LUISS

Corso di laurea in Marketing analytics & metrics

Cattedra Customer intelligence e logiche di analisi dei big data

Predictive analytics using Artificial Intelligence algorithms for Customer Loyalty

Prof. Marina Paolanti

RELATORE

Prof. Luca Romeo

CORRELATORE

Chiara Iaccarino
Matr. 749121

CANDIDATO

Anno Accademico 2022/2023

Indice

Capitolo 1 – Introduzione

1.1 Contesto di riferimento

“La moda passa, lo stile resta”. Quest'affermazione di Coco Chanel, regina dell'industria della moda di tutti i tempi, una figura iconica e tuttora fonte di ispirazione, coglie in modo impeccabile una delle caratteristiche intrinseche di questo settore: la sua natura effimera, passeggera, in costante evoluzione e mutamento, ancora più evidente ai giorni nostri.

L'avvento della quarta rivoluzione industriale, portata dalla trasformazione digitale, sta consentendo all'industria della moda, così come ad altri settori, di incrementare la propria capacità di generare e utilizzare i dati che precedentemente non erano tecnicamente o finanziariamente realizzabili. Tuttavia, l'impatto più significativo sulla produzione e sulla distribuzione deve ancora manifestarsi appieno. In un'intervista datata 6 marzo 2019, Federico Marchetti, CIO e fondatore di YOOX, espone come l'Intelligenza Artificiale (IA) stia contribuendo a rivoluzionare non solo il processo produttivo, ad esempio, il marchio "8" è interamente creato grazie all'impiego dell'IA, ma anche le strategie di vendita e di marketing, in quanto l'impiego dei *big data* e dell'IA permetterà a ciascun cliente di avere una homepage personalizzata (Marchetti, F., 2019).

Prima di analizzare la *Fashion Industry*, è fondamentale comprendere la complessa definizione del termine, che presenta profonde implicazioni sia a livello sociologico che economico.
Il Presidente della *Costume Society* della Gran Bretagna, Collen McDowell, ha definito la moda come “quella forma di arte che, sebbene minore, reagisce più rapidamente di qualsiasi altra a qualsiasi sfumatura sociale, politica e culturale del nostro tempo” (McDowell, C., 2000). In tempi molto rapidi, la moda ha influenzato quasi tutti gli aspetti del nostro essere, partendo dalla musica che ascoltiamo, dalle automobili che preferiamo guidare, dall’architettura fino ad arrivare, ai vestiti e agli accessori. Poiché viviamo in contesti sociali piuttosto che isolati, l'abbigliamento oggi non solo soddisfa un bisogno primario, ma è anche una vera e propria forma di comunicazione che aiuta ad esprimere chi siamo o chi vorremmo essere. Il modo di vestire aiuta a conoscere l’etnia di una persona, lo stile di vita, l’occupazione e, naturalmente, la posizione sociale. La moda è quindi un'area strategica dal punto di vista sociale ed economico in quanto influenza le decisioni di acquisto delle persone e modifica le loro preferenze.
Inoltre, la moda può essere definita come "l'insieme delle attività creative, economiche e gestionali finalizzate al design, alla produzione, alla promozione e alla vendita dei prodotti" (Andreeva, A., 2006).

La presenza della tecnologia è stata sempre intrinseca nella produzione di abbigliamento e, in generale, in tutte le attività umane fin dagli albori della storia. Come qualsiasi forma di artigianato, la creazione di indumenti richiede l'utilizzo di strumenti appositi e una conoscenza specifica su come utilizzarli in maniera adeguata. Fin dai tempi preistorici, l'umanità ha avvertito l'esigenza di preservarsi dagli effetti del clima, sia dal freddo che

dal caldo. Tuttavia, la produzione di abiti richiedeva una tecnologia relativamente avanzata, nonché l'impiego di strumenti specializzati e sofisticati. Durante il periodo del Neolitico, noto anche come rivoluzione agricola (7000-3000 a.C.), l'adozione di uno stile di vita sedentario caratterizzato da insediamenti in villaggi, pratiche agricole e la disponibilità di piante utilizzate per il cibo e la produzione di fibre, favorì lo sviluppo dell'industria tessile, compresa la tessitura di tessuti come il lino e la lana. Successivamente, tra la fine del XVII e l'inizio del XVIII secolo, emersero diverse innovazioni e brevetti che apportarono notevoli miglioramenti nella produzione e nell'uso degli indumenti. Fu in questo contesto che nacque l'industria tessile moderna nel Regno Unito, grazie ai significativi progressi tecnologici raggiunti.

L'innovazione si diffuse presto negli Stati Uniti e nel 1847, il numero di americani che lavoravano nel settore tessile era superiore a quello di qualsiasi altro settore. Grazie all'introduzione delle macchine nel processo produttivo, furono gettate le basi per trasformare il settore da piccolo sistema manifatturiero a grande industria. Tuttavia, la macchina da cucire ha rappresentato il più grande progresso nella tecnologia della moda brevettata da Isaac Singer nel 1851; l'innovazione permise di produrre capi a un ritmo molto più veloce, riducendo i costi di produzione degli abiti e permettendo così ai consumatori di acquistarne di più. La macchina da cucire fu ampiamente adottata non solo nelle fabbriche ma anche tra le cucitrici domestiche grazie all'innovativa campagna di marketing di Singer e alla possibilità di effettuare l'acquisto a rate. Per secoli, il vestire alla moda è stato strutturato in modo classista, in quanto privilegio delle classi dominanti e superiori. L'innovazione e la tecnologia hanno contribuito alla democratizzazione della moda e alla creazione di un mercato di massa. La macchina da cucire, insieme al metro (1820) e ai brevetti per i cartamodelli (1850), ha democratizzato la moda e reso possibile la produzione di massa. Mentre gli abiti su misura richiedevano molto tempo per essere misurati, scegliere un modello e un tessuto e sottoporsi a numerose prove, gli abiti pronti per l'uso erano disponibili per essere provati e acquistati nei grandi magazzini (1850, USA). Nello stesso periodo si diffusero le prime riviste di moda, dando vita al *fashion system*.

I primi creatori di tendenze della moda hanno avuto sede a Parigi: Paul Poiret, Gabrielle Chanel e Jean Patou. Nel 1920, alcuni famosi stilisti aggiunsero linee di *prêt-à-porter* alle loro collezioni su ordinazione. I centri di moda si estesero presto all'Italia, al Regno Unito e agli Stati Uniti, e più tardi al Giappone. Negli anni '60 i laboratori si trasformarono sempre più in fabbriche; gli stilisti si concentrarono maggiormente sulle tecnologie e sullo sviluppo di macchine per automatizzare i processi. Courrèges fu tra i primi a dare valore all'innovazione nella produzione, considerando le linee di *prêt-à-porter* importanti quanto la "*haute couture*".

I tessuti sono stati un'altra area di grande innovazione. Storicamente, i tessuti sono stati ottenuti dalla lavorazione di un numero relativamente limitato di fibre naturali: lana, cotone, seta e lino. L'introduzione delle fibre artificiali (acetato di cellulosa, nylon, poliestere) ha rivoluzionato completamente l'industria tessile, aprendo una miriade di possibilità inesplorate per l'abbigliamento.

Negli anni '80, la globalizzazione ha favorito le importazioni di tessuti e abbigliamento, soprattutto dai Paesi in via di sviluppo. Per gestire in modo più efficiente la grande varietà di prodotti, i grandi volumi di

informazioni e per connettersi con le aziende di qualsiasi parte del mondo, la moda ha adottato rapidamente processi tecnologici e l'uso dell'EDI (Electronic Data Interchange) per le operazioni logistiche e lo scambio di documenti commerciali. Da allora si è assistito a una costante evoluzione delle attrezzature, dei sistemi informativi e della tecnologia (Tortora, P.G., 2015).

L'industria della moda rappresenta un settore di fondamentale importanza a livello globale, caratterizzato da una vasta portata economica, sociale e culturale.
La dinamicità della moda è uno dei suoi tratti distintivi più significativi. Le tendenze e i gusti dei consumatori sono in continua evoluzione e spesso sono guidati da influenze culturali, sociali ed economiche. Ciò significa che le aziende del settore devono essere costantemente attente a cogliere e ad interpretare questi mutamenti per rimanere al passo con le preferenze dei consumatori e per soddisfare le loro esigenze.
Questo settore ha un impatto economico considerevole, generando entrate significative a livello globale. Le spese dei consumatori nell'acquisto di abbigliamento, accessori e prodotti correlati costituiscono una fetta significativa dell'economia di molti paesi. Inoltre, l'industria della moda crea un gran numero di posti di lavoro, sia direttamente nella produzione e nella vendita di prodotti, che nelle attività di marketing, design, comunicazione e distribuzione.
L'industria della moda è un ambiente altamente competitivo, in cui le aziende devono adattarsi rapidamente ai cambiamenti e alle sfide del mercato per mantenere la propria rilevanza e sostenibilità. L'adozione di nuove tecnologie e strategie innovative è diventata essenziale per affrontare le mutevoli esigenze dei consumatori e per anticipare le tendenze future. Le aziende devono essere in grado di interpretare i segnali provenienti dal mercato e di adattare le loro offerte di prodotti e servizi al fine di mantenere una posizione competitiva e di soddisfare le aspettative dei consumatori.

La trasformazione digitale sta permettendo all'industria della moda, come a molti altri settori, di aumentare la propria capacità di produrre e utilizzare dati. La *datafication* si riferisce alla generazione di dati attraverso la digitalizzazione dei contenuti e il monitoraggio delle attività, comprese le attività e i fenomeni del mondo reale, tramite i sensori (OECD, 2015).
La *datafication* della moda si riferisce alla capacità di creare dati digitali in relazione ai prodotti e ai processi produttivi, consentendo di monitorarli, tracciarli, analizzarli e ottimizzarli. I dati, raccolti attraverso l'*e-commerce,* le vendite dirette, quelle *online* (fast fashion come Zara e H&M vendono in negozio e online) e i social media (Facebook, Blog, Instagram, ecc.), stanno rendendo il settore della moda sia un importante consumatore che un fornitore di dati che possono essere utili per ottenere dei *feedback* più diretti per cambiare le decisioni di produzione, migliorare la pianificazione e la logistica e personalizzare i prodotti.
Inoltre, i dati sulla moda possono essere utilizzati per alimentare il resto della catena del valore: grossisti, dettaglianti, finanza e settori di input del prodotto come il tessile e il cuoio (OECD, 2019).

L'accesso e l'utilizzo dei dati lungo la catena del valore possono aumentare l'efficienza e la resilienza della catena del valore della moda, ad esempio attraverso la tracciabilità, l'assistenza nella certificazione degli standard e la facilitazione delle catene logistiche commerciali. La domanda di tracciabilità e di trasparenza è in crescita e serve per monitorare le merci pirata, il commercio illecito e per supportare la tracciabilità e la rintracciabilità.
La creazione di nuovi dati e di conoscenze grazie alle tecnologie digitali consente di comprendere e gestire al meglio il settore della moda, riducendo l'incertezza e aumentando il coordinamento. La trasformazione digitale fornisce molti strumenti per ridurre le asimmetrie informative e creare conoscenze sui mercati, sui prodotti e sulle opportunità; può quindi favorire una migliore differenziazione dei prodotti e aprire nuovi mercati.
Grazie alle tecnologie digitali, i consumatori svolgono un ruolo attivo nella trasformazione della moda; non si tratta solo di ciò che i clienti acquistano, ma anche di ciò che postano su Instagram mostrando ciò che indossano a milioni di potenziali clienti. All'inizio era il re a dettare lo stile poi sono arrivate le riviste di moda e lo *star system* (star del cinema, cantanti, artisti, ecc.), oggi tutti possono essere esperti di moda e influencer (Taroy, D., 2015).

Tuttavia, l'impatto più importante sulla produzione e sulla distribuzione deve ancora arrivare: l'intelligenza artificiale potrebbe trasformare completamente il settore della moda. Essa appare la più significativa non solo perché è fortemente interdipendente con molte altre tecnologie come i *Big Data*, i *TAG*, l'*IoT* e altre ancora, ma anche perché gioca un ruolo fondamentale in tutti gli anelli della catena del valore.

1.2 Problema di ricerca e Obiettivi dello studio

Nonostante la fidelizzazione dei clienti nel settore della moda è molto rilevante, le aziende spesso si trovano di fronte alla sfida di comprendere i consumatori, prevedere le loro esigenze e preferenze in modo accurato e tempestivo. Le tradizionali tecniche di analisi dei dati potrebbero ormai non essere più sufficienti per affrontare la complessità e la vastità delle informazioni disponibili. Pertanto, sorge la necessità di esplorare l'applicazione di algoritmi di intelligenza artificiale per l'analisi predittiva dei dati dei clienti al fine di massimizzare l'efficacia delle iniziative di fidelizzazione in questo settore.

Nel contesto dinamico e altamente competitivo dell'industria della moda, la fidelizzazione dei clienti è diventata una priorità strategica per tutte le aziende. Con l'aumento della concorrenza e delle opzioni di acquisto disponibili per i consumatori, è fondamentale per le aziende sviluppare strategie efficaci per mantenere e aumentare la fedeltà dei clienti.
L'obiettivo di questo studio è quello di esplorare l'applicazione dell'analisi predittiva con algoritmi di intelligenza artificiale per migliorare la fidelizzazione dei clienti nel settore della moda. Si intende valutare l'efficacia di tali algoritmi nel prevedere il comportamento dei clienti, consentendo alle aziende di anticipare le loro esigenze, le preferenze e i comportamenti di acquisto.

In particolare, si mira a selezionare le caratteristiche rilevanti dei clienti e ad addestrare i modelli predittivi in grado di classificare i clienti in base al loro stato di fedeltà. L'obiettivo è anche quello di fornire delle raccomandazioni pratiche alle aziende e ai *manager* per l'implementazione dei sistemi di analisi predittiva basati sull'intelligenza artificiale. Attraverso l'utilizzo di dati storici e attuali sui clienti, questi algoritmi possono identificare i modelli e le correlazioni significative che permettono di fare delle previsioni accurate sulle azioni future dei clienti. Ad esempio, è possibile identificare quali prodotti o servizi potrebbero interessare maggiormente a un determinato cliente in base ai suoi acquisti precedenti, alle sue interazioni sui canali digitali o alle sue preferenze. Ciò consentirà alle aziende di personalizzare l'offerta e le comunicazioni in modo mirato, offrendo prodotti e servizi rilevanti e aumentando così le probabilità di fidelizzazione del cliente. Inoltre, l'analisi predittiva può anche contribuire a individuare segnali di allarme che indicano un potenziale rischio di *churn*, consentendo alle aziende di adottare misure preventive per mantenere la fedeltà dei clienti.

Attraverso questo studio, inoltre, si intende contribuire alla comprensione delle potenzialità dell'analisi predittiva e dell'intelligenza artificiale nell'ambito della fidelizzazione dei clienti nel settore della moda. I risultati e le conclusioni dello studio potranno fornire delle indicazioni preziose per le decisioni strategiche delle aziende, aiutandole ad ottimizzare le loro risorse e ad adattarsi alle esigenze mutevoli dei clienti, promuovendo relazioni durature e proficue con la clientela.

1.3 Gap di ricerca

Il gap di ricerca identificato in questa tesi riguarda la mancanza di studi precedentemente condotti che indaghino specificamente sull'applicazione del *machine learning* per la classificazione della *customer loyalty* nel settore della moda. Come emergerà nel capitolo successivo, quello della literature review, si rileva una carenza di ricerche che abbiano esaminato in modo esaustivo e approfondito l'utilizzo di tali approcci predittivi in questo ambito. Pertanto, si può affermare che esiste un vuoto di conoscenza riguardo all'efficacia e alle potenzialità dell'applicazione di tali algoritmi nell'industria della moda.

Questo gap di ricerca è rilevante perché questo settore è caratterizzato da una rapida evoluzione delle tendenze e dei gusti dei consumatori e le aziende sono costantemente alla ricerca di nuovi modi per comprendere al meglio i desideri e le preferenze dei loro clienti al fine di offrire esperienze personalizzate e costruire relazioni solide e durature.

1.4 Metodologia di ricerca

Per raggiungere gli obiettivi dello studio, è stata adottata la seguente metodologia di ricerca.

Inizialmente, è stato condotto un ampio esame della letteratura scientifica e dei lavori di ricerca esistenti riguardanti l’utilizzo del *machine learning* per classificare la *customer loyalty* nei diversi settori. Ciò ha permesso di acquisire una solida base teorica e comprensione dei concetti chiave in questo campo.

Successivamente, sono stati selezionati due set di dati appropriati per l'analisi, in particolare il "Transaction Details" e “Contact Active”, il primo contenente le informazioni dettagliate sulle transazioni nel periodo 2015-2022 e il secondo riguardante le informazioni sui clienti. Questi due dataset sono stati sottoposto ad un'adeguata preparazione e pulizia per garantire l'integrità e la qualità dei dati.
Per il compito di classificazione dei clienti in base allo stato di fedeltà, sono stati selezionati diversi algoritmi di intelligenza artificiale, tra cui algoritmi di apprendimento supervisionato come Random Forest, Decision Tree, Logistic Regression, Naïve Bayes, K-Neighbors e XGBoost.
Successivamente è stato eseguito l'addestramento dei modelli utilizzando due set di dati (uno con tutte le *features* e uno eliminando quelle più correlate tra di loro) e una volta completato, è stata effettuata un'attenta valutazione delle prestazioni dei modelli utilizzando le metriche appropriate. Tali metriche sono state selezionate in base agli obiettivi dello studio e alle caratteristiche specifiche del problema di classificazione dei clienti. L'obiettivo principale era determinare l'accuratezza e l'efficacia della soluzione di apprendimento automatico nel classificare correttamente i clienti in base al loro stato di fedeltà.

1.5 Struttura della tesi

Questa tesi è strutturata in diversi capitoli, ognuno dei quali affronta una specifica sezione di ricerca. Il Capitolo 1, ovvero l'introduzione, fornisce una panoramica del contesto, del problema di ricerca, degli obiettivi dello studio e della metodologia di ricerca adottata.
Il Capitolo 2 presenta una revisione approfondita della letteratura, esaminando le teorie, i modelli e le ricerche precedenti relative all'analisi predittiva e all'utilizzo di algoritmi di machine learning nei diversi settori.
Il Capitolo 3 illustra la metodologia di ricerca utilizzata nello studio, comprese le tecniche di raccolta dei dati, l'elaborazione e la preparazione dei dati, le analisi statistiche, nonché l'implementazione degli algoritmi di intelligenza artificiale.
Nel Capitolo 4 vengono riportati i risultati delle valutazioni delle prestazioni dei modelli e vengono discussi i principali *insights* ottenuti dall'analisi dei dati e discusse le implicazioni pratiche dei risultati.
Il Capitolo 5 fornisce le conclusioni dello studio dove vengono riassunti i principali risultati ottenuti, i limiti dello studio e vengono suggerite possibili direzioni future per la ricerca.
Infine, l'Appendice contiene il *link* che ricondurrà allo script dello studio condotto attraverso la piattaforma Spyder.

La struttura della tesi è stata concepita in modo tale da fornire una presentazione chiara e coerente dei risultati e delle conclusioni dello studio, guidando il lettore attraverso un percorso di ricerca e consentendo una comprensione completa del contesto, dei metodi e dei risultati della seguente analisi predittiva.

Capitolo 2 – Literature Review

2.1 Intelligenza Artificiale applicata al mondo Fashion

Da tempo immemorabile, la moda è stata intimamente associata all'essere umano.

Nella società contemporanea, la moda ha avuto un effetto significativo su ogni aspetto della vita sociale, causando e riflettendo cambiamenti nel panorama sociale, economico, politico e culturale. L'industria della moda è diventata uno dei più grandi segmenti dell'economia mondiale, con una stima di 3.000 miliardi di dollari nel 2018, che rappresenta il 2% del PIL globale.

D'altra parte, la crescente popolarità dei social media e la prosperità dell'*e-commerce* hanno prodotto enormi quantità di dati crossmediali sulla moda, come i dati condivisi dagli utenti, i dati delle sfilate rilasciati dai *brand* e i dati dei prodotti forniti dagli *e-commerce*, mostrando un insieme ricco e complesso di contenuti multimediali. Pertanto, la comprensione e l'analisi della semantica dei dati di moda su larga scala attraverso tecniche di *machine learning* e di *computer vision* è uno degli strumenti tecnologici e di analisi aziendale essenziali per rivoluzionare il settore e ridisegnare la meccanica della moda. Ad esempio, un numero crescente di stilisti e marchi famosi stanno sfruttando i principali social network per sondare le preferenze dei clienti, come opinioni, idee, *feedback* e tendenze.

A causa del suo impatto sociale ed economico, la gestione di dati crossmediali sulla moda con nuove tecniche è diventata una sfida interessante per gli informatici. Fortunatamente, negli ultimi anni gli studi sul mondo *fashion* hanno ricevuto una crescente attenzione da parte delle comunità di *computer vision* e *machine learning*. I ricercatori delle più importanti aziende tecnologiche stanno trasformando la moda a un ritmo più veloce che mai (Gu, X. et al., 2020).

L'industria della moda e dell'abbigliamento (F&A) è una delle maggiori economie che contribuisce per il 38% all'Asia Pacifica, per il 26% all'Europa e per il 22% al Nord America (Statista, 2019). Secondo Business of Fashion (2022), per le vendite di F&A nel 2023 è stata prevista una crescita potenzialmente molto lenta in Europa, Stati Uniti e Cina e secondo una previsione più ottimistica, la crescita in Cina dovrebbe essere del 7%, ma potrebbe scendere fino al 2%. Questa stima è leggermente inferiore negli Stati Uniti e più bassa ancora in Europa (Statista, 2023).

Con l'emergere della globalizzazione e della digitalizzazione, l'IA ha guadagnato attenzione per connettere le aziende a livello globale e nell'ultimo decennio, l'industria F&A ha utilizzato quest'ultima per migliorare i processi della catena di fornitura. Questo è stato importante perché il settore F&A è volatile ed è sempre impegnativo rispondere rapidamente ai cambiamenti delle tendenze e alle richieste dei consumatori in continua evoluzione (Giri, C. et al., 2019).

Secondo l'IA in Fashion Market Research Report 2022, sotto l'impatto cumulativo di COVID-19 (ReportLinker, 2022), la spesa globale per l'IA nel mercato della moda è stata stimata a 419,70 milioni di

dollari nel 2021 e dovrebbe raggiungere 500,66 milioni di dollari nel 2022 e si prevede che crescerà a un tasso di crescita annuale composto (CAGR) del 19,46% per raggiungere 1.220,11 milioni di dollari entro il 2027.

L'industria della moda è sull'orlo di un cambiamento senza precedenti e l'implementazione del *machine learning*, della *computer vision* e dell'intelligenza artificiale (IA) nelle applicazioni di moda sta aprendo molte nuove opportunità per questo settore (Mohammadi, S.O. et al., 2021).
Negli ultimi anni lo *shopping online* di abbigliamento è cresciuto a una velocità sorprendente soprattutto in relazione alla situazione del Coronavirus. Le persone di tutto il mondo hanno iniziato a vedere il potenziale dell'*e-commerce*, un settore in evoluzione che ha visto notevoli progressi ma che è ancora lontano dalla perfezione ma queste tecnologie avanzate possono influenzarlo ora più che mai (Mohammadi, S.O. et al., 2022).
Un ulteriore impatto della digitalizzazione si nota nel comportamento dei consumatori nel settore F&A. L'aumento della consapevolezza e l'avvento di nuovi mezzi di comunicazione sia offline che online ha cambiato il modello decisionale del consumatore contemporaneo, influenzato ormai da tutti i vari tipi di mezzi di comunicazione. È quindi importante creare delle piattaforme digitali per raccogliere efficientemente i dati. Questo obiettivo può essere raggiunto utilizzando i vantaggi offerti dalle tecnologie dell'informazione (IT), dalle tecniche di intelligenza artificiale (AI), dagli strumenti di analisi dei *big data* e da altre tecnologie attuali. È evidente che il settore F&A è uno dei più dinamici, con nuovi dati generati ogni volta che un nuovo capo di abbigliamento viene progettato, prodotto e poi venduto. Tuttavia, questo settore non ha ancora adottato in modo esteso i metodi di IA. L'industria utilizza ancora strumenti di calcolo basati su algoritmi classici e le moderne tecniche di IA sono confinate alla ricerca accademica. È quindi necessario che l'industria adotti queste nuove tecniche per avere un vantaggio competitivo e migliorare la redditività aziendale (Giri, C. et al., 2019).

Indubbiamente i *brand* ci credono e vogliono intraprendere questo nuovo percorso. Ad esempio, un team di Amazon ha sviluppato un algoritmo che impara a conoscere un particolare stile di moda e crea immagini simili da zero. Alibaba ha collaborato con GUESS per lanciare un *concept shop* di *FashionAI*, offrendo ai clienti un'esperienza di acquisto più ricca che combina i comportamenti di acquisto *online* e *offline*. Asos, ormai da anni utilizza l'intelligenza artificiale per gestire i dati dei consumatori ed indirizzarli negli acquisti facendo forza sulle loro preferenze; Zara e H&M usano l'IA per molti scopi legati alle catene di approvvigionamento, utilizzano gli algoritmi per prevedere le tendenze future e inseriscono i microchip nelle etichette per poter trovare i vari modelli e taglie per ottenere la piena trasparenza sull'inventario. Inoltre, molti *brand* sperano che entro il 2025 riusciranno ad utilizzare i *big data*, l'intelligenza artificiale e l'apprendimento automatico per poter creare le "*smart warehouse*" (Gu, X. et al., 2020).

Quando si parla di IA applicata al mondo del *fashion* il grande protagonista è il cliente, la sua buona esperienza d'acquisto, il momento dell'incontro con il *brand*: in altre parole il *customer relationship management*.

L’intelligenza artificiale è in grado di perfezionare tutto il *customer journey*. Quasi tutti gli *e-commerce* utilizzano i *tag* e il riconoscimento visivo per proporre ai clienti prodotti simili o correlati a quelli già acquistati. Inoltre, lo stesso meccanismo potrebbe essere in grado di suggerire agli utenti *outfit* adatti per ogni occasione.

Il successo dell’intelligenza artificiale nel *fashion marketing* e nel *fashion retail* richiede di tenere in considerazione un consumatore medio che, soprattutto nel caso del *fast fashion*, ma più in generale anche per quanto riguarda l’acquisto *online*, è completamente cambiato rispetto qualche anno fa.

Ad oggi i Millennial e la Gen Z rappresentano la generazione di consumatori più numerosa di chi acquista; sono attenti al denaro, sono alla ricerca di prodotti che soddisfino le loro aspettative, amano la personalizzazione e sono molto attenti agli aspetti esperienziali dell’acquisto.

Non dovrebbe sorprendere, quindi, che la maggior parte delle applicazioni dell’IA nel mondo della moda si concentri principalmente sul momento in cui i consumatori incontrano i *brand*.

L'intelligenza artificiale viene utilizzata nei negozi per semplificare il processo di vendita e garantire un'esperienza di acquisto personalizzata, attirando più persone nel negozio. Infatti, con i *software* basati su meccanismi di intelligenza artificiale, si possono eseguire diverse azioni come: visualizzare o implementare i dati dei clienti in tempo reale, ottimizzare l’inventario e gestire in modo efficace i vari servizi *omnichannel*.

Il mondo della moda sta sempre più orientandosi verso delle soluzioni legate alla tecnologia e all'intelligenza artificiale, anche per la crescente consapevolezza dell'impatto ambientale dei processi produttivi e di distribuzione. Oggi, sia i consumatori che i *brand* di moda sono sempre più attenti all'etica coinvolta nella realizzazione di un prodotto, soprattutto per l'impatto notevole sull’inquinamento che hanno gli acquisti di moda *online*. Per questo motivo, Burberry ha sviluppato "*Voyage*", un’applicazione per permettere ai consumatori di seguire l'intero ciclo di vita di un capo, dalla raccolta della materia prima fino alla vendita. Attualmente, il 44% dei rivenditori europei sta adottando *l'Internet of Things* (IoT) per migliorare la visibilità e il controllo della loro catena di approvvigionamento, mentre un altro 36% prevede di integrare questa tecnologia entro i prossimi tre anni.

Altri *brand*, come H&M e Adidas, hanno utilizzato la tecnologia come fattore di crescita sostenibile, diventando un punto chiave della loro strategia. Inoltre, Inditex ha annunciato obiettivi molto ambiziosi per diventare più sostenibile, utilizzando materie prime "*green*", come cotone biologico, poliestere riciclato e lyocell e implementando processi più rispettosi dei consumi di acqua ed energia entro il 2025. Inoltre, vi sono iniziative di economia circolare, come la raccolta di abiti usati per il successivo riutilizzo o di riciclaggio per beneficenza, e il lancio di una piattaforma di negozi eco-efficiente. È fondamentale che la trasformazione digitale e il deciso progresso verso *standard* di sostenibilità più esigenti siano supportati dall'efficienza del modello di *business*, che si basi sull'offrire ai clienti moda di qualità. In questo modo, la tecnologia e l'IA possono contribuire a rendere l'industria della moda più sostenibile ed etica (Jin, B.E. et al., 2020).

2.2 La Customer Loyalty e il suo ruolo strategico

La fedeltà è una tematica rilevante nell'ambito del marketing e la rilevanza del concetto è determinata dai vantaggi legati alla conservazione dei clienti (McMullan, R., 2005).

Le ricerche effettuate sulla *customer loyalty* hanno da tempo messo in risalto la rilevanza dei clienti fedeli in quanto tendono a spendere di più rispetto agli occasionali o ai nuovi. Questo è dovuto al fatto che essi hanno una maggiore fiducia nella marca e nei suoi prodotti, il che li spinge ad acquistare di più; essi sono anche più propensi ad acquistare una maggiore varietà di prodotti rispetto ai clienti occasionali o ai nuovi. Questo è dovuto al fatto che hanno già sperimentato la qualità dei prodotti e sono convinti della loro efficacia, sono anche disposti a pagare un prezzo più alto per i prodotti del *brand* ed infine, sono spesso i migliori ambasciatori. Questi clienti sono molto soddisfatti dei prodotti dell'impresa e tendono a raccomandarli ad amici, familiari e conoscenti, generando un passaparola positivo per l'impresa (Zeithaml V.A., et al., 1996).

Jacoby e Kyner (1973), hanno dato un contributo importante allo studio della *customer loyalty* e la loro definizione rappresenta uno dei concetti maggiormente condivisi e noti in letteratura che la definisce come "una risposta comportamentale, premeditata, espressa nel tempo da un'unità decisionale di acquisto, rispetto a una o più marche alternative, dipendente da un processo psicologico".

Nella letteratura di marketing è possibile trovare diverse definizioni utilizzate per descrivere la fedeltà dei clienti, così come molteplici approcci per quantificarla. Secondo quanto affermato da Engel e Blackwell (1982), la fedeltà del consumatore è rappresentata dalla preferenza, sia a livello di atteggiamento che comportamentale, verso uno o più marchi in una specifica categoria di prodotto dimostrata da un consumatore entro un determinato periodo di tempo. Assael (1992) definisce la fedeltà come l'atteggiamento positivo verso un marchio che si esprime attraverso l'acquisto regolare nel corso del tempo dei prodotti dell'azienda in questione.

Per valutare la fedeltà dei clienti, sia studiosi che professionisti hanno utilizzato nel corso del tempo misure sia comportamentali che attitudinali (Oliver, R.L., 1999). In termini comportamentali, la *customer loyalty* è caratterizzata dall'acquirente che fornisce un supporto costante e promuove continuamente l'impresa (Yang, Z. et al., 2004) si può anche definire come la frequenza con cui un consumatore seleziona un particolare prodotto o servizio all'interno di una specifica categoria merceologica rispetto al totale degli acquisti effettuati in quella stessa categoria di prodotti (Neal, W.D., 1999). Tuttavia, questa misura della fedeltà dei clienti presenta due problemi principali. Il primo è correlato al fatto che il ripetere gli acquisti non è sempre dovuto ad un legame psicologico con l'azienda (Tepeci, M., 1999) ed il secondo determinato dal fatto che gli acquisti ripetuti non necessariamente riflettono le intenzioni del cliente (Yang, Z. et al.,2004).

Focalizzando l'analisi sull'atteggiamento del consumatore, si valuta la *customer loyalty* in base alla forza del legame emotivo e psicologico tra l'azienda e il cliente (Bowen J.T. et al., 2001). In base a quanto detto, l'*attitudinal loyalty* viene quindi definita come la volontà del consumatore di rafforzare la relazione con l'impresa (Czepiel, J.A et al., 1987).

L'attenzione rivolta alla *customer loyalty* deriva dall'importanza che tale fattore riveste nella gestione delle relazioni tra l'impresa e il cliente. Il consumatore cerca di stabilire un rapporto di fiducia con l'azienda presso cui effettua i propri acquisti, in modo da ridurre i costi legati alla ricerca di informazioni sui prodotti e il rischio di dover acquistare da un'azienda sconosciuta. In questo modo, si riducono i tempi di valutazione e si minimizza il rischio associato alla scelta di un nuovo fornitore e questi sono solo alcuni dei vantaggi per il consumatore. Tuttavia, il comportamento d'acquisto ripetuto è solo un prerequisito per la fedeltà, che non può essere garantita se la fedeltà stessa è il risultato di comportamenti inerziali o di una mancanza di alternative valide (Jacoby, J. et al., 1973).

