LUISS

Corso di Laurea in Marketing Analytics and Metrics

Cattedra Customer Intelligence e Logiche di Analisi dei Big Data

ARTIFICIAL INTELLIGENCE FOR EXPLORING HOW THE EFFECT OF ATTRIBUTES VARIES WITH E-TAILER FASHION PRODUCTS

Marina Paolanti

RELATORE

Luca Romeo

CORRELATORE

Matricola 750011

Sophia Stefani

CANDIDATO

Anno Accademico 2022/2023

INDICE

Capitolo 1. Introduzione

1.1 Analisi del contesto di riferimento

1.1.1 Come l'introduzione di Internet ha dato vita allo shopping online

Negli ultimi anni, il rapido sviluppo della tecnologia informatica e di Internet, ha creato un mercato molto favorevole allo sviluppo dello shopping online facendo di Internet lo strumento principale per accedere a varie risorse e informazioni (Xiong, Y., 2022).

L'uso diffuso di Internet segna l'ingresso della società umana nella "economia di rete" permettendo alle aziende, attraverso il commercio elettronico, di creare una connessione più diretta tra venditori e acquirenti (Xiong, Y., 2022). Quindi, per innovarsi e sopravvivere all'interno del mercato, le aziende non si sono potute limitare solo a vendere tramite punti vendita fisici ma anche tramite Internet.

Tutto questo ha permesso l'incremento di plurime attività commerciali consentendo alle aziende di vendere in tutto il mondo con un semplice clic del mouse (Gunasekaran et al., 2002).

Grazie all'introduzione di Internet, la comunicazione tra aziende e consumatori ha consentito alle aziende di fornire ai clienti informazioni adeguate ed efficaci, riducendo così il rischio di un processo decisionale asimmetrico. Quando i consumatori raccolgono e analizzano attivamente le informazioni, possono prendere decisioni di acquisto psicologicamente equilibrate, ridurre il senso di rischio e aumentare la fiducia nei prodotti (Xiong, Y., 2022).

Alcuni attori aziendali come produttori, grossisti, dettaglianti e fornitori di servizi hanno assistito a miglioramenti significativi, con un conseguente aumento dell'impatto delle operazioni commerciali, che ha accresciuto notevolmente l'efficienza del Marketing e delle transazioni (Xiong, Y., 2022). L'invenzione e l'innovazione della tecnologia e il modo in cui Internet ha influenzato le aziende Business-to-Consumer (B2C) e Business-to- Business (B2B) (exelab.com, 2023) a investire nel Marketing digitale ("Internet World Stats", 2014) ha permesso di migliorare la redditività al minor costo possibile (Ngo, 2015).

L'invenzione di Internet non ha mai cambiato i concetti di Marketing, ma ha reso più facile per le aziende utilizzare questo spazio tecnologico per servire meglio i clienti (Orapin, 2009). Difatti, come è stato appena detto, metà della popolazione mondiale è oggi utente di Internet attraverso telefoni cellulari, computer, tablet, laptop, ecc.; quindi, con l'avvento del Marketing digitale, i clienti possono ora acquistare beni e servizi di loro scelta con la massima comodità, risparmio di tempo, una selezione più ampia, maggiori informazioni, efficienza, confrontare i prezzi ecc. (Anwuri, P. N., & Eke., 2020).

Per effettuare i loro acquisti online, i consumatori si affidano sempre di più agli E-tailers che negli ultimi tempi stanno diventando protagonisti del commercio elettronico, offrendo ai consumatori la possibilità di acquistare comodamente online e aprendo così nuove opportunità di business (Kautish, P., Paul, J., & Sharma, R., 2021). Gli analisti di mercato prevedono che il mercato globale dell'E-tailing raggiungerà i 4,11 trilioni di dollari nel

2023 (Statista, 2023). Il settore moda è il più grande segmento di E-commerce del mercato al dettaglio, che si stima cresca dell'11,5% annuo (Statista, 2023). Con i rapidi progressi tecnologici guidati da Internet (Singh e Söderlund, 2020), le aziende hanno avuto prospettive di business senza precedenti oltre ai confini fisici dei tradizionali modelli di vendita al dettaglio (Amin et al., 2015; Shi et al., 2020).
Nel prossimo paragrafo esploreremo più in dettaglio il mondo dell'E-tailer e dell'E-commerce.

1.1.2 La nascita dell'E-tailer

l'E-tailer (o anche vendita al dettaglio elettronica) si riferisce ad un'azienda o ad un rivenditore che effettua la vendita di beni al dettaglio tramite transazioni Business-to-Consumer (B2C) su Internet (Kautish, 2011; FagerstrØm et al., 2020).
La vendita al dettaglio online è un formato abbastanza nuovo che alcuni rivenditori hanno adottato come conseguenza dell'evoluzione dell'E-commerce, tuttavia, anche se molto spesso i due termini vengono utilizzati come sinonimi, i concetti di E-commerce ed E-tailer non sono proprio la stessa cosa.
L'E-commerce (o commercio elettronico) è un termine più ampio rispetto a quello di E-tailer che comprende molte più attività commerciali condotte tramite Internet includendo altre forme di transazioni e attività connesse (come il Marketing, la pubblicità, la gestione delle scorte, il supporto ai clienti) e che quindi vanno oltre la semplice vendita al dettaglio online (Judy Howell, 2021, 25 July. Differenza tra E Tailing ed E Commerce. https://it.strephonsays.com/e-tailing-and-vs-e-commerce-15337). Inoltre l'E-commerce può coinvolgere sia aziende Business-to-Consumer (B2C) che aziende Business-to-Business (B2B) (exelab.com, 2023), dove le transazioni possono avvenire tra aziende o tra aziende e consumatori.
Certamente vi è una grande correlazione tra E-tailing ed E-commerce ma più precisamente l'E-tailing è una sottocategoria specifica del commercio elettronico.
Esistono due diversi tipi di E-tailer online, ovvero l'E-tailer "puro" e l'E-tailer multicanale. Gli E-tailer definiti "puri" sono quelli presenti solo online e quindi non hanno negozi fisici,perciò tutte le loro transazioni e i processi commerciali sono svolti completamente online come Amazon.com, dell.com, E-bay.com o nel settore della moda Zalando.com (Huang, H., & He, S., 2011, August).
Questa tipologia di E-tailers può formare delle alleanze strategiche o delle partnership con altre società al fine di fornire migliori servizi ai clienti come ad esempio stabilire partnership strategiche con corrieri, come FedEx, per consegnare i prodotti in modo efficiente ai clienti (Huang, H., & He, S., 2011, August).
Mentre gli E-tailer multicanale si riferiscono a quei rivenditori online che hanno anche una presenza fisica nei negozi. Questo tipo di rivenditori elettronici è noto anche come rivenditori click-and-mortar. Gli E-tailer multicanale hanno alcuni vantaggi rispetto agli E-tailer "puri", come l'esistenza di un canale di distribuzione completo, una Brand Awareness consolidata, una forte Brand Image e Customer Loyalty (Huang, H., & He, S., 2011, August). Alcuni esempi di rivenditori elettronici multicanale sono Walmart.com, Chapters.ca e Futureshop.ca (Huang, H., & He, S., 2011, August).

La proliferazione di Internet è stata la spina dorsale del mercato delle soluzioni di E-tailing, consentendo ai fornitori di servizi di raggiungere un pubblico più ampio anno dopo anno. Con il sostegno incrollabile del settore dell'E-commerce, i marchi hanno soddisfatto con successo ogni segmento di consumatori attraverso la vendita al dettaglio online.

Ma che cosa fa dell'E-tailer un modello di successo? L'E-tailing può portare vantaggi sia ai rivenditori che ai consumatori.

L'E-tailing richiede alle aziende di adattare i propri modelli di business per catturare le vendite su Internet, che possono includere la creazione di canali di distribuzione come magazzini, pagine Web e centri di spedizione dei prodotti. In particolare, i canali di distribuzione sono fondamentali per la vendita al dettaglio elettronica poiché permettono di portare il prodotto a casa del cliente (KamilTaylan.blog, economia finanziaria, 2021). Diventare un E-tailer offre quindi ai rivenditori l'opportunità di ampliare la propria base di clienti aiutandoli a raggiungerne di più rispetto a prima (Raffaele Felaco, 2020).

La vendita al dettaglio online comprende un'ampia gamma di aziende e settori (KamilTaylan.blog, economia finanziaria,2021) come quello della moda che è analizzato in questo lavoro di tesi. Tuttavia, ci sono somiglianze tra la maggior parte delle società di E-tailing che includono un sito Web accattivante, una strategia di marketing online, una distribuzione efficiente di prodotti o servizi e un'analisi dei dati dei clienti (KamilTaylan.blog, economia finanziaria,2021).

Un altro fattore di successo dell'Etailer è sicuramente la possibilità di ridurre i costi soprattutto per gli E-tailer "puri". I rivenditori online possono tagliare molto i costi perché non hanno bisogno di costruire negozi fisici e di solito hanno meno lavoro di ufficio rispetto ai rivenditori tradizionali (Huang, H., & He, S., 2011, August). Mentre i vantaggi per i consumatori sono sicuramente la possibilità di avere più alternative e fornitori disponibili in quanto i clienti possono essere raggiunti dagli E-tailer di tutto il mondo. Poi il confronto sul prezzo e le caratteristiche dei prodotti per i clienti è molto più facile in quanto è velocissimo passare da un sito Web all'altro dei vari rivenditori, difatti il fattore comodità e risparmio di tempo è al terzo posto tra i motivi per cui le persone acquistano online (Huang, H., & He, S., 2011, August).

Quindi il confronto tra i prezzi consente di poter scegliere il prezzo più conveniente in quanto la merce venduta online ha il potenziale per essere più economica di quella venduta in modo tradizionale data la grande concorrenza online.

E infine, un ulteriore vantaggio, consiste nell' avere accesso a maggiori informazioni da parte dei consumatori: una delle principali caratteristiche di Internet è l'empowerment delle informazioni. Internet può fornire molte più informazioni al pubblico su prodotti che non è probabile o non facile da ottenere rispetto a qualsiasi altro media. Tali informazioni possono aumentare il potere contrattuale quando i clienti stanno negoziando con i rivenditori (Huang, H., & He, S., 2011, August).

1.2 Obiettivi, il Gap nello stato dell'arte e la Metodologia di ricerca

Per gli E-tailers di moda sta diventando sempre più importante riuscire a catalogare velocemente i nuovi prodotti sul loro sito online in quanto ogni articolo è provvisto di caratteristiche diverse che lo contraddistingue.
Questo lavoro di tesi ha permesso la realizzazione di un modello predittivo che affronta un problema di classificazione multiclasse, utilizzando diversi algoritmi di Machine Learning, in grado di assegnare correttamente le etichette delle diverse classi di borse dell'E-tailer di moda MyTheresa in base alle loro caratteristiche.

L'Obiettivo di ricerca di questo lavoro di tesi è quello di permettere ad un E-tailer, più precisamente un E-tailer di moda, di non dover categorizzare manualmente ogni prodotto presente nel catalogo online ma di poter velocizzare questa attività aziendale tenendo conto delle features più rilevanti mediante l'Intelligenza Artificiale così da poter utilizzare questo tempo risparmiato e le informazioni raccolte in altre attività aziendali. La rilevanza di questa ricerca è inoltre data dalla possibilità di poter essere estesa non solo alle borse ma anche a qualsiasi altro prodotto di ogni settore di mercato.

Questo lavoro di tesi, ha evidenziato un Gap nella letteratura riguardante l'applicazione dell'Intelligenza Artificiale, e quindi del Machine Learning, in maniera approfondita, nel settore degli E-tailers di borse e accessori in base agli attributi. Le applicazioni più rilevanti, sviluppate in campo fashion riguardano per lo più la categoria abbigliamento per la quale sono state utilizzate tecnologie di AI per problemi di classificazione multiclasse. Si sono rilevate comunque carenze di studio e di applicazioni in questo ambito specifico.
Data la grande necessità di innovare e stare sempre di più al passo con i cambiamenti nel settore della moda, e quindi di borse e accessori, si propone con questo lavoro di superare il Gap creando un modello predittivo che consenta di categorizzare le borse di un E-tailer in base alle loro caratteristiche, evidenziandone quelle più rilevanti, per dare la possibilità alle aziende non solo di velocizzare questi processi aziendali, ma anche per arricchirle di dati per offrire ai consumatori prodotti sempre più personalizzati e su misura per loro.

Il lavoro di ricerca è stato svolto su Anaconda, una distribuzione dei linguaggi di programmazione Python e R per il calcolo scientifico (scienza dei dati, applicazioni di apprendimento automatico, elaborazione dati su larga scala, analisi predittiva, ecc.), che mira a semplificare l'installazione e la gestione delle librerie necessarie per l'analisi dei dati. Contiene un insieme di strumenti preinstallati, tra cui l'ambiente di sviluppo integrato (IDE) Spyder su cui sono state svolte le analisi del lavoro di ricerca.

Per la creazione e l'addestramento del modello predittivo si sono utilizzati 6 algoritmi di Machine Learning, ovvero il Random Forest, l'XG-Boost, il KNN, il Decision Tree, la Logistic Regression e il Naive Bayes i cui

risultati, per decretare quali siano gli algoritmi migliori per questo problema di classificazione multiclasse, sono stati valutati tramite la misurazione delle metriche di Accuracy, Precision, Recall e F1-score.
In questo lavoro di tesi, si desidera inoltre evidenziare che, al fine di effettuare un migliore confronto dei risultati ottenuti, sono stati attuati due bilanciamenti diversi del dataset, ovvero il Bilanciamento della variabile Y e il Bilanciamento del Train set.
È stato effettuato un Bilanciamento della variabile Y in quanto mostra le performance del modello a livello ideale e teorico, ovvero permette di dire quali sarebbero stati i risultati se fosse stato un caso puramente concettuale, mentre il Bilanciamento del Train è stato fatto perché consente di costruire un modello predittivo rappresentativo del problema reale che si vuole risolvere, riflettendo la natura dei dati che effettivamente si incontreranno nel mondo reale. L'obiettivo principale dei dati di Train è quello di fornire al modello di Machine Learning esempi realistici che gli consentano di imparare e generalizzare correttamente su nuovi dati.
Tra i due diversi bilanciamenti i risultati più corretti e rilevanti per questo lavoro di tesi sono naturalmente quelli derivanti dal Bilanciamento del Train.

1.3 Struttura del lavoro

Il presente lavoro di tesi consta di 5 capitoli, che verranno brevemente illustrati di seguito.
Nel capitolo 2, ovvero quello dedicato alla Literature Review, si andranno ad esplorare le applicazioni degli E-tailers nel mondo dei Big Data e dell'Intelligenza Artificiale riportando, inoltre, alcuni esempi di problemi di classificazione multiclasse di Machine Learning sia nel settore moda sia in altri ambiti.
Il Capitolo 3 descrive i materiali e i metodi che sono stati adottati, in particolare verrà descritto che cos'è un modello predittivo, verrà delineato il Dataset utilizzato (il Dataset dell'Etailer Mytheresa), la fase di Data-pre processing in cui si effettuerà il Data Cleaning e le Analisi Statistiche in cui verrà fatto un confronto tra i risultati delle statistiche descrittive di tutte le features (ovvero un grafico di distribuzione ed un boxplot) per entrambi i bilanciamenti (ovvero il Bilanciamento del Train e della Y), le matrici di correlazione per entrambi i bilanciamenti prima e dopo il Data Cleaning e per finire la descrizione degli algoritmi utilizzati in questo lavoro di tesi.
Il Capitolo 4 illustrerà i risultati ottenuti da tutti gli algoritmi per entrambi i bilanciamenti, analizzando le metriche di Accuracy, Precision, Recall, F1-score e la matrice di confusione e la curva Roc, delineando poi quali sono state le caratteristiche più rilevanti per ogni algoritmo, per poi passare al paragrafo delle Applicazioni di Marketing nell'E-tailing.
Infine, il Capitolo 5 mostrerà le conclusioni del lavoro, le implicazioni manageriali e gli sviluppi futuri del presente lavoro di tesi.

Capitolo 2. Stato dell'arte

2.1 Big Data ed E-tailer

Le moderne tecnologie di raccolta e analisi dei Big Data hanno portato a un'esplosione sia della portata che del volume dei dati di mercato, così da permettere alle aziende di ottenere informazioni preziose per poi produrre beni e servizi da vendere sul mercato (Dai, B., Liu, N., & Jian, Z.,2021).

Il concetto di Big Data esiste da anni (sas.com, 2023), già dagli anni Cinquanta. Decenni prima che qualcuno pronunciasse il termine (sas.com, 2023), le aziende utilizzavano l'analitica di base (essenzialmente numeri in un foglio di calcolo che venivano esaminati manualmente) per scoprire intuizioni e tendenze (Nadikattu, R. R., 2020).

Oggi l'analisi dei Big Data, è comunemente nota come la tecnologia che impiega dati massicci e metodi di analisi scientifica per aiutare le imprese a effettuare operazioni ottimali come lo sviluppo di prodotti sulla base di informazioni accurate (Dai, B., Liu, N., & Jian, Z., 2021) permettendo di offrire nuovi vantaggi come la velocità e l'efficienza. Mentre qualche anno fa un'azienda raccoglieva dati, eseguiva analisi e scopriva informazioni che potevano essere utilizzate per decisioni future, oggi può identificare intuizioni per decisioni immediate (Nadikattu, R. R., 2020).

I Big Data sono spesso caratterizzati da tre V: volume, velocità e varietà (Meta Group, 2001). Il volume si riferisce alla proprietà di "grandezza", la velocità si riferisce alla estrema velocità con cui i processi digitali rendono i Big Data ancora più grandi, la varietà si riferisce a nuovi formati e tipi di dati. Si tratta di dati che non sono nella forma adatta per l'analisi statistica tradizionale e che spesso consistono anche in parole, immagini, video o altri output di consumo non numerici (Hofacker, C. F., Malthouse, E. C., & Sultan, F., 2016).

Nel Marketing, il principale motore di interesse per i Big Data è la loro potenziale utilità per informare le decisioni ed eseguire campagne di Marketing. Alcuni autori hanno suggerito che l'analisi dei Big Data ha trasformato in modo significativo il modo in cui il Marketing viene condotto oggi (Erevelles et al., 2015).

EMarketer riferisce che in un sondaggio del 2013, condotto tra esperti di Marketing negli Stati Uniti, "l'85% dei dirigenti di agenzie e marchi statunitensi ha affermato che i Big Data hanno prodotto più della metà delle iniziative di Marketing quando si trattava di aumentare le informazioni sul comportamento dei consumatori" (emarketer.com, 2013).

La Big Data Analytics (BDA) si riferisce all' "analisi quantitativa basata sui fatti" che aiuta il processo decisionale (Davenport, 2006). La BDA genera valore attraverso la trasparenza, la scoperta, l'esposizione alla variabilità e il miglioramento delle prestazioni (Manyika, 2011).

Numerosi studi hanno rilevato una conoscenza sufficiente del servizio clienti attraverso i metodi e i sistemi di Big Data Analytics per rispondere meglio alle esigenze e alle aspettative dei consumatori. Inoltre, il supporto alle decisioni svolge un ruolo importante nella gestione delle relazioni con i clienti (Hui, S. C. S., Dastane, O., Johari, Z., & Roslee, M., 2021). Pertanto, l'analisi predittiva aiuterà a gestire le relazioni nell'ambiente

aziendale. Con le tecnologie dei Big Data vengono utilizzati molti strumenti che consentono di migliorare la capacità di fornire più rapidamente valore al cliente (Romika, 2015; Tomas, 2017; Li Q. C., 2017). Akter (2016) ha classificato il valore commerciale della BDA come vantaggi transazionali, informativi e strategici per le aziende di E-tailer. Di conseguenza, i Big Data Analytics aiutano le aziende, soprattutto gli E-tailer, a generare approfondimenti incentrati sul consumatore per migliorare le operazioni commerciali (Avinash, B.M., A.B., 2017). Le aziende di E-tailer sono una tipologia di azienda che adotta più rapidamente la BDA perché ha bisogno di rimanere molto competitiva sul mercato di riferimento (Koirala, 2012).

L'analisi dei Big Data consente alle aziende di accedere e analizzare rapidamente più fonti di dati, migliorare il processo decisionale, sviluppare servizi su misura che soddisfino le esigenze dei consumatori e valutare meglio le attività di prestito per ridurre al minimo le perdite (Jiwat, R.H., 2017).

I consumatori scelgono spesso le marche e i prodotti tramite Internet dando la possibilità ai clienti di cercare informazioni importanti e di acquistare i prodotti desiderati (Moon, 2004). Moon ha inoltre affermato che lo shopping online è quindi molto efficace rispetto ai canali di Marketing tradizionali (Moon, 2004). Oltre a ciò, i clienti possono anche cercare informazioni e opzioni appropriate per la scelta dei prodotti (Fiona, 2005). La ricerca di informazioni è quindi la prima fase del processo di acquisto del cliente. Di conseguenza, Adeline ha affermato che la ricerca di informazioni è uno stato significativo perché i potenziali clienti ricevono messaggi pubblicitari che possono influenzare le loro scelte e le loro capacità decisionali (Adeline, C.P., 2006). Farhang ha invece affermato che la ricerca di dati è un processo che aiuta i clienti a fare delle scelte nella piattaforma di shopping online (Farhang, 2012). La ricerca di dati è uno strumento eccellente per aiutare gli acquirenti online a scoprire i prodotti o i servizi più appropriati. Per questo motivo, i rivenditori online devono migliorare il supporto dei dati e offrirne un elenco completo sui prodotti e utilizzare i motori di ricerca interni per migliorare l'efficacia della ricerca (Mittal, 2013). È evidente che il commercio elettronico aumenta la disponibilità di informazioni in modo esponenziale, fornisce ai clienti l'accesso alla conoscenza del prodotto-servizio, possiede una qualità migliore ed è più veloce di prima (Fawzy, 2018).

Lo studio condotto da Hui, S. C. S., Dastane, O., Johari, Z., & Roslee, M. (2021) dal titolo “Enhancing Online Repurchase Intention via Application of Big Data Analytics in E-Commerce” analizza l'impatto di applicazioni selezionate di Big-Data Analytics sull'intenzione di riacquisto. La discussione è citata per ogni fattore di Big-Data Analytics, ovvero la ricerca di informazioni, il dynamic prizing, i sistemi di raccomandazione e la personalizzazione. I risultati hanno rivelato che la ricerca di informazioni influenza l'intenzione di riacquisto dei consumatori nel settore dello shopping online in Malesia. Questo conferma i risultati delineati dai ricercatori Andrew e Adeline, che indicano che la ricerca di informazioni e l'usabilità del sito di shopping online hanno una relazione positiva con l'intenzione di riacquisto dei consumatori (Andrew, 2014; Adeline, 2006). Inoltre, se il sito di shopping online fornisce le informazioni necessarie per la ricerca, può modificare le decisioni di acquisto degli acquirenti online (Pee, 2018). Pertanto, per migliorare l'intenzione di riacquisto, questo parametro giocherà un ruolo fondamentale. I risultati hanno anche mostrato che il sistema di raccomandazioni ha un impatto positivo sull'intenzione di riacquisto dei consumatori. Precedenti lavori di

letteratura e studi empirici (Yan, 2018; Ankush, 2017; Zhang, 2016) hanno convalidato il sistema di raccomandazioni come uno dei fattori critici che ha influenzato positivamente e significativamente la soddisfazione e l'intenzione di riacquisto dei clienti dello shopping online. Studi precedenti (Vijay, 2018; Chaitanya, 2018; Shin, 2017) hanno riconosciuto che il prezzo dinamico è un fattore critico che ha un impatto positivo sull'intenzione di riacquisto degli acquisti online. Un argomento che ha guadagnato un'attenzione imperativa negli ultimi anni è il dynamic pricing con prezzi competitivi dei prodotti. Il dynamic pricing possiamo definirlo come il libero adeguamento del prezzo dei prodotti in base alla domanda e all'offerta, anche a livello individuale della transazione (Wim, 2014). In altre parole, adegua i prezzi di beni identici online in base alla disponibilità dei clienti a pagare (Robert M., 2001). La maggiore capacità dei rivenditori online di conoscere il comportamento d'acquisto dei propri clienti ha fatto sorgere il problema del dynamic pricing, una pratica in cui il venditore fissa i prezzi in base alla disponibilità dell'acquirente a pagare (Oliver, 2011). Pertanto, in caso di variazione della domanda o di altri parametri, il dynamic pricing, che consiste in una semplice variazione del prezzo, potrebbe essere utile per determinare il prezzo adeguato. Inoltre, il dynamic pricing comprende una domanda misurabile, la sensibilità del cliente e altri fattori (Başak, 2013).

Sarà quindi possibile stabilire il prezzo corretto per i vari profili di clienti per migliorare l'utilità complessiva (Hui, S. C. S., Dastane, O., Johari, Z., & Roslee, M., 2021). Di conseguenza, è necessario raccogliere tutte le informazioni necessarie sui loro acquisti, sulla loro sensibilità a determinate variabili e su ciò che cercano nel settore per capire il cliente (Hui, S. C. S., Dastane, O., Johari, Z., & Roslee, M., 2021).

È difficile raccomandare una strategia di prezzi dinamici, perché ha un impatto diretto sui profitti. È fondamentale considerare gli aspetti legati ai prezzi della concorrenza, ai prezzi differenziali, ai prezzi situazionali, ecc. L'e-commerce è noto per proporre offerte convenienti, sconti, coupon, ecc. e il pricing dinamico dovrebbe includere anche queste considerazioni (Hui, S. C. S., Dastane, O., Johari, Z., & Roslee, M., 2021). I risultati della ricerca condotta da Hui, S. C. S., Dastane, O., Johari, Z., & Roslee, M. (2021) hanno dimostrato che la personalizzazione non ha avuto un impatto significativo e positivo sull'intenzione di riacquisto e che gli annunci pubblicitari sul sito di shopping online che sono rilevanti per le preferenze del cliente scatenano l'intenzione di riacquisto del consumatore (Chou, 2016). Inoltre, la notifica via e-mail di una promozione di un prodotto per lo shopping online innesca l'intenzione di riacquisto dei clienti se questi sono soddisfatti del rivenditore online (Thi & Shu, 2017; Chong, 2012).

La personalizzazione può riguardare l'interfaccia, l'elenco delle informazioni, la procedura di prenotazione e vari suggerimenti. Esiste un ambito specifico per sviluppare la sequenza delle fasi in base agli aspetti della personalizzazione. In questo paper, si raccomanda alle aziende di commercio online di identificare le scelte personalizzate dei clienti sulla base dei dati relativi agli acquisti passati o di un sondaggio e di progettare la strategia di personalizzazione di conseguenza (Hui, S. C. S., Dastane, O., Johari, Z., & Roslee, M., 2021).

Precedenti lavori di letteratura (Yusepaldo, 2018; Xin, K L, J. L. 2018; Wajeeha, 2018) hanno confermato che la soddisfazione del cliente ha mediato in modo imperativo e positivo l'intenzione di riacquisto e altri fattori nello shopping online. Inoltre, vi è stata anche la dimostrazione che la ricerca di informazioni gioca un ruolo

fondamentale nell'influenzare la soddisfazione degli acquirenti online (Nebojša, 2019; Davis, 1989; Gupta, 2013). Lo studio di Radu (2015) sostiene che la soddisfazione del cliente per la personalizzazione ha un effetto di mediazione positivo sull'intenzione di riacquisto. Sulla base dei risultati ottenuti dalla sua ricerca, alcune caratteristiche dei negozi online sono state chiaramente preferite da alcuni gruppi di utenti. Questi risultati hanno aiutato a comprendere gli stili decisionali dei clienti e hanno reso più fattibile la personalizzazione dell'interfaccia utente/negozio online per soddisfare meglio le aspettative degli utenti (Hui, S. C. S., Dastane, O., Johari, Z., & Roslee, M., 2021). L'applicazione dell'analisi dei dati ha permesso di ottenere una soddisfazione positiva dei clienti grazie a un'interfaccia e a una notifica di acquisto online personalizzate. Tuttavia, i risultati hanno rivelato che fattori come la ricerca di informazioni, i prezzi dinamici e il sistema di raccomandazione non hanno avuto l'effetto di mediazione della soddisfazione del cliente sull'intenzione di riacquisto. Pertanto, un'azienda di shopping online non deve concentrarsi sull'effetto di mediazione della soddisfazione del cliente sulla ricerca di informazioni, sui prezzi dinamici e sul sistema di raccomandazione per migliorare l'intenzione di riacquisto (Hui, S. C. S., Dastane, O., Johari, Z., & Roslee, M., 2021).

Lo studio di Hui, S. C. S., Dastane, O., Johari, Z., & Roslee, M., (2021) ha fornito un prezioso contributo di informazioni sui quattro fattori analitici dei Big Data, quali la ricerca di informazioni, il sistema di raccomandazione, il prezzo dinamico e la personalizzazione, e sul loro impatto sull'intenzione di riacquisto. Inoltre, la loro ricerca intendeva anche esaminare l'effetto di mediazione della soddisfazione del cliente tra i fattori analitici dei Big Data e l'intenzione di riacquisto. I risultati della loro ricerca hanno concluso che i quattro fattori analitici dei Big Data hanno incoraggiato direttamente l'intenzione di riacquisto dei clienti e la loro soddisfazione complessiva, suggerendo che le aziende possono migliorare l'intenzione di riacquisto attraverso l'applicazione di queste tecniche analitiche dei big data. La personalizzazione ha avuto un impatto secondario e significativo sull'effetto di mediazione tra la soddisfazione dei clienti e l'intenzione di riacquisto. Pertanto, le aziende di shopping online devono considerare l'importanza della personalizzazione nella soddisfazione dei clienti per aumentare la loro intenzione di riacquisto. Tuttavia, la ricerca di informazioni, il sistema di raccomandazione e i prezzi dinamici non hanno un effetto di mediazione della soddisfazione del cliente sull'intenzione di riacquisto nel settore dello shopping online in Malesia (Hui, S. C. S., Dastane, O., Johari, Z., & Roslee, M., 2021).

Analizzare i dati storici da una prospettiva moderna, individuare scenari aziendali nuovi (e stimolanti) e applicare metodologie per trovare una soluzione migliore sono le principali preoccupazioni di un analista di dati (Nadikattu, R. R., 2020). Gli scienziati dei dati hanno il compito di scoprire i fatti nascosti nella complessa rete di dati non strutturati sviluppando algoritmi e modelli euristici in modo da poter essere utilizzati per prendere decisioni aziendali. L'analisi dei dati ha mostrato una crescita così straordinaria in tutto il mondo che si prevede che presto il fatturato del mercato dei Big Data aumenterà del 50% (Nadikattu, R. R.,2020).

La maggior parte delle organizzazioni ha ormai capito che se catturano tutti i dati che confluiscono nelle loro aziende possono, applicando tecnologia, matematica e tecniche statistiche, ricavarne un valore significativo (Nadikattu, R. R., 2020). La capacità di lavorare più velocemente (rimanendo agili) offre alle organizzazioni

un vantaggio competitivo che prima non avevano (Nadikattu, R. R.,2020). Possiamo affermare che nel corso degli anni l'orientamento delle aziende è cambiato, passando da un orientamento al prodotto ad un orientamento ai dati (Nadikattu, R. R., 2020).

Oggi anche una piccola informazione è preziosa per le aziende, il che rende indispensabile ricavare sempre più informazioni possibili. Questa necessità ha fatto nascere l'esigenza di esperti in grado di fornire approfondimenti significativi. In realtà, le tecnologie stanno aiutando in modo eccellente diversi settori, consentendo loro di mettere a frutto ogni singola intuizione (Nadikattu, R. R., 2020).

I dati sono la base di quasi tutte le attività svolte oggi, sia nel campo dell'istruzione, della ricerca, della sanità, della tecnologia, ma soprattutto degli E-tailer (Nadikattu, R. R., 2020). Sfruttare queste tecnologie richiederà ai rivenditori di pensare in modo creativo all'offerta di nuovi servizi e allo sviluppo di nuovi modelli di business. Alcuni rivenditori, come Amazon.com (Amazon), sembrano particolarmente abili nello sviluppo di tali servizi. Oggi, la maggior parte dei prodotti viene esaminata ampiamente su Amazon e l'azienda utilizza la sua vasta quantità di dati per identificare per tutti i consumatori "articoli che vengono acquistati frequentemente insieme" e "anche i clienti che hanno acquistato questo articolo " (Fisher, M., & Raman, A., 2018).

In un E-tailer, la risposta per rimanere in gioco ed essere competitivi è capire meglio il cliente per servirlo nel modo migliore possibile. Ciò naturalmente richiede la capacità di analizzare tutte le fonti di dati disparate con cui le aziende hanno a che fare ogni giorno, compresi i weblog, i dati delle transazioni dei clienti, i social media, i dati delle carte di credito e i dati dei programmi di fidelizzazione (Fisher, M., & Raman, A., 2018).

2.1.1 L'utilizzo dei Big Data dagli E-tailers

I Big Data possono aiutare gli E-tailers non solo a esplorare nuove opportunità e innovazioni, ma anche a mettere in atto i loro attuali modelli di business e liquidare parti dell'attività che non funzionano per migliorare le loro prestazioni (Fisher, M., & Raman, A., 2018).

Nel 1995, Marshall Fisher e Ananth Raman hanno iniziato a ricercare come i rivenditori potessero utilizzare meglio i dati. C'è da dire che i Big Data non erano così grandi come lo sono oggi e il loro uso era limitato. I dati principalmente a disposizione dei rivenditori erano dati del punto vendita (POS), integrati da elementi come le carte fedeltà, che fornivano informazioni sui dati demografici dei clienti che acquistavano i prodotti del rivenditore (Fisher, M., & Raman, A., 2018). Oggi, i dati POS sono stati aumentati in numerosi modi. In molti contesti di E-tailer e fisici, i rivenditori possono tenere traccia non solo di ciò che viene venduto in luoghi e orari diversi (attraverso i sistemi POS), ma anche di chi sta comprando questi articoli. Inoltre, la tecnologia consente potenzialmente ai rivenditori di osservare ciò che un cliente sfoglia o prova nel camerino prima dell'acquisto. È facile visualizzare come gli E-tailers possono monitorare il comportamento di navigazione dei propri clienti (Fisher, M., & Raman, A., 2018).

Durante i 15 anni che hanno preceduto la pubblicazione di The New Science of Retailing, Marshall Fisher e Ananth Raman hanno scoperto che anche solo i dati di vendita, opportunamente analizzati, possono essere estremamente utili a un rivenditore per migliorare il processo decisionale. La maggior parte degli E-tailer ora osserva molto i siti Web dei propri concorrenti per ottenere informazioni sulla concorrenza, come i prezzi che stanno applicando sui loro prodotti in quel momento (Fisher, M., & Raman, A., 2018).
Ora gli analisti dispongono di molti nuovi dati, inclusi quelli in continua crescita da E-tailing, video in-store, tracciamento dei clienti all'interno del negozio tramite telefoni cellulari e camerini intelligenti (Fisher, M., & Raman, A., 2018).

2.1.2 Big Data utilizzati per la gestione dei resi di un E-tailer di abbigliamento

Più le persone fanno acquisti online, più i rivenditori online si trovano ad affrontare un numero elevato di resi da parte dei consumatori. Per pianificare adeguatamente la capacità del processo di gestione dei resi di un E-tailer, è necessario prevedere la quantità prevista di pacchi restituiti. L'analisi dei Big Data offre un gran numero di metodi per svolgere tali compiti. Tuttavia, va notato che soprattutto gli E-tailer di piccole e medie dimensioni non hanno le capacità e le risorse per impiegare tecniche così complesse (Asdecker, B., & Karl, D., 2018, September).
Il lavoro di ricerca condotto da Asdecker Björn and Karl David (2018) analizza le prestazioni di diversi metodi di analisi dei dati che differiscono per complessità di applicazione, utilizzando dati reali di un E-tailer di abbigliamento.
I risultati indicano che l'uso dei Big Data Analytics è di grande utilità per gestire in modo efficace ed efficiente i resi dei consumatori. D'altra parte, e dal punto di vista dei professionisti probabilmente ancora più interessante, si conclude anche che una regressione logistica binaria come metodo analizzato più semplice può già fornire risultati soddisfacenti (Asdecker, B., & Karl, D., 2018, September).
Per accelerare le restituzioni, le operazioni richiedono una pianificazione della capacità più accurata, che a sua volta si basa su previsioni. L'importanza di questo compito non può essere sottovalutata: quanto migliori sono le previsioni, tanto più efficace ed efficiente sarà l'elaborazione dei resi dei consumatori.
La letteratura fornisce diversi metodi econometrici, statistici e/o di data mining che possono essere impiegati per prevedere i resi. Alcuni sono molto complessi, mentre altri sono più semplici e facili da applicare (Asdecker, B., & Karl, D., 2018, September).
In un mondo in cui molti decisori cercano l'unica alternativa ottimale, i metodi complessi e all'avanguardia sembrano essere la scelta migliore. Tuttavia, essi richiedono anche competenze sofisticate, capacità aggiuntive e risorse finanziarie che sono limitate, soprattutto nelle piccole e medie imprese di vendita al dettaglio elettronico (Coleman et al., 2016).
Molto simile al lavoro di ricerca condotto da Asdecker Björn and Karl David (2018), Urbanke et al. (2015) hanno sviluppato un sistema di supporto alle decisioni che identifica le transazioni con un'alta probabilità di ritorno prima che avvenga la vendita effettiva e hanno dimostrato l'applicabilità dell'approccio utilizzando un

ampio set di dati di un E-tailer di moda tedesco (Asdecker, B., & Karl, D., 2018, September). Nel loro studio hanno confrontato sette tecniche di previsione, ovvero l'analisi dei componenti principali, l'analisi discriminante lineare, la decomposizione randomizzata del valore singolare troncato, la selezione delle caratteristiche basata sulla statistica univariata del chi-quadro, la proiezione casuale, la fattorizzazione della matrice non negativa e una tecnica specifica di estrazione delle caratteristiche che ignora gli indicatori nominali. Pur cercando la tecnica con la massima precisione, Urbanke et al. non confrontano gli approcci semplici con quelli complessi che è quello che invece hanno fatto Asdecker Björn and Karl David (2018).
Poiché i consumatori decidono di restituire o tenere gli articoli consegnati, la variabile dipendente è di natura binaria. Questo articolo prende in considerazione cinque approcci, descritti brevemente di seguito, ovvero:

1. Regressione logistica binaria.
   La regressione logistica binaria è il metodo più semplice da prendere in considerazione. È un'estensione della regressione lineare, in cui la variabile dipendente è binaria (1=ritorno, 0=mantenimento). Le variabili indipendenti possono essere continue (intervallo/rapporto) o categoriali (ordinali/nominali) (Asdecker, B., & Karl, D., 2018, September).

2. Analisi della funzione discriminante lineare.
   L'analisi della funzione discriminante lineare, eseguita per la prima volta da Fisher (1936), presenta grandi analogie con la regressione logistica. L'idea di base è quella di creare una combinazione lineare di variabili indipendenti che classifichi al meglio i dati disponibili. In questo modo, si determina un punteggio per ogni osservazione che viene poi confrontato con un punteggio discriminante critico per effettuare la classificazione (restituire o mantenere) (Asdecker, B., & Karl, D., 2018, September).

3. Rete neurale artificiale: perceptron multistrato
   Le reti neurali artificiali si basano su un insieme di nodi connessi, i cosiddetti neuroni artificiali, organizzati in strati. I neuroni artificiali connessi possono scambiare segnali tra loro. Il neurone artificiale ricevente lo elabora e, a sua volta, invia segnali ai neuroni artificiali ad esso collegati (Asdecker, B., & Karl, D., 2018, September). L'obiettivo finale è quello di trovare una funzione che assegni al meglio i dati di ingresso all'uscita corretta. Per raggiungere questo obiettivo nel contesto della gestione dei rendimenti, questo studio utilizza i perceptron multistrato, una classe di reti neurali feed forward (Asdecker, B., & Karl, D., 2018, September). In questo caso, le informazioni fluiscono esclusivamente dallo strato di ingresso attraverso strati nascosti con una certa quantità di unità fino allo strato di uscita, senza alcun flusso di feedback. L'addestramento avviene tramite back propagation, una tecnica di apprendimento supervisionato che confronta le uscite della rete con i valori reali noti (Hastie et al., 2009).

4. Apprendimento ad albero decisionale: Algoritmo C5.0

   Gli alberi decisionali sono strutture gerarchiche di rami, che rappresentano le congiunzioni di determinate caratteristiche, e di foglie, che rappresentano le etichette delle classi (Asdecker, B., & Karl, D., 2018, September). L'obiettivo di questa tecnica è creare un albero decisionale che classifichi al meglio le osservazioni disponibili. A questo scopo sono stati presentati molti algoritmi di alberi decisionali. In questa analisi si utilizza l'algoritmo C5.0 che è il successore più veloce ed efficiente dell'algoritmo C4.5, ampiamente utilizzato (Pandya, 2015).

5. Tecnica di apprendimento in ensemble

   Il metodo dell'apprendimento in ensemble utilizza diversi algoritmi per migliorare le prestazioni predittive. Determina il risultato di ogni singolo algoritmo e lo interpreta come ipotesi per il verdetto finale (Polikar, 2006).

Nel complesso, sorprende la buona performance di classificazione dei metodi parametrici (regressione logistica binaria e analisi discriminante lineare). Infatti, le loro prestazioni sono peggiori di soli 1,66/1,85 punti percentuali rispetto alla tecnica ensemble come miglior metodo non parametrico (Asdecker, B., & Karl, D., 2018, September).
Pertanto, modelli semplici come la regressione logistica binaria potrebbero essere la scelta migliore nella pratica commerciale, soprattutto per gli E-tailer di piccole e medie dimensioni che devono affrontare capacità di Data mining e risorse finanziarie limitate. Ciò è particolarmente vero perché possono essere utilizzati anche per l'avvio di misure preventive di gestione dei resi (Asdecker, B., & Karl, D., 2018, September).

2.1.3 Big Data per aggiornare gli assortimenti degli E-tailer

Un altro modo in cui vengono utilizzati i Big Data è per aggiornare l'assortimento dei prodotti presenti nelle varie categorie degli E-tailer.
Periodicamente, i retailer aggiornano l'assortimento dei prodotti presenti nelle varie categorie dei loro E-tailer, eliminando alcuni prodotti e aggiungendone altri, in risposta all'evoluzione dei modelli di domanda e ai nuovi prodotti entrati sul mercato (Fisher, M., & Raman, A.,2018). La parte più difficile di questo processo è sapere quanto un potenziale nuovo prodotto venderà se aggiunto all'assortimento di un E-tailer, e quanto di queste vendite sarà incrementale e quanto cannibalizzerà le vendite dei prodotti esistenti (Fisher, M., & Raman, A.,2018). Fisher e Vaidyanathan (2012, 2014) descrivono un modo per farlo, identificando innanzitutto gli attributi dei prodotti di una categoria, utilizzando poi le vendite dei prodotti esistenti per stimare la domanda di attributi e, infine, stimando la domanda di un potenziale nuovo prodotto dalla domanda degli attributi che lo compongono. Esempi di attributi sono le dimensioni e il prezzo/valore del prodotto, il materiale, la pietra preziosa primaria e il prezzo/valore per i gioielli (Fisher, M., & Raman, A.,2018). I parametri di un modello di domanda, le quote di domanda a livello di attributo in un E-tailer per gli attributi e la probabilità che un

cliente si sostituisca a un altro livello di attributo se la sua prima scelta non è nell'assortimento, sono scelti per minimizzare la deviazione tra le vendite effettive e le previsioni del modello di domanda per i prodotti esistenti. L'implementazione di questo approccio ha prodotto un aumento dei ricavi compreso tra il 3% e il 6% (Fisher, M., & Raman, A.,2018).

2.1.4 Big Data utilizzati per monitorare i prezzi dei concorrenti online

I Big Data possono essere impiegati anche per tenere traccia dei prezzi dei concorrenti degli E-tailer e, in effetti, molti E-tailer utilizzeranno software per "raccogliere" quotidianamente i siti Web della concorrenza per scaricare l'assortimento di prodotti che trasportano, i loro prezzi e se sono disponibili o meno (Fisher, M., & Raman, A.,2018).

Fisher et al. (2017) descrivono uno studio svolto con un E-tailer che raccoglieva quotidianamente dati sui prezzi da diversi concorrenti e si chiedeva come rispondere con i propri prezzi. Hanno così formulato un modello che stimava la domanda per ogni prodotto venduto dall'E-tailer in base ai loro prezzi e a quelli della concorrenza, un esperimento con prezzi randomizzati per stimare i parametri nel modello e un algoritmo di determinazione del prezzo con la migliore risposta che tiene conto della scelta del comportamento del consumatore, azioni dei concorrenti e parametri di fornitura (costi di approvvigionamento, obiettivo di margine e restrizioni sui prezzi del produttore) (Fisher, M., & Raman, A.,2018).

L'implementazione iniziale dell'algoritmo ha prodotto un aumento dei ricavi dell'11% in una categoria di prodotto e del 19% in una seconda, mantenendo al contempo un margine superiore a un obiettivo specificato dal rivenditore (Fisher, M., & Raman, A.,2018).

2.2 Intelligenza Artificiale nell'E-tailer

Nel corso degli anni, il ruolo dell'Intelligenza Artificiale nello sviluppo del mercato dell'E-tailing è stato straordinario in quanto ha osservato attentamente le intenzioni, i desideri e le esigenze umane e lo ha valutato per creare un impatto positivo sul business (Fan, X., Ning, N., & Deng, N., 2020). L'IA "si riferisce a programmi, algoritmi, sistemi e macchine che dimostrano intelligenza" (Shankar 2018, p. vi), è "manifestata da macchine che mostrano aspetti dell'intelligenza umana" (Huang e Rust 2018, p. 155) e involge a macchine che imitano "comportamenti umani intelligenti" (Syam e Sharma 2018, p. 136). Si basa su diverse tecnologie chiave, come l'apprendimento automatico, l'elaborazione del linguaggio naturale, i sistemi esperti basati su regole, le reti neurali, l'apprendimento profondo, i robot fisici e l'automazione dei processi robotici (Davenport 2018). In altre parole, il software IA è progettato per sfruttare l'apprendimento automatico per migliorare l'etichettatura e l'organizzazione per fornire ricerche visive correttamente etichettate (Fan, X., Ning, N., & Deng, N., 2020).

Utilizzando questi strumenti, l'IA fornisce un mezzo per "interpretare correttamente i dati esterni, imparare da tali dati e mostrare un adattamento flessibile" (Kaplan e Haenlein 2019, p. 17).

Per automatizzare i processi aziendali, gli algoritmi di IA eseguono compiti ben definiti con un intervento umano minimo o nullo, come il trasferimento di dati da e-mail o call center a sistemi di archiviazione (aggiornamento dei file dei clienti), la sostituzione di carte bancomat smarrite, l'esecuzione di semplici operazioni di mercato o la "lettura" di documenti per estrarre disposizioni chiave utilizzando l'elaborazione del linguaggio naturale e anche la classificazione dei prodotti in base agli attributi (Davenport, T., Guha, A., Grewal, D., & Bressgott, T., 2020).

L'IA è stata impiegata anche per ricavare informazioni da vasti volumi di dati relativi a clienti e transazioni che comprendono non solo dati numerici, ma anche testi, voci, immagini ed espressioni facciali (Davenport, T., Guha, A., Grewal, D., & Bressgott, T., 2020).

Utilizzando le analisi abilitate dall'IA, le aziende possono prevedere cosa un cliente probabilmente acquisterà, anticipare le frodi creditizie prima che si verifichino o distribuire pubblicità digitale mirata in tempo reale.

Ad esempio, gli stilisti di Stitch Fix (caso che verrà discusso nel corso del paragrafo), utilizzano già l'Intelligenza Artificiale per identificare i modelli di abbigliamento più adatti ai diversi clienti (Davenport, T., Guha, A., Grewal, D., & Bressgott, T., 2020) integrando dati forniti dalle preferenze espresse dai clienti, dalle loro bacheche Pinterest, dagli appunti scritti a mano, dalle preferenze di clienti simili e dalle tendenze generali di stile (Davenport, T., Guha, A., Grewal, D., & Bressgott, T., 2020).

Un altro modo di descrivere l'IA non dipende dalla sua tecnologia di base, ma piuttosto dalle sue applicazioni di Marketing e aziendali, come l'automazione dei processi aziendali, l'acquisizione di informazioni dai dati o il coinvolgimento di clienti e dipendenti (Davenport e Ronanki 2018).

In un'analisi di oltre 400 casi d'uso dell'IA in 19 settori e 9 funzioni aziendali, McKinsey & Co. (2018) indica che il maggior valore potenziale dell'IA riguarda i settori legati al Marketing e alle vendite (Chui et al. 2018), attraverso l'impatto su attività di Marketing come le offerte migliori per i clienti (Davenport et al. 2011), l'acquisto programmatico di annunci digitali (Parekh 2018) e il lead scoring predittivo (Harding 2017). L'impatto dell'IA sul Marketing è maggiore in settori come i beni di consumo confezionati, la vendita al dettaglio, le banche e i viaggi. Questi settori implicano intrinsecamente un contatto frequente con un gran numero di clienti e producono grandi quantità di dati sulle transazioni e sugli attributi dei clienti (Davenport, T., Guha, A., Grewal, D., & Bressgott, T., 2020).

Inoltre, le informazioni provenienti da fonti esterne, come i media o i rapporti dei broker, possono aumentare questi dati. Successivamente, l'IA può essere sfruttata per analizzare tali dati e definire raccomandazioni personalizzate (relative al prossimo prodotto da acquistare, al prezzo ottimale, ecc.) in tempo reale (Mehta et al. 2018). Inoltre i Marketer prevedono di utilizzare l'IA in aree come la segmentazione e l'analisi (legate alla strategia di Marketing) e la messaggistica, la personalizzazione e i comportamenti predittivi (legati ai comportamenti dei clienti) (Columbus 2019).

L'Intelligenza Artificiale può essere adottata per coinvolgere di più i clienti, prima e dopo il processo di vendita come ad esempio l'impiego da parte di Conversica.com di bot AI che lavorano per spostare le transazioni dei clienti lungo la pipeline di Marketing, mentre il bot AI, utilizzato da 1-800-Flowers, fornisce assistenza sia

alle vendite che al servizio clienti. I bot AI offrono vantaggi che vanno oltre la semplice disponibilità 24/7. Non solo i bot AI hanno tassi di errore più bassi, ma consentono anche agli agenti umani di occuparsi di casi più complessi. Inoltre, l'implementazione dei bot AI può essere aumentata o ridotta a seconda delle necessità, quando la domanda aumenta o diminuisce (Davenport, T., Guha, A., Grewal, D., & Bressgott, T., 2020). I diversi esempi citati sopra indicano che, piuttosto che sostituire gli esseri umani, le aziende in genere utilizzano l'IA per aumentare le capacità dei loro dipendenti umani, come nel caso di Stitch Fix che utilizza l'IA per aumentare gli sforzi dei suoi stilisti nel fare scelte appropriate per i clienti (Gaudin 2016). Questo punto si allinea bene con i sentimenti espressi da Ginni Rometty, CEO di IBM, che ha proposto che l'IA non porterà a un mondo di uomini "contro" macchine, ma piuttosto a un mondo di uomini "più" macchine (Carpenter 2015).
In futuro, l'Intelligenza Artificiale (IA) sembra destinata a influenzare le strategie di Marketing, compresi i modelli di business, i processi di vendita e le opzioni di assistenza ai clienti, nonché i loro comportamenti (Davenport, T., Guha, A., Grewal, D., & Bressgott, T., 2020). l'IA offre il potenziale per aumentare i ricavi e ridurre i costi. I ricavi possono aumentare grazie al miglioramento delle decisioni di Marketing (ad esempio, prezzi, promozioni, raccomandazioni sui prodotti, maggiore coinvolgimento dei clienti); i costi possono diminuire grazie all'automazione di semplici attività di Marketing, del servizio clienti e delle transazioni di mercato (strutturate) (Davenport, T., Guha, A., Grewal, D., & Bressgott, T., 2020).
Il settore della moda al dettaglio ha assistito a cambiamenti significativi negli ultimi anni attraverso la trasformazione digitale. Ad eccezione della fase finale, il trasferimento fisico dei beni, ogni parte del percorso di acquisto di un consumatore viene digitalizzata (Baird 2018). I grandi rivenditori ed E-tailers di oggi si sono trasformati in aziende multicanale, in cui lo stesso cliente può visitare il rivenditore tramite diversi canali per scopi diversi come navigare tra i prodotti online ed effettuare acquisti offline (Modi, D., & Zhao, L., 2019). La maggior parte ha anche ampliato la propria attenzione dalla vendita di prodotti al coinvolgimento e all'empowerment dei clienti, con l'obiettivo di creare un'esperienza-cliente unica e gratificante. Di conseguenza, la pratica della vendita al dettaglio sta abbracciando una gamma sempre più ampia di attività man mano che i rivenditori espandono i confini dei loro mercati target e sviluppano modi innovativi per interagire con i loro clienti (Modi, D., & Zhao, L., 2019).
Il modello di business attualmente utilizzato dagli E-tailer prevede generalmente che i clienti effettuino gli ordini, dopodiché il rivenditore online spedisce i prodotti (il modello shopping-then-shipping) – (Agrawal et al. 2018; Gans et al. 2017). Con l'IA, gli E-tailers sono in grado di prevedere ciò che i clienti desiderano identificando le preferenze dei clienti spedendo gli articoli ai clienti senza un'offerta formale, con la possibilità di restituire ciò che non serve (Agrawal et al. 2018; Gans et al. 2017). Questo cambiamento trasformerebbe le strategie di Marketing dei rivenditori, i modelli di business e i comportamenti dei clienti (ad esempio, la ricerca di informazioni) (Davenport, T., Guha, A., Grewal, D., & Bressgott, T., 2020).
Aziende come Stitch Fix e Trendy Butler utilizzano già l'IA per cercare di prevedere ciò che i loro clienti desiderano, con vari livelli di successo (Davenport, T., Guha, A., Grewal, D., & Bressgott, T., 2020).

Molti esperti e professionisti del settore prevedono che l'IA cambierà il volto delle strategie di Marketing e i comportamenti dei clienti. Infatti, un'indagine di Salesforce mostra che l'IA sarà la tecnologia più adottata dai Marketer nei prossimi anni (Columbus 2019). E dai casi di maggior successo si può trarre almeno una lezione generale: L'intelligenza artificiale può essere il motore cognitivo di un approccio originale e personalizzato al consumatore, in grado di raggiungere alti livelli di soddisfazione e fedeltà dei clienti (Davenport et al., 2020). Gli esempi sopra citati dimostrano i vantaggi e le soluzioni offerte da sistemi di AI. Nonostante ciò è necessario un approccio critico a queste tecnologie in quanto possono comportare delle sfide di carattere etico e giuridico in materia di privacy dei dati e pregiudizi algoritmici (Larson 2019).

2.2.1 Come l'intelligenza Artificiale è stata applicata nell'E-tailer

Ci sono molti esempi in cui è stata adottata l'Intelligenza Artificiale nel mondo dell'E-tailer, ci sono difatti studi precedenti che hanno dimostrato che la tecnologia di vendita al dettaglio intelligente riduce i costi operativi e migliora la redditività delle imprese (Vazquez et al., 2017; Dacko, 2017; Renko and Druzijanic, 2014).

L'azienda Stitch Fix (fondata nel 2011), un E-tailer multimarca di articoli di moda, ha compiuto un significativo passo in avanti nel suo modello di "shipping-then-shopping", combinando un sistema di Intelligenza Artificiale molto sofisticato e complesso con il lavoro di stilisti umani (Roberto Grandinetti, 2020).

La fondatrice di Stitch Fix, Katrina Lake, ha deciso di utilizzare un sistema di Intelligenza Artificiale che utilizza una grande quantità di informazioni per selezionare una serie di cinque articoli da inserire in ogni spedizione. Queste informazioni sono fornite in gran parte dai clienti che rispondono a un questionario dettagliato sulle loro preferenze di stile, taglia e prezzo (che possono essere indicate utilizzando un formato tabellare), oltre a immagini o altri dati non numerici su di loro (dalle pagine Pinterest e dai like dei clienti) (Roberto Grandinetti, 2020).

Altre informazioni sono tratte in particolare dal portafoglio clienti di Stitch Fix, ormai molto ampio.

Per un articolo notoriamente difficile da adattare come i jeans, ad esempio, "gli algoritmi sono in grado di selezionare per ogni cliente una varietà di jeans che altri clienti con misure simili hanno deciso di tenere" (Malone, 2018, p. 37).

Katrina Lake, definisce il modello commerciale che ha inventato come semplice: "Noi ti inviamo capi di abbigliamento e accessori che pensiamo ti piacciano; tu tieni i capi che vuoi e rispedisci gli altri" (Lake, 2018, p. 35). Naturalmente, indovinare cosa piacerà ai clienti (cosa che l'azienda è riuscita a fare, come dimostrano le sue vendite) è tutt'altro che facile.

Dopo ogni spedizione, i clienti forniscono anche un feedback informativo che il sistema utilizza per migliorare le sue scelte nel tempo (Lake, 2018; Luce, 2019). Tuttavia, la preparazione di una serie di articoli personalizzati non è solo opera degli algoritmi. L'ultima parola spetta sempre agli stilisti umani che possono relazionarsi con i consumatori in modo più personale (Malone, 2018). Il loro intervento giustifica il fatto che se un cliente non

tiene nessuno degli articoli di una spedizione, paga comunque 20 dollari. Lo straordinario successo dell'azienda ha spinto altre imprese del mondo della moda, tra cui Trunk Club (Modi & Zhao, 2019), a muoversi nella stessa direzione, con risultati diversi (Davenport et al., 2020).

Trunk Club, una filiale della società Nordstrom, è nata come startup con sede a Chicago e sta sconvolgendo il settore della moda al dettaglio grazie al suo format innovativo e incentrato sul cliente (Modi, D., & Zhao, L., 2019).

L'azienda, uno dei primi operatori nel settore della vendita al dettaglio di moda maschile e femminile di fascia alta su abbonamento, ha progettato il proprio modello di business per consentire ai propri clienti (ad esempio, giovani dirigenti americani con una disponibilità di tempo ridotta, che desiderano vestirsi alla moda ma non sanno esattamente cosa starebbe loro bene) di esternalizzare le proprie esigenze di moda (Modi, D., & Zhao, L., 2019).

L'azienda innova il processo di acquisto dei consumatori fornendo servizi di personal styling basati sulle loro esigenze, sulla taglia, sul budget e sulle preferenze di stile, con l'aiuto di un team di esperti di styling dedicati, attraverso il suo sito web o i suoi negozi in tutti gli Stati Uniti (Modi, D., & Zhao, L., 2019).

Ma tutto ciò non finisce qui perché nella selezione dei prodotti per i suoi clienti, l'azienda utilizza l'Intelligenza Artificiale adottando l'apprendimento automatico e gli algoritmi di raccomandazione personalizzati insieme a stilisti personali, che costituiscono un buon esempio di innovazione di processo sconvolgendo così il tradizionale modello di vendita al dettaglio (Reid 2012).

A differenza della maggior parte dei suoi concorrenti, Trunk Club si rivolge al mercato degli abbonamenti di fascia alta, dimostrando di essere un leader delle innovazioni in termini di vendita al dettaglio omnicanale perché offre un'esperienza non solo online ma anche offline (Modi, D., & Zhao, L., 2019).

Ogni cliente, lavora con un personal stylist che sceglie abbigliamento e accessori per la sua scatola (chiamata Trunk), che viene poi spedita a casa sua. I clienti possono acquistare gli articoli che preferiscono e rispedire il resto a Trunk Club senza alcun costo se non gli piace. Questi servizi possono essere forniti online o in una delle sei sedi dell'azienda chiamate Clubhouse (Modi, D., & Zhao, L., 2019). Ogni Clubhouse è personalizzata con arredi unici, camerini e un bar che offre birra artigianale, vino, champagne e liquori per offrire ai clienti un'esperienza di shopping di lusso. I clienti che visitano le Clubhouse possono acquistare l'abbigliamento direttamente dal sito o chiedere al loro stilista di curare un baule e spedirlo a casa loro. Quindi, Trunk Club offre flessibilità ai suoi clienti in termini di come vogliono fare acquisti e dove vogliono fare acquisti (Modi, D., & Zhao, L., 2019).

Per soddisfare il proprio segmento di clientela, Trunk Club ha utilizzato questo tipo di modello di convenienza, in cui i clienti si abbonano per acquistare periodicamente i loro prodotti su base mensile o stagionale senza dover perdere tempo per visitare negozi e siti Web per trovare abbigliamento (Modi, D., & Zhao, L., 2019).

Inoltre, Trunk Club esternalizza e dipende dai membri del suo team di analisi che unificano e analizzano il comportamento del cliente e il comportamento dello stilista da varie fonti di dati. Trunk Club impiega più di 500 stilisti che costruiscono relazioni speciali con i propri clienti, facendo l'inventario dei loro gusti personali,

preferenze e misure (Modi, D., & Zhao, L., 2019). Gli stilisti esaminano l'inventario del Trunk Club con l'aiuto di un motore di raccomandazione basato sull'Intelligenza Artificiale e raccomandano gli articoli ai loro clienti. La comunicazione tra stilisti e clienti avviene per telefono, e-mail e tramite l'app Trunk Club (Modi, D., & Zhao, L., 2019). Quindi, enormi dati su quali articoli gli stilisti scelgono per quale cliente, e quindi quali prodotti tra quelli inviati vengono conservati o restituiti dai clienti, forniscono informazioni utili per vari team durante la progettazione dell'inventario o nel formulare raccomandazioni agli stilisti per il loro clienti (Segment 2018).

2.2.2 Best prectices: Amazon e Netflix

Possiamo affermare che anche altri due grandi E-tailer come Amazon e Netflix avevano già avviato i loro servizi per fornire ai clienti consigli per gli acquisti (Shen, 2014).

Amazon, attore chiave nel mercato dell'E-tailing, è stato uno dei primi ad adottare la tecnologia AI per accrescere l'efficienza aziendale (Roberto Grandinetti, 2020). L'azienda ha implementato l'Intelligenza Artificiale per comprendere la volontà dei consumatori di acquistare un prodotto specifico per consentire transazioni senza contanti sulle consegne. L'utilizzo di tecnologie intelligenti per comprendere il contesto del comportamento dei consumatori è stato l'obiettivo principale delle aziende che operano nel settore dell'E-tailer (Roberto Grandinetti, 2020).

Entro la fine del 2030, si prevede che le aziende dedicheranno il 10-12% dell'intera spesa in ricerca e sviluppo alle capacità di blockchain e intelligenza artificiale per migliorare la penetrazione nel mercato (Roberto Grandinetti, 2020).

Il fulcro di Netflix è il suo modello operativo incentrato sui dati e sull'Intelligenza Artificiale. È alimentato da un'infrastruttura software che raccoglie dati, addestra ed esegue algoritmi che guidano praticamente ogni aspetto del business, dalla personalizzazione dell'esperienza dell'utente alla scelta di concetti cinematografici vincenti per le sue prossime produzioni (Verganti, R., Vendraminelli, L., & Iansiti, M., 2020).

Netflix ha iniziato a sfruttare l'Intelligenza Artificiale nel 2010, per alimentare il suo sistema di raccomandazione. Nel 2014, ha ampliato il proprio approccio investendo ampiamente nella comprensione del comportamento degli utenti e sviluppando un'esperienza di streaming personalizzata per ciascun utente (Verganti, R., Vendraminelli, L., & Iansiti, M., 2020).

Le schermate dell'applicazione che un utente vede oggi sono "disegnate in tempo reale" da una macchina. Molti confini e parametri sono specificati dai progettisti umani all'inizio del processo. Ma le decisioni su quali film mostrare, come mostrarli, con quali immagini rappresentarli e molte altre decisioni di progettazione vengono prese da algoritmi incorporati nei cicli di risoluzione dei problemi dell'IA (Verganti, R., Vendraminelli, L., & Iansiti, M., 2020).

I problemi di base che la maggior parte dei sistemi di Intelligenza Artificiale cerca di risolvere per dare forma a un'esperienza di progettazione riguardano la previsione di un risultato (Verganti, R., Vendraminelli, L., & Iansiti, M., 2020).

Lo strumento per fare quella previsione è un algoritmo: l'insieme di regole che una macchina segue per risolvere un particolare problema. L'intelligenza artificiale può incorporare molti tipi di algoritmi (Domingos, 2012). Alcuni di loro hanno un processo integrato per l'aggiornamento e il miglioramento, il più delle volte basato su "processi decisionali di Markov", che cercano di modellare una sequenza di azioni, ciascuna modellata da una politica e seguita da una ricompensa (Verganti, R., Vendraminelli, L., & Iansiti, M., 2020). Un esempio potrebbero essere gli algoritmi di Netflix che aggiornano dinamicamente la sua interfaccia utente, in base al comportamento effettivo dell'utente, come indicato dai suoi clic (mentre la politica decide cosa viene visualizzato, il clic è la "ricompensa") (Verganti, R., Vendraminelli, L., & Iansiti, M., 2020).

2.3 Modelli di classificazione di Machine Learning adottati in diversi settori

Questa sezione del capitolo propone un'analisi della letteratura relativa ai classificatori di Machine Learning utilizzati per categorizzare diverse tipologie di prodotti non solo nel settore moda e dell'e-commerce ma anche in altri ambiti come il settore musicale, agroalimentare e politico.
Gli E-tailer offrono tantissimi prodotti ai consumatori e molti di questi presentano un numero sempre crescente di diverse categorie che richiedono un enorme sforzo di manodopera per quanto riguarda la loro categorizzazione. In recenti studi è emerso che questa attività può arrivare a far impiegare complessivamente fino al 25% del tempo speso per la gestione dei contenuti (Ding, Y., Korotkiy, M., Omelayenko, B., Kartseva, V., Zykov, V., Klein, M., ... & Fensel, D., 2002, April).
Dato che la categorizzazione dei prodotti per gli E-tailer è così costosa, complicata, dispendiosa in termini di tempo e soggetta ad errori, la gestione dei contenuti necessita il supporto di processi di Machine Learning per consentire una più rapida categorizzazione dei prodotti nelle diverse categorie.

2.3.1 Modelli di classificazione di Machine Learning impiegati nel mondo della moda

La crescente popolarità dei social media e la prosperità dell'E-commerce hanno prodotto enormi quantità di dati crossmediali sulla moda, come i dati di strada condivisi dagli utenti, i dati delle sfilate rilasciati dai marchi di moda e i dati dei prodotti forniti dai siti di 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 apprendimento automatico e di visione computerizzata è diventata sempre più importante per rivoluzionare il settore e ridisegnare la meccanica della moda.
Diversi ricercatori si sono cimentati in queste analisi come Kiapour, M. H., Yamaguchi, K., Berg, A. C., & Berg, T. L. (2014) che hanno definito cinque stili di moda e hanno addestrato un classificatore SVM su un insieme di caratteristiche create a mano per classificare gli articoli in stili di abbigliamento. Altri studiosi come Ma, Jia, Zhou et al., a differenza dei metodi precedenti che costruiscono una rete di classificazione, hanno creato uno spazio semantico della moda per descrivere gli stili di abbigliamento.

Un altro studio è stato condotto da Rachel Rose Getman, Denise Nicole Green, Kavita Bala, Utkarsh Mall, Nehal Rawat, Sonia Appasamy, and Bharath Hariharan dal titolo “Machine Learning (ML) for Tracking Fashion Trends: Documenting the Frequency of the Baseball Cap on Social Media and the Runway”. In questo studio, storici della moda e informatici hanno collaborato per esplorare il potenziale pratico di questo metodo emergente esaminando una tendenza relativa a un particolare articolo di moda, ovvero il berretto da baseball, utilizzando due grandi set di dati: il database Vogue Runway (2000-2018) e il set di dati Streetstyle-27K di Matzen et al. (2013-2016) (Getman, R. R., Green, D. N., Bala, K., Mall, U., Rawat, N., Appasamy, S., & Hariharan, B., 2021). E’ stato adottato il Machine Learning per il riconoscimento dei pattern attraverso una ricerca interdisciplinare che ha mostrato un aumento dei cappellini da baseball sulle passerelle di moda mai visto prima del 2008 (Getman, R. R., Green, D. N., Bala, K., Mall, U., Rawat, N., Appasamy, S., & Hariharan, B., 2021). Questo aumento di cappellini sportivi casual sulle passerelle di moda si è allineato con diverse tendenze popolari dello street-style e con la campagna presidenziale del 2016, oltre ad altre influenze economiche e sociali. Il cappellino da baseball viene utilizzato come proof of concept per dimostrare come la computer vision e l'analisi automatizzata di molte migliaia di immagini possano offrire nuovi metodi per studiare il riconoscimento delle tendenze della moda (Getman, R. R., Green, D. N., Bala, K., Mall, U., Rawat, N., Appasamy, S., & Hariharan, B., 2021). I tag o le etichette degli attributi dell'abbigliamento sono le caratteristiche identificate all'interno dell'immagine durante il processo di addestramento in cui i ricercatori hanno etichettato le immagini nel set di dati di addestramento iniziale come aventi un "cappello da baseball" o "altro cappello" o "nessun cappello"(Getman, R. R., Green, D. N., Bala, K., Mall, U., Rawat, N., Appasamy, S., & Hariharan, B., 2021). I tassi di accuratezza del classificatore tra il set di dati dei social media di Instagram e quello delle passerelle di moda di Vogue Runway sono stati entrambi elevati rispetto ai tassi di accuratezza di ricerche simili nel campo della CS (Al-Halah et al., 2017, Matzen et al., 2017). I social media hanno registrato un tasso di accuratezza complessivo del 92,18%, mentre il set di dati di Vogue Runway ha registrato un tasso di accuratezza complessivo dell'87,96%. Il tasso di accuratezza è stato calcolato attraverso una matrice di confusione, confrontando l'accuratezza delle identificazioni effettuate dal classificatore con quelle effettuate dai ricercatori. Per raggiungere un tasso di accuratezza statisticamente significativo, è stato necessario identificare circa 1.000 immagini (Getman, R. R., Green, D. N., Bala, K., Mall, U., Rawat, N., Appasamy, S., & Hariharan, B., 2021). Il classificatore dei social media e i ricercatori hanno identificato le immagini con "nessun cappello" nel 96,38% dei casi, "altri cappelli (non da baseball)" nell'88,88% dei casi e "cappelli da baseball" nel 91,05% dei casi. Per la matrice di confusione di Vogue Runway, sia il classificatore che i ricercatori hanno identificato "nessun cappello" (93,84%) e hanno etichettato le immagini come "altri cappelli" per l'87,87% e "cappellini da baseball" per il 92,18% del tempo (Getman, R. R., Green, D. N., Bala, K., Mall, U., Rawat, N., Appasamy, S., & Hariharan, B., 2021).

Un altro studio condotto da Chloe Satinet e François Fouss dal titolo “A Supervised Machine Learning Classification Framework for Clothing Products’ Sustainability” utilizza il Machine Learning per sviluppare

un modello predittivo, applicato alla categoria dei prodotti di abbigliamento, che possa valutare facilmente e rapidamente la sostenibilità ambientale dei prodotti durante il loro ciclo di vita in base alle caratteristiche.
L'industria della moda manca davvero di un chiaro sistema di etichettatura ambientale. Di conseguenza, i consumatori attenti alla sostenibilità affrontano un grosso problema quando cercano di identificare prodotti sostenibili dal punto di vista ambientale. Come evidenziato nei lavori di ricerca di Brad, A., Delemare, A., Hurley, N., Lenikus, V., Mulrenan, R., Nemes, N., Trunk, U., Urbancic, N. (2018) e Gaasbeek, A., Golsteijn, L., Vieira, M. (2015), ci sono una varietà di certificazioni di sostenibilità confuse e poche etichette che catturano l'impatto ambientale complessivo dei prodotti, poiché le procedure esistenti per valutare l'impatto ambientale dei prodotti durante il loro ciclo di vita richiedono tempo e sono costose.
L'utilizzo di strumenti di apprendimento automatico per trovare un modello che possa valutare facilmente e rapidamente la sostenibilità di un gran numero di prodotti durante il loro ciclo di vita non è stato tuttavia ancora testato per tutte le categorie di prodotti (ad esempio, prodotti di abbigliamento invece di articoli per la casa, elettrodomestici o prodotti chimici) o per contesti più specifici (ad esempio, per l'uso da parte di E-tailer anziché di progettisti di prodotti) (Satinet, C., & Fouss, F., 2022).
Per costruire un modello di valutazione dell'impatto ambientale, sono state applicate tecniche di apprendimento automatico supervisionato a un problema di classificazione. In un problema di classificazione, la variabile di output è categorica e il modello ottenuto tenta di prevedere come dovrebbe essere etichettato un nuovo oggetto di input (Satinet, C., & Fouss, F., 2022).
In questo studio sono stati testati nove strumenti di Machine Learning supervisionati per classificare i prodotti di abbigliamento in base al loro livello di sostenibilità: un algoritmo k-nearest neighbors (k-NN), una regressione logistica (LR), una Support Vector Machine (SVM), una rete neurale artificiale (ANN) con un singolo strato nascosto, un Decision Tree (DT) e metodi di ensemble come Bagging, Random Forest, Boosting e Gradient Boosting (Satinet, C., & Fouss, F., 2022).
Dai risultati è emerso che ci sono diversi algoritmi che funzionano bene, ma quello che ha dato i risultati migliori è stato il Random Forest in quanto oltre a fare pochi errori di previsione (precisione media di 0,91 su 5 volte), ha anche prestazioni relativamente stabili su diversi set di dati di addestramento (precisione con deviazione standard di 0,01 su 5 volte), rispetta l'ordine di classe relativo dei prodotti (Tau-b di Kendall di 0,92), classifica pochi prodotti sostenibili come non sostenibili e viceversa (MSE di 0,11 e OCI di 0,14) e sembra formare modelli con regole decisionali rappresentativo della realtà (Satinet, C., & Fouss, F., 2022).

2.3.2 Modelli di classificazione di Machine Learning impiegati nel mondo dell'E-commerce

Il lavoro di ricerca condotto da Mieczysław Pawłowski dal titolo "Machine Learning Based Product Classification for E-Commerce" colma alcune lacune della letteratura relativa alla classificazione dei prodotti e al mantenimento della coerenza dei dati dei prodotti per i negozi online con l'applicazione di classificatori di Machine Learning.

I nomi dei prodotti per i negozi online o le frasi di ricerca degli utenti sono solitamente testi di breve durata che consentono di essere presentati sullo schermo del laptop, dello smartphone o stampati sullo scontrino fiscale (Pawłowski, M., 2022). La classificazione dei prodotti fa parte dei campi di ricerca noti come corrispondenza di ontologie e tassonomie. Nella corrispondenza della tassonomia, i dati sono annotati per relazione (non per significato) (Pawłowski, M., 2022). Se stai per trovare un oggetto in una determinata categoria, un algoritmo di corrispondenza deve stabilire il significato nei dati esterni o nel contesto all'interno dello schema. Al contrario, i sistemi logici delle ontologie degli assiomi lavorano per l'annotazione dei dati secondo il significato funzionale (Shvaiko & Euzenat, 2005). Quando la classificazione viene creata nel modo dell'ontologia, elementi simili vengono raggruppati in base alla funzionalità. Durante un processo di ricerca, l'utente si concentra sulla corrispondenza delle aspettative con le funzionalità, e quindi per questo motivo si può affermare che le fasi di ricerca collegano le variabili nascoste sul sito dell'utente con le caratteristiche dei prodotti rappresentate da oggetti di testo (nomi di prodotti) (Pawłowski, M., 2022). L'assistenza automatizzata dal lato utente può migliorare le prestazioni del ricercatore (misurate dal numero di documenti pertinenti trovati) di circa il 20% (Jansen & McNeese, 2005).
Nel presente studio si sono testati ben 12 classificatori: Bernoulli NB, Complement NB, K-Neighbors Class, Linear SVC, Logistic Regression, Multinomial NB, Nearest Centroid, Pass-Aggressive, Perceptron, Random Forest, Ridge Classifier, SGD Classifier ma per ulteriori esperimenti sono stati selezionati solo Linear SVC, Ridge Classifier e Random Forest Classifier in quanto hanno ottenuto un'elevata accuratezza, una buona efficienza in termini di tempo e utilizzano metodi di classificazione diversi (Pawłowski, M., 2022).
Dai risultati emersi si può quindi riassumere che i classificatori non riconoscono sufficientemente bene il vocabolario specifico degli utenti finché non vengono fornite nuove conoscenze nel file Wysz per l'addestramento (Pawłowski, M., 2022). Questo si ripercuote non solo sulla classificazione di cataloghi, ma anche sui prodotti di nicchia. L'accuratezza della classificazione durante i test su Kw4, che è un insieme di dati di prodotti di coda, è aumentata da 0, quando i modelli sono stati addestrati con i dati pre-elaborati Org2, Kw17, Kw17s, a 0,94 quando i classificatori sono stati addestrati su KwSem (Pawłowski, M., 2022). Questo dato può essere interpretato come il fatto che i nomi dei prodotti di coda non si correlano con i file preelaborati, ma si correlano molto bene con le frasi di ricerca degli utenti (Pawłowski, M., 2022). Riassumendo, possiamo affermare che l'ipotesi H1, secondo cui l'integrazione dei nomi dei prodotti come dati testuali che rappresentano le caratteristiche e le funzionalità di un oggetto reale con i contenuti degli utenti come vocabolario aggiuntivo della descrizione del prodotto può migliorare significativamente l'accuratezza della classificazione, è verificata positivamente (Pawłowski, M., 2022).

Un altro studio sull'E-commerce condotto da Y. Ding, M. Korotkiy, B. Omelayenko, V. Kartseva, V. Zykov, M. Klein, E. Schulten, and D. Fensel dal titolo "GoldenBullet: Automated Classification of Product Data in E-commerce" ha descritto un sistema chiamato GoldenBullet che applica tecniche di information retrieval e Machine Learning al problema della classificazione dei dati di prodotto.

Il problema principale è causato dall'eterogeneità delle descrizioni delle informazioni utilizzate da venditori e clienti e quindi soluzioni intelligenti che aiutino a meccanizzare il processo di strutturazione, classificazione, allineamento e personalizzazione sono un requisito fondamentale per superare con successo gli attuali colli di bottiglia del commercio elettronico B2B (Ding, Y., Korotkiy, M., Omelayenko, B., Kartseva, V., Zykov, V., Klein, M., ... & Fensel, D., 2002, April). Una gestione efficace dei contenuti per il commercio elettronico B2B deve affrontare diversi aspetti: l'estrazione delle informazioni da fonti approssimative, la classificazione delle informazioni per rendere i dati dei prodotti manutenibili e accessibili, la riclassificazione dei dati dei prodotti, la personalizzazione delle informazioni e la mappatura tra le diverse presentazioni delle informazioni (D. Fensel, Y. Ding, E. Schulten, B. Omelayenko, G. Botquin, M. Brown, and A. Flett, 2001). Tutti questi compiti secondari sono ostacolati dalla mancanza di standard adeguati e per tale motivo i ricercatori sono ricorsi all'impiego di GoldenBullet che aiuta a meccanizzare il processo di classificazione dei prodotti. Nello studio viene anche menzionato uno schema di classificazione ampiamente utilizzato negli Stati Uniti simile al GoldenBullet che è l'UNSPSC: è un tipico esempio di standard orizzontale che copre tutti i possibili ambiti di prodotto, ma non è molto dettagliato in nessun ambito. Un altro esempio di standard di questo tipo è l'Universal Content Extended Classification (UCEC), che prende l'UNSPSC come punto di partenza e lo perfeziona con gli attributi. Un altro esempio di standard verticale è Rosetta Net4 che delinea dettagliatamente i prodotti dell'industria hardware e software che descrivono un certo dominio di prodotti in modo più dettagliato rispetto ai comuni standard orizzontali (D. Fensel, Y. Ding, E. Schulten, B. Omelayenko, G. Botquin, M. Brown, and A. Flett, 2001). GoldenBullet è un ambiente software progettato per classificare automaticamente i prodotti, sulla base delle loro descrizioni originali e degli standard di classificazione esistenti (come UNSPSC), integrando diversi algoritmi di classificazione dalle aree di recupero delle informazioni e apprendimento automatico e alcune tecniche di elaborazione del linguaggio naturale per pre-elaborare i dati e indicizzare UNSPSC (D. Fensel, Y. Ding, E. Schulten, B. Omelayenko, G. Botquin, M. Brown, and A. Flett, 2001). Un metodo standard nell'Information Retrieval è il noto Vector space model (VSM). Il "Salton's Vector Space Model" (cf. [Salton et al., 1975]) utilizza un vettore di parole per rappresentare il documento e l'interrogazione dell'utente, quindi applica la formula della somiglianza del coseno per calcolare la somiglianza tra il documento e la query in modo da recuperare il documento più rilevante per la query dell'utente. J. M. Gomez-Hidalgo and M. B. Rodriguez (1997, July) ha utilizzato il "Salton's Vector Space Model" (cf. [Salton et al., 1975]) per rappresentare il documento (nel nostro caso la descrizione del prodotto) e le categorie esistenti (ad esempio in questo caso UNSPSC). Quindi la categoria (UNSPSC) può essere assegnata a un documento (prodotto) quando la somiglianza del coseno tra loro supera una certa soglia. L'idea di base è rappresentare ogni documento come un vettore di determinate frequenze di parole ponderate (Ding, Y., Korotkiy, M., Omelayenko, B., Kartseva, V., Zykov, V., Klein, M., ... & Fensel, D., 2002, April).

VSM è adottato quindi anche in questo studio per trovare la corrispondenza tra i prodotti UNSPSC e le descrizioni dei prodotti. Un altro classificatore basato su istanze che è stato implementato è basato sul metodo k-Nearest Neighbor (KNN). Anche in questo caso, l'algoritmo utilizza direttamente l'insieme di esempi pre-

classificati per classificare un esempio (Ding, Y., Korotkiy, M., Omelayenko, B., Kartseva, V., Zykov, V., Klein, M., ... & Fensel, D., 2002, April). L'algoritmo passa l'intero set di esempi di addestramento e cerca quello più simile, quindi assegna la classe al nuovo esempio, uguale alla classe di quello più simile. KNN è computazionalmente costoso e richiede molta memoria per funzionare a seconda del numero di esempi pre-classificati (Ding, Y., Korotkiy, M., Omelayenko, B., Kartseva, V., Zykov, V., Klein, M., ... & Fensel, D., 2002, April). Il classificatore Naïve-Bayes (NB) (C. J. C. Burges, 1998) è un algoritmo standard di classificazione del testo, con una lunga storia di applicazioni di successo e assegna la merce che ha la più alta probabilità di essere corretta. UNSPSC fornisce una gerarchia di quattro livelli per la classificazione dei prodotti: Segmento, Famiglia, Classe e Prodotto (Ding, Y., Korotkiy, M., Omelayenko, B., Kartseva, V., Zykov, V., Klein, M., ... & Fensel, D., 2002, April). Inoltre, è stata fatta esperienza che molti prodotti pre-classificati che sono stati ricevuti per le loro valutazioni non sono classificati fino al livello Commodity, ma solo fino ai livelli Classe o Famiglia (Ding, Y., Korotkiy, M., Omelayenko, B., Kartseva, V., Zykov, V., Klein, M., ... & Fensel, D., 2002, April). Pertanto, è stato costruito un classificatore gerarchico, che in realtà consiste di quattro classificatori, ognuno dei quali lavora a un livello corrispondente (vedi anche Chakrabarti, S. et al. (1997), Koller D. and Sahami M. (1997, July), Dumais S. and Chen H. (2000, August), e Agrawal R. and Srikant R. (2001, May)). Per addestrare gli algoritmi è stato scelto di adottare un campione casuale del 60% dal set di dati delle descrizioni dei prodotti è stato utilizzato come set di addestramento e il restante 40% di dati come set di test. È possibile ottenere una precisione fino al 90% in base all'apprendimento da un esempio pre-classificato (Ding, Y., Korotkiy, M., Omelayenko, B., Kartseva, V., Zykov, V., Klein, M., ... & Fensel, D., 2002, April).
GoldenBullet mira a meccanizzare il processo di classificazione dei dati di prodotto. I tassi di accuratezza compresi tra il 70% e il 98% indicano che questo processo può essere meccanizzato a un livello tale da poter ottenere una forte riduzione dei costi, il che è un prerequisito per un E-commerce scalabile. Recentemente, il Support Vector Machines si è dimostrato uno strumento molto utile per la categorizzazione del testo (cfr. C. J. C. Burges (1998), C. Cortes and V. Vapnik (1995), Joachims T. (1998), e Dumais S. and Chen H. (2000, August).

2.3.3 Modelli di classificazione di Machine Learning impiegati nell'ambito musicale

Anche in ambito musicale sono state impiegate tecniche di Machine Learning in cui negli ultimi anni la ricerca nel campo del Music Information Retrieval ha prodotto metodi di classificazione automatica per far fronte alla quantità di musica digitale disponibile. Un problema recente è la classificazione automatica dell'umore della musica, che consiste in un sistema che prende la forma d'onda di un brano musicale e produce etichette di testo che descrivono l'umore della musica (come felice, triste, ecc…). Utilizzando alcune semplici categorie di umore, la classificazione automatica basata su caratteristiche audio dà risultati promettenti in quanto studi psicologici precedenti (Besson M., Faita F., Peretz I., A.-M.Bonnel, and J. Requin, 1998) hanno mostrato che parte dell'informazione semantica delle canzoni risiede esclusivamente nei testi. Ciò significa che i testi contengono informazioni emotive importanti che non sono incluse nell'audio.

A tal proposito è stato riportato lo studio di Cyril Laurier, Jens Grivolla e Perfecto Herrera dal titolo "Multimodal Music Mood Classification using Audio and Lyrics" in cui propongono un lavoro di classificazione automatica per categorizzare le canzoni in base all'umore che combina le informazioni del testo e quelle dell'audio. Per questo studio è stato utilizzato un approccio categorico (categorical approach) per rappresentare l'umore con le seguenti categorie: felice, triste, arrabbiato, rilassato. Per comodità è stata condotta una classificazione binaria ovvero (felice/ non felice) e (arrabbiato/non arrabbiato), mentre per classificare la musica sono state estratte caratteristiche audio di diverso tipo ovvero timbriche, ritmiche, tonali e descrittori temporali (Laurier, C., Grivolla, J., & Herrera, P., 2008, December).

Per quanto concerne l'audio sono stati utilizzati l'SVM, la Logistic Regression e il Random Forest. Il miglior risultato è stato dato dall'SVM con un'accuratezza in tutte e 4 le categorie dell'80%, in particolare la categoria con il risultato migliore è stata "arrabbiato" del 98% con parametri predefiniti, normalizzando le caratteristiche e utilizzando un kernel polinomiale (a seguire poi i risultati migliori sono stati del Random Forest e della Logistic Regression). Per quanto riguarda i testi è stato utilizzato il KNN in cui si cercava di testare la somiglianza sui testi. Purtroppo però quest'approccio non ha dato grandi risultati in quanto l'accuratezza media è stata di circa il 60% (Laurier, C., Grivolla, J., & Herrera, P., 2008, December).

Infine, cercando di combinare informazioni audio e testo sono stati utilizzati diversi classificatori come Decision Tree, KNN, la Logistic Regression, il Random Forest e l'Svm (Laurier, C., Grivolla, J., & Herrera, P., 2008, December).

I risultati migliori li ha dati anche in questo caso l'Svm per le categorie "felice" e "triste (o non felice)" in cui la complementarietà di testo e audio aumenta significativamente l'accuratezza complessiva, mentre per la categoria "arrabbiato" e "rilassato (o non arrabbiato)" hanno ottenuto un'accuratezza migliore quando si è studiato solo l'audio rispetto ai risultati di testo e audio. I risultati confermano la rilevanza dei testi per trasmettere le emozioni o anche che lo stato d'animo espresso dalla musica e dai dati acustici viene correlato alle informazioni contenute nel testo (Laurier, C., Grivolla, J., & Herrera, P., 2008, December).

2.3.4 Modelli di classificazione di Machine Learning impiegati in ambito agroalimentare

Anche in campo agroalimentare sono stati utilizzati modelli di classificazione di Machine Learning come nello studio di Wilson Castro, Jimy Oblitas, Miguel De La Torre, Carlos Cotrina, Karen Bazán e Himer Avila-George (Senior Member, IEEE) dal titolo " Classification of Cape Gooseberry Fruit According to Its Level of Ripeness Using Machine Learning Techniques and Different Color Spaces", in cui si è voluto studiare che l'elaborazione e la classificazione combinata di immagini ottenute dalla spettroscopia del dominio del tempo dei Terahertz e l'utilizzo di algoritmi di apprendimento automatico possono essere utilizzati per classificare i diversi stati di maturazione dell'uva spina. La maggior parte di questi metodi richiede molto tempo, strutture di laboratorio sofisticate, operatori specializzati e molti reagenti chimici (Luchese C.L., P. D. Gurak, and L. D. F. Marczak, 2015). Per questo motivo l'industria ha bisogno di sviluppare metodi semplici, accurati, veloci, non distruttivi e online per valutare la qualità e la sicurezza dei prodotti agroalimentari, anche perché la

determinazione non distruttiva del composto dell'uva spina con il metodo di rilevazione spettrale è ancora una sfida a causa della variazione spettrale THZ causata da abbondanti variazioni biologiche, come le origini geografiche e le stagioni di raccolta (Castro, W., Oblitas, J., De-La-Torre, M., Cotrina, C., Bazán, K., & Avila-George, H., 2019).

Al fine di indagare il potenziale della spettroscopia Terahertz nel dominio del tempo per classificare gli stati di maturità dei frutti, sono stati esaminati gli spettri Terahertz (0,5-10 THz) di 4 stati di maturità dell'uva spina. L'uva spina continua a maturare dopo essere stata rimossa dalla pianta e ciò è evidenziato principalmente dai cambiamenti di colore della buccia (Brosnan T. and D.-W. Sun, 2004). Pertanto, per questa ricerca sono stati utilizzati 4 stati di maturazione, caratterizzati in base al loro colore sulla scala Cielab e al loro contenuto totale di antociani (Castro, W., Oblitas, J., De-La-Torre, M., Cotrina, C., Bazán, K., & Avila-George, H., 2019).

Le matrici di dati acquisite sono state sottoposte all'applicazione di MATLAB 2019b Classification Learner utilizzando 24 modelli di classificatori. Per ottenere una classificazione adeguata, sono stati utilizzati modelli di classificazione lineari (LDA) e non lineari (SVM). Questi modelli sono stati realizzati utilizzando l'applicazione Matlab Machine Learning, che ha permesso di esplorare il set di dati in modo interattivo, di selezionare le caratteristiche e di specificare gli schemi di validazione (Castro, W., Oblitas, J., De-La-Torre, M., Cotrina, C., Bazán, K., & Avila-George, H., 2019).

Tutti i modelli hanno utilizzato una validazione incrociata (15 pieghe). Quello migliore è stato il Fine Gaussian SVM, che ha ottenuto un'accuratezza dell'84,3%, con un Kernel Scale di 0,35 e un metodo multiclasse one-to-one. Questo tipo di modello è stato riportato più volte in lavori di ricerca sull'uso del Machine Learning per il riconoscimento delle immagini (Zakaluk R. and R. S. Ranjan, 2006), inoltre questo tipo di algoritmo è stato utilizzato con successo anche nel settore dell'imaging sanitario, dove insieme a KNN sono modelli di classificazione di successo (Du C.-J and D.-W. Sun, 2008). Nel caso specifico dell'uva spina, gli studi dimostrano che i diversi stadi di maturità sono legati alla concentrazione di antociani presenti nello sviluppo del frutto, tra cui la cianidina 3-galattoside, la petunidina 3-glicoside, la cianidina-3-acetil-esoside Arabasadi Z. et al. (2013, October).

2.3.5 Modelli di classificazione di Machine Learning impiegati in ambito politico

Lo studio condotto in ambito politico da parte di Konrad A. Ciecierski and Mariusz Kamola dal titolo "Comparison of Text Classification Methods for Government Documents" affronta la classificazione di documenti civili, legali e governativi in ambito ministeriale implementando tecniche di Machine Learning per attività come il rilevamento di spam, il rilevamento di tentativi di phishing o per il rilevamento di messaggi falsi. In particolare, grazie all'implementazione della classificazione automatizzata, un'istituzione governativa può eseguire la scansione di tutta la corrispondenza in arrivo e assegnarla automaticamente ai dipartimenti appropriati (Ciecierski, K. A., & Kamola, M., 2020).

I risultati presentati in questo lavoro di ricerca si basano sulla classificazione di due gruppi di documenti: il primo, dove una data classe di documenti è derivata dalla loro origine comune, il secondo, dove la classe è

determinata dalla destinazione comune dei documenti. L'obiettivo di questo lavoro è confrontare i risultati della classificazione dei documenti ottenuti utilizzando diversi approcci di apprendimento automatico utilizzati per due serie selezionate di documenti ufficiali del governo. Il problema della classificazione del testo è noto da tempo nel mondo del Machine Learning (Ciecierski, K. A., & Kamola, M., 2020). Esistono vari approcci e algoritmi ideati per questo scopo: da una parte i classificatori si concentrano sulle parole presenti nel testo o anche sull'ordine di queste parole, mentre i metodi basati su statistiche di occorrenza, come ad esempio TF-IDF, trattano il testo come una quantità di parole. Altri come LSTM (Long Short-Term Memory (LSTM) (Gers, F.A., Schmidhuber, J., Cummins,F., 1999) o (BERT) Bidirectional Encoder Representations from Transformers (Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018), si concentrano maggiormente sull'ordine delle parole nel testo (Aggarwal, C.C., Zhai, C., 2012). I modelli di classificazione del testo possono essere suddivisi in due grandi tipi: white-box e black-box. I primi consentono una facile spiegazione dei motivi che hanno portato a classificare un testo in un modo particolare, ma spesso hanno prestazioni inferiori a quelli della black-box. Questi ultimi offrono una precisione senza precedenti se sono disponibili molti dati invarianti nel tempo (Ciecierski, K. A., & Kamola, M., 2020). Il loro principale svantaggio, nonostante gli sforzi attuali, è l'opacità. Prima della vettorializzazione, tutti i documenti erano sottoposti a una procedura standard di tokenizzazione, ovvero estrazione delle parole e conversione delle lettere minuscole, rimozione della punteggiatura e delle stopword (le parole più frequenti) utilizzando uno strumento online, WoSeDon (Kedzia, P., Piasecki, M., Orlin´ska, M., 2016), che ha portato tutte le parole nel testo alla loro forma base.

Una volta vettorializzato, un documento è stato classificato in modo indipendente mediante due metodi cardine: Naive Bayes (NB) e Logistic Regression (LR) (Jurafsky, D., Martin, J.H., 2009), and Pedregosa, F., et al., 2011)). Il Naive Bayes cerca di ricostruire il modello che genererà i campioni osservati con la massima probabilità, mentre l'obiettivo della Logistic Regression è quello di identificare le caratteristiche (elementi di w) che contraddistinguono maggiormente le classi.

In altre parole, i tentativi della Logistic Regression (LR) descrivono in modo efficiente le differenze tra le classi in termini di caratteristiche del documento selezionate (Ciecierski, K. A., & Kamola, M., 2020).

Gli esperimenti di addestramento e classificazione del modello sono stati eseguiti in ciascun caso su un set di dati di input con un numero uguale di campioni in ciascuna classe. Questi dati di input sono stati suddivisi in modo casuale in set di train, sviluppo e test di dimensioni predefinite (Ciecierski, K. A., & Kamola, M., 2020). Mentre per i modelli basati su black-box, ovvero reti neurali, non vi è alcuna spiegazione del motivo per cui una determinata classe è stata scelta per un determinato documento, ma si ottiene non solo la classe ritenuta più adatta, ma anche altre classi che potrebbero essere di scelta alternativa. In conclusione, i risultati ottenuti utilizzando la Logistic Regression sono migliori di quelli ottenuti utilizzando la rete LSTM, da 0,823 a 0,805 per il corpus di interpellanza e da 0,988 a 0,920 per Journal of Laws (Ciecierski, K. A., & Kamola, M., 2020).

Capitolo 3. Materiali e metodi

3.1 Obiettivo della ricerca

Per gli E-tailer di moda sta diventando sempre più importante riuscire a catalogare velocemente i nuovi prodotti sul loro sito online in quanto ogni articolo è provvisto di caratteristiche diverse che lo contraddistingue.

Il presente lavoro di tesi si pone come obiettivo quello di permettere ad un E-tailer, più precisamente un E-tailer del settore fashion, di non dover categorizzare manualmente ogni prodotto presente nel catalogo online ma di poter velocizzare questa attività aziendale mediante l'Intelligenza Artificiale.

Verrà affrontato un problema di classificazione multiclasse utilizzando un modello di Machine Learning con apprendimento supervisionato, che viene addestrato su un set di dati contenente esempi di input e le relative etichette di classe corrispondenti.

In altre parole, un problema di classificazione consiste nel cercare di categorizzare le diverse tipologie di borse in diverse categorie o classi in base alle diverse caratteristiche.

L'obiettivo è sviluppare un modello di Machine Learning che possa assegnare correttamente un'etichetta di classe a una borsa a partire dalle sue caratteristiche. Durante il processo di addestramento del modello, vengono utilizzate diverse tipologie di borse con le rispettive etichette di classe così da imparare a riconoscere i modelli e le caratteristiche che distinguono le diverse categorie di prodotto.

La rilevanza di questa ricerca è inoltre data dalla sua scalabilità, ovvero dalla possibilità di essere estesa non solo alle borse ma anche a qualsiasi altro prodotto di ogni settore di mercato.

3.2 Strumenti utilizzati per la ricerca e descrizione del modello predittivo

Il lavoro di ricerca è stato condotto su Anaconda, ovvero una distribuzione dei linguaggi di programmazione Python e R per il calcolo scientifico (scienza dei dati, applicazioni di apprendimento automatico, elaborazione dati su larga scala, analisi predittiva, ecc.), che mira a semplificare l'installazione e la gestione delle librerie necessarie per l'analisi dei dati, offrendo un ambiente di sviluppo integrato completo per i professionisti del settore. Contiene un insieme di strumenti, tra cui l'ambiente di sviluppo integrato (IDE) (Red hat, 2019) Spyder, Python, Jupyter Notebook, JupyterLab, Oracle Data Science Service oltre ad altre librerie e strumenti (Anaconda.com).

In questa tesi le analisi si sono svolte utilizzando Spyder, ovvero un ambiente di sviluppo integrato (IDE) multipiattaforma open source per la programmazione scientifica nel linguaggio Python che viene fornito con l'installazione di Anaconda. È appositamente progettato per la programmazione scientifica e offre un'interfaccia utente che integra un editor di testo, un ambiente di esecuzione del codice, un visualizzatore di variabili e plot e strumenti per l'analisi dei dati. Spyder si integra con una serie di pacchetti importanti nello

stack scientifico di Python, tra cui NumPy, SciPy, Matplotlib, Pandas, Seaborn e Scikit-learn così come altri software open source.

3.2.1 Descrizione del modello predittivo

Prima di entrare nel vivo della ricerca, di seguito verrà descritto il processo di un modello predittivo che permette di riconoscere e classificare gli item mai visti prima:

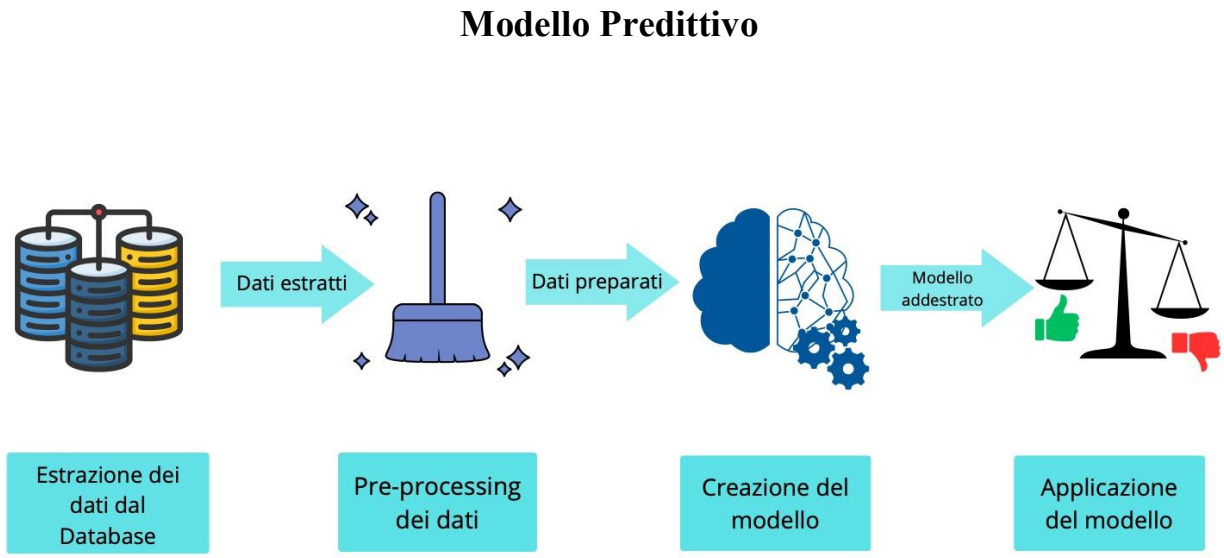


Figura 1: Modello predittivo

1) **Estrazione dei dati dal Database**

In questa prima fase vengono individuati i dati di interesse, per poi effettuare l'estrazione dei dati da un database, file csv...

L'output di questa fase iniziale è un insieme di dati che prende il nome di dataset composto da due parti distinte (che verranno splittate in seguito):

a) Trainig Set: parte del dataset che è utilizzata per addestrare un algoritmo di Machine Learning
b) Test Set: parte del dataset utilizzata per verificare se il modello utilizzato può funzionare bene nel riconoscere nuovi dati, e quindi di conseguenza la precisione dell'algoritmo.

2) **Pre-processing dei dati**

Nella seconda fase i dati estratti vengono così puliti e pre-processati: i valori mancanti vengono sostituiti con la media nel caso di features numeriche, mentre con la moda (ovvero il valore più frequente) nel caso di features categoriche per poi procedere con l'eliminazione dei valori duplicati e dei valori anomali.

Vengono selezionati ed estratti gli attributi utili ai fini della costruzione del modello sulla base di analisi statistiche utilizzando le tecniche di feature selection. L'output risultante, è un dataset pronto per la prossima fase, ovvero per essere utilizzato nella fase di creazione del modello.

3) **Creazione del modello**

Nella terza fase viene generato e addestrato il modello predittivo utilizzando i classificatori di Machine Learning.

In questa ricerca sono stati utilizzati e confrontati 6 algoritmi:

a) Random Forest
b) XG-Boost
c) KNN
d) Decision Tree
e) Logistic Regression
f) Naive Bayes

4) **Applicazione del modello**

Nell'ultima fase, il modello addestrato viene testato sulle features sconosciute e per ciascuno di essi il modello fornisce la sua predizione.

Successivamente si valuta la precisione del modello attraverso il calcolo delle metriche di Accuracy, Precision, Recall e F1 score. Vengono, poi, rappresentati i grafici della Confusion Matrix e della curva ROC. E infine, saranno mostrate le features più rilevanti per ogni algoritmo.

3.3 Descrizione del Dataset e delle Features

Dopo questa panoramica generale, possiamo passare alla descrizione del dataset utilizzato e delle features.

Il Dataset utilizzato per questo lavoro di ricerca proviene dall'Etailer di moda Mytheresa.

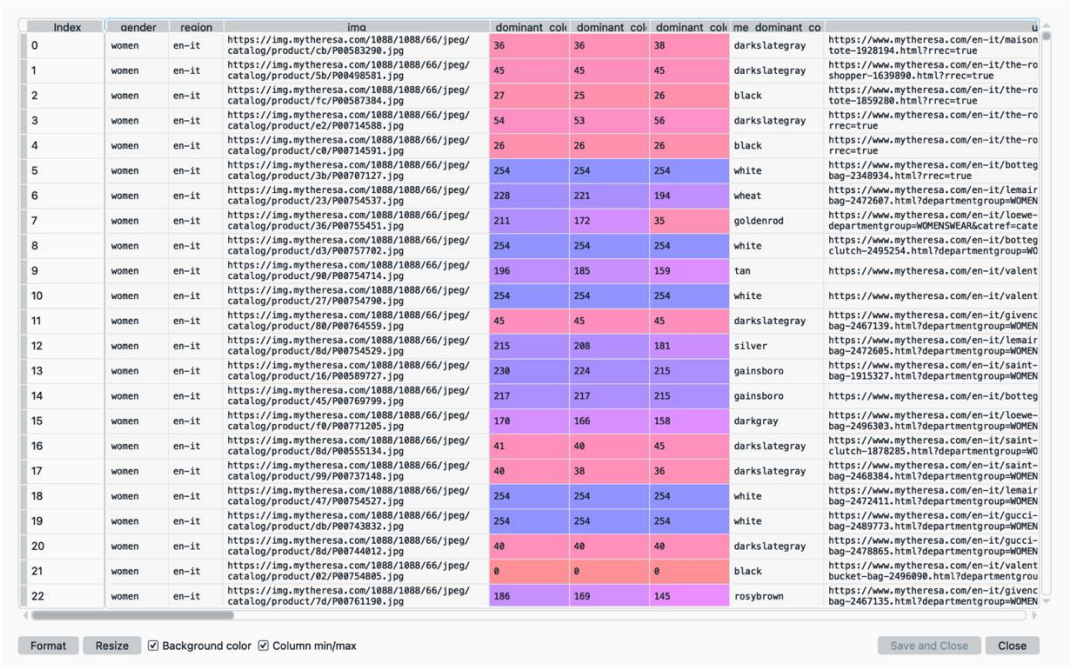
Mytheresa è un E-tailer di moda di lusso specializzato nella vendita di collezioni firmate di moda donna, uomo, bambino/a e articoli lifestyle. L'Etailer vanta una consolidata esperienza nel mondo della moda, frutto di oltre 30 anni di successi. Tutto è iniziato in una boutique nel centro di Monaco di Baviera specializzata in collezioni e designer internazionali che è diventata uno dei più innovativi e lussuosi E-tailer al mondo (Mytheresa.com, 2023). La loro accurata selezione permette di scegliere fra oltre 200 brand internazionali del lusso tra cui Bottega Veneta, Givenchy e Burberry e offre una piattaforma online semplice e intuitiva per un'esperienza di shopping proprio come in boutique che ti permette ogni giorno di essere sempre al passo con gli ultimi trend del mondo della moda (Mytheresa.com, 2023).

Inizialmente sono state fornite 10 cartelle (ognuna riferita a giorni diversi) ed ognuna di queste conteneva 8 database diversi:

1) Uomo Italia
2) Donna Italia
3) Uomo Giappone

4) Donna Giappone
5) Uomo Korea del Sud
6) Donna Korea del Sud
7) Uomo Stati Uniti d'America
8) Donna Stati Uniti d'America

Si è scelto quindi di condurre l'analisi unendo il database Uomo e Donna per la nazione Italia comprendente un totale di 4906 prodotti e 49 features che verranno descritte di seguito:



Index	gender	region	img	dominant col	dominant col	dominant col	me dominant co	u
0	women	en-it	https://img.mytheresa.com/1088/1088/66/jpeg/catalog/product/cb/P00583290.jpg	36	36	38	darkslategray	https://www.mytheresa.com/en-it/maison tote-1928194.html?rrec=true
1	women	en-it	https://img.mytheresa.com/1088/1088/66/jpeg/catalog/product/5b/P00498581.jpg	45	45	45	darkslategray	https://www.mytheresa.com/en-it/the-ro shopper-1639890.html?rrec=true
2	women	en-it	https://img.mytheresa.com/1088/1088/66/jpeg/catalog/product/fc/P00587384.jpg	27	25	26	black	https://www.mytheresa.com/en-it/the-ro tote-1859280.html?rrec=true
3	women	en-it	https://img.mytheresa.com/1088/1088/66/jpeg/catalog/product/e2/P00714588.jpg	54	53	56	darkslategray	https://www.mytheresa.com/en-it/the-ro rrec=true
4	women	en-it	https://img.mytheresa.com/1088/1088/66/jpeg/catalog/product/c0/P00714591.jpg	26	26	26	black	https://www.mytheresa.com/en-it/the-ro rrec=true
5	women	en-it	https://img.mytheresa.com/1088/1088/66/jpeg/catalog/product/3b/P00707127.jpg	254	254	254	white	https://www.mytheresa.com/en-it/botteg bag-2348934.html?rrec=true
6	women	en-it	https://img.mytheresa.com/1088/1088/66/jpeg/catalog/product/23/P00754537.jpg	228	221	194	wheat	https://www.mytheresa.com/en-it/lemair bag-2472607.html?departmentgroup=WOMEN
7	women	en-it	https://img.mytheresa.com/1088/1088/66/jpeg/catalog/product/36/P00755451.jpg	211	172	35	goldenrod	https://www.mytheresa.com/en-it/loewe- departmentgroup=WOMENSWEAR&catref=cate
8	women	en-it	https://img.mytheresa.com/1088/1088/66/jpeg/catalog/product/d3/P00757702.jpg	254	254	254	white	https://www.mytheresa.com/en-it/botteg clutch-2495254.html?departmentgroup=WO
9	women	en-it	https://img.mytheresa.com/1088/1088/66/jpeg/catalog/product/90/P00754714.jpg	196	185	159	tan	https://www.mytheresa.com/en-it/valent
10	women	en-it	https://img.mytheresa.com/1088/1088/66/jpeg/catalog/product/27/P00754790.jpg	254	254	254	white	https://www.mytheresa.com/en-it/valent
11	women	en-it	https://img.mytheresa.com/1088/1088/66/jpeg/catalog/product/80/P00764559.jpg	45	45	45	darkslategray	https://www.mytheresa.com/en-it/givenc bag-2467139.html?departmentgroup=WOMEN
12	women	en-it	https://img.mytheresa.com/1088/1088/66/jpeg/catalog/product/8d/P00754529.jpg	215	208	181	silver	https://www.mytheresa.com/en-it/lemair bag-2472605.html?departmentgroup=WOMEN
13	women	en-it	https://img.mytheresa.com/1088/1088/66/jpeg/catalog/product/16/P00589727.jpg	230	224	215	gainsboro	https://www.mytheresa.com/en-it/saint- bag-1915327.html?departmentgroup=WOMEN
14	women	en-it	https://img.mytheresa.com/1088/1088/66/jpeg/catalog/product/45/P00769799.jpg	217	217	215	gainsboro	https://www.mytheresa.com/en-it/botteg
15	women	en-it	https://img.mytheresa.com/1088/1088/66/jpeg/catalog/product/f0/P00771205.jpg	170	166	158	darkgray	https://www.mytheresa.com/en-it/loewe- bag-2496303.html?departmentgroup=WOMEN
16	women	en-it	https://img.mytheresa.com/1088/1088/66/jpeg/catalog/product/8d/P00555134.jpg	41	40	45	darkslategray	https://www.mytheresa.com/en-it/saint- clutch-1878285.html?departmentgroup=WO
17	women	en-it	https://img.mytheresa.com/1088/1088/66/jpeg/catalog/product/99/P00737148.jpg	40	38	36	darkslategray	https://www.mytheresa.com/en-it/saint- bag-2468384.html?departmentgroup=WOMEN
18	women	en-it	https://img.mytheresa.com/1088/1088/66/jpeg/catalog/product/47/P00754527.jpg	254	254	254	white	https://www.mytheresa.com/en-it/lemair bag-2472411.html?departmentgroup=WOMEN
19	women	en-it	https://img.mytheresa.com/1088/1088/66/jpeg/catalog/product/db/P00743832.jpg	254	254	254	white	https://www.mytheresa.com/en-it/gucci- bag-2489773.html?departmentgroup=WOMEN
20	women	en-it	https://img.mytheresa.com/1088/1088/66/jpeg/catalog/product/8d/P00744012.jpg	40	40	40	darkslategray	https://www.mytheresa.com/en-it/gucci- bag-2478865.html?departmentgroup=WOMEN
21	women	en-it	https://img.mytheresa.com/1088/1088/66/jpeg/catalog/product/02/P00754805.jpg	0	0	0	black	https://www.mytheresa.com/en-it/valent bucket-bag-2496090.html?departmentgrou
22	women	en-it	https://img.mytheresa.com/1088/1088/66/jpeg/catalog/product/7d/P00761190.jpg	186	169	145	rosybrown	https://www.mytheresa.com/en-it/givenc bag-2467135.html?departmentgroup=WOMEN

Format Resize ☑ Background color ☑ Column min/max Save and Close Close

Figura 2: Estratto del dataset di Mytheresa

- gender = genere uomo e donna
- region = regione Italia
- img = qui si trova l'url dell'immagine
- rgb_dominant_color/0, rgb_dominant_color/1, rgb_dominant_color/2 = numero corrispondente ai colori dominanti
- name_dominant_color = nome del colore dominante
- url = url del prodotto sul sito
- date_time = data
- number = codice del prodotto
- brand = nome del brand
- title = titolo con cui viene presentato il prodotto sul sito
- type = tipo di prodotto
- subtype = diverse tipologie di borse

- currency= simbolo della valuta utilizzata
- price = prezzo finale
- original_price = prezzo originale
- discount_percentage = percentuale di sconto applicata
- description = descrizione del prodotto
- status = nuovo arrivo o vecchia collezione
- materials/material = materiali utilizzati
- materials/internal details = materiali/ dettagli interni
- info_materials/0, info_materials/1, info_materials/2, info_materials/3, info_materials/4 = informazioni sulla provenienza dei materiali
- info_size&fit/0, info_size&fit/1, info_size&fit/2, info_size&fit/3, info_size&fit/4, info_size&fit/5, info_size&fit/6, info_size&fit/7, info_size&fit/8 = informazioni sulle misure e le taglie delle borse
- cluster_price = prezzo del cluster
- starting_date = data in cui è stato aggiunto il prodotto nel catalogo dell'E-tailer
- materials/colour of fastening, materials/colour of chain strap, materials/colour of strap= materiali/colore della chiusura, materiali/colore della cinghia a catena, materiali/colore della cinghia
- materials/trim= materiale finiture
- materials/Designer colour name = nome dei colori dei materiali
- materials/colour of fastening and strap = materiali/colore della chiusura e del cinturino
- materials/upper = materiali/superiore
- materials/lining = materiali/fodera
- materials/Material = materiali
- materials/care instructions = materiali/istruzioni per la cura e la manutenzione
- materials/details = dettagli dei materiali

3.4 Data Pre-processing

3.4.1 Analisi esplorativa del dataset

Una volta importate le librerie e i pacchetti su Python importiamo il dataset per poi analizzarlo.
Per prima cosa viene fatto un check del numero di variabili categoriche e numeriche presenti nel dataset e quindi di queste 49 features vediamo che 42 sono categoriche e 7 sono numeriche.
Il passaggio successivo consisterà nell'andare a pulire i nostri dati in quanto capita spesso che siano mancanti, incoerenti e rumorosi a causa della loro origine eterogenea.

3.4.2 Data-cleaning

Il dataset non presenta valori duplicati ma purtroppo alcune features presentano dei missing value (ovvero valori mancanti).
Si sceglie quindi di eliminare le features che presentano una percentuale di valori mancanti superiore al 50% e questo porta ad eliminare ben 19 features quindi da 49 features siamo passati a 30 features nel dataset.
Per le variabili rimanenti che presentano una percentuale di missing value minore del 50% scegliamo di tenerle nelle nostre analisi sostituendo questi valori con la media per le variabili numeriche, mentre per le variabili categoriche scegliamo di sostituirli con il valore più frequente, ovvero la moda.

Sempre in questa fase, molto importante poi è andare a vedere come sono distribuite le classi all'interno della variabile Y, ovvero la variabile 'Subtype' (anche detta variabile Target) soprattutto per constatare che le classi non siano sbilanciate tra loro.

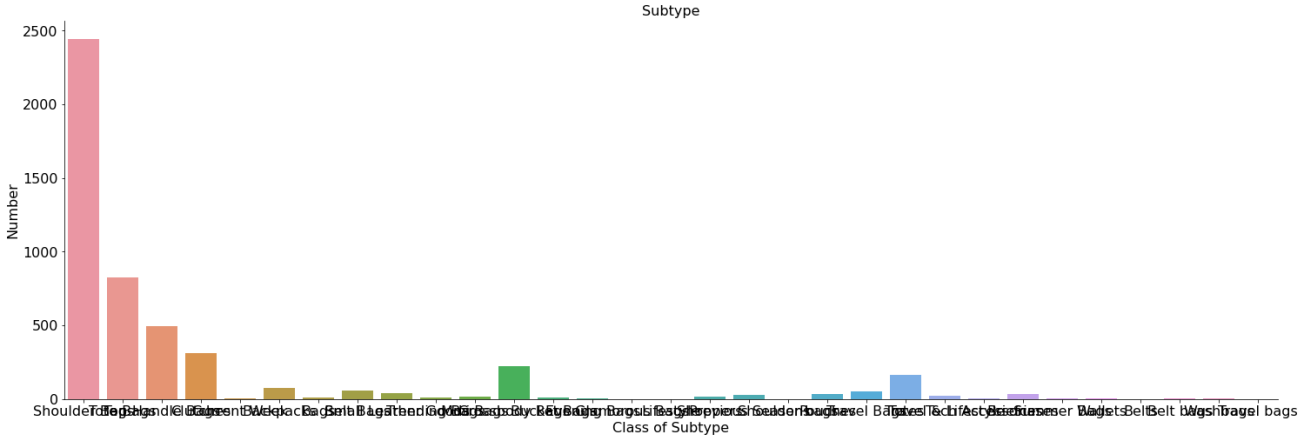


Figura 3: Rappresentazione Grafica della Variabile Y "Subtype"

Rappresentando graficamente la variabile 'Subtype', notiamo che le 31 classi che fanno parte della feature sono molto sbilanciate tra loro ed è quindi necessario correggere questo sbilanciamento perché se non viene fatto si rischia di arrivare a dei valori non buoni in termini di risultati.

In questo lavoro di tesi sono stati effettuati e messi a confronto due bilanciamenti:

1) Bilanciamento della Y (Variabile Target)
2) Bilanciamento del Train set

I risultati derivanti dal Bilanciamento della Y (1) mostrano le performance del modello predittivo a livello ideale e teorico, ovvero quali sarebbero stati se si fosse trattato di un caso concettuale, però i risultati corretti che consideriamo in questo lavoro di tesi, sono naturalmente quelli derivanti dal Bilanciamento del Train (2) perché qui viene costruito un modello predittivo rappresentativo del problema reale che si vuole risolvere riflettendo la natura dei dati che si incontreranno nel mondo reale. L'obiettivo principale dei dati di Train è

quello di fornire al modello di Machine Learning esempi realistici che gli consentano di imparare e generalizzare correttamente su nuovi dati.

3.4.3 Bilanciamento della Y (Variabile Target)

Per bilanciare la variabile Target si è scelto, come prima cosa, di eliminare le classi che contengono meno di 50 elementi e quindi le classi di borse rimanenti saranno solo 9:

- Shoulder Bags = 2446 elementi
- Tote Bags = 827 elementi
- Top-Handle Bags = 495 elementi
- Clutches = 313 elementi
- Crossbody Bags = 225 elementi
- Totes = 165 elementi
- Backpacks = 78 elementi
- Belt Bags = 55 elementi
- Travel Bags = 54 elementi

Per bilanciare la variabile Y viene utilizzata la tecnica del Resampling (Oversampling e Undersampling) definendo la classe Shoulder Bags come classe maggioritaria in quanto è quella che contiene più elementi di tutte (ovvero 2446 elementi), mentre tutte le altre come classi minoritarie.
E' stata utilizzata la tecnica dell'Oversampling per creare dati artificiali o duplicati o del campione di classe di minoranza per bilanciare l'etichetta di classe (Satyam Kumar, 2021, Sep 19). In altre parole, significa portare tutte le classi allo stesso numero della classe di minoranza o quella con il minor numero di righe che in questo caso si è scelto di metterla pari a 1000 (Nour Al-Rahman Al-Serw, 2021, Feb 21). Così poi possiamo utilizzare la tecnica dell'Undersampling per poter rimuovere o ridurre la maggior parte dei campioni di classe per bilanciare l'etichetta della classe maggioritaria (Satyam Kumar, 2021, Sep 19).

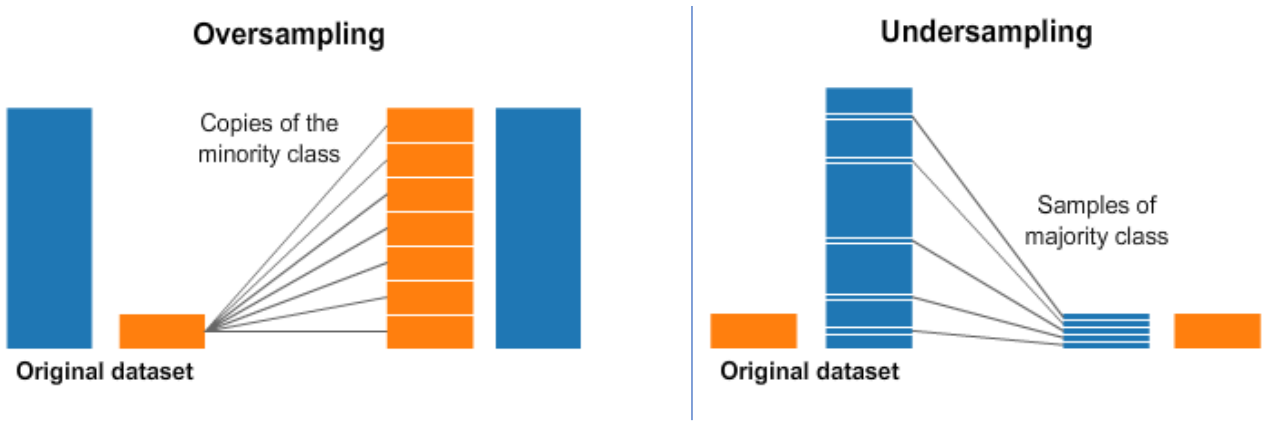


Figura 4: Oversampling e Undersampling

Fonte: https://medium.com/analytics-vidhya/undersampling-and-oversampling-an-old-and-a-new-approach-4f984a0e8392

Alla fine di tutto vengono unite tutte le classi bilanciate in un unico Dataframe così da poterlo avere finalmente bilanciato e poter procedere con le prossime analisi. Il Dataframe finale sarà composto da 9000 elementi totali (ovvero 1000 elementi per ognuna delle 9 classi della variabile Y) come possiamo notare dall'immagine seguente.

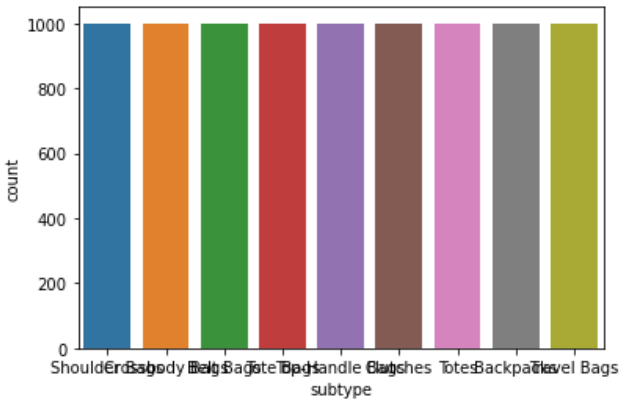


Figura 5: Variabile Subtype bilanciata

3.4.4 Bilanciamento del Train set

Per bilanciare il Train set, bisogna per prima cosa procedere a convertire tutte le features da categoriche in numeriche per poter fare un'analisi più accurata per poi procedere alla scelta di quelle che saranno le features che saranno scelte per la X. Per fare questa conversione utilizziamo il Label Encoder, ovvero una tecnica di codifica utilizzata per convertire le variabili categoriche in valori numerici.

Questo è spesso necessario perché molti algoritmi di Machine Learning richiedono dati di input numerici piuttosto che categorici. Quindi prima di splittare il dataset in Train set e Test set verrà definita la variabile X e la variabile Y.

La variabile X sarà composta da 26 features dopo aver eliminato 3 features in quanto non erano rilevanti per l'analisi (nei prossimi paragrafi verrà spiegato più nel dettaglio come si è arrivati alla scelta delle features).

Procediamo poi con lo split del dataset in Train set e Test set che in questo lavoro di ricerca è stato diviso in 80% in Train e 20% in Test.

(trainX, testX, trainY, testY) = train_test_split (X, Y, random_state =1, test_size= 0.20)

La procedura di suddivisione del dataset in Train set e Test set viene utilizzata per stimare le prestazioni degli algoritmi di apprendimento automatico quando vengono utilizzati per effettuare previsioni sui dati non utilizzati per addestrare il modello (Jason Brownlee, 2020, July 24). Come possiamo notare dalla figura 5,

anche in questo caso le classi della Variabile Y del Dataframe di Train sono sbilanciate tra di loro e dovremmo quindi procedere con il bilanciamento della Variabile Target.

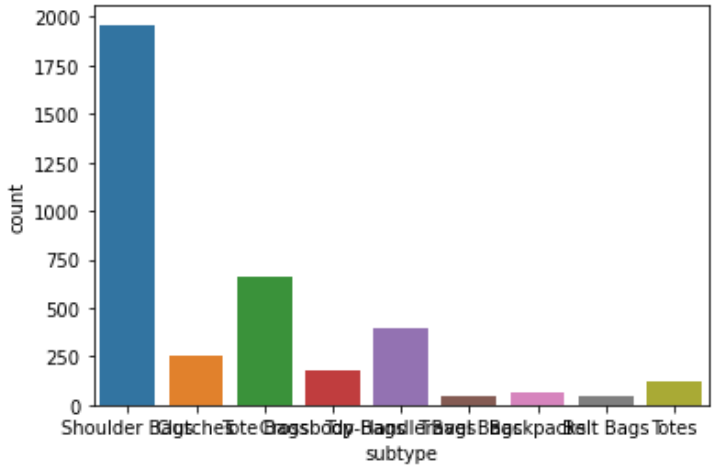


Figura 6: Variabile 'Subtype' sbilanciata

In dettaglio si è proceduto ad unire i due DataFrame, trainX e trainY, in un unico DataFrame, poi è stato contato il numero di campioni per classe per poi andare a determinare il numero di campioni per la classe più grande selezionando lo stesso numero di campioni casuali da ogni classe. Infine si è diviso il DataFrame bilanciato in trainX e trainY.

Il Dataframe finale (balanced_trainX e balanced_trainY) sarà composto da 17640 elementi totali (ovvero 1960 elementi per ognuna delle 9 classi della variabile Y).

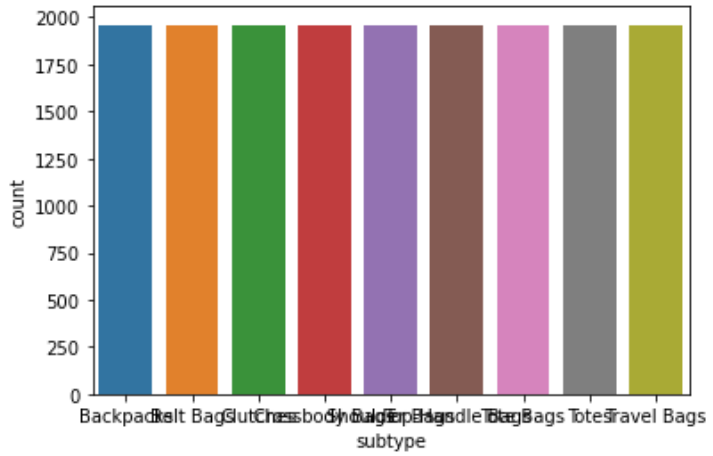


Figura 7: Variabile 'Subtype' bilanciata

3.5 Analisi statistiche

3.5.1 Analisi statistiche descrittive

Le analisi di statistica descrittiva sono molto importanti in quanto permettono di vedere come sono distribuite le variabili, ma soprattutto per capire se ci sono degli outlier, ovvero dei valori anomali da dover rimuovere dalla variabile presa in esame.
In questo lavoro è stato presentato un confronto tra i grafici di distribuzione e del boxplot dei due diversi bilanciamenti effettuati e quindi questo ci serve per vedere se da un bilanciamento all'altro ci sono stati dei risultati diversi per il grafico della distribuzione e del boxplot.
In statistica, un grafico di distribuzione è uno strumento visuale utilizzato per rappresentare la distribuzione dei dati di un determinato insieme di osservazioni. Esso fornisce una rappresentazione grafica della frequenza con cui si verificano diversi valori all'interno di un insieme di dati.
Un tipo comune di grafico di distribuzione è l'istogramma che è particolarmente utile per visualizzare la distribuzione di dati numerici o quantitativi.
L'obiettivo principale di un grafico di distribuzione è quello di fornire una rappresentazione visiva chiara della distribuzione dei dati, permettendo di identificare pattern, valori anomali, asimmetrie o altri comportamenti rilevanti. Questo aiuta a comprendere meglio le caratteristiche e le proprietà dei dati, fornendo informazioni utili per l'analisi statistica e la presa di decisioni.
Il boxplot, anche chiamato diagramma a scatola e baffi, è un tipo di grafico statistico utilizzato per rappresentare la distribuzione di un insieme di dati numerici. E' composto da un rettangolo (il "box") che rappresenta il range interquartile (IQR), ovvero l'intervallo tra il primo quartile (Q1) e il terzo quartile (Q3) dei dati. La mediana, che rappresenta il valore centrale della distribuzione, è rappresentata da una linea all'interno del rettangolo.
I "baffi" del boxplot si estendono dai lati del rettangolo e indicano l'estensione dei dati, escludendo i valori considerati valori anomali, più precisamente si estendono fino al valore minimo e massimo all'interno di un intervallo definito. I valori anomali, detti "outlier", vengono rappresentati come punti singoli al di fuori dei baffi. Questi punti potrebbero indicare valori che si discostano significativamente dalla distribuzione principale dei dati e per tale ragione bisogna rimuoverli.
In sintesi, il boxplot è un grafico che sintetizza le informazioni chiave sulla distribuzione dei dati, rendendo facile e veloce l'analisi e il confronto delle distribuzioni numeriche.

Dopo aver analizzato per ogni variabile il grafico di distribuzione e il boxplot per entrambi i bilanciamenti, si effettuerà un'analisi della matrice di correlazione mostrando se ci sono stati miglioramenti o peggioramenti nei valori degli indici di correlazione tra le variabili prima e dopo aver effettuato il Data-cleaning.

1. Gender

Bilanciamento Y

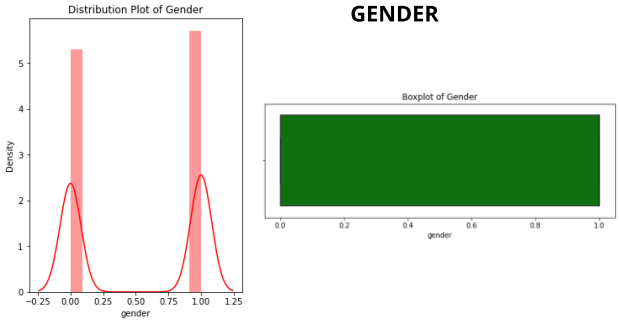


Figura 8

Bilanciamento Train set

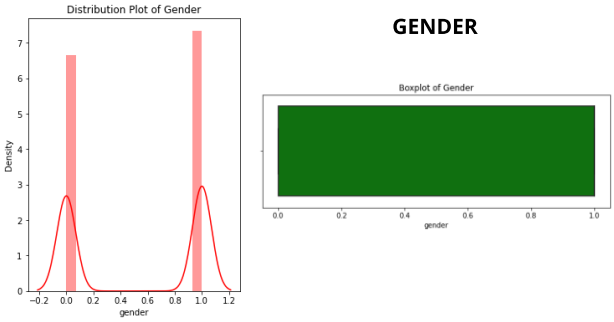


Figura 9

Nel grafico di distribuzione, che rappresenta la distribuzione dei valori numerici di un insieme di dati, possiamo notare che la variabile gender (uomo e donna) presenta un maggior numero di dati nella regione di valore 1, mentre si osserva una minore presenza di dati nella regione di valore 0. Questo ci indica che le tipologie di borse destinate alle donne sono leggermente più elevate (regione di valore 1) rispetto a quelle destinate agli uomini (regione di valore zero). L'andamento è molto simile per entrambi i grafici dei due diversi bilanciamenti.

I grafici boxplot di entrambi i bilanciamenti sono composti da una scatola rettangolare centrata sull'intervallo che va da 0 a 1 e per entrambi i grafici notiamo che non ci sono differenze significative. La maggior parte dei dati si concentra sulla zona centrale dell'intervallo senza presentare valori estremi o anomali.

Più nel dettaglio possiamo notare che il primo quartile, la linea mediana del rettangolo, il terzo quartile e il valore massimo del boxplot si posizionano su entrambi i grafici sul valore 1. Entrambi i boxplot non presentano valori anomali.

2. IMG

Bilanciamento Y

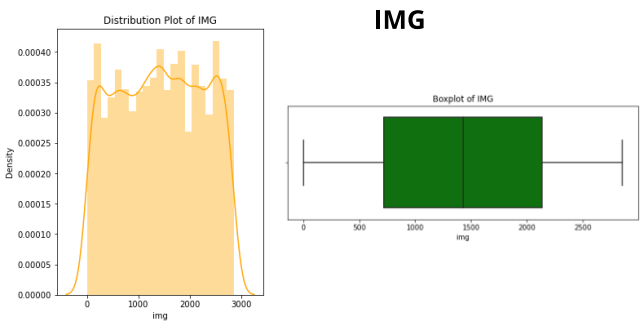


Figura 10

Bilanciamento Train set

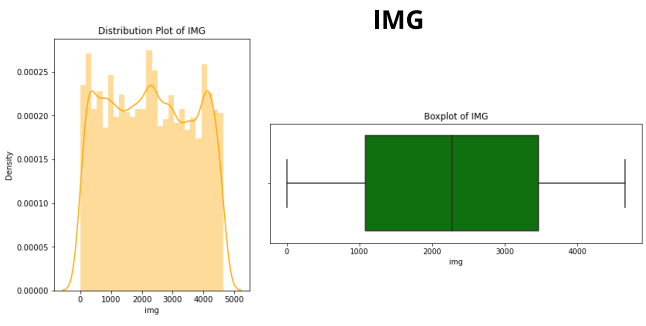


Figura 11

Nel grafico di distribuzione, che rappresenta la distribuzione dei valori numerici di un insieme di dati, possiamo notare che per il bilanciamento della Y i valori della variabile Img sono maggiormente concentrati nella regione centrale ovvero tra 0 e 2854, mentre nel grafico di distribuzione dove è stato effettuato il bilanciamento del Train i valori sono maggiormente concentrati nella regione compresa tra 1 e 4651.
I due grafici boxplot sono composti da una scatola rettangolare centrata sull'intervallo di valori che va da 0 a 2854 per quanto riguarda il grafico riferito al bilanciamento della Y, mentre sull'intervallo di valori che va da 0 a 4651 per il grafico riferito al bilanciamento del Train.
Il boxplot sulla sinistra presenta il baffo inferiore posizionato sullo 0, mentre il primo quartile si trova posizionato sul valore 717, la linea mediana del rettangolo su 1434, il terzo quartile su 2138 e in fine il baffo superiore è posizionato su 2854.
Mentre il boxplot sulla destra presenta il baffo inferiore posizionato sul valore 1, mentre il primo quartile si trova posizionato sul 717, la linea mediana del rettangolo su 1434, il terzo quartile su 2138 e in fine il baffo superiore è posizionato su 2854. Entrambi i grafici boxplot non presentano outlier.

3. Rgb_dominant_color_0

Bilanciamento Y

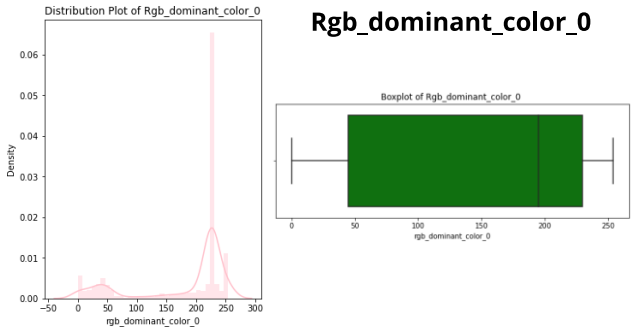


Figura 12

Bilanciamento Train set

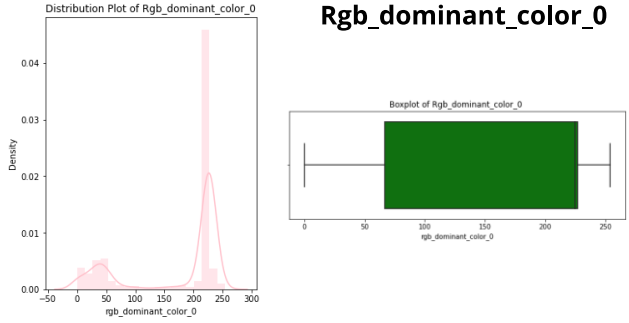


Figura 13

Nel grafico di distribuzione, possiamo notare che per il bilanciamento della Y della variabile rgb_dominant_color_0 i valori sono maggiormente concentrati nella regione di valore 225, mentre si osserva una minore presenza di dati nella regione compresa tra 0 e 200. Nel grafico di distribuzione dove è stato effettuato il bilanciamento del train i valori sono maggiormente concentrati nella regione di valore compresa tra 200 e 250, mentre si osserva una minore presenza di dati nella regione compresa tra 100 e 200.
Per entrambi i bilanciamenti i due grafici boxplot sono composti da una scatola rettangolare centrata sull'intervallo di valori che va da 0 a 254.
Il boxplot sulla sinistra presenta il baffo inferiore posizionato sullo 0, mentre il baffo superiore si posiziona sul valore 254. Il primo quartile si trova posizionato sul 49, la mediana su 197, il terzo quartile su 230.

Mentre il boxplot sulla destra presenta il baffo inferiore posizionato sul valore 1, mentre il primo quartile si trova posizionato sul 67. La linea mediana si trova molto vicina al terzo quartile, essa si trova sul valore 225, mentre il terzo quartile su 227. Entrambi i grafici boxplot non presentano outlier.

4. Rgb_dominant_color_1

Bilanciamento Y

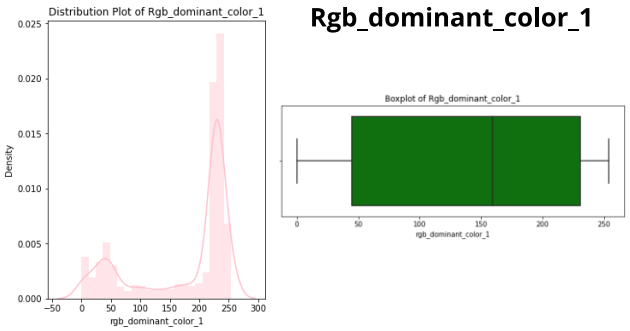


Figura 14

Bilanciamento Train set

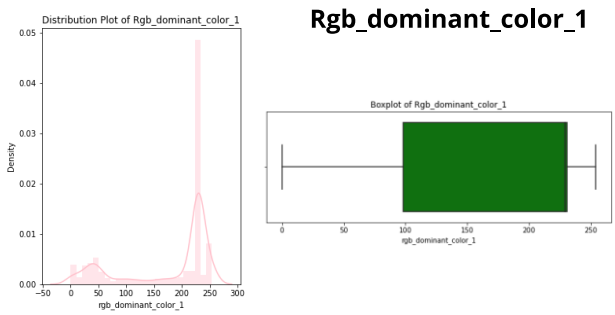


Figura 15

Nel grafico di distribuzione, possiamo notare che per il bilanciamento della Y i valori sono maggiormente concentrati nella regione compresa tra il valore 200 e 250, mentre si osserva una minore presenza di dati nella regione compresa tra 80 e 170. Nel grafico di distribuzione dove è stato effettuato il bilanciamento del train i valori sono maggiormente concentrati nella regione di valore 225, mentre si osserva una minore presenza di dati nella regione compresa tra 70 e 200.

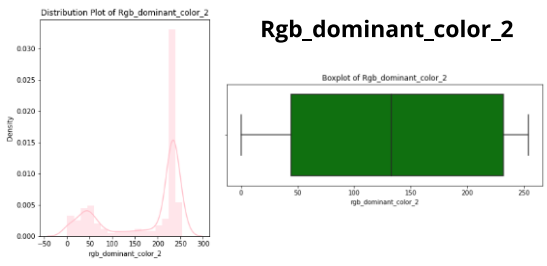
Per entrambi i bilanciamenti, i due grafici boxplot sono composti da una scatola rettangolare centrata sull'intervallo di valori che va da 0 a 254 come per la feature rgb_dominant_color_1.

Il boxplot sulla sinistra presenta il baffo inferiore posizionato sullo 0, mentre il baffo superiore si posiziona sul valore 254. Il primo quartile si trova posizionato sul 49, la mediana su 160, il terzo quartile su 235.

Mentre il boxplot sulla destra presenta il baffo inferiore posizionato sul valore 0, mentre il baffo superiore si posiziona sul valore 254. Il primo quartile si trova posizionato sul 98, la mediana si trova molto vicina al terzo quartile infatti la mediana sta sul valore 229 mentre il terzo quartile su 231. Entrambi i grafici boxplot non presentano outlier.

5. Rgb_dominant_color_2

Bilanciamento Y



Bilanciamento Train set

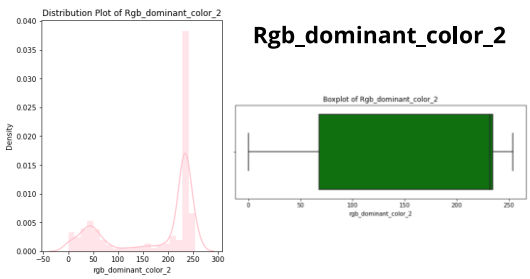


Figura 16 *Figura 17*

Nel grafico di distribuzione, possiamo notare che per il bilanciamento della Y i valori sono maggiormente concentrati nella regione di valore 225, mentre si osserva una minore presenza di dati nella regione compresa tra 70 e 200. Nel grafico di distribuzione dove è stato effettuato il bilanciamento del train i valori sono maggiormente concentrati nella regione di valore 225, mentre si osserva una minore presenza di dati nella regione compresa tra 70 e 200.

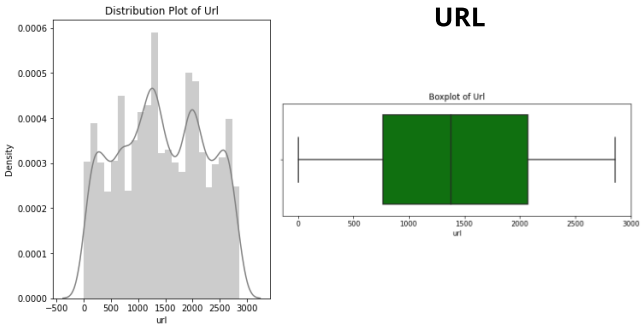
Per entrambi i bilanciamenti, i due grafici boxplot sono composti da una scatola rettangolare centrata sull'intervallo di valori che va da 0 a 254 come per la feature rgb_dominant_color_1 e la feature rgb_dominant_color_2.

Il boxplot sulla sinistra presenta il baffo inferiore posizionato sullo 0, mentre il baffo superiore si posiziona sul valore 254. Il primo quartile si trova posizionato sul 49, la mediana su 140, il terzo quartile su 235.

Mentre il boxplot sulla destra presenta il baffo inferiore posizionato sul valore 0, mentre il baffo superiore si posiziona sul valore 254. Il primo quartile si trova posizionato sul 68, la mediana si trova molto vicina al terzo quartile infatti la mediana sta sul valore 232 mentre il terzo quartile su 235. Entrambi i grafici boxplot non presentano outlier.

6. Url

Bilanciamento Y **Bilanciamento Train set**



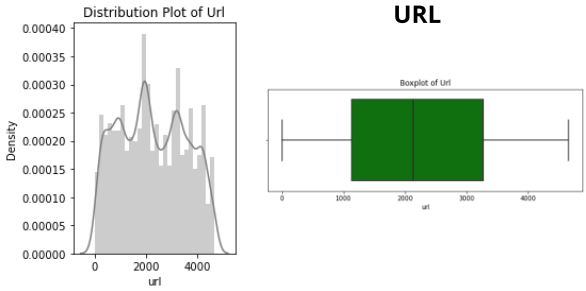


Figura 18 *Figura 19*

Nel grafico di distribuzione, possiamo notare che per il bilanciamento della Y i valori sono maggiormente concentrati nella regione compresa tra 0 e 2858. Nel grafico di distribuzione dove è stato effettuato il bilanciamento del train i valori sono maggiormente concentrati nella regione compresa tra 1 e 4657.

I due grafici boxplot sono composti da una scatola rettangolare centrata sull'intervallo di valori che va da 0 a 2858 per quanto riguarda il grafico riferito al bilanciamento della Y, mentre sull'intervallo di valori che va da 0 a 4657 per il grafico riferito al bilanciamento del Train.

Il boxplot sulla sinistra presenta il baffo inferiore posizionato sullo 0, mentre il baffo superiore si posiziona sul valore 2858. il primo quartile si trova posizionato sul valore 762, la mediana su 1377, il terzo quartile su 2069.

Mentre il boxplot sulla destra presenta il baffo inferiore posizionato sul valore 1, mentre il baffo superiore si posiziona sul valore 4657. Il primo quartile si trova posizionato su 1117.25, la mediana si trova sul valore 2137 mentre il terzo quartile su 3287. Entrambi i grafici boxplot non presentano outlier.

7. Date_time

Bilanciamento Y

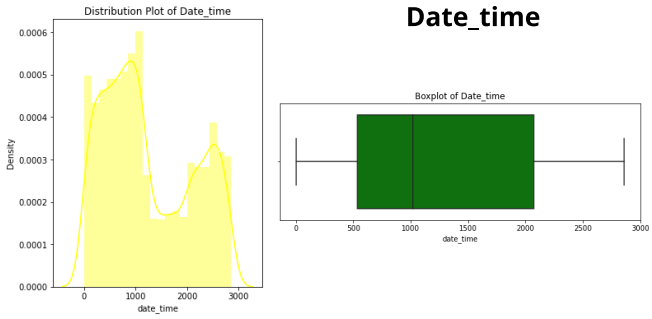


Figura 20

Bilanciamento Train set

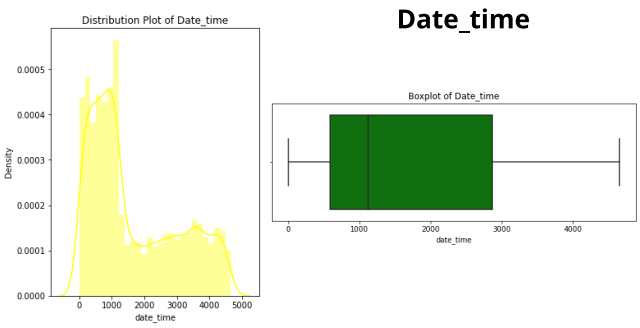


Figura 21

Nel grafico di distribuzione, possiamo notare che per il bilanciamento della Y i valori sono maggiormente concentrati nella regione compresa tra 0 e 2858. Nel grafico di distribuzione dove è stato effettuato il bilanciamento del train i valori sono maggiormente concentrati nella regione compresa tra 2 e 4656.

I due grafici boxplot sono composti da una scatola rettangolare centrata sull'intervallo di valori che va da 0 a 2858 per quanto riguarda il grafico riferito al bilanciamento della Y, mentre sull'intervallo di valori che va da 0 a 4657 per il grafico riferito al bilanciamento del Train.

Il boxplot sulla sinistra (riferito al bilanciamento della Y) presenta il baffo inferiore posizionato sullo 0, mentre il baffo superiore si posiziona sul valore 2858. Il primo quartile si trova posizionato sul valore 535, la mediana su 1020, il terzo quartile su 2069.

Mentre il boxplot sulla destra (riferito al bilanciamento del Train) presenta il baffo inferiore posizionato sul valore 0, mentre il baffo superiore si posiziona sul valore 4657. Il primo quartile si trova posizionato su 587, la mediana si trova sul valore 1120 mentre il terzo quartile su 2875. Entrambi i grafici boxplot non presentano outlier.

8. Number

Bilanciamento Y

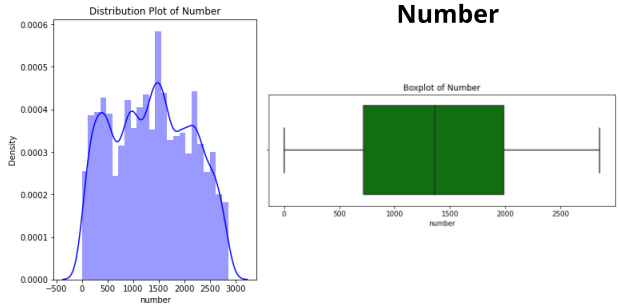


Figura 22

Bilanciamento Train set

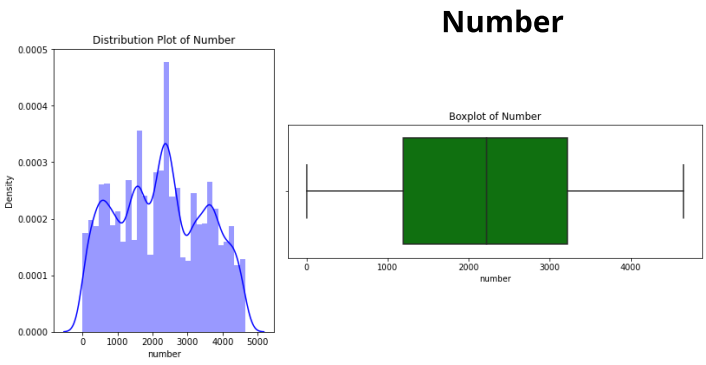


Figura 23

Nel grafico di distribuzione, possiamo notare che per il bilanciamento della Y i valori sono maggiormente concentrati nella regione compresa tra 0 e 2854. Nel grafico di distribuzione dove è stato effettuato il bilanciamento del Train i valori sono maggiormente concentrati nella regione compresa tra 2 e 4648.

I due grafici boxplot sono composti da una scatola rettangolare centrata sull'intervallo di valori che va da 0 a 2854 per quanto riguarda il grafico riferito al bilanciamento della Y, mentre sull'intervallo di valori che va da 0 a 4647 per il grafico riferito al bilanciamento del Train.

Il boxplot sulla sinistra presenta il baffo inferiore posizionato sullo 0, mentre il baffo superiore si posiziona sul valore 2854. Il primo quartile si trova posizionato sul valore 718,75, la mediana su 1365, il terzo quartile su 1986.

Mentre il boxplot sulla destra presenta il baffo inferiore posizionato sul valore 0, mentre il baffo superiore si posiziona sul valore 4647. Il primo quartile si trova posizionato su 1228, la mediana si trova sul valore 2234 mentre il terzo quartile su 3306. Entrambi i grafici boxplot non presentano outlier.

9. Brand

Bilanciamento Y

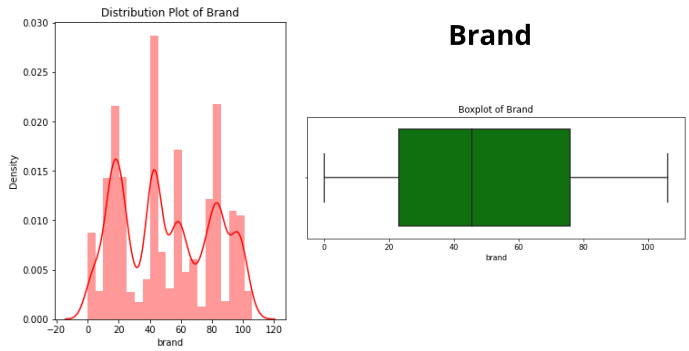


Figura 24

Bilanciamento Train set

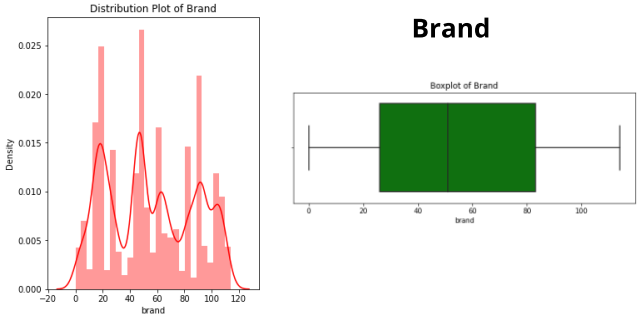


Figura 25

Nel grafico di distribuzione, possiamo notare che per il bilanciamento della Y i valori sono maggiormente concentrati nella regione compresa tra 0 e 106. Nel grafico di distribuzione dove è stato effettuato il bilanciamento del train i valori sono maggiormente concentrati nella regione compresa tra 0 e 114.
I due grafici boxplot sono composti da una scatola rettangolare centrata sull'intervallo di valori che va da 0 a 106 per quanto riguarda il grafico riferito al bilanciamento della Y, mentre sull'intervallo di valori che va da 1 a 114 per il grafico riferito al bilanciamento del Train.
Il boxplot sulla sinistra presenta il baffo inferiore posizionato sullo 0, mentre il baffo superiore si posiziona sul valore 106. Il primo quartile si trova posizionato sul valore 23, la mediana su 45.5, il terzo quartile su 76.
Mentre il boxplot sulla destra presenta il baffo inferiore posizionato sul valore 0, mentre il baffo superiore si posiziona sul valore 114. Il primo quartile si trova posizionato su 26, la linea mediana si trova sul valore 51 mentre il terzo quartile su 83. Entrambi i grafici boxplot non presentano outlier.

10. Title

Bilanciamento Y

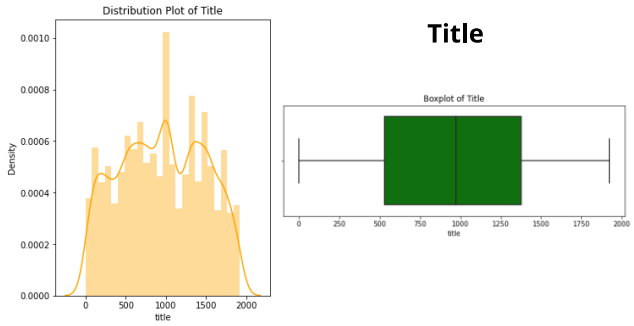


Figura 26

Bilanciamento Train set

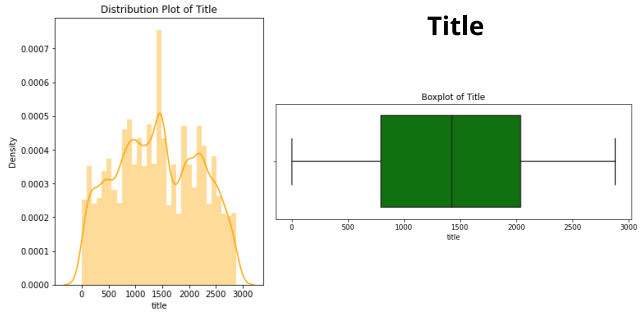


Figura 27

Nel grafico di distribuzione, possiamo notare che per il bilanciamento della Y i valori sono maggiormente concentrati nella regione compresa tra 0 e 1923 così come nel grafico nel grafico di distribuzione dove è stato effettuato il bilanciamento del Train.
I due grafici boxplot sono composti da una scatola rettangolare centrata sull'intervallo di valori che va da 0 a 1923 per quanto riguarda il grafico riferito al bilanciamento della Y, mentre sull'intervallo di valori che va da 0 a 2881 per il grafico riferito al bilanciamento del Train.
Il boxplot sulla sinistra presenta il baffo inferiore posizionato sullo 0, mentre il baffo superiore si posiziona sul valore 1923. Il primo quartile si trova posizionato sul valore 532, la mediana su 974.5, il terzo quartile su 1377.
Il boxplot sulla destra presenta il baffo inferiore posizionato sullo 0, mentre il baffo superiore si posiziona sul valore 2881. Il primo quartile si trova posizionato sul valore 812, la mediana su1429, il terzo quartile su 2034.
Entrambi i grafici boxplot non presentano outlier.

11. Type

Bilanciamento Y

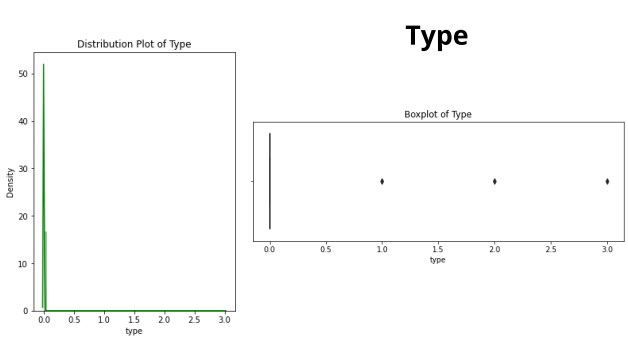


Figura 28

Bilanciamento Train set

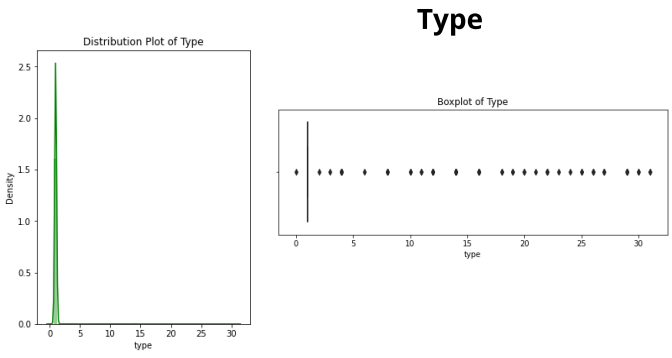


Figura 29

Nel grafico di distribuzione, possiamo notare che per il bilanciamento della Y i valori sono maggiormente concentrati nella regione di valore 0 con valore massimo 3. Nel grafico di distribuzione dove è stato effettuato il bilanciamento del Train i valori sono sempre concentrati nella regione di valore 0 ma la distribuzione si estende fino al valore 31. Anche se i due grafici sembrano avere una distribuzione molto simile, possiamo notare che il grafico a sinistra arriva fino a 50 mentre il grafico a destra arriva fino a 2,5.

I due grafici boxplot sono composti da una scatola rettangolare centrata sull'intervallo di valori che va da 0 a 3 per quanto riguarda il grafico riferito al bilanciamento della Y, mentre sull'intervallo di valori che va da 1 a 31 per il grafico riferito al bilanciamento del Train.

Il boxplot sulla sinistra presenta il baffo inferiore posizionato sullo 0, mentre il baffo superiore si posiziona sul valore3. Il primo quartile si trova posizionato sul valore 0, la mediana su 0, il terzo quartile su 0.

Mentre il boxplot sulla destra presenta il baffo inferiore posizionato sul valore 0, mentre il baffo superiore si posiziona sul valore 31. Il primo quartile si trova posizionato su 1, la mediana si trova sul valore 1 mentre il terzo quartile su 1. Entrambi i grafici boxplot presentano degli outlier.

La scelta finale è stata quella di non eliminarli, in quanto essendo la maggior parte dei valori concentrati sullo 0, cancellare i pochi valori positivi si sarebbe rilevata una scelta poco saggia in quanto poi avremmo dovuto eliminare la feature essendo che avrebbe contenuto solo valori pari a zero.

12. Discount_percentage

Bilanciamento Y

Figura 30

Bilanciamento Train set

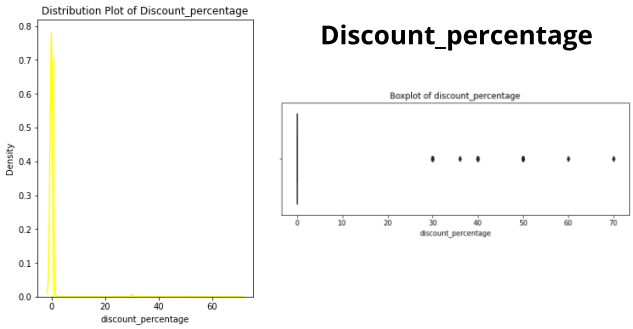


Figura 31

Nel grafico di distribuzione, possiamo notare che per il bilanciamento della Y i valori sono maggiormente concentrati nella regione di valore 0 con valore massimo 50. Nel grafico di distribuzione dove è stato effettuato il bilanciamento del Train i valori sono sempre concentrati nella regione di valore 0 ma la distribuzione si estende fino al valore 70. Anche se i due grafici sembrano avere una distribuzione molto simile, possiamo notare che il grafico a sinistra arriva fino a 1 mentre il grafico a destra arriva fino a 0.8 verticalmente.

I due grafici boxplot sono composti da una scatola rettangolare centrata sull'intervallo di valori che va da 0 a 50 per quanto riguarda il grafico riferito al bilanciamento della Y, mentre sull'intervallo di valori che va da 0 a 70 per il grafico riferito al bilanciamento del Train.

Il boxplot sulla sinistra presenta il baffo inferiore posizionato sullo 0, mentre il baffo superiore si posiziona sul valore 3. Il primo quartile si trova posizionato sul valore 0, la mediana su 0, il terzo quartile su 0.

Mentre il boxplot sulla destra presenta il baffo inferiore posizionato sul valore 0, mentre il baffo superiore si posiziona sul valore 70. Il primo quartile si trova posizionato su 0, la mediana si trova sul valore 0 mentre il terzo quartile su 0. Entrambi i grafici presentano degli outlier. La scelta finale è stata quella di non eliminarli, in quanto essendo la maggior parte dei valori concentrati sullo 0, cancellare i pochi valori positivi si sarebbe rilevata una scelta poco saggia perché poi avremmo dovuto eliminare tutta la variabile essendo che avrebbe contenuto solo valori pari a zero.

13. Description

Bilanciamento Y

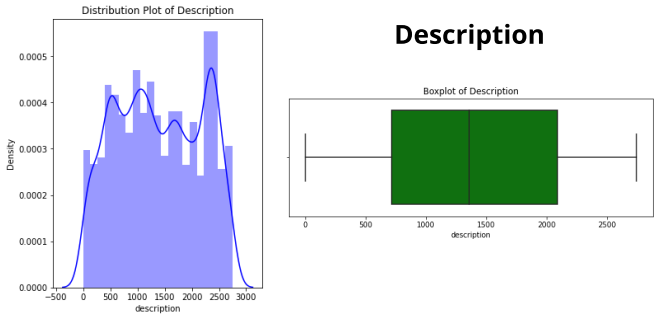


Figura 32

Bilanciamento Train set

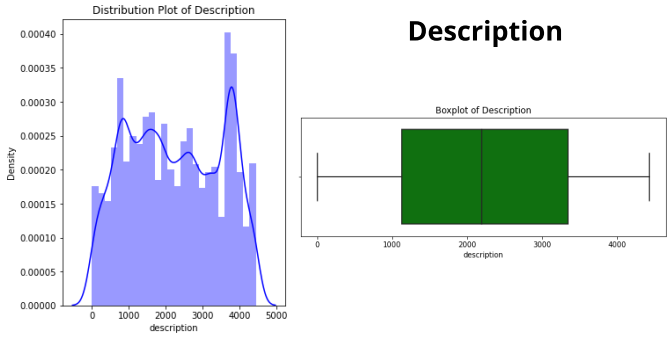


Figura 33

Nel grafico di distribuzione, possiamo notare che per il bilanciamento della Y i valori sono maggiormente concentrati nella regione compresa tra 0 e 2744. Nel grafico di distribuzione dove è stato effettuato il bilanciamento del train i valori sono maggiormente concentrati nella regione compresa tra 0 e 4431.

I due grafici boxplot sono composti da una scatola rettangolare centrata sull'intervallo di valori che va da 0 a 2744 per quanto riguarda il grafico riferito al bilanciamento della Y, mentre sull'intervallo di valori che va da 0 a 4431 per il grafico riferito al bilanciamento del Train.

Il boxplot sulla sinistra presenta il baffo inferiore posizionato sullo 0, mentre il baffo superiore si posiziona sul valore 2744. Il primo quartile si trova posizionato sul valore 715, la mediana su 1359, il terzo quartile su 2090.

Mentre il boxplot sulla destra presenta il baffo inferiore posizionato sul valore 0, mentre il baffo superiore si posiziona sul valore 4431. Il primo quartile si trova posizionato su 1124, la mediana si trova sul valore 2192 mentre il terzo quartile su 3373. Entrambi i grafici boxplot non presentano outlier.

14. Status

Bilanciamento Y

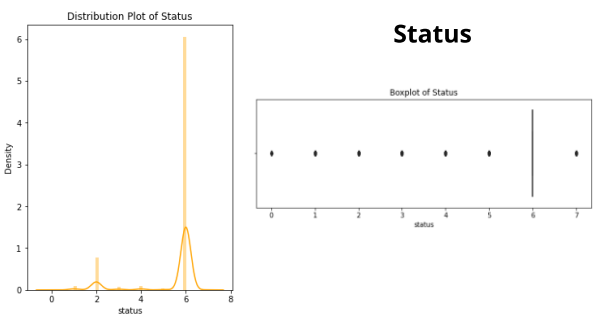


Figura 34

Bilanciamento Train set

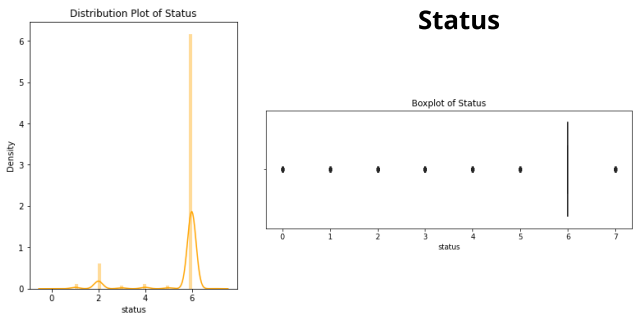


Figura 35

La variabile Status nei grafici dei due diversi bilanciamenti possiamo notare che presenta la stessa distribuzione e gli stessi valori. I valori sono maggiormente concentrati nella regione di valore 6 mentre sono meno concentrati nella regione di valore 2.

Entrambi i grafici boxplot sono composti da una scatola rettangolare centrata sull'intervallo di valori che va da 0 a 7.

Per entrambi i grafici il baffo inferiore è posizionato sullo 0, mentre il baffo superiore si posiziona sul valore 7. Il primo quartile si trova posizionato sul valore 6, così come anche la mediana e il terzo quartile.

Anche in questo caso il grafico presenta degli outlier e la scelta è stata quella di non eliminarli, in quanto essendo la maggior parte dei valori concentrati sul valore 6, cancellare i pochi valori positivi si sarebbe rilevata una scelta poco saggia perché poi avremmo dovuto eliminare tutta la variabile essendo che avrebbe contenuto solo valori pari a zero.

15. Info_materials_0

Bilanciamento Y

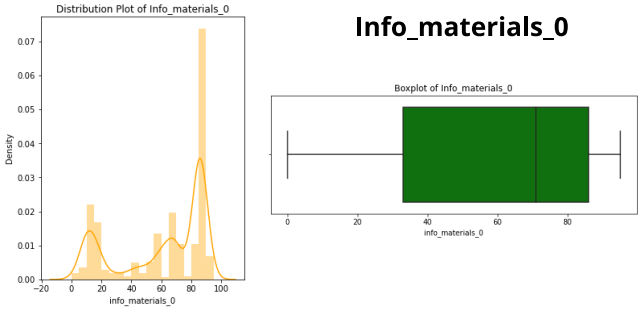


Figura 36

Bilanciamento Train set

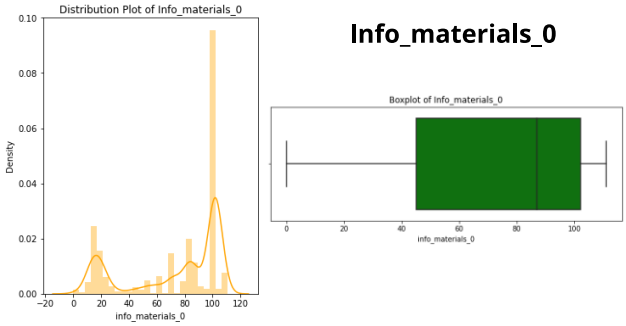


Figura 37

Nel grafico di distribuzione, possiamo notare che per il bilanciamento della Y i valori sono maggiormente concentrati nella regione compresa tra 0 e 95 con un picchio attorno all'85. Nel grafico di distribuzione dove è stato effettuato il bilanciamento del train i valori sono maggiormente concentrati nella regione compresa tra 0 e 111 con un picchio intorno al 100.

I due grafici boxplot sono composti da una scatola rettangolare centrata sull'intervallo di valori che va da 0 a 95 per quanto riguarda il grafico riferito al bilanciamento della Y, mentre sull'intervallo di valori che va da 1 a 111 per il grafico riferito al bilanciamento del Train.

Il boxplot sulla sinistra presenta il baffo inferiore posizionato sullo 0, mentre il baffo superiore si posiziona sul valore 95. Il primo quartile si trova posizionato sul valore 33, la mediana su 71, il terzo quartile su 86.

Mentre il boxplot sulla destra presenta il baffo inferiore posizionato sul valore 0, mentre il baffo superiore si posiziona sul valore 111. Il primo quartile si trova posizionato su 45, la mediana si trova sul valore 87 mentre il terzo quartile su 102. Entrambi i grafici boxplot non presentano outlier.

16. Info_size&fit_0

Bilanciamento Y

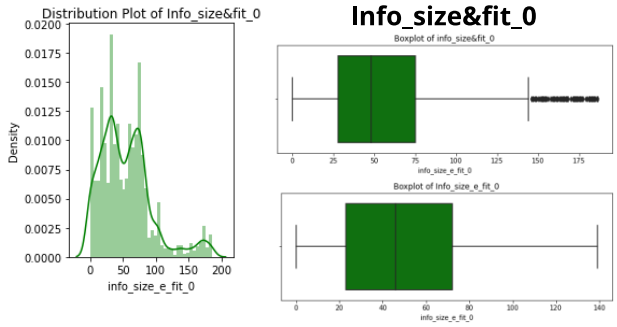


Figura 38

Bilanciamento Train set

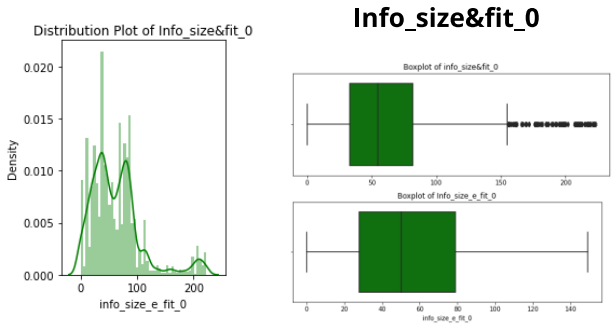


Figura 39

Nel grafico di distribuzione, possiamo notare che per il bilanciamento della Y i valori sono maggiormente concentrati nella regione compresa tra 0 e 186 con un picchio attorno al valore 30. Nel grafico di distribuzione dove è stato effettuato il bilanciamento del train i valori sono maggiormente concentrati nella regione compresa tra 0 e 223 con un picchio intorno al 50.

I due grafici boxplot sono composti da una scatola rettangolare centrata sull'intervallo di valori che va da 0 a 186 per quanto riguarda il grafico riferito al bilanciamento della Y, mentre sull'intervallo di valori che va da 0 a 223 per il grafico riferito al bilanciamento del Train. Notiamo che per entrambi i bilanciamenti i due boxplot presentano degli outlier che andremo opportunamente a rimuovere.

Partendo con il boxplot sulla sinistra presenta il baffo inferiore posizionato sullo 0, mentre il baffo superiore si posiziona sul valore 186. Il primo quartile si trova posizionato sul valore 28, la mediana su 48, il terzo quartile su 75. Procediamo rimuovendo dal boxplot i valori maggiori di 140 (escluso) e notiamo che sono stati tolti tutti gli outlier quindi il boxplot definitivo avrà una distribuzione di dati compresi nell'intervallo di valori tra 0 e 139, in particolare il baffo inferiore posizionato sullo 0, mentre il baffo superiore si posiziona sul valore 139. Il primo quartile si trova posizionato sul valore 23, la mediana su 46, il terzo quartile su 72 passando così da 6937 dati totali a 6591 valori totali.

Procediamo anche con il boxplot di destra che presenta il baffo inferiore posizionato sullo 0, mentre il baffo superiore si posiziona sul valore 223. Il primo quartile si trova posizionato sul valore 33, la mediana su 55, il terzo quartile su 82. Procediamo rimuovendo dal boxplot i valori maggiori di 150 (escluso) e notiamo che sono stati tolti tutti gli outlier quindi il boxplot definitivo avrà una distribuzione di dati compresi nell'intervallo di valori tra 0 e 149, in particolare il baffo inferiore posizionato sullo 0, mentre il baffo superiore si posiziona sul valore 149. Il primo quartile si trova posizionato sul valore 28, la mediana su 50, il terzo quartile su 79 passando così da 13528 dati totali a 12716 valori totali.

17. Info_size&fit_1

Bilanciamento Y

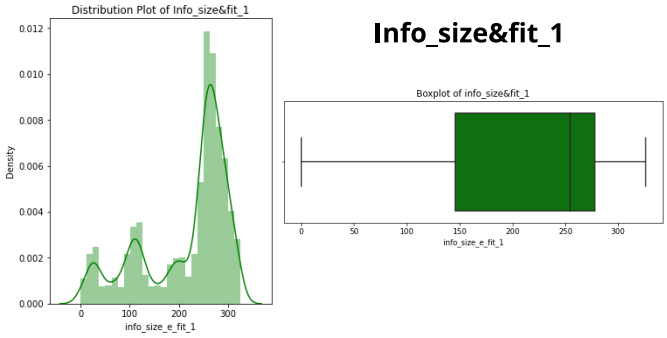


Figura 40

Bilanciamento Train set

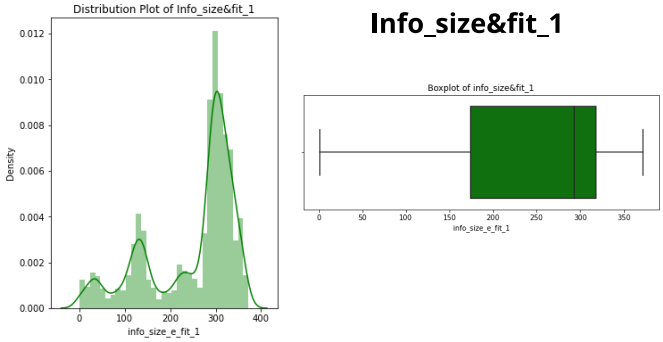


Figura 41

Nel grafico di distribuzione, possiamo notare che per il bilanciamento della Y i valori sono maggiormente concentrati nella regione compresa tra 200 e 326 con un picchio attorno all'250. Nel grafico di distribuzione dove è stato effettuato il bilanciamento del train i valori sono maggiormente concentrati nella regione compresa tra 300 e 372 con un picchio intorno al 300.

I due grafici boxplot sono composti da una scatola rettangolare centrata sull'intervallo di valori che va da 0 a 326 per quanto riguarda il grafico riferito al bilanciamento della Y, mentre sull'intervallo di valori che va da 1 a 372 per il grafico riferito al bilanciamento del Train.

Il boxplot sulla sinistra presenta il baffo inferiore posizionato sullo 0, mentre il baffo superiore si posiziona sul valore 326. Il primo quartile si trova posizionato sul valore 146, la mediana su 255, il terzo quartile su 278. Mentre il boxplot sulla destra presenta il baffo inferiore posizionato sul valore 1, mentre il baffo superiore si posiziona sul valore 372. Il primo quartile si trova posizionato su 168, la mediana si trova sul valore 293 mentre il terzo quartile su 318. Entrambi i grafici boxplot non presentano outlier.

18. Info_size&fit_2

Bilanciamento Y

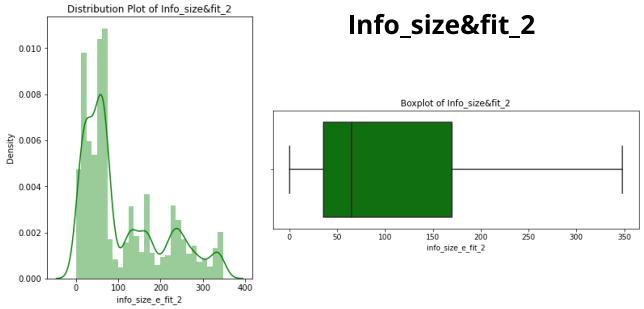


Figura 42

Bilanciamento Train set

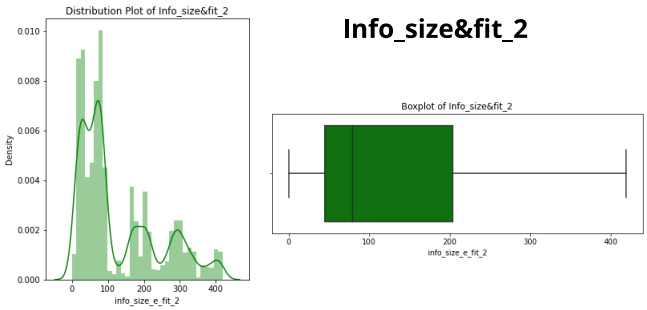


Figura 43

Nel grafico di distribuzione, possiamo notare che per il bilanciamento della Y i valori sono maggiormente concentrati nella regione compresa tra 0 e 100. Anche nel grafico di distribuzione dove è stato effettuato il bilanciamento del train i valori sono maggiormente concentrati nella regione compresa tra 0 e 100.
I due grafici boxplot sono composti da una scatola rettangolare centrata sull'intervallo di valori che va da 0 a 348 per quanto riguarda il grafico riferito al bilanciamento della Y, mentre sull'intervallo di valori che va da 0 a 419 per il grafico riferito al bilanciamento del Train.
Il boxplot sulla sinistra presenta il baffo inferiore posizionato sullo 0, mentre il baffo superiore si posiziona sul valore 348. Il primo quartile si trova posizionato sul valore 36, la mediana su 65, il terzo quartile su 170.
Mentre il boxplot sulla destra presenta il baffo inferiore posizionato sul valore 0, mentre il baffo superiore si posiziona sul valore 419. Il primo quartile si trova posizionato su 43, la mediana si trova sul valore 79 mentre il terzo quartile su 204. Entrambi i grafici boxplot non presentano outlier.

19. Info_size&fit_3

Bilanciamento Y

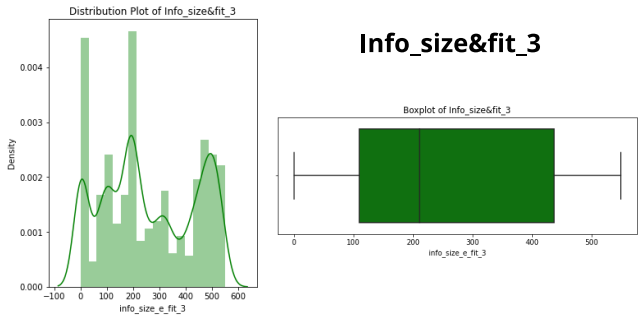


Figura 44

Bilanciamento Train set

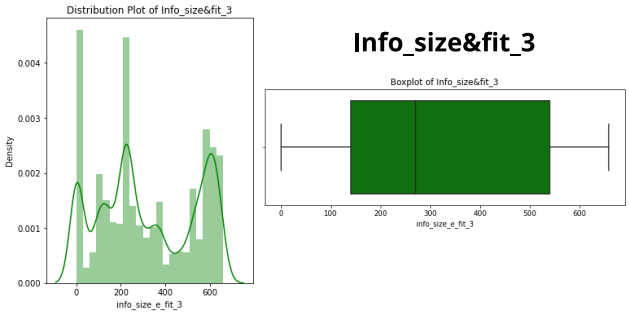


Figura 45

Nel grafico di distribuzione, possiamo notare che per il bilanciamento della Y i valori sono distribuiti nella regione compresa tra 0 e 549. Anche nel grafico di distribuzione dove è stato effettuato il bilanciamento del train i valori sono maggiormente concentrati nella regione compresa tra 0 e 658.
I due grafici boxplot sono composti da una scatola rettangolare centrata sull'intervallo di valori che va da 0 a 549 per quanto riguarda il grafico riferito al bilanciamento della Y, mentre sull'intervallo di valori che va da 0 a 658 per il grafico riferito al bilanciamento del Train.
Il boxplot sulla sinistra presenta il baffo inferiore posizionato sullo 0, mentre il baffo superiore si posiziona sul valore 549. Il primo quartile si trova posizionato sul valore 110, la mediana su 211, il terzo quartile su 437.
Mentre il boxplot sulla destra presenta il baffo inferiore posizionato sul valore 0, mentre il baffo superiore si posiziona sul valore 658. Il primo quartile si trova posizionato su 138, la mediana si trova sul valore 262.50 mentre il terzo quartile su 539. Entrambi i grafici boxplot non presentano outlier.

20. Info_size&fit_4

Bilanciamento Y

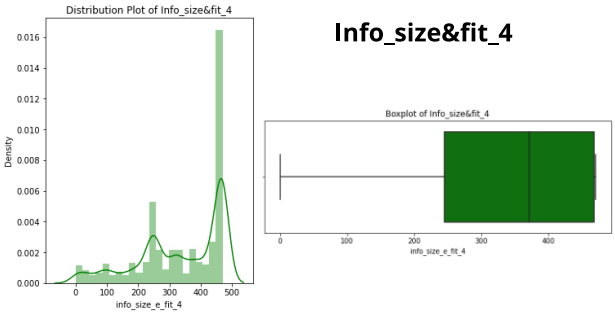


Figura 46

Bilanciamento Train set

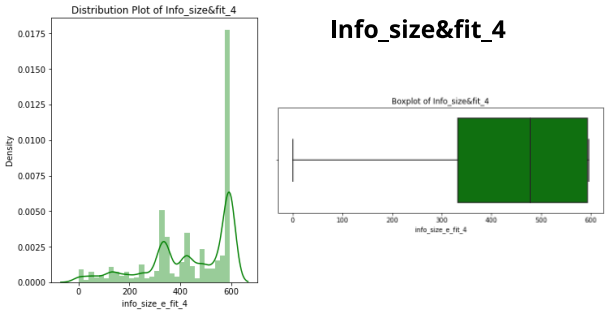


Figura 47

Nel grafico di distribuzione, possiamo notare che per il bilanciamento della Y i valori sono maggiormente concentrati nella regione di valore 470 e 250. Nel grafico di distribuzione dove è stato effettuato il bilanciamento del train i valori sono maggiormente concentrati nella regione di valore 596 e 300 (circa).

I due grafici boxplot sono composti da una scatola rettangolare centrata sull'intervallo di valori che va da 0 a 470 per quanto riguarda il grafico riferito al bilanciamento della Y, mentre sull'intervallo di valori che va da 0 a 596 per il grafico riferito al bilanciamento del Train.

Il boxplot sulla sinistra presenta il baffo inferiore posizionato sullo 0, mentre il baffo superiore si posiziona sul valore 470. Il primo quartile si trova posizionato su 245, la mediana su 371, il terzo quartile su 468.

Mentre il boxplot sulla destra presenta il baffo inferiore posizionato sul valore 0, mentre il baffo superiore è posizionato su 596. Il primo quartile si trova posizionato su 332, la mediana si trova molto vicina al terzo quartile infatti la mediana sta sul valore 480 mentre il terzo quartile su 594. Entrambi i grafici boxplot non presentano outlier.

21. Cluster Price

Bilanciamento Y

Figura 48

Bilanciamento Train set

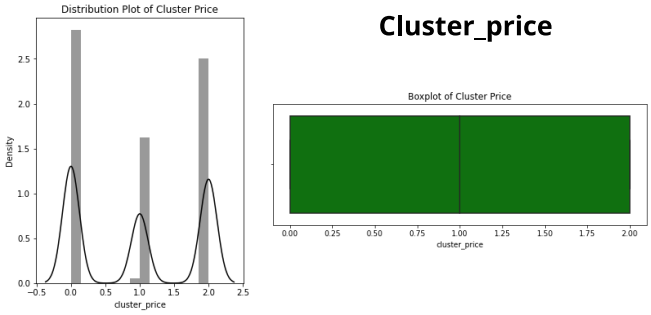


Figura 49

Nel grafico di distribuzione, possiamo notare che per il bilanciamento della Y i valori sono maggiormente concentrati nella regione di valore 0 e 2, mentre il picchio più basso si trova nella regione 1. Nel grafico di distribuzione dove è stato effettuato il bilanciamento del train i valori sono maggiormente concentrati nella regione di valore 0 e 2, mentre il picchio più basso si trova sempre nella regione 1.

Entrambi i due grafici boxplot sono composti da una scatola rettangolare centrata sull'intervallo di valori che va da 0 a 2 e per entrambi i grafici notiamo che non ci sono differenze significative. La maggior parte dei dati si concentra sulla zona centrale dell'intervallo. Il baffo inferiore di entrambi i boxplot si trova sullo 0, la mediana si trova posizionata sul valore 1, mentre il terzo quartile e il baffo superiore del boxplot si posizionano sul valore 2. Entrambi i grafici boxplot non presentano outlier.

22. Price

Bilanciamento Y

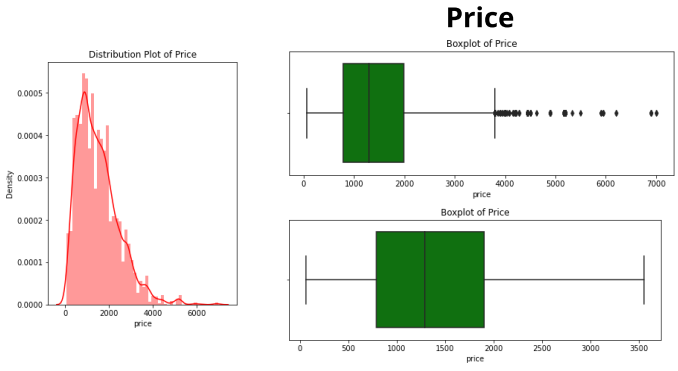


Figura 50

Bilanciamento Train set

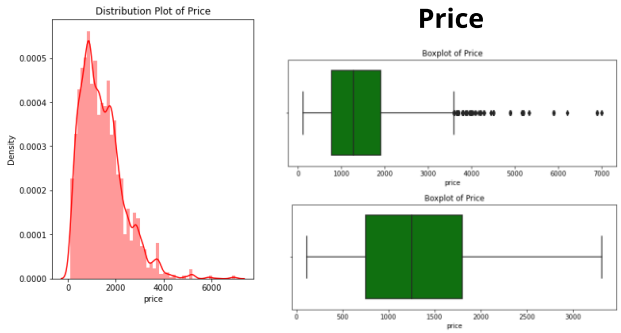


Figura 51

Nel grafico di distribuzione, possiamo notare che per il bilanciamento della Y i valori sono maggiormente concentrati nella regione compresa tra 60 e 7000. Nel grafico di distribuzione dove è stato effettuato il bilanciamento del train i valori sono maggiormente concentrati nella regione compresa tra 105 e 7000.

I due grafici boxplot sono composti da una scatola rettangolare centrata sull'intervallo di valori che va da 60 a 7000 per quanto riguarda il grafico riferito al bilanciamento della Y, mentre sull'intervallo di valori che va da 105 a 7000 per il grafico riferito al bilanciamento del Train.Notiamo che per entrambi i bilanciamenti i due boxplot presentano degli outlier che andremo opportunamente a rimuovere.

Partendo con il boxplot sulla sinistra il baffo inferiore è posizionato su 60, mentre il baffo superiore si posiziona sul valore 7000. Il primo quartile si trova posizionato sul valore 790, la mediana su 1300, il terzo quartile su 1990. Procediamo rimuovendo dal boxplot i valori maggiori di 3500 (escluso) e notiamo che sono stati tolti tutti gli outlier quindi il boxplot definitivo avrà una distribuzione di dati compresi nell'intervallo di valori tra 60 e 3550, in particolare il baffo inferiore posizionato su 60, mentre il baffo superiore si posiziona sul valore 3550. il primo quartile si trova posizionato sul valore 790, la mediana su 1290, il terzo quartile su 1900 passando così da 9000 dati totali a 8703 valori totali.

Procediamo anche con il boxplot di destra che presenta il baffo inferiore posizionato su 105, mentre il baffo superiore si posiziona sul valore 7000. Il primo quartile si trova posizionato sul valore 765, la mediana su 1280, il terzo quartile su 1900. Procediamo rimuovendo dal boxplot i valori maggiori di 3000 (escluso) e notiamo che sono stati tolti tutti gli outlier quindi il boxplot definitivo avrà una distribuzione di dati compresi nell'intervallo di valori tra 105 e 3315, in particolare il baffo inferiore posizionato su 105, mentre il baffo superiore si posiziona sul valore 3315. Il primo quartile si trova posizionato sul valore 750, la mediana su 1250, il terzo quartile su 1850 passando così da 17640 dati totali a 17026 valori totali.

23. Original price

Bilanciamento Y

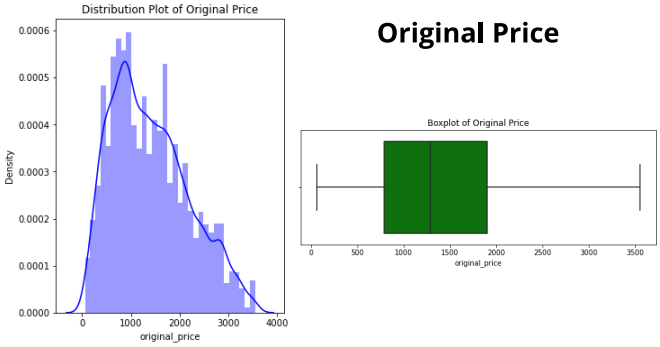


Figura 52

Bilanciamento Train set

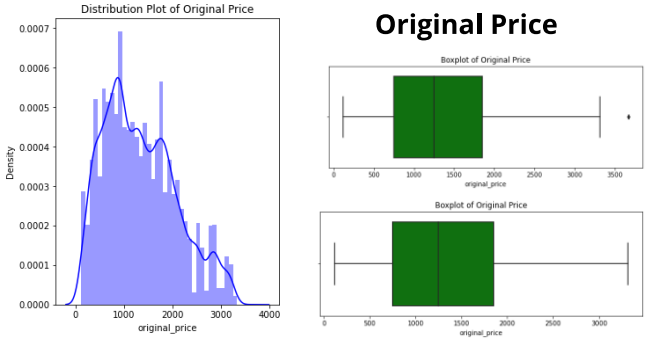


Figura 53

Nel grafico di distribuzione, possiamo notare che per il bilanciamento della Y i valori sono maggiormente concentrati nella regione compresa tra 60 e 3555. Nel grafico di distribuzione dove è stato effettuato il bilanciamento del train i valori sono maggiormente concentrati nella regione compresa tra 115 e 3670.

I due grafici boxplot sono composti da una scatola rettangolare centrata sull'intervallo di valori che va da 60 a 3555 per quanto riguarda il grafico riferito al bilanciamento della Y, mentre sull'intervallo di valori che va da 115 a 3670 per il grafico riferito al bilanciamento del Train. Notiamo che per il grafico con il bilanciamento del train è presente un valore anomalo che andremo opportunamente a rimuovere.

Il boxplot sulla sinistra non presenta valori anomali quindi andremo semplicemente a descrivere la sua distribuzione il baffo inferiore è posizionato su 790, mentre il baffo superiore si posiziona sul valore 3550.

Il primo quartile si trova posizionato sul valore 790, la mediana su 1290, il terzo quartile su 1990.

Procediamo anche con il boxplot di destra che presenta il baffo inferiore posizionato su 115, mentre il baffo superiore si posiziona sul valore 3670. Il primo quartile si trova posizionato sul valore 750, la mediana su 1250, il terzo quartile su 1850. Procediamo rimuovendo dal boxplot il valore anomalo e notiamo che sono stati tolti tutti gli outlier quindi il boxplot definitivo avrà una distribuzione di dati compresi nell'intervallo di valori tra 115 e 3315, in particolare il baffo inferiore posizionato su 115, mentre il baffo superiore si posiziona sul valore 3315. Il primo quartile si trova posizionato sul valore 750, la mediana su 1250, il terzo quartile su 1850 passando così da 17026 dati totali a 17025 valori totali.

24. Name_dominant_color

Bilanciamento Y

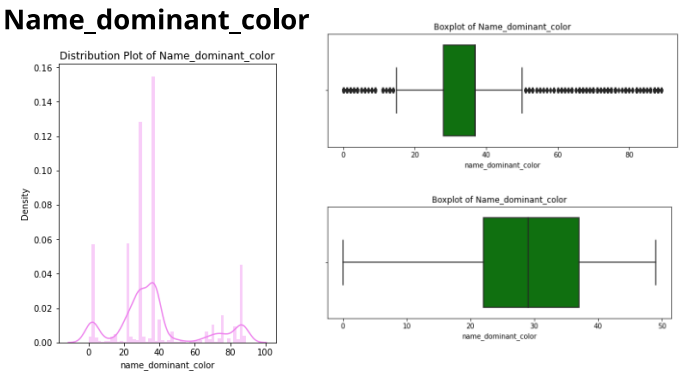


Figura 54

Bilanciamento Train set

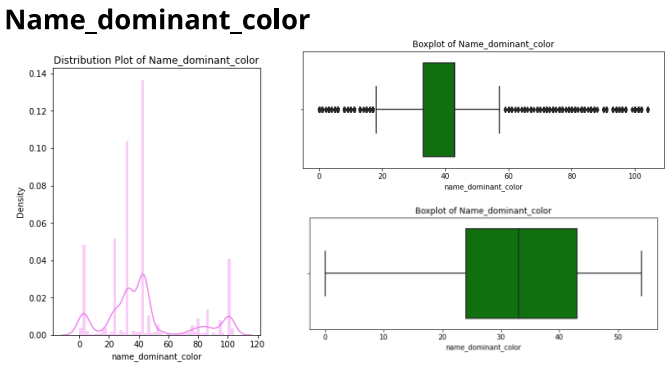


Figura 55

Nel grafico di distribuzione, possiamo notare che per il bilanciamento della Y i valori sono distribuiti nella regione compresa tra 0 e 89. Nel grafico di distribuzione dove è stato effettuato il bilanciamento del train i valori sono maggiormente concentrati nella regione compresa tra 0 e 104.

I due grafici boxplot sono composti da una scatola rettangolare centrata sull'intervallo di valori che va da 0 a 89 per quanto riguarda il grafico riferito al bilanciamento della Y, mentre sull'intervallo di valori che va da 0 a 104 per il grafico riferito al bilanciamento del Train. Notiamo che per entrambi i bilanciamenti i due boxplot presentano degli outlier che andremo opportunamente a rimuovere.

Partendo dal boxplot sulla sinistra il baffo inferiore è posizionato su 0, mentre il baffo superiore si posiziona sul valore 89. Il primo quartile si trova posizionato sul valore 28, la mediana su 37, il terzo quartile su 37. Procediamo rimuovendo dal boxplot i valori maggiori di 50 (escluso) e notiamo che sono stati tolti tutti gli outlier quindi il boxplot definitivo avrà una distribuzione di dati compresi nell'intervallo di valori tra 0 e 49, in particolare il baffo inferiore posizionato su 0, mentre il baffo superiore si posiziona sul valore 49. Il primo quartile si trova posizionato sul valore 22, la mediana su 29, il terzo quartile su 37 passando così da 8703 dati totali a 7050 valori totali.

Procediamo anche con il boxplot di destra che presenta il baffo inferiore posizionato su 0, mentre il baffo superiore si posiziona sul valore 104. Il primo quartile si trova posizionato sul valore 33, la mediana su 43, il terzo quartile su 43. Procediamo rimuovendo dal boxplot i valori maggiori di 50 (escluso) e notiamo che sono stati tolti tutti gli outlier quindi il boxplot definitivo avrà una distribuzione di dati compresi nell'intervallo di valori tra 0 e 54, in particolare il baffo inferiore posizionato su 0, mentre il baffo superiore si posiziona sul valore 54. Il primo quartile si trova posizionato sul valore 24, la mediana su 33, il terzo quartile su 43 passando così da 17025 dati totali a 13745 valori totali.

25. Materials_internal_details

Bilanciamento Y

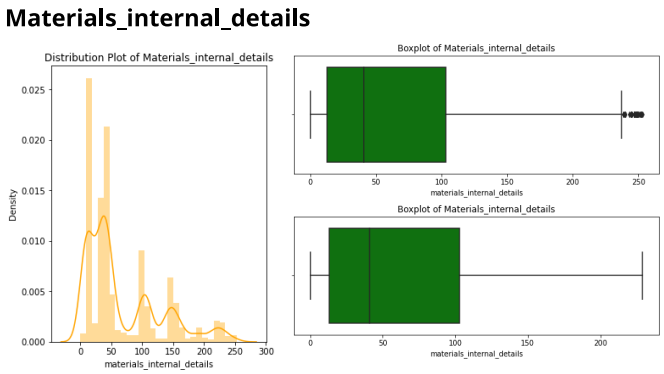


Figura 56

Bilanciamento Train set

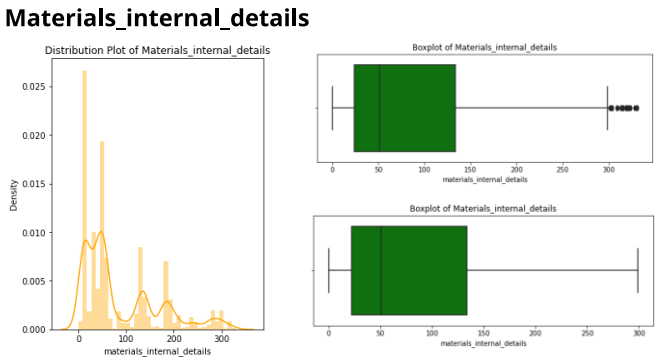


Figura 57

Nel grafico di distribuzione, possiamo notare che per il bilanciamento della Y i valori sono distribuiti nella regione compresa tra 0 e 253. Nel grafico di distribuzione dove è stato effettuato il bilanciamento del train i valori sono maggiormente concentrati nella regione compresa tra 0 e 331.

I due grafici boxplot sono composti da una scatola rettangolare centrata sull'intervallo di valori che va da 0 a 253 per quanto riguarda il grafico riferito al bilanciamento della Y, mentre sull'intervallo di valori che va da 0 a 331 per il grafico riferito al bilanciamento del Train. Notiamo che per entrambi i bilanciamenti i due boxplot presentano degli outlier che andremo opportunamente a rimuovere.

Partendo dal boxplot sulla sinistra il baffo inferiore è posizionato su 0, mentre il baffo superiore si posiziona sul valore 253. il primo quartile si trova posizionato sul valore 13, la mediana su 41, il terzo quartile su 103. Procediamo rimuovendo dal boxplot i valori maggiori di 230 (escluso) e notiamo che sono stati tolti tutti gli outlier quindi il boxplot definitivo avrà una distribuzione di dati compresi nell'intervallo di valori tra 0 e 229, in particolare il baffo inferiore posizionato su 0, mentre il baffo superiore si posiziona sul valore 229. Il primo quartile si trova posizionato sul valore 13, la mediana su 41, il terzo quartile su 103 passando così da 7050 dati totali a 6937 valori totali.

Procediamo anche con il boxplot di destra che presenta il baffo inferiore posizionato su 0, mentre il baffo superiore si posiziona sul valore 331. Il primo quartile si trova posizionato sul valore 24, la mediana su 51, il terzo quartile su 134. Procediamo rimuovendo dal boxplot i valori maggiori di 300 (escluso) e notiamo che sono stati tolti tutti gli outlier quindi il boxplot definitivo avrà una distribuzione di dati compresi nell'intervallo di valori tra 0 e 299, in particolare il baffo inferiore posizionato su 0, mentre il baffo superiore si posiziona sul valore 299. Il primo quartile si trova posizionato sul valore 24, la mediana su 51, il terzo quartile su 134 passando così da 13745 dati totali a 13572 valori totali.

26. Materials_material

Bilanciamento Y

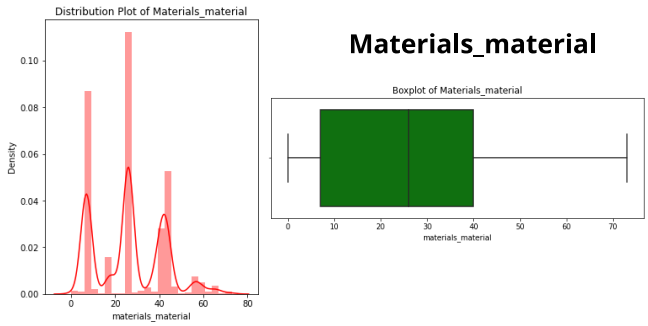


Figura 58

Bilanciamento Train set

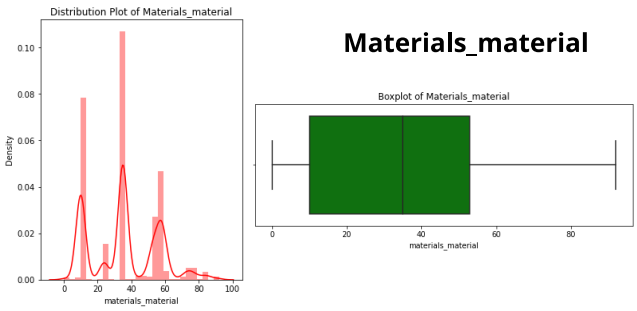


Figura 59

Nel grafico di distribuzione, possiamo notare che per il bilanciamento della Y i valori sono maggiormente concentrati nella regione compresa tra il valore 10, 25 e 45, mentre si osserva una minore presenza di dati nella regione di valore 18, 30 e da 50 in poi fino a 73. Nel grafico di distribuzione dove è stato effettuato il bilanciamento del train i valori sono maggiormente concentrati nella regione di valore10, 35 e 58, mentre si osserva una minore presenza di dati nella regione compresa tra 10 e 35, tra 37 e 57 e da 62 a 93.

I due grafici boxplot sono composti da una scatola rettangolare centrata sull'intervallo di valori che va da 0 a 73 per quanto riguarda il grafico riferito al bilanciamento della Y, mentre sull'intervallo di valori che va da 0 a 92 per il grafico riferito al bilanciamento del Train.

Il boxplot sulla sinistra presenta il baffo inferiore posizionato sullo 0, mentre il baffo superiore si posiziona sul valore 73. Il primo quartile si trova posizionato su 7, la mediana su 26, il terzo quartile su 40.

Mentre il boxplot sulla destra presenta il baffo inferiore posizionato sul valore 0, mentre il baffo superiore si posiziona su 92. Il primo quartile si trova posizionato su10, la mediana si trova posizionata sul valore 35 mentre il terzo quartile su53. Entrambi i grafici boxplot non presentano outlier.

3.5.2 Matrice di correlazione

In statistica, una matrice di correlazione è una tabella che mostra i coefficienti di correlazione tra le variabili. La matrice rappresenta la correlazione tra tutte le possibili coppie di valori presenti in una tabella ed è composta da righe e colonne che mostrano le variabili. Assume generalmente la forma quadrata, in cui il numero di righe e colonne coincide e può essere simmetrica, con le stesse variabili indicate nelle righe e nelle colonne (la scuola dei dati, yimp.it, 2023). Il coefficiente di correlazione è una misura che indica la forza e la direzione della relazione lineare tra due variabili. La matrice, può essere calcolata utilizzando il coefficiente di correlazione di Pearson per le variabili continue e il coefficiente di correlazione di Spearman o Kendall per le variabili ordinali o categoriche.

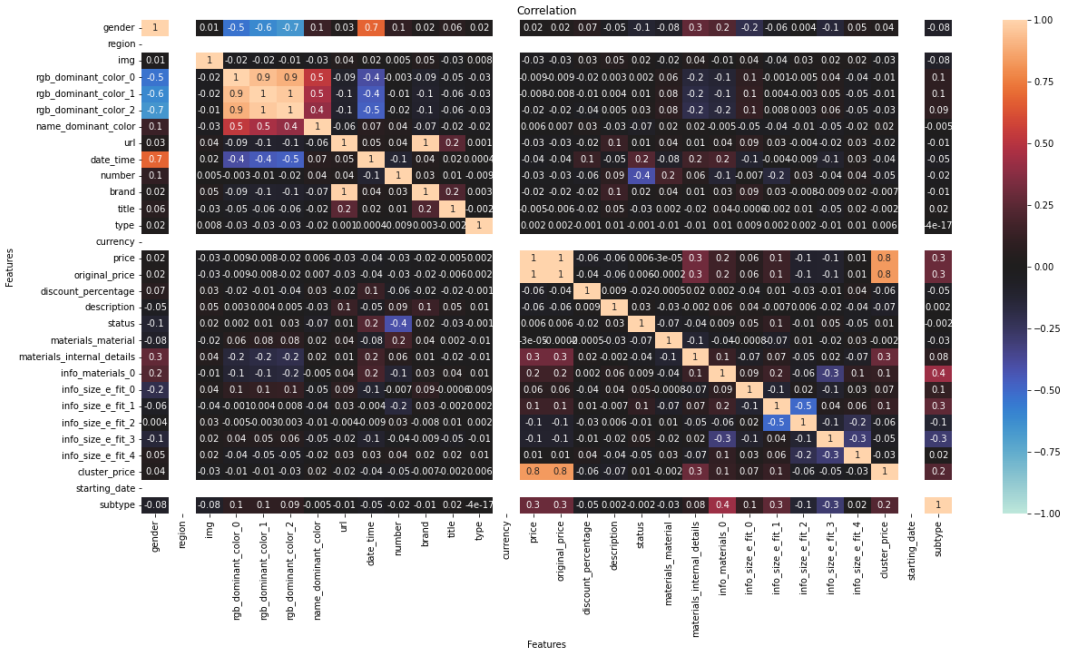
Viene utilizzata per studiare le relazioni tra le variabili di un dataset e selezionare quelle fortemente correlate per un'analisi successiva. Inoltre, può essere utile per individuare variabili ridondanti o multicollineari, ovvero variabili che sono altamente correlate tra loro e che possono quindi influenzarsi reciprocamente.
Per scegliere quali variabili utilizzare nella nostra analisi è stata fatta la matrice di correlazione per entrambi i bilanciamenti prima e dopo aver effettuato le procedure di Data cleaning per vedere anche se gli indici di correlazione sono migliorati o peggiorati dopo il processo di pulizia delle variabili.

3.5.2.1 Matrice di correlazione riferita al bilanciamento della Y

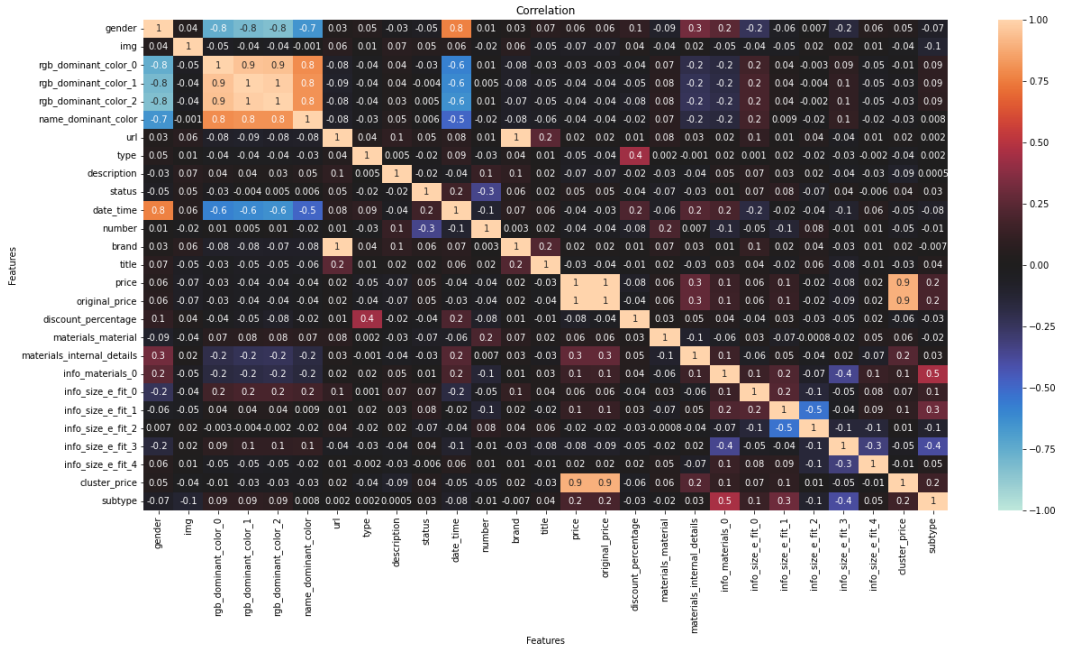
Dalla rappresentazione dei due grafici, si può notare che tutto il processo di pulizia del dataset ha aumentato la correlazione tra alcune variabili e questo è certamente positivo ai fini dell'analisi.
Più un indice di correlazione è vicino all'1, più la correlazione tra le due variabili sarà alta. Si noti, in primo luogo, che i valori sulla diagonale principale sono tutti pari a 1, poiché rappresentano la correlazione di ogni variabile con sé stessa. Possiamo notare graficamente che le variabili con l'indice di correlazione più alto sono Original Prize e prize, Brand e Url, rgb_dominant_color_2 e rgb_dominant_color_1 che risultano correlate positivamente con un indice di correlazione pari a 1, questo valore resta invariato anche dopo la pulizia del dataset sempre pari a 1. Altri indici di correlazione alti partendo da gender e andando verso destra nel descrivere le variabili, li vediamo tra le variabili rgb_dominant_color_0 e gender che risultano essere correlate negativamente di 0,5, che poi dopo la pulizia del dataset arriva ad un valore di ben -0,8; tra rgb_dominant_color_1 e gender anch'esse correlate negativamente di 0,6 che poi dopo la pulizia arriva ad un valore di -0,8; tra rgb_dominant_color_2 e gender correlate negativamente di 0,7 che poi dopo la pulizia resta pari allo stesso valore di -0,7; tra name_dominant_color e gender si può notare una forte correlazione negativa tra le variabili dopo la pulizia perché l'indice di correlazione passa da +0,1 a -0,7; e infine tra date_time e gender si può osservare una correlazione positiva di +0,7 che dopo la pulizia si alza ad un valore di ben +0,8. Vediamo una forte correlazione positiva tra le variabili rgb_dominant_color_1 e rgb_dominant_color_0 di ben 0,9 che però dopo il data cleaning resta pari allo stesso valore senza avere né miglioramenti, né peggioramenti; stessa cosa per rgb_dominant_color_2 e rgb_dominant_color_0 che sono correlate positivamente di valore 0,9 e resta invariato di valore anche dopo la pulizia. La variabile name_dominant_color è correlata positivamente con le variabili: rgb_dominant_color_0 (+0,5), rgb_dominant_color_1 (+0,5) e rgb_dominant_color_2 (+0,4) che dopo la pulizia notiamo che la loro correlazione si è fatta negativamente più intensa con rgb_dominant_color_0 (-0,6), rgb_dominant_color_1 (-0,6) e rgb_dominant_color_2 (-0,6). Notiamo che anche la correlazione tra le variabili date_time e name_dominant_color è migliorata dopo il data cleaning passando da una correlazione molto debole di + 0,07 ad una correlazione negativa di valore -0,5. Infine tra le variabili cluster_prize e price, cluster_prize e original_prize vi è una correlazione positiva di 0,8 che poi dopo la pulizia diventa di 0,9 e questo sicuramente indica ancora di più che c'è una forte correlazione positiva tra le variabili che riguardano il prezzo. Queste che sono state descritte sono le correlazioni tra le variabili che compongono la X.

Se si va ad osservare l'ultima riga, ovvero quella corrispondente alla variabile 'Subtype' (ovvero la variaibile Y) possiamo notare che non c'è un'alta correlazione tra le variabili della X e la variabile Y, infatti la più alta correlazione che si può constatare è tra la variabile subtype e info_materials_0, ovvero una correlazione positiva di 0,5 dopo che è stata effettuata la pulizia. Data la bassa correlazione tra le variabili della X e Subtype si è preferito considerarle tutte senza escluderne nessuna in quanto da diverse prove effettuate è emerso che i risultati migliori in termini di Accuracy, Recall, Precision e F1-score si raggiungono considerandole tutte senza escluderne nessuna.

Matrice di correlazione riferita al bilanciamento della Y prima del data cleaning



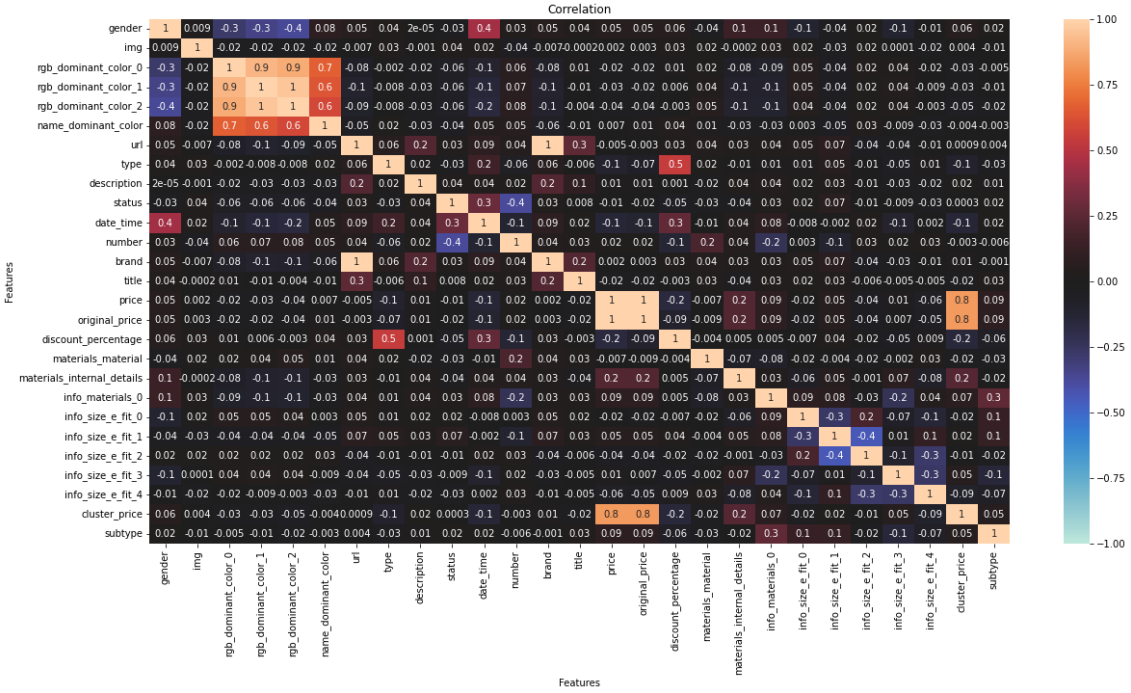
Matrice di correlazione riferita al bilanciamento della Y dopo il data cleaning



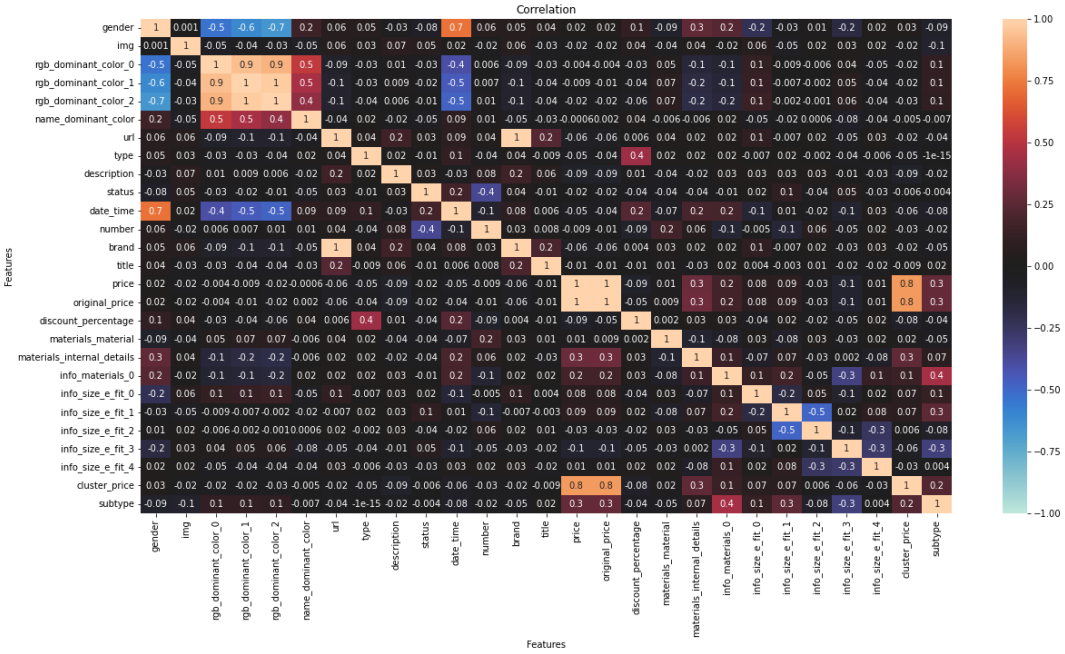
3.5.2.2. Matrice di correlazione riferita al bilanciamento del Train set

Dalla rappresentazione di entrambe le matrici di correlazione, constatiamo anche qui che c'è stato un miglioramento dei valori degli indici di correlazione tra le variabili.
Si può notare graficamente che le variabili con l'indice di correlazione più alto sono Original Prize e Prize, Brand e Url, rgb_dominant_color_2 e rgb_dominant_color_1. Tutte e tre le coppie di variabili di fatto sono correlate positivamente con un indice di correlazione pari a 1, questo valore resta invariato anche dopo la pulizia del dataset sempre pari a 1. Altre coppie di variabili che risultano essere altamente correlate sono (partendo da gender e andando verso destra nel descrivere le variabili) rgb_dominant_color_0 e gender di -0,3 che poi dopo la pulizia aumentano il loro indice di correlazione, infatti risultano correlate negativamente di 0,7; tra rgb_dominant_color_1 e gender correlate negativamente di 0,3 che poi dopo la pulizia risultano correlate negativamente di 0,6; tra rgb_dominant_color_2 correlate negativamente di 0,4 che poi dopo la pulizia resta a 0,7; e infine tra date_time e gender correlate positivamente di 0,4 prima della pulizia del dataset, per poi alzarsi a 0,7.
Vediamo una forte correlazione positiva tra rgb_dominant_color_1 e rgb_dominant_color_0 di 0,9 che però dopo il data cleaning resta pari allo stesso valore. Lo stesso valore di correlazione lo vediamo anche tra le variabili rgb_dominant_color_2 e rgb_dominant_color_0 che sono correlate positivamente di valore 0,9 che resta di valore costante anche dopo la pulizia. La variabile name_dominant_color è correlata positivamente con le variabili: rgb_dominant_color_0 (+0,7), rgb_dominant_color_1 (+0,6) e rgb_dominant_color_2 (+0,6) che però dopo la pulizia notiamo che il loro indice di correlazione si è abbassato con rgb_dominant_color_0 (+0,5), rgb_dominant_color_1 (+0,5) e rgb_dominant_color_2 (+0,4). Infine, tra le variabili cluster_prize e price, cluster_prize e original_prize vi è una correlazione positiva di 0,8 che poi dopo la pulizia resta invariata sempre di valore pari a 0,8 e questo sicuramente indica che c'è una forte correlazione positiva tra le variabili che riguardano il prezzo. Queste che sono appena state descritte sono le correlazioni tra le variabili che compongono la X.
Se si osserva l'ultima riga corrispondente alla variabile 'Subtype' (ovvero la variaibile Y) si può notare che anche in questo bilanciamento purtroppo non c'è un'alta correlazione tra le variabili della X e la variabile Y, infatti la più alta correlazione che si può constatare è tra la variabile Subtype e info_materials_0 di 0,4 dopo che è stata effettuata la pulizia. Data la bassa correlazione tra le variabili della X e Subtype si è preferito considerare tutte le features senza escluderne nessuna in quanto da diverse prove effettuate è emerso che i risultati migliori in termini di Accuracy, Recall, Precision e F1- score si raggiungono solamente considerandole tutte.

Matrice di correlazione riferita al bilanciamento del Train prima del data cleaning



Matrice di correlazione riferita al bilanciamento della Y dopo il data cleaning



3.6 Descrizione degli Algoritmi di classificazione utilizzati

Gli algoritmi di classificazione, o classificatori come sono anche conosciuti, rientrano nel ramo dell'apprendimento automatico. Come suggerisce il nome, questi modelli predittivi sono progettati per determinare la classe a cui appartiene un determinato soggetto. Poiché utilizzano l'apprendimento supervisionato, richiedono dati etichettati che includano una colonna contenente la loro classe.

Il processo di modellazione della classificazione di base comporta l'ottenimento di un set di dati, la creazione di funzionalità di variabili indipendenti e il loro utilizzo per prevedere una variabile dipendente o una classe di destinazione (Matt Clarke, 2021, Saturday, May 29).
Nelle analisi effettuate in questo lavoro di tesi sono stati utilizzati 6 diversi tipi di algoritmi di classificazione ovvero il Random Forest, XG-Boost, Knn, Decision Tree, Logistic Regression, Naive Bayes. In questa sezione finale del capitolo 3 verranno descritti brevemente.

- **Algoritmo Random Forest**

Il Random Forest è un algoritmo di Machine Learning ed è comunemente utilizzata per lo sviluppo di modelli predittivi. Questo metodo combina più Decision Tree generati in modo casuale e fornisce la modalità delle loro previsioni (nella classificazione) (Hastie et al., 2008). In altre parole, il Random Forest è un meta stimatore che adatta una serie di classificatori ad albero decisionale su vari sottocampioni del set di dati e utilizza la media per migliorare l'accuratezza predittiva e controllare l'over-fitting (Scikit learn, Pedregosa et al., 2011).
Rispetto ai singoli Decision Tree, il Random Forest presenta un overfit molto minore a causa della legge dei grandi numeri e una varianza significativamente inferiore. Nonostante si tratti di una tecnica che generalmente produce buoni risultati in termini di performance predittiva, il Random Forest tende comunque a sovra adattare i dati rispetto ad altri algoritmi ed è anche molto impegnativo dal punto di vista computazionale. Un altro punto negativo è il fatto che funziona in modo simile a una black-box, con l'output fornito senza alcuna interpretabilità (Breiman, 2001).

I principali iperparametri di un modello Random Forests che sono spesso soggetti a regolazione sono (Breiman, 2001):
- Numero di caratteristiche di divisione per nodo: questo parametro si riferisce al numero di variabili di input selezionate in modo casuale che vengono confrontate durante la decisione di divisione di un nodo. Quando il numero di variabili di divisione aumenta, aumenta anche la probabilità di trovare una divisione migliore. Tuttavia, aumentando il numero di variabili aumenta anche la correlazione tra gli alberi, amplificando così la varianza dei risultati (da Silva, D. V., 2018).
- Numero di alberi: il numero di alberi utilizzati nel modello non contribuisce all'overfitting e riduce la varianza del modello. Sebbene l'uso di un maggior numero di alberi migliori l'accuratezza del modello, ciò avviene a un ritmo decrescente e al costo di un tempo di elaborazione significativamente più lento (da Silva, D. V., 2018).
- Dimensione degli alberi: gli alberi di grandi dimensioni sono quelli che comprendono più suddivisioni. Un albero di grandi dimensioni è in grado di avere una maggiore capacità discriminatoria e di comprendere strutture di dati più complesse. Tuttavia, i modelli che li utilizzano sono anche più inclini all'overfitting dei dati (da Silva, D. V., 2018).

In questo lavoro di tesi per il Random Forest è stato utilizzato l'algoritmo RandomForestClassifier che è basato su un insieme di alberi decisionali impostando n_estimators = 200 come parametro che indica il numero di alberi decisionali da creare nell'ensemble del Random Forest. In questo caso specifico, vengono creati 200 alberi decisionali.
In altre parole, il RandomForestClassifier (n_estimators = 200) indica che si desidera creare un insieme di 200 alberi decisionali nel Random Forest. Il metodo fit viene utilizzato per addestrare il modello utilizzando i dati ditrain, mentre il metodo predict viene utilizzato per fare previsioni sulle nuove istanze di test.

- **Algoritmo XG-Boost**

L'algoritmo XGBoost o Extreme Gradient Boosting è un algoritmo di Machine Learning basato su albero decisionale che utilizza un processo chiamato boosting per migliorare le prestazioni (Data Science Team, 2020, 15 May). Dalla sua introduzione, è diventato uno degli algoritmi di apprendimento automatico più efficaci e produce regolarmente risultati che superano la maggior parte degli altri algoritmi, come la Logistic Regression, il modello di Random Forest e il Decision Tree (Matt Clarke, 2021, Saturday, May 29).
Nel lavoro di tesi si è partiti inizialmente con l'inserimento del cosiddetto "modello base" e dopo averne definito i parametri, è stato assegnato l'output a un oggetto chiamato model. Successivamente, è stata utilizzata la funzione fit () dell'oggetto model per addestrare il modello sui dati di Train. Al modello di Train viene assegnata una porzione dell'80% selezionata in modo casuale dell'intero set di dati che, con i dati X ed Y separati. Quando la funzione fit () viene eseguita, l'algoritmo XGBoost esaminerà i dati e cercherà le correlazioni tra le caratteristiche e la variabile target. Rieseguirà il processo di addestramento più e più volte fino a quando non sarà più accurato nel fare previsioni. Infine, viene utilizzato il modello addestrato sui dati di Train per fare previsioni sul set di dati di test o convalida utilizzando la funzione predict () (Matt Clarke, 2021, Saturday, May 29).

- **Algoritmo KNN**

L'algoritmo KNN può essere considerato un sistema di votazione, in cui l'etichetta della classe maggioritaria determina l'etichetta della classe di un nuovo punto dati tra i suoi "k" più vicini (dove k è un numero intero) nello spazio delle caratteristiche. Nell'algoritmo KNN, dove l'etichetta della classe di maggioranza determina l'etichetta della classe di un nuovo punto dati tra i suoi k vicini più prossimi (Adam Shafi, Feb 2023).

In questo lavoro di tesi è stato implementato l'algoritmo di classificazione KNeighborsClassifier (il codice utilizzato è stato più precisamernte KNeighborsClassifier(n_neighbors=3) che cerca di predire l'etichetta di classe di un'istanza di test basandosi sulle etichette delle istanze più vicine nel set di addestramento. Il parametro n_neighbors specifica il numero di vicini più prossimi che saranno presi in considerazione per la predizione che nel nostro caso sono stati pari a 3, quindi il modello considererà le etichette delle tre istanze più vicine per fare una previsione.

Per utilizzare il classificatore KNeighbors, si utilizza il metodo fit (più in dettaglio è stato utilizzato il codice knn.fit(balanced_trainX, balanced_trainY)), dove balanced_trainX rappresenta la matrice delle caratteristiche delle istanze di addestramento e balanced_trainY rappresenta i corrispondenti valori di classe. Questo metodo addestra il modello utilizzando i dati di addestramento forniti. Durante il processo di addestramento, il modello memorizza le istanze di addestramento nel suo spazio di feature e le etichette di classe associate ad esse.
Successivamente, dopo aver addestrato il modello, è possibile utilizzare il metodo predict (knn.predict(X_test)) per effettuare previsioni sui nuovi dati di test (testX). L'algoritmo KNeighbors troverà i n_neighbors vicini più prossimi a ciascuna istanza di test nel set di addestramento e utilizzerà le loro etichette di classe per fare una previsione per ciascuna istanza di test.

- **Algoritmo Decision Tree**

Denominato anche Albero decisionale è uno dei metodi più diffusi per la classificazione. Gli alberi decisionali sono ampiamente utilizzati poiché sono facili da