesso qui realizzato. Oltre alle tecniche specifiche, sono state considerate anche tecniche statistiche secondarie come l'analisi in componenti principali (Tu, 2010).

Un ulteriore studio realizzato con l'obiettivo di realizzare le *marketing personas* è di Quan et al. intitolato "*Research on the construction of Personas model based on K-Means*". Nonostante l'algoritmo utilizzato sia già stato definito leggendo il titolo, il lavoro approfondisce una tecnica vista per la prima volta finora per i lavori considerati precedentemente. Nello specifico i dati dei diversi utenti sono stati raccolti mediante tecniche di *web scraping* mediante *Python* sul social asiatico *Weibo* in questo caso. Il *dataset* ultimo era costituito da 32,915 unità statistiche rappresentanti ognuna uno user.

Il numero di *cluster* ottenuto è pari a 5 per i quali sono state definite le *personas* in base ad un'analisi descrittiva di ogni variabile per singolo *cluster* (Quan, 2015).

Lo studio "*Personas: a market segmentation approach for transportation behavior change*" di Arian et al. prende in considerazione il settore del trasporto dei consumatori segmentandoli in base ai modelli di viaggio e alla sensibilità agli incentivi. L'obiettivo ultimo è quello di creare delle *personas*. L'algoritmo di *machine learning K-Means* ha generato dieci *Personas* analizzando apriori il numero di *cluster* mediante il punteggio silhouette. Le variabili utilizzate per il contributo sono state principalmente di due categorie: sociodemografiche e le traiettorie GPS.

Altre tecniche statistiche utilizzate sono state il *DBSCAN*, il DTW e il *signal processing methods*. Nello specifico queste hanno reso possibile l'estrazione di diversi attributi per ogni singolo utente. Il *dataset* di riferimento conteneva le osservazioni di 24 utenti (Arian, 2021).

Il lavoro di Xin et al. intitolato "*Building up Personas by Clustering Behavior Motivation from Extreme Users*" propone la generazione di *personas* partendo da dati raccolti medianti metodologie quali le osservazioni e le interviste in profondità. In particolare, le variabili sono principalmente di tipo comportamentale. Il risultato finale è stato la generazione di cinque *personas*. La metodologia di *clustering* scelta è di tipo manuale e, nello specifico, è stato utilizzato il modello *Goodwin*. Oltre al *clustering*, è stata realizzata una *behavioral dimension analysis* (Xin, 2022).

Di seguito, è stata riportata una sintesi dello stato dell'arte finora elaborato (Tabella 2).

Titolo	Autori	Features	Dataset	Dominio	Tecnologia	Risultati
Using Customer Knowledge Surveys to Explain Sales of Postgraduate Programs: A Machine Learning Approach	Asensio, E. et al.	Descrittive, Occupazionali, Economiche, Comportamentali, Geografiche	16.272 form risposti	Istruzione	K-Medoids con distanza Gower	Tre cluster
A case study of customer segmentation with the use of hierarchical cluster analysis of categorical *data*	Cibulková, J. A. N. A., & Sulc, Z.	Descrittive, Geografiche, Comportamentali, Economiche	5755 viaggi dei clienti	Viaggi e Intrattenimento	Clustering gerarchico aggregativo (legame completo, medio e singolo)	2-4 cluster a seconda delle distanze e del legame.

Customer segmentation in a large database of an online customized fashion business	Brito, P.Q. et al.	Descrittive, Demografiche, Biometriche, Psicografiche, Comportamentali	10775 righe, 29 variabili categoriche e numeriche	Fashion	K-Medoids e CN2-SD	6 cluster e 6 sottogruppi
Customers segmentation in eco-friendly hotels using multi-criteria and machine learning techniques	Yadegaridehkordi, E. et al.	Descrittive, Economiche	Recensioni online su TripAdvisor per ogni singolo hotel eco-friendly	Alberghiero/turismo	K-Means	Cinque cluster
Segmenting retail customers with an enhanced RFM and a hybrid regression/clustering method	Yoseph, F., & Heikkila, M.	Anagrafiche, Descrittive, Economiche, Comportamentali	30240 osservazioni, 192 clienti registrati, 37 tipologie di prodotto	Retail	K-Means	Quattro cluster
Fashion retail recommendation system by multiple clustering	Bellini, P. et al.	Demografiche, Economiche, Comportamentali	-	Fashion	K-Means, K-Medoids, Clara, Self-organizing map (SOM)	14 cluster
Clustering Consumers Through their Consumption Behavior: Analysis on the Fashion Industry	Puiu A.	Descrittive, Comportamentali, Economiche	Non specificato	Fashion	K-Means	Otto cluster
Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach	Wang, C.	Comportamentali, Economiche	-	Digital Marketing	Self-Organizing Neural Network (SONN), Deep Neural Network (DNN)	6 clusters e 5 segmenti
Market segmentation for profit maximization using machine learning algorithms	Janardhanan, S., & Muthalagu, R.	Economiche	Vendite settimanali, 91 prodotti, 45 filiali	-	K-Means	Quattro cluster
Customer Clustering Using The K-Means Clustering Algorithm in shopping mall in Indonesia	Jerry Heikal e Iksan Adityo Mulyo	Demografiche, Economiche, Comportamentali	-	Retail	K-Means	Tre cluster
Multi Clustering Recommendation System for Fashion Retail	Bellini, P. et al.	Demografiche, Economiche, Comportamentali, Geografiche	50000 item, 608447 osservazioni	Fashion	K-Medoids con distanza Gower ed Euclidea	13 cluster/ 5 cluster

A two-phase clustering method for intelligent customer segmentation	Namvar, M., Gholamian, M. R., & KhakAbi, S.	Demografiche, Comportamentali	25 feature, 491 clienti	Banking	K-Means (two-phase clustering) - Deep Learning	3 cluster/ 9 cluster
Developing Marketing Personas with Machine Learning for Educational Program Finder	Koponen M.	Comportamentali	-	Istruzione	-	Lavoro teorico
Towards automatic persona generation using social media	An, J. et al.	Comportamentali, Descrittive	Dati raccolti su Twitter, YouTube e Facebook	Informazione e Intrattenimento	K-Means++	Sette cluster
Customer segmentation using real transactional *data* in *e-commerce* platform: A case of online fashion bags shop	Yan, Z., & Zhao, Y.	Comportamentali, Geografiche, Descrittive, Logistiche, Economiche	-	Fashion	Fuzzy C-Means	Tre cluster
Product recommendation based on shared customer's behaviour	Rodrigues, F., & Ferreira, B.	Comportamentali, Economiche	3245 clienti	Cosmetica	K-Means	Tre cluster
Using cluster analysis in persona development	Tu, N. et al.	Descrittive, Comportamentali, Psicografiche	-	-	-	-
Research on the construction of Personas model based on K-Means Clustering Algorithm	Quan, C., Liu, D., & Zhou, L.	Descrittive, Psicografiche	32,915 unità statistiche	Social Network - Weibo	K-Means	Cinque cluster
Personas: a market segmentation approach for transportation behavior change	Arian, A., Meiyu Pan, M., & Chiu, Y. C.	Sociodemografiche, Geografiche	24 utenti	Trasporto	K-Means	Dieci Personas
Building up Personas by Clustering Behavior Motivation from Extreme Users	Xin, X. et al.	Comportamentali	-	-	Manuale (modello Goodwin)	Cinque Personas

Tabella 2 Overview dello stato dell'arte

4 Research Gap e Metodologia

Sulla base dei contributi riportati si evincono diverse mancanze sotto molteplici profili. In primo luogo, analizzando l'attuale letteratura da un punto di visto macro, l'industria del *fashion* presenta un ridotto numero di elaborati sull'utilizzo del *machine learning* e della *clustering analysis* sia in generale che per lo sviluppo di *marketing personas*. Allo stesso modo, la quantità di studi volti a generare delle *buyer personas* utilizzando algoritmi di apprendimento automatico o apprendimento profondo è davvero limitata. I pochi casi riscontrati in tal direzione sono stati applicati su delle *feature* di natura diversa rispetto a quella transazionale come, ad esempio, in "*Towards automatic persona generation using social media"* di An, J. et al. Come già discusso, lo studio presenta la realizzazione di *marketing personas* basandosi su dati ottenuti da *social media* come gli interessi e altre metriche di riferimento (An, 2016).

In aggiunta, il numero di *feature* utilizzate è sempre inferiore a quello necessario per la realizzazione di una *marketing personas* così come definita da Cooper e, di conseguenza, maggiormente utile in ottica di *business*. Una *marketing persona* ha come aspetto distintivo esattamente quello di essere un'accurata descrizione di un archetipo di uno o più segmenti. La peculiarità appena definita è l'elemento che effettivamente determina un distacco tra *customer segmentation* e *buyer persona*.

Sviluppare le *personas* mediante *machine learning* è un obiettivo ambizioso e che necessiterà di innumerevoli tentativi e studi prima che venga raggiunto un risultato realmente applicabile in questa direzione nei contesti aziendali. Lo studio presentato di seguito è un primo tentativo per l'industria del *fashion*, ambiente che così come gli altri, si caratterizza per aver visto delle *personas* basate esclusivamente su delle idee e delle esperienze proprie dei *marketing manager* nel corso degli ultimi decenni.

Poter affiancare ad un lavoro da anni prettamente manageriale, degli algoritmi di apprendimento automatico per generare dei risultati precisi circa gli archetipi *target*, è la l'obiettivo dello studio. Di conseguenza si può derivare la domanda di ricerca:

"È possibile utilizzare il machine learning per la creazione di marketing personas nell'industria del luxury fashion?".

Lo studio è un'analisi realizzata con *Python*, tra i linguaggi di programmazione maggiormente utilizzati in ambito di *data science*. Nello specifico, grazie a diverse librerie di codice, funzionalità del linguaggio e algoritmi di intelligenza artificiale è stato

posto in essere un contributo che attraversa diverse analisi statistiche, tecniche di esplorazione e aggregazione dei dati. Lo *script* è stato inserito in Appendice.

Ai fini dell'obiettivo presentato, sono stati individuati due *dataset* contenenti svariati dati transazionali, demografici, comportamentali, ecc. dei clienti di una famosa azienda internazionale nel settore del *luxury fashion*. Queste osservazioni potranno generare una risposta sufficientemente pertinente per la domanda di ricerca definita precedentemente. Come si approfondirà di seguito, una *marketing persona* è costituita da determinate variabili dalle quali non si può prescindere altrimenti risulterebbe essere carente in termini di valore apportato nonché significato nei contesti dove è utilizzata. Di conseguenza, la realizzazione del *task* è stata fortemente condizionata dalla presenza di specifiche *feature* nel *set* di dati. In aggiunta, anche la scelta dell'azienda dalla quale sono stati ricevuti i dati è stata fortemente limitata dall'*industry* selezionata siccome soggetta a delle dinamiche di acquisto proprie come specifici *decision journey*, *touching point*, *stakeholder*, approcci comunicativi e così via. È evidente che un prodotto con un prezzo superiore al prezzo medio di mercato per la medesima tipologia di prodotto così come una scelta dei materiali, in termini di qualità e ricercatezza, superiore e delle strategie di acquisizione clienti focalizzate su delle leve di altra natura rispetto alle controparti non *luxury* determinino la necessità di uno studio mirato per il caso specifico.

5 Exploratory Data Analysis

I *dataset* su cui è stato realizzato lo studio sono di una nota azienda italiana nel settore del *luxury fashion* con quasi cento anni di storia. Oltre alla vendita di abbigliamento esclusivamente per uomo, si occupa anche della produzione dei capi. Circa l'80% dei capi di abbigliamento sono esportati in diversi paesi in tutto il mondo. I prodotti realizzati dall'azienda sono sia capi con misure standard sia capi su misura "*Made to Measure*".

L'insieme dei dati utilizzato per questo contributo si divide in due tabelle. La prima, chiamata *Transactions Details*, riguarda dati transazionali, di conseguenza, ogni riga rappresenta una transazione. Il secondo *set* di dati, chiamato *Contact Active*, comprende informazioni su clienti pertanto, ogni riga, rappresenta un acquirente considerato come singolo. Le due tabelle hanno come *primary key*[1] il *customer ID*.

Transactions Details ha una *shape* pari a (1021739, 65) ed è bidimensionale; le variabili contenute sono 38 qualitative e 27 quantitative (16 *float64* e 11 *int64*). Il numero di *missing value* cambia in base alla variabile considerata; ad esempio, il 95,73% dei valori era mancante nel caso della variabile *loyalty_membership_no*, mentre, per altre variabili, come la *primary key* "*customer*" non era presente nessun dato mancante.

Contact Active ha una *shape* formata da 384352 righe e da 135 colonne; anche questo *set* di dati è bidimensionale. Le variabili presenti sono in parte qualitative e in parte quantitative, rispettivamente 72 e 63 di cui 50 *float64*[2] e 13 *int64*[3]. Così come per *Transactions Details*, anche per *Contact Active* mancano dei dati in diverse variabili; in particolare alcune variabili non contengono alcun dato ed è il caso di *cross-selling* mentre, anche in questo caso, non mancano dati per la colonna in comune ossia *customer_id*.

Dato l'ingente numero di variabili considerando entrambi i *set* di dati, saranno presentate solo quelle contenute all'interno del *dataframe*[4] oggetto degli algoritmi di *clustering*.

[1] Una *primary key* (o chiave primaria) è un concetto fondamentale dei *database* relazionali. Il termine indica un campo o un insieme di campi che identificano univocamente ogni record all'interno della tabella. Essa funziona come un identificativo unico per ogni osservazione del dataset in modo tale da garantire l'integrità dei dati e il collegamento attraverso diverse tabelle che contengono la medesima "chiave".

[2] Un *float64* è un tipo di dato utilizzato presente nella libreria *NumPy* il quale rappresenta numeri a virgola mobile in un formato 64 *bit*.

[3] Un *int64*, così come il *float64*, è un tipo di dato della libreria *NumPy* utilizzato per rappresentare numeri interi in un formato a 64 *bit.*

[4] Un *DataFrame* è una rappresentazione bidimensionale dei dati, simile a una tabella di un database con righe e colonne. Ogni colonna può avere un tipo di dato diverso (interi, stringhe, numeri in virgola mobile, ecc.) e ogni riga rappresenta un record o un'osservazione.

Tutte le tecniche statistiche utilizzate per arrivare ad avere il *dataset* descritto di seguito sono state riportate ed elaborate nel capitolo 6 che concerne la parte di *pre-processing*.
Il *dataset* utilizzato per l'analisi di *clustering* mediante *machine learning* è costituito da 146778 righe e 26 colonne. Ogni riga rappresenta un cliente e le variabili sono informazioni sullo stesso o sulle transazioni effettuate dal medesimo.
Le *feature* si dividono in 10 qualitative e 16 quantitative di cui la maggior parte sono *int64* e la minoranza *float64*, rispettivamente, 9 e 7. Per facilitare la trattazione delle prossime parti dell'elaborato, quest'ultimo *dataset* presentato sarà "*ca_updated*". Ai fini di una esaustiva descrizione delle variabili, verranno qui presentate con un grafico annesso laddove possibile.
La distribuzione della variabile *gender* (Figura 13) risulta essere non proporzionata e con un peso maggiore per la categoria "*Male*", in linea con il *target* dell'azienda a cui i dati si riferiscono siccome, come precisato in diverse paragrafi precedenti, si tratta di un'impresa di abbigliamento di lusso unicamente maschile. Come si evince dal relativo grafico, la *feature* presenta tre modalità *male*, *female* e *not declared*.

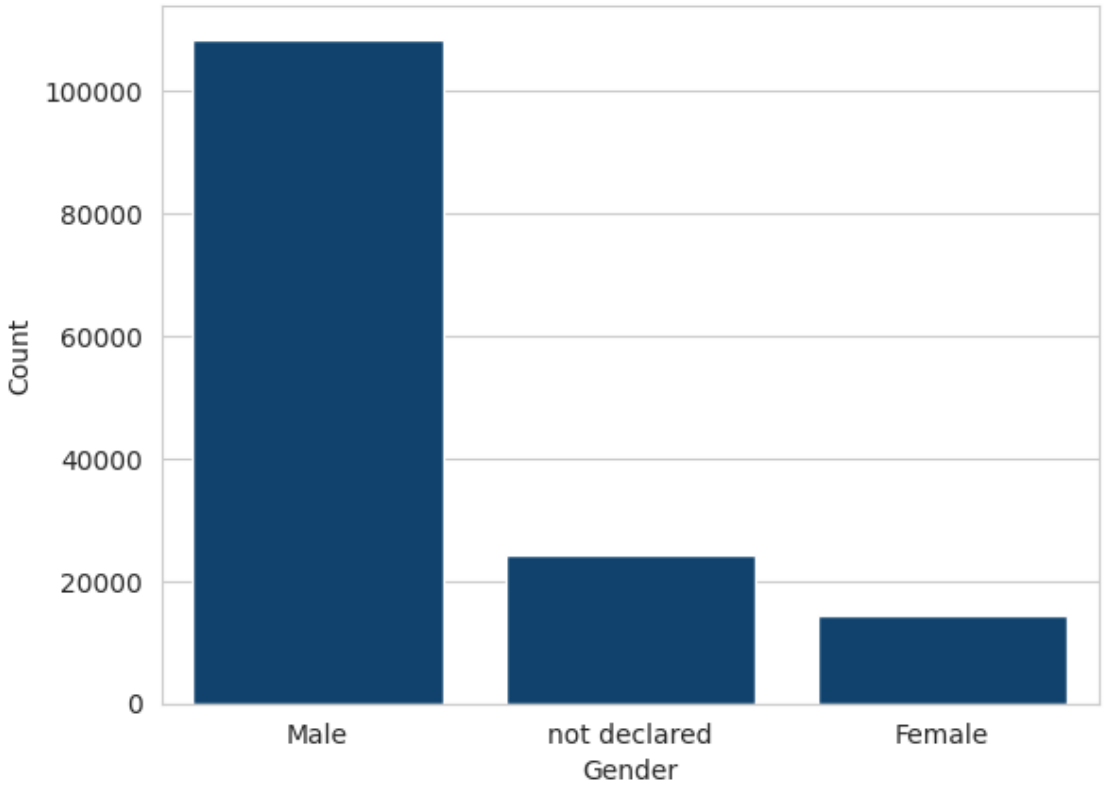

Figura 13 Grafico a barre per variabile gender

La variabile *Sales Habit* è tra le poche variabili comportamentali presenti nel *dataset* ed è costituita da tre modalità: *Regular*, *On* Sale e *Both* (Figura 14). La variabile qualitativa in questione evidenzia le abitudini di acquisto. Dal grafico riportato si evince come più del 40% dei clienti siano acquirenti di prodotti non in saldo. La differenza tra la percentuale di acquirenti *Both* e *On Sale* è decisamente inferiore rispetto a quella che si ha tra clienti *Regular* e *On Sale*. Tuttavia, il numero di consumatori che hanno acquistato prodotti in saldo (considerando *Both* e *On Sale*) è superiore a chi ha acquistato solo prodotti non scontati.

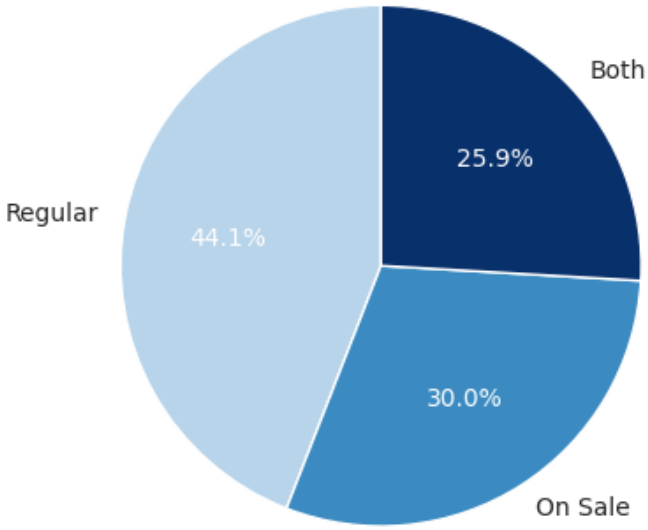

Figura 14 Grafico a torta per variabile sales habit

Customer Habit è una variabile di natura demografica e comportamentale; contiene informazioni circa la classificazione del cliente come residente, turista, viaggiatore oppure cliente internazionale. La sua distribuzione è stata riportata con la Figura 15. La modalità *not defined* indica quegli individui per i quali questa informazione risultava essere mancante prima della risoluzione dei *missing value*. Sono contenute cinque modalità; la più frequente è *Resident* con circa 80000 casi mentre i clienti internazionali sono una minoranza. In aggiunta, i clienti classificati come residenti superano tutti gli altri clienti considerati anche cumulativamente.

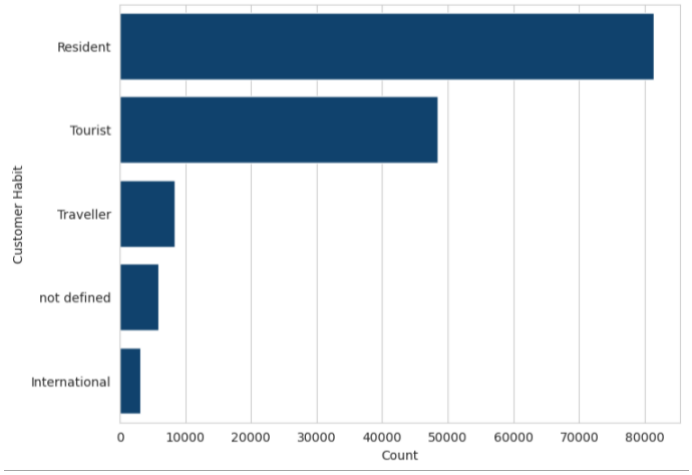

Figura 15 Grafico a barre per variabile customer habit

Le variabili legate alla *privacy* e alle preferenze di fruizione delle informazioni come *Preference Mail*, *Preference Telephone*, *Preference SMS* e *Preference E-mail* sono state inserite nel grafico a barre sottostante (Figura 16). La preferenza maggiore ricade sull'utilizzo delle e-mail mentre solo pochi acquirenti hanno votato "Sì" per indicare di ricevere comunicazioni via posta tradizionale. Un'altra tendenza interessante è che la comunicazione digitale scritta ossia *SMS* ed *e-mail* è tra quelle maggiormente preferite.
Il consenso è stato dato per una misura pari a circa all'80% del totale per il *Marketing*, *Newsletter* e *SMS*. Non sono state riscontrate notevoli variazioni tra gli strumenti di comunicazione per cui è stato misurato il consenso.

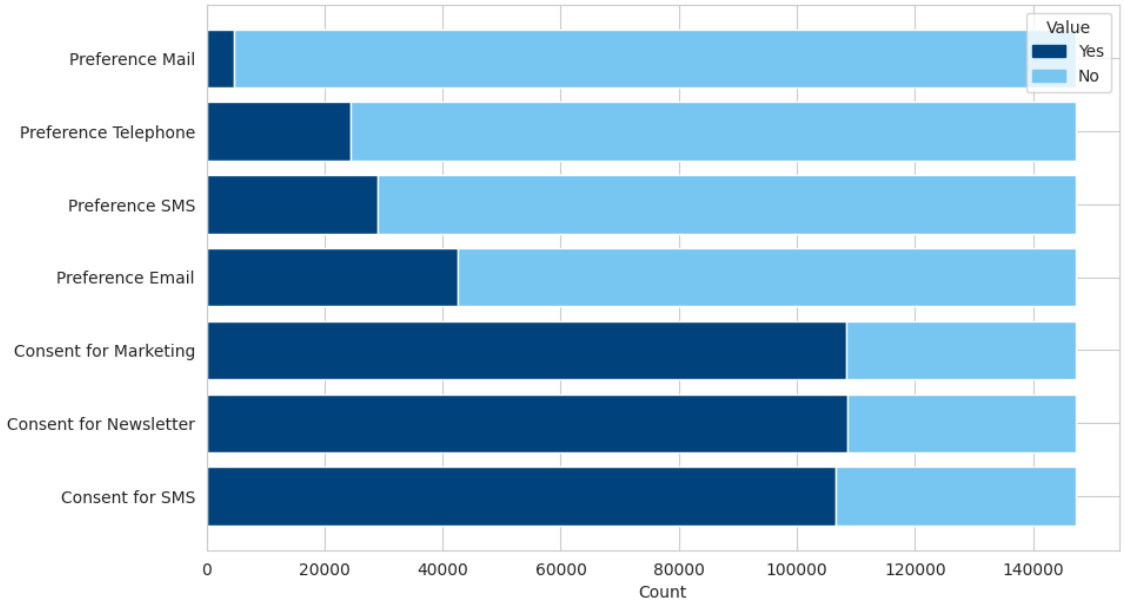

Figura 16 Grafico a barre per variabili su preferenza di comunicazione e consenso

La nuvola di parole riportata di seguito (Figura 17) è una rappresentazione grafica dei prodotti maggiormente acquistati dai clienti contenuti nella variabile *Top Products*. In questo caso la frequenza è stata visualizzata considerando la dimensione della parola; più sarà grande, maggiore sarà la frequenza. Si evince che le camicie sono il prodotto maggiormente scelto con il 18,93% del totale. Successivamente si posizionano le giacche (14,51%), gli abiti (10,27%), le cravatte (7,35%) e i pantaloni (7,32%). In totale, il numero di prodotti presenti in questa variabile è 35. Tra questi, 19 rappresentano meno dell'1% ciascuno sul totale. Un dettaglio interessante che si evidenza con questa analisi della variabile è che gli impermeabili, tra i prodotti di punta dell'azienda di abbigliamento a cui i dati appartengono, rappresentano solo il 0,20% del totale.

Figura 17 Nuvola di parole per variabile top products

Continuando l'analisi per quanto concerne le transazioni, il *dataset* contiene una variabile relativa ai colori dei prodotti maggiormente frequenti di cui sotto (Figura 18). La variabile categorica riportata di seguito è stata rappresentata con una *treemap*. I colori blu (31,16%) e bianco (14,78%) risultano essere quelli con ricorrenza più elevata in linea con i colori utilizzati in larga misura nel settore dell'abbigliamento più classico e formale generalmente inteso in cui, l'azienda qui considerata, difatti si posiziona. Le frequenze minori sono state registrate per il verde (3,12%) e il rosa (1,01%).

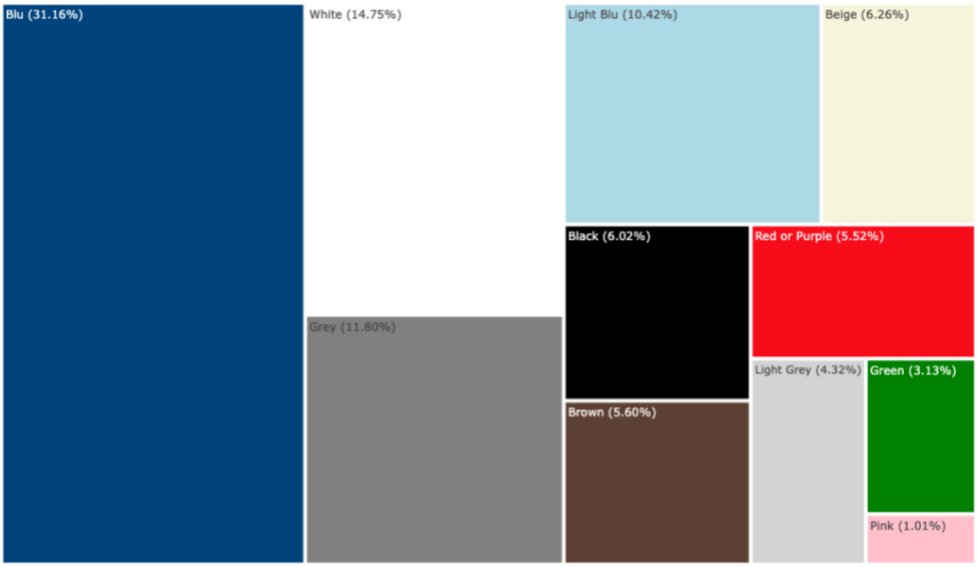

Figura 18 Treemap per variabile top colors

L'ultima variabile che è stata considerata per quanto concerne i prodotti acquistati è la variabile *top materials* (Figura 19) la quale contiene nove diverse modalità di materiali utilizzati nella produzione dei prodotti acquistati. Tra queste compaiono le fibre naturali e fibre sintetiche maggiormente utilizzate, fibre naturali e sintetiche considerabili di nicchia, la pelle, i metalli, i materiali esotici e tutto il resto. Ogni modalità raggruppa diversi materiali presenti nel *dataset* di partenza *Transactions Details*. Per ragioni computazionali e di interpretazione dei risultati sono state raggruppate come è stato descritto nel capitolo che concerne le parti di *preprocessing*.

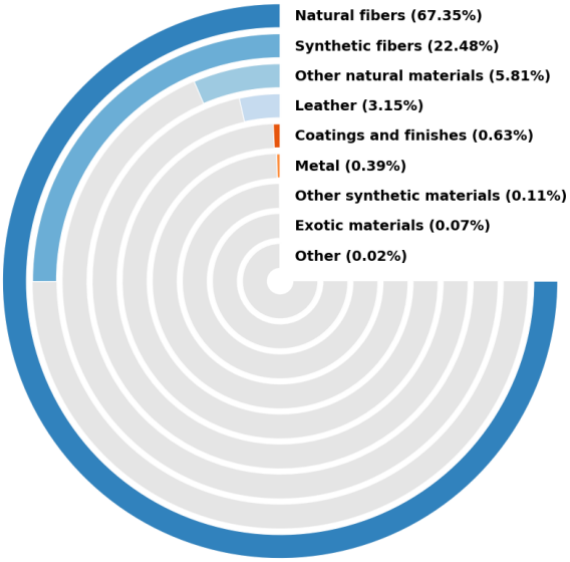

Figura 19 Grafico a barre per variabile top materials

Il grafico sottostante, nella Figura 20, è una rappresentazione interattiva visionabile tramite il codice QR inserito in basso a sinistra. Il mappamondo interattivo raggruppa due genere di dati; il primo qualitativo è il paese in cui sono presente gli *store* dell'azienda destinati alla rivendita al dettaglio. Il secondo dato è il numero di clienti per ogni paese. Quest'ultimo aspetto è stato raffigurato con dei fasci bianchi semi trasparenti. Il numero più elevato di clienti proviene dalla Cina, in particolare *Beijing*. A seguire figurano, *London*, *Los Angeles*, *Washington D.C.* e *Las Vegas*. Questo permette di asserire che, almeno per quanto concerne il numero di clienti, i mercati più importanti per l'impresa saranno quello cinese, il mercato inglese e infine, quello americano. Inoltre, configurandosi questo scenario, l'azienda esporta molto in paesi situati in altri continenti, esponendosi alle dinamiche, ai bisogni e alle necessità di altre culture e popolazioni.

Figura 20 Grafico per distribuzione geografica nel mondo

Un'altra variabile qualitativa connessa alla dimensione della distribuzione è la tipologia di *store*, la quale è presente nel *dataset* con il nome di *Most Frequent Store Type* (Figura 21). Il numero di modalità della variabile è nove e di seguito è stata riportata la sua distribuzione sotto forma di *donut plot*. Questa *feature* rappresenta il luogo di acquisto più frequente per ogni cliente. Il 50% degli acquisti è stato realizzato in una delle boutique

dell'azienda. In linea con il posizionamento dell'impresa, *l'e-commerce* non figura tra le prime voci per frequenza. Questo risultato è dovuto sia al posizionamento esclusivo del *brand* sia alla tipologia di prodotto commercializzato il quale, in alcune ipotesi, necessita di un'interazione fisica per poter essere realizzato su misura. La seconda tipologia di *store* per ricorrenza è il *company store. Outlet*, *temporary shop* e *private sale* rappresentano le categorie con minor frequenza.

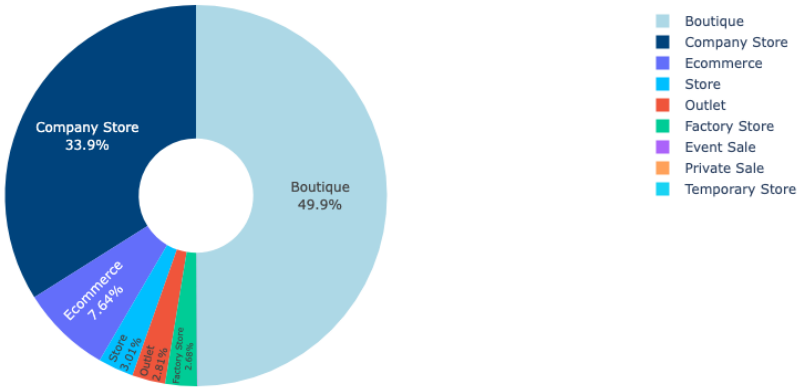

Figura 21 Donut chart per tipologia di store

In merito alla classificazione dei prodotti, da *Transactions Details*, è stata presa una variabile che classifica il genere di prodotti acquistati. Questa è costituita da cinque modalità ossia *formal*, *sportswear*, *accessories*, *made to measure e made to order*, la cui distribuzione è riportata nella Figura 22. La tipologia di prodotti venduta con più frequenza è quella degli abiti formali in linea con la tipologia di prodotti per cui l'azienda si è collocata.

Il *make-to-order* (*MTO*) è una strategia di realizzazione di indumenti, la quale si basa sulla produzione dell'ordine solo dopo che lo stesso sia stato ricevuto. Di conseguenza, non è previsto uno *stock* di prodotti come nella maggior parte delle aziende del *fast fashion* odierno (Kabaivanova, 2015). Questo approccio produttivo è difatti l'approccio considerato storicamente come classico in quanto, prima dell'avvento del fast *fashion*, le modalità con cui si ottenevano dei vestiti erano principalmente l'acquisto presso una sartoria commissionando il capo su misura oppure la produzione autonoma dell'indumento (Thomas, 2019).

Il *made to measure* è definito come il processo produttivo mediante il quale, il capo è realizzato su misura. Tuttavia, non si tratta della realizzazione dell'indumento partendo dalla realizzazione di un nuovo modello per ogni singolo acquisto. L'attività sartoriale,

in questo caso, si limita alla realizzazione delle modifiche su un modello di vestito già esistente. In altre parole, partendo da un indumento con una taglia vicina a quella che si vuole avere, si realizzano le opportune modifiche affinché il capo da standardizzato diventi personalizzato per le esigenze del cliente (Forbes, 2013).

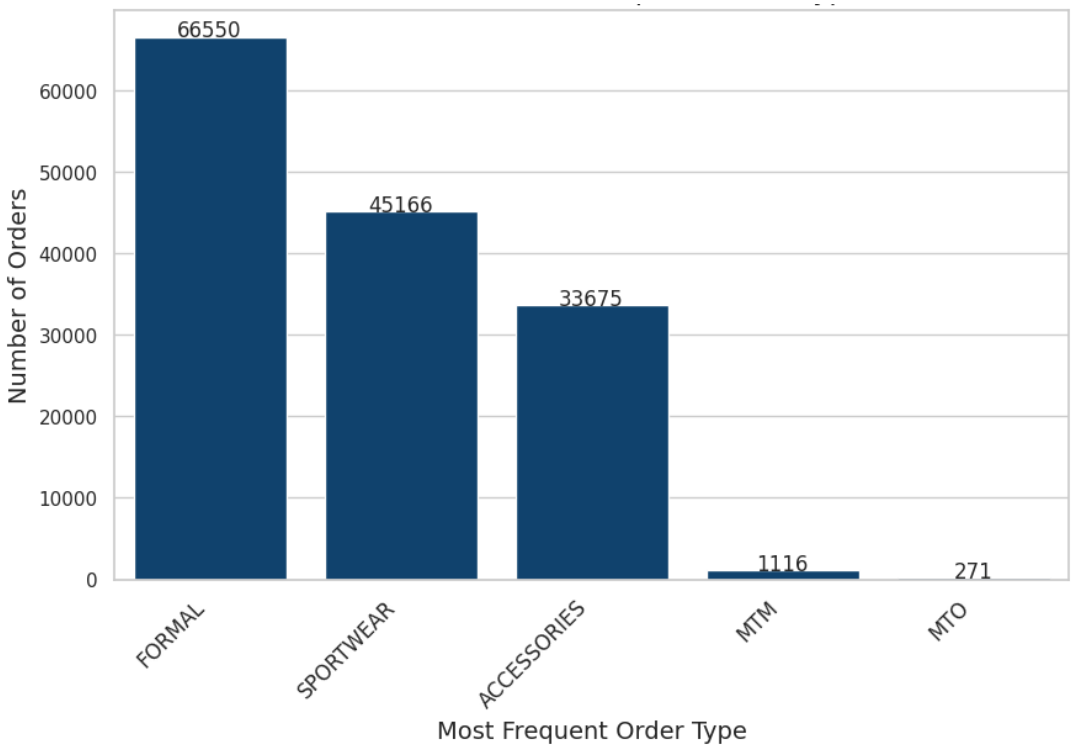

Figura 22 Grafico a barre per distribuzione di frequenza della tipologia di ordine

Nell'ambito delle variabili categoriche del *dataset*, *Latest Year* si compone di diverse categorie qui rappresentate per numero di clienti (Figura 23). Questa colonna del *set* di dati rappresenta l'anno più recente di acquisto per il cliente. La distribuzione ha il suo massimo tra il 2016 e 2017, toccando il suo minimo nel 2015. La linea rappresentante la frequenza segue quelle che sono state le dinamiche dovute alla pandemia. Infatti, in corrispondenza dei periodi di *lockdown*, i clienti che hanno acquistato sono diminuiti. Negli ultimi 3 anni, infatti, il 2020 è l'anno con il numero di clienti minore mentre, tra il 2021 e il 2022, c'è stato l'incremento massimo degli anni presi in considerazione prima. Questa informazione sui clienti determina la possibilità di analizzare la *recency* dei clienti.

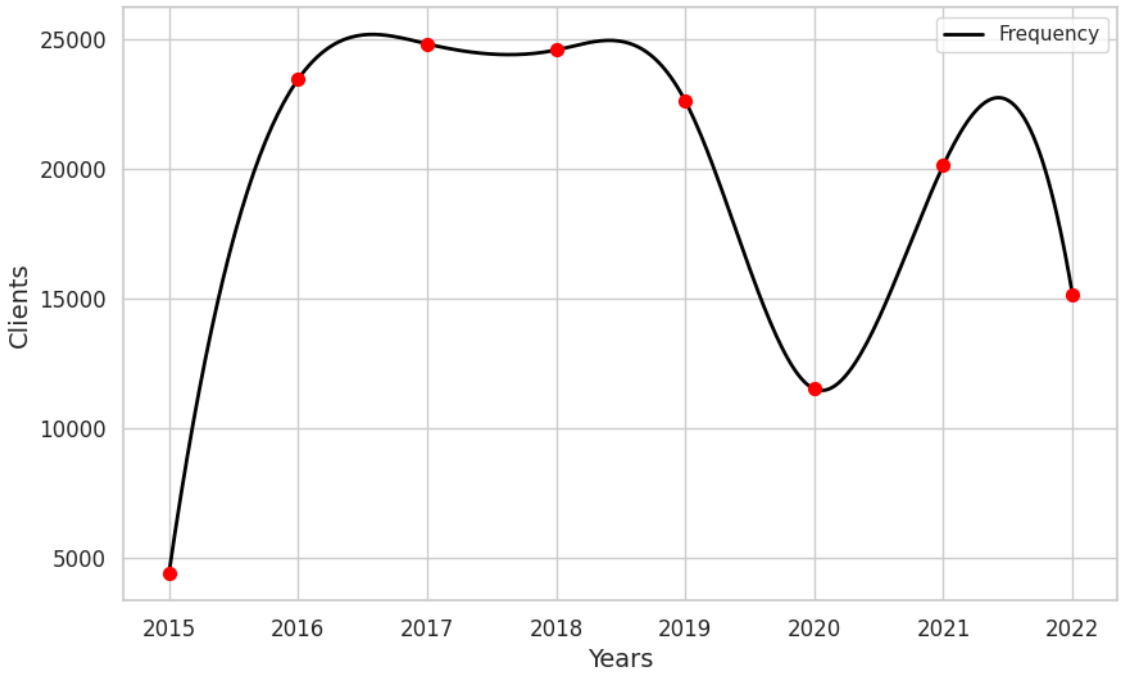

Figura 23 Grafico per distribuzione dei clienti nel corso degli anni

Il *dataset* analizzato include anche la varibile categorica *Newsletter Subscription Form* suddivisa in sei categorie: *not defined*, *newsletter*, *account*, *contact-us*, *book-appointment* e *facebook* (Figura 24). Queste rappresentano tutte i percorsi di iscrizione alla newsletter dell'azienda. Sebbene per quasi l'80% della distribuzione non è possibile definire la modalità di iscrizione alla newsletter, il 20% restante è stato attribuito alla *newsletter*. Quest'ultimo dato indicherebbe l'iscrizione di utenti alla *newsletter* provenendo da un'altra. Infine, l'1,50% degli utenti si è iscritto mediante *account* personale.

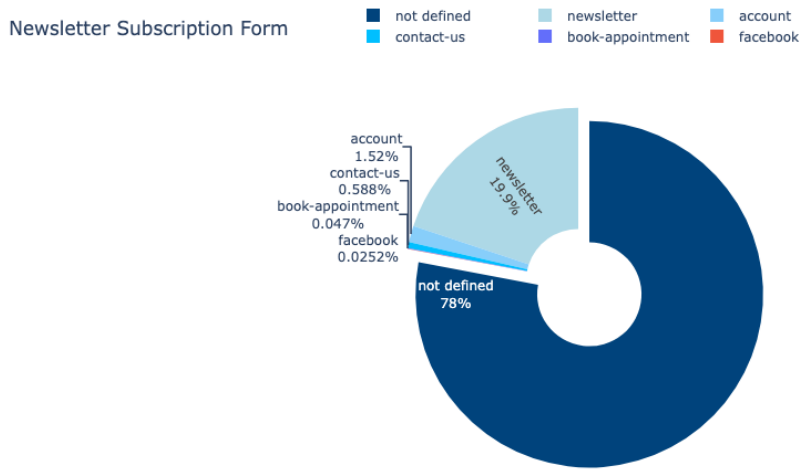

Figura 24 Donut chart per distribuzione di frequenza del metodo di iscrizione alla newsletter

In merito alla frequenza di acquisto ossia il numero di interazioni che vi sono tra cliente e azienda, è presente una metrica nominata *rfm_frequency* relativa al modello di valutazione del comportamento dei clienti "*Recency, Frequency, Monetary Value*" (Figura 25). Sebbene sia molto simile al numero di ordini e con differenze minime, risulta essere solo moderatamente correlata con quest'ultima (0,51 *coefficiente di correlazione di Pearson*). La variabile in questione è formata da cinque modalità numerate da 1 a 5. Dove 1 è la categoria di clienti con minori interazioni (*very low frequency*) e 5 la categoria di clienti con il numero di interazioni maggiori (*very high frequency*). Come si evince dalla Figura 25, l'81% dei clienti è stato classificato con una frequenza di interazioni bassa, il numero diventa sempre via via minore considerando gli altri *score* attribuiti ai clienti. Nello specifico la frequenza 5 è stata attribuita solo all'1% dei clienti.

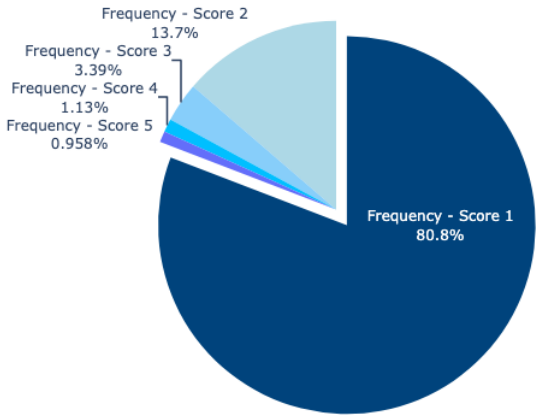

Figura 25 Grafico a torta per rfm frequency

Per quanto concerne l'ultimo grafico per le variabili categoriche, nella Figura 26 sono stati riportati due dimensioni insieme ossia la *Most Frequent Day of Week* e il *Most Frequent Month*. Sulla base dei dati dei clienti, i mesi con più clienti sono dicembre e gennaio. Inoltre, per ogni mese è stato riscontrato un numero elevato di clienti nel fine settimana, in particolare ogni sabato. Nell'arco dei giorni feriali, la distribuzione è abbastanza uniforme con qualche minima differenza solo per il venerdì in alcuni mesi come aprile e novembre.

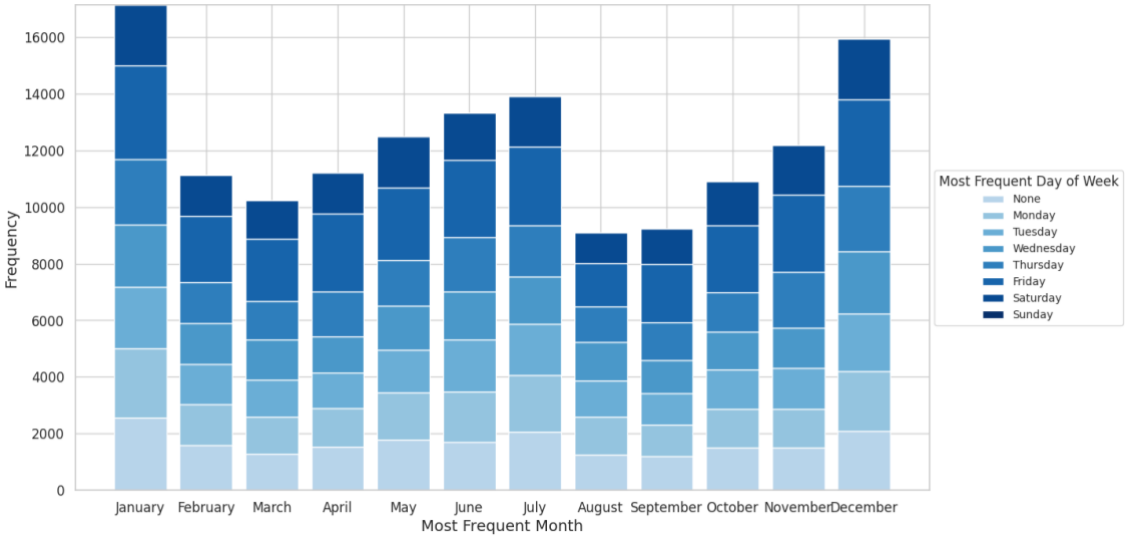

Figura 26 Grafico a barre per distribuzione di frequenza combinata di mese e giorno della settimana

Le variabili quantitative (non binarie) del *dataset* sono il numero di *e-mail* ricevute, il numero di *e-mail cliccate*, i ricavi per singolo cliente e il numero di acquisti per singolo cliente. Di seguito è stata riportata la Tabella 3 riassuntiva delle statistiche principali per le quattro variabili quantitative presenti nel *dataset*.

Metriche	**Ricavi**	**Ordini**	**E-mail Cliccate**	**E-mail Ricevute**
Media	2579,68	2	0	13
STD	9607,49	3	2	25
Min	-14043,22	1	0	0
25,00%	380	1	0	0
50,00%	1038,97	1	0	0
75,00%	2446,92	2	0	12
Max	1669502,12	231	130	143

Tabella 3 Statistiche descrittive delle variabili quantitative non binarie

In termini di ricavi, il paese che risulta avere il valore maggiore è la Cina con 155 milioni di euro (41% del totale). A seguire, vi sono *USA* (30%), Inghilterra (15%) e l'Italia (12%). Il valore medio per paese è di undici milioni. Il paese con i ricavi minori è Malta.

Il cliente da cui l'impresa ha ricevuto il numero maggiore di ricavi è un turista, il quale ha effettuato 45 ordini in Inghilterra determinando circa 1,7 milioni di euro per l'impresa ossia circa 38000 euro in media per ordine. La distribuzione è stata inserita nella Figura 27.

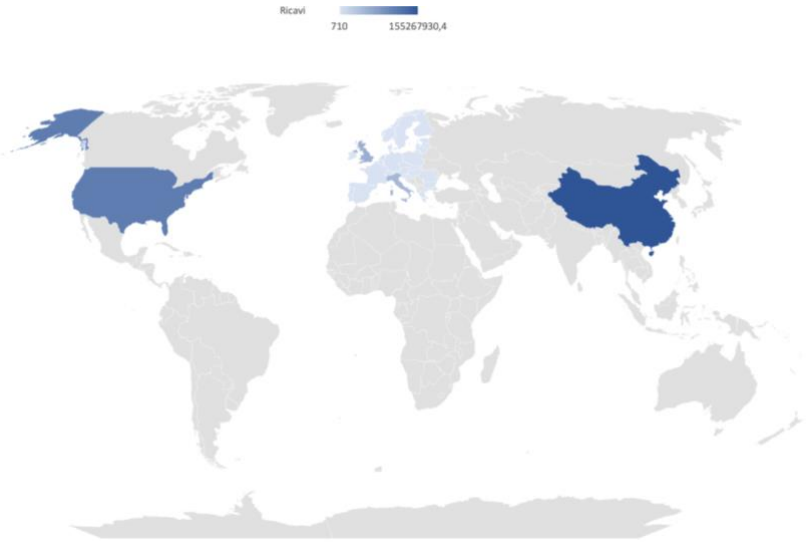

Figura 27 Ricavi per paese

Per quanto concerne la variabile relativa al numero di ordini, questa è stata riportata graficamente come segue (Figura 28). La Cina e gli Stati Uniti d'America risultano essere i primi due mercati anche in termini di numero d'ordine; rispettivamente 34,30% e 33,73%. L'Italia, invece, supera l'Inghilterra con un 16,74% rispetto al 12,19% dell'Inghilterra. Questa evidenza ci permette di asserire che, sebbene in Italia ci sia un numero di ordini maggiore, non risulta essere così tanto remunerativo quanto l'Inghilterra considerando i dati riportati precedentemente per i ricavi. Il numero di ordini medio per cliente è pari a 2.

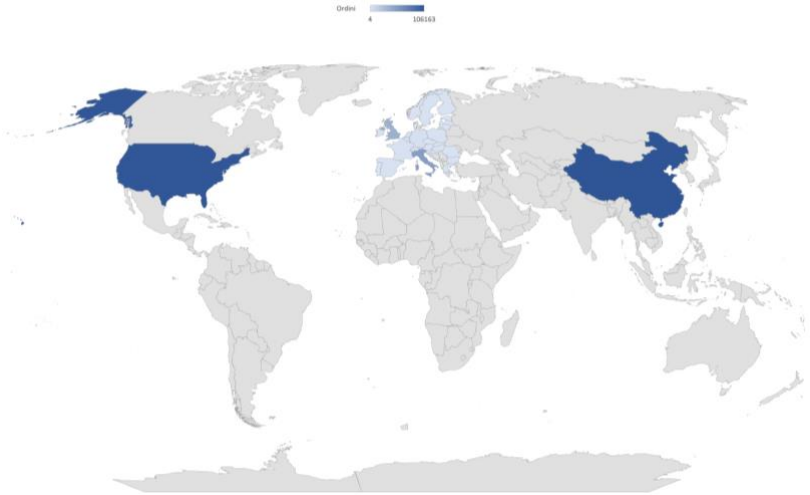

Figura 28 Numero di ordini per paese

Con le *feature* relative alle e-mail ossia il numero di *e-mail* cliccate e il numero di *e-mail* ricevute è stato calcolato l'*open rate* per paese (Figura 29). Questa metrica può essere

utilizzata come un indice del livello di digitalizzazione per paese. In particolare, si evince come i paesi dove il numero di ordini e i ricavi sono maggiori, l'*open rate* sia il più basso. Questo fenomeno è dovuto al numero di *e-mail* ricevute dai clienti di gran lunga maggiore rispetto agli altri paesi. Nonostante ciò, ai primi posti figurano principalmente paesi del continente europeo e le prime cinque posizioni sono occupate da paesi dell'est Europa dove, negli ultimi anni, si sta verificando una forte crescita in termini di digitalizzazione.

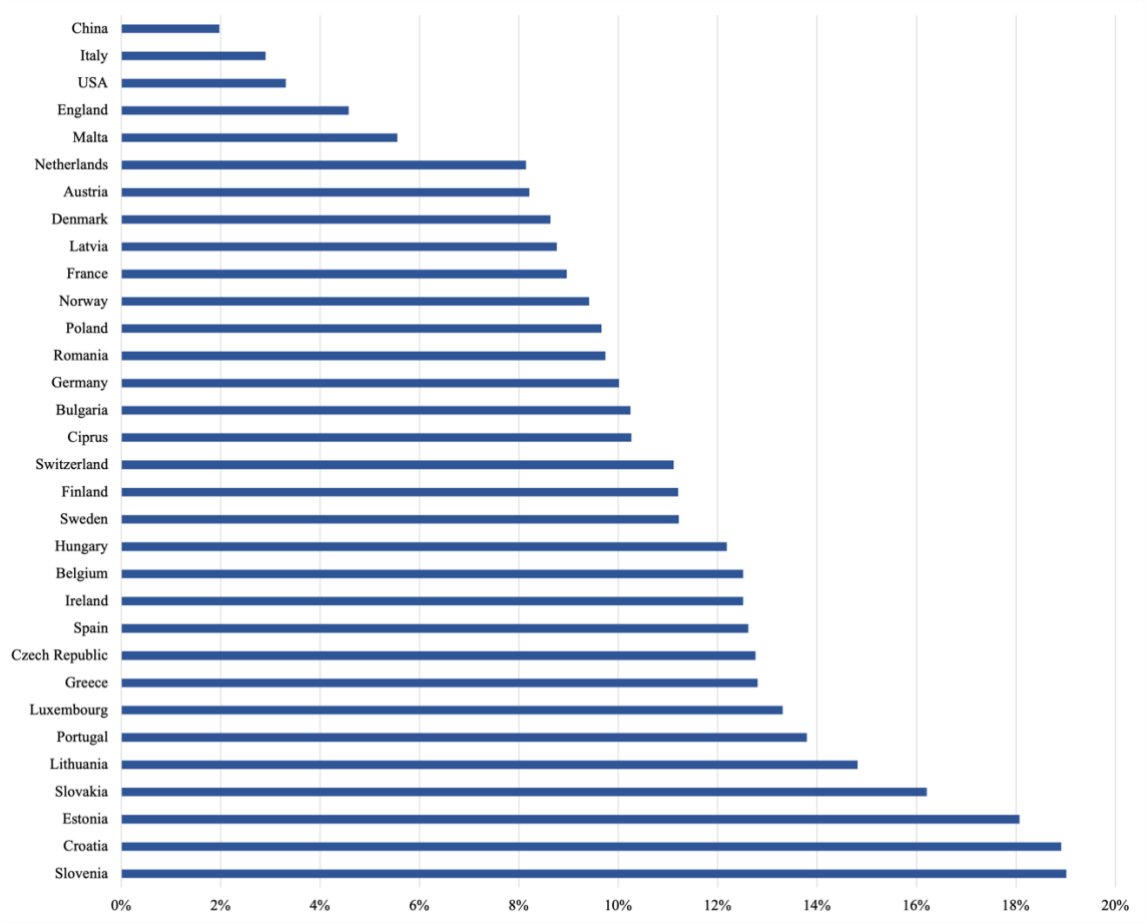

Figura 29 E-mail Open Rate per paese

Per quanto concerne l'analisi degli *outliers* per le ultime quattro variabili, sono stati utilizzati i *boxplot* come strumento esplorativo (Figura 30; Figura 31).

Sebbene siano stati trovati molteplici valori anomali, dopo diverse valutazioni è stato deciso di non eliminarli in quanto non sono stati identificati come errori di misurazione bensì come osservazioni valide.

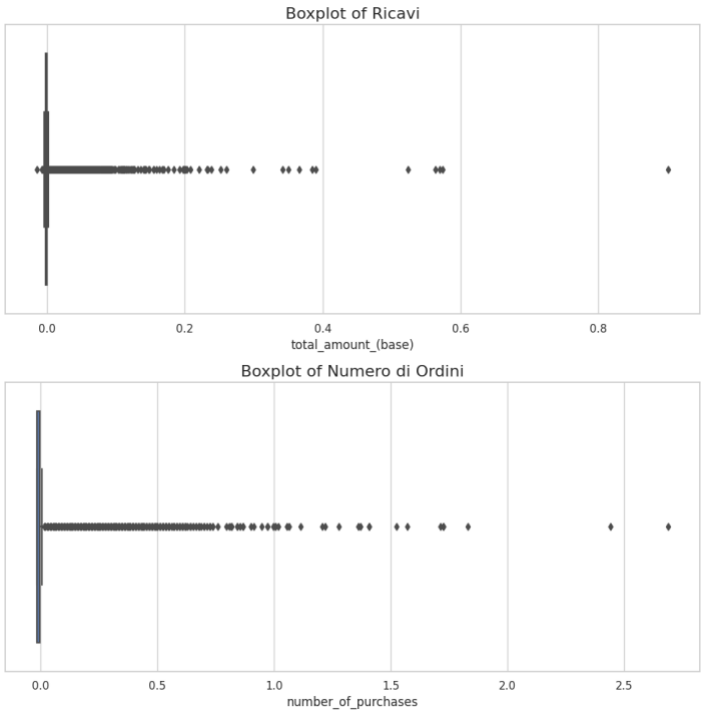

Figura 30 Boxplot per gli outlier della variabile ricavi e numero di acquisti

Infatti, i valori sono parte della variabilità della distribuzione e sono anche fonte di spunti interessanti per valutazioni ulteriori. Considerando che il *dataset* riguarda molteplici paesi, clienti di un'azienda del *fashion* di lusso e raccoglie le osservazioni verificatesi in sette anni, gli *outliers* individuati risultano essere parte integrante della struttura dei dati e di conseguenza da mantenere per le finalità dello studio.

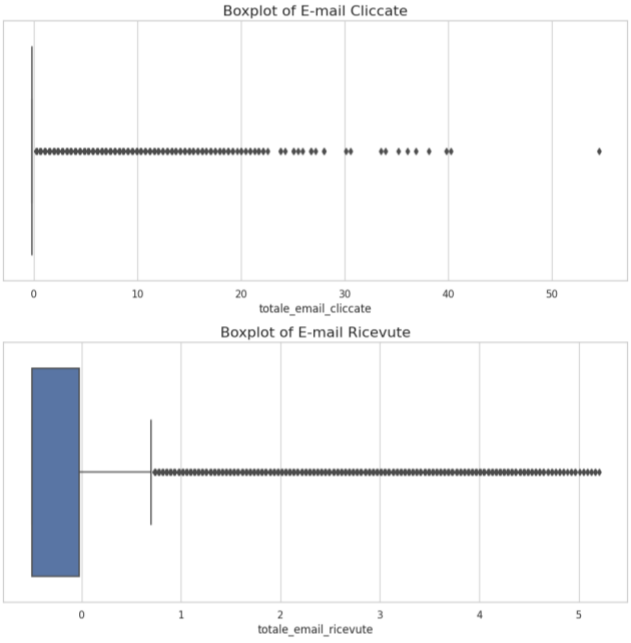

Figura 31 Boxplot per gli outlier delle variabili e-mail cliccate ed e-mail ricevute

Di seguito, una breve tabella riassuntiva di tutte le variabili finora descritte (Tabella 4).

N	Variabile	Modalità	Tipo	Descrizione
1	*Gender*	3	Demografica	Distribuzione del genere dei clienti: "Male", "Female", "Not declared".
2	*Sales Habit*	3	Comportamentale	Abitudini di acquisto dei clienti: "Regular", "On Sale", "Both".
3	*Customer Habit*	5	Demografica	Classificazione dei clienti: "Resident", "Turist", "Viaggiatore", "Cliente Internazionale", "Not defined".
4	Preferenza *Mail*	2	Comportamentale	Preferenza per le comunicazioni via posta tradizionale: "Yes", "No".
5	Preferenza Telefono	2	Comportamentale	Preferenza per le comunicazioni telefoniche: "Yes", "No".
6	Preferenza *SMS*	2	Comportamentale	Preferenza per le comunicazioni via SMS: "Yes", "No".
7	Preferenza *E-mail*	2	Comportamentale	Preferenza per le comunicazioni via e-mail: "Yes", "No".
8	Consenso SMS	2	Comportamentale	Consenso per le comunicazioni via SMS: “Yes”, “No”
9	Consenso Newsletter	2	Comportamentale	Consenso per le comunicazioni via newsletter: “Yes”, “No”
10	Consenso Marketing	2	Comportamentale	Consenso per le attività di marketing: “Yes”, “No”
11	*Top Products*	35	Comportamentale	Prodotti maggiormente acquistati dai clienti.
12	*Top Colors*	11	Comportamentale	Colori dei prodotti maggiormente frequenti: "Blu", "Bianco", "Verde", "Rosa", ecc.
13	*Top Materials*	9	Psicografica	Materiali utilizzati nella produzione dei prodotti acquistati.
14	*Most Frequent Store*	N/A	Demografica	Distribuzione geografica degli *store* dell'azienda per la rivendita al dettaglio.
15	*Most Frequent Store Type*	9	Comportamentale	Tipologia di negozio più frequente per ogni cliente.
16	*Order Type*	5	Comportamentale	Classificazione dei prodotti acquistati: "Formal", "Sportswear", "Accessories", "Made to Measure", "Made to Order".

	Newsletter Subscription Form	6	Comportamentale	Metodo di iscrizione alla newsletter dell'azienda.
17				
18	*Newsletter Subscription*	2	Comportamentale	Iscrizione alla newsletter: “Yes”, “No”.
19	*RFM Frequency*	5	Comportamentale	Valutazione del comportamento dei clienti basata su *Recency*, *Frequency*, *Monetary Value*.
20	*Most Frequent Day of Week*	7	Comportamentale	Giorno della settimana più frequente per ogni cliente.
21	*Most Frequent Month*	12	Comportamentale	Mese più frequente per ogni cliente.
22	*Latest Year*	N/A	Comportamentale	Anno più recente di acquisto per il cliente.
23	*Revenue*	N/A	Economica	Ricavi per singolo cliente.
24	*Orders*	N/A	Economica	Numero di ordini.
25	*E-mail Clicked*	N/A	Comportamentale	Numero di e-mail cliccate.
26	*E-mail Received*	N/A	Comportamentale	Numero di e-mail ricevute.

Tabella 4 Overview delle variabili utilizzate per la clustering analysis

6 Data Pre-Processing

Il *dataset* descritto nel capitolo 5 è il risultato di diverse tecniche statistiche, analisi e valutazioni applicate mediante *python*. Precedentemente sono stati descritti i *dataset Transactions Details e Contact Active*, in questo capitolo verranno approfondite le relative *exploratory data analysis*, i *missing value* e la parte di *pre-processing*.

Entrambi i *set* di dati sono stati inseriti in un *pandas dataframe* al fine di semplificare ed efficientare le analisi e le metodologie applicate di seguito. Per la parte di *EDA* così come anche per le successive, sono stati utilizzati diversi pacchetti; tra questi, oltre *pandas*, vi sono anche *numpy* e *scikit-learn* i quali offrono una vasta gamma di funzioni e modelli per manipolare i dati. Per la parte di visualizzazione sono stati utilizzati molteplici pacchetti come *matplotlib, seaborn, plotnine, re, WordCloud, squarify* e *plotly*. Dopo l'import dei pacchetti principali, sono stati osservati i *set* di dati per capirne le variabili e organizzare il lavoro al fine di raggiungere l'obiettivo nel modo più efficiente possibile.

Il lavoro è iniziato come l'*EDA* per *Transactions Details*, per il quale sono state rinominate le variabili eliminando gli spazi che vi erano nell'*header* di ciascuna colonna. Successivamente sono state richieste le statistiche descrittive del *dataset* come *shape, size*, dimensioni e numero variabili quantitative e qualitative. Per valutare la correlazione delle variabili quantitative è stata usata una matrice di correlazione di *Pearson* riportata con la Figura 32, dalla quale si evincono che alcune variabili sono molto correlate tra loro in modo positivo. Sebbene non siano del tutto assenti, elevati coefficienti di correlazione negativa non sono stati riscontrati. Le correlazioni più elevate sono state osservate tra le variabili riconducibili ai ricavi e ai prezzi nelle loro diverse forme (ad esempio *net*) e valute (*base* indica che l'ammontare è in euro). Sebbene non siano stati gestiti nella EDA, sono stati esplorati anche i missing *value*. In particolare, per alcune variabili si registrava anche fino al 96% di valori mancanti. In media, tra la totalità di variabili il 14% dei dati era mancante. In questa prima fase sono stati utilizzati anche diversi grafici al fine di effettuare le dovute valutazioni circa le distribuzioni delle molteplici variabili. Questa prima fase si è conclusa con l'esplorazione dei valori unici di alcune variabili come la colonna "*Transaction*" e "*Customer*", rispettivamente i codici alfanumerici per identificare le transazioni e per identificare i clienti.

L'*exploratory data analysis* realizzata per *Contact Active* è stata realizzata in modo molto simile, di conseguenza dopo le fondamentali descrittive per inquadrare opportunamente il *dataset*, è stata realizzata la matrice di correlazione (Figura 33) e l'esplorazione dei dati

mancanti. Così come per *Transactions Details*, anche qui, per *Contact Active*, è stato utilizzato il pacchetto *sidetable* per ottenere una descrizione globale dei valori mancanti.

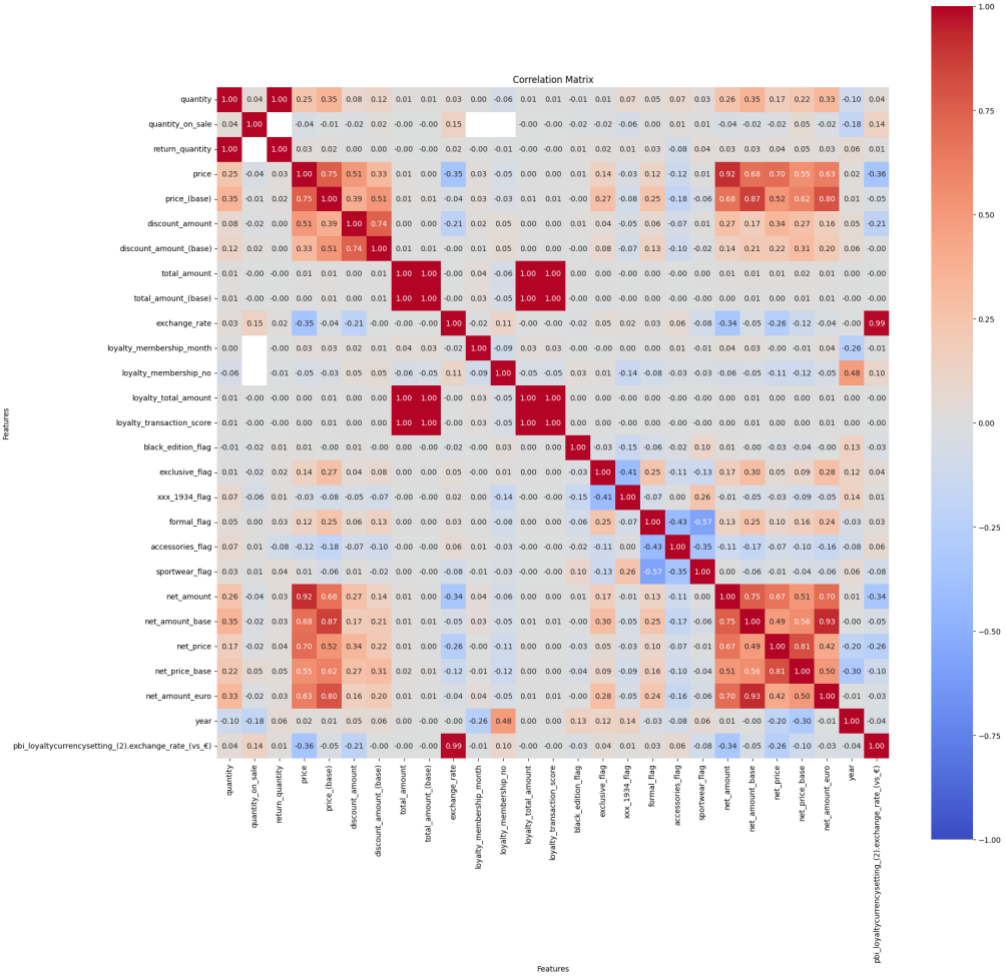

Figura 32 Matrice di correlazione per Transactions Details

Nel *dataset* Contact Active, vi è un numero di variabili avente più del 95% di dati assenti pari a 35. La media dei valori mancanti tra le diverse colonne del *set* di dati è del 53%. Di conseguenza, per entrambe le matrici di dati, la risoluzione del problema dei dati mancanti era cruciale al fine di ottenere un lavoro accurato.

Con l'*EDA* svolta e sulla base delle metriche necessarie per le finalità dello studio è stato possibile eliminare alcune svariate variabili in quanto non avrebbero aggiunto valore e informazioni utili per la realizzazione di *marketing personas*. Pertanto, in *Transactions Details*, nella parte di *data manipulation* relativa ai *missing value*, sono state eliminate 39 variabili ottenendo così un *dataset* di 26 colonne. Per alcune variabili, come quella contenente i materiali utilizzati nel prodotto acquistato, sono state create nuove modalità. In questo caso specifico, i valori assenti sono stati sostituiti con la categoria "*not defined*"

in quanto, sostituendo con la media, si sarebbero potute generare delle osservazioni anomale come, ad esempio, un accessorio o un indumento con un materiale difficilmente utilizzato se non proprio impossibile da realizzare.

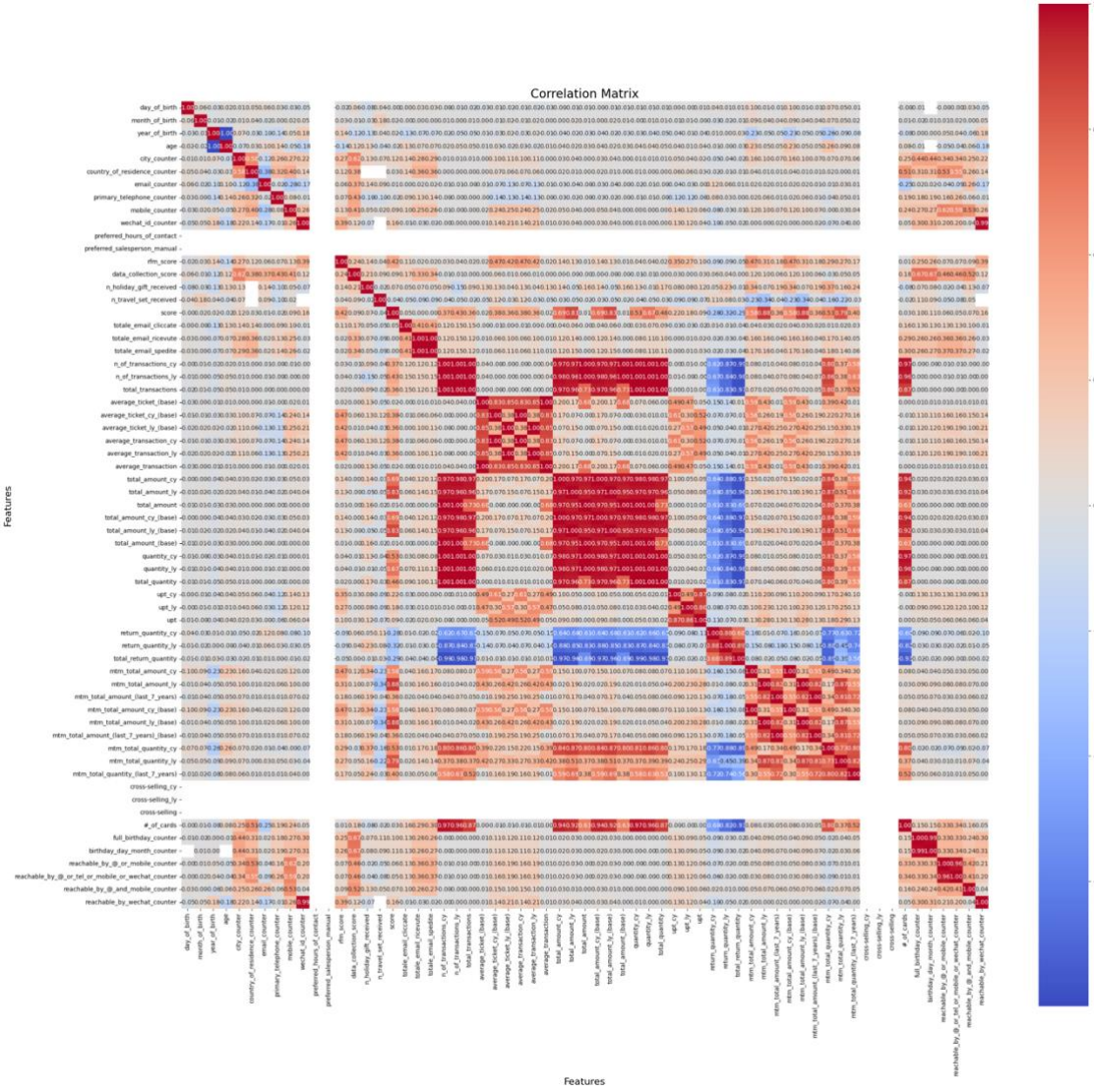

Figura 33 Matrice di correlazione per Contact Active

Per altre variabili quantitative, sono stati sostituiti i valori mancanti con uno zero laddove avesse significato come, per esempio, nel numero di *loyalty membership* sebbene queste variabili non siano state utilizzate nel *dataset* finale. Infine, per le variabili con la percentuale di valori mancanti più bassa, è stato deciso di eliminare direttamente le unità statistiche che non avessero valori nelle rispettive colonne indicate. Grazie a queste valutazioni, decisioni e tecniche di risoluzione dei dati mancanti, il *dataset Transactions Details* aveva raggiunto una situazione per cui tutte le *feature* contenevano un dato per ogni unità. In questa fase sono stati effettuati ulteriori controlli legati alla fase di *data*

cleaning che avrebbe seguito nella fase successiva di *pre-processing*; fondamentalmente, per avere un risultato globalmente consistente, è stato controllato se ci fossero degli spazi anomali per ogni valore della colonna "*customer*" (contenente l'*ID* dei clienti) così come se ci fossero dei segni di punteggiatura non in linea con quelli utilizzati per ogni valore di ciascuna colonna del *dataset*. Tale controllo è stato effettuato anche per il *dataset Contact Active*. Riguardo a quest'ultimo, i valori assenti sono stati gestiti con le medesime procedure e valutazioni. In primo luogo, sono state eliminate tutte le *feature* che non sarebbero servite per le finalità del contributo qui presentato. Dopo ciò il *dataset* ottenuto era costituito da 22 variabili. Per le variabili con un numero di valori mancanti irrisorio sono state eliminate le righe contenenti i valori mancanti. Al contrario per le restanti variabili incomplete sono state realizzate nuove categorie oppure con categorie che non generassero distorsione a livello informativo, ed è questo il caso, per le variabili dicotomiche legate al consenso o alle preferenze dello strumento di comunicazione per le quali è stata inserita la modalità indicante il mancato consenso oppure la mancata preferenza. In altre ipotesi, così come per *Transactions Details*, per le variabili che si prestavano a questa tecnica di sostituzione, è stata creata e aggiunta un'altra modalità. Sulla base di queste decisioni, anche questa matrice di dati presentava delle colonne prive di valori mancanti.

La parte di pre-processamento è iniziata con lo studio dei valori unici di ogni variabile sulla base del quale sono stati riscontrate diverse considerazioni. Per alcune variabili, infatti, vi erano degli evidenti errori di codifica dei valori delle variabili. In altre parole, le variabili contenevano valori unici duplicati per via di errori di battitura, registrazione oppure codifica. Questo scenario si è configurato in particolar modo per la *feature* "*composition_item*" contenente i materiali costituenti i prodotti di ogni transazione. Questa variabile conteneva sia le percentuali che il nome dei materiali. Per ragioni computazionali e di interpretazione, è stata creata una nuova variabile con un numero di modalità inferiori e raggruppando i diversi materiali in dei gruppi con le medesime caratteristiche. Le nuove modalità della variabile contenente i materiali di ogni prodotto sono "*Natural fibers"* (contenente seta, cotone, lana, cashmere, lino ecc.), "*Synthetic fibers"* (contenente elastane, viscosa, poliestere, PVC, poliacrilico, nylon ecc.), "*Other natural fibers"* (contenente diverse tipologie più particolari di lana, la gomma,), "*Other synthetic fibers* (come la bioceramica, la microfibra, ecc.), "*Exotic materials"* (come la pelle di opossum, pitone, di coccodrillo o di altri rettili, ecc.), "*Leather"* (come la pelle

bovina, dell'ermellino, ecc.), "*Metal"* (come l'argento, l'oro e così via), "*Coating and finishes"* (come *CottonwithPUcoating* e altre valori che contenessero questa peculiarità nel nome) e "*Other"* per tutto il resto qualora non potesse essere classificato in una delle categorie definite precedentemente. La categorizzazione dei valori si è basata sullo studio dei materiali e del loro tipico utilizzo per prodotti di abbigliamento. Con il medesimo approccio, è stato realizzato lo stesso lavoro per le variabili *category_item* riguardante la categoria di prodotto acquistato e *style_item*, feature contenente lo stile di prodotto. In questo caso, le due colonne sono state unite, separando i valori con una virgola e successivamente sono stati codificati al fine di ridurre il numero di modalità della variabile. La variabile finale rinominata *category_style_item* conteneva i seguenti valori unici *"accessories", "bags", "belts", "blazer", "bow ties", "cargo trousers", "chinos", "coats", "cuffs", "cummerbund", "field jackets", "gloves", "hats", "home", "jackets", "jeans", "knitwear", "overcoat", "polo", "raincoats", "scarves", "shearling", "jackets", "shirts", "shoes", "shorts", "sports accessories", "sports trousers", "suits", "sweaters", "sweatshirt", "swimwear", "ties", "trousers", "waistcoats"* e *"wallets"*. Infine, per la variabile *store*, contenente sia la tipologia di negozio che la città in cui è situato, è stata codificata ottenendo due variabili. La prima è lo *store_state* ossia il paese in cui lo *store* è situato oppure il luogo dove è stato effettuato l'acquisto nel caso in cui fosse l'*e-commerce*. L'altra variabile è *store_type* la quale prevede le seguenti modalità *"Company Store", "Boutique", "E-commerce", "Store", "Outlet", "Factory Store", "Event Sale", "Private Sale"* e *"Temporary Store"*. Nello specifico, tutte le codifiche o estrazioni di variabili riportate finora sono state realizzate sempre partendo dalla creazione di un *dictionary* da utilizzare poi con i codici *replace* oppure *split* in dei cicli di *for*. La variabile *date* è stata scomposta in quattro nuove variabili per semplificare la lettura del dato negli algoritmi nonché l'esplorazione dei dati in fase di interpretazione. Le nuove feature ottenute sono state *year*, *month*, *day* e *day_of_the_week.* La *shape* finale del *dataset Transactions Details* è di 878462 x 27.

La parte di pre-processamento per *Contact Active* è iniziata con la codifica delle variabili qualitative dicotomiche ("*Yes*" o "*No*") mediante *LabelEncoder*. In seguito, così come visto precedentemente, sono stati esplorati tutti i valori unici del *dataset* per verificare la presenza di eventuali errori e porre in essere le dovute considerazioni sulle dimensioni delle variabili. Non avendo riscontrato nessun problema in termini di registrazione, codifica o di scrittura del dato, mediante la funzione *groupby* sono stati uniti i due *dataset*

Transaction Details e *Contact Active*, sulla base della colonna "*customer ID*" in quanto *primary key*. Il *dataset* risultante è stato nominato *Contact Active Updated*; questo è costituito dai valori più frequenti nelle transazioni di ogni cliente per ogni variabile. Inoltre, per alcune variabili, per non perdere eccessivamente il quantitativo di informazioni presente nelle variabili di partenza, sono stati riportati i tre valori più frequenti. Queste variabili sono "*top_materials*"*, "top_colors", "top_products".* Per *Contact Active Updated*, sono stati selezionati solo le unità statistiche classificate come *Customer* nella variabile *crm_type*, contenente la tipologia di acquirente, il quale poteva essere anche un dipendente dell'azienda o altro. La fase di *preprocessing* si è evoluta con ulteriori controlli circa i valori unici delle variabili e le distribuzioni di alcune *feature* per controllare che non si fossero generati errori con l'operazione di *groupby*. Per la *feature selection* sono stati utilizzati due approcci. Il primo basato sulla conoscenza del dominio e di conseguenza, la maggior parte delle variabili sono state necessariamente mantenute nel *dataset* dato l'obiettivo definito a priori. In secondo luogo, anche per il nuovo *dataset Contact Active Updated*, è stata realizzata una matrice di correlazione di *Pearson* riportata con la Figura 34.

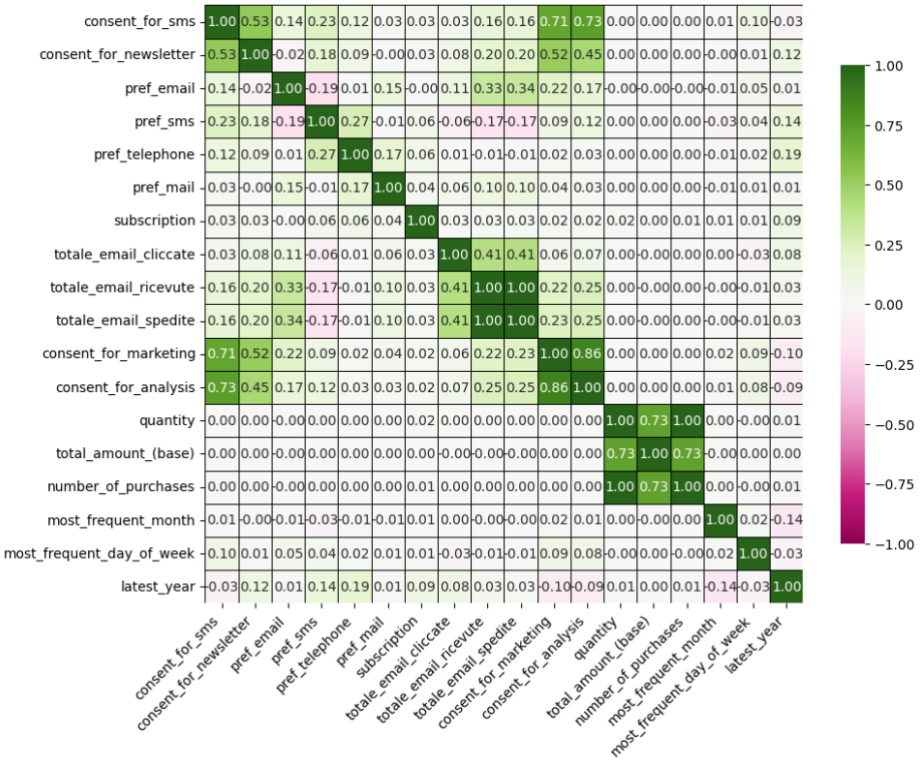

Figura 34 Matrice di correlazione per Contact Active Updated

Sulla base dei coefficienti di correlazione sono state rimosse le variabili *"consent_for_analysis", "consent_for_sms", "consent_for_newsletter", "quantity", "totale_email_spedite"* e *"rfm_recency".* Con quest'ultima operazione il *dataset* era stato processato completamente, pertanto, è stata creata una variante dello stesso contenente tutte le variabili codificate. La codifica è avvenuta mediate le funzioni *LabelEncoder* e soprattutto *get_dummies*. Infine, sono state standardizzate le variabili di natura puramente quantitativa come *"total_amount_(base)", "number_of_purchases", "totale_email_cliccate"* e *"totale_email_ricevute"*.

Infine, per completezza di analisi e come ulteriore approfondimento delle variabili contenute nel *dataset* finale, è stata realizzata un'analisi esplorativa per ogni variabili riportata in buona parte del capitolo 5.

7 Clustering Analysis

La *cluster analysis*, un'analisi multivariata, rappresenta una tecnica mediante la quale è possibile suddividere un insieme di *pattern*, cioè unità statistiche, in sottoinsiemi caratterizzati da proprietà o caratteristiche simili (Figura 35). Il risultato che si intende raggiungere è quello di ridurre al minimo la devianza all'interno di ciascun *cluster* (devianza *within-groups*), massimizzando al tempo stesso la devianza tra i gruppi (devianza *between-groups*), così da minimizzare la "lontananza logica". In altre parole, questa analisi esplorativa consiste nel ricercare tra le *n* osservazioni *p*-dimensionali, gruppi di unità tra loro simili, senza certezza e conoscenza iniziale che tali gruppi omogenei esistano effettivamente tra i dati oggetto di analisi. L'obiettivo dell'esplorazione così descritta è dunque quello di identificare dei gruppi che appaiono con naturalezza nelle osservazioni.

La misurazione della lontananza logica si basa sull'applicazione di misure di similarità tra le unità statistiche: per dati quantitativi si utilizzano misure di distanza, definite metriche, mentre per dati qualitativi si adottano misure di tipo *matching*, che possono indicare similarità o dissimilarità. Dopo aver selezionato la misura di similarità adeguata, è necessario decidere l'algoritmo di *clustering* e, eventualmente, il metodo di aggregazione o suddivisione.

La conseguenza naturale per quanto concerne l'analisi dei *cluster* finora trattata è che i *pattern* delle unità statistiche all'interno di ogni *cluster* saranno il più simile possibile mentre con maggiore diversità se considerati tra *cluster*.

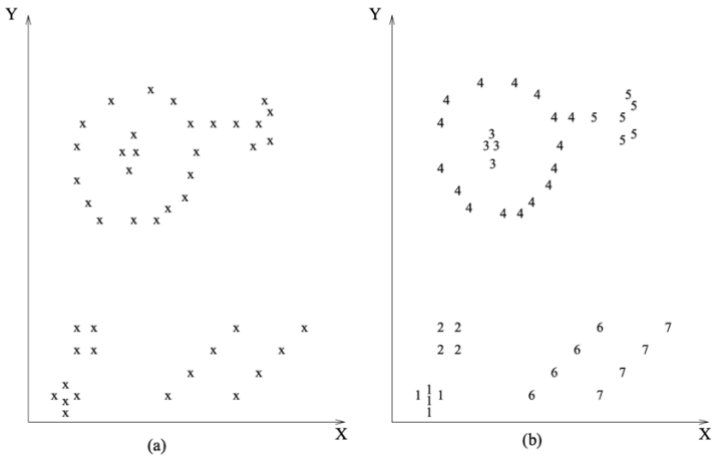

Figura 35 Esempio di clustering

Il problema del *clustering* è stato trattato in diversi contesti e discipline, il che riflette l'ampia diffusione e l'utilità significativa di questa tecnica nell'analisi esplorativa dei dati. Il *clustering*, come metodo di *unsupervised machine learning*, rappresenta un problema complesso in termini computazionali. Le differenze nelle ipotesi e nei contesti delle diverse comunità hanno rallentato il trasferimento di concetti e metodologie generiche. La grande varietà di tecniche per misurare la prossimità (similarità) tra gli elementi, raggrupparli e rappresentarli ha generato un assortimento ricco ma spesso confuso di metodi di *clustering*.

A differenza dei metodi supervisionati, nel *clustering*, si utilizzano dei dati non etichettati ossia delle informazioni per le quali non sono stati identificati dei *pattern* apriori. In questa tecnica di classificazione non supervisionata, gli algoritmi di apprendimento automatico opereranno al fine di identificare e assegnare un pattern per ogni unità statistica. Ne consegue che la difficoltà è nella ricerca dei *pattern* non essendoci un addestramento dell'algoritmo al fine di determinare un apprendimento degli stessi nonché un confronto successivo (Jain, 1999).

Sebbene ci siano studi e contributi precedenti al 1939, quest'anno rappresenta tutt'oggi la data in cui uno dei primi pionieri nel campo dell'analisi dei *cluster* ha pubblicato un contributo divenuto poi uno dei più significativi nel suo settore.

Nel 1939 *Robert C. Tyron* pubblicò un'opera intitolata *"Cluster Analysis: Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality"* la quale si concentrava principalmente sull'approccio statistico alla comprensione e all'interpretazione del comportamento umano. L'obiettivo di ricerca era quello di esplorare come i tratti della personalità potessero essere raggruppati in *cluster* determinando così una comprensione completa della psicologia umana. Nel suo contributo impiegò un metodo noto come *orthometric (factor) analysis* determinando le basi per l'attuale *clustering analysis* (Tyron, 1939).

Le tipologie dei metodi di *clustering* sono divise in gerarchiche e partizionali (non gerarchici). Un'*overview* delle principali tipologie è riportata nella Figura 36.

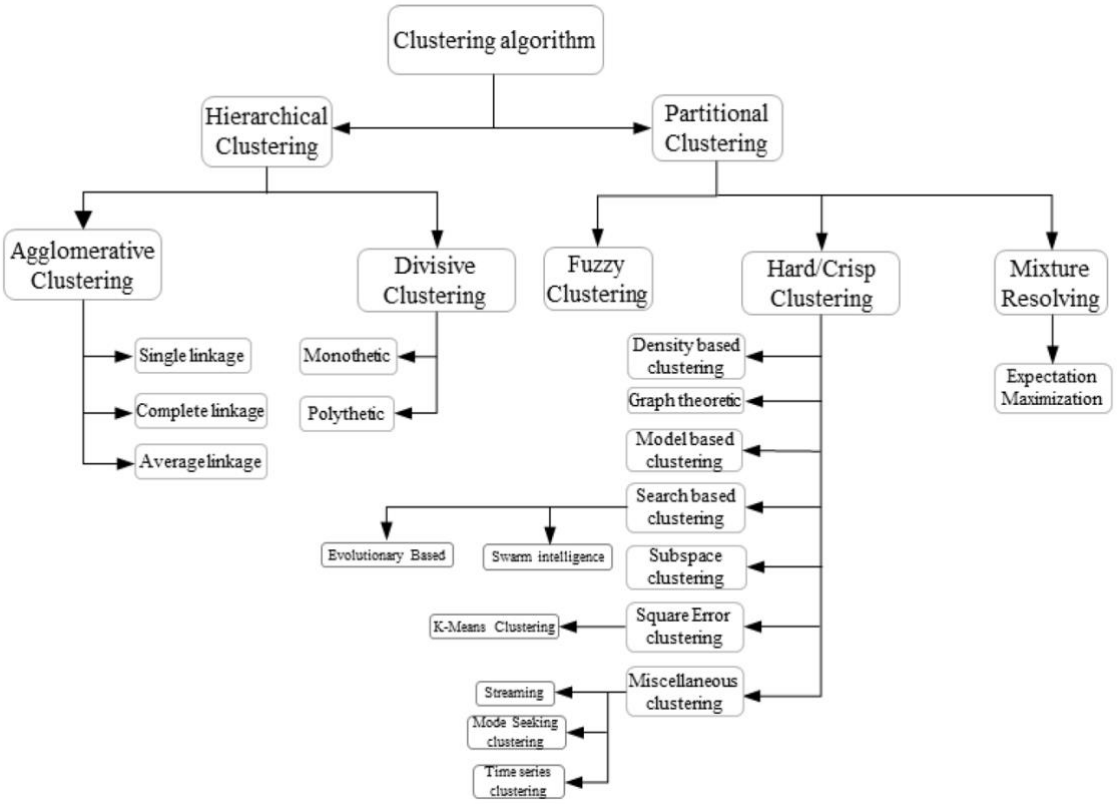

Figura 36 Tassonomia di clustering

Il *clustering* gerarchico si divide a sua volta in aggregativo (o agglomerativo) e scissorio (Pitafi, 2023).

7.1 Algoritmi di clustering gerarchici

Il *clustering* gerarchico aggregativo permette di ottenere una famiglia di partizioni che va da *n* a 1 (approccio *bottom-up*) partendo da una situazione iniziale in cui ogni unità statistica forma un gruppo distinto per conseguire, mediante aggregazioni successive, a quella in cui tutte le unità sono un unico gruppo.

Il *clustering* gerarchico scissorio, viceversa, si basa su una situazione di partenza in cui tutte le n unità sono riunione in un unico gruppo, per giungere via separazioni successive, a delle scissioni volte a definire tanti *cluster* quante sono le n unità statistiche nel caso estremo. Di conseguenza, il risultato sarà opposto, da 1 a *n cluster* (approccio *top-down*).

L'analisi dei *cluster* di tipo gerarchica agglomerativa inizia con la definizione della matrice dei dati *n x p* dove *n* sono le unità statistiche e *p* sono le variabili ossia l'insieme di osservazioni per ogni unità.

La seconda fase si caratterizza per la definizione del concetto di prossimità e dell'applicazione della matrice di prossimità di dimensioni *n x n*. Le matrici di prossimità possono contenere delle distanze o degli indici di distanze per i fenomeni quantitativi mentre indici di similarità o dissimilarità per fenomeni qualitativi.

Prendendo in considerazione la matrice delle distanze, queste possono essere calcolate con diversi approcci. La distanza euclidea tra due unità statistiche *x* e *y* avviene mediante il calcolo della radice quadrata della sommatoria tra le differenze al quadrato dei valori delle unità statistiche per ogni *p* variabile.

$$d(x,y) = \sqrt{\Sigma(x_i - y_i)^2}$$

La distanza di *Manhattan*, detta anche della città a blocchi, si basa sulla sommatoria delle differenze in valore assoluto. tra i valori di ogni *p* variabile tra due unità statistiche.

$$d(x,y) = \Sigma|(x_i - y_i)|$$

La distanza di *Chebychev* (o di *Lagrange*) è pari al valore maggiore tra le differenze in valore assoluto dei valori per ogni *p* variabile tra le unità statistiche.

$$d(x,y) = \max\left(|x_i - y_i|\right)$$

Infine, la distanza di *Canberra*, proposta da *Lance* e *Williams*, si caratterizza per essere la sommatoria di un rapporto per ogni valore delle *p* variabili attribuito alle *n* unità statistiche. Al numeratore vi è la differenza in valore assoluto mentre, al denominatore, la somma dei valori; in formula (Lance, 1966).

$$d(x,y) = \Sigma\frac{|x_i - y_i|}{|x_i + x_i|}$$

La distanza euclidea ha delle buone performance nel momento in cui i *cluster* sono densi e ben separati in quanto, a causa della potenza al quadrato, è fortemente influenzata da differenze elevate; peculiarità che non si verifica nel caso della città a blocchi (Jain, 1999).

Le distanze riportate finora sono adeguate a fenomeni quantitativi. Quando si hanno variabili binarie o qualitative è necessario utilizzare altri indici di prossimità. Per le variabili booleane è possibile utilizzare il coefficiente di *Jaccard*, la distanza di *Hamming*, il coefficiente di corrispondenza semplice (SMC) oppure il coefficiente di *Russel* e *Rao* (Lourenco, 2004).

Per variabili qualitative categoriche vi è la distanza di *Gower* (appropriata anche per matrici di dati contenenti sia dati continui che categorici), la misura di *Goodall*, il coefficiente di *Dice* oppure la corrispondenza semplice (Le, 2005).

Successivamente, ottenuta la matrice di prossimità, sarà necessario verificare l'esistenza di *cluster* naturali nella struttura dei dati. Questo può essere fatto sia con procedure inferenziali che con procedure esplorative. In quest'ultimo caso si realizza uno *scatterplot* partendo dalla matrice dei dati e visualizzando la densità di frequenza.

L'analisi di *cluster* gerarchica aggregativa ha come passaggio successivo quello di vedere applicare un algoritmo. Gli algoritmi di *clustering* gerarchico agglomerativo più diffusi sono il *single linkage*, il *complete linkage*, l'*average linkage*, il *Ward's method*, il *centroid method* e il *median method*.

Tra questi, la caratteristica principale che li differenzia è il modo in cui è definita la distanza, la similarità o la dissimilarità tra i *cluster* aggregati.

Il metodo del legame singolo (o *single linkage*) misura la distanza tra due *cluster* come la distanza minima tra un punto di un *cluster* e un punto nell'altro *cluster* (Figura 37). In principio, ogni unità statistica è considerata come un *cluster* a sé stante. Sulla base della matrice di prossimità basata su distanze o indici di similarità si calcola in modo iterativo la distanza tra i *cluster* unendo quelli con distanza minore o con similarità maggiore, fino a raggiungere una situazione per la quale si avrà un solo *cluster*. Un metodo di visualizzazione di questo genere di algoritmo è il dendogramma. Per definire il numero di *cluster* ideale, si realizza un "taglio" laddove, tra le diverse distanze minime, si riscontra quella maggiore. Questo algoritmo presenta un criticità nota come effetto catena in quanto può determinare un risultato per il quale vi sono dei *cluster* con unità molto distanti fra loro, quindi, non densi e compatti. In questi casi, graficamente, la distribuzione delle unità è molto simile a una "catena" o, in altre parole, a una successione di punti linearmente distribuiti. Tuttavia, qualora ci dovessero essere dei gruppi di dati distribuiti in lunghezza, questo modello potrebbe performare meglio di altri.

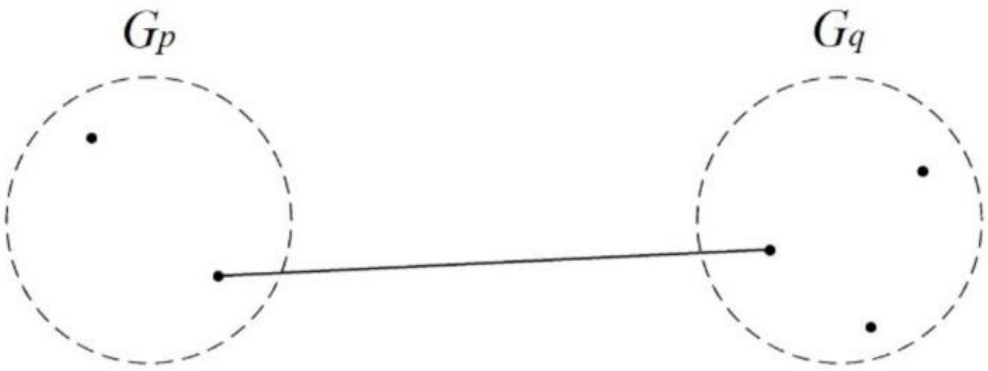

Figura 37 Metodo del legame singolo

L'algoritmo del legame completo (o *complete linkage*) misura la distanza tra due *cluster* come la distanza massima tra un punto in un *cluster* e un punto nell'altro *cluster* (Figura 38). In altre parole, si prende la distanza tra due punti più lontani, uno in ciascun *cluster*. Così come per il metodo del legame singolo, anche in questo caso, la genesi si basa sulla condizione per cui ogni unità statistica rappresenta un *cluster*. Aggregando sempre i gruppi in cui la distanza massima è minore, si giunge alla situazione finale in cui si ha un solo *cluster*. Allo stesso modo, si può usare il dendogramma sia per visualizzare il processo dell'algoritmo sia per definire il numero di *cluster* ottimale laddove ci sia la distanza maggiore tra i rami. Questo, così come per il legame singolo, può essere una significativa differenza o dissimilarità tra *cluster*. Una particolarità di questo algoritmo è l'ottima capacità che ha nel determinare dei gruppi molto compatti soprattutto di forma circolare.

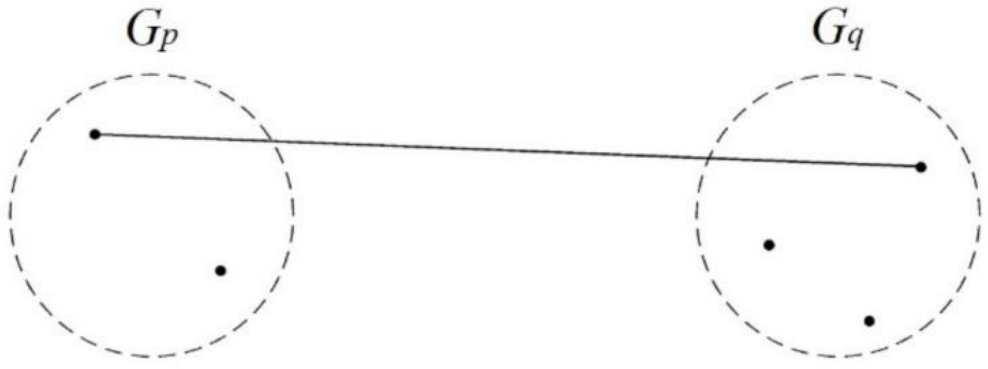

Figura 38 Metodo del legame completo

Il metodo del legame medio (*between-group linkage*) si basa sulla medesima procedura vista finora con il *single linkage* e il *complete linkage* ma le distanze tra *cluster* sono definite come la media aritmetica dele distanze di ciascuna unità di un gruppo con ciascuna unità dell'altro gruppo (Figura 39).

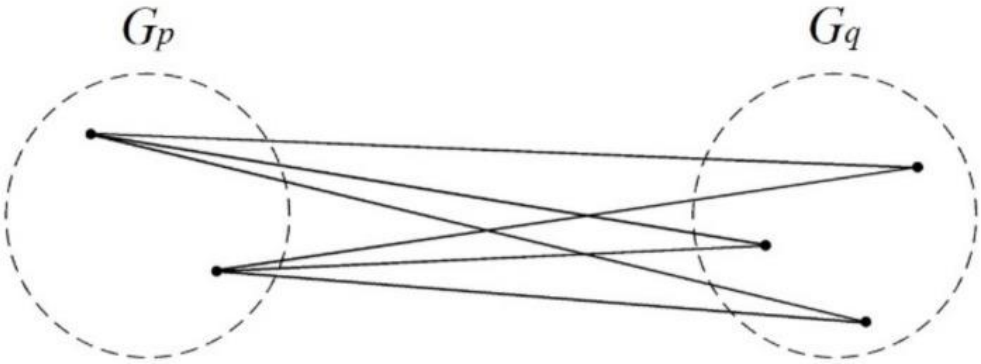

Figura 39 Metodo del legame medio (between-group linkage)

Il metodo del legame medio (*within-group linkage*) si caratterizza per essere basato su di una distanza tra due *cluster* pari alla distanza media tra ogni punto dei *cluster*, considerando anche la distanza media che intercorre tra due unità presenti nello stesso *cluster* (Figura 40).

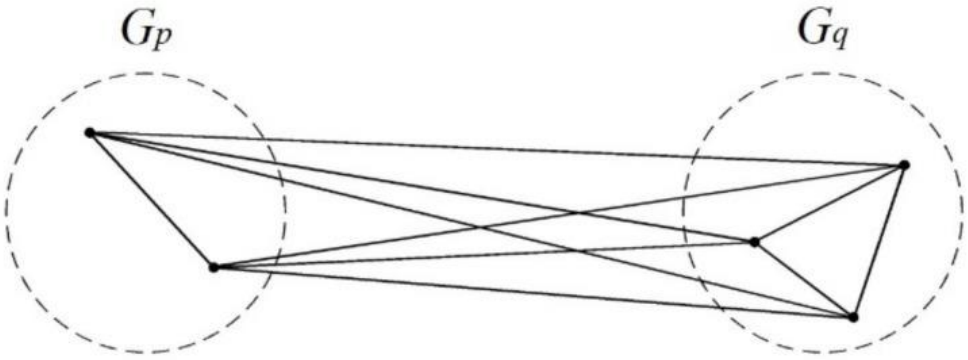

Figura 40 Metodo del legame medio (within-group linkage)

Il metodo del centroide (Figura 41) si basa sulla distanza tra gruppi definita come la distanza tra i rispettivi centroidi. Il centroide è il punto medio di un insieme di punti in uno spazio multidimensionale e può essere diverso a seconda di quale si considera. Fondamentalmente può essere definito in base a diversi approcci.

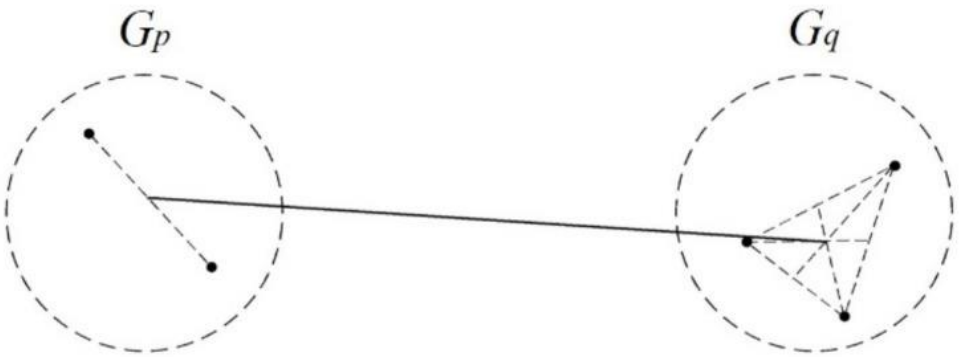

Figura 41 Metodo del centroide

Il modello di *clustering* della mediana (Figura 42) si distingue dagli altri algoritmi in quanto è previsto il calcolo della mediana per i *cluster* che si formano, questi diventano il nuovo centro del *cluster*. Successivamente l'algoritmo unirà i due *cluster* che hanno la mediana più vicina.

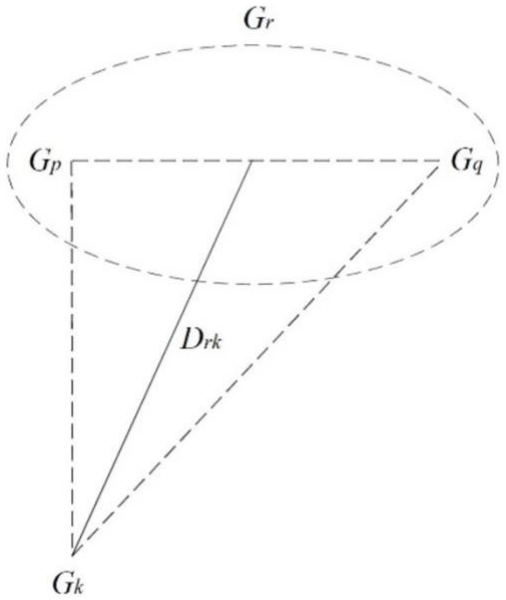

Figura 42 Modello della mediana

L'algoritmo di *Ward* (o della varianza minima) deriva dall'*ANOVA* (analisi della varianza) e ha come obiettivo quello di definire dei *cluster* per i quali ogni *step* aggregativo del processo sia basato sulla minimizzazione della varianza *within* (indice di coesione interna dei *cluster*). Per spiegare in modo più chiaro quello che il modello fa è necessario comprendere la scomposizione della varianza totale ossia un misura che descrive la variabilità dei dati. La varianza totale (T) delle *p* variabili si scompone in varianza *within* (W) e varianza *between* (B).

$$T = W + B$$

La varianza *within* nei *cluster* è data dalla sommatoria delle distanze quadrate tra ogni punto del *dataset* e il centroide del *cluster* a cui appartiene. Una bassa varianza within indica che i punti all'interno dei rispettivi *cluster* sono molto simili tra loro.

La varianza *between* è pari alla somma delle distanze quadrate tra i centroidi dei *cluster* e il centroide di tutti i dati. Tale misura indica quanto sono distati tra loro i *cluster*. Nel momento in cui si dovesse configurare un *clustering analysis* con un alto valore di varianza *between* questo indicherebbe che i *cluster* definiti sono molto diversi fra loro.

Il procedimento esplorativo volto a definire il numero di *cluster* ideale nonché la loro composizione è sempre quello visto per il metodo del legame singolo e del legame completo. Ciò per il quale il metodo di *Ward* differisce è la considerazione della distanza tra *cluster* la quale, in questo algoritmo, equivale alla minimizzazione delle varianza *within*. Quindi, ogni qual volta che si aggregano due *cluster* si procede cercando la combinazione di *cluster* che determini il minor incremento della *Within-cluster Sum of Squares* (WSS) ossia la varianza *within* (Bu, 2020; Murtagh, 2014; Ward Jr, 1963).

In questo lavoro di tesi non sono stati applicati algoritmi di *clustering* gerarchico scissorio, di conseguenza, non saranno oggetto di approfondimento. Per determinare il numero ottimale di *cluster* nel caso in cui siano stati applicati degli algoritmi di tipo *bottom-up* è possibile utilizzare tre procedure.

Il primo riguarda l'applicazione di test di separazione tra *cluster*, i quali sono atti a verificare la significatività della distanza tra i centroidi.

Come è già stato anticipato, un altro modo per poter definire il partizionamento ottimale è l'utilizzo del dendogramma con la tecnica dell'α-taglio riportato in forma esemplificativa nella Figura 43. Mediante l'ispezione del dendogramma si appone un taglio in corrispondenza della distanza maggiore che intercorre tra un'aggregazione e quella successiva.

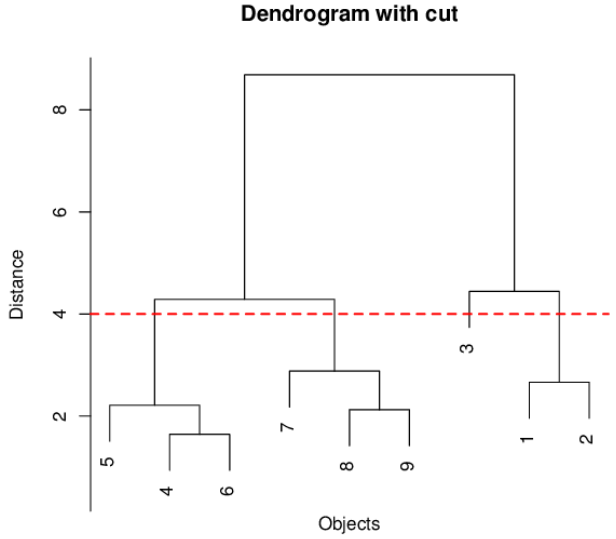

Figura 43 Esempio di dendogramma con alpha-taglio

Con questo strumento sarà necessario comunque dover fare ulteriori considerazioni valutando l'eventuale presenza di *cluster* naturali nonché ulteriori ed eventuali esigenze di interpretabilità (Pereira, 2013).

Infine, anche per poter valutare il modello con una metrica, utile soprattutto per effettuare confronti tra algoritmi, è possibile utilizzare degli indici sintetici come il *Calinski-Harabasz* il quale valuta la qualità del *clustering* calcolando un rapporto tra la dispersione tra *cluster* e la dispersione all'interno del *cluster* moltiplicato per un fattore di scala relativo al numero di *cluster* e al numero di punti. Questo indice è simile all'indice R^2 di varianza spiegata ma aggiunge un fattore di scala. Un valore più altro di questo indice indica un *clustering* migliore. Pertanto, maggiore sarà il valore dell'indice, migliore sarà la scelta del numero di partizioni.

L'indice di varianza spiegata R^2 si calcola mediante la seguente formula:

$$R^2 = \frac{SSB}{SST}$$

dove SSB è l'acronimo di *between sum of squares* e SST *total sum of squares*. Questo indice misura la proporzione della varianza totale dei dati spiegata trai i *cluster*. Anche in questo caso, valori maggiori indicheranno un *clustering* più efficace in quanto ogni *cluster* risulterà essere più compatto e isolato.

Altre metriche di valutazione sono il *Silhouette score*, l'indice di *Dunn* e l'indice di *Gap*. Per la valutazione degli algoritmi di *clustering* in questo elaborato è stato utilizzato l'indice di *silhouette,* il quale definisce quanto bene ogni punto si adatta al suo *cluster* rispetto ai *cluster* più vicini. Rispetto all'indice di *Calinski-Harabasz,* questo *score* ha dei limiti in quanto varia tra -1 e 1 con valori più alti che indicano un *clustering* migliore.

La formula del coefficiente di *silhouette* è la seguente:

$$s = \frac{b - a}{\max(a, b)}$$

dove a e la distanza media tra un'unità e tutti gli altri punti nello stesso *cluster*. Questo può indicare quando un'unità sia adeguata a quel *cluster*. *b* è la distanza media tra un punto e tutte le unità del *cluster* più vicino. Questo elemento del coefficiente mira a quantificare quanto adeguato fosse stato, se l'unità oggetto di calcolo fosse stata in un altro *cluster* ad essa più vicina. Minore è questo valore, maggiore sarà la prossimità con un altro *cluster*. Il coefficiente di *silhouette* per un intero *dataset* è dato dalla media dei coefficienti calcolati per ogni punto dello stesso (Liu, 2022).

7.2 Algoritmi di clustering partizionali

Sebbene il *clustering* gerarchico sia comunemente considerato superiore in termini di qualità dei risultati, la sua applicabilità è ostacolata dalla sua *quadratic time complexity*. La criticità sottolineata è dovuta ai calcoli richiesti per giungere a un risultato ottimale (Steinbach, 2000).

In risposta al problema, i metodi di *clustering* non gerarchici (partizionali) si caratterizzano proprio per avere un solo partizionamento in *g cluster* delle *n* unità, dove *g* è fissato a priori. L'assegnazione delle entità ai vari gruppi viene realizzata ottimizzando una funzione obiettivo, la quale è comunemente espressa attraverso la suddivisione della devianza totale. L'obiettivo è generalmente quello di trovare una suddivisione che massimizzi la coesione interna, ovvero minimizzi la devianza all'interno del gruppo. Inoltre, gli algoritmi tendono ad ottenere un risultato ottimale a livello locale.

Gli algoritmi di *clustering* partizionale si divide in *fuzzy clustering*, *hard o crisp clustering* e *mixture resolving*. L'approccio di *clustering* partizionale presenta una serie di potenziali benefici. In primo luogo, viene meno il limite tale per cui, nei metodi gerarchici, non è possibile separare due unità che sono state precedentemente aggregate. Con gli algoritmi di *clustering* partizionali, ad ogni nuovo *step*, verranno assegnate le unità ai *cluster* basandosi esclusivamente sulla funzione di minimizzazione della varianza *within-groups*, il che può portare a risultati diversi al cambiare del numero di *cluster*. Questo supera le difficoltà che possono derivare dal raggruppamento "errato" di unità disomogenee nei primi stadi di un processo gerarchico. Grazie alla loro natura iterativa e all'ottimizzazione per un singolo valore di *k*, questi metodi sono computazionalmente efficienti e non necessitano della realizzazione preliminare di una matrice di prossimità tra le unità. Pertanto, si rivelano particolarmente adatti quando il numero di unità da analizzare è molto grande, situazione in cui l'impiego di metodi gerarchici risulterebbe estremamente dispendioso in termini computazionali. Inoltre, gli algoritmi non gerarchici possono essere la scelta ideale quando l'obiettivo della ricerca è la caratterizzazione delle specificità dei *cluster* (ad esempio, attraverso il calcolo dei centroidi e le misure di variabilità dei *cluster*) piuttosto che l'analisi del comportamento delle singole unità durante i successivi stadi dell'aggregazione gerarchica.

La tecnica di *clustering* nota come *fuzzy* (sfocato), introdotta da *Zadeh* con "*Fuzzy Sets*" nel 1965, si caratterizza per raggruppare in *cluster* in modo tale che ogni elemento appartenga simultaneamente a più insiemi *fuzzy* ossia i *cluster*. La distribuzione dei punti

dati a più *cluster*, ognuno con diversi gradi di coinvolgimento, determina quindi una relazione non binaria. Questo metodo di raggruppamento si dimostra efficace per l'aggregazione di unità distribuite in modo incerto e in gruppi con confini poco delineati e quindi sovrapposti. Le informazioni relative al grado di adesione di un punto a un *cluster* possono illuminare la relazione intrinseca del punto con i vari *cluster* (Zadeh, 1965).

I modelli di *clustering Fuzzy* possono essere divisi a loro volta in gerarchici e non gerarchici. Quelli gerarchici si distinguono per un procedimento a due fasi: nella prima, analogamente ai metodi di *hard clustering* si quantifica la misura di somiglianza per le coppie di elementi; nel secondo *step* vi è l'assegnazione delle unità ai gruppi formati attribuendo anche un livello di appartenenza per ognuno. Sono esempi di algoritmi di *clustering fuzzy* gerarchici l'algoritmo della sintesi di più partizioni, l'algoritmo dei ricoprimenti sfocati e l'algoritmo del legame medio sfocato. I modelli di *clustering* non gerarchico *fuzzy*, come quelli gerarchici, sono molto simile alla loro controparte non *fuzzy*. In particolare, il processo alla base è costituito dalla definizione a priori del numero di partizioni per poi attribuire a ogni gruppo ciascuna unità mediante un iterazione. Con l'approccio *fuzzy*, gli algoritmi variano in base alla funzione obiettivo selezionata e, pertanto, alla procedura iterativa utilizzata per determinare i livelli di appartenenza di ciascuna unità per ogni *cluster*. La funzione obiettivo dà luogo a un valore di errore per ogni possibile soluzione. Quest'ultimo è espresso sotto forma di efficacia o costo considerando la distanza che intercorre tra i punti dati e i rappresentativi delle partizioni. Questo approccio, essendo un problema di ottimizzazione, determina delle limitazioni strutturali in quanto la soluzione ottimale è proprio il risultato ottimale raggiungibile con la funzione stessa. Tra i modelli di *clusterizzazione* partizionale *fuzzy* spicca il metodo delle *k* medie o *fuzzy k-means* in quanto è uno dei più utilizzati (La Rocca, 2006).

Di seguito, verranno approfonditi i modelli di *clustering* classificati come *hard* oppure *crisp clustering* in quanto ogni unità statistica è attribuita esclusivamente ad una partizione.

Le tecniche di *clustering* basate sulla densità si basano sulla caratteristica per la quale nello spazio, le unità statistiche possono essere inquadrate come delle zone dense circondate da spazi a densità minore. Proprio grazie a ciò funzionano gli algoritmi di *clustering* rientranti in questa categoria. Fondamentalmente sono in grado di rilevare le aree in cui vi è una concentrazione di punti. Gli elementi che sono situati al di fuori delle aree ad alta densità sono categorizzati come rumore. Il *DBSCAN* (*Density-Based Spatial*

Clustering of Applications with Noise) opera mediante l'esplorazione ed esaminazione dell'ambiente per ciascuna unità statistica nello spazio. Fondamentalmente, per ogni punto viene realizzato un conteggio di quanti punti si trovano in un'area delimitata in partenza. Se, intorno al punto in considerazione si osservano un numero di punti superiore a una certa soglia ossia un parametro del modello allora sono definiti come punti *core* considerabili come parte di un partizionamento. In base a questo concetto di densità si formano i *cluster*; infatti, qualora un punto non sia definibile come *core* ma è nell'area considerata per il punto *core,* allora questa unità statistica verrà attribuita a quel *cluster*. Altrimenti, se non si verifica neanche questo, come già detto precedentemente, l'unità statistica non verrà assegnata ad alcun gruppo. Altri modelli di *clustering* basati sula densità sono l'*OPTICS* (*Ordering Points to Identify the Clustering Structure*) e l'*HDBSCAN* (*Hierarchical Density-Based Spatial Clustering*) (Ester, 1996; Birant, 2007).

Il metodo *K-Means*, inizialmente concepito da *MacQueen,* è una delle tecniche di partizionamento non gerarchico più diffusa. Il fine principale è quello di suddividere le n entità statistiche in un predeterminato numero di *cluster*. Questo si verifica grazie alla riduzione al valore minimo possibile della varianza all'interno del *cluster* intesa come la distanza che intercorre tra le unità e il centroide calcolato come media (per questo *means*) della partizione. Rientra tra le tecniche di *square error clustering* in quanto l'obiettivo finale è il raggiungimento di un partizionamento tale per cui ci sia il valore minore dell'errore quadratico totale.

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

Nella sua forma base, l'algoritmo si basa sulla distanza euclidea. Nel processo di *clustering K-Means,* si inizia formando *k* centroidi, definiti come semi, a partire dal *set* di dati iniziale. I centroidi possono essere scelti in base a diverse metodologie, tendenzialmente, il criterio principale di scelta è la distanza che vi è tra gli stessi la quale si prova a massimizzare. Successivamente, si forma una suddivisione inziale formata da *k* gruppi in cui ogni dato viene assegnato al *cluster* più vicino basandosi sulla distanza euclidea minore rispetto ai centroidi. Una volta che i dati sono stati assegnati ai *cluster*, i centroidi vengono aggiornati in base ai risultati del raggruppamento. In altre parole, si ricalcola il punto medio tra le diverse unità della partizione così come la distanza euclidea tra i nuovi centroidi e gli elementi del *cluster*. Se si dovesse verificare lo scenario per il

quale la distanza euclidea tra l'elemento e il centroide del *cluster* nel quale è stato assegnato in fase iniziale, verrà ricollocato laddove sia stata calcolata la minor distanza. Avvenuto ciò, il modello ricalcola il centroide per entrambi i *cluster*, il primo che ha visto perdere questa entità nonché la partizione che ha guadagnato l'unità statistica. Questa procedura viene iterata fino a quando non si verifica alcuna modifica significativa nell'assegnazione dei dati ai *cluster*, ossia fino a quando il modello non converge e quindi non osservando più un passaggio di dati da un *cluster* all'altro. Per ridurre il quantitativo di tempo conseguente al processo iterativo dell'algoritmo è possibile stabilire una regola d'arresto la quale determinerà la fine del processo quando si verifica la convergenza del modello, se la distanza ricalcolata tra ogni unità e i centroidi non supera un certo valore oppure ancora definendo a priori un numero di iterazioni oltre il quale il modello non dovrà andare. La distanza euclidea è la distanza più comunemente utilizzata in quanto non determina limiti al conseguimento della convergenza da parte dell'algoritmo mediante le iterazioni. Inoltre, un criticità del metodo delle *k* medie è l'influenza dell'ordine dei dati nel *dataset* iniziale in merito al risultato di *clustering* finale (Anderberg, 1973). La modalità più comune per la ricerca del numero ottimale di *cluster* è la ripetizione del modello cambiando il numero di *k cluster* ogni volta per poi utilizzare un indice sintetico di valutazione della bontà del clusterizzazione ottenuta. Ovviamente, il valore ottimale di *cluster* sarà il risultato della valutazione migliore. Altre metodologie sono l'utilizzo del test di significatività (inferenza statistica), l'applicazione di un algoritmo di clusterizzazione gerarchico oppure la ricerca delle mode mediante un istogramma *p*-dimensionale. Per quanto concerne la scelta dei centroidi iniziali, questa, può essere attuata mediante diverse tecniche. In primo luogo, possono considerarsi le prime unità statistiche del *dataset* come poli del partizionamento base. In alternativa, si può porre in essere un campionamento casuale semplice mirato a selezionare delle osservazioni utilizzate come poli iniziali. Entrambe le tecniche definite finora non risulterebbero essere rappresentativamente adeguate all'intera matrice dei dati. Per risolvere il problema, sarebbe necessario ricercare dei semi iniziali debitamente distanti. Per questo sono state strutturate delle procedure per cui è calcolata la distanza euclidea tra i semi, la quale non deve essere inferiore a un determinato limite (MacQueen, 1967; Anderberg, 1973; Kutbay, 2018; Pitafi, 2023).

Il *Bisecting K-means* è una variante del tradizionale algoritmo *k* medie e si basa su un approccio diverso. Il modello inizia con tutte le unità statistiche in un unico *cluster* per

poi, in modo iterativo, definisce un *cluster* da dividere in due gruppi. La divisione avviene mediante il processo del *k-means* finora discusso con un numero di *cluster* pari a due. Questo approccio di natura scissoria è ripetuto finché non si raggiunge il numero di *cluster* indicato a priori. Il modo con cui viene scelto il *cluster* può essere scelto; quello più frequente è il criterio della maggiore somma delle distanze quadrata (SSQ) ossia una misura dell'eterogeneità del *cluster* (Ansarifar, 2018; Steinbach, 2000)

Il *K-Prototype* è un'estensione del modello *K-Means* ed è stato progettato per clusterizzare dei *dataset* costituiti sia da dati categorici che da dati numerici. Dato l'approccio su cui si basa il *k-means*, le performance migliori si avranno con dati numerici, in particolare con quelli continui. Il modello di *Huang*, invece, divide i dati in numerici e in categorici. Mentre i primi verranno clusterizzati con un algoritmo *k-means*, le variabili qualitative della matrice dei dati saranno gestire da un altro algoritmo chiamato *k-modes* (k-mode) il quale si basa non sul valore medio come centroide di ogni *cluster* ma sulla moda ossia la categoria con frequenza maggiore. L'algoritmo ha un processo simile a quello del *k-means* qui riportato, ciò che cambia sono i punti di partenza e parte del processo iterativo con il quale si ricalcola il centroide di ogni *cluster*. All'inizio, ogni punto centrale del *cluster* (chiamato prototipo) contiene sia dati numerici che categorici. L'assegnazione di ogni punto al *cluster* avviene sia calcolando la distanza euclidea per le variabili quantitative sia considerando la discrepanza categorica che vi è tra le *feature* qualitative, la quale può essere ottenuta in vario modo contando il numero di categorie che non corrispondono. Infine, l'aggiornamento dei prototipi avviene calcolando la media dei valori numerici dei punti in ogni *cluster* come per il *k-means* e scegliendo la modalità più frequente delle variabili qualitative dei punti in ogni *cluster*. Il processo si ripete in modo iterativo fino a raggiungere la regola di arresto come per l'algoritmo *k-means* (Huang, 1997).

8 Model Selection e Valutazione

Gli algoritmi testati per questo lavoro di tesi sono stati 13: nello specifico sono il *K-Means*, il *K-Prototypes*, il *Bisecting K-Means*, il *K-Medoids*, l'*Affinity Propagation*, il *Mean-shift*, lo *Spectral clustering*, il modello gerarchico di *Ward*, *l'Agglomerative clustering*, il *DBSCAN*, l'*OPTICS*, il *Gaussian Mixtures* e il *BIRCH* (Scikit-learn).

Tra questi, molti non sono andati a buon fine a causa di problemi computazionali. Fondamentalmente il quantitativo di RAM utilizzata (25 gb con *Colab Pro*) non era sufficiente per il processo di diversi algoritmi di *clustering*. Gli unici risultati ottenuti dai modelli di *clustering* applicati arrivano dal *K-Means*, *K-Prototypes*, *Bisecting K-Means*, *DBSCAN* e *Gaussian Mixtures*. Tra questi, dei risultati effettivamente validi e utilizzabili ai fini dell'obiettivo di ricerca di questo contributo sono principalmente tre. Un'ulteriore considerazione che ha determinato la scelta del modello per questo lavoro è stata la struttura della matrice di dati sulla quale sono stati applicati i modelli di *machine learning* per la clusterizzazione. Come già presentato e ribadito, il *dataset* comprendeva sia variabili quantitative che qualitative. Nonostante queste ultime siano state opportunamente codificate come riportato nel capitolo 6 riguardante la parte di pre-processamento, la natura delle variabili restava tuttavia qualitativa. Inoltre, la maggior parte delle variabili del *dataset* codificato erano di natura dicotomica quindi non variabili quantitative continue. La distanza euclidea utilizzata come distanza standard nei modelli partizionali di *hard/crisp clustering* come il k-medie, non è indicata per variabili quantitative non continue.

Nonostante ciò, i risultati che hanno senso per le operazioni di *business* nonché per l'obiettivo di questo studio sono dati dal *K-Means*, *Bisecting K-Means* e dal *K-Prototypes*. Per il *K-Means* sono stati effettuati anche ulteriori test mirati ad utilizzare distanze o misure di similarità preferibili con un *dataset* di questo genere; ad esempio, la distanza di *Gower*. Quest'ultima, così come tutte le altre (ad eccezione della distanza euclidea), ha determinato un problema computazionale dovuto alla *RAM* insufficiente come già esposto prima per gli altri algoritmi.

L'algoritmo *K-Means* è stato realizzato partendo dalla scelta ottimale del numero di *cluster* basata sull'*elbow method* con la *Within-cluster Sum of Squares* (*WCSS*) ossia la somma delle distanze al quadrato tra ogni punto del *cluster* e il suo centroide. Essa è una misura comunemente utilizzata per valutare la varianza in funzione del numero di *cluster*. Valori di *WCSS* più bassi indicano che i punti all'interno del *cluster* son più vicini tra loro

e di conseguenza, c'è una migliore coesione interna. Il metodo del gomito si applica al grafico delle *WCSS* in funzione del numero di *cluster*. Il punto di svolta viene scelto considerando il decremento della varianza nel caso in cui diventa meno significativo rispetto ai *cluster* aggiuntivi. In altre parole, il numero di *cluster* ideale corrisponde al punto sul grafico in cui un incremento ulteriore di *k cluster* non determina un miglioramento significativo nella riduzione del valore della *WCSS* (Thorndike, 1953). Considerando il grafico per il *K-Means* (Figura 44), il numero di *cluster* ideale è stato scelto in corrispondenza di due.

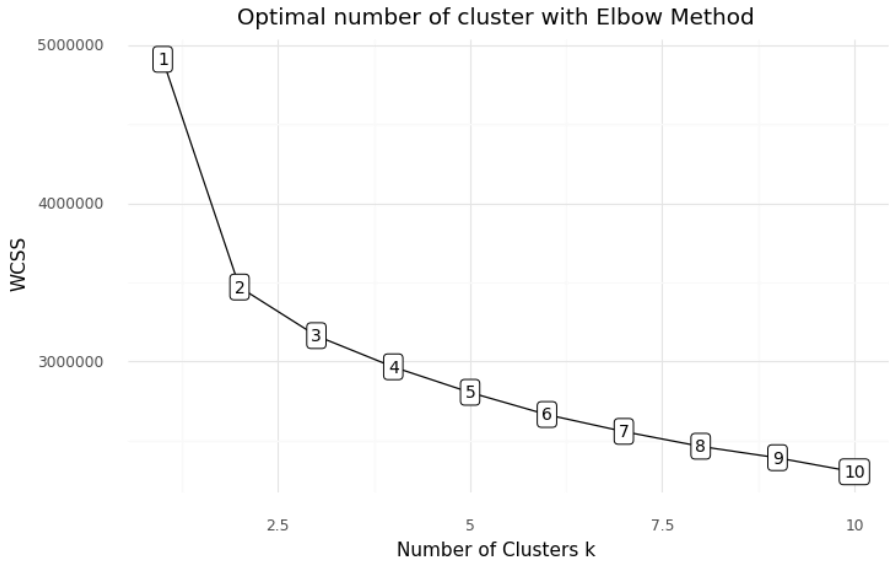

Figura 44 K-Means e metodo del gomito

L'algoritmo è stato poi applicato con i seguenti iperparameri: un numero di *cluster* pari a due, l'inizializzazione *k-means++* per la scelta dei centroidi iniziali al fine di evitare la convergenza subottimale, un numero di iterazioni massimo pari a 300 e *n_init* pari a dieci ossia il numero massimo di iterazioni iniziali con centroidi diversi.

Il *silhouette score* medio è pari a 0,26 mentre l'indice *Calinski-Harabasz* è di 59318,57. È possibile asserire che il modello raggruppi in modo ragionevole le unità statistiche sebbene ci siano delle marginali assegnazioni errate. In altre parole, è presente un sufficiente grado di separazione dei *cluster* sebbene ci sia possibilità di miglioramento. Per un ulteriore valutazione, è stata applicata la tecnica statistica *t-SNE* per visualizzare un *dataset* multidimensionale con uno *scatterplot* bidimensionale. Sebbene la precisione della tecnica possa essere discutibile, il risultato è sufficientemente soddisfacente data una divisione abbastanza netta malgrado una sovrapposizione tra *cluster* (Figura 45).

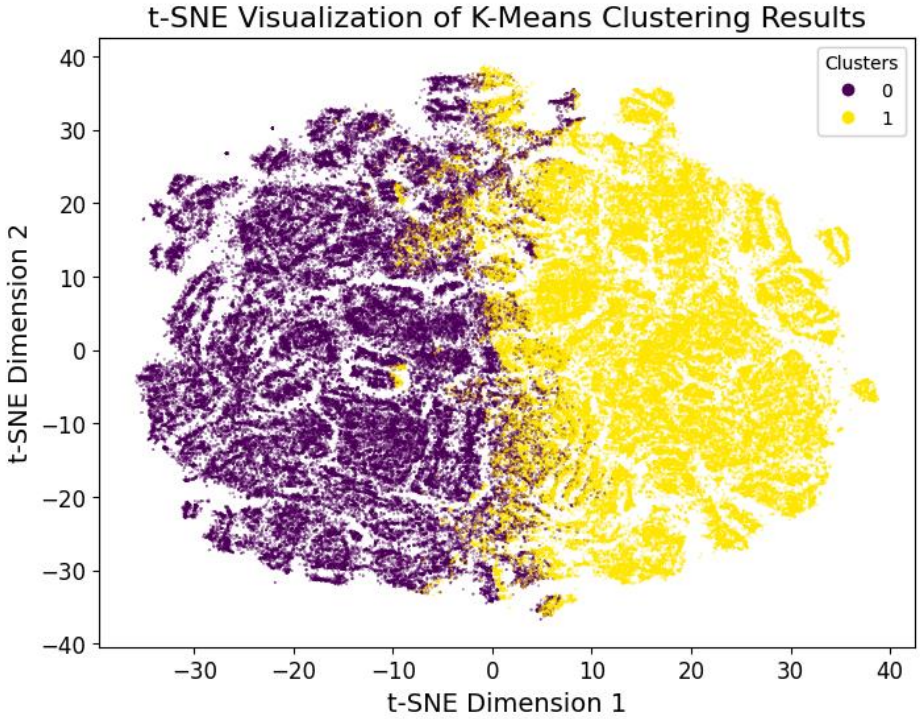

Figura 45 Scatterplot t-SNE del K-Means

I risultati del *Bisecting K-Means* risultano essere molto simili sia in termini di metriche di valutazione che di visualizzazione dei *cluster*.

L'algoritmo *K-Prototypes* è stato realizzato con il medesimo processo. Pertanto, partendo dal grafico su cui è stato applicato il metodo del gomito, è stato definito il numero ottimale di *cluster* che in questo caso, come per il risultato dell'algoritmo k-medie, è stato pari a due. A differenza dell'algoritmo *k-means* in cui è stata utilizzata la *WCSS*, qui è possibile utilizzare la funzione di costo propria del modello, la quale tiene conto sia delle variabili categoriche (considerando le distanze) che delle variabili numeriche (in base alle similitudini). La funzione di costo è una funzione di dissimilarità tra i punti del *cluster* e il suo centroide. Pertanto, l'obiettivo è quella di minimizzarla e anche in questo caso, viene scelto un numero di *cluster* per il quale, un aggiunta di *k cluster* non determini un miglioramento significativo in questa direzione. Come si evince dal grafico (Figura 46), anche qui il numero ottimale di *cluster* è pari a due. Sebbene sia il modello statisticamente più rigoroso per l'obiettivo finale del contributo, presenta delle criticità nella valutazione dei risultati in quanto gli indici sintetici riportati per il metodo delle *k-medie* sono adeguati solo per le variabili numeriche. Tra le possibili soluzioni utilizzate per la valutazione di questo modello vi è la distanza di *Gower* la quale, tuttavia, come già riportato, presenta

dei problemi computazionali di *RAM*. Pertanto, la valutazione è avvenuta su base grafica mediante un altro tentativo di *data* visualisation realizzato con la tecnica *t-SNE*.

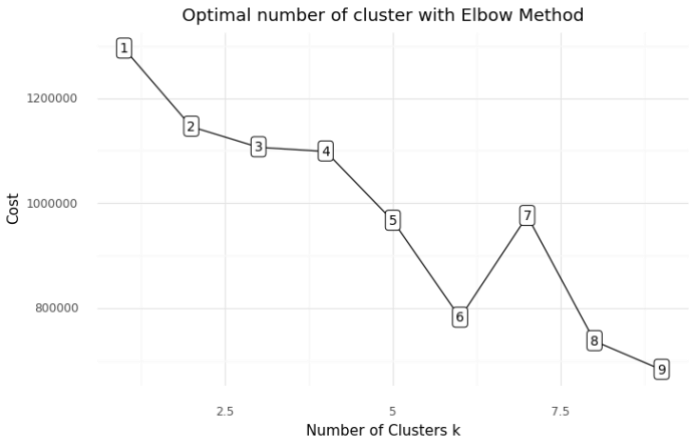

Figura 46 K-Prototype e metodo del gomito

Gli iperparametri utilizzati in definitiva dopo il *fine tuning* sono stati un numero di *cluster* pari a due, un numero di inizializzazioni pari a tre, il metodo di inizializzazione "*Huang*" e *verbose* pari a uno. Sebbene sia stato testata anche l'inizializzazione "*Cao*" basata sulla densità, questa non ha determinato miglioramenti. Inoltre, l'iterazione migliore è stata la seconda restituendo anche i seguenti centroidi *[['-0.1676', '-0.3403', ' 0.0003', '-0.0042', 'Male', '1', '1', '0', '0', '0', '0', 'Regular', 'Resident', '0', 'not defined', '1', 'Natural fibers', 'shirts', 'Blu', 'FORMAL', 'Boutique', 'China', '1.0', '5.0', '2016.0', '0']; ['1.0275', '2.2317', '0.0009', '0.0053', 'Male', '1', '1', '1', '0', '0', '0', 'Regular', 'Resident', '0', 'newsletter', '1', 'Natural fibers', 'suits', 'Blu', 'FORMAL', 'Boutique', 'Italy', '1.0', '5.0', '2019.0', '0']].*
In termini grafici (Figura 47), la divisione dei *cluster* sembrerebbe essere piuttosto complessa, frammentata, sovrapposta e quindi con dei *cluster* poco isolati e sensi. Tuttavia, risulta essere ancora possibile identificare due aree rappresentative dei *cluster*.

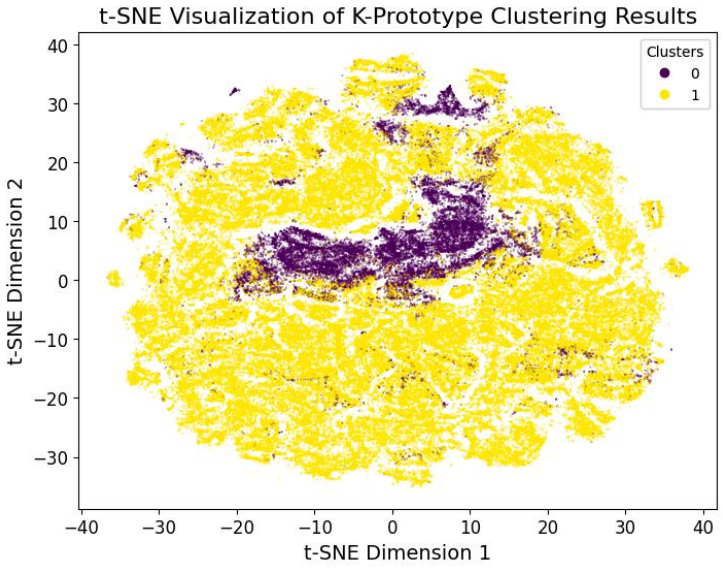

Figura 47 Scatterplot t-SNE del K-Prototypes

Il modello selezionato per la realizzazione della domanda di ricerca è stato il *K-Prototypes* sia per i risultati positivi sia per il maggiore rigore statistico presente con l'applicazione dell'algoritmo su di un *dataset* come questo.

9 Interpretazione e Limiti

Analizzando i dati per ogni *cluster* generato dal *K-Prototypes* mediante valori medi, il calcolo delle mode e rapporti di dati ne risultano due *marketing personas* con caratteristiche sufficientemente diverse. La prima *marketing personas* si basa su di un consumatore residente in Cina descritto dalla seguente infografica (Figura 48).

Figura 48 I[a] Marketing Persona generata con Machine Learning

Invece, la seconda *marketing personas* è descrittiva di un consumatore europeo e sebbene potrà sembrare molto simile alla prima è importante considerare che diverse caratteristiche sono spesso correlate con comportamenti d'acquisto completamente diversi. Basti pensare alla provenienza culturale; è dimostrato da diversi studi che un consumatore asiatico non sia spinto dalle medesime esigenze e dai medesimi bisogni, così come non abbia le stesse motivazioni alla base delle proprie preferenze rispetto a un consumatore europeo o americano (Manzato, 2019). Le similitudini tra le *marketing personas* sono anche figlie di un'analisi di *clustering* realizzata con dei *dataset* non nati e sviluppati per questo obiettivo ultimo. L'adattamento e l'operazione di *feature selection*

per la scelta di varabili utili al fine di generare delle *marketing personas* sono comunque condizionati e limitati dai dati di partenza.

La seconda *marketing persona* è descritta dalla seguente infografica (Figura 49).

Figura 49 II[a] Marketing Persona generata con Machine Learning

Le *marketing persona* di cui sopra sono state realizzate con i soli dati presenti nella matrice dell'analisi di *clustering*. Come è stato elaborato nel capitolo 2.3, questo strumento strategico di *marketing* è costituito soprattutto da variabili comportamentali, dai bisogni, dai *pain points* e tutti ciò che concerne il *decision journey* del consumatore. Il *dataset* è oggettivamente limitante in quanto contiene solo alcune di queste variabili mancando di informazioni sull'età, l'occupazione, il salario, ecc. Quindi, uno sviluppo futuro potrebbe essere la ripetizione dell'analisi con dei dati raccolti e manipolati proprio per questo fine. In aggiunta, si potrebbero realizzare delle *marketing personas* dividendo la base di clienti in relazione della variabile geografica, della disponibilità a pagare, ecc. facendo così un clusterizzazione a due fasi. Un altro limite del modello è il numero ristretto di *test* che sono stati attuati cambiando gli iperparametri dell'algoritmo (*fine tuning*). Questo è avvenuto sia per ragioni di calcolo dell'*hardware* sia per cambiamenti poco significativi in termini di *performance* dei modelli. Verificare quali valori degli

iperparametri sono maggiormente adeguati e determinanti allo scenario di applicazione potrebbe essere fondamentale ai fini delle *performance*; in quanto, migliorare il processo scegliendo i giusti iperparametri rilevanti per l'algoritmo, potrebbe migliorarne la definizione dei *cluster*.

Allo stesso modo, l'elevato quantitativo di *missing value* può essere considerato come un'altra restrizione del contributo qui riportato.

La determinazione di due *cluster* sulla base dei risultati degli algoritmi di *machine learning* rende possibile un tentativo di confronto con l'azienda e quanto fatto negli anni precedenti. Se, in passato, l'impresa avesse definito uno o più archetipi di cliente, potrebbe verificare il grado di similitudine tra quanto previsto e quanto raggiunto. Questo risulterebbe essere anche una metodologia di valutazione delle strategie passate di *marketing* confrontando le *personas* utilizzate in principio con quanto risulta dagli algoritmi di apprendimento automatico da questo contributo. Un numero di *cluster* pari a due fa sì che ci sia una focalizzazione mirata su clienti specifici. Riducendo i rischi di aumentare i costi di acquisizione, *retention* e sviluppo su dei clienti non particolarmente remunerativi. In aggiunta, le *personas* riportate potrebbero essere un cliente tipo non ancora studiato e valutato come *target* di campagne specifiche. Infatti, ciò potrebbe essere rappresentativo di un gruppo di individui che sono diventati clienti dell'azienda non come risultato di azioni di *marketing*. Allo stesso modo, il numero di *cluster* ottimale potrebbe essere solo parzialmente identificativo dei clienti tipo dell'azienda. In questo caso, si potrebbe indagare l'attrattività degli altri valutando, di conseguenza, una riorganizzazione del *budget* di *marketing* al fine di accrescere i *KPI* (*Key Performance Indicator*) relativi alle operazoni destinate ai *target* presenti sia nell'analisi dei *cluster* che nella presente strategia del *business*.

10 Implicazioni Manageriali

Lo studio è riuscito a determinare una risposta alla domanda di ricerca e pertanto, risulta essere possibile utilizzare il *machine learning* per realizzare delle *marketing personas*. Essendo uno strumento strategico e di conseguenza, con un impatto rilevante nelle decisioni dell'intera realtà aziendale, è fondamentale però che tale processo venga realizzato nel modo più rigoroso possibile partendo dallo studio delle variabili per concludersi con un test approfondito sui diversi algoritmi e le loro peculiarità.

Le implicazioni manageriali risultano essere molteplici in quanto, come detto, lo strumento delle *marketing personas* aiuta l'azienda non solo nel prendere decisioni di importante rilievo ma anche in diversi fasi della vita di un'azienda. Sono state individuate cinque principali implicazioni di *business*. La prima riguarda un miglioramento della segmentazione di mercato poiché approfondiscono le caratteristiche dei segmenti per cui sono realizzate creando dei profili di consumatori e di conseguenza, facilitando anche la comprensione e l'esecuzione di diverse operazioni all'interno dell'azienda. Inoltre, nel momento in cui ciò viene realizzato mediante algoritmi, si riducono tutti gli errori che derivano dalla generalizzazione dell'esperienza dei *decision maker* all'interno della realtà aziendale. Un'altra implicazione manageriale, naturale conseguenza della precedente, è la possibilità di personalizzare le strategie di *marketing* in modo più puntuale e accurato frutto di applicazione matematica e statistica nella realizzazione degli strumenti di *business* quali le *marketing personas*.

Ottimizzare le strategie di prodotto, creando delle varianti più accurate e cucite sui bisogni e caratteristiche del cliente è un ulteriore aspetto che può essere raggiunto con queste approccio basato sull'intelligenza artificiale applicata nelle dinamiche aziendali. Naturale conseguenza della personalizzazione della strategia di *marketing* e dell'ottimizzazione del prodotto, è il miglioramento dell'esperienza del cliente. In generale, conoscere i bisogni dei clienti, il *decision journey* e i punti di contatto permette di soddisfare il cliente in modo migliore, strutturando ogni singolo aspetto operativo in risposta alle sue esigenze e necessità. In ultima analisi, rendere più precisi i processi determina anche un'ottimizzazione delle risorse in quanto, così facendo, l'azienda può concentrare i suoi sforzi, il *budget* e tutte le altre risorse, in primo luogo, sui clienti più promettenti nonché in modo più profittevole per ognuno conoscendo, con un margine di errore inferiore, le peculiarità che lo contraddistinguono. Le *marketing personas* finora descritte potranno essere un ottimo strumento anche per migliorare la previsione della domanda attesa

conoscendo il *target* in modo più approfondito. Per le medesime motivazioni, possono essere sviluppate nuove strategie di acquisizione clienti. E ancora, questi strumenti semplificano le operazioni aziendali riducendo il tempo necessario per prendere decisioni soprattutto sapendo che quanto utilizzato sia stato ottenuto con un approccio privo di una componente soggettiva ed esclusivamente basato sull'elaborazione di calcoli matematici e statistici. In conclusione, tutte queste attività, potrebbero aiutare i manager non solo ad aumentare i ricavi ma anche il margine che viene realizzato per ogni prodotto destinato ai diversi *target* dell'azienda.

11 Conclusioni

L'utilizzo delle tecniche di *clustering*, comprese quelle basate su algoritmi di *machine learning*, ha registrato un crescente interesse negli ultimi anni. Questa direzione è stata sostenuta sia dagli studiosi che si occupano della componente metodologica, sia da coloro che applicano tali tecniche nei vari ambiti aziendali. Il successo di queste tecniche è stato facilitato dall'enorme sviluppo di tecniche statistiche che sono legate ad algoritmi di analisi multivariata dei dati e dall'evoluzione costante dell'*hardware* che li supporta. L'analisi dei dati, spesso, richiede l'utilizzo di tecniche di *clustering* per identificare *pattern* nascosti che potrebbero non emergere da una semplice ispezione o da statistiche descrittive tradizionali. Proprio su queste premesse si basa questo contributo, il quale vuole essere un primo tentativo per la parte strategica e operativa nel mondo del *luxury fashion*. Lo sviluppo delle *marketing personas* per un'azienda dell'*Industry* di cui sopra mediante degli algoritmi di *machine learning* risulta essere ancora oggi una sfida e un argomento su ci sarà molto spazio per la ricerca, l'implementazione e l'ottimizzazione. L'approccio di *crisp clustering* ha permesso di ottenere risultati coerenti e facilmente interpretabili, contribuendo a identificare gruppi omogenei di consumatori con preferenze e comportamenti simili. Tale metodologia di *clustering* offre un valido supporto per l'azienda nel comprendere meglio i suoi clienti e adattare le proprie strategie di vendita e di comunicazione, garantendo così una migliore esperienza di acquisto nel settore del *luxury fashion*.

Le implicazioni manageriali derivanti dall'utilizzo di algoritmi di *clustering* di *machine learning* nella creazione di *marketing personas* per un'azienda del *luxury fashion* sono significative. Questi risultati forniscono un quadro più dettagliato e approfondito dei clienti, consentendo all'azienda di adattare le proprie strategie di *marketing* in modo più mirato ed efficace. L'ottimizzazione delle strategie, il miglioramento dell'esperienza del cliente con la realtà aziendale e l'efficientamento delle risorse economiche e umane dell'impresa sono attualmente e saranno sempre in misura maggiore un tema di importante interesse e sviluppo. Ottenere miglioramenti in queste aree è possibile sicuramente partendo da un approccio fortemente analitico. Questo studio vuole essere una risposta in questa direzione affinché ci sia un'evoluzione in questo ambito.

In aggiunta, comprendere i bisogni in modo profondo così come un minimo livello di errore potrebbe essere cruciale nel prossimo futuro ai fini della competizione. La democratizzazione dell'intelligenza artificiale e quindi, di queste elaborate tecnologie

computazionali, porterà numerose aziende ad avere dei processi automatizzati, precisi e basati sulle concrete necessità aziendali. In un prossimo futuro sarà verosimile vedere processi aziendali, soprattutto quelli più strategici, gestiti da complesse componenti *hardware*.

In conclusione, l'utilizzo di algoritmi di *machine learning* e tecniche di *clustering* nel contesto del *luxury fashion* offre nuove opportunità per comprendere meglio i clienti, adattare le strategie di *marketing* e migliorare l'esperienza di acquisto. L'evoluzione tecnologica continua a fornire strumenti sempre più potenti e accessibili, aprendo la strada a un futuro in cui la gestione dei processi aziendali sarà fortemente supportata da algoritmi basati su intelligenza artificiale e analisi avanzate. La sfida per le aziende del *luxury fashion* sarà quella di abbracciare queste tecnologie e sfruttarle in modo efficace per rimanere competitive e soddisfare le esigenze mutevoli dei clienti.

12 Bibliografia

An, J. C. (2016). Towards automatic persona generation using social media. *2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)* (p. 206-211). IEEE.

Anderberg, M. R. (1973). *The broad view of cluster analysis. Cluster analysis for applications* . Elsevier.

Andreea-Ionela, P. U. (2020). Clustering Consumers Through their Consumption Behavior: Analysis on the Fashion Industry. *Management and Economics Review, 5(1)*, 23-32.

Anisin, A. (2022, Novembre 8). *Introduction To Machine Learning For Marketing*. Tratto da Forbes: https://www.forbes.com/sites/theyec/2022/11/08/introduction-to-machine-learning-for-marketing/?sh=9aa5f5617b09

Ansarifar, F. &. (2018). A novel algorithm for adaptive data stream clustering. *Electrical Engineering (ICEE), Iranian Conference* (p. 1542-1546). IEEE.

Arian, A. M. (2021). Personas: a market segmentation approach for transportation behavior change. *Transportation research record, 2675(11)*, 172-185.

Asensio, E. A.-Á. (2022). Using Customer Knowledge Surveys to Explain Sales of Postgraduate Programs: A Machine Learning Approach.

Šulc, Z. (2016). Similarity measures for nominal data in hierarchical clustering (Doctoral dissertation). *Similarity measures for nominal data in hierarchical clustering*.

Bellini P., P. L. (2022). Multi clustering recommendation system for fashion retail. *Multimedia Tools and Applications*, 1-28.

Bellini, P. N. (2021). Fashion retail recommendation system by multiple clustering. *Proceedings of the 27th International DMS Conference on Visualization and Visual Languages (DMSVIVA)*, (p. 29-30). Pittsburgh, USA.

Birant, D. &. (2007). ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & knowledge engineering, 60(1)*, 208-221.

Brito, P. Q. (2015). Customer segmentation in a large database of an online customized fashion business. *Robotics and Computer-Integrated Manufacturing, 36*, 93-100.

Bu, J. L. (2020). Comparative Study of Hydrochemical Classification Based on Different Hierarchical Cluster Analysis Methods. *International journal of environmental research and public health, 17(24)*.

Cabigiosu, A. (2020). An Overview of the Luxury Fashion Industry. In *Digitalization in the Luxury Fashion Industry* (p. 9–31). Palgrave Macmillan, Cham. Tratto da https://doi.org/10.1007/978-3-030-48810-9_2

Cibulková, J. A. (2018). A case study of customer segmentation with the use of hierarchical cluster analysis of categorical data. *Proceedings of the Applications of Mathematics and Statistics in Economics, AMSE.* Kutná Hora, Czech Republic.

Cooper, A. (1999). *The inmates are running the asylum.* Indianapolis: SAMS - Pearson.

Dolan, R. (2014, 06 30). Framework for Marketing Strategy. Harvard Business Publishing.

Eskin, E. A. (2002). A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. *Applications of data mining in computer security*, 77-101.

Ester, M. K. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *dd (Vol. 96, No. 34)*, 226-231.

Forbes. (2013, Gennaio 16). *What Is The Difference Between Made To Measure And Bespoke?* Tratto da Forbes: https://www.forbes.com/sites/quora/2013/01/16/what-is-the-difference-between-made-to-measure-and-bespoke/?sh=8409f6e5352a

Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values. *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD)*, 21-34.

Jain, A. K. (1999). Data clustering: a review. *ACM computing surveys (CSUR), 31(3)*, 264-323.

Janardhanan, S. &. (2020). Market segmentation for profit maximization using machine learning algorithms. *ournal of Physics: Conference Series (Vol. 1706, No. 1, p. 012160)*.

Jansen, B. J.-g. (2020). Data-driven personas for enhanced user understanding: Combining empathy with rationality for better insights to analytics. *Data and Information Management Vol. 4. ScienceDirect*, 1-17.

Jung, S. G. (2017). Persona generation from aggregated social media data. *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*, 1748-1755.

Jung, S. G. (2018). Automatically conceptualizing social media analytics data via personas. *Proceedings of the International AAAI Conference on Web and Social Media (Vol. 12, No. 1)*.

Kabaivanova, S. (2015, Febbraio 10). *Make-to-order vs Make-to-stock in Fashion Industry*. Tratto da Be Global Fashion Netword: https://made-to-measure-suits.bgfashion.net/article/10455/58/Make-to-order-vs-Make-to-stock-in-Fashion-Industry

Koponen, M. (2017). Developing Marketing Personas with Machine Learning for Educational Program Finder (Master's thesis).

Kotler P., K. L. (2016). *Marketing Management.* Pearson.

Kutbay, U. (2018). Partitional Clustering. *Tech*.

La Rocca, A. (2006). uzzy clustering: la logica, i metodi. *ISTAT 5*.

Lance, G. N. (1966). Computer programs for hierarchical polythetic classification ("similarity analyses"). *The Computer Journal, 9(1)*, 60-64.

Le, S. Q. (2005). An association-based dissimilarity measure for categorical data. *Pattern Recognition Letters, 26(16)*, 2549-2557.

Lee, D. D. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 788-791.

Li, S. (2020, Febbraio 25). *Using user interviews to create "personas"*. Tratto da The Elements of Product Management: https://shawli.substack.com/p/using-user-interviews-to-create-personas

Lin, D. (1998). An information-theoretic definition of similarity. *Icml Vol. 98*, 296-304.

Lipovetsky, G. &. (2003). *Le luxe éternel.* Parigi: Gallimard.

Liu, G. (2022). New Index for Clustering Evaluation Based on Density Estimation. *arXiv preprint arXiv:2207.01294*.

Lourenco, F. L. (2004). inary-based similarity measures for categorical data and their application in Self-Organizing Maps.

MacQueen, J. (1967). Classification and analysis of multivariate observations. *5th Berkeley Symp. Math. Statist. Probability*, 281-297.

Mameli, M. P. (2021). Deep learning approaches for fashion knowledge extraction from social media: a review. *IEEE Access*, 1545 - 1576.

Manzato, R. (2019). The Impact of Culture on Luxury Consumption Behaviour: An Empirical Study on Chinese'and Italians' Behaviours.

MarketLine. (2021, Novemebre). *Global - Apparel Manufacturing*. Tratto da MarketLine: https://advantage.marketline.com/Analysis/ViewasPDF/global-apparel-manufacturing-145098

MarketLine. (2021, Novembre). *Global - Luxury Goods*. Tratto da MarketLine: https://advantage.marketline.com/Analysis/ViewasPDF/global-luxury-goods-145093

McKinsey. (2008, Settembre 1). *Enduring Ideas: The GE–McKinsey nine-box matrix.* Tratto da McKinsey: https://www.mckinsey.com/capabilities/strategy-and-corporate-finance/our-insights/enduring-ideas-the-ge-and-mckinsey-nine-box-matrix#

McKinsey. (2021, Novembre 15). *The State of Fashion 2022*. Tratto da McKinsey: https://www.mckinsey.com/~/media/mckinsey/industries/retail/our%20insights/state%20of%20fashion/2022/the-state-of-fashion-2022.pdf

Mongay, J. (2006). Strategic Marketing. A literature review on definitions, concepts and boundaries. Autonomous University of Barcelona (UAB) & SBS Swiss Business School.

Mulyo, I. A. (2022). Customer Clustering Using The K-Means Clustering Algorithm in Shopping Mall in Indonesia. *Management Analysis Journal, 11(4)*.

Murtagh, F. &. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of classification, 31*, 274-295.

Namvar, M. G. (2010). A two phase clustering method for intelligent customer segmentation. *010 International Conference on Intelligent Systems, Modelling and Simulation* (p. 215-219). IEEE.

Pereira, C. M. (2013). Common dissimilarity measures are inappropriate for time series clustering. *Revista de Informática Teórica e Aplicada, 20(1)*, 25-48.

Pitafi, S. A. (2023). A Taxonomy of Machine Learning Clustering Algorithms, Challenges, and Future Realms. *Applied Sciences, 13(6), 3529*.

Porter, M. E. (1980). *Competitive Strategy: Techniques for Analyzing Industries and Competitors.* New York: Free Press.

Quan, C. L. (2015). Research On The Construction Of Personas Model Based On K-Means Clustering Algorithm.

Revella, A. (2011). *The Buyer Persona Manifesto. E-book op www. buyerpersona. com. isbn, 978(90), 8965.*

Rezankova, H. L. (2011). Evaluation of categorical data clustering. In S. P. Mugellini E., *Advances in Intelligent Web Mastering–3: Proceedings of the 7th Atlantic Web Intelligence Conference, AWIC 2011* (p. 173-182). Fribourg: Springer Berlin Heidelberg.

Rodrigues, F. &. (2016). Product recommendation based on shared customer's behaviour. *Procedia Computer Science, 100,* 136-146.

Scikit-learn. (s.d.). *2.3 Clustering*. Tratto da Scikit-learn.org: https://scikit-learn.org/stable/modules/clustering.html

Siegel, E. (2023, Marzo 24). *How Machine Learning Can Improve the Customer Experience*. Tratto da Harvard Business Review: https://hbr.org/2023/03/how-machine-learning-can-improve-the-customer-experience

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation, 28(1),* 11-21.

Statista. (2023, Maggio 9). *Luxury Fashion - Worldwide.* Tratto da Statista: https://www.statista.com/outlook/cmo/luxury-goods/luxury-fashion/worldwide?currency=EUR

Steinbach, M. K. (2000). A comparison of document clustering techniques. *Computer Science & Engineering (CS&E) Technical Reports [749]*.

Strategic Business Insights. (2009). *VALS*. Tratto da strategicbusinessinsights: https://www.strategicbusinessinsights.com/vals/

Thomas, D. (2019). *Fashionopolis: The Price of Fast Fashion and the Future of Clothes.* Penguin Press.

Thorndike, R. (1953). Who belongs in the family? *Psychometrika, 18(4),* 267-276.

Tu, N. D. (2010). Using cluster analysis in persona development. *2010 8th International Conference on Supply Chain Management and Information* (p. 1-5). IEEE.

Tyron, R. C. (1939). Cluster analysis: correlation profile and orthometric (factor) analysis for the isolation of unities in mind and personality. *Edwards brother, Incorporated*.

Varian, H. R. (1992). *Microeconomic analysis (Vol. 3).* New York: Norton.

Wang, C. (2022). Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach. *Information Processing & Management, 59(6)*.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association 58.301*, 236-244.

Xin, X. W. (2022). Building up Personas by Clustering Behavior Motivation from Extreme Users. *Design, User Experience, and Usability: UX Research, Design, and Assessment: 11th International Conference, DUXU 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings, Part I* (p. 120-131). Springer International Publishing.

Yadegaridehkordi, E. N. (2021). Customers segmentation in eco-friendly hotels using multi-criteria and machine learning techniques. *Technology in Society, 65*.

Yan, Z. &. (2021). Customer segmentation using real transactional data in e-commerce platform: A case of online fashion bags shop. *Proceedings of the International Conference on Electronic Business*, (p. 90-99).

Yoseph, F. &. (2018). Segmenting retail customers with an enhanced RFM and a hybrid regression/clustering method. *2018 International Conference on Machine Learning and Data Engineering (iCMLDE)* (p. 108-116). IEEE.

Zadeh, L. (1965). Fuzzy sets. *Inform Control, 8*, 338-353.

Appendice

Python script

Il seguente *QR code* (o *link* sottostante) rimanda allo *script python*.

bit.ly/piero-battistel-masters-thesis-colab-python-script

Introduzione

L'utilizzo dell'intelligenza artificiale nel *marketing* è solo uno dei possibili campi di applicazione di questo insieme di tecnologie per i diversi domini. Data l'esperienza pregressa, è ormai ben fondata l'idea che l'utilizzo di metodologie basate su *machine learning* e *deep learning* abbiano migliorato diverse realtà aziendali. Nel dominio del *marketing*, considerandolo nell'industria del *fashion*, l'apprendimento automatico è stato implementato per diverse finalità finora. Il caso più ricorrente è la segmentazione dei clienti; avere la possibilità di segmentare i clienti in modo accurato e in grado di poter considerare molteplici *pattern* tra le loro informazioni ne migliora di conseguenza i risultati delle campagne di *marketing* e, più in generale, i risultati di *business* semplificando anche le scelte strategiche. Proprio per questo lo studio che qui si presenta vuole essere un contributo per dimostrare ulteriormente come porre in essere complessi algoritmi di intelligenza artificiale in chiave di *business*. Nello specifico, l'elaborato riguarderà la possibilità di sviluppare delle *marketing personas* mediante l'apprendimento automatico. Queste sono delle elaborate descrizioni di clienti di riferimento che aiutano i dipartimenti di *marketing* nel prendere decisioni strategiche ed operative considerando il *target* di riferimento. Il tentativo di innovazione qui proposto è dovuto all'intenzione di voler sviluppare degli strumenti di *marketing* con degli algoritmi di intelligenza artificiale per la creazione di strumenti di *strategic marketing* quando, questi, sono sempre stati posti in essere in base alle esperienze e percezioni dei *marketing* manager o *team* di *marketing*. L'analisi qui riportata riguarderà l'applicazione di modelli di *clustering* su di un *dataset* di un importante azienda nel panorama della moda di lusso. Nello specifico, i dati riguarderanno clienti internazionali e molteplici caratteristiche di diversa natura relativi a questi ultimi. La finalità ultima non è meramente quella di ottenere profili di acquirenti ma in particolar modo anche di stressare e validare la possibilità di applicazione del *machine learning* in un processo volto a migliorare l'esperienza del cliente finale come obiettivo (Siegel, 2023).

Luxury Fashion Industry

L'industria del *luxury fashion* è uno dei settori in più rapida espansione e con le migliori *performance*, con aziende leader che hanno registrato una crescita rilevante negli ultimi

anni. Il mercato mondiale dei prodotti di lusso è dominato dalla moda, che si prevede crescerà a un ritmo sostenuto nel prossimo lustro. Il *luxury fashion* è solo uno delle classificazioni di prodotti nel più ampio mercato dei beni di lusso composti anche da orologi e gioielli, prodotti in pelle, cosmetici e fragranze e occhialeria (Figura 1).

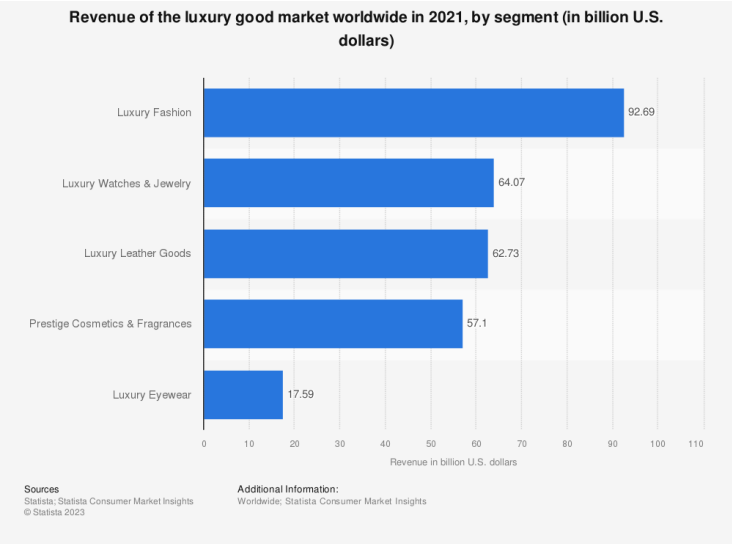

Figura 1 Ricavi del mercato globale dei beni di lusso nel 2021, per segmento (in milioni di dollari U.S.).

In economia, un bene può essere definito di lusso nel momento in cui la domanda aumenta più che proporzionalmente all'aumentare del reddito (Varian, 1992).

LVMH, *Kering*, *Ralph Lauren*, *PVH*, *Canali*, *Armani* e tante altre, sono le aziende più note nella *luxury fashion industry*. Negli ultimi anni il settore ha registrato un consolidamento significativo, con grandi gruppi come *LVMH* e *Kering* che hanno ampliato il loro portafoglio attraverso acquisizioni di marchi di lusso più piccoli. Si prevede che questa tendenza continui, poiché il settore cerca di massimizzare le economie di scala e di competere in un mercato sempre più affollato. Nonostante l'impatto della pandemia sul settore produttivo, la moda di lusso ha continuato a registrare buoni risultati, trainata dalla forte domanda in Asia e Nord America. Nel 2020, secondo Statista, il mercato globale dei beni di lusso è stato valutato 256 miliardi di euro, con l'Asia che ha rappresentato la quota di mercato maggiore, pari al 38%.

Il mercato del *luxury fashion* genererà globalmente circa 108,70 miliardi di euro nel 2023 (Figura 2). Valore in crescita dal momento che, globalmente, si prevede che arrivi a 125 miliardi di euro entro il 2028; il *CAGR* (2023-2028) si attesta intorno al 3%.

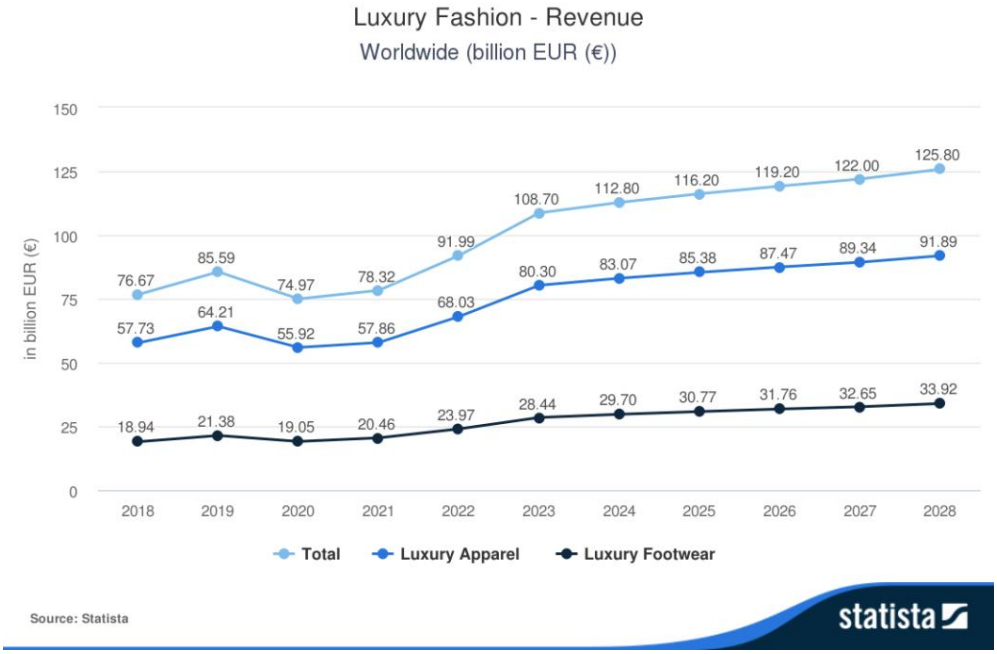

Figura 2 Ricavi dell'industria globale della moda di lusso 2018-28 (in milioni di euro).

Nello specifico, quest'anno così come per i precedenti, *USA*, Cina, Giappone e *UK* sono i paesi con maggiore fatturato generato dal mercato (Figura 3).

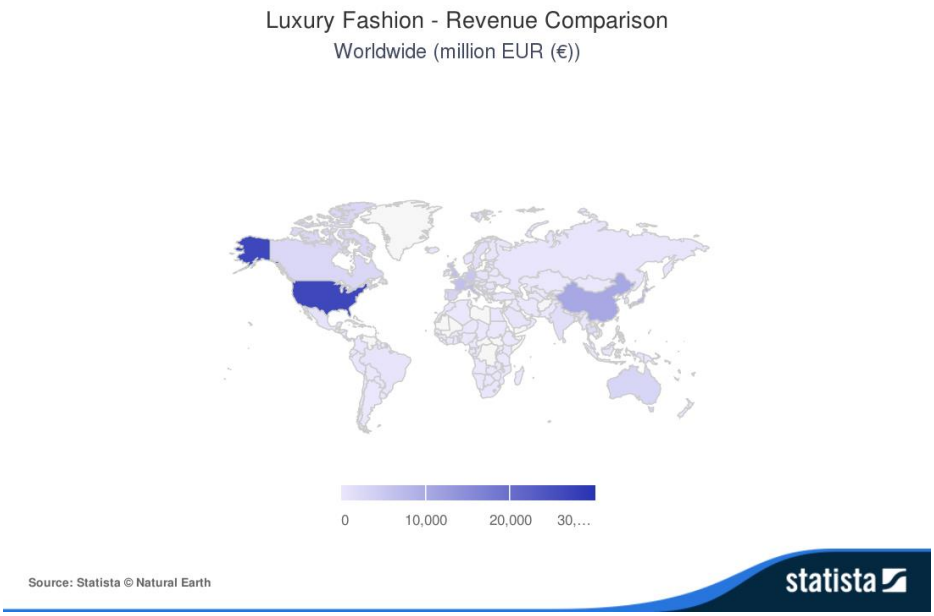

Figura 3 Confronto dei ricavi 2023 per le nazioni nel luxury fashion (in milioni di euro).

In aggiunta è previsto che i canali di vendita digitali siano al 22% del totale ma il dato è in aumento di circa due punti percentuali per anno. Durante il 2025 si prevede che ancora il 74% delle vendite avverrà offline (Figura 4). Un altro aspetto inerente ai canali di vendita è la distinzione tra gli acquisti mediante cellulare o via computer. Per il 2023, Statista ha previsto una distinzione ancora molto vicina al 50% per ciascuno di questi. Più in particolare, la percentuale di *revenue* generata via cellulare è leggermente superiore con un valore pari al 57%. Anch'esso in aumento di un punto percentuale ogni anno (Statista, 2023).

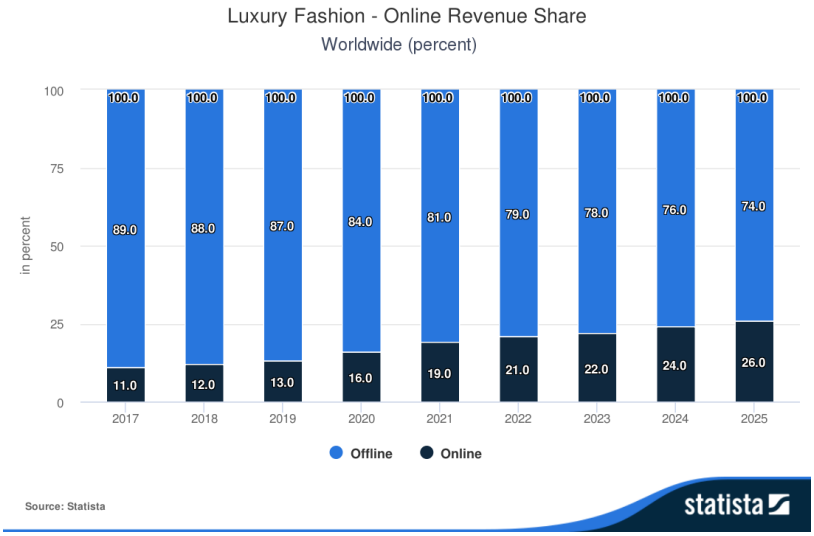

Figura 4 Quote di ricavi online e offline nel mercato globale della moda di lusso.

Marketing Personas

Le *personas* sono uno strumento nato per diverse finalità; per quanto concerne lo *strategic marketing*, possono essere definite come uno strumento volto a migliorare l'approfondimento del *target* di riferimento mediante un'accurata descrizione di un archetipo. Da ciò ne deriva un miglioramento della conoscenza delle peculiarità del consumatore finale, ergo la possibilità di prendere decisioni molto più precise sulla base dei bisogni, attitudini, esigenze, problematiche e comportamenti del consumatore finale.
Una credenza comune è quella per cui i dati demografici siano sufficienti per definire in modo omogeneo e rilevante i segmenti. Tuttavia, la correlazione tra le variabili descrittive della dimensione demografica di un individuo e i bisogni dello stesso è molto bassa.
Due individui con le medesime caratteristiche demografiche, possono avere bisogni completamente diversi o addirittura opposti (Kotler P., 2016).

L'idea di creare delle *personas* è nata nell'industria del *software* con Alan Cooper, il quale ha riconosciuto che le aziende non riuscivano a produrre *software* di alta qualità perché non prendevano in considerazione le *personas* degli utenti durante il processo di progettazione. Il concetto di creazione di *persona* di Cooper non si limitava allo sviluppo di *software*, ma poteva essere applicato anche alle vendite e al *marketing*. Questo ha portato all'era in cui i *marketer* hanno iniziato a creare delle *personas* per comprendere e spiegare le esigenze degli acquirenti (Cooper, 1999). Pur essendoci delle analogie tra *personas* del *software* e del *marketing* in quanto queste condividono le stesse componenti di base come familiarità, facilità di riconoscimento e tentativo di creare un legame emotivo con il gruppo di utenti che rappresentano, differiscono nel loro approccio. Le *personas* nell'industria del *software* mirano a raccontare la vita degli utenti e il modo in cui il prodotto potrebbe essere utilizzato, mentre le *personas* in ottica di *marketing* considerano il motivo per cui gli utenti hanno bisogno del prodotto, cosa scatena la realizzazione dei loro bisogni e quali fattori portano alla decisione finale di acquisto.

Riportando la definizione di Adele Ravella fondatrice del *Buyer Persona Institute*, una *buyer persona* è:

"It's an archetype, a composite picture of the real people who buy, or might buy, products like the ones you sell.

It's an avatar you craft from what you learn in direct interviews with as many buyers as possible. And from behavior observed anywhere else: at industry conferences; in online forums; through social media.

If crafted with skill and insight, the person who emerges in this picture may become as real to you as anyone you can ever remember meeting. In your mind's eye, this person becomes three-dimensional to the point that you can see the world through his or her eyes." (Revella, 2011).

La Figura 5 riporta un esempio di *marketing personas*.

Figura 5 Esempi di Personas

Literature Review, Research Gap e Metodologia

Sulla base dei contributi studiati e riportati nel dettaglio nel lavoro di tesi, si evincono diverse mancanze sotto molteplici profili. In primo luogo, analizzando l'attuale letteratura da un punto di visto macro, l'industria del *fashion* presenta un ridotto numero di elaborati sull'utilizzo del *machine learning* e della *clustering analysis* sia in generale che per lo sviluppo di *marketing persona.* Allo stesso modo, la quantità di studi volti a generare delle *buyer personas* utilizzando algoritmi di apprendimento automatico o apprendimento profondo è davvero limitata. I pochi casi riscontrati in tal direzione sono stati applicati su delle feature di natura diversa rispetto a quella transazionale come, ad esempio, in "*Towards automatic persona generation using social media"* di An, J. et al. Quest'ultimo studio presenta la realizzazione di *marketing personas* basandosi su dati ottenuti da *social media* come gli interessi, *vanity metric* e altre metriche di riferimento (An, 2016). In aggiunta, il numero di *feature* utilizzate è sempre inferiore a quello necessario per la realizzazione di una *marketing personas* così come definita da Cooper e, di conseguenza, maggiormente utile in ottica di *business*. Una *marketing personas* ha come aspetto distintivo esattamente quello di essere un'accurata descrizione di un archetipo di uno o più segmenti. La peculiarità appena definita è l'elemento che effettivamente determina un distacco netto tra *customer segmentation* e *buyer personas*.

Sviluppare le *personas* mediante *machine learning* è un obiettivo ambizioso e che necessiterà di innumerevoli tentativi e studi prima che venga raggiunto un risultato realmente applicabile in questa direzione nei contesti aziendali. Lo studio presentato di seguito è un primo tentativo per l'industria del *fashion*, ambiente che così come gli altri, si caratterizza per aver visto delle *personas* basate esclusivamente su delle idee e delle esperienze proprie dei *marketing manager* nel corso degli ultimi decenni. Poter affiancare ad un lavoro da anni prettamente manageriale, degli algoritmi di apprendimento automatico per generare dei risultati precisi circa gli archetipi *target*, è l'obiettivo dello studio. Di conseguenza si può derivare la domanda di ricerca:

"È possibile utilizzare il machine learning per la creazione di marketing personas nell'industria del fashion?".

Lo studio è un'analisi realizzata con *Python*, tra i linguaggi di programmazione maggiormente utilizzati in ambito di *data science*. Nello specifico, grazie a diverse librerie di codice, funzionalità del linguaggio e algoritmi di intelligenza artificiale è stato posto in essere uno studio che attraversa diverse analisi statistiche, tecniche di esplorazione e aggregazione dei dati. Lo *script* completo è stato inserito in Appendice.

Exploratory Data Analysis

I *dataset* su cui è stato realizzato lo studio sono di una nota azienda italiana nel settore del *luxury fashion* con quasi cento anni di storia. Oltre alla vendita di abbigliamento esclusivamente per uomo, si occupa anche della produzione dei capi. Circa l'80% dei capi di abbigliamento sono esportati in diversi paesi in tutto il mondo. I prodotti realizzati dall'azienda sono sia capi con misure standard sia capi su misura "*Made to Measure*".

L'insieme dei dati utilizzato per questo contributo si divide in due tabelle. La prima, chiamata *Transactions Details*, riguarda dati transazionali, di conseguenza, ogni riga rappresenta una transazione. Il secondo *set* di dati, chiamato *Contact Active*, comprende informazioni su dei clienti; pertanto, ogni riga, rappresenta un acquirente considerato come singolo. Le due tabelle hanno come *primary key* il *customer ID*.

Transactions Details ha una *shape* pari a (1021739, 65) ed è bidimensionale; le variabili contenute sono 38 qualitative e 27 quantitative (16 *float64* e 11 *int64*). Il numero di *missing value* cambia in base alla variabile considerata; ad esempio, il 95,73% dei valori

era mancante nel caso della variabile *loyalty_membership_no*, mentre, per altre variabili, come la *primary key* "*customer*" non era presente nessun dato mancante.

Contact Active ha una *shape* formata da 384352 righe e da 135 colonne; anche questo *set* di dati è bidimensionale. Le variabili presenti sono in parte qualitative e in parte quantitative, rispettivamente 72 e 63 di cui 50 *float64* e 13 *int64*. Così come per *Transactions Details*, anche per *Contact Active* mancano dei dati in diverse variabili; in particolare, alcune variabili non contengono alcun dato ed è il caso di *cross-selling* mentre, anche in questo caso, non mancano dati per la colonna in comune ossia *customer_id*. Dato l'ingente numero di variabili considerando entrambi i *set* di dati, saranno presentate solo quelle contenute all'interno del *dataframe* oggetto degli algoritmi di *clustering*. Tutte le tecniche statistiche utilizzate per arrivare ad avere il *dataset* descritto di seguito sono state riportate ed elaborate nel capitolo 6 che concerne la parte di *pre-processing*. Il *dataset* utilizzato per l'analisi di *clustering* mediante *machine learning* è costituito da 146778 righe e 26 colonne. Ogni riga rappresenta un cliente e le variabili sono informazioni sullo stesso o sulle transazioni effettuate dal medesimo. Le *feature* si dividono in 10 qualitative e 16 quantitative di cui la maggior parte sono *int64* e la minoranza *float64*, rispettivamente, 9 e 7. Per facilitare la trattazione delle prossime parti dell'elaborato, quest'ultimo *dataset* presentato sarà richiamato come "*ca_updated*". Qui di seguito è stato riportato uno dei grafici realizzati (Figura 6), il quale permette di visualizzare in modo interattivo il numero di clienti per le migliori città in cui è presente il *brand.*

Figura 6 Grafico clienti per città

Clustering Analysis

La *cluster analysis*, un'analisi multivariata, rappresenta una tecnica mediante la quale è possibile suddividere un insieme di *pattern*, cioè unità statistiche, in sottoinsiemi caratterizzati da proprietà o caratteristiche simili. Il risultato che si intende raggiungere è quello di ridurre al minimo la devianza all'interno di ciascun *cluster* (devianza *within-groups*), massimizzando al tempo stesso la devianza tra i gruppi (devianza *between-groups*), così da minimizzare la "lontananza logica". In altre parole, questa analisi esplorativa consiste nel ricercare tra le *n* osservazioni *p*-dimensionali, gruppi di unità tra loro simili, senza certezza e conoscenza iniziale che tali gruppi omogenei esistano effettivamente tra i dati oggetto di analisi. L'obiettivo dell'esplorazione così descritta è dunque quello di identificare dei gruppi che appaiono con naturalezza nelle osservazioni. La misurazione della lontananza logica si basa sull'applicazione di misure di similarità tra le unità statistiche: per dati quantitativi si utilizzano misure di distanza, definite metriche, mentre per dati qualitativi si adottano misure di tipo *matching*, che possono indicare similarità o dissimilarità. Dopo aver selezionato la misura di similarità adeguata, è necessario decidere l'algoritmo di *clustering* e, eventualmente, il metodo di aggregazione o suddivisione.

La conseguenza naturale per quanto concerne l'analisi dei *cluster* finora trattata è che i *pattern* delle unità statistiche all'interno di ogni *cluster* saranno il più simile possibile mentre con maggiore diversità se considerati tra *cluster*.

Il problema del *clustering* è stato trattato in diversi contesti e discipline, il che riflette l'ampia diffusione e l'utilità significativa di questa tecnica nell'analisi esplorativa dei dati. Il *clustering*, come metodo di *unsupervised machine learning*, rappresenta un problema complesso in termini computazionali. Le differenze nelle ipotesi e nei contesti delle diverse comunità hanno rallentato il trasferimento di concetti e metodologie generiche. La grande varietà di tecniche per misurare la prossimità (similarità) tra gli elementi, raggrupparli e rappresentarli ha generato un assortimento ricco ma spesso confuso di metodi di *clustering*. A differenza dei metodi supervisionati, nel *clustering*, si utilizzano dei dati non etichettati ossia delle informazioni per le quali non sono stati identificati dei *pattern* apriori. In questa tecnica di classificazione non supervisionata, gli algoritmi di apprendimento automatico opereranno al fine di identificare e assegnare un *pattern* per ogni unità statistica. Ne consegue che la difficoltà è nella ricerca dei *pattern*, non

essendoci un addestramento dell'algoritmo al fine di determinare un apprendimento degli stessi nonché un confronto successivo (Jain, 1999).

Sebbene ci siano studi e contributi precedenti al 1939, quest'anno rappresenta tutt'oggi la data in cui uno dei primi pionieri nel campo dell'analisi dei *cluster* ha pubblicato un contributo divenuto poi uno dei più significativi nel suo settore.

Nel 1939 *Robert C. Tyron* pubblicò un'opera intitolata *"Cluster Analysis: Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality"* la quale si concentrava principalmente sull'approccio statistico alla comprensione e all'interpretazione del comportamento umano. L'obiettivo di ricerca era quello di esplorare come i tratti della personalità potessero essere raggruppati in *cluster* determinando così una comprensione completa della psicologia umana. Nel suo contributo impiegò un metodo noto come *orthometric (factor) analysis* determinando le basi per l'attuale *clustering analysis* (Tyron, 1939).

Le tipologie dei metodi di *clustering* sono divise in gerarchiche e partizionali (non gerarchici). La Figura 7 presenta uno schema riassuntivo delle diverse classificazioni. Il *clustering* gerarchico si divide a sua volta in aggregativo (o agglomerativo) e scissorio (Pitafi, 2023).

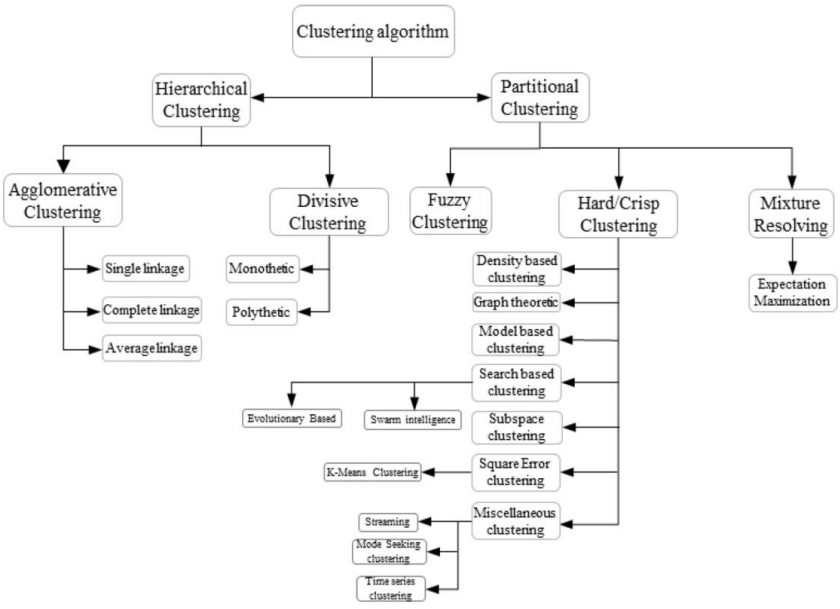

Figura 7 Tassonomia di clustering

Model Selection e Valutazione

Gli algoritmi testati per questo lavoro di tesi sono stati 13: nello specifico sono il *K-Means*, il *K-Prototypes*, il *Bisecting K-Means*, il *K-Medoids*, l'*Affinity Propagation*, il *Mean-shift*, lo *Spectral clustering*, il modello gerarchico di *Ward*, *l'Agglomerative clustering*, il *DBSCAN*, l'*OPTICS*, il *Gaussian Mixtures* e il *BIRCH* (Scikit-learn). Gli algoritmi con i risultati migliori sono stati il *K-Means* e il *K-Prototypes*.

L'algoritmo *K-Means* è stato realizzato partendo dalla scelta ottimale del numero di *cluster* basata sull'*elbow method* con la *Within-cluster Sum of Squares* (*WCSS*) ossia la somma delle distanze al quadrato tra ogni punto del *cluster* e il suo centroide. Il numero di *cluster* ideale è stato scelto in corrispondenza di due per quanto concerne i risultati di questo modello. Il *silhouette score* medio è pari a 0,26 mentre l'indice *Calinski-Harabasz* è di 59318,57. La Figura 8 rappresenta i due cluster mediante uno *scatterplot*.

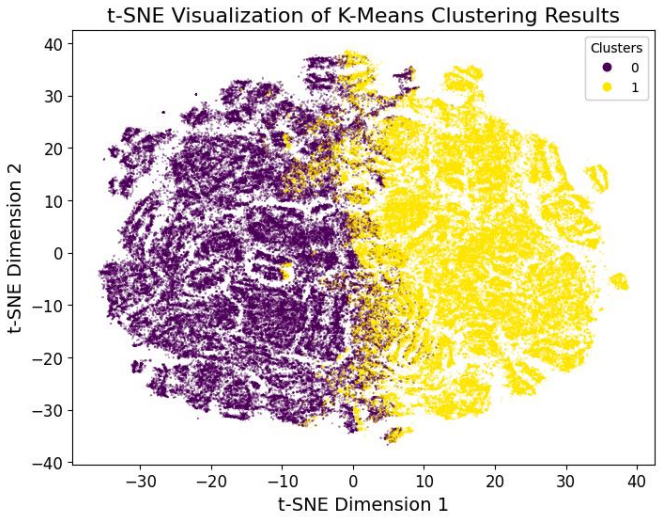

Figura 8 Scatterplot t-SNE del K-Means

L'algoritmo *K-Prototypes* è stato realizzato con il medesimo processo. A differenza però dell'algoritmo *K-Means* in cui è stata utilizzata la *WCSS*, qui è possibile utilizzare la funzione di costo propria del modello, la quale tiene conto sia delle variabili categoriche (considerando le distanze) che delle variabili numeriche (in base alle similitudini). La funzione di costo è una funzione di dissimilarità tra i punti del *cluster* e il suo centroide. Pertanto, l'obiettivo è quella di minimizzarla e anche in questo caso, viene scelto un numero di *cluster* per il quale, un aggiunta di *k cluster* non determini un miglioramento significativo in questa direzione. Come si evince dal grafico (Figura 9), anche qui il numero ottimale di *cluster* è pari a due.

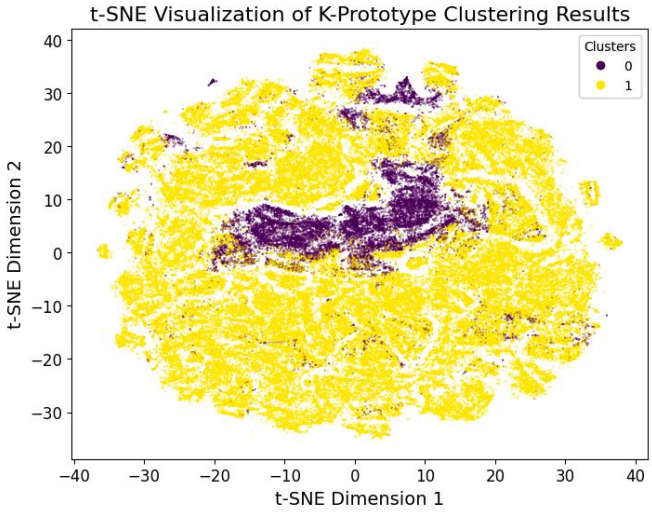

Figura 9 Scatterplot t-SNE del K-Prototypes

Interpretazione e Limiti

Analizzando i dati per ogni *cluster* generato dal *K-Prototypes* mediante valori medi, il calcolo delle mode e rapporti di dati ne risultano due *marketing personas* con caratteristiche sufficientemente diverse. Le *marketing personas* si basano su di un consumatore residente in Cina, la prima (Figura 10), e su di un consumatore europeo per quanto concerne la seconda (Figura 11).

Figura 10 I^a Marketing Personas generata con Machine Learning

Figura 11 II[a] Marketing Personas generata con Machine Learning

I limiti principali del contributo sono l'utilizzo di un *dataset* non nato per questo *task*, diversi problemi computazionali circa l'applicazione di alcuni modelli di *machine learning*, un numero elevato di *missing value* e un'assenza di miglioramento delle performance degli algoritmi in seguito al *fine tuning* degli iperparametri.

Implicazioni Manageriali

Lo studio è riuscito a determinare una risposta alla domanda di ricerca e pertanto, risulta essere possibile utilizzare il *machine learning* per realizzare delle *marketing personas*. Essendo uno strumento strategico e di conseguenza, con un impatto rilevante nelle decisioni dell'intera realtà aziendale, è fondamentale però che tale processo venga realizzato nel modo più rigoroso possibile partendo dallo studio delle variabili per concludersi con un test approfondito sui diversi algoritmi e le loro peculiarità.

Le implicazioni manageriali risultano essere molteplici in quanto, come detto, lo strumento delle *marketing personas* aiuta l'azienda non solo nel prendere decisioni di importante rilievo ma anche in diversi fasi della vita di un'azienda. Sono state individuate cinque principali implicazioni di *business*. La prima riguarda un miglioramento della segmentazione di mercato poiché approfondiscono le caratteristiche dei segmenti per cui sono realizzate creando dei profili di consumatori e di conseguenza, facilitando anche la comprensione e l'esecuzione di diverse operazioni all'interno dell'azienda. Inoltre, nel

momento in cui ciò viene realizzato mediante algoritmi, si riducono tutti gli errori che derivano dalla generalizzazione dell'esperienza dei *decision maker* all'interno della realtà aziendale. Un'altra implicazione manageriale, naturale conseguenza della precedente, è la possibilità di personalizzare le strategie di *marketing* in modo più puntuale e accurato frutto di applicazione matematica e statistica nella realizzazione degli strumenti di *business* quali le *marketing personas*.

Ottimizzare le strategie di prodotto, creando delle varianti più accurate e cucite sui bisogni e caratteristiche del cliente è un ulteriore aspetto che può essere raggiunto con questo approccio basato sull'intelligenza artificiale applicata nelle dinamiche aziendali. Naturale conseguenza della personalizzazione della strategia di *marketing* e dell'ottimizzazione del prodotto, è il miglioramento dell'esperienza del cliente. In generale, conoscere i bisogni dei clienti, il *decision journey* e i punti di contatto permette di soddisfare il cliente in modo migliore, strutturando ogni singolo aspetto operativo in risposta alle sue esigenze e necessità. In ultima analisi, rendere più precisi i processi determina anche un'ottimizzazione delle risorse in quanto, così facendo, l'azienda può concentrare i suoi sforzi, il *budget* e tutte le altre risorse, in primo luogo, sui clienti più promettenti nonché in modo più profittevole per ognuno conoscendo, con un margine di errore inferiore, le peculiarità che lo contraddistinguono. Le *marketing personas* finora descritte potranno essere un valido strumento anche per migliorare la previsione della domanda attesa conoscendo il *target* in modo più approfondito. Per le medesime motivazioni, possono essere sviluppate nuove strategie di acquisizione clienti. E ancora, questi strumenti semplificano le operazioni aziendali riducendo il tempo necessario per prendere decisioni soprattutto sapendo che quanto utilizzato sia stato ottenuto con un approccio privo di una componente soggettiva ed esclusivamente basato sull'elaborazione di calcoli matematici e statistici. In conclusione, tutte queste attività, potrebbero aiutare i manager non solo ad aumentare i ricavi ma anche il margine che viene realizzato per ogni prodotto destinato ai diversi *target* dell'azienda.

Conclusioni

L'utilizzo delle tecniche di *clustering*, comprese quelle basate su algoritmi di *machine learning*, ha registrato un crescente interesse negli ultimi anni. Questo contributo vuole essere un primo tentativo di applicazione del *machine learning* per la parte strategica e

operativa dello sviluppo di una *marketing personas* nel mondo del *luxury fashion*. La realizzazione di questi strumenti strategici posta in essere come riportato in questo studio risulta essere ancora oggi una sfida e un argomento su ci sarà molto spazio per la ricerca, l'implementazione e l'ottimizzazione in futuro. L'approccio di *crisp clustering* ha permesso di ottenere risultati coerenti e facilmente interpretabili, contribuendo a identificare gruppi omogenei di consumatori con preferenze e comportamenti simili. Tale metodologia di *clustering* offre un valido supporto per l'azienda nel comprendere meglio i suoi clienti e adattare le proprie strategie di vendita e di comunicazione, garantendo così una migliore esperienza di acquisto nel settore del *luxury fashion*. La democratizzazione dell'intelligenza artificiale e quindi, di queste elaborate tecnologie computazionali, porterà numerose aziende ad avere dei processi automatizzati, precisi e basati sulle concrete necessità aziendali. In un prossimo futuro sarà verosimile vedere processi aziendali, soprattutto quelli più strategici, gestiti da complesse componenti *hardware* e *software*. In conclusione, la sfida per le aziende del *luxury fashion* sarà quella di abbracciare queste tecnologie e sfruttarle in modo efficace per rimanere competitive e soddisfare le esigenze mutevoli dei clienti.