In questa prospettiva, la fedeltà del cliente si sviluppa solo se è soddisfatto dell'offerta, rendendo il comportamento d'acquisto ripetitivo un fattore determinante. Infatti, la *customer satisfaction* è universalmente riconosciuta come uno dei principali indicatori della fedeltà (Garbarino, E. et al., 1999). La letteratura riconosce ampliamente i benefici che le imprese possono ottenere dalla fidelizzazione della clientela. Tra i principali benefici vi sono innanzitutto i minori costi associati al mantenimento della clientela esistente rispetto a quelli necessari per acquisirne dei nuovi, specialmente in mercati maturi e competitivi (Ehrenberg A.S.C. et al., 2000). I clienti fedeli sono inoltre meno sensibili al prezzo (Krishnamurthi, L. et al., 1991), ad eventuali esperienze negative, alle politiche pubblicitarie/promozionali e alle proposte dai *competitor* (Jensen, J.M. et al., 2006). Inoltre, è stato riscontrato che i clienti fedeli sono maggiormente inclini a estendere la loro fiducia verso la marca su tutta la gamma di prodotti da essa supportati (Grayson, K. et al., 1999). I clienti fedeli diventano dei veri e propri canali di informazione per il *brand*, diffondendo informazioni in modo informale all'interno dei loro *network* di amici, familiari e conoscenti, creando un effetto positivo di *Word Of Mounth* a vantaggio del *brand* (Shoemaker, S. et al. 1999). Possedere una base solida di clienti fedeli può rappresentare un vantaggio competitivo duraturo e sostenibile per le aziende, con possibili effetti positivi sulla redditività (Reichheld, F. 1993). Per tali motivi, la *customer loyalty* rappresenta una fonte di profitto e un importante *asset* per l'impresa (Anderson, E.W. et al., 2000).

Esistono diverse teorie che hanno influenzato la *customer loyalty*. Ad esempio, la teoria del comportamento pianificato di Fishbein e Ajzen (1975) sostiene che la comprensione del comportamento dei consumatori è fondamentale per il mantenimento della loro fedeltà. Inoltre, le teorie dell'attenzione normativa e della psicologia individuale vengono utilizzate per spiegare aspetti come l'acquisto d'impulso, la dissonanza cognitiva e la soddisfazione dei consumatori, che alla fine contribuiscono alla creazione della loro fedeltà (Lin, C.T. et al., 2018). Un recente studio ha dimostrato che le teorie del flusso e del valore percepito hanno un effetto significativo sullo sviluppo di strategie volte ad aumentare la fedeltà dei consumatori verso le PMI (Guerra-Tamez, C.R. et al., 2021).

2.3 E-Loyalty

L'*e-commerce* è un paradigma commerciale che digitalizza la comunicazione, l'integrazione delle informazioni, le transazioni e la condivisione dei dati, riducendo i vincoli di tempo e di spazio.

Questo paradigma presenta molti vantaggi significativi rispetto alle forme più convenzionali di organizzazione aziendale. Ad esempio, poiché l'intero processo, che include la distribuzione di beni e servizi, è automatizzato, i costi possono essere drasticamente ridotti. L'*e-commerce* può assumere diverse forme, ma lo shopping *online* è una delle più diffuse. Pertanto, è essenziale avere una buona comprensione delle variabili coinvolte nell'acquisto *online* ed avere un modello in grado di prevedere come l'*e-loyalty* dei clienti sarà influenzata dai vari elementi. La così chiamata *e-loyalty* presenta somiglianze significative con la *loyalty in store*, poiché entrambe implicano due azioni fondamentali compiute dal consumatore nei confronti dell'azienda o di un brand specifico: la frequente visita e l'acquisto ripetuto dei prodotti disponibili.

L'avvento e la crescita del commercio elettronico "*Business to Consumer*" (B2C) hanno amplificato l'importanza di costruire una base di visitatori fedeli di un *e-commerce*. La maggior parte dei modelli di *e-commerce* B2C si basava inizialmente su uno sforzo intensivo per generare una base di clienti sufficientemente ampia e, in seguito, ottenere una redditività basata su un "potenziale di reddito a vita" da parte di ciascun cliente fedele. Oggi, alcune aziende hanno un numero enorme di transazioni all'ora e l'uso di tecnologie come l'*e-commerce* è indispensabile.

La riduzione dei tempi, dei costi, degli errori, degli svantaggi del denaro cartaceo e delle autorizzazioni nel commercio elettronico incoraggiano molto le aziende e i loro clienti ad utilizzare quest'ultimo sempre di più. Tuttavia, le aziende *online* perdono clienti a causa dell'ambiente commerciale competitivo poichè l'*e-commerce* è diventato ormai un bene essenziale per tutte le aziende (Saibaba, S., 2023).

Secondo eMarketer (2022), si prevede che le vendite al dettaglio degli *e-commerce* in tutto il mondo raggiungeranno i 7,391 trilioni di dollari entro il 2025, rispetto ai 4,938 trilioni di dollari del 2021, mentre la cifra esatta era di soli 1,336 trilioni di dollari nel 2014. Ciò indica il rapido tasso di crescita dell'*e-commerce* dovuto alla maggiore adozione di Internet e alla propensione ad acquistare beni e servizi *online*.

Inoltre, il contributo delle vendite degli *e-commerce* al totale delle vendite al dettaglio a livello mondiale era solo del 7,4% nel 2014, ma è aumentato più del doppio fino a raggiungere il 18,8% nel 2021, e si stima che lo stesso raggiungerà il 24% entro il 2026.

Secondo il rapporto di Dynata (2022), la possibilità di fare acquisti in qualsiasi momento è stato il vantaggio più citato dai consumatori di tutto il mondo. Gran parte degli intervistati ha espresso la propria preferenza per questo canale data la sua flessibilità e "i prezzi migliori" sono stati il secondo vantaggio principale degli acquisti online, con quasi quattro risposte su dieci. Il rapporto afferma inoltre che, secondo i consumatori globali, gli svantaggi più significativi dell'*e-commerce* sono legati alle proprietà fisiche o alla qualità del prodotto. Il 46% degli intervistati in alcuni Paesi del mondo ritiene che l'*e-commerce* non offra la possibilità di toccare o sentire gli articoli.

Kumar e Ayodeji (2021) hanno affermato che i rivenditori *online* dovrebbero dare priorità alla qualità delle informazioni e alla qualità del sistema dei loro siti web per aumentare la soddisfazione dei clienti, che porta a una decisione di riacquisto. A causa delle caratteristiche multicanale, come l'integrazione dei social media, essi sostengono che l'interazione degli acquirenti *online* con gli *e-commerce* dovrebbe essere misurata utilizzando una metodologia basata sull'analisi digitale.

2.4 AI e Customer Loyalty

Ad oggi, ci sono opinioni ottimistiche secondo cui l'apprendimento automatico avrà un impatto significativo sull'*e-commerce* nei prossimi cinque anni. Gli *e-commerce* registrano ogni giorno un numero enorme di visite degli utenti, ma un gran numero di visitatori viene perso. Se si riesce a combinare le informazioni sulle visite degli utenti e sul loro consumo sul sito è possibile ricavarne alcune informazioni utili che possono aiutare le aziende a capirne rapidamente la perdita così da poter successivamente apportare le modifiche giuste, migliorare l'esperienza degli utenti del *brand* e fidelizzarne un maggior numero sempre maggiore. Lo sviluppo dello *smart shopping* e la sua rapida diffusione hanno portato alla generazione di grandi quantità di dati a una velocità senza precedenti (Mohammadi, M. et al., 2018).
Purtroppo, a causa della mancanza di meccanismi e di *standard* consolidati che traggano vantaggio dalla disponibilità di tali dati, la maggior parte di essi viene sprecata senza estrarre informazioni e conoscenze potenzialmente utili (N. Taherkhani et al., 2016). Inoltre, la natura altamente dinamica dello *smart shopping* richiede una nuova generazione di metodi di *machine learning* in grado di adattarsi in modo flessibile alla natura dinamica dei dati per eseguire analisi e imparare da essi in tempo reale (Lemley, J. et al., 2017).

Sembra che tutti stiano cercando di identificare in che modo l'intelligenza artificiale (IA) possa avvantaggiare il proprio business. Uno studio (Forbs, 2019) ha rilevato che nel 2017 l'80% delle imprese aveva "una qualche forma di intelligenza artificiale in produzione" e il 30% "stava pianificando di espandere i propri investimenti in intelligenza artificiale". Lo stesso studio ha rilevato che il miglioramento delle esperienze dei clienti ha motivato il 62% delle imprese a investire nell'intelligenza artificiale.
Il marketing focalizzato sui clienti e basato sui dati ed è progettato per modificare il comportamento dei clienti in modo positivo sia per il marchio che per il cliente. L'obiettivo a breve termine è raggiungere obiettivi finanziari come l'aumento delle vendite e delle visite, mentre l'obiettivo a lungo termine è stabilire relazioni preziose, durature e reciprocamente vantaggiose che isolano il marchio dalle offerte competitive del mercato. Questa disciplina di marketing è senza dubbio la più importante dell'era moderna, poiché il marketing di fidelizzazione è altamente misurabile e finanziariamente responsabile. Le tattiche delle campagne e lo sviluppo delle offerte sono entrambi ispirati da ciò che si può apprendere analizzando i dati dei clienti. Pianifichiamo, eseguiamo e quindi misuriamo i risultati per assicurarci di raggiungere gli obiettivi e di essere in grado di riferire chiaramente agli *stakeholder*.

È noto che le persone generano impronte digitali attraverso le visite, gli acquisti, le ricerche, i post, le recensioni e altre attività quotidiane. I *brand* si sono distinti nella raccolta di questi dati e hanno aggiunto dei metadati esterni per aggiungere un contesto ricco ad un semplice acquisto o transazione con l'obiettivo di prevedere la prossima mossa di un cliente.

La segmentazione è una tecnica che fa parte del set di strumenti di marketing focalizzato sui clienti sin dall'inizio del marketing di fidelizzazione di oltre 20 anni fa. I metodi di segmentazione sono diventati sempre più sofisticati e sono degli indispensabili trampolini di lancio nel viaggio verso il marketing *one-to-one*.

Anche se la segmentazione è stata efficace in passato ma è tempo di riconoscere i limiti intrinseci di questo processo. I marketer riconoscono che ormai i segmenti risultanti sono delle generalizzazioni eccessive o forniscono una visione statica del cliente. Al suo livello più elementare, l'intelligenza artificiale può automatizzare delle attività semplici e ripetibili e accelerare il riconoscimento dei modelli. Questo apre le porte all'efficienza del marketing e accelera l'adempimento della promessa di fornire "l'offerta giusta alla persona giusta al momento giusto e nel canale giusto". Fornisce anche personalizzazione, ovvero modi pratici in cui un marchio può dimostrare che sta ascoltando i propri clienti, sta imparando di più su ciò che vogliono e sta formulando offerte per i prodotti che desiderano effettivamente invece di ciò che il marchio vuole vendere. Questa è esattamente l'essenza di come il marketing basato sui dati può creare fiducia con i clienti e fedeltà a lungo termine. Inoltre, migliorare l'analisi con l'intelligenza artificiale può ottimizzare le singole esperienze su larga scala e può rendere possibile l'ipersegmentazione ovvero, puntare su gruppi più precisi di clienti che condividono attributi e comportamenti specifici.

Come passaggio successivo, si possono ottimizzare i contenuti dinamici per la distribuzione su tutti i canali in modo tempestivo, con il risultato di un'esperienza del cliente piacevole e di livelli di soddisfazione più elevati. I clienti stanno diventando ogni giorno sempre più consapevoli del valore dei loro dati e di ciò che rappresentano per i marchi ed offrire un'esperienza altamente personalizzata comunica chiaramente al cliente che il suo marchio preferito "capisce" i suoi bisogni. Il marketing di fidelizzazione si concentra sulla creazione di relazioni a lungo termine reciprocamente vantaggiose e creare redditività da una vera relazione *one-to-one* con un cliente è l'apice dell'esecuzione; tuttavia, farlo non è possibile, tanto meno sostenibile, su larga scala senza IA (Hanifin, B., 2019).

Il cambiamento in corso causato dal *machine learning* ha gettato le basi per aumentare il valore dei dati dei clienti ed anticipare i modelli di comportamento degli acquirenti, al fine di prevedere le future decisioni di acquisto e creare un'esperienza personalizzata.

Se i rivenditori e i *brand* investono nel *machine Learning*, i benefici si estenderanno ben oltre le prestazioni aziendali. Nel frattempo, gli esperti di programmi di fidelizzazione dei clienti prevedono che, grazie all'utilizzo di strumenti di marketing automatizzati, il servizio clienti sarà elevato e il *customer journey* sarà ottimizzato. Data la natura dinamica del *machine learning*, gli addetti al marketing si imbatteranno nelle seguenti sfide per passare dal tradizionale modello di database dei clienti a un CRM avanzato, alimentato da soluzioni IA e

quindi gestire la raccolta dei dati dei clienti attraverso tutti i punti di contatto in tempo reale, analizzare i preziosi *insight* dei clienti, le attività comportamentali e transazionali per trasformarli in conoscenze di *business*, mantenere e migliorare le prestazioni dell'algoritmo e gestire allo stesso tempo enormi volumi di dati sui clienti ed infine, applicare la potenza del *machine learning* ai programmi di fidelizzazione dei clienti.
Gli strumenti di automazione forniscono una conoscenza più approfondita e precisa dei dati demografici, delle preferenze e di altre importanti informazioni sui clienti. Pertanto, è di grande interesse per i rivenditori e i marchi conquistare le conoscenze di cui sopra e riuscire finalmente ad ottimizzare le campagne di marketing, modificare la messaggistica di comunicazione, quando è necessario, e fornire contenuti mirati più velocemente e in modo più accurato. Inoltre, gli strumenti di *machine learning* sono in grado di prevedere le frodi nei programmi di fidelizzazione, rilevando incongruenze e attività anomale.
Il *machine learning* svolge quindi un ruolo significativo nella fidelizzazione del marchio, aiuta a creare un *customer journey* che rassicuri la fidelizzazione dei clienti esistenti e l'acquisizione di nuovi, aiuta ad eseguire campagne di fidelizzazione di successo sfruttando i dati dei clienti per sbloccare il loro comportamento.
La fornitura di un servizio clienti olistico su tutti i diversi canali può incrementare KPI significativi: tassi di conversione/ritenzione, visite al negozio e carrello medio. Riconoscendo l'importanza delle decisioni basate sui dati, gli strumenti di *machine learning* non solo aiuteranno i *marketer* a guadagnare più tempo per pensare ad alto livello, ma anche a trasformare il tradizionale rapporto dare-avere in un percorso di coinvolgimento dei clienti di successo (Siti Zulaikha, S. et al., 2020).

2.5 Modelli di classificazione della Customer Loyalty utilizzati in diverse industry

Il concetto di *customer loyalty* ha suscitato l'interesse di numerose ricerche negli ultimi anni, portando alla realizzazione di numerosi studi. In questa sezione, sono state selezionate e discusse diverse ricerche esistenti per ottenere maggiori informazioni che saranno utilizzate per condurre la ricerca proposta. I seguenti *paper* sono stati suddivisi in base al contesto di applicazione per fornire una panoramica completa delle ricerche esistenti su questo argomento. In particolare, sono stati identificati cinque ambiti principali di utilizzo di questi algoritmi per la classificazione della *customer loyalty*: (1) il settore di credito, (2) quello delle telecomunicazioni, (3) quello alberghiero, (4) del fast moving consumer goods ed infine (5) quello dei servizi multimediali. In ogni *paper* analizzato, si sono riscontrati differenti algoritmi di *machine learning* impiegati per raggiungere l'obiettivo di classificare la *customer loyalty*.

Nel settore delle telecomunicazioni, Wissam Nazeer Wassouf et al. (2020) hanno analizzato l'applicazione di quattro algoritmi di *machine learning* per la classificazione della *customer loyalty*, ovvero il Multilayer Perception Classifier, il Decision Tree Classifier, il Random Forest Classifier e il Gradient-Boosted-Tree. Gli studiosi hanno fatto affidamento alla segmentazione TFM e alla definizione dei livelli di fedeltà. Gli algoritmi di classificazione sono stati applicati utilizzando i livelli di fedeltà come categorie di classificazione e le caratteristiche comportamentali. I risultati sono stati comparati e il modello di classificazione con l'accuratezza

migliore, il Gradient-Boosted-Tree, è stato selezionato. Da questo modello sono state identificate le regole di previsione della fedeltà, che esprimono la correlazione tra le caratteristiche comportamentali e le categorie di classificazione, consentendo l'identificazione delle cause della fedeltà in ogni segmento. Un ulteriore vantaggio dell'utilizzo degli algoritmi di classificazione è stato la costruzione di un modello predittivo preciso per la classificazione dei nuovi utenti sulla base della loro fedeltà.
Un altro documento che ha esaminato la classificazione della *customer loyalty* nel settore delle telecomunicazioni è stato redatto da Oladapo K. A. et al. (2018) che hanno condotto un'analisi di regressione logistica per valutare gli effetti delle questioni riguardanti la fatturazione, i servizi a valore aggiunto e il servizio di messaggistica breve sulla probabilità di mantenere i clienti. Il modello di regressione logistica ha dimostrato di essere statisticamente significativo ed ha spiegato l'89,3% della varianza nella fidelizzazione dei clienti e ha correttamente classificato il 95,5% dei problemi. Si è riscontrato che risolvendo i problemi di fatturazione si ha una maggiore probabilità di fidelizzare i clienti, mentre i servizi a valore aggiunto e il servizio di messaggistica breve sono associati alla probabilità di fidelizzazione dei clienti.

Nel settore alberghiero, l'articolo di Youngkeun Choi e Jae Won Choi (2020) ha impiegato esclusivamente il Decision Tree per prevedere e classificare i *big data*. Il presente lavoro ha adottato due approcci principali: in primo luogo, si è inteso approfondire la comprensione del ruolo delle variabili nella modellazione della previsione della fedeltà dei clienti degli hotel e in secondo luogo, lo studio ha mirato a valutare le prestazioni predittive degli alberi decisionali. Sulla base dei risultati ottenuti, è stato possibile prevedere la fedeltà dei clienti dell'hotel considerando diversi fattori individuali. L'algoritmo ha ottenuto un tasso di accuratezza pari a 0,989, il che implica che il tasso di errore è stato pari a 0,011. Questa applicazione è in grado di aiutare le aziende alberghiere a gestire i dati personali dei clienti e a prendere decisioni più rapide se hanno già a disposizione il profilo dell'utente.
Per quanto concerne lo studio condotto da Iskandar Zul Putera Hamdan e Muhaini Othman (2022) hanno analizzando i dati delle prenotazioni alberghiere per prevedere la fedeltà dei clienti nel settore alberghiero utilizzando tre algoritmi di classificazione selezionati: Logistic Regression, Decision Tree e Random Forest. Confrontando i risultati dei tre algoritmi, si può concludere che l'algoritmo Decision Tree si è dimostrato il migliore da utilizzare nell'analisi del set di dati delle prenotazioni alberghiere, poiché ha generato un punteggio del 71,44% in termini di accuratezza, il più alto tra gli algoritmi selezionati.

Nel settore del fast moving consumer goods, lo studio condotto da H. Sulistiani et al. (2019) ha confrontato l'accuratezza di due metodi di classificazione, il Support Vector Machine (SVM) e il Naïve Bayes, per identificare i fattori rilevanti che influenzano le prestazioni della classificazione della fedeltà dei clienti.
È stato dimostrato che il metodo SVM ha valori più alti *accuracy*, *precision*, *recall* e *f-1 Score*. Applicando il DMI come metodo di selezione delle caratteristiche, si sono ottenute cinque caratteristiche selezionate, ovvero Età, Commento, Spese totali, Consumo medio e Indirizzo. Il valore di accuratezza del metodo SVM è del

76,42% quando si utilizzano tutte le caratteristiche, mentre il metodo Naïve Bayes ha solo un valore del 72,54%. Applicando il metodo di selezione delle caratteristiche DMI, il valore di accuratezza di SVM è stato del 73,57% e quello di Naive Bayes solo del 70,46%. I risultati dell'implementazione di questo studio mostrano una diminuzione dell'accuratezza dei classificatori e l'SVM ha portato a una maggiore precisione delle prestazioni rispetto al Naïve Bayes.

Un altro campo di applicazione che merita di essere considerato è quello delle società di servizi multimediali. Sardjoeni Moedjiono et al. (2016) hanno redatto questa ricerca basata sulla scelta di un modello che utilizza gli algoritmi di segmentazione k-means e di classificazione C4.5. Oltre a generare un modello di semplice comprensione, questo modello ha anche un alto livello di accuratezza nella classificazione dei clienti. Da questa ricerca, con l'applicazione del solo modello dell'algoritmo C4.5, l'accuratezza è abbastanza buona, circa il 69,23% senza segmentazione k-means, mentre con la segmentazione k-means l'accuratezza è del 79,33%. Il valore di AUC prima dell'ottimizzazione è pari a 0,723 mentre dopo l'ottimizzazione, il valore AUC è aumentato a 0,831.

Infine, con riferimento al settore di fornitura dei servizi credito, sulla base dei risultati delle ricerche che i ricercatori Ridlo Muttaqien, et al. (2021) hanno effettuato sui dati raccolti in merito all'utilizzo dell'algoritmo C4.5 per la previsione della fedeltà dei clienti presso questo tipo di società, si può concludere che dai risultati ottenuti mediante la matrice di confusione è stato raggiunto un elevato livello di accuratezza, ovvero il 90% nei dati del test 1 e il 94% nei dati del test 2.

L’analisi degli studi esistenti mostra che, in ambito accademico, la ricerca sulla *customer loyalty* è un tema importante nella gestione delle relazioni con i clienti e i diversi *paper* analizzati hanno evidenziato la validità degli algoritmi di *machine learning* per la classificazione della *customer loyalty* in diversi ambiti di applicazione.

Capitolo 3 – Materiali e Metodi

3.1 Ambiente di lavoro

Prima di focalizzarci sullo sviluppo effettivo, è necessario considerare un ulteriore aspetto, ovvero le tecnologie utilizzate. Esistono innumerevoli linguaggi di programmazione disponibili per progetti di *machine learning*, ma questo capitolo si concentrerà su quelli selezionati. Inoltre, saranno descritte solo le librerie più importanti, per comprendere meglio il loro utilizzo.

Python è un linguaggio di programmazione ad alto livello, orientato agli oggetti, che può essere utilizzato per una vasta gamma di applicazioni, tra cui lo sviluppo di applicazioni distribuite, *scripting*, computazione numerica e *system testing* ed è stato scelto perché, per quanto riguarda il *machine learning*, è il linguaggio ideale grazie alle sue librerie per l'analisi dei dati e il calcolo numerico, come Numpy, Pandas, Scikit-learn.

Anaconda è una distribuzione gratuita e open-source dei linguaggi di programmazione Python e R.

Pandas è una libreria software scritta in Python per la manipolazione e l'analisi dei dati. In particolare, offre strutture dati e operazioni per la manipolazione di tabelle numeriche e serie temporali. È stato fondamentale sin dall'inizio dello sviluppo del progetto per creare un *dataframe* iniziale, nonché per fornire altre funzionalità necessarie.
NumPy è una libreria *open source* e fornisce supporto per le grandi matrici e per gli *array* multidimensionali, insieme ad una vasta gamma di funzioni matematiche per poter operare efficientemente su queste strutture di dati. Nel progetto in questione, NumPy è stata ampiamente utilizzata per la gestione degli *array*, nonché per le funzioni matematiche messe a disposizione per calcolare velocemente medie, deviazioni standard e altre operazioni matematiche.
Scikit-learn è una libreria *open source* di *machine learning* per il linguaggio di programmazione Python, che comprende algoritmi di classificazione, regressione, clustering, macchine a vettori di supporto, regressione logistica, classificazione bayesiana, k-means e DBSCAN. Questa libreria rappresenta il nucleo fondamentale poiché contiene tutti gli algoritmi utilizzati, la cui implementazione e gli *import* specifici saranno descritti dettagliatamente in questo capitolo.
Matplotlib è una libreria per la creazione di grafici nel linguaggio di programmazione Python, che utilizza anche la libreria matematica NumPy. È possibile generare vari tipi di grafici, come grafici a barre, istogrammi, diagrammi a dispersione e altri ancora. Grazie all'interfaccia orientata agli oggetti, l'utente ha il pieno controllo degli stili del grafico.
Ed infine, Seaborn, una libreria di visualizzazione dati in python che potenzia gli strumenti di matplotlib e fornisce un'interfaccia di alto livello per disegnare grafici statistici attraenti e informativi.

3.2 Dataset e Features

I dati utilizzati per la presente tesi fanno riferimento a due dataset “Transaction fashion brand” e “Contact Active”. Entrambi i dataset utilizzati riguardano il periodo che va dal 2015 al 2022 e presentano informazioni di grande interesse per l'analisi. Il primo *dataset*, "Transaction fashion brand", si riferisce alle transazioni effettuate dai clienti di un noto *brand* di moda e si compone di un elevato numero di osservazioni (1.021.739) e 65 features principalmente in formato categorico (es. uomo/donna). La vasta quantità di informazioni presenti in questo *dataset* permette di analizzare in modo dettagliato i comportamenti degli utenti nei confronti del *brand*, e di individuare eventuali *pattern* o tendenze che possano essere utili per la pianificazione di future strategie di marketing. In particolare, l'analisi di questo *dataset* ha permesso di individuare alcune tendenze di acquisto tra i clienti del *brand*, come ad esempio i prodotti più venduti, le categorie di prodotti preferite e il valore medio degli acquisti.

Il secondo *dataset*, "Contact Active", riguarda invece informazioni personali degli utenti e anch’esso presenta un elevato numero di osservazioni (384.352) e 135 features in formato categorico. Questo *dataset* consente di analizzare le caratteristiche degli utenti che hanno interagito con il *brand*, e di individuare eventuali fattori determinanti per l'acquisizione di nuovi clienti o per la fidelizzazione di quelli già esistenti come l’area in cui effettuano gli acquisti, le preferenze di acquisto dei clienti e le loro preferenze riguardo ai trattamenti dei dati. Inoltre, all’interno di questo *dataset* è presente la *lable* (y) denominata “Loyalty”.

La scelta di utilizzare questi due *dataset* è stata motivata dall'obiettivo di fornire una panoramica completa e dettagliata del contesto in cui si muove il *brand* analizzato, in modo da identificare eventuali criticità ed individuare eventuali strategie di miglioramento. La presenza di due *dataset* distinti ha permesso inoltre di integrare informazioni diverse, offrendo una visione a 360 gradi del contesto in cui si muovono i clienti del *brand*.

La successiva unione dei due *dataset* mediante la colonna in comune "Customer ID" ha rappresentato un passaggio fondamentale per l'analisi dei dati. Grazie a questa operazione, infatti, è stato possibile integrare le informazioni relative alle transazioni effettuate dai clienti con quelle relative alle loro caratteristiche personali, consentendo di ottenere un quadro ancora più completo e preciso della situazione analizzata.

Di seguito viene fornita un'ampia panoramica delle caratteristiche (*features*) al fine di favorire una comprensione più approfondita dell'analisi condotta successivamente.

Gender	Rappresenta il sesso del consumatore
Macro_Area	Suddivisione geografica più ampia e generalizzata, utilizzata per raggruppare le aree geografiche
Loyalty	Livelli della fedeltà dei consumatori

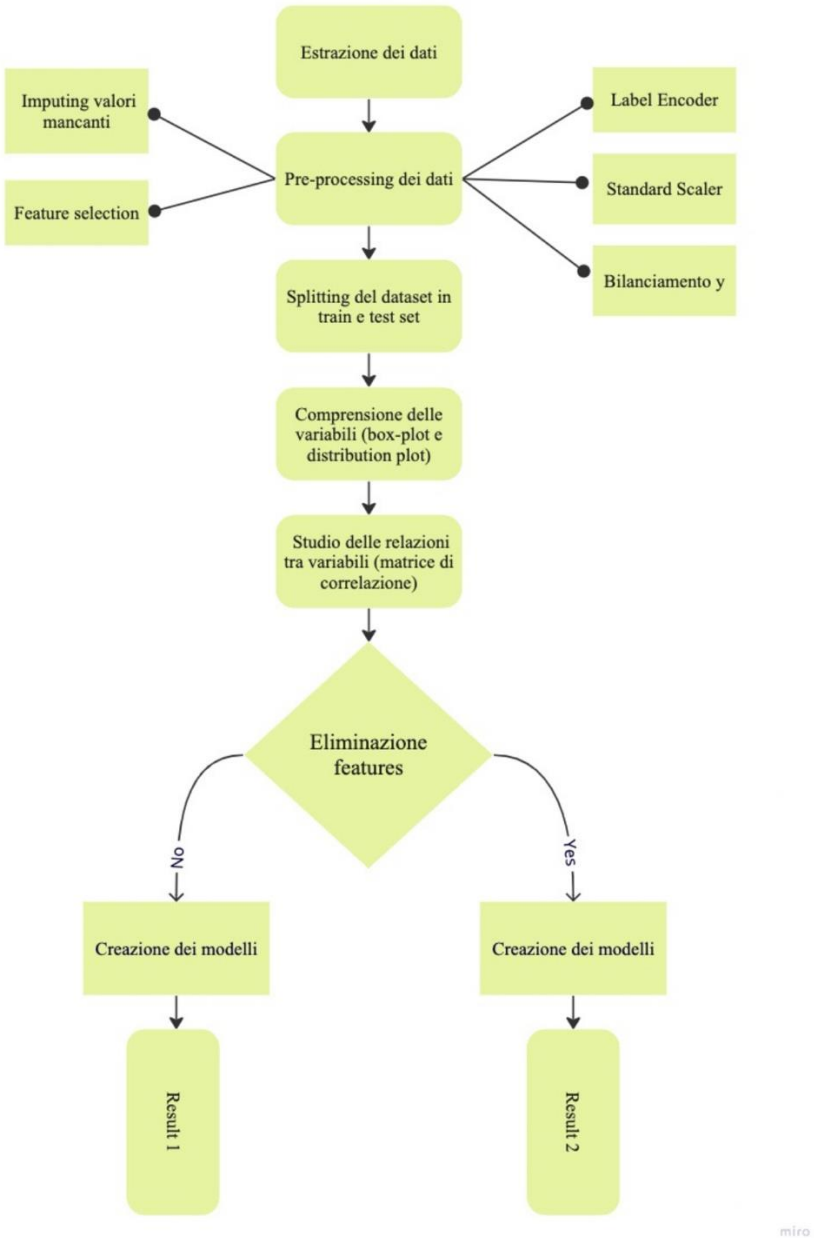
Consent_for_Marketing	L'autorizzazione da parte di un individuo affinché le sue informazioni personali vengano utilizzate per finalità di marketing
Consent_for_Analysis	L'autorizzazione da parte di un individuo affinché le sue informazioni personali vengano utilizzate per scopi di analisi da parte dell'azienda
Consent_for_SMS	Si riferisce all'autorizzazione di un individuo a ricevere messaggi SMS da parte dell'azienda
Consent_for_Newsletter	Si riferisce all'autorizzazione di un individuo a ricevere la newsletter dell'azienda, che è solitamente una forma di comunicazione periodica inviata via e-mail contenente informazioni, notizie o promozioni
Pref_Email	La preferenza di un individuo riguardo all'utilizzo dell'e-mail come mezzo di comunicazione
Pref_Sms	La preferenza di un individuo riguardo all'utilizzo degli SMS come mezzo di comunicazione
Pref_Telephone	La preferenza di un individuo riguardo all'utilizzo del telefono come mezzo di comunicazione
Pref_Mail	La preferenza di un individuo riguardo all'utilizzo della posta cartacea come mezzo di comunicazione
Sales_Habit	La propensione di un individuo ad acquistare prodotti o servizi in saldo o a prezzo pieno
Customer_Habit	Dove i clienti hanno effettuato gli acquisti (Tourist/Resident)
Data Collection Score	Percentuale di dati personali che i consumatori hanno fornito all'azienda
Unsubscribed_Magnews	Indica se i consumatori si sono cancellati dalla Newsletter
Quantity	La quantità di prodotti acquistate
Total Amount (Base)	L'importo totale di una transazione in euro
number_of_purchases	Numero di acquisti effettuati dai consumatori
most_frequent_order_type	Il tipo di prodotto acquistato più frequentemente (accessories, MTM, MTO, services, sportwear)
latest_year	L'ultimo anno in cui i consumatori hanno effettuato un acquisto

Il *dataset* riveste un ruolo centrale, poiché un *dataset* ben strutturato è in grado di generare risultati migliori rispetto ad un *dataset* poco curato. A tal fine, è stato necessario effettuare un lavoro di pre-processamento dei dati.

3.3 Definizione del lavoro svolto

Una volta acquisite le conoscenze teoriche necessarie per il progetto, si procede con la descrizione dettagliata della parte pratica, la quale illustra i metodi utilizzati per raggiungere gli obiettivi prefissati.

Il primo passo, essenziale per garantire il successo del progetto, consiste nell'organizzazione e nella preparazione dei dati in modo da poter eseguire l'analisi in modo efficiente. Il presente paragrafo è interamente dedicato a fornire informazioni sulle azioni svolte e sulle scelte compiute durante questa fase, inoltre, si è ritenuto opportuno iniziare con un *flowchart* per facilitare la comprensione della metodologia di *data analysis* implementata.



Dalle informazioni precedenti si deduce che i *database* in questione presentano una notevole mole di dati. Pertanto, al fine di gestire in modo efficiente il *dataset* Contact Active (d'ora in avanti anche CA), è stato effettuato un processo di limitazione dell'analisi ai soli clienti tramite l'eliminazione di tutti gli elementi non appartenenti alla categoria "Customer", individuata nella *feature* "CRM Type". In un secondo momento, tale variabile verrà rimossa poiché superflua, in quanto tutti i rispondenti rimanenti faranno parte di questa categoria.
Analogamente, per il *dataset* Transactions Details (d'ora in avanti anche TD), si è seguito un processo simile per la *feature* " most_frequent_order_type ", eliminando gli elementi appartenenti alla categoria "Services" perché non sono oggetti reali ma riguardano persone che hanno chiesto un servizio (es. aggiustare un capo) e quindi non sono utili per l'analisi.

Il passo successivo compiuto verso l'analisi dei dati è stato quello di eseguire un processo di valutazione della distribuzione dei valori di ciascuna variabile. Tuttavia, è stato riscontrato che alcune informazioni selezionate non erano disponibili per tutti i soggetti scelti. La mancanza di informazioni è dovuta al processo di archiviazione dei dati; un'archiviazione dei dati incompleta o errata o una mancanza di informazione si riflette nella presenza di valori mancanti nel *dataset*. È stata calcolata la quantità di valori mancanti per ciascuna variabile al fine di determinare la fattibilità di effettuare una procedura di *imputing* per completare il *dataset*.
Per gestire questi valori mancanti, è stata stabilita una soglia di accettabilità del 50% per ogni variabile e nel caso in cui la percentuale di valori mancanti fosse stata superiore, la variabile sarà eliminata dal *dataset*. Successivamente, attraverso il linguaggio di programmazione Python, è stato eseguito un processo di *imputing* dei valori mancanti assegnando il valore più frequente alle *features* categoriche e la media dei valori a quelle numeriche. In aggiunta all'eliminazione delle *features* che superavano la soglia di valori mancanti prestabilita, sono state rimosse anche altre che non erano rilevanti per l'analisi.
L'approccio di analisi dei valori mancanti è stato esteso ad entrambi i *dataset* e i risultati ottenuti indicano che il *dataset* CA risulta composto da 16 feature e 260.601 rispondenti, mentre il *dataset* TD risulta composto da 8 features e 158.593 rispondenti.

Dopo aver ridotto il numero di colonne, è stato possibile applicare la tecnica del Label Encoder per *l'encoding* delle *features* categoriche del *dataset* CA. Il label encoder è una tecnica utilizzata nell'ambito del *machine learning* per la trasformazione di variabili categoriche in variabili numeriche, in modo da renderle utilizzabili da algoritmi di apprendimento automatico. L'utilizzo del label encoder è importante perché molti algoritmi di apprendimento automatico richiedono dati numerici come *input* ed essendo le variabili categoriche descrizioni testuali, per poterle utilizzare è necessario convertirle in numeri. Esso assegna un numero univoco a ciascuna categoria di una variabile categorica ed in questo modo, le variabili categoriche vengono trasformate in numeri che possono essere utilizzati come *input* per un algoritmo di apprendimento automatico.

Il label encoder viene utilizzato in alternativa ad altre tecniche di codifica, come la one-hot encoding, che consiste nell'assegnare un valore binario (0 o 1) a ciascuna categoria di una variabile categorica. Tuttavia, la one-hot encoding può creare un numero elevato di colonne nel *dataset*, soprattutto se la variabile categorica ha molte categorie diverse. Il label encoder, invece, assegna un unico valore intero a ciascuna categoria, riducendo il numero di colonne e semplificando la gestione dei dati.

Passando al dataset TD, la colonna “Total Amount (Base)”, presentava valori in formato categorico a causa della presenza del simbolo dell'euro. Pertanto, è stata necessaria una manipolazione di questa *feature* al fine di trasformarla in un formato numerico.
A causa della presenza di tutte le transazioni, sia *online* che *offline*, effettuate dai consumatori nel *dataset* TD, l'analisi è risultata più complessa e per garantire un'analisi corretta, è stato necessario raggruppare, attraverso l’utilizzo di groupby(), ogni colonna per "Customer" poiché ogni consumatore aveva effettuato molteplici transazioni.
È stata eseguita un'analisi più approfondita del *dataset* TD ed è stata calcolata per ogni cliente la somma della colonna "Quantity" e "total_amount_(base)", inoltre è stato contato il numero di transazioni uniche per ogni cliente utilizzando il metodo nunique(), fornito dalla libreria Pandas, per la *feature* "Transaction", rinominandola successivamente "number_of_purchases" per renderla più intuitiva. Utilizzando il metodo value_counts() è stato calcolato il tipo di ordine più frequente per ogni cliente per la colonna "PBI Item.cnl_ordertype". Inoltre, è stato trovato per la colonna "Year" l'ultimo anno di acquisto per ogni cliente. Successivamente, queste nuove *features* sono state aggiunte al dataset CA utilizzando la funzione merge(), generando così un nuovo *dataset* denominato "ca_updated" composto da 260.601 e 20 *features*.

In seguito, è stata effettuata la definizione delle variabili indipendenti (x) e della variabile dipendente (y), che costituiscono la base per l'analisi dei dati. La variabile x rappresenta l'insieme di tutte le variabili indipendenti (feature o caratteristiche) utilizzate per spiegare o predire il valore della variabile dipendente (y), ed è fondamentale per l'addestramento di un modello di apprendimento automatico efficace. La variabile y, invece, rappresenta la variabile dipendente da modellare o prevedere sulla base delle informazioni contenute nella variabile indipendente x. La corretta definizione di x e y consente un'analisi accurata dei dati e una comprensione chiara del problema, permettendo di individuare le relazioni tra le variabili e di identificare i fattori più significativi per la predizione della variabile y.

A seguire è stata implementata la tecnica di StandardScaler per effettuare la scalatura dei dati, in modo che essi siano distribuiti attorno a una media pari zero e con una deviazione standard pari ad uno. Tale operazione è necessaria per evitare che le variabili con valori molto grandi abbiano un peso maggiore rispetto a quelle con valori più piccoli durante la fase di addestramento del modello. Inoltre, scalare i dati può favorire la convergenza più rapida di alcuni algoritmi di apprendimento automatico. Grazie a questa operazione, che

rendere i dati comparabili e normalizzati in modo da evitare distorsioni causate da differenze di scala tra le variabili, il modello diventa in grado di generalizzare su nuovi dati e di fornire previsioni accurate, migliorando così la sua capacità predittiva.

La suddivisione del *dataset* rappresenta una fase essenziale del processo di sviluppo di un modello di apprendimento automatico. La funzione train_test_split è stata impiegata per separare il *dataset* in due parti disgiunte: il *train set*, utilizzato per l'addestramento del modello, e il *test set*, utilizzato per valutare le prestazioni del modello su nuovi dati.
La dimensione del *test set* è stata scelta con cura, poiché essa deve essere sufficientemente grande da consentire una valutazione accurata del modello, ma anche sufficientemente piccola da evitare che il modello venga addestrato su informazioni del *test set*. Infatti, ciò potrebbe generare una sovrastima delle prestazioni del modello.
Nel caso specifico, è stata scelta una dimensione del *test set* pari al 20% del *dataset* originale (test_size=0.20). Tale scelta rappresenta una pratica comune per ottenere un equilibrio adeguato tra le dimensioni del *train set* e del *test set*. Di conseguenza, l'80% rimanente del *dataset* è stato utilizzato per l'addestramento del modello, garantendo una buona capacità di generalizzazione del modello su nuovi dati.

La variabile *target* "Loyalty" contenuta nel *dataset* CA presentava numerose categorie. Poiché era necessario condurre una classificazione binaria, è stata effettuata una selezione delle categorie da assegnare alla classe "Leale" e quelle da assegnare alla classe "Infedele". Le categorie attribuite alla classe "Infedele" sono state Lost, Inactive, Prospect Store, Sleeper, New, Occasional Reactivated e Occasional Retained, mentre le categorie Loyal Retained, New Loyal e Loyal Reactivated sono state assegnate alla classe "Leale".
Come precedentemente menzionato, attraverso l'utilizzo del LabelEncoder, è stato attribuito il valore 0 alla classe negativa (ovvero l'assenza della caratteristica) e il valore 1 alla classe positiva (ovvero la presenza della caratteristica).
La variabile è stata analizzata tramite l'utilizzo della funzione sns.countplot, la quale consente di contare il numero di osservazioni per categoria e di visualizzarli mediante un grafico a barre (Figura 1). Attraverso questa analisi è stato possibile notare che la *feature* presenta un forte sbilanciamento tra le due categorie.
La distribuzione sbilanciata dei dati riduce l'accuratezza della classificazione, il che rappresenta una grande preoccupazione in molte applicazioni del mondo reale.

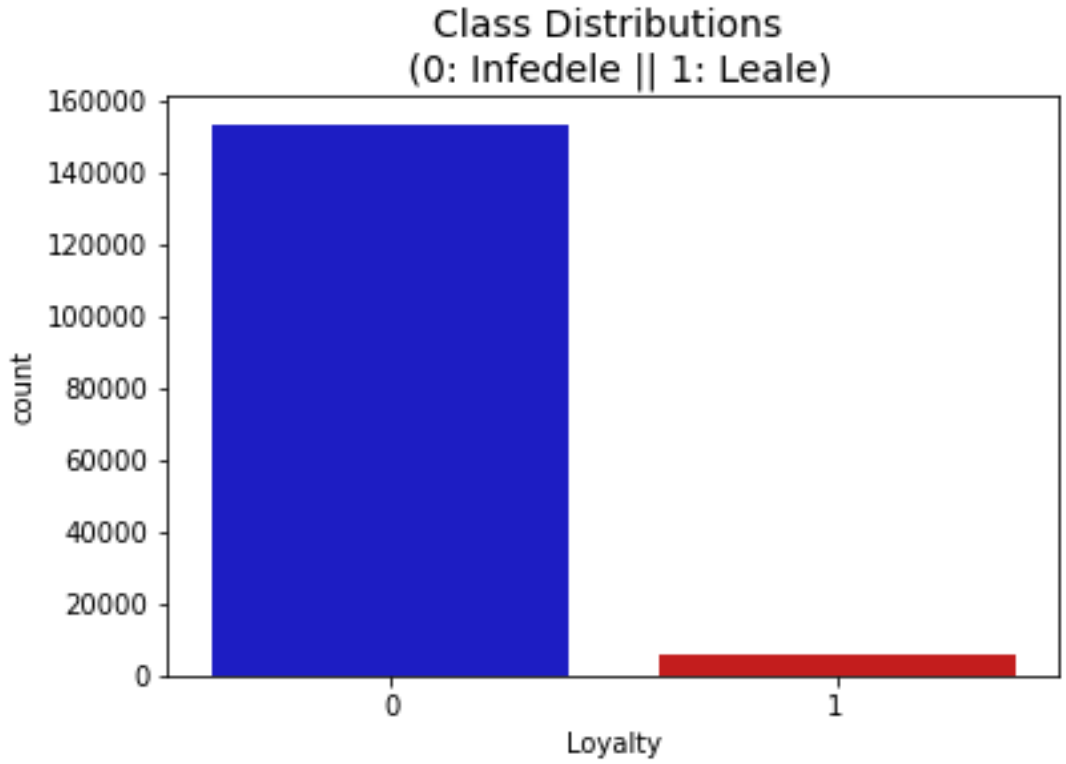


Figura 1: distribuzione della "Loyalty" (y) prima del bilanciamento

Prima di illustrare la soluzione adottata per affrontare il problema di sbilanciamento delle classi, è opportuno precisare che questa procedura è stata eseguita unicamente sulla porzione di *training* del *dataset*.
Bilanciare la variabile *target* solo sulla parte del *training* è una buona pratica per evitare di introdurre informazioni del *test* nel processo di apprendimento del modello.
Come detto in precedenza, il *training set* viene utilizzato per addestrare il modello, mentre il *test set* viene utilizzato per valutarne le prestazioni. Qualora il bilanciamento della variabile *target* fosse applicato sull'intero *dataset*, il modello risulterebbe addestrato su informazioni del *test set*, producendo così una stima inaccurata delle sue prestazioni. È quindi preferibile bilanciare solo il *training set* in quanto ciò permette di ottenere una valutazione più realistica dell'accuratezza del modello in situazioni reali, dove la distribuzione delle classi potrebbe variare rispetto al *training set*. Bilanciare su tutto il *dataset* potrebbe generare un modello che funziona correttamente soltanto su un sottoinsieme e che non sarebbe in grado di generalizzare bene su nuovi dati.

Per ovviare quindi, al problema dello sbilanciamento della y in cui, una classe è rappresentata da un numero significativamente inferiore di osservazioni rispetto all'altra, è stata utilizzata la tecnica SMOTEENN (Synthetic Minority Over-sampling Technique Edited Nearest Neighbors) sviluppata da Batista et al. (2004). Questa tecnica combina SMOTE e Edited Nearest Neighbours (ENN) ed esegue contemporaneamente un *upsampling* e un *downsampling*.

L'obiettivo di SMOTE è generare nuovi campioni sintetici della classe minoritaria, aumentando così il numero di esempi per quella classe. Tuttavia, questa tecnica potrebbe portare alla generazione di campioni che non sono rappresentativi della classe minoritaria, ma che sono estremamente simili ad esempi già esistenti.
L'aggiunta di ENN, invece, serve a rimuovere i campioni indesiderati che potrebbero essere stati generati da SMOTE. In particolare, ENN rimuove i campioni della classe maggioritaria che sono troppo vicini a quelli della classe minoritaria. La combinazione di queste tecniche aiuta a migliorare il bilanciamento delle classi e ridurre *l'overfitting* e quindi a migliorare la capacità del modello di generalizzare su dati non visti in precedenza. In questo modo, si cerca di bilanciare le classi senza alterare la distribuzione originale dei dati ma è possibile che ci sia comunque una leggera differenza tra le due distribuzioni.

Figura 2: distribuzione della "Loyalty" (y) dopo il bilanciamento

Come si può osservare dal grafico a barre (Figura 2), la variabile y presenta una distribuzione bilanciata. Pertanto, la suddivisione della "Loyalty" in classi è equamente distribuita. Successivamente, per l'addestramento degli algoritmi, si è deciso di sostituire le variabili x e y con le corrispondenti versioni bilanciate, ovvero x_res e y_res, in modo da evitare eventuali problemi dovuti a squilibri di classe.

3.4 Analisi statistiche

L'analisi statistica esplorativa dei dati nel contesto del *machine learning* si riferisce alla fase del processo in cui i dati di *input* vengono esplorati e compresi al fine di identificare eventuali problemi, comprendere la distribuzione delle variabili ed individuare eventuali relazioni tra esse. Questa fase è essenziale per garantire la qualità dei dati e la validità dei modelli di *machine learning* che verranno creati e valutati successivamente. Di seguito saranno presentate le descrizioni delle *features* una per una.

Most_frequent_order_type

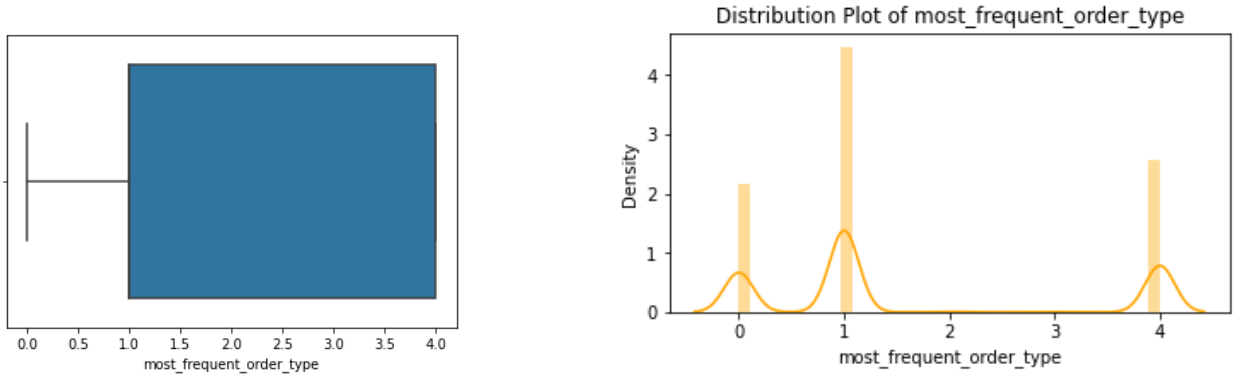


Figura 3: box plot e distribution plot di Most_frequent_order_type

Il grafico rappresenta una distribuzione di dati, i quali sono compresi all'interno di un intervallo che va da 0 a 4. Il *box plot* è composto da una scatola rettangolare centrata sull'intervallo che va da 1 a 4, la linea mediana del rettangolo coincide con il valore del primo quartile e il baffo inferiore, che indica la variazione dei dati al di fuori della mediana, che si estende dal bordo inferiore della scatola fino a 0 mentre non è presente un baffo superiore; infatti, il terzo quartile coincide con il valore massimo. La visualizzazione dei dati permette di osservare che la maggior parte dei valori si concentra nella parte superiore dell'intervallo di riferimento (oltre 1), mentre non sono presenti valori estremamente alti (oltre 4).

Il grafico di distribuzione in questione rappresenta la distribuzione dei valori ed è possibile notare che la maggior parte dei dati è concentrata nella regione di valore 1, mentre si osserva una minore presenza di dati nelle regioni di valore 4 e 0. Inoltre, si rileva una bassa frequenza di dati nelle regioni di valore 2 e 3.

Latest_year

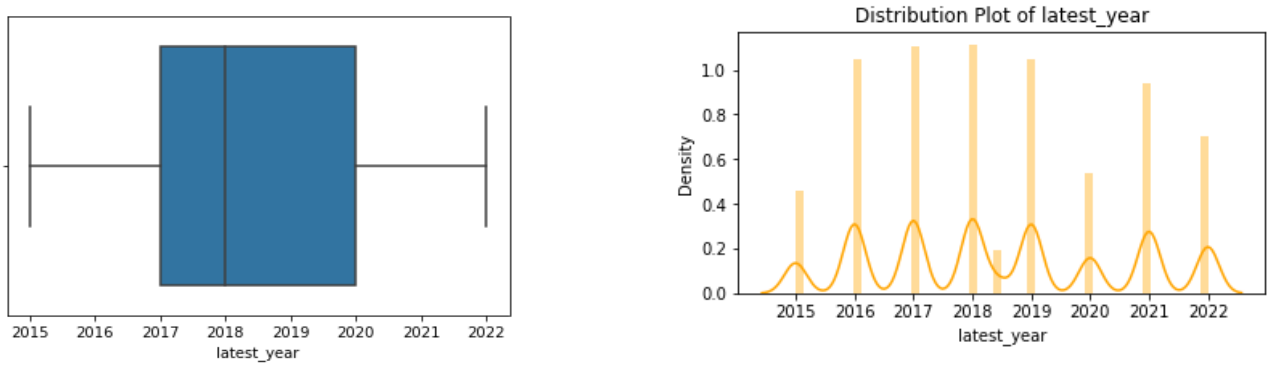


Figura 4: box plot e distribution plot di latest_year

Il *box plot* in questione rappresenta la distribuzione di un insieme di dati che si estende nel periodo dal 2015 al 2022. La scatola centrale del *box plot* si estende dal 2017 al 2020, la linea mediana del rettangolo è posizionata sul 2018, mentre il baffo inferiore parte dal bordo inferiore della scatola centrale e si estende dal 2017 al 2015 mentre il baffo superiore parte dal bordo superiore della scatola centrale e si estende dal 2020 al 2022.

Dalla rappresentazione grafica della distribuzione dei dati è possibile osservare che la distribuzione risulta essere relativamente uniforme in tutte le regioni, ad eccezione di due decrementi localizzati rispettivamente nel 2015 e nel 2020.

Data_Collection_Score

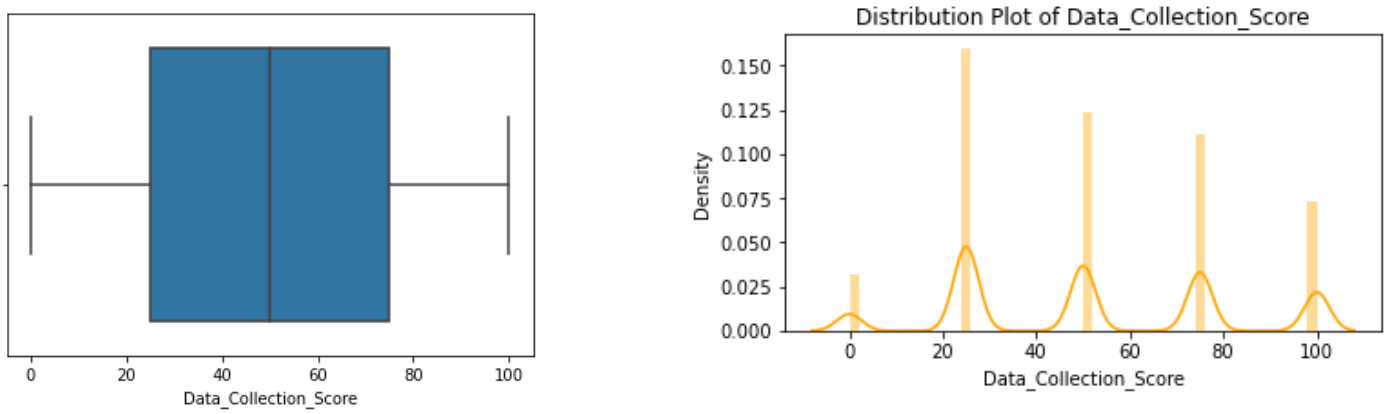


Figura 5: box plot e distribution plot di Data_Collection_Score

Il *box plot* rappresenta la distribuzione di un insieme di dati compresi all'interno di un intervallo che va da 0 a 100. La scatola rettangolare è centrata sull'intervallo che va da 25 a 75, con la mediana che si posiziona al centro del rettangolo sul valore 50. Il baffo inferiore, si estende dal bordo inferiore della scatola fino a 0, mentre il baffo superiore, si estende dal bordo superiore della scatola fino a 100. Tale distribuzione indica che il 25% dei dati si trova al di sotto del valore 25 e il 25% dei dati si trova al di sopra del valore 75.
Dal *distribution plot* si può notare un leggero picco nella regione di valore 20, una distribuzione leggermente inferiore nella regione tra 40 e 60 e una distribuzione uguale tra 60 e 80. Inoltre, sono presenti meno dati nella regione di valore 100 e una bassa densità nella regione di valore 0.

Total_Amount_Base

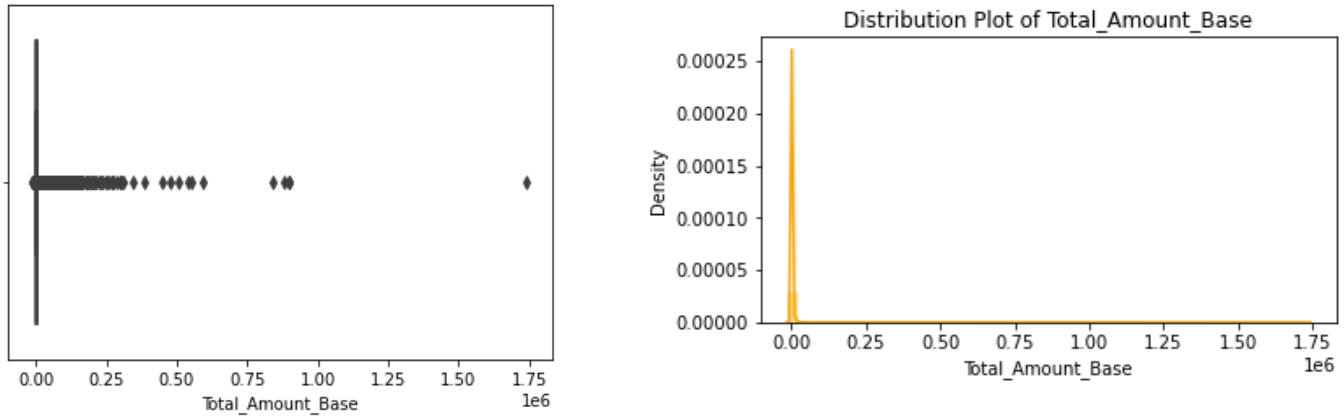


Figura 6: box plot e distribution plot di Total_Amount_Base

Dai dati rappresentati nel *box plot*, si può osservare che è presente un baffo inferiore che corrisponde al valore 0 e una serie di *outliers* che si estendono da 0 a 1.75, con una maggior frequenza di osservazioni intorno allo 0, come evidenziato anche dal *distribution plot*. Non è stata ritenuta opportuna l'eliminazione degli *outliers* in quanto la loro presenza potrebbe fornire informazioni significative sulla distribuzione dei dati. In particolare, l'osservazione che la maggior parte degli acquisti effettuati dai consumatori sono di piccolo taglio è ritenuta rilevante per lo studio in questione e quindi la loro rimozione potrebbe comportare una perdita di tali informazioni.

Quantity

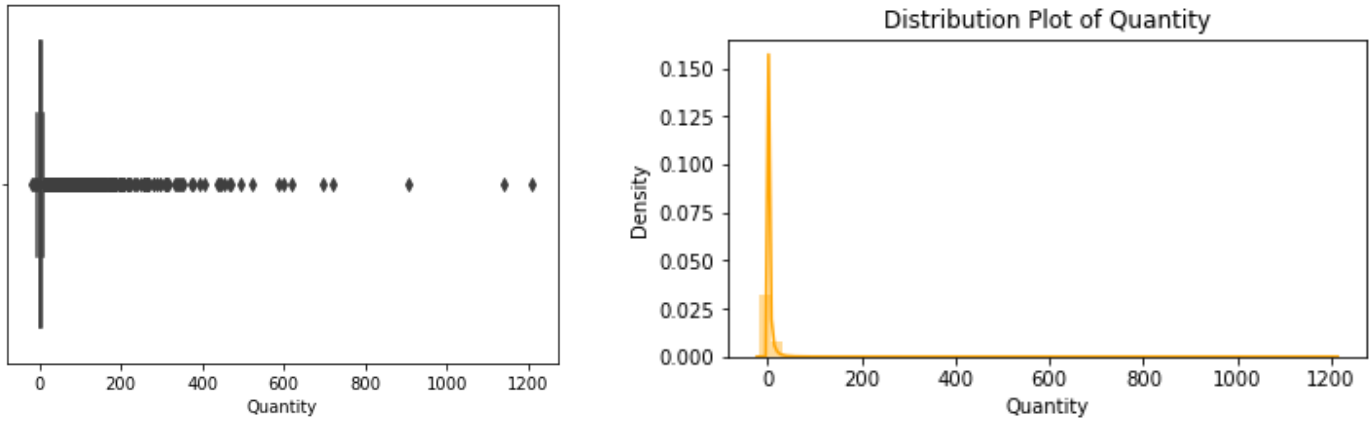


Figura 7: box plot e distribution plot di Quantity

Sia il *box plot* che il *distribution plot* mostrano somiglianze rispetto a quelli presentati precedentemente e alla luce del ragionamento fatto prima si è deciso di non escludere questi dati dalla ricerca, poiché hanno evidenziato che i consumatori hanno effettuato ordini principalmente con piccole quantità di prodotti.

Sales_Habit

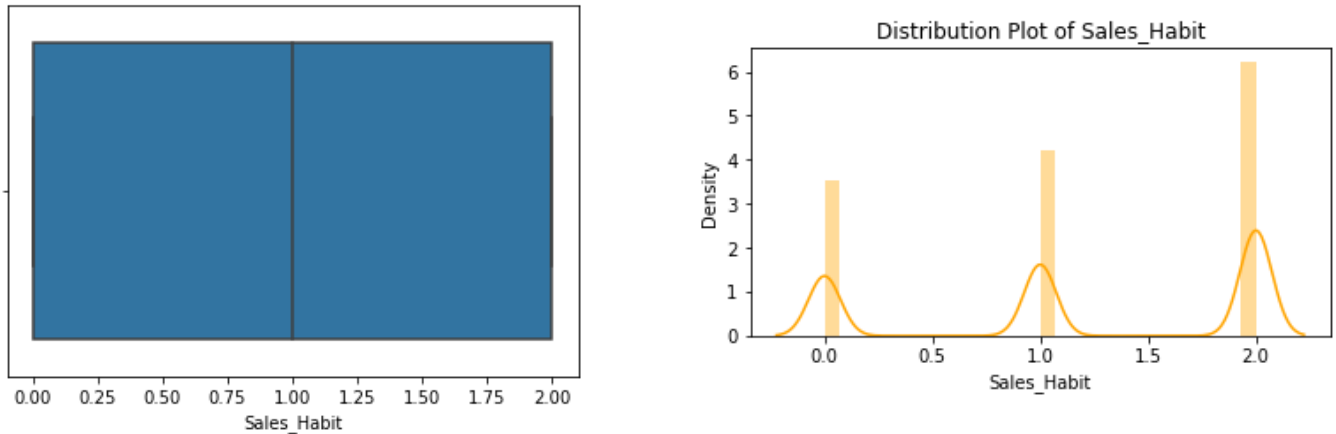


Figura 8: box plot e distribution plot di Sales_Habit

Il *box plot* in questione rappresenta la distribuzione di un insieme di dati compresi all'interno di un intervallo che va da 0 a 2. La scatola del *box plot* copre l'intero *range*, senza la presenza di baffi inferiore e superiore e la mediana è posizionata in corrispondenza del valore 1. La maggior parte dei dati si concentrano nella zona centrale dell'intervallo, senza presentare valori estremi al di fuori dell'intervallo di riferimento.
Il *distribution plot* rappresenta la distribuzione di un insieme di dati, i quali sono maggiormente concentrati nella regione di valore 2 mentre la distribuzione è quasi equamente distribuita nelle regioni di valore 1 e 0.

Customer_Habit

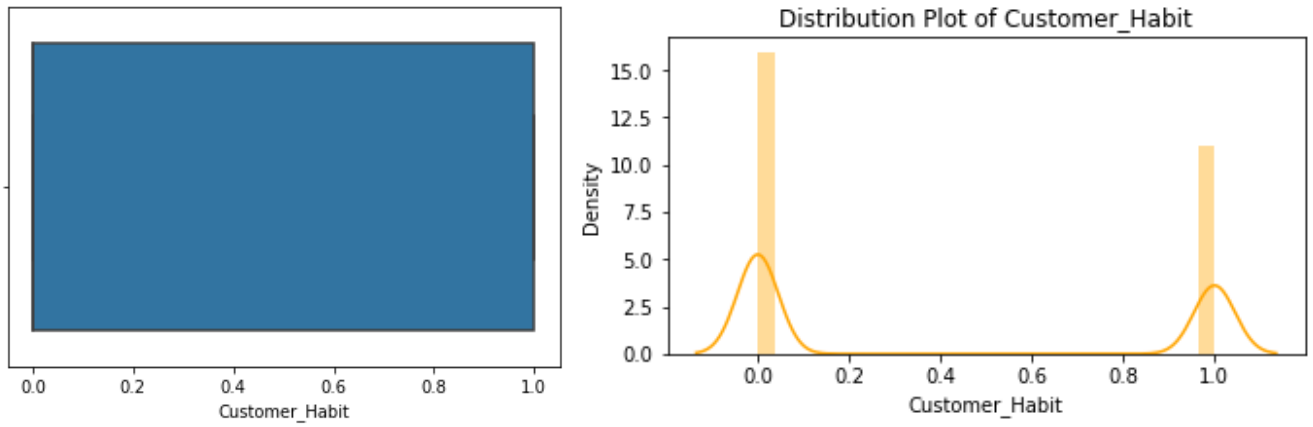


Figura 9: box plot e distribution plot di Customer_Habit

Il *box plot* mostra la distribuzione di dati racchiusi in un intervallo da 0 a 1, con una scatola rettangolare centrata sull'intervallo stesso e senza la presenza di baffi inferiore e superiore mentre, la mediana dei dati si trova sullo 0.
Nel *distribution plot* si può osservare un picco sul valore 0 e una leggera diminuzione sull'1.
Inoltre, per quanto riguarda questa *feature* è stata effettuata un'analisi di modellazione delle classi al fine di evitare la presenza di ridondanze.
Macro_Area

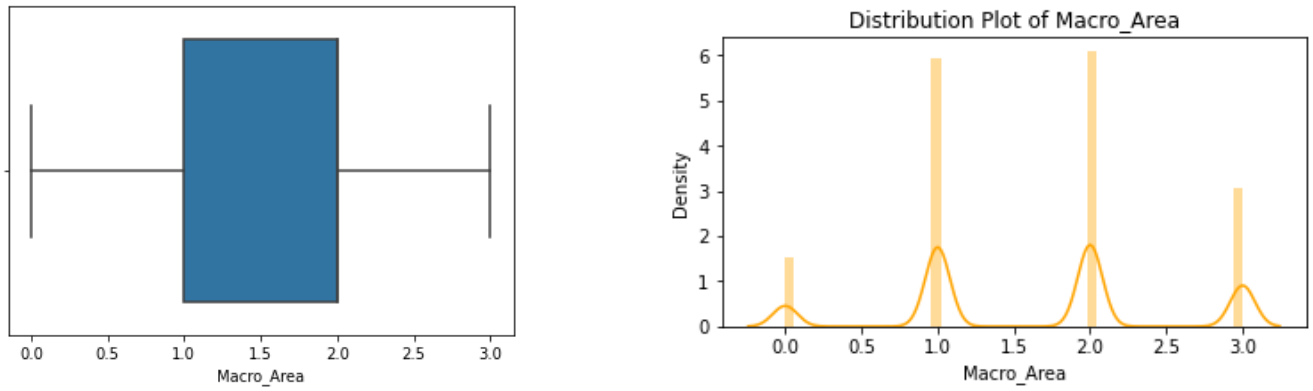


Figura 10: box plot e distribution plot di Macro_Area

Anche in questo caso è stata condotta una modellazione delle classi al fine di evitare la presenza di ridondanze all'interno della seguente *feature*.

Il *box plot* rappresenta la distribuzione di un insieme di dati compresi all'interno di un intervallo che va da 0 a 3. La scatola rettangolare del *box plot* è centrata sull'intervallo che va da 1 a 2, con il baffo inferiore che si estende dal bordo inferiore della scatola fino a 0 e il baffo superiore che si estende dal bordo superiore della scatola fino a 3 con la rispettiva mediana dei dati si colloca sul 2.

Dal *distribution plo*t si evinche che i dati si concentrano principalmente nei valori 1 e 2 con una distribuzione leggermente inferiore sul valore 3 e una presenza molto ridotta di dati sul valore 0.

Number_of_purchases

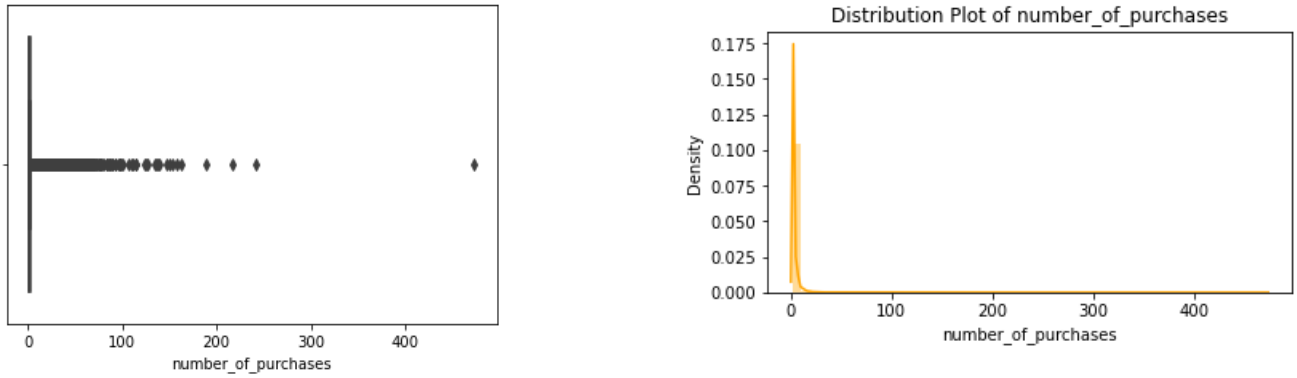


Figura 11: box plot e distribution plot di Number_of_purchases

Come già espresso in precedenza per le altre due *feature*, anche in questo caso si può notare sia dal *box plot* che dal *distribution plot*, la presenza di numerosi *outliers*. Tuttavia, in considerazione del fatto che la rimozione di tali valori potrebbe comportare la perdita di informazioni di rilievo ai fini del presente studio, si è deciso di mantenere tali dati.

Di seguito sono riportate le *features* binarie in cui si è analizzato solo il *distribution plot*, in quanto presentano solamente classi con valore pari a 0 e 1.

Pref_Email, Pref_Sms, Pref_Mail, Pref_Telephone

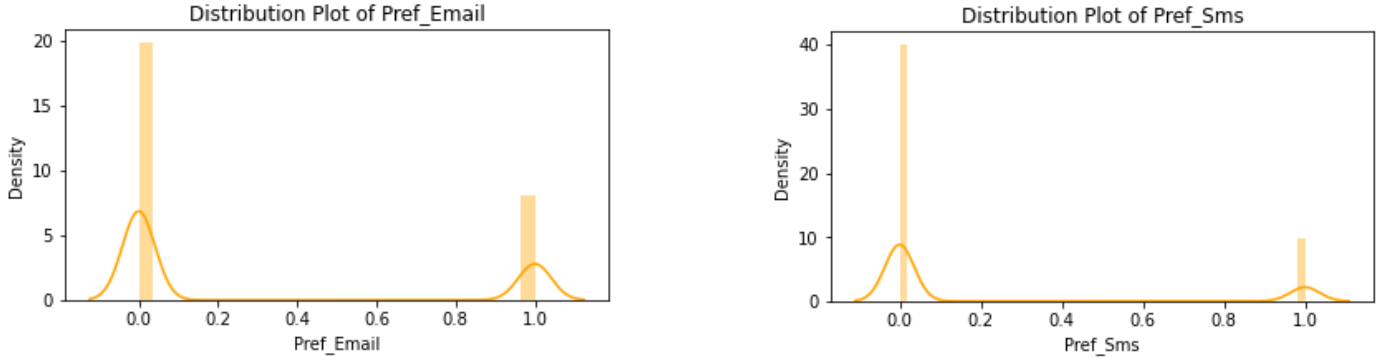


Figura 12: distribution plot di Pref_Email

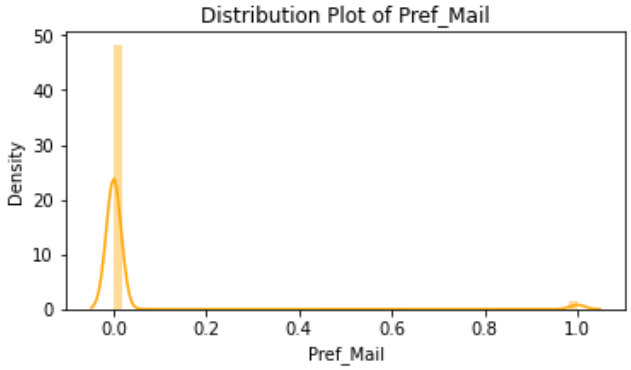


Figura 13: distribution plot di Pref_Sms

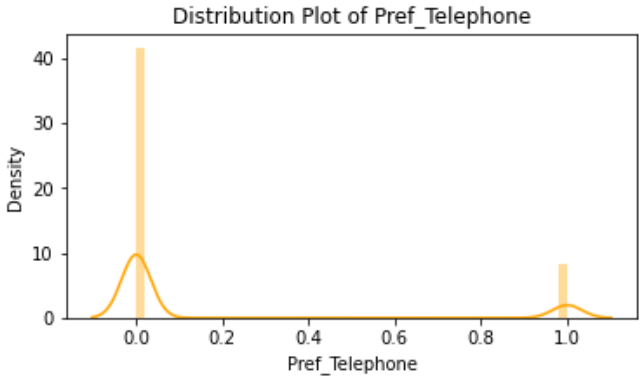
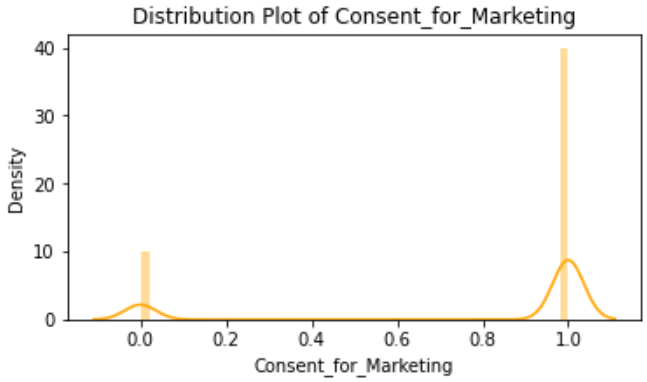


Figura 14: distribution plot di Pref_Mail

Figura 15: distribution plot di Pref_Telephone

Dai risultati dell'analisi delle *features* Pref_Email, Pref_Sms, Pref_Mail e Pref_Telephone si può notare che i rispondenti hanno dato maggiormente risposta negativa (No). Tali informazioni potrebbero essere utilizzate per personalizzare le modalità di contatto in base alle preferenze individuali dei soggetti considerati e migliorando in questo modo l'efficacia delle strategie di comunicazione adottate.

Consent_for_Marketing, Consent_for_Analysis, Consent_for_SMS, Consent_for_Newsletter



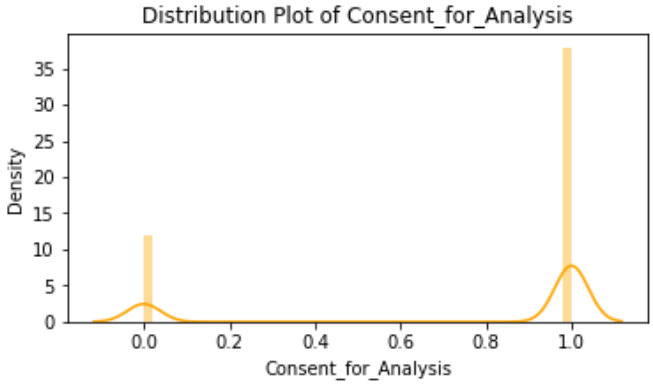
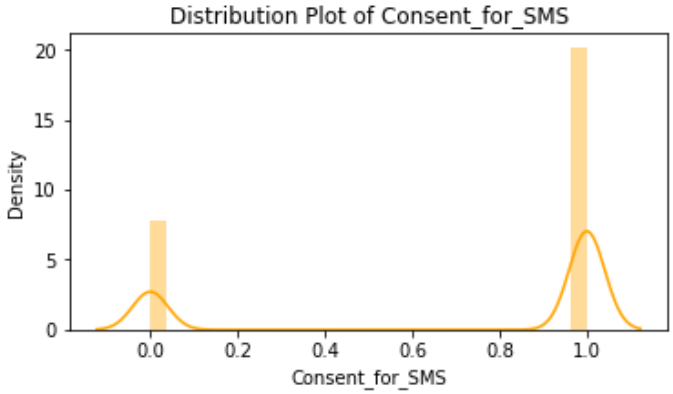


Figura 16: distribution plot di Consent_for_Marketing

Figura 17: distribution plot di Consent_for_Analysis



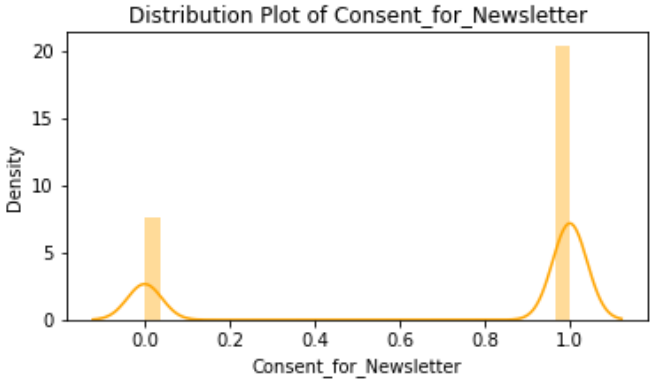


Figura 18: distribution plot di Consent_for_SMS *Figura 19: distribution plot di Consent_for_Newsletter*

La valutazione delle *features* Consent_for_Marketing, Consent_for_Analysis, Consent_for_SMS e Consent_for_Newsletter ha mostrato che la maggioranza dei rispondenti ha indicato una risposta positiva (Yes). Questa informazione potrebbe risultare significativa per comprendere l'atteggiamento che hanno i soggetti rispetto all'utilizzo dei propri dati personali da parte del *brand* preso in esame.
I risultati potrebbero essere utili per la pianificazione di campagne di marketing personalizzate e mirate, basate sulle preferenze dei consumatori. Tuttavia, queste informazioni sottolineano anche l'importanza di garantire un adeguato livello di trasparenza e rispetto della *privacy* nell'uso dei dati personali, al fine di mantenere la loro fiducia e collaborazione nel tempo.

Gender

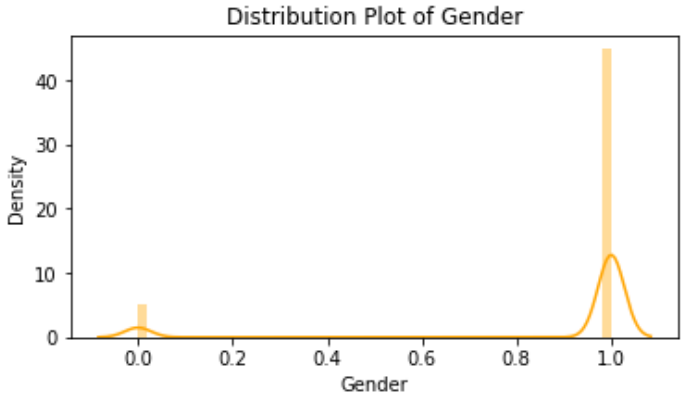


Figura 20: distribution plot di Gender

È possibile notare, dai risultati dell'analisi della *feature* Gender, che la maggioranza dei soggetti considerati si identifica con il genere maschile. Questa informazione può essere di grande rilevanza per il *brand* al fine di compiere scelte strategiche mirate e personalizzate. Potrebbe essere opportuno sviluppare campagne pubblicitarie con messaggi specifici e canali di comunicazione mirati per il pubblico maschile, allo scopo di massimizzare l'impatto delle attività di marketing e quindi cercare di attrarre più clienti maschi. Tuttavia, non è necessario trascurare la parte del pubblico femminile, anche esso infatti, potrebbe rappresentare una parte rilevante dei consumatori e quindi, il *brand* potrebbe considerare l'opzione di sviluppare strategie di marketing anche per il pubblico femminile.

Unsubscribed_Magnews

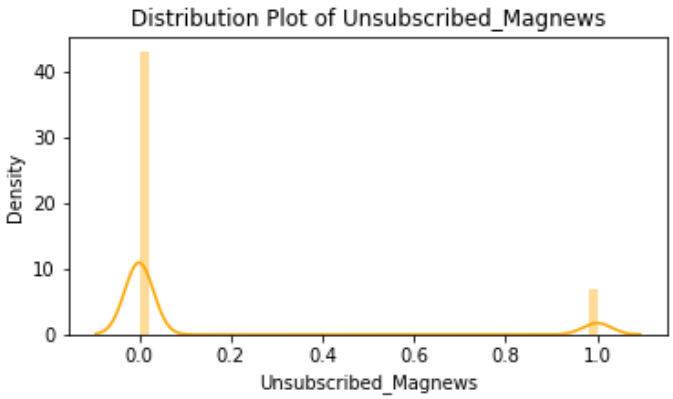


Figura 21: distribution plot di Unsubscribed_Magnews

Dai risultati dell'analisi della *feature* Unsubscribed Magnews si può notare che la maggioranza dei soggetti considerati non si è disiscritta dalla *newsletter* dell'azienda, come indicato dalla maggioranza di risposte "No". Questa informazione può essere utile per il *brand* al fine di comprendere meglio il livello di interesse dei propri clienti verso le attività di comunicazione dell'azienda. Inoltre, potrebbe essere interessante indagare ulteriormente sui motivi alla base della decisione di disiscriversi dalla *newsletter*, nel caso in cui ci siano risposte affermative. Questi dati potrebbero essere utilizzati per migliorare la qualità della *newsletter* e renderla più interessante e pertinente per il pubblico di riferimento, al fine di ridurre il tasso di disiscrizioni e migliorare l'efficacia delle attività di comunicazione dell'azienda.

L'obiettivo delle analisi statistiche effettuate (*box-plot* e *distribution plot*) è stato quello di visualizzare i dati presenti all'interno delle *features* e la loro distribuzione. Era importante anche capire se fosse stato il caso di eliminare eventuali *outliers;* tuttavia, come precedentemente indicato nei tre casi specifici in cui sono stati rilevati gli *outliers*, è stato deciso di non eliminarli in quanto ciò avrebbe comportato la perdita di informazioni significative per l'analisi.

In seguito, si è proceduto all'analisi della matrice di correlazione al fine di individuare le relazioni tra le variabili presenti nel *dataset*. La matrice di correlazione rappresenta uno strumento utile per visualizzare graficamente le associazioni tra le variabili del *dataset* e facilitare l'identificazione di eventuali relazioni significative.

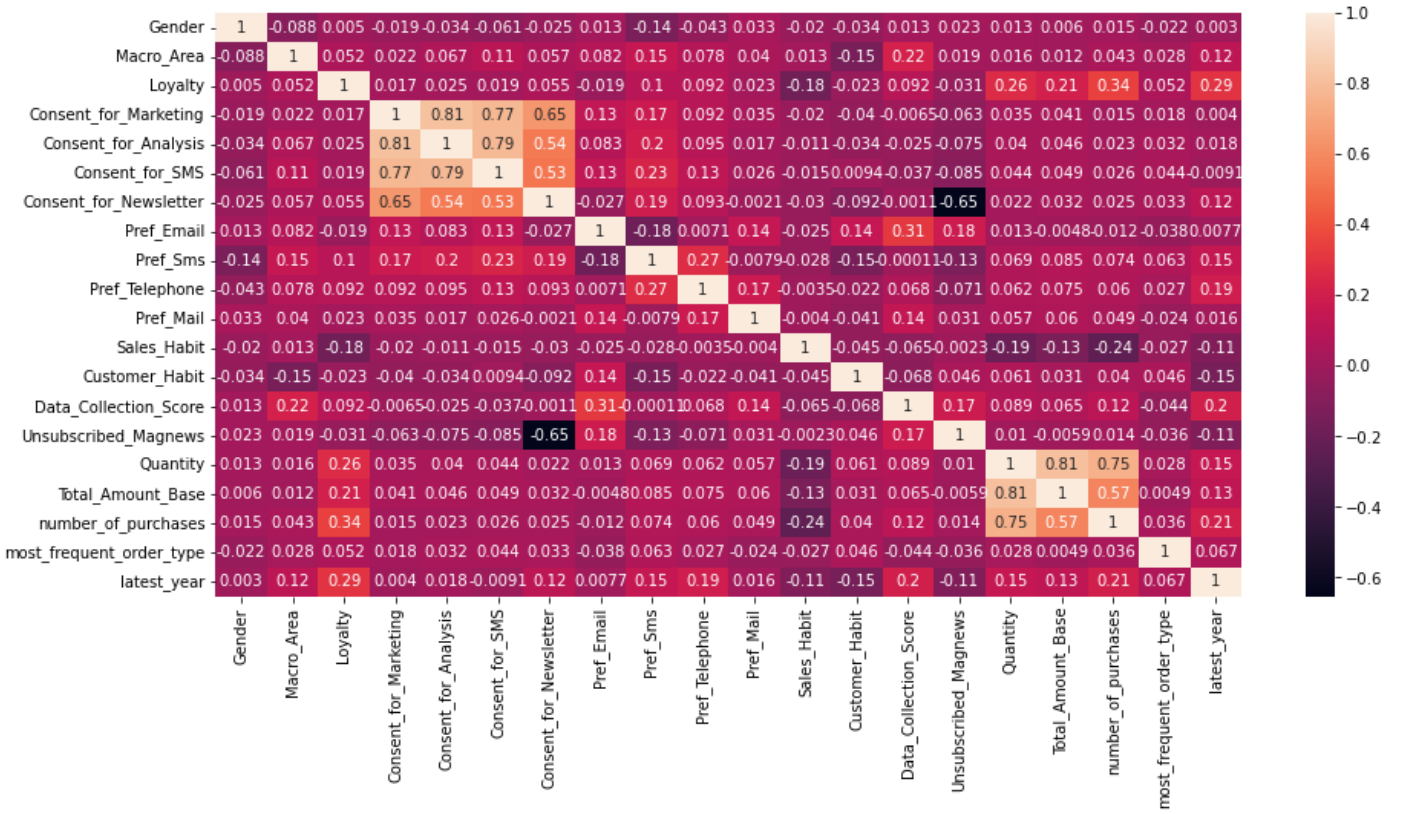


Figura 22: Matrice di correlazione dataset ca_updated

La variabile “Loyalty” presenta una correlazione positiva con la variabile "number_of_purchases" (0.341750) e una correlazione moderata positiva con "latest_year" (0.292818) e "Total_Amount_Base" (0.206928). Questo suggerisce che i clienti che hanno fatto un maggior numero di acquisti e quelli che hanno speso di più nel tempo sono anche quelli più fedeli al *brand*.

Ci sono alcune variabili che mostrano una debole correlazione positiva con la “Loyalty”, come "Pref_Sms" (0.100980), "Pref_Telephone" (0.092458) e "Data_Collection_Score" (0.091763), suggerendo che i clienti che preferiscono essere contattati tramite SMS o telefono e quelli che hanno una maggiore partecipazione alle attività di raccolta dati potrebbero essere leggermente più fedeli al *brand*.

D'altra parte, la variabile "Sales_Habit" mostra una correlazione moderata negativa con la “Loyalty” (-0.176023), suggerendo che i clienti hanno una maggiore tendenza ad acquistare prodotti in sconto potrebbero essere meno fedeli all'azienda. Infine, le altre variabili presentano correlazioni molto deboli con la y.

La forte correlazione tra le *features* "Consent_for_Marketing", "Consent_for_Analysis", "Consent_for_SMS" e "Consent_for_Newsletter" indicata dalla matrice di correlazione suggerisce che queste potrebbero essere dipendenti l'una dall'altra e avere un effetto simile sulla variabile *target*. Potrebbe essere il caso che un cliente che ha dato il proprio consenso per le analisi di marketing, abbia anche dato il consenso per le *newsletter* o per le comunicazioni tramite SMS e ciò potrebbe anche suggerire che la presenza di una di queste *features* potrebbe essere un buon indicatore per prevedere la presenza delle altre.

Inoltre, la forte correlazione anche tra le *features* "Number_of_purchases", "Total_Amount_Base" e "Quantity" suggerisce che queste potrebbero rappresentare un unico concetto o costrutto, come ad esempio l'attitudine all'acquisto o il livello di coinvolgimento con l'azienda.

Tuttavia, dal punto di vista della costruzione del modello, la forte correlazione tra queste *features* potrebbe portare a problemi di multicollinearità, cioè la presenza di una forte relazione lineare tra le caratteristiche (x), che potrebbe rendere difficile l'individuazione dell'effetto indipendente di ciascuna di esse sulla variabile *target* (y).

Lo scopo della seguente matrice di correlazione è stato quello di studiare e capire la relazione tra la variabile dipendente y e le variabili indipendenti x, nonché di identificare eventuali *features* ridondanti, infatti, grazie all'analisi della seguente matrice è stato possibile eliminare le seguenti caratteristiche: "Consent_for_Analysis", "Consent_for_SMS", "Consent_for_Newsletter", "Total_Amount_Base", "Quantity" al fine di valutare un possibile miglioramento delle *performance* degli algoritmi.

Nel capitolo successivo saranno presentati i risultati dei classificatori sia utilizzando tutte le caratteristiche, sia con l'eliminazione di quelle precedentemente elencate.

3.5 Algoritmi

In accordo con quanto riportato nella revisione della letteratura riguardante la parte degli algoritmi, è stato deciso di implementare nello studio quelli che hanno dimostrato le prestazioni più elevate, al fine di valutare se, in questo contesto, avessero lo stesso impatto. L'unica differenza è che l'algoritmo C4.5 non trova implementazione all'interno della libreria scikit-learn, ma al suo posto vi è l'algoritmo CART, che presenta notevoli somiglianze con il primo.
In particolare, i modelli di classificazione adottati per lo studio, a cui dedicheremo particolare attenzione, sono il Random Forest, Logistic Regression, Decision Tree, Naïve Bayes, K-neighbors e XGBoost Classifier.

Random Forest

Il classificatore Random Forest è un algoritmo di apprendimento automatico che genera un insieme di alberi decisionali mediante la creazione di una foresta di alberi. Ogni albero decisionale nell'insieme viene costruito su un sottoinsieme di dati di addestramento selezionati casualmente.
La previsione per un campione di *input* viene generata combinando le previsioni di tutti gli alberi decisionali mediante il voto maggioritario. Il risultato è un classificatore ad alta accuratezza e generalizzazione, che riduce la varianza del modello e migliora le prestazioni rispetto all'utilizzo di un singolo albero decisionale.
I parametri utilizzati dal RandomForestClassifier sono:

- n_estimators=20: specifica il numero di alberi decisionali (o estimatori) che verranno creati nella foresta casuale. In questo caso, verranno creati 20 alberi di decisione.
- random_state=1: determina il seme casuale che viene utilizzato per riprodurre gli stessi risultati ogni volta che il codice viene eseguito.
- class_weight="balanced": un parametro opzionale che specifica il peso da assegnare alle classi durante l'addestramento del modello. La stringa "balanced" indica che il peso delle classi verrà assegnato in modo tale da bilanciare il numero di campioni in ogni classe.

Logistic Regression

L'algoritmo di regressione logistica è una tecnica di classificazione che stima la probabilità di appartenenza a una determinata classe, sulla base di un insieme di variabili indipendenti, utilizzando una funzione logistica che misura la relazione tra la variabile dipendente categorica e una o più variabili indipendenti.
In particolare, la regressione logistica stima la probabilità di accadimento di un evento binario (per esempio, "sì"/"no" o "vero"/"falso"), valutando la relazione tra la variabile dipendente e le variabili indipendenti. L'algoritmo utilizza la funzione logistica per convertire la somma pesata delle variabili indipendenti in una probabilità, che viene utilizzata per la classificazione.

Decision Tree

Il Decision Tree è un algoritmo di apprendimento supervisionato che può essere utilizzato per la classificazione o la regressione, ma è particolarmente adatto alla classificazione. Si tratta di un classificatore strutturato a forma di albero, in cui i nodi interni rappresentano le caratteristiche dei dati, i rami rappresentano le regole decisionali e ogni nodo foglia rappresenta una categoria di classificazione o un valore di regressione. I nodi decisionali sono utilizzati per prendere decisioni e possono avere più rami, mentre i nodi foglia sono *l'output* di tali decisioni e non hanno ulteriori rami. Per la costruzione dell'albero, si utilizza l'algoritmo CART (Classification and Regression Tree algorithm), che si occupa della selezione dei nodi decisionali e della definizione delle regole di suddivisione dell'albero. Il Decision Tree si basa su una serie di domande che, a seconda della risposta (sì o no), dividono l'albero in sottoalberi sempre più specifici.

Il particolare, il modello di classificazione Decision Tree implementato nello studio applica:

- random_state=1 è un parametro opzionale utilizzato per riprodurre i risultati in modo deterministico. Impostando questo parametro su un valore fisso, si garantisce che il modello venga addestrato nello stesso modo ogni volta che il codice viene eseguito. In questo caso, il valore 1 è stato scelto casualmente come seme casuale per la riproducibilità dei risultati.
- p_grid = {"max_depth": [2, 3, 4, 5]} definisce una griglia di parametri (p_grid) che specifica i valori che il parametro max_depth del classificatore decision tree deve assumere durante la ricerca dei migliori parametri tramite GridSearchCV. In particolare, la griglia di parametri specifica che il

parametro max_depth può assumere i valori [2, 3, 4, 5]. Durante la ricerca dei migliori parametri, GridSearchCV testerà il classificatore decision tree con ogni valore possibile di max_depth all'interno della griglia di parametri specificata e restituirà la combinazione di parametri che produce la migliore performance sul dataset di validazione.

- inner_cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=1) definisce una validazione incrociata stratificata con 5 fold. La validazione incrociata è una tecnica utilizzata per valutare le prestazioni di un modello di machine learning. In particolare, la validazione incrociata suddivide il set di dati in k fold (in questo caso, k = 5), utilizzando uno dei fold come set di test e gli altri k-1 fold come set di addestramento. Questa operazione viene ripetuta k volte, in modo che ogni fold sia utilizzato come set di test una volta. La validazione incrociata stratificata garantisce che la distribuzione delle classi all'interno dei fold sia simile alla distribuzione delle classi nell'intero set di dati. Ciò è particolarmente utile quando si lavora con set di dati sbilanciati, in cui una classe è rappresentata da un numero significativamente inferiore di campioni rispetto alle altre classi. Il parametro shuffle=True indica che i dati verranno mischiati prima della suddivisione in fold, mentre il parametro random_state=1 garantisce la riproducibilità dei risultati.
- clf = GridSearchCV(estimator=dt, param_grid=p_grid, cv=inner_cv, verbose=0) crea un oggetto GridSearchCV che viene utilizzato per eseguire una ricerca a griglia per trovare i migliori iperparametri per il modello di classificazione basato su albero decisionale, utilizzando una cross-validation stratificata (inner_cv) con 5 fold. Il parametro param_grid specifica un dizionario di iperparametri (max_depth) da testare durante la ricerca a griglia. La GridSearchCV itera su tutte le possibili combinazioni di iperparametri specificati in param_grid, addestra un modello con ciascuna combinazione e valuta le prestazioni del modello con la cross-validation. Il parametro verbose è impostato su 0, il che significa che non verrà fornita alcuna stampa di output durante l'addestramento del modello.

Naïve Bayes

L'algoritmo classificatore Naïve Bayes si basa sul teorema di Bayes ed è ampiamente utilizzato nell'apprendimento automatico. Si tratta di una famiglia di classificatori statistici che si caratterizzano per l'adozione di ipotesi semplificate, tanto che vengono definiti "naive". In particolare, tali ipotesi considerano le varie caratteristiche (*features*) del modello come tra loro indipendenti.

In particolare, nello studio è stata importata la classe GaussianNB dal modulo naive_bayes della libreria sklearn in Python. La classe GaussianNB implementa l'algoritmo Naïve Bayes con l'assunzione di distribuzione normale dei dati e viene utilizzata per risolvere problemi di classificazione supervisionata.

K-neighbors

L'algoritmo dei k-nearest neighbors (K-NN) è un modello di classificazione di apprendimento supervisionato non parametrico che utilizza la vicinanza per effettuare predizioni sul raggruppamento di un singolo punto dati. Il modello è fondato sull'assunzione che oggetti simili si trovino vicini l'uno all'altro. Il K-NN cerca i k punti più vicini al punto di dati di interesse e li utilizza per determinare la classe o il valore di previsione del punto in questione.

XGBoost Classifier

XGBoost è un robusto algoritmo di apprendimento automatico che può aiutare a comprendere i dati e a prendere decisioni migliori ed è un'implementazione degli alberi decisionali a gradiente.
XGBoost è stato progettato per garantire velocità, facilità d'uso e prestazioni su grandi insiemi di dati.

Di seguito sono riportati gli iperparametri utilizzati nel processo di addestramento:

- base_score=0.5: Il punteggio base per il calcolo dei punteggi predetti è impostato su 0.5.
- booster='gbtree': Viene utilizzato il booster "gbtree", che indica che gli alberi di decisione sono utilizzati come booster.
- colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1: Non viene effettuata alcuna sottocampionatura delle colonne durante la costruzione degli alberi. Tutte le colonne sono considerate in ogni livello, nodo e albero.
- eval_metric='mlogloss': La metrica utilizzata per valutare il modello durante l'addestramento è la perdita logaritmica multinomiale.
- gamma=0: Non viene impostata alcuna riduzione minima richiesta della funzione di perdita per ulteriori partizionamenti degli alberi.
- gpu_id=-1: L'addestramento del modello viene eseguito sulla CPU, poiché l'ID della GPU è impostato su -1.
- importance_type='gain': Viene calcolata l'importanza delle feature basata sul guadagno di informazione.
- interaction_constraints='': Non sono imposti vincoli di interazione tra le feature.
- learning_rate=0.3: Il tasso di apprendimento del modello è impostato su 0.3. Questo indica la velocità con cui il modello si adatta ai dati durante l'addestramento.
- max_delta_step=0: Non viene impostato alcun limite massimo per il passo di aggiornamento degli alberi.
- max_depth=6: La profondità massima degli alberi è impostata su 6. Gli alberi non potranno superare questa profondità.
- min_child_weight=1: Viene richiesto un peso minimo di 1 per creare ulteriori partizioni di un nodo dell'albero durante la costruzione.

- missing=nan: Non viene specificato alcun valore specifico per i dati mancanti. Potrebbe essere necessario impostarlo in base al contesto dei dati utilizzati.
- monotone_constraints='()': Non sono imposti vincoli monotoni sulle feature.
- n_estimators=100: Vengono creati 100 alberi nell'ensemble.
- n_jobs=16: L'addestramento del modello viene eseguito utilizzando 16 thread.
- num_parallel_tree=1: Viene utilizzato un solo albero parallelo.
- random_state=0: Il seme per la generazione dei numeri casuali è impostato su 0, garantendo la riproducibilità dei risultati.
- reg_alpha=0, reg_lambda=1: Non viene applicata alcuna regolarizzazione L1 (norma L1) o L2 (norma L2) durante l'addestramento.
- scale_pos_weight=None: Non viene specificato alcun peso per le istanze positive nei dati di addestramento.
- subsample=1: Tutti i campioni vengono utilizzati durante la costruzione di ciascun albero, senza sottocampionatura.
- tree_method='exact': Gli alberi vengono costruiti utilizzando il metodo "exact", che esegue una ricerca esatta per trovare le partizioni migliori.
- validate_parameters=1: I parametri passati al modello vengono convalidati.
- verbosity=None: Non vengono forniti messaggi di output durante l'addestramento.

4. Risultati e Discussioni

Dopo aver descritto nel dettaglio la metodologia utilizzata, è giunto il momento di passare all'analisi dei dati e alla loro interpretazione. Questo capitolo è dunque dedicato alla presentazione dei risultati emersi dalla ricerca e alla discussione dei principali risultati.

Come precedentemente enunciato, verrà effettuato un confronto tra i risultati ottenuti tramite l'utilizzo del *dataframe* contenente tutte le 20 *features* e quello ottenuto tramite l'eliminazione delle *features* (x) altamente correlate tra loro.

4.1 Descrizioni dei risultati

Random Forest

Le metriche di valutazione prese in considerazione sono *Precision, Recall, F1-score* e Accuracy.

La *Precision* è la proporzione di veri positivi rispetto ai veri positivi e falsi positivi. In questo caso, il valore di precisione del modello è di 0.97, il che significa che il 97% delle previsioni positive del modello sono corrette. La *Recall* è la proporzione di veri positivi rispetto ai veri positivi e falsi negativi. In questo caso, il valore di *recall* del modello è di 0.95, il che significa che il 95% dei veri positivi sono stati correttamente identificati dal modello. L'*F1-score* è una media armonica di *Precision* e *Recall* ed è utilizzata per valutare il bilanciamento tra queste due metriche. In questo caso, l'*F1-score* del modello è di 0.96, il che indica una buona combinazione tra *Precision* e *Recall*.

L'*Accuracy* è la proporzione di istanze correttamente classificate rispetto al totale delle istanze e in questo caso è di 0.95, il che significa che il 95% delle istanze del *dataset* di *test* sono state classificate correttamente dal modello.

Per quanto concerne il *dataframe* con un numero inferiore di features, si può constatare che i risultati delle metriche non hanno riscontrato alcuna variazione.

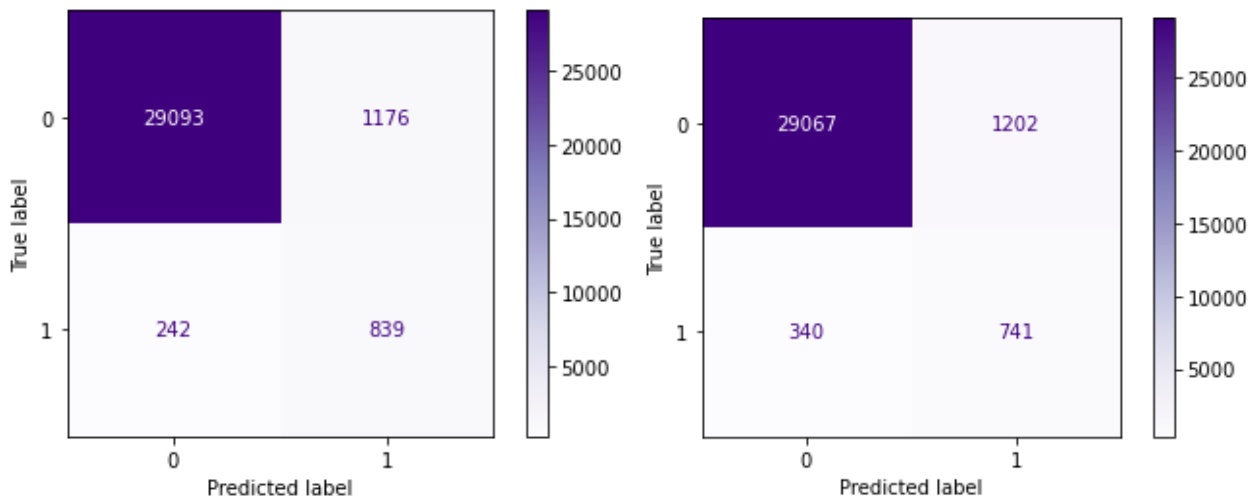


Figura 23: Confronto Confusion Matrix Random Forest (sinistra: dataframe completo, destra: dataframe ridotto)

La prima matrice di confusione mostra un alto numero di veri negativi (VN), il che significa che il modello ha correttamente classificato la maggior parte delle istanze negative. Tuttavia, il numero di falsi positivi (FP) e falsi negativi (FN) è relativamente alto rispetto al numero di veri positivi (VP). Ciò suggerisce che il modello potrebbe avere problemi a distinguere tra le istanze positive e negative.
La seconda matrice di confusione, basata su un sottoinsieme ridotto di *features*, mostra un numero ancora più alto di falsi negativi (FN) rispetto alla prima matrice. Ciò indica che la riduzione delle *features* ha influenzato negativamente le prestazioni del modello, poiché il modello sta perdendo informazioni importanti per la classificazione corretta. Questo modello, inoltre, ha prodotto solo 741 veri positivi, il che potrebbe suggerire che la selezione di *features* utilizzata non è stata in grado di cogliere in modo efficace la complessità della relazione tra i dati e la *customer loyalty*.
In generale, un modello con un numero maggiore di veri positivi e veri negativi e un numero inferiore di falsi positivi e falsi negativi è considerato migliore.

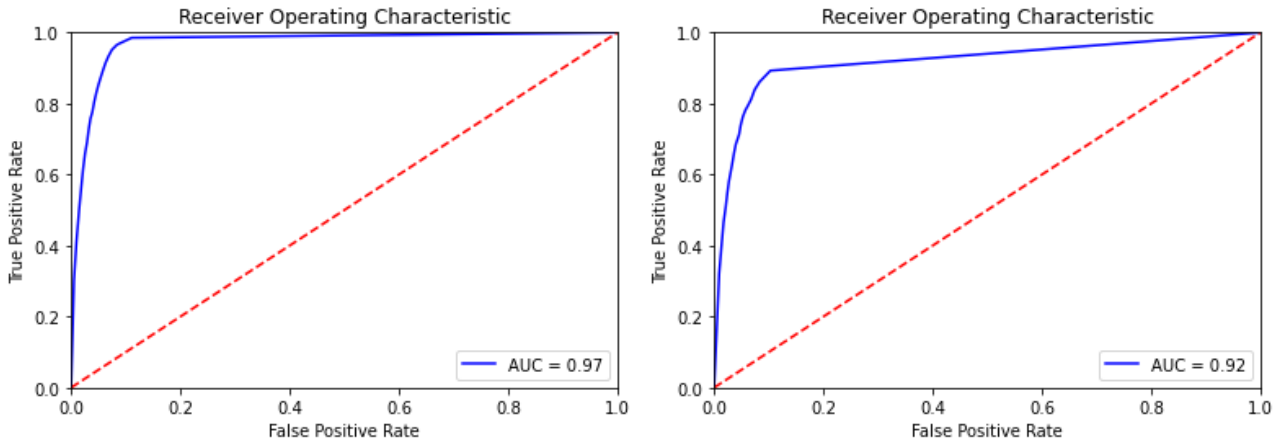


Figura 24: Confronto Curva ROC - AUC Random Forest (sinistra: dataframe completo, destra: dataframe ridotto)

Confrontando le due curve ROC-AUC si evince che la prima curva (con tutte le *features*) ha un'area sottesa della curva ROC pari a 0.97, mentre la seconda curva (con la riduzione delle *features*) ha un'area sottesa della curva ROC pari a 0.92. Ciò suggerisce che il modello con tutte le *features* ha una maggiore capacità di discriminazione rispetto al modello con la riduzione.

Logistic Regression

Il modello con il *dataframe* completo ha ottenuto un'*accuracy* dello 0.88, una *precision* dello 0.97, una *recall* dello 0.88 e un *f1-score* del 0.91. Il modello con il *dataframe* ridotto ha invece ottenuto un'*accuracy* pari 0.87, una *precision* dello 0.97, una *recall* dello 0.87 e un *f1-score* dello 0. 91.
Dai risultati delle metriche di valutazione, si può notare che il modello con il *dataframe* completo ha una *recall* leggermente superiore rispetto al modello ridotto, indicando che è in grado di identificare un maggior numero

di casi positivi. Tuttavia, la differenza tra le due *recall* è piuttosto piccola e le *precision* e l'*f1-score* sono simili tra i due modelli.

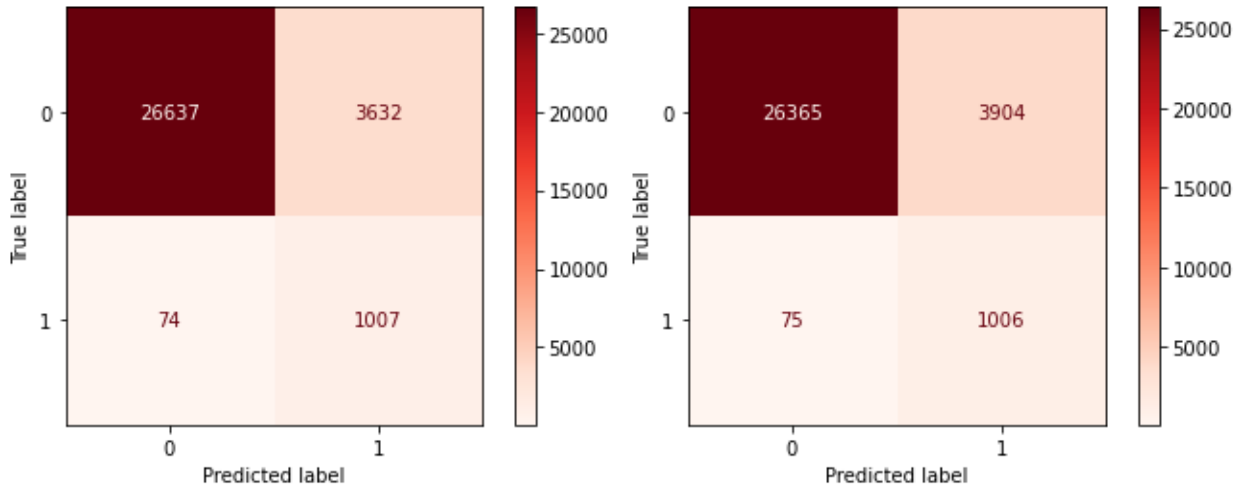


Figura 25: Confronto Confusion Matrix Logistic Regression (sinistra: dataframe completo, destra: dataframe ridotto)

Si può notare che entrambi i modelli hanno ottenuto un numero alto di veri negativi, indicando una buona capacità di classificare correttamente i casi negativi. Tuttavia, i modelli hanno un numero significativo di falsi positivi, il che significa che sono stati classificati erroneamente come positivi un certo numero di casi negativi. Il modello con tutte le *features* ha ottenuto un numero leggermente inferiore di falsi positivi rispetto al modello con il *dataframe* ridotto, ma ha anche un numero leggermente inferiore di falsi negativi, il che significa che è stato più preciso nel classificare i casi positivi. In generale, il confronto tra le *confusion matrix* dei due modelli indica che entrambi hanno una buona capacità di classificazione.

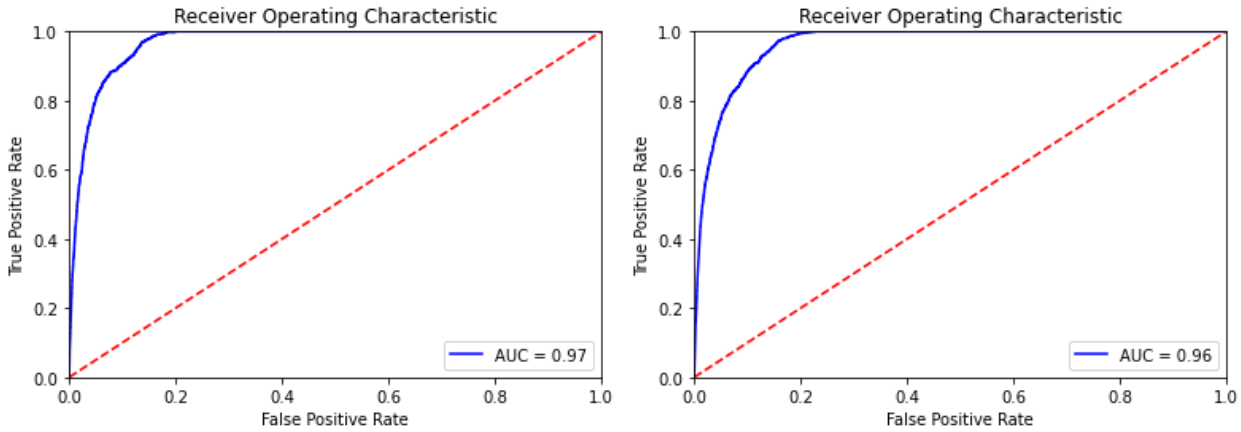


Figura 26: Confronto Curva ROC - AUC Logistic Regression (sinistra: dataframe completo, destra: dataframe ridotto)

Il confronto tra le due Curve ROC-AUC indica che il modello con tutte le *features* ha ottenuto un'*Area Under the Curve* (AUC) leggermente superiore (0.97) rispetto al modello con riduzione *features* (0.96).
Va notato inoltre, che la differenza tra i valori di AUC dei due modelli è relativamente piccola e potrebbe non essere significativa in termini di prestazioni del modello.

Decision Tree

L'algoritmo Decision Tree addestrato con il *dataframe* completo presenta un'*accuracy* dello 0.93, una *precision* pari a 0.97, una *recall* dello 0.93 e un *f1-score* dello 0.94. D'altra parte, il modello addestrato con il *dataframe* ridotto presenta un'*accuracy* dello 0.91, una precision dello 0.97, una *recall* dello 0.91 e un *f1-score* pari a 0.93.
Dal confronto delle due configurazioni si può notare che il modello addestrato con il *dataframe* completo ha una *recall* leggermente superiore rispetto al modello addestrato con quello ridotto, il che significa che il modello completo è in grado di identificare un maggior numero di casi positivi. Tuttavia, la differenza nella *recall* tra i due modelli è relativamente piccola. Inoltre, l'*accuracy* e la *precision* sono simili tra i due modelli.

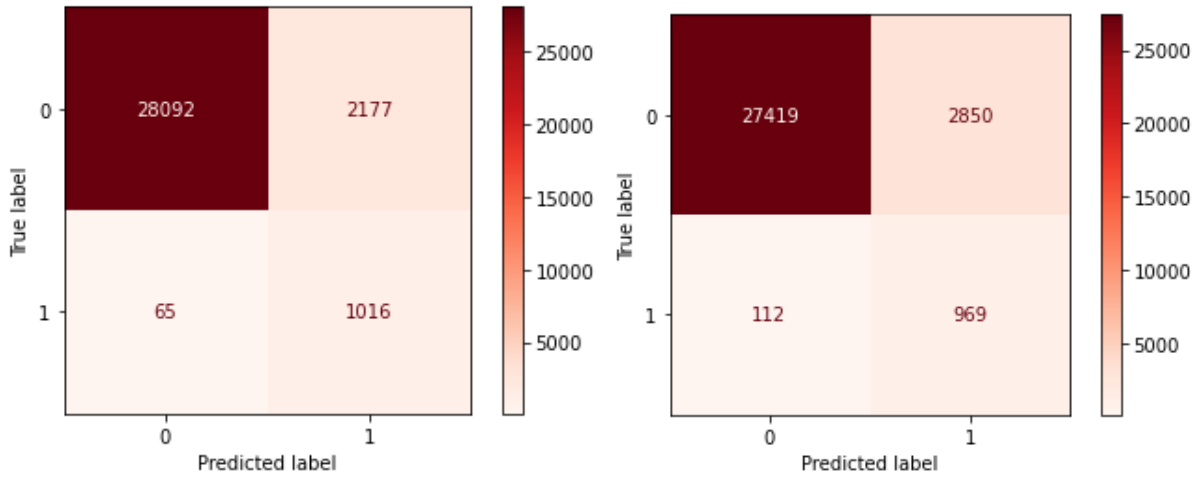


Figura 27: Confronto Confusion Matrix Decision Tree (sinistra: dataframe completo, destra: dataframe ridotto)

Questi risultati indicano che il modello con tutte le *features* ha identificato correttamente un maggior numero di casi positivi (1016) rispetto al modello con il *dataframe* ridotto (969). I risultati suggeriscono che il modello con il *dataframe* completo potrebbe essere leggermente migliore nell'identificare i casi positivi e più preciso nel prevedere la *customer loyalty*, evitando di identificare erroneamente alcuni casi negativi come positivi.

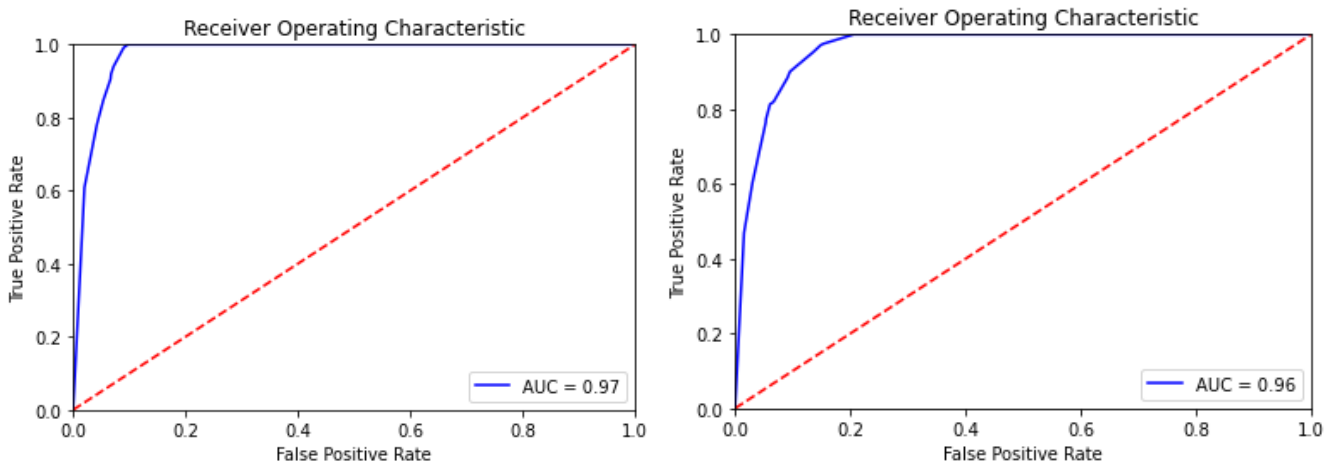


Figura 28: Confronto Curva ROC - AUC Decision Tree (sinistra: dataframe completo, destra: dataframe ridotto)

La differenza di AUC tra il modello con tutte le *features* (AUC=0.97) e il modello con riduzione (AUC=0.96) è di 0.01. Sebbene il modello con tutte le *features* abbia ottenuto un valore di AUC leggermente superiore, questa differenza è relativamente piccola.

Naïve Bayes

La valutazione delle metriche del Naïve Bayes indica che, in entrambi i casi, la *precision* del modello è stata elevata, raggiungendo un valore pari a 0.97, il che suggerisce che il modello ha prodotto un elevato numero di previsioni corrette mentre l'*accuracy*, sempre in entrambi i casi, è pari allo 0.93. Tuttavia, si può notare una differenza tra le due configurazioni: il modello con tutte le *features* ha prodotto una *recall* dello 0.93, che indica la capacità del modello di identificare correttamente la maggior parte dei casi positivi, mentre la *recall* del modello con riduzione delle *features* è stata dello 0.91, leggermente inferiore.

Inoltre, l'*F1-score* è stato dello 0. 94 per il modello completo e dello 0.93 per l'altro. In generale, si può notare che il modello con tutte le *features* ha prestazioni leggermente superiori rispetto al modello con la riduzione, soprattutto in termini di *recall* e *F1-score*.

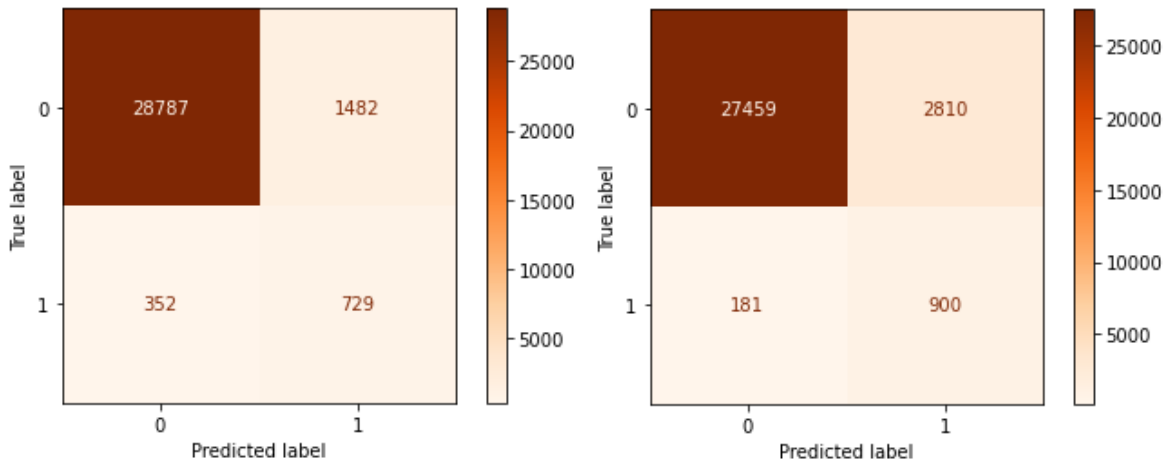


Figura 29: Confronto Confusion Matrix Naïve Bayes (sinistra: dataframe completo, destra: dataframe ridotto)

In entrambe le matrici di confusione, il numero di veri negativi (TN) è elevato, il che indica una buona capacità del modello di classificare correttamente le istanze negative (ovvero i clienti che non hanno fedeltà nei confronti dell'azienda).
Tuttavia, i numeri di falsi positivi (FP) e falsi negativi (FN) differiscono tra le due *confusion matrix*. Nel modello addestrato con tutte le *features*, il numero di falsi positivi è inferiore rispetto al *dataset* ridotto, ma il numero di falsi negativi è maggiore. Ciò potrebbe indicare che il modello addestrato con tutte le *features* ha una maggiore capacità di distinguere tra le istanze positive e negative, ma allo stesso tempo potrebbe perdere alcune istanze positive.
D'altra parte, il modello addestrato sul *dataset* ridotto ha un maggior numero di falsi positivi, il che potrebbe significare una maggiore inclinazione a classificare erroneamente alcune istanze come positive, ma al tempo stesso ha un minore numero di falsi negativi, che potrebbe indicare una maggiore sensibilità alle istanze positive.

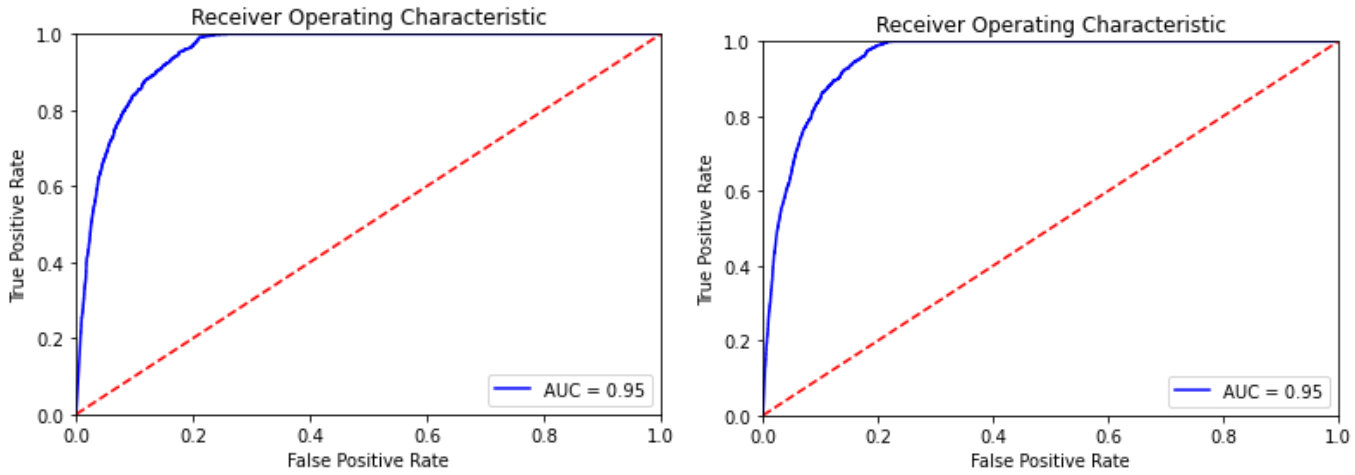


Figura 30: Confronto Curva ROC - AUC Naïve Bayes (sinistra: dataframe completo, destra: dataframe ridotto)

Il confronto tra due curve ROC - AUC con valori simili come quelle mostrate nella Figura 30 (0.95 per entrambe) suggerisce che il modello ha una buona capacità di distinguere tra le classi positive e negative, indipendentemente dal numero di *features* utilizzate.

K-neighbors

Entrambi i casi presentano gli stessi risultati: una *precision* dello 0.97 indica che il modello è molto preciso nella classificazione dei dati, mentre una *recall* dello 0.92 indica che il modello è in grado di identificare correttamente il 92% dei casi positivi, l'*F1-score* è dello 0.94 ed infine, l'*accuracy* pari a 0.92 indica che il

modello è preciso nel classificare correttamente i dati, ma come si evince della *confusion matrics*, non è perfettamente in grado di individuare tutti i casi positivi.

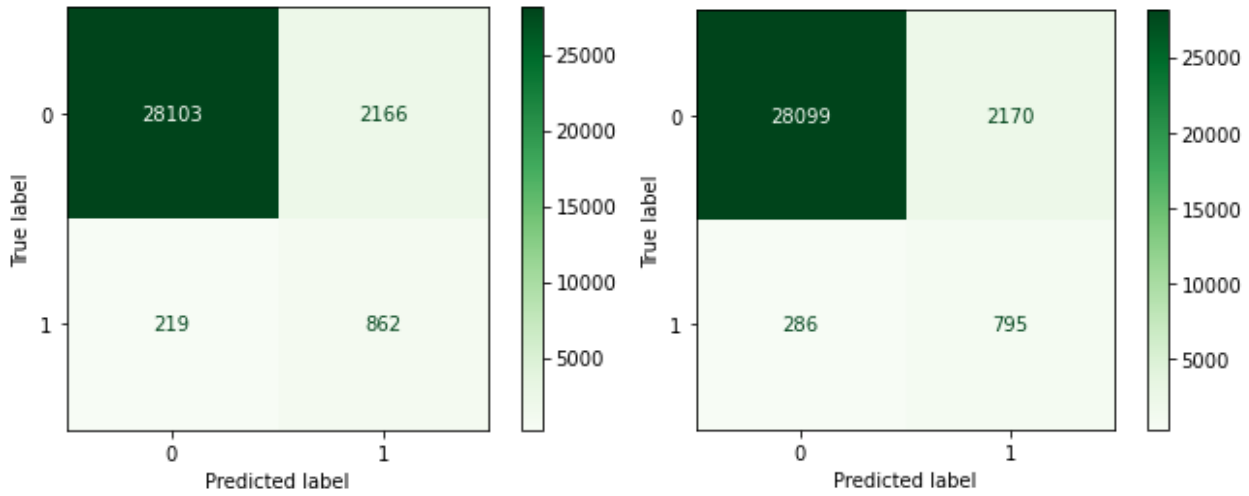


Figura 31: Confronto Confusion Matrix K-neighbors (sinistra: dataframe completo, destra: dataframe ridotto)

Possiamo notare che, la matrice di confusione ottenuta dal *dataframe* completo mostra un maggior numero di veri negativi e veri positivi rispetto alla matrice di confusione del *dataframe* ridotto. Ciò suggerisce che il modello con tutte le *features* ha una maggiore capacità di distinguere tra le classi positive e negative rispetto al modello con riduzione di *features*.

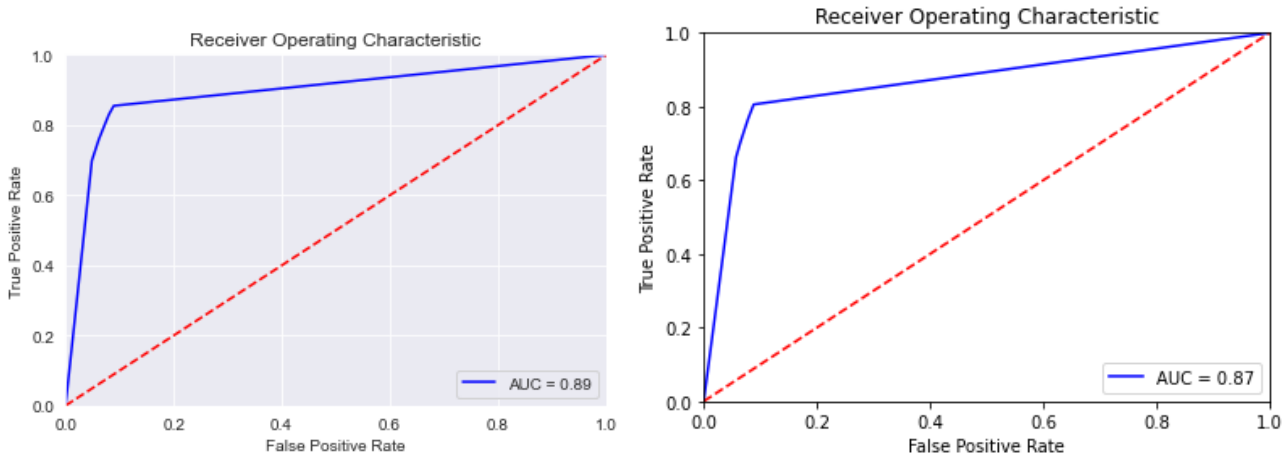


Figura 32: Confronto Curva ROC – AUC K-neighbors (sinistra: dataframe completo, destra: dataframe ridotto)

Il confronto tra le due curve ROC - AUC, con valori leggermente differenti (0.89 per la prima e 0.87 per la seconda), implica che il modello addestrato con tutte le *features* potrebbe presentare una migliore capacità di discriminazione tra le classi positive e negative rispetto al modello addestrato con una la riduzione.

XGBoost Classifier

Anche con questo algoritmo è possibile notare che i risultati delle metriche di valutazione sono le stesse per entrambi i casi. *Precision* dello 0.97, una *recall* dello 0.92, l'*F1-score* è dello 0.94 ed infine, *l'accuracy* dello 0.92.

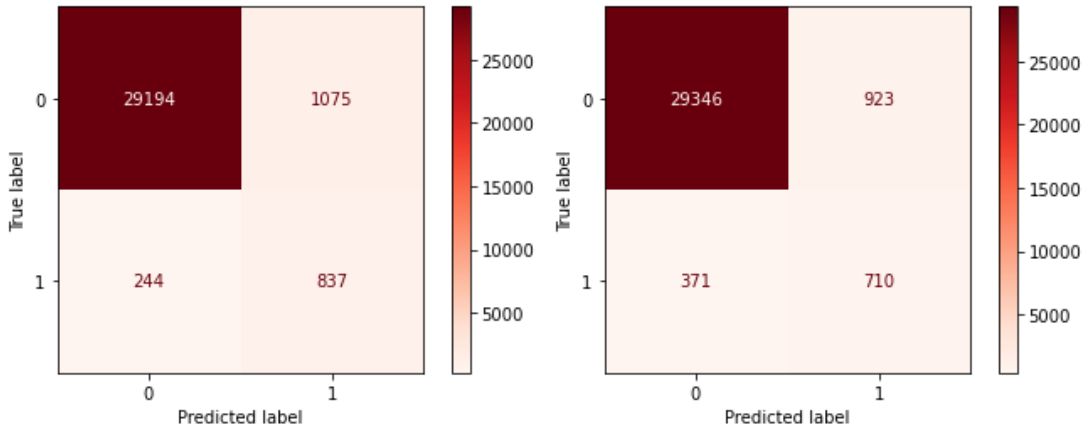


Figura 33: Confronto Confusion Matrix XGBoost Classifier (sinistra: dataframe completo, destra: dataframe ridotto)

La prima matrice di confusione presenta un numero di falsi positivi maggiore e un numero di falsi negativi inferiore rispetto alla seconda. In un contesto di previsione della *customer loyalty* dei consumatori, i falsi negativi sono considerati più rilevanti dei falsi positivi. Ciò implica che il modello addestrato con il *dataframe* completo potrebbe essere preferibile in situazioni in cui è necessario minimizzare il numero di falsi negativi, ovvero quando si desidera prevedere correttamente la *customer loyalty* dei clienti, anche se ciò comporta un aumento dei falsi positivi.

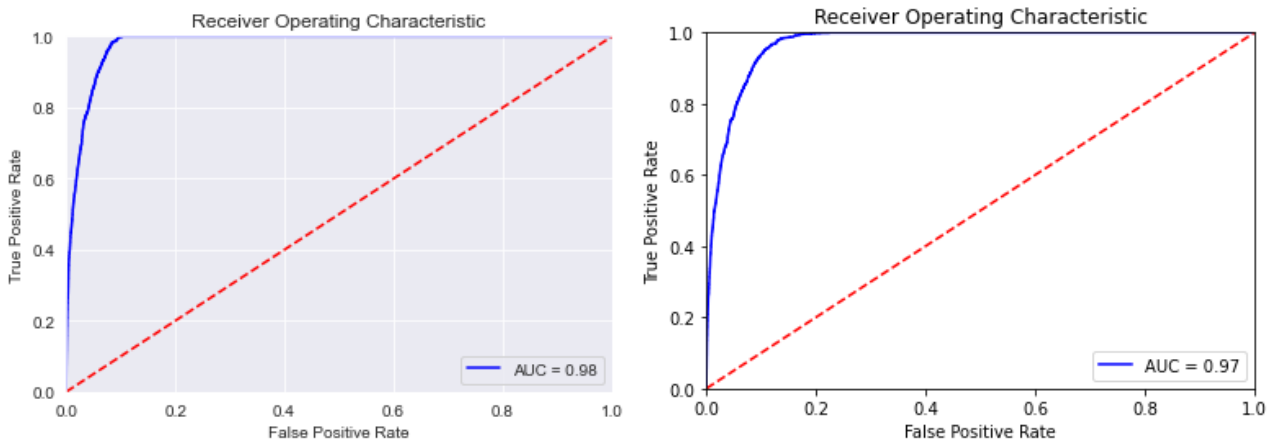


Figura 34: Confronto Curva ROC – AUC XGBoost Classifier (sinistra: dataframe completo, destra: dataframe ridotto)

Il confronto tra due curve ROC - AUC con valori elevati (0.98 per la prima e 0.97 per la seconda) suggerisce che entrambi i modelli hanno una buona capacità di distinguere tra le classi positive e negative.

Completo	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Random Forest	0,97	0,95	0,96	0,95
Logistic Regression	0,97	0,88	0,91	0,88
Decision Tree	0,97	0,93	0,94	0,93
Naïve Bayes	0,97	0,93	0,94	0,93
K-Neighbors	0,97	0,92	0,94	0,92
Ridotto	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Random Forest	0,97	0,95	0,96	0,95
Logistic Regression	0,97	0,87	0,91	0,87
Decision Tree	0,97	0,91	0,93	0,91
Naïve Bayes	0,97	0,91	0,93	0,93
K-Neighbors	0,97	0,92	0,94	0,92

Figura 35: Tabella riassuntiva dei risultati sia con il dataframe completo che ridotto

Date le buone *performance* ottenute dei modelli di *machine learning* utilizzati è stato ritenuto interessante andare a vedere quali sono state le *features* più rilevanti.

L'importanza della caratteristica indica quanto ciascuna contribuisca alla previsione del modello e viene rappresentata utilizzando un valore numerico chiamato punteggio, più è alto questo valore più la *feature* è rilevante per l'analisi.

L'utilizzo della *feature importance* è utile poiché ne accelera il funzionamento aiutando a capire quali caratteristiche sono irrilevanti per il modello e potrebbe migliorarne le prestazioni.

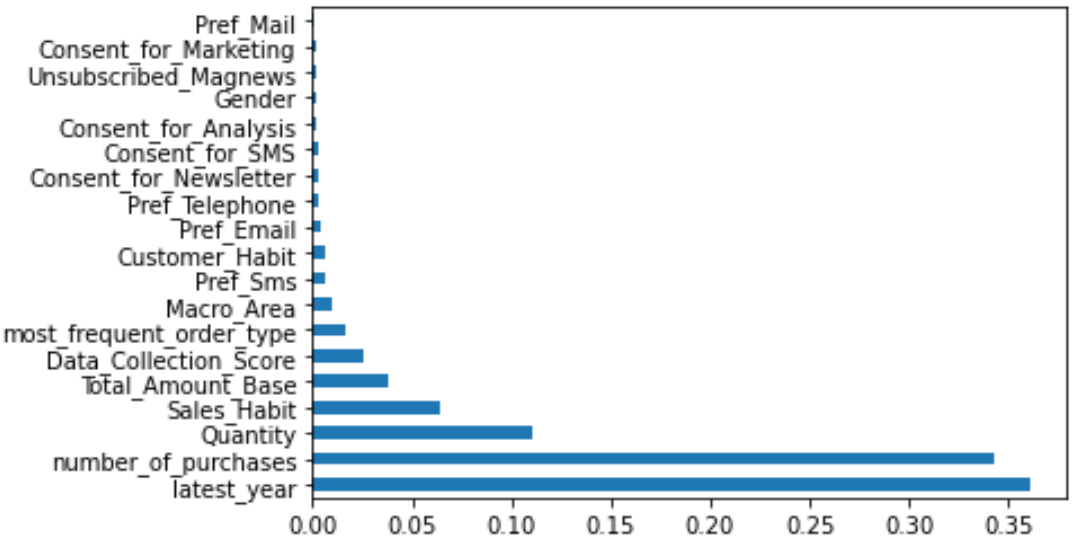


Figura 36: feature importance Random Forest

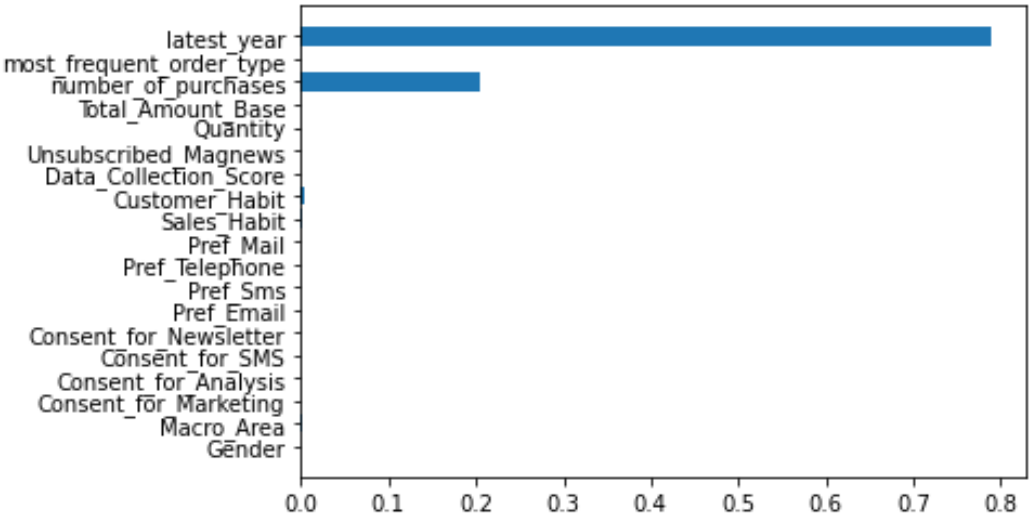


Figura 37: feature importance Decision Tree

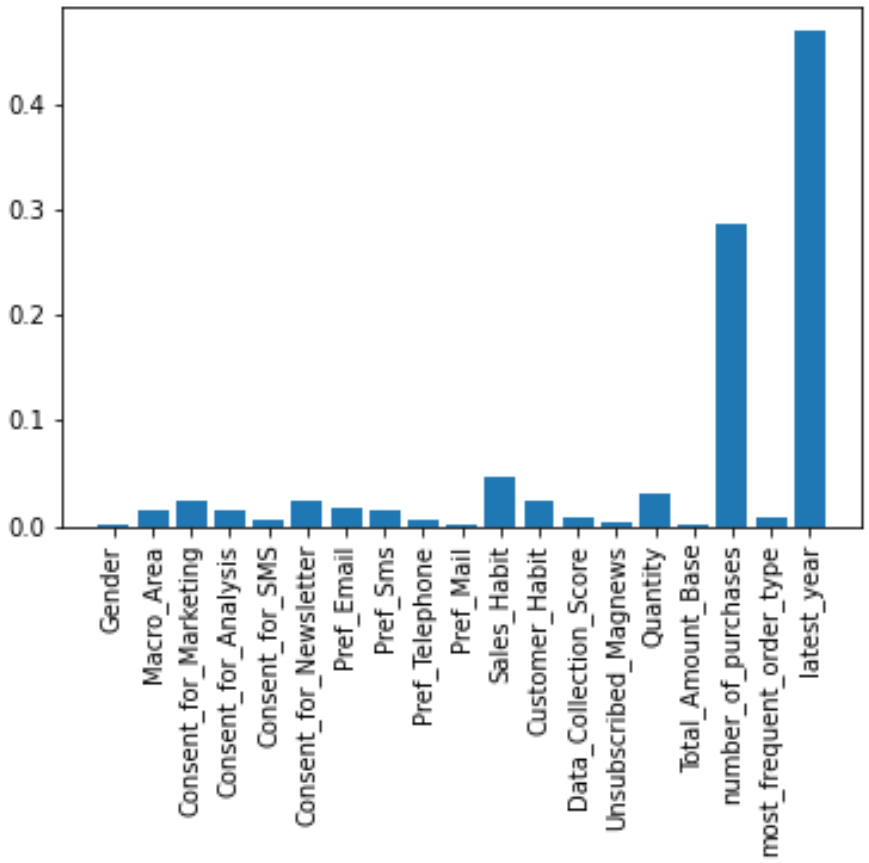


Figura 38: feature importance Logistic Recression

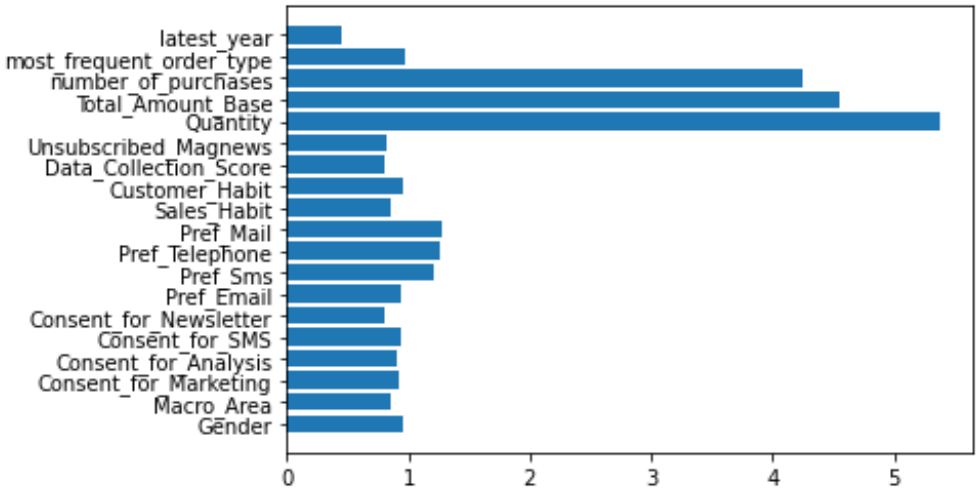


Figura 39: feature importance Naïve Bayes

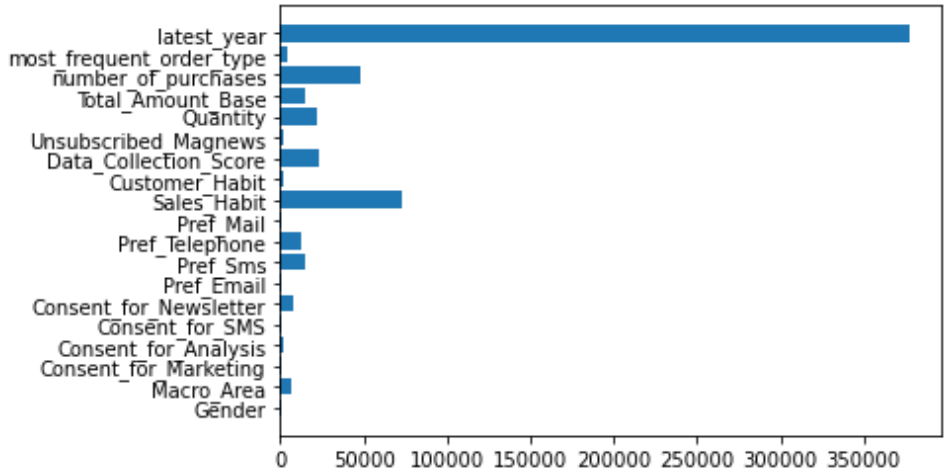


Figura 40: feature importance K-neighbors

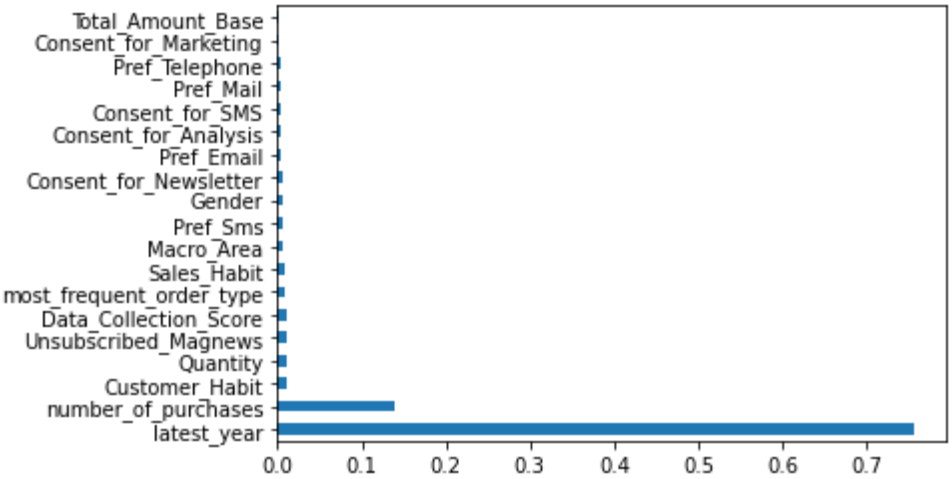


Figura 41: feature importance XGBoost Classifier

È stata dunque stimata l'importanza delle caratteristiche di tutti gli algoritmi applicati e i risultati sono riportati nelle figure 36 - 41 dove le caratteristiche sono classificate e rappresentate graficamente in base al punteggio di importanza. Questa analisi ha identificato *features* altamente valide per la previsione della *customer loyalty* che potrebbero mostrare una potenziale utilità per i *marketer* che cercano di prevederla.

Lo studio mira a trovare le migliori tecniche di *machine learning* per la classificazione della *customer loyalty* tra una serie di algoritmi ben accettati e facili da implementare, scoprendo che, almeno per questo *set* di dati, hanno dato buoni risultati. Si tratta di una fase iniziale dell'utilizzo di approcci di *machine learning* in questo contesto ed un possibile sviluppo futuro potrebbe essere quello di testare se riducendo le caratteristiche si potrebbero raggiungere delle migliori prestazioni predittive.

4.2 Applicazioni nel marketing

Nell'attuale ambiente competitivo e omnicanale, in cui i costi di cambiamento sono bassi e i clienti possono confrontare in modo più trasparente le offerte e i livelli di prezzo dei rivenditori, la fedeltà dei clienti sembra sempre più difficile da raggiungere e nel contesto della vendita al dettaglio, in cui i clienti acquistano ripetutamente, comprendere la fedeltà dei clienti e come influenzarla è essenziale.
Incrementare la *customer loyalty* rappresenta quindi un obiettivo cruciale per le aziende, poiché una clientela fedele può generare maggiori profitti e vantaggi competitivi. Tuttavia, identificare i clienti fedeli non è un compito facile e le aziende spesso si affidano a metriche e indicatori per monitorare la *customer loyalty*. In questo contesto, la classificazione binaria della *customer loyalty* rappresenta un'importante metodologia per suddividere i clienti in due gruppi: fedeli e infedeli e tale modello ha implicazioni significative per le strategie di fidelizzazione dei clienti adottate dalle aziende e per le decisioni manageriali riguardanti la gestione del rapporto con la clientela. Le aziende possono creare strategie di marketing mirate e personalizzate in grado di soddisfare le esigenze dei clienti più fedeli, fidelizzandoli ulteriormente e aumentando la probabilità di acquisto da parte loro. In questo paragrafo, analizzeremo le implicazioni manageriali della classificazione della *customer loyalty*, discutendo le possibili azioni che le aziende possono intraprendere.

I consumatori moderni sono sommersi dal marketing e dalla messaggistica dei vari *brand* e di conseguenza, sono diventati sempre più esigenti riguardo ai messaggi con cui interagire. Quando si utilizza una strategia basata sui dati, i *team* di marketing possono aumentare drasticamente le possibilità che il loro pubblico di destinazione faccia *clic* sul loro annuncio, legga un *post* sul *blog* o esegua un'altra azione che guida un obiettivo di conversione.
Le strategie basate sui dati migliorano l'esperienza del cliente, la percezione del marchio e la fedeltà dei clienti in quanto offrono alle organizzazioni una comprensione delle esigenze e degli interessi dei consumatori, migliorano anche i tassi di conversione perché è più probabile che la messaggistica altamente mirata abilitata

da questo tipo di marketing attiri l'attenzione degli utenti ed infatti queste strategie si concentrano sull'utilizzo di profili approfonditi dei consumatori per migliorare la loro esperienza.

Un elemento fondamentale per i professionisti del marketing è determinare dove viene sprecato il loro *budget* pubblicitario. Il *data-driven* marketing, guidato da strumenti di analisi, consente ai *team* di marketing di scoprire quale parte del *budget* pubblicitario sta avendo il maggiore impatto sulle conversioni o sulla consapevolezza del marchio e questo viene fatto valutando i percorsi dei clienti utilizzando modelli di attribuzione, come la misurazione del marketing unificato che esamina l'attribuzione *multitouch* e la modellazione del *media mix* per fornire una visione completa del percorso di acquisto. Le organizzazioni possono determinare cosa sposta i potenziali clienti e i clienti lungo la canalizzazione e quindi allocare i fondi di conseguenza.

Valutare i dati dei consumatori offre ai *team* di marketing informazioni sui tipi di creatività, su elementi visivi, i testi e i contenuti con cui il pubblico di destinazione preferisce interagire. Trasmettere il messaggio giusto, che soddisfi gli interessi personali e crei valore, al momento giusto è essenziale per entrare in contatto con i consumatori. Tale approccio può condurre a decisioni più efficaci riguardo ai prodotti e può fornire un'analisi approfondita dei clienti per l'azienda.

Questo tipo di strategie sono positive sia per i *marketer* che per i consumatori. Tuttavia, ci sono alcune sfide che possono impedire ai professionisti del marketing di trarre tutti i vantaggi dai dati o di raggiungere i clienti in modo efficace. In prima istanza i *brand* dovrebbero evitare di essere troppo invasivi; nonostante il desiderio dei consumatori di fruire di esperienze personalizzate, vi è una reticenza nei confronti delle organizzazioni che acquisiscono una conoscenza eccessiva su di loro. In particolare, se i consumatori scelgono di condividere informazioni personali, desiderano avere chiarezza sul modo in cui tali dati saranno utilizzati a loro beneficio. Un ulteriore aspetto da considerare riguarda la qualità insufficiente dei dati; al fine di implementare una strategia *data-driven* efficiente, è necessario avere adeguati processi di gestione dei dati che garantirà la possibilità di prendere decisioni e sviluppare delle strategie basate sui dati di alta qualità che riflettano accuratamente le necessità dei clienti. Nel caso in cui i dati a disposizione non soddisfino criteri di qualità quali tempestività, accuratezza, completezza e rappresentatività, vi è il rischio di basare le decisioni su dati che offrono una visibilità limitata sulle reali esigenze dei clienti.
Il marketing incentrato sui dati mira a potenziare l'efficacia del marketing attraverso un'ottimizzazione dell'esperienza del cliente che riveste un ruolo chiave in questa dinamica. Pertanto, ogni campagna sviluppata con l'ausilio dei dati dovrebbe fornire una chiara dimostrazione dei benefici che il cliente può ottenere.

La predizione del comportamento dei clienti in relazione alla loro fedeltà rappresenta un importante strumento competitivo per le aziende, in grado di garantire una maggiore efficacia delle strategie di marketing e di

fidelizzazione. La classificazione dei clienti in base alla loro fedeltà rappresenta la prima fase di questo processo ma la previsione del loro comportamento futuro è l'elemento chiave per adattare le strategie di marketing e di *retention* alle loro esigenze. Ciò, infatti, consente di concentrare le risorse sui clienti più importanti e di offrire loro un'esperienza di acquisto personalizzata, che può portare a un aumento della fedeltà e, di conseguenza, della redditività del marchio.

Questa capacità predittiva del comportamento dei clienti conferisce un vantaggio competitivo alle imprese, consentendo loro di (1) adottare misure preventive per ridurre il *churn* dei clienti; tale defezione rappresenta un costo significativo per le aziende, sia in termini di perdita di fatturato che di risorse impiegate per rimpiazzare i clienti persi. Attraverso l'analisi predittiva, gli algoritmi di classificazione possono individuare dei segnali di imminente defezione, identificando i clienti che potrebbero essere più propensi ad abbandonare l'azienda e questo consente ai dirigenti di adottare misure preventive mirate per mantenere i clienti a rischio. Ad esempio, possono essere implementate strategie di fidelizzazione personalizzate, come offerte speciali, assistenza dedicata, programmi di fedeltà o miglioramenti del servizio. Ridurre il tasso di defezione dei clienti è fondamentale per mantenere una base di clientela stabile e sostenibile nel lungo termine.

(2) Identificare le opportunità di *cross-selling* e *up-selling* in base alle preferenze individuali dei clienti; gli algoritmi di classificazione possono individuare modelli e associazioni tra i diversi prodotti o servizi che i clienti tendono ad acquistare o utilizzare insieme. Questo consente ai *manager* di identificare opportunità di *cross-selling*, che consistono nel proporre ai clienti prodotti o servizi complementari a quelli che hanno già acquistato. Ad esempio, se un cliente ha acquistato un computer portatile, l'azienda potrebbe suggerire l'acquisto di una borsa per il trasporto o di un *software* aggiuntivo. Inoltre, l'analisi predittiva può rivelare anche opportunità di *up-selling*, che consistono nel proporre ai clienti prodotti o servizi di fascia superiore o con maggiori funzionalità rispetto a quelli che hanno giò considerato. Questo può essere basato sulla segmentazione dei clienti e sulla comprensione delle loro caratteristiche e preferenze specifiche.

L'identificazione delle opportunità di *cross-selling* e *up-selling* consente alle aziende di massimizzare il valore dei propri clienti esistenti offrendo loro prodotti o servizi aggiuntivi che sono rilevanti e di interesse e quindi aumentano il valore medio delle transazioni e stimolano ulteriori acquisti che si traduce in un incremento delle entrate e della redditività complessiva dell'azienda. Inoltre, favorisce l'esperienza del cliente, poiché gli viene presentata una proposta su misura, rispondendo alle sue esigenze specifiche e migliorando la sua soddisfazione complessiva.

(3) Ottimizzazione dell'allocazione delle risorse di marketing; le aziende dedicano una parte significativa del loro *budget* alle attività di marketing, come pubblicità, promozioni, eventi e altro ancora. Tuttavia, l'allocazione di queste risorse non può essere casuale o basata solo su supposizioni. È fondamentale investire le risorse di marketing in modo mirato ed efficiente per massimizzare il ritorno sull'investimento infatti, la previsione accurata del comportamento dei clienti consente alle aziende di ottenere una comprensione approfondita di come i clienti risponderanno alle diverse iniziative di marketing. Gli algoritmi di intelligenza

artificiale possono analizzare dati storici, modelli di acquisto, preferenze e altre variabili per prevedere quali clienti sono più inclini a rispondere positivamente a una determinata campagna o offerta.
Utilizzando queste previsioni, le aziende possono ottimizzare l'allocazione delle risorse di marketing in modo da concentrare gli sforzi su quei segmenti di clientela che hanno maggiori probabilità di rispondere positivamente.
Questa ottimizzazione dell'allocazione delle risorse porta a diversi vantaggi: innanzitutto, consente di evitare sprechi di risorse su segmenti che hanno scarse probabilità di risposta positiva; in secondo luogo, consente di massimizzare l'impatto delle iniziative di marketing, concentrandosi su clienti con una maggiore probabilità di conversione ed infine, aiuta a migliorare l'efficienza complessiva delle attività di marketing, consentendo un utilizzo più strategico e mirato delle risorse disponibili.
(4) Migliorare il *customer journey*, ovvero l'intero percorso che un cliente compie dal momento in cui scopre un prodotto o servizio fino all'acquisto e oltre. Identificando i punti critici del percorso del cliente e prevedendo le loro azioni e preferenze, le aziende possono personalizzare l'esperienza del cliente in modo più efficace.

L'emergere della pandemia da COVID-19 ha avuto un impatto significativo sul modo di comprare e sul comportamento dei consumatori, accelerando l'adozione e il potenziamento dell'utilizzo dei *big data* e del *machine learning* per analizzare le abitudini di consumo. Con l'aumento delle restrizioni e la necessità di distanziamento sociale, i consumatori si sono rivolti sempre di più agli acquisti *online* per soddisfare le loro esigenze e questo ha creato una crescente quantità di dati digitali, che servono per comprendere le preferenze dei consumatori e personalizzare l'esperienza di acquisto.
Dato che i modelli di acquisto sono stati ridefiniti e molte aziende all'avanguardia hanno adottato strategie basate sui dati per anticipare le tendenze e offrire prodotti e servizi rilevanti.
La pandemia ha anche spinto i consumatori ad essere più consapevoli delle loro scelte e a cercare prodotti e servizi che soddisfino criteri di sicurezza, sostenibilità e benessere. Gli acquirenti, infatti, cercano informazioni dettagliate sui prodotti, valutazioni dei clienti e *feedback* prima di prendere una decisione d'acquisto e in risposta a questa richiesta, i rivenditori e gli *e-commerce* devono sviluppare algoritmi di raccomandazione personalizzati basati sui *big data*, che suggeriscono prodotti pertinenti e migliorano l'esperienza d'acquisto.
Nel panorama post-pandemico, l'uso dei dati giocherà un ruolo fondamentale nel soddisfare le aspettative dei consumatori e nel guidare il successo delle imprese.

Pertanto, risulta evidente come l'abilità di sfruttare il *machine learning* e l'analisi dei *big data* per anticipare il comportamento dei consumatori non comporti soltanto un aumento delle conversioni, obiettivo di per sé significativo ma implica altresì il potenziamento della *brand awareness* e la fidelizzazione dei clienti, generando l'immagine di un marchio in grado non solo di ascoltare, ma addirittura di suggerire soluzioni perfette. Considerando che le imprese si adoperano per mantenere la competitività in un mercato in costante evoluzione, la previsione accurata e tempestiva della domanda assume sempre maggiore importanza.

I modelli di *machine learning* possono essere costantemente aggiornati e ottimizzati, consentendo alle aziende di adeguare prontamente le proprie previsioni per rispecchiare le mutevoli condizioni di mercato e ciò permetterà alle aziende di ottenere previsioni più precise con una velocità ed efficienza senza precedenti.
Inoltre, questo approccio può essere applicato anche al marketing di prodotto, consentendo lo sviluppo di nuovi articoli o servizi sempre più personalizzati e in linea con le preferenze del target di riferimento.

Una strategia orientata al futuro che si traduce in risultati a lungo termine, grazie a ottimizzazioni continue che considerano l'evoluzione dei *trend* nel corso del tempo.
I vantaggi del marketing predittivo si riflettono chiaramente sulle vendite e sulla crescita, ma possono spingere le aziende ancora oltre. La capacità di anticipare con precisione le tendenze future può influenzare ogni aspetto del marketing aziendale.
Il marketing predittivo rappresenta il risultato finale e naturale delle strategie di marketing basate sui dati che sono state sviluppate nel tempo, partendo dalla definizione di una strategia dati complessiva e ben strutturata e dallo sviluppo delle competenze necessarie, sia a livello tecnico che manageriale. Per ottenere risultati effettivi, il marketing predittivo non può prescindere dalla mappatura dei percorsi dei clienti. L'identificazione di tutti i punti di contatto attraversati e della loro sequenza è fondamentale, poiché consente di valutare localmente e globalmente l'accessibilità, l'ergonomia e l'esperienza utente di ciascun punto di contatto e del flusso complessivo, nonché di raccogliere dati ad ogni *touchpoint*.

Il cambiamento del comportamento dei clienti è influenzato da tre aspettative che i consumatori hanno: personalizzazione, previsione e adattabilità e tali aspettative stanno promuovendo una tendenza che va verso l'instaurazione di una relazione intima con il cliente. Considerando che le aspettative di intimità sono destinate a crescere ulteriormente, le aziende devono concentrarsi sulla creazione di una connessione significativa e questo implica l'attribuzione di un ruolo centrale degli individui, l'accelerazione nell'adozione di analisi e intelligenza artificiale, nonché l'innovazione su vasta scala.

L'intimità con il cliente rappresenta una strategia volta a comprenderne le specifiche esigenze al fine di fornire la "migliore soluzione", ovvero prodotti o servizi adattati in modo continuativo alle situazioni e alle nicchie specifiche dei clienti. L'intimità è intesa come "vicinanza" e richiede una completa ristrutturazione dell'intera organizzazione per avere successo.
Al centro di questa tendenza, come già accennato, si trovano tre aspettative predominanti dei clienti:

1. Personalizzazione: i clienti si aspettano che ogni interazione con un'azienda sia personalizzata. Desiderano che le aziende conoscano la loro identità, le loro preferenze e il modo in cui desiderano essere serviti.
2. Previsione: i clienti si aspettano che, grazie alla loro relazione con un'azienda e alla dimostrazione delle loro preferenze e comportamenti d'acquisto, quest'ultima sia in grado di anticipare opportunità di

prodotti o servizi che potrebbero risultare vantaggiosi e offrirli prima ancora che i clienti ne esprimano esplicitamente il desiderio.

3. Adattabilità: i clienti vivono oggi in un ambiente sempre più intelligente e si aspettano che le aziende dimostrino la stessa sensibilità nei confronti del loro ambiente e si adattino alle specifiche esigenze di quest'ultimo. Vogliono che le aziende siano consapevoli della loro posizione e dei loro desideri in termini di coinvolgimento.

La personalizzazione rappresenta una capacità essenziale che ha un impatto significativo, indipendentemente dalla natura dell'azienda, infatti, i consumatori non solo la desiderano ma la considerano una necessità imprescindibile. I vantaggi del marketing personalizzato sono molteplici, sia per le aziende che per i consumatori ma solo quando le strategie vengono implementate con successo, si possono ottenere risultati significativi.

La creazione di un'esperienza coinvolgente e rilevante per i clienti rappresenta uno degli obiettivi fondamentali della personalizzazione delle risorse, infatti, quando viene offerta un'esperienza personalizzata, i clienti percepiscono un coinvolgimento e un apprezzamento che aumentano il loro legame con il *brand* e la loro probabilità di soddisfazione. Questa esperienza coinvolgente si basa sull'interazione costante tra l'azienda e il cliente, che può avvenire attraverso diverse piattaforme come *siti web*, applicazioni o *social media*.

Le aziende possono sfruttare le informazioni raccolte dall'analisi predittiva per personalizzare l'esperienza del cliente su tali canali, ad esempio fornendo contenuti pertinenti, raccomandazioni personalizzate o promozioni speciali.

Quando i clienti si sentono coinvolti e ricevono un'esperienza rilevante, sono più inclini a sviluppare un legame emotivo con il *brand* e a rimanere fedeli nel tempo e le aziende che dimostrano di essere capaci di creare questa vicinanza con i clienti registrano tassi di crescita dei ricavi più rapidi rispetto ai loro concorrenti.

Con l'uso appropriato della tecnologia di automazione, i professionisti del marketing possono identificare il canale con cui i clienti interagiscono e automatizzare il *follow-up* su diversi canali come parte di un approccio omnicanale.

Inoltre, il marketing personalizzato favorisce l'aumento della fedeltà al marchio; quando i consumatori forniscono informazioni e dati, si aspettano di essere trattati come individui unici con preferenze specifiche e le aziende che dedicano tempo e risorse all'implementazione di strategie di marketing personalizzate di successo beneficeranno di un vantaggio competitivo sia in termini di fedeltà al marchio che di soddisfazione del cliente. Infine, questa tipologia di marketing aiuta a creare coerenza tra i vari canali. I consumatori interagiscono con i marchi attraverso diversi canali come e-mail, social media, dispositivi mobili, ecc ed è quindi fondamentale che i marchi creino un'esperienza coerente tra questi ultimi. L'esperienza in negozio dovrebbe essere allineata con quella dell'*app*, che a sua volta dovrebbe essere coerente con le e-mail, così facendo i clienti possono riprendere la conversazione da dove l'hanno interrotta e indipendentemente dal canale utilizzato.

Non è mai stato un momento più opportuno per sfruttare la personalizzazione dato che attualmente, le persone desiderano esprimere la propria individualità e personalità attraverso tutto ciò che possiedono, inclusi abbigliamento, accessori e persino l'arredamento domestico. Inoltre, la personalizzazione è diventata un elemento cruciale nel marketing dei *brand*, con numerose aziende che cercano di personalizzare i propri prodotti per attrarre il loro pubblico di riferimento poiché adattando i prodotti alle esigenze individuali dei clienti, si garantisce che essi ottengano esattamente ciò che desiderano. Ciò non solo contribuirà alla loro soddisfazione, ma, soprattutto, aumenterà la probabilità che ritornino per future transazioni. La creazione di una base di clienti fedeli è fondamentale per mantenere alti i profitti e la personalizzazione rappresenta un ottimo strumento per raggiungere questo obiettivo.

Ci troviamo in un periodo in cui la capacità di un'azienda di dimostrare un'elevata empatia, instaurare una forte intimità e offrire totale affidabilità assume un'importanza cruciale e solo quando le aziende pongono i clienti al centro delle proprie attività, mostrando empatia, interagendo in modo risonante e gestendo eticamente i dati forniti, i clienti tendono a ricordare e a rimanere fedeli. Concentrando l'attenzione sull'individuo, le aziende possono creare una forma di scambio di valore in cui i clienti forniscono i propri dati in cambio di garanzie sulla *privacy*, costruzione di rapporti di fiducia e offerta di un'esperienza unica e personalizzata che anticipa le loro esigenze e si adatta al loro contesto ambientale.

I dirigenti aziendali devono porre maggiore attenzione al programma di trasformazione aziendale al fine di creare intimità con i clienti e questo implica l'accelerazione dell'uso dell'analisi e dell'intelligenza artificiale per l'iperpersonalizzazione e la previsione, nonché l'innovazione su larga scala per adattarsi ai cambiamenti delle esigenze dei clienti. È necessario quindi abbandonare le tradizionali fonti di vantaggio competitivo a favore di tre dinamici fattori di valore: il posizionamento dei consumatori al centro, l'utilizzo della tecnologia e l'innovazione su larga scala.
In questo modo, i *manager* potranno dimostrare che le loro aziende creano un valore a lungo termine, misurabile attraverso le *performance* ottenute con i clienti, i dipendenti e la società nel suo complesso e se gestito in modo adeguato, questo approccio si ripagherà migliorando le performance aziendali e garantendo la sostenibilità nel mercato.

Capitolo 5 - Conclusioni

Dai capitoli relativi all'Introduzione e alla Literature Review è emerso come le aziende di oggi, sia per volontà che per necessità, stiano sempre più orientandosi verso l'adozione dell'intelligenza artificiale al fine di migliorare le relazioni con i propri clienti.
Il settore della moda, oggetto del presente elaborato, ha avviato da tempo tale transizione; tuttavia, nonostante i progressi compiuti, vi è ancora un ampio margine di miglioramento da percorrere.

In questo elaborato, è stato condotto uno studio sull'analisi predittiva degli algoritmi di intelligenza artificiale per la fidelizzazione dei clienti nel settore della moda. Lo scopo principale è stato valutare l'efficacia di questi algoritmi nel prevedere il comportamento dei clienti e successivamente fornire raccomandazioni pratiche per l'implementazione di sistemi basati sull'intelligenza artificiale per massimizzare l'efficacia dei programmi di fidelizzazione dei clienti e ottenere un vantaggio competitivo nel mercato.

Tutti i modelli predittivi hanno ottenuto risultati incoraggianti e nessuno si è distinto chiaramente poiché tutti hanno ottenuto risultati simili.

1. *Accuracy*: I modelli hanno raggiunto un'accuratezza elevata, con valori che variano tra lo 0.87 e lo 0.97. Ciò indica che i modelli hanno una buona capacità di classificare correttamente le istanze dei dati nel contesto della *customer loyalty*.
2. *Precision*: Tutti i modelli hanno mostrato una *precision* elevata, con valori parti allo 0.97. In questo caso, i modelli sono in grado di identificare correttamente la *customer loyalty* con una percentuale molto alta.
3. *Recall*: I modelli hanno ottenuto valori compresi tra lo 0.88 e lo 0.95, indicando anche in questo caso una buona capacità di rilevare correttamente la *customer loyalty*.
4. *F1-score*: Tutti i modelli hanno ottenuto valori che si aggirano tra lo 0.91 e lo 0.96 e ciò sta ad indicare un buon equilibrio tra la capacità di identificare correttamente la *customer loyalty* e di minimizzare i falsi positivi.

I risultati ottenuti indicano che tutti i modelli hanno dimostrato una buona capacità di classificazione della *customer loyalty* nell'industria del *fashion*. Il motivo per cui potrebbe non emergere un metodo nettamente migliore potrebbe dipendere dalla natura del problema e delle caratteristiche dei dati utilizzati.
Il fatto che i risultati siano simili può indicare che i modelli stiano approssimando bene la relazione tra le *features* e la *customer loyalty*.

4.1 Limitazioni

Nel corso di questo studio, è importante considerare alcune limitazioni che potrebbero influire sia sull'interpretazione e sia sulla generalizzazione dei risultati ottenuti. I risultati si basano su un set di dati specifico e potrebbero non essere generalizzabili in tutti i contesti del settore della moda. È quindi importante considerare che la natura stessa dei dati utilizzati, come la loro qualità e disponibilità, potrebbe influenzare direttamente i risultati dell'analisi predittiva e nel caso in cui i dati siano incompleti, errati o non rappresentativi della popolazione di clienti, le previsioni potrebbero non essere accurate o affidabili. Inoltre, se l'azienda ha una base di clienti limitata o ha raccolto dati solo per un breve periodo, le previsioni potrebbero non essere generalizzabili nel lungo termine. L'analisi predittiva, infatti si basa sul presupposto che i comportamenti dei clienti rimangano stabili nel tempo, tuttavia, questi ultimi possono essere influenzati da una serie di fattori esterni e interni, quali i cambiamenti nel mercato, la concorrenza, le tendenze sociali e i gusti individuali.

Anche se gli algoritmi di classificazione possono fornire previsioni e segmentazioni utili, l'interpretazione dei risultati può essere complessa. I modelli di *machine learning* possono essere opachi e non fornire spiegazioni dettagliate su come siano state raggiunte le previsioni e questo può limitare la comprensione delle ragioni sottostanti dei comportamenti dei clienti e rendere difficile l'adattamento delle strategie di marketing in modo significativo.

L'utilizzo dei dati dei clienti per l'analisi predittiva solleva questioni legate anche all'etica e alla *privacy* che vanno considerate attentamente ed infatti è fondamentale garantire che i dati siano raccolti, conservati e utilizzati in conformità alle leggi e alle normative vigenti. Inoltre, l'uso dei dati dei clienti potrebbe suscitare preoccupazioni sulla sicurezza e la potenziale condivisione non autorizzata delle informazioni personali.

Infine, le previsioni basate sull'analisi predittiva possono essere influenzate dalle condizioni di mercato in continua evoluzione e nel caso in cui il mercato subisca improvvisi cambiamenti o turbolenze, le previsioni potrebbero non essere più valide o potrebbe essere necessario un aggiornamento per riflettere la nuova realtà.

Considerando queste limitazioni, è importante adottare un approccio consapevole nell'interpretazione e nell'applicazione dei risultati dell'analisi predittiva, tenendo conto del contesto specifico dell'azienda e delle dinamiche del mercato della moda.

5.2 Direzioni future

È opportuno considerare possibili direzioni future di ricerca che potrebbero contribuire a estendere e ad arricchire le conoscenze acquisite nell'ambito di questo studio.

Come espresso in precedenza nella sezione 4.1, è possibile condurre un ulteriore analisi per valutare se una riduzione delle caratteristiche, limitandosi a includere solo quelle considerate più rilevanti in base all'importanza attribuita dalla *feature importance*, possa portare a miglioramenti nelle prestazioni predittive.

Inoltre, con l'avanzamento della tecnologia e la crescente disponibilità di dati, le aziende avranno accesso a una vasta gamma di informazioni sui clienti e questo potrebbe includere dati provenienti da social media, dispositivi mobili ed altri canali. L'integrazione di tali dati diversificati consentirà un'analisi più approfondita e una previsione ancora più accurata del comportamento dei clienti.
L'evoluzione delle tecniche di *machine learning*, come il *deep learning* e l'apprendimento rinforzato, potrebbe migliorare ulteriormente la capacità di previsione del comportamento dei clienti. Queste tecniche consentono di gestire dati complessi e di identificare *pattern* e relazioni più sottili, fornendo previsioni ancora più precise e dettagliate.
Attualmente, molte analisi predittive si basano su dati storici per fare previsioni sul comportamento futuro dei clienti. Tuttavia, uno sviluppo futuro potrebbe essere l'utilizzo di dati in tempo reale per adattare immediatamente le risorse dedicate ai clienti e questo consentirebbe alle aziende di rispondere in tempo reale ai cambiamenti nel comportamento dei clienti e di offrire esperienze personalizzate in tempo reale.
Con la crescente attenzione sulla *privacy* dei dati e l'etica nell'utilizzo delle informazioni personali dei clienti, gli sviluppi futuri dovranno tener conto di tali considerazioni. Le aziende dovranno adottare pratiche di analisi responsabili e garantire la protezione dei dati dei clienti, rispettando le normative e le aspettative degli utenti.

Questi sviluppi consentiranno alle aziende di offrire esperienze ancora più personalizzate e rilevanti per i propri clienti, migliorando la loro fedeltà e il successo dell’azienda a lungo termine.

Appendice

https://github.com/Chi0000/Classificazione-Customer-Loyalty.git

Bibliografia

Andreeva, A. (2006). Designer Brands in Fashion Business. SPbU Publishing House, Graduate School of Management

Anderson, E. W., Mittal, V. (2000). Strengthening the Satisfaction-Profit Chain. Journal of Service Research.

Assael, H. (1992). Consumer behavior and marketing action. PWS-KENT Publishing Company.

Bowen, J. T., Chen, S.-L. (2001). The relationship between customer loyalty and customer satisfaction. International Journal of Contemporary Hospitality Management.

Choi, Y., Jae Won Choi, W.J. (2020). The Prediction of Hotel Customer Loyalty using Machine Learning Technique. International Journal of Advanced Trends in Computer Science and Engineering.

Czepiel, J. A., Congram, C. A., Shanahan J., Shanahan J. B. (Eds.). (1987). The services challenge: Integrating for competitive advantage. American Marketing Association.

Dynata. (2022). Benefits of e-commerce among global consumers as of February 2022 [Graph]. Statista.
Ehrenberg, A.S.C., Goodhardt, G.J. (2000). New brands: near instant loyalty. Journal of Marketing Management.

eMarketer. (2022). Retail e-commerce sales worldwide from 2014 to 2026 (in billion U.S. dollars) [Graph]. Statista.

Engel, J. F., Blackwell, R. D. (1982). Consumer behavior. The Dryden Press.

Fishbein, M., & Ajzen. (1975). Belief, attitude, intention, and behavior: An introduction to theory and research. Reading, MA.

Garbarino, E., Johnson, M.S. (1999). The different roles of satisfaction, trust and commitment in customer relationships. Journal of Marketing.

Giri, C., Jain, S., Zeng, X., Bruniaux, P. (2019). A detailed review of artificial intelligence applied in the fashion and apparel industry. IEEE Access.

Grayson, K., Ambler, T. (1999). The dark side of long term relationships in marketing services. Journal of Marketing Research.

Gu, X., Gao, F., Tan, M., Peng, P. (2020). Fashion analysis and understanding with artificial intelligence. Elsevier.

Guerra-Tamez, C. R., Dávila-Aguirre, M. C., Barragán Codina, J. N., Guerra Rodríguez, P. (2021). Analysis of the Elements of the Theory of Flow and Perceived Value and Their Influence in Craft Beer Consumer Loyalty. Journal of International Food and Agribusiness Marketing.

Hamdan, I.Z.P., Othman, M. (2022). Predicting Customer Loyalty Using Machine Learning for Hotel Industry. Journal of Soft Computing and Data Mining.

Hanifin, B. (2019). What Artificial Intelligence Means For Customer Loyalty Marketing. Forbs.

Jacoby, J., Kyner, D.B. (1973). Brand loyalty vs repeat purchasing behaviour. Journal of Marketing Research.

Jensen, J.M., Hansen, T. (2006). An empirical examination of brand loyalty. Journal of Product & Brand Management.

Jin, B.E., Shin, D.C. (2020). Changing the game to compete: Innovations in the fashion retail industry from the disruptive business model. Elsevier.

Krishnamurthi, L., Raj, S.P. (1991). An empirical analysis of the relationship between brand loyalty and consumer price elasticity. Marketing Science.

Kumar, V., Ayodeji, O. G. (2021). E-retail factors for customer activation and retention: An empirical study from Indian e-commerce customers. Journal of Retailing and Consumer Services.

Lemley, J., Bazrafkan, S., Corcoran, P. (2017). Deep learning for consumer devices and services: pushing the limits for machine learning, artificial intelligence, and computer vision. IEEE Consumer Electronics Magazine.

Lenzing. (2019). Demand share of apparel market, by region worldwide 2005-2020 [Graph]. Statista.

Lin, C. T., Chen, C. W., Wang, S. J., Lin, C. C. (2018). The influence of impulse buying toward consumer loyalty in online shopping: a regulatory focus theory perspective. Journal of Ambient Intelligence and Humanized Computing.

Marchetti, F. (2019). How A.I. is shaping fashion. CNBC.

McDowell, C. (2000). Fashion Today. Phaidon.

McMullan, R. (2005). A multiple-item scale for measuring customer loyalty development. Journal of Services Marketing.

Moedjiono, S., Isak, Y.R., Kusdaryono, A. (2016). Customer Loyalty Prediction In Multimedia Service Provider Company With K-Means Segmentation And C4.5 Algorithm). International Conference on Informatics and Computing (ICIC).

Mohammadi, M., Al-Fuqaha, A. (2018). Enabling cognitive smart cities using big data and machine learning: approaches and challenges. IEEE Communications Magazine.

Mohammadi, S.O., Kalhor, A. (2021). Smart Fashion: A Review of AI Applications in Virtual Try-On & Fashion Synthesis. Journal of Artificial Intelligence and Capsule Networks.

Mohammadi, S.O., Kalhor, A. (2022). Smart Fashion: A Review of AI Applications in the Fashion & Apparel Industry. Journal of Artificial Intelligence and Capsule Networks.

Muttaqien, R., Pradana, M.G., Pramuntadi A. (2021). Implementation of Data Mining Using C4.5 Algorithm for Predicting Customer Loyalty of PT. Pegadaian (Persero) Pati Area office. International Journal of Computer and Information System (IJCIS).

Neal, W. D. (1999). Satisfaction is nice, but value drives loyalty. Marketing Research.

OECD 2015. Data-driven Innovation: Big Data for Growth and Well-being. OECD Publishing.

OECD 2019. Going Digital: Shaping Policies, Improving Lives. OECD Publishing.

Oladapo, K. A., Omotosho, O. J., Adeduro, O. A. (2018). Predictive Analytics for Increased Loyalty and Customer Retention in Telecommunication Industry. International Journal of Computer Applications.

Oliver, R. L. (1999). Whence Consumer Loyalty? Journal of Marketing.

Reichheld, F. (1993). Loyalty-Based Management. Harvard Business Review.

Saibaba, S. (2023). Customer loyalty in e-commerce: a review and bibliometric analysis. Korea review of international studies.

Shoemaker, S., Lewis, R.C. (1999). Customer loyalty: the future of hospitality marketing. International Journal of Hospitality Management.

Sulistiani, H., Muludi, K., Syarif, A. (2019). Implementation of Dynamic Mutual Information and Support Vector Machine for Customer Loyalty Classification. Journal of Physics: Conference Series.

Taherkhani, N., Pierre, S. (2016). Centralized and localized data congestion control strategy for vehicular ad hoc networks using a machine learning clustering algorithm. IEEE Transactions on Intelligent Transportation Systems.

Taroy, D. (2015). How Instagram is Democratizing Fashion. Fast Company.

Tepeci, M. (1999). Increasing brand loyalty in the hospitality industry. International Journal of Contemporary Hospitality Management.

The Business of Fashion, & McKinsey. (2023). Worldwide forecasted sales growth in the fashion industry in 2023, by region [Graph]. Statista.

Tortora, P. G. (2015). Dress, Fashion and Technology – From Prehistory to the Present. Bloomsbury Academic.

Wassouf, W.N., Alkhatib, R., Salloum, K., Balloul S. (2020). Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study. Journal of Big Data.

Yang, Z., Peterson, R. T. (2004). Customer perceived value, satisfaction, and loyalty: The role of switching costs. Psychology and Marketing.

Zeithaml, V.A., Berry, L.L., Parasuraman, A. (1996). The behavioural consequences of service quality. Journal of Marketing.

Zulaikha, S., Mohamed, H., Kurniawati M., Rusgianto S., Rusmita S.A. (2020). Customer predictive analytics using artificial intelligence. The Singapore Economic Review.

360iResearch. (2022). AI in Fashion Market Research Report by Product Type, Component, Deployment, Application, End User, Region - Global Forecast to 2027 - Cumulative Impact of COVID-19. ReportLinker.

Extended Abstract

Nonostante la fidelizzazione dei clienti nel settore della moda è molto rilevante, le aziende spesso si trovano di fronte alla sfida di comprendere i consumatori, prevedere le loro esigenze e preferenze in modo accurato e tempestivo. Le tradizionali tecniche di analisi dei dati potrebbero ormai non essere più sufficienti per affrontare la complessità e la vastità delle informazioni disponibili. Pertanto, sorge la necessità di esplorare l'applicazione di algoritmi di intelligenza artificiale per l'analisi predittiva dei dati dei clienti al fine di massimizzare l'efficacia delle iniziative di fidelizzazione in questo settore.

Nel contesto dinamico e altamente competitivo dell'industria della moda, la fidelizzazione dei clienti è diventata una priorità strategica per tutte le aziende. Con l'aumento della concorrenza e delle opzioni di acquisto disponibili per i consumatori, è fondamentale per le aziende sviluppare strategie efficaci per mantenere e aumentare la fedeltà dei clienti.

L'obiettivo di questo studio è quello di esplorare l'applicazione dell'analisi predittiva con algoritmi di intelligenza artificiale per migliorare la fidelizzazione dei clienti nel settore della moda. Si intende valutare l'efficacia di tali algoritmi nel prevedere il comportamento dei clienti, consentendo alle aziende di anticipare le loro esigenze, le preferenze e i comportamenti di acquisto.

In particolare, si mira a selezionare le caratteristiche rilevanti dei clienti e ad addestrare i modelli predittivi in grado di classificare i clienti in base al loro stato di fedeltà. L'obiettivo è anche quello di fornire delle raccomandazioni pratiche alle aziende e ai *manager* per l'implementazione dei sistemi di analisi predittiva basati sull'intelligenza artificiale. Attraverso l'utilizzo di dati storici e attuali sui clienti, questi algoritmi possono identificare i modelli e le correlazioni significative che permettono di fare delle previsioni accurate sulle azioni future dei clienti. Ad esempio, è possibile identificare quali prodotti o servizi potrebbero interessare maggiormente a un determinato cliente in base ai suoi acquisti precedenti, alle sue interazioni sui canali digitali o alle sue preferenze. Ciò consentirà alle aziende di personalizzare l'offerta e le comunicazioni in modo mirato, offrendo prodotti e servizi rilevanti e aumentando così le probabilità di fidelizzazione del cliente. Inoltre, l'analisi predittiva può anche contribuire a individuare segnali di allarme che indicano un potenziale rischio di *churn*, consentendo alle aziende di adottare misure preventive per mantenere la fedeltà dei clienti. Attraverso questo studio, inoltre, si intende contribuire alla comprensione delle potenzialità dell'analisi predittiva e dell'intelligenza artificiale nell'ambito della fidelizzazione dei clienti nel settore della moda. I risultati e le conclusioni dello studio potranno fornire delle indicazioni preziose per le decisioni strategiche delle aziende, aiutandole ad ottimizzare le loro risorse e ad adattarsi alle esigenze mutevoli dei clienti, promuovendo relazioni durature e proficue con la clientela.

Il gap di ricerca identificato in questa tesi riguarda la mancanza di studi precedentemente condotti che indaghino specificamente sull'applicazione del *machine learning* per la classificazione della *customer loyalty* nel settore della moda. Come emergerà successivamente si rileva una carenza di ricerche che abbiano esaminato in modo esaustivo e approfondito l'utilizzo di tali approcci predittivi in questo ambito. Pertanto, si

può affermare che esiste un vuoto di conoscenza riguardo all'efficacia e alle potenzialità dell'applicazione di tali algoritmi nell'industria della moda.
Questo gap di ricerca è rilevante perché questo settore è caratterizzato da una rapida evoluzione delle tendenze e dei gusti dei consumatori e le aziende sono costantemente alla ricerca di nuovi modi per comprendere al meglio i desideri e le preferenze dei loro clienti al fine di offrire esperienze personalizzate e costruire relazioni solide e durature.

La fedeltà è una tematica rilevante nell'ambito del marketing e la rilevanza del concetto è determinata dai vantaggi legati alla conservazione dei clienti (McMullan, R., 2005). Le ricerche effettuate sulla *customer loyalty* hanno da tempo messo in risalto la rilevanza dei clienti fedeli in quanto tendono a spendere di più rispetto agli occasionali o ai nuovi. Questo è dovuto al fatto che essi hanno una maggiore fiducia nella marca e nei suoi prodotti, il che li spinge ad acquistare di più; essi sono anche più propensi ad acquistare una maggiore varietà di prodotti rispetto ai clienti occasionali o ai nuovi. Questo è dovuto al fatto che hanno già sperimentato la qualità dei prodotti e sono convinti della loro efficacia, sono anche disposti a pagare un prezzo più alto per i prodotti del *brand* ed infine, sono spesso i migliori ambasciatori. Questi clienti sono molto soddisfatti dei prodotti dell'impresa e tendono a raccomandarli ad amici, familiari e conoscenti, generando un passaparola positivo per l'impresa (Zeithaml V.A., et al., 1996). Jacoby e Kyner (1973), hanno dato un contributo importante allo studio della *customer loyalty* e la loro definizione rappresenta uno dei concetti maggiormente condivisi e noti in letteratura che la definisce come "una risposta comportamentale, premeditata, espressa nel tempo da un'unità decisionale di acquisto, rispetto a una o più marche alternative, dipendente da un processo psicologico". L'attenzione rivolta alla *customer loyalty* deriva dall'importanza che tale fattore riveste nella gestione delle relazioni tra l'impresa e il cliente. Il consumatore cerca di stabilire un rapporto di fiducia con l'azienda presso cui effettua i propri acquisti, in modo da ridurre i costi legati alla ricerca di informazioni sui prodotti e il rischio di dover acquistare da un'azienda sconosciuta. In questo modo, si riducono i tempi di valutazione e si minimizza il rischio associato alla scelta di un nuovo fornitore e questi sono solo alcuni dei vantaggi per il consumatore. Tuttavia, il comportamento d'acquisto ripetuto è solo un prerequisito per la fedeltà, che non può essere garantita se la fedeltà stessa è il risultato di comportamenti inerziali o di una mancanza di alternative valide (Jacoby, J. et al., 1973).
Il cambiamento in corso causato dal *machine learning* ha gettato le basi per aumentare il valore dei dati dei clienti ed anticipare i modelli di comportamento degli acquirenti, al fine di prevedere le future decisioni di acquisto e creare un'esperienza personalizzata. Se i rivenditori e i *brand* investono nel *machine Learning*, i benefici si estenderanno ben oltre le prestazioni aziendali. Nel frattempo, gli esperti di programmi di fidelizzazione dei clienti prevedono che, grazie all'utilizzo di strumenti di marketing automatizzati, il servizio clienti sarà elevato e il *customer journey* sarà ottimizzato. Data la natura dinamica del *machine learning*, gli addetti al marketing si imbatteranno nelle seguenti sfide per passare dal tradizionale modello di database dei clienti a un CRM avanzato, alimentato da soluzioni IA e quindi gestire la raccolta dei dati dei clienti attraverso

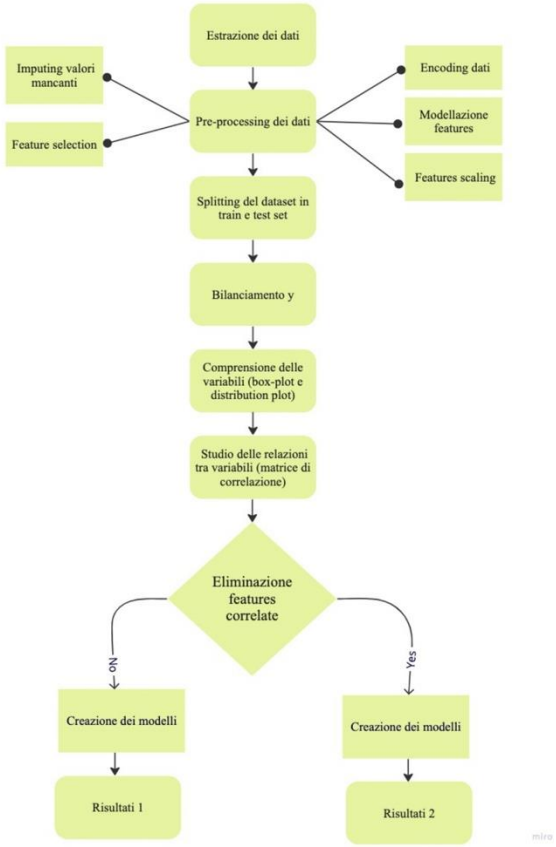
tutti i punti di contatto in tempo reale, analizzare i preziosi *insight* dei clienti, le attività comportamentali e transazionali per trasformarli in conoscenze di *business*, mantenere e migliorare le prestazioni dell'algoritmo e gestire allo stesso tempo enormi volumi di dati sui clienti ed infine, applicare la potenza del *machine learning* ai programmi di fidelizzazione dei clienti. Gli strumenti di automazione forniscono una conoscenza più approfondita e precisa dei dati demografici, delle preferenze e di altre importanti informazioni sui clienti. Pertanto, è di grande interesse per i rivenditori e i marchi conquistare le conoscenze di cui sopra e riuscire finalmente ad ottimizzare le campagne di marketing, modificare la messaggistica di comunicazione, quando è necessario, e fornire contenuti mirati più velocemente e in modo più accurato. Inoltre, gli strumenti di *machine learning* sono in grado di prevedere le frodi nei programmi di fidelizzazione, rilevando incongruenze e attività anomale. Il *machine learning* svolge quindi un ruolo significativo nella fidelizzazione del marchio, aiuta a creare un *customer journey* che rassicuri la fidelizzazione dei clienti esistenti e l'acquisizione di nuovi, aiuta ad eseguire campagne di fidelizzazione di successo sfruttando i dati dei clienti per sbloccare il loro comportamento. La fornitura di un servizio clienti olistico su tutti i diversi canali può incrementare KPI significativi: tassi di conversione/ritenzione, visite al negozio e carrello medio. Riconoscendo l'importanza delle decisioni basate sui dati, gli strumenti di *machine learning* non solo aiuteranno i *marketer* a guadagnare più tempo per pensare ad alto livello, ma anche a trasformare il tradizionale rapporto dare-avere in un percorso di coinvolgimento dei clienti di successo (Siti Zulaikha, S. et al., 2020).

Il concetto di *customer loyalty* ha suscitato l'interesse di numerose ricerche negli ultimi anni, portando alla realizzazione di numerosi studi. Sono state selezionate e discusse diverse ricerche esistenti per ottenere maggiori informazioni che saranno utilizzate per condurre la ricerca proposta. I *paper* sono stati suddivisi in base al contesto di applicazione per fornire una panoramica completa delle ricerche esistenti su questo argomento. In particolare, sono stati identificati cinque ambiti principali di utilizzo di questi algoritmi per la classificazione della *customer loyalty*: (1) il settore di credito, (2) quello delle telecomunicazioni, (3) quello alberghiero, (4) del fast moving consumer goods ed infine (5) quello dei servizi multimediali. In ogni *paper* analizzato, si sono riscontrati differenti algoritmi di *machine learning* impiegati per raggiungere l'obiettivo di classificare la *customer loyalty*. L'analisi degli studi esistenti mostra che, in ambito accademico, la ricerca sulla *customer loyalty* è un tema importante nella gestione delle relazioni con i clienti e i diversi *paper* analizzati hanno evidenziato la validità degli algoritmi di *machine learning* per la classificazione della *customer loyalty* in diversi ambiti di applicazione.

I dati utilizzati per la presente tesi fanno riferimento a due dataset "Transaction fashion brand" e "Contact Active". Entrambi i dataset utilizzati riguardano il periodo che va dal 2015 al 2022 e presentano informazioni di grande interesse per l'analisi. Il primo *dataset*, "Transaction fashion brand", si riferisce alle transazioni effettuate dai clienti di un noto *brand* di moda e si compone di un elevato numero di osservazioni (1.021.739) e 65 features principalmente in formato categorico (es. uomo/donna). La vasta quantità di informazioni presenti in questo *dataset* permette di analizzare in modo dettagliato i comportamenti degli utenti nei confronti del *brand*, e di individuare eventuali *pattern* o tendenze che possano essere utili per la pianificazione di future

strategie di marketing. In particolare, l'analisi di questo *dataset* ha permesso di individuare alcune tendenze di acquisto tra i clienti del *brand*, come ad esempio i prodotti più venduti, le categorie di prodotti preferite e il valore medio degli acquisti. Il secondo *dataset*, "Contact Active", riguarda invece informazioni personali degli utenti e anch'esso presenta un elevato numero di osservazioni (384.352) e 135 features in formato categorico. Questo *dataset* consente di analizzare le caratteristiche degli utenti che hanno interagito con il *brand*, e di individuare eventuali fattori determinanti per l'acquisizione di nuovi clienti o per la fidelizzazione di quelli già esistenti come l'area in cui effettuano gli acquisti, le preferenze di acquisto dei clienti e le loro preferenze riguardo ai trattamenti dei dati. Inoltre, all'interno di questo *dataset* è presente la *lable* (y) denominata "Loyalty". La scelta di utilizzare questi due *dataset* è stata motivata dall'obiettivo di fornire una panoramica completa e dettagliata del contesto in cui si muove il *brand* analizzato, in modo da identificare eventuali criticità ed individuare eventuali strategie di miglioramento. La presenza di due *dataset* distinti ha permesso inoltre di integrare informazioni diverse, offrendo una visione a 360 gradi del contesto in cui si muovono i clienti del *brand*. La successiva unione dei due *dataset* mediante la colonna in comune "Customer ID" ha rappresentato un passaggio fondamentale per l'analisi dei dati. Grazie a questa operazione, infatti, è stato possibile integrare le informazioni relative alle transazioni effettuate dai clienti con quelle relative alle loro caratteristiche personali, consentendo di ottenere un quadro ancora più completo e preciso della situazione analizzata.

Il seguente *flowchart* è stato inserito per facilitare la comprensione della metodologia di *data analysis* implementata.



Come si evince dalla figura dopo aver esportato i dati è stato effettuato il *pre – processing* che consiste nel: (1) valutare la distribuzione dei valori di ciascuna variabile per poter calcolare la quantità di valori mancanti e per poterli gestire è stata stabilita una soglia di accettabilità del 50%. Se la percentuale di valori mancanti fosse stata superiore, la variabile verrebbe eliminata dal *dataset* al fine di consentire l'esecuzione del processo di *imputing* assegnando il valore più frequente alle *features* categoriche e la media dei valori a quelle numeriche. (2) In aggiunta sono state rimosse anche altre che non erano rilevanti per l'analisi. Dopo aver ridotto il numero di colonne, è stato possibile (3) applicare la tecnica del Label Encoder per *l'encoding* delle *features* categoriche del *dataset* Contact Active. (4) Alcune colonne del *dataset* "Transaction fashion brand" hanno subito una modellazione al fine di rendere l'analisi più lineare. A seguire (5) è stata implementata la tecnica di StandardScaler per effettuare la scalatura dei dati, in modo che essi siano distribuiti attorno a una media pari zero e con una deviazione standard pari ad uno. Tale operazione è necessaria per evitare che le variabili con valori molto grandi abbiano un peso maggiore rispetto a quelle con valori più piccoli durante la fase di addestramento del modello.

La suddivisione del *dataset* rappresenta una fase essenziale del processo di sviluppo di un modello di apprendimento automatico. La funzione train_test_split è stata impiegata per separare il *dataset* in due parti disgiunte: il *train set*, utilizzato per l'addestramento del modello, e il *test set*, utilizzato per valutare le prestazioni del modello su nuovi dati. Nel caso specifico, è stata scelta una dimensione del *test set* pari al 20% del *dataset* originale (test_size=0.20). Tale scelta rappresenta una pratica comune per ottenere un equilibrio adeguato tra le dimensioni del *train set* e del *test set*. Di conseguenza, l'80% rimanente del *dataset* è stato utilizzato per l'addestramento del modello, garantendo una buona capacità di generalizzazione del modello su nuovi dati.

La variabile *target* "Loyalty" contenuta nel *dataset* Contact Active presentava numerose categorie. Poiché era necessario condurre una classificazione binaria, è stata effettuata una selezione delle categorie da assegnare alla classe "Leale" e quelle da assegnare alla classe "Infedele". Le categorie attribuite alla classe "Infedele" sono state Lost, Inactive, Prospect Store, Sleeper, New, Occasional Reactivated e Occasional Retained, mentre le categorie Loyal Retained, New Loyal e Loyal Reactivated sono state assegnate alla classe "Leale".

Come precedentemente menzionato, attraverso l'utilizzo del LabelEncoder, è stato attribuito il valore 0 alla classe negativa (ovvero l'assenza della caratteristica) e il valore 1 alla classe positiva (ovvero la presenza della caratteristica). La variabile è stata analizzata tramite l'utilizzo della funzione sns.countplot, la quale consente di contare il numero di osservazioni per categoria e di visualizzarli mediante un grafico a barre. Attraverso questa analisi è stato possibile notare che la *feature* presenta un forte sbilanciamento tra le due categorie. È stato necessario procedere al bilanciamento della y, eseguito unicamente sulla porzione di *training* del *dataset*. Bilanciare la variabile *target* solo sulla parte del *training* è una buona pratica per evitare di introdurre informazioni del *test* nel processo di apprendimento del modello. Il bilanciamento è stato effettuato tramite la

tecnica SMOTEENN (Synthetic Minority Over-sampling Technique Edited Nearest Neighbors) sviluppata da Batista et al. (2004).

Questa tecnica combina SMOTE e Edited Nearest Neighbours (ENN) ed esegue contemporaneamente un *upsampling* e un *downsampling*. L'obiettivo di SMOTE è generare nuovi campioni sintetici della classe minoritaria, aumentando così il numero di esempi per quella classe. Tuttavia, questa tecnica potrebbe portare alla generazione di campioni che non sono rappresentativi della classe minoritaria, ma che sono estremamente simili ad esempi già esistenti. L'aggiunta di ENN, invece, serve a rimuovere i campioni indesiderati che potrebbero essere stati generati da SMOTE.

Successivamente è stata eseguita l'analisi statistica esplorativa dei dati che nel contesto del *machine learning* si riferisce alla fase del processo in cui i dati di *input* vengono esplorati e compresi al fine di identificare eventuali problemi, comprendere la distribuzione delle variabili ed individuare eventuali relazioni tra esse. Questa fase è essenziale per garantire la qualità dei dati e la validità dei modelli di *machine learning* che verranno creati e valutati successivamente. L'obiettivo delle analisi statistiche effettuate (*box-plot* e *distribution plot*) è stato quello di visualizzare i dati presenti all'interno delle *features* e la loro distribuzione. Era importante anche capire se fosse stato il caso di eliminare eventuali *outliers;* tuttavia, nei tre casi specifici in cui sono stati rilevati gli *outliers*, è stato deciso di non eliminarli in quanto ciò avrebbe comportato la perdita di informazioni significative per l'analisi. In seguito, si è proceduto all'analisi della matrice di correlazione. Lo scopo della seguente matrice è stato quello di studiare e capire la relazione tra la variabile dipendente y e le variabili indipendenti x, nonché di identificare eventuali *features* ridondanti, infatti, grazie all'analisi è stato possibile eliminare le seguenti caratteristiche: "Consent_for_Analysis", "Consent_for_SMS", "Consent_for_Newsletter", "Total_Amount_Base", "Quantity" al fine di valutare un possibile miglioramento delle *performance* degli algoritmi.

In accordo con quanto riportato nella revisione della letteratura riguardante la parte degli algoritmi, è stato deciso di implementare nello studio quelli che hanno dimostrato le prestazioni più elevate, al fine di valutare se, in questo contesto, avessero lo stesso impatto. L'unica differenza è che l'algoritmo C4.5 non trova implementazione all'interno della libreria scikit-learn, ma al suo posto vi è l'algoritmo CART, che presenta notevoli somiglianze con il primo. In particolare, i modelli di classificazione adottati per lo studio, a cui dedicheremo particolare attenzione, sono il Random Forest, Logistic Regression, Decision Tree, Naïve Bayes, K-neighbors e XGBoost Classifier.

Lo scopo principale è stato valutare l'efficacia di questi algoritmi nel prevedere il comportamento dei clienti, attraverso le metriche, Confusion Matrix e Curva ROC - AUC e successivamente fornire raccomandazioni pratiche per l'implementazione di sistemi basati sull'intelligenza artificiale per massimizzare l'efficacia dei programmi di fidelizzazione dei clienti e ottenere un vantaggio competitivo nel mercato.

Tutti i modelli predittivi hanno ottenuto risultati incoraggianti e nessuno si è distinto chiaramente poiché tutti hanno ottenuto risultati simili.

1. *Accuracy*: I modelli hanno raggiunto un'accuratezza elevata, con valori che variano tra lo 0.87 e lo 0.97. Ciò indica che i modelli hanno una buona capacità di classificare correttamente le istanze dei dati nel contesto della *customer loyalty*.
2. *Precision*: Tutti i modelli hanno mostrato una *precision* elevata, con valori parti allo 0.97. In questo caso, i modelli sono in grado di identificare correttamente la *customer loyalty* con una percentuale molto alta.
3. *Recall*: I modelli hanno ottenuto valori compresi tra lo 0.88 e lo 0.95, indicando anche in questo caso una buona capacità di rilevare correttamente la *customer loyalty*.
4. *F1-score*: Tutti i modelli hanno ottenuto valori che si aggirano tra lo 0.91 e lo 0.96 e ciò sta ad indicare un buon equilibrio tra la capacità di identificare correttamente la *customer loyalty* e di minimizzare i falsi positivi.

I risultati ottenuti indicano che tutti i modelli hanno dimostrato una buona capacità di classificazione della *customer loyalty* nell'industria del *fashion*. Il motivo per cui potrebbe non emergere un metodo nettamente migliore potrebbe dipendere dalla natura del problema e delle caratteristiche dei dati utilizzati.
Il fatto che i risultati siano simili può indicare che i modelli stiano approssimando bene la relazione tra le *features* e la *customer loyalty*. Date le buone *performance* ottenute dei modelli di *machine learning* utilizzati è stato ritenuto interessante andare a vedere quali sono state le *features* più rilevanti per il seguente *task*. L'importanza della caratteristica indica quanto ciascuna contribuisca alla previsione del modello e viene rappresentata utilizzando un valore numerico chiamato punteggio, più è alto questo valore più la *feature* è rilevante per l'analisi. L'utilizzo della *feature importance* è utile poiché ne accelera il funzionamento aiutando a capire quali caratteristiche sono irrilevanti per il modello e potrebbe migliorarne le prestazioni. Lo studio, come già precedentemente detto, mira a trovare le migliori tecniche di *machine learning* per la classificazione della *customer loyalty* tra una serie di algoritmi ben accettati e facili da implementare, scoprendo che, almeno per questo *set* di dati, hanno dato buoni risultati. Si tratta di una fase iniziale dell'utilizzo di approcci di *machine learning* in questo contesto ed un possibile sviluppo futuro potrebbe essere quello di testare se riducendo le caratteristiche si potrebbero raggiungere delle migliori prestazioni predittive.

Nell'attuale ambiente competitivo e omnicanale, in cui i costi di cambiamento sono bassi e i clienti possono confrontare in modo più trasparente le offerte e i livelli di prezzo dei rivenditori, la fedeltà dei clienti sembra sempre più difficile da raggiungere e nel contesto della vendita al dettaglio, in cui i clienti acquistano ripetutamente, comprendere la fedeltà dei clienti e come influenzarla è essenziale. Incrementare la *customer loyalty* rappresenta quindi un obiettivo cruciale per le aziende, poiché una clientela fedele può generare maggiori profitti e vantaggi competitivi. Tuttavia, identificare i clienti fedeli non è un compito facile e le aziende spesso si affidano a metriche e indicatori per monitorare la *customer loyalty*. In questo contesto, la classificazione binaria della *customer loyalty* rappresenta un'importante metodologia per suddividere i clienti

in due gruppi: fedeli e infedeli e tale modello ha implicazioni significative per le strategie di fidelizzazione dei clienti adottate dalle aziende e per le decisioni manageriali riguardanti la gestione del rapporto con la clientela. Le aziende possono creare strategie di marketing mirate e personalizzate in grado di soddisfare le esigenze dei clienti più fedeli, fidelizzandoli ulteriormente e aumentando la probabilità di acquisto da parte loro. In questo paragrafo, analizzeremo le implicazioni manageriali della classificazione della *customer loyalty*, discutendo le possibili azioni che le aziende possono intraprendere.

I consumatori moderni sono sommersi dal marketing e dalla messaggistica dei vari *brand* e di conseguenza, sono diventati sempre più esigenti riguardo ai messaggi con cui interagire. Quando si utilizza una strategia basata sui dati, i *team* di marketing possono aumentare drasticamente le possibilità che il loro pubblico di destinazione faccia *clic* sul loro annuncio, legga un *post* sul *blog* o esegua un'altra azione che guida un obiettivo di conversione. Le strategie basate sui dati migliorano l'esperienza del cliente, la percezione del marchio e la fedeltà dei clienti in quanto offrono alle organizzazioni una comprensione delle esigenze e degli interessi dei consumatori, migliorano anche i tassi di conversione perché è più probabile che la messaggistica altamente mirata abilitata da questo tipo di marketing attiri l'attenzione degli utenti ed infatti queste strategie si concentrano sull'utilizzo di profili approfonditi dei consumatori per migliorare la loro esperienza.

Un elemento fondamentale per i professionisti del marketing è determinare dove viene sprecato il loro *budget* pubblicitario. Il *data-driven* marketing, guidato da strumenti di analisi, consente ai *team* di marketing di scoprire quale parte del *budget* pubblicitario sta avendo il maggiore impatto sulle conversioni o sulla consapevolezza del marchio e questo viene fatto valutando i percorsi dei clienti utilizzando modelli di attribuzione, come la misurazione del marketing unificato che esamina l'attribuzione *multitouch* e la modellazione del *media mix* per fornire una visione completa del percorso di acquisto. Le organizzazioni possono determinare cosa sposta i potenziali clienti e i clienti lungo la canalizzazione e quindi allocare i fondi di conseguenza.

Valutare i dati dei consumatori offre ai *team* di marketing informazioni sui tipi di creatività, su elementi visivi, i testi e i contenuti con cui il pubblico di destinazione preferisce interagire. Trasmettere il messaggio giusto, che soddisfi gli interessi personali e crei valore, al momento giusto è essenziale per entrare in contatto con i consumatori. Tale approccio può condurre a decisioni più efficaci riguardo ai prodotti e può fornire un'analisi approfondita dei clienti per l'azienda.

Questo tipo di strategie sono positive sia per i *marketer* che per i consumatori. Tuttavia, ci sono alcune sfide che possono impedire ai professionisti del marketing di trarre tutti i vantaggi dai dati o di raggiungere i clienti in modo efficace. In prima istanza i *brand* dovrebbero evitare di essere troppo invasivi; nonostante il desiderio dei consumatori di fruire di esperienze personalizzate, vi è una reticenza nei confronti delle organizzazioni che acquisiscono una conoscenza eccessiva su di loro. In particolare, se i consumatori scelgono di condividere informazioni personali, desiderano avere chiarezza sul modo in cui tali dati saranno utilizzati a loro beneficio. Un ulteriore aspetto da considerare riguarda la qualità insufficiente dei dati; al fine di implementare una strategia *data-driven* efficiente, è necessario avere adeguati processi di gestione dei dati che garantirà la

possibilità di prendere decisioni e sviluppare delle strategie basate sui dati di alta qualità che riflettano accuratamente le necessità dei clienti. Nel caso in cui i dati a disposizione non soddisfino criteri di qualità quali tempestività, accuratezza, completezza e rappresentatività, vi è il rischio di basare le decisioni su dati che offrono una visibilità limitata sulle reali esigenze dei clienti.

Il marketing incentrato sui dati mira a potenziare l'efficacia del marketing attraverso un'ottimizzazione dell'esperienza del cliente che riveste un ruolo chiave in questa dinamica. Pertanto, ogni campagna sviluppata con l'ausilio dei dati dovrebbe fornire una chiara dimostrazione dei benefici che il cliente può ottenere.

La predizione del comportamento dei clienti in relazione alla loro fedeltà rappresenta un importante strumento competitivo per le aziende, in grado di garantire una maggiore efficacia delle strategie di marketing e di fidelizzazione. La classificazione dei clienti in base alla loro fedeltà rappresenta la prima fase di questo processo ma la previsione del loro comportamento futuro è l'elemento chiave per adattare le strategie di marketing e di *retention* alle loro esigenze. Ciò, infatti, consente di concentrare le risorse sui clienti più importanti e di offrire loro un'esperienza di acquisto personalizzata, che può portare a un aumento della fedeltà e, di conseguenza, della redditività del marchio.

Questa capacità predittiva del comportamento dei clienti conferisce un vantaggio competitivo alle imprese, consentendo loro di (1) adottare misure preventive per ridurre il *churn* dei clienti; tale defezione rappresenta un costo significativo per le aziende, sia in termini di perdita di fatturato che di risorse impiegate per rimpiazzare i clienti persi. Attraverso l'analisi predittiva, gli algoritmi di classificazione possono individuare dei segnali di imminente defezione, identificando i clienti che potrebbero essere più propensi ad abbandonare l'azienda e questo consente ai dirigenti di adottare misure preventive mirate per mantenere i clienti a rischio. Ad esempio, possono essere implementate strategie di fidelizzazione personalizzate, come offerte speciali, assistenza dedicata, programmi di fedeltà o miglioramenti del servizio. Ridurre il tasso di defezione dei clienti è fondamentale per mantenere una base di clientela stabile e sostenibile nel lungo termine.

(2) Identificare le opportunità di *cross-selling* e *up-selling* in base alle preferenze individuali dei clienti; gli algoritmi di classificazione possono individuare modelli e associazioni tra i diversi prodotti o servizi che i clienti tendono ad acquistare o utilizzare insieme. Questo consente ai *manager* di identificare opportunità di *cross-selling*, che consistono nel proporre ai clienti prodotti o servizi complementari a quelli che hanno già acquistato. Ad esempio, se un cliente ha acquistato un computer portatile, l'azienda potrebbe suggerire l'acquisto di una borsa per il trasporto o di un *software* aggiuntivo. Inoltre, l'analisi predittiva può rivelare anche opportunità di *up-selling*, che consistono nel proporre ai clienti prodotti o servizi di fascia superiore o con maggiori funzionalità rispetto a quelli che hanno giò considerato. Questo può essere basato sulla segmentazione dei clienti e sulla comprensione delle loro caratteristiche e preferenze specifiche.

L'identificazione delle opportunità di *cross-selling* e *up-selling* consente alle aziende di massimizzare il valore dei propri clienti esistenti offrendo loro prodotti o servizi aggiuntivi che sono rilevanti e di interesse e quindi aumentano il valore medio delle transazioni e stimolano ulteriori acquisti che si traduce in un incremento delle entrate e della redditività complessiva dell'azienda. Inoltre, favorisce l'esperienza del cliente, poiché gli viene

presentata una proposta su misura, rispondendo alle sue esigenze specifiche e migliorando la sua soddisfazione complessiva.

(3) Ottimizzazione dell'allocazione delle risorse di marketing; le aziende dedicano una parte significativa del loro *budget* alle attività di marketing, come pubblicità, promozioni, eventi e altro ancora. Tuttavia, l'allocazione di queste risorse non può essere casuale o basata solo su supposizioni. È fondamentale investire le risorse di marketing in modo mirato ed efficiente per massimizzare il ritorno sull'investimento infatti, la previsione accurata del comportamento dei clienti consente alle aziende di ottenere una comprensione approfondita di come i clienti risponderanno alle diverse iniziative di marketing. Gli algoritmi di intelligenza artificiale possono analizzare dati storici, modelli di acquisto, preferenze e altre variabili per prevedere quali clienti sono più inclini a rispondere positivamente a una determinata campagna o offerta.

Utilizzando queste previsioni, le aziende possono ottimizzare l'allocazione delle risorse di marketing in modo da concentrare gli sforzi su quei segmenti di clientela che hanno maggiori probabilità di rispondere positivamente. Questa ottimizzazione dell'allocazione delle risorse porta a diversi vantaggi: innanzitutto, consente di evitare sprechi di risorse su segmenti che hanno scarse probabilità di risposta positiva; in secondo luogo, consente di massimizzare l'impatto delle iniziative di marketing, concentrandosi su clienti con una maggiore probabilità di conversione ed infine, aiuta a migliorare l'efficienza complessiva delle attività di marketing, consentendo un utilizzo più strategico e mirato delle risorse disponibili.

(4) Migliorare il *customer journey*, ovvero l'intero percorso che un cliente compie dal momento in cui scopre un prodotto o servizio fino all'acquisto e oltre. Identificando i punti critici del percorso del cliente e prevedendo le loro azioni e preferenze, le aziende possono personalizzare l'esperienza del cliente in modo più efficace.

L'emergere della pandemia da COVID-19 ha avuto un impatto significativo sul modo di comprare e sul comportamento dei consumatori, accelerando l'adozione e il potenziamento dell'utilizzo dei *big data* e del *machine learning* per analizzare le abitudini di consumo. Con l'aumento delle restrizioni e la necessità di distanziamento sociale, i consumatori si sono rivolti sempre di più agli acquisti *online* per soddisfare le loro esigenze e questo ha creato una crescente quantità di dati digitali, che servono per comprendere le preferenze dei consumatori e personalizzare l'esperienza di acquisto. Dato che i modelli di acquisto sono stati ridefiniti e molte aziende all'avanguardia hanno adottato strategie basate sui dati per anticipare le tendenze e offrire prodotti e servizi rilevanti. La pandemia ha anche spinto i consumatori ad essere più consapevoli delle loro scelte e a cercare prodotti e servizi che soddisfino criteri di sicurezza, sostenibilità e benessere. Gli acquirenti, infatti, cercano informazioni dettagliate sui prodotti, valutazioni dei clienti e *feedback* prima di prendere una decisione d'acquisto e in risposta a questa richiesta, i rivenditori e gli *e-commerce* devono sviluppare algoritmi di raccomandazione personalizzati basati sui *big data*, che suggeriscono prodotti pertinenti e migliorano l'esperienza d'acquisto. Nel panorama post-pandemico, l'uso dei dati giocherà un ruolo fondamentale nel soddisfare le aspettative dei consumatori e nel guidare il successo delle imprese.

Pertanto, risulta evidente come l'abilità di sfruttare il *machine learning* e l'analisi dei *big data* per anticipare il comportamento dei consumatori non comporti soltanto un aumento delle conversioni, obiettivo di per sé significativo ma implica altresì il potenziamento della *brand awareness* e la fidelizzazione dei clienti, generando l'immagine di un marchio in grado non solo di ascoltare, ma addirittura di suggerire soluzioni perfette. Considerando che le imprese si adoperano per mantenere la competitività in un mercato in costante evoluzione, la previsione accurata e tempestiva della domanda assume sempre maggiore importanza.

I modelli di *machine learning* possono essere costantemente aggiornati e ottimizzati, consentendo alle aziende di adeguare prontamente le proprie previsioni per rispecchiare le mutevoli condizioni di mercato e ciò permetterà alle aziende di ottenere previsioni più precise con una velocità ed efficienza senza precedenti.

Inoltre, questo approccio può essere applicato anche al marketing di prodotto, consentendo lo sviluppo di nuovi articoli o servizi sempre più personalizzati e in linea con le preferenze del target di riferimento.

Una strategia orientata al futuro che si traduce in risultati a lungo termine, grazie a ottimizzazioni continue che considerano l'evoluzione dei *trend* nel corso del tempo.

I vantaggi del marketing predittivo si riflettono chiaramente sulle vendite e sulla crescita, ma possono spingere le aziende ancora oltre. La capacità di anticipare con precisione le tendenze future può influenzare ogni aspetto del marketing aziendale.

Il marketing predittivo rappresenta il risultato finale e naturale delle strategie di marketing basate sui dati che sono state sviluppate nel tempo, partendo dalla definizione di una strategia dati complessiva e ben strutturata e dallo sviluppo delle competenze necessarie, sia a livello tecnico che manageriale. Per ottenere risultati effettivi, il marketing predittivo non può prescindere dalla mappatura dei percorsi dei clienti. L'identificazione di tutti i punti di contatto attraversati e della loro sequenza è fondamentale, poiché consente di valutare localmente e globalmente l'accessibilità, l'ergonomia e l'esperienza utente di ciascun punto di contatto e del flusso complessivo, nonché di raccogliere dati ad ogni *touchpoint*.

Il cambiamento del comportamento dei clienti è influenzato da tre aspettative che i consumatori hanno: personalizzazione, previsione e adattabilità e tali aspettative stanno promuovendo una tendenza che va verso l'instaurazione di una relazione intima con il cliente. Considerando che le aspettative di intimità sono destinate a crescere ulteriormente, le aziende devono concentrarsi sulla creazione di una connessione significativa e questo implica l'attribuzione di un ruolo centrale degli individui, l'accelerazione nell'adozione di analisi e intelligenza artificiale, nonché l'innovazione su vasta scala.

L'intimità con il cliente rappresenta una strategia volta a comprenderne le specifiche esigenze al fine di fornire la "migliore soluzione", ovvero prodotti o servizi adattati in modo continuativo alle situazioni e alle nicchie specifiche dei clienti. L'intimità è intesa come "vicinanza" e richiede una completa ristrutturazione dell'intera organizzazione per avere successo.

Al centro di questa tendenza, come già accennato, si trovano tre aspettative predominanti dei clienti:

1. Personalizzazione: i clienti si aspettano che ogni interazione con un'azienda sia personalizzata. Desiderano che le aziende conoscano la loro identità, le loro preferenze e il modo in cui desiderano essere serviti.
2. Previsione: i clienti si aspettano che, grazie alla loro relazione con un'azienda e alla dimostrazione delle loro preferenze e comportamenti d'acquisto, quest'ultima sia in grado di anticipare opportunità di prodotti o servizi che potrebbero risultare vantaggiosi e offrirli prima ancora che i clienti ne esprimano esplicitamente il desiderio.
3. Adattabilità: i clienti vivono oggi in un ambiente sempre più intelligente e si aspettano che le aziende dimostrino la stessa sensibilità nei confronti del loro ambiente e si adattino alle specifiche esigenze di quest'ultimo. Vogliono che le aziende siano consapevoli della loro posizione e dei loro desideri in termini di coinvolgimento.

La personalizzazione rappresenta una capacità essenziale che ha un impatto significativo, indipendentemente dalla natura dell'azienda, infatti, i consumatori non solo la desiderano ma la considerano una necessità imprescindibile. I vantaggi del marketing personalizzato sono molteplici, sia per le aziende che per i consumatori ma solo quando le strategie vengono implementate con successo, si possono ottenere risultati significativi.

La creazione di un'esperienza coinvolgente e rilevante per i clienti rappresenta uno degli obiettivi fondamentali della personalizzazione delle risorse, infatti, quando viene offerta un'esperienza personalizzata, i clienti percepiscono un coinvolgimento e un apprezzamento che aumentano il loro legame con il *brand* e la loro probabilità di soddisfazione. Questa esperienza coinvolgente si basa sull'interazione costante tra l'azienda e il cliente, che può avvenire attraverso diverse piattaforme come *siti web*, applicazioni o *social media*.

Le aziende possono sfruttare le informazioni raccolte dall'analisi predittiva per personalizzare l'esperienza del cliente su tali canali, ad esempio fornendo contenuti pertinenti, raccomandazioni personalizzate o promozioni speciali.

Quando i clienti si sentono coinvolti e ricevono un'esperienza rilevante, sono più inclini a sviluppare un legame emotivo con il *brand* e a rimanere fedeli nel tempo e le aziende che dimostrano di essere capaci di creare questa vicinanza con i clienti registrano tassi di crescita dei ricavi più rapidi rispetto ai loro concorrenti.

Con l'uso appropriato della tecnologia di automazione, i professionisti del marketing possono identificare il canale con cui i clienti interagiscono e automatizzare il *follow-up* su diversi canali come parte di un approccio omnicanale.

Inoltre, il marketing personalizzato favorisce l'aumento della fedeltà al marchio; quando i consumatori forniscono informazioni e dati, si aspettano di essere trattati come individui unici con preferenze specifiche e le aziende che dedicano tempo e risorse all'implementazione di strategie di marketing personalizzate di successo beneficeranno di un vantaggio competitivo sia in termini di fedeltà al marchio che di soddisfazione del cliente. Infine, questa tipologia di marketing aiuta a creare coerenza tra i vari canali. I consumatori

interagiscono con i marchi attraverso diversi canali come e-mail, social media, dispositivi mobili, ecc ed è quindi fondamentale che i marchi creino un'esperienza coerente tra questi ultimi. L'esperienza in negozio dovrebbe essere allineata con quella dell'*app*, che a sua volta dovrebbe essere coerente con le e-mail, così facendo i clienti possono riprendere la conversazione da dove l'hanno interrotta e indipendentemente dal canale utilizzato.

Non è mai stato un momento più opportuno per sfruttare la personalizzazione dato che attualmente, le persone desiderano esprimere la propria individualità e personalità attraverso tutto ciò che possiedono, inclusi abbigliamento, accessori e persino l'arredamento domestico. Inoltre, la personalizzazione è diventata un elemento cruciale nel marketing dei *brand*, con numerose aziende che cercano di personalizzare i propri prodotti per attrarre il loro pubblico di riferimento poiché adattando i prodotti alle esigenze individuali dei clienti, si garantisce che essi ottengano esattamente ciò che desiderano. Ciò non solo contribuirà alla loro soddisfazione, ma, soprattutto, aumenterà la probabilità che ritornino per future transazioni. La creazione di una base di clienti fedeli è fondamentale per mantenere alti i profitti e la personalizzazione rappresenta un ottimo strumento per raggiungere questo obiettivo. Ci troviamo in un periodo in cui la capacità di un'azienda di dimostrare un'elevata empatia, instaurare una forte intimità e offrire totale affidabilità assume un'importanza cruciale e solo quando le aziende pongono i clienti al centro delle proprie attività, mostrando empatia, interagendo in modo risonante e gestendo eticamente i dati forniti, i clienti tendono a ricordare e a rimanere fedeli. Concentrando l'attenzione sull'individuo, le aziende possono creare una forma di scambio di valore in cui i clienti forniscono i propri dati in cambio di garanzie sulla *privacy*, costruzione di rapporti di fiducia e offerta di un'esperienza unica e personalizzata che anticipa le loro esigenze e si adatta al loro contesto ambientale. I dirigenti aziendali devono porre maggiore attenzione al programma di trasformazione aziendale al fine di creare intimità con i clienti e questo implica l'accelerazione dell'uso dell'analisi e dell'intelligenza artificiale per l'iperpersonalizzazione e la previsione, nonché l'innovazione su larga scala per adattarsi ai cambiamenti delle esigenze dei clienti. È necessario quindi abbandonare le tradizionali fonti di vantaggio competitivo a favore di tre dinamici fattori di valore: il posizionamento dei consumatori al centro, l'utilizzo della tecnologia e l'innovazione su larga scala.
In questo modo, i *manager* potranno dimostrare che le loro aziende creano un valore a lungo termine, misurabile attraverso le *performance* ottenute con i clienti, i dipendenti e la società nel suo complesso e se gestito in modo adeguato, questo approccio si ripagherà migliorando le performance aziendali e garantendo la sostenibilità nel mercato.

Nel corso di questo studio, è importante considerare alcune limitazioni che potrebbero influire sia sull'interpretazione e sia sulla generalizzazione dei risultati ottenuti. I risultati si basano su un set di dati specifico e potrebbero non essere generalizzabili in tutti i contesti del settore della moda. È quindi importante considerare che la natura stessa dei dati utilizzati, come la loro qualità e disponibilità, potrebbe influenzare

direttamente i risultati dell'analisi predittiva e nel caso in cui i dati siano incompleti, errati o non rappresentativi della popolazione di clienti, le previsioni potrebbero non essere accurate o affidabili. Inoltre, se l'azienda ha una base di clienti limitata o ha raccolto dati solo per un breve periodo, le previsioni potrebbero non essere generalizzabili nel lungo termine. L'analisi predittiva, infatti si basa sul presupposto che i comportamenti dei clienti rimangano stabili nel tempo, tuttavia, questi ultimi possono essere influenzati da una serie di fattori esterni e interni, quali i cambiamenti nel mercato, la concorrenza, le tendenze sociali e i gusti individuali.
Anche se gli algoritmi di classificazione possono fornire previsioni e segmentazioni utili, l'interpretazione dei risultati può essere complessa. I modelli di *machine learning* possono essere opachi e non fornire spiegazioni dettagliate su come siano state raggiunte le previsioni e questo può limitare la comprensione delle ragioni sottostanti dei comportamenti dei clienti e rendere difficile l'adattamento delle strategie di marketing in modo significativo.
L'utilizzo dei dati dei clienti per l'analisi predittiva solleva questioni legate anche all'etica e alla *privacy* che vanno considerate attentamente ed infatti è fondamentale garantire che i dati siano raccolti, conservati e utilizzati in conformità alle leggi e alle normative vigenti. Inoltre, l'uso dei dati dei clienti potrebbe suscitare preoccupazioni sulla sicurezza e la potenziale condivisione non autorizzata delle informazioni personali.
Infine, le previsioni basate sull'analisi predittiva possono essere influenzate dalle condizioni di mercato in continua evoluzione e nel caso in cui il mercato subisca improvvisi cambiamenti o turbolenze, le previsioni potrebbero non essere più valide o potrebbe essere necessario un aggiornamento per riflettere la nuova realtà.
Considerando queste limitazioni, è importante adottare un approccio consapevole nell'interpretazione e nell'applicazione dei risultati dell'analisi predittiva, tenendo conto del contesto specifico dell'azienda e delle dinamiche del mercato della moda.

È opportuno considerare possibili direzioni future di ricerca che potrebbero contribuire a estendere e ad arricchire le conoscenze acquisite nell'ambito di questo studio.
Come espresso in precedenza nella sezione 4.1, è possibile condurre un ulteriore analisi per valutare se una riduzione delle caratteristiche, limitandosi a includere solo quelle considerate più rilevanti in base all'importanza attribuita dalla *feature importance*, possa portare a miglioramenti nelle prestazioni predittive.
Inoltre, con l'avanzamento della tecnologia e la crescente disponibilità di dati, le aziende avranno accesso a una vasta gamma di informazioni sui clienti e questo potrebbe includere dati provenienti da social media, dispositivi mobili ed altri canali. L'integrazione di tali dati diversificati consentirà un'analisi più approfondita e una previsione ancora più accurata del comportamento dei clienti.
L'evoluzione delle tecniche di *machine learning*, come il *deep learning* e l'apprendimento rinforzato, potrebbe migliorare ulteriormente la capacità di previsione del comportamento dei clienti. Queste tecniche consentono di gestire dati complessi e di identificare *pattern* e relazioni più sottili, fornendo previsioni ancora più precise e dettagliate.

Attualmente, molte analisi predittive si basano su dati storici per fare previsioni sul comportamento futuro dei clienti. Tuttavia, uno sviluppo futuro potrebbe essere l'utilizzo di dati in tempo reale per adattare immediatamente le risorse dedicate ai clienti e questo consentirebbe alle aziende di rispondere in tempo reale ai cambiamenti nel comportamento dei clienti e di offrire esperienze personalizzate in tempo reale.

Con la crescente attenzione sulla *privacy* dei dati e l'etica nell'utilizzo delle informazioni personali dei clienti, gli sviluppi futuri dovranno tener conto di tali considerazioni. Le aziende dovranno adottare pratiche di analisi responsabili e garantire la protezione dei dati dei clienti, rispettando le normative e le aspettative degli utenti.

Questi sviluppi consentiranno alle aziende di offrire esperienze ancora più personalizzate e rilevanti per i propri clienti, migliorando la loro fedeltà e il successo dell'azienda a lungo termine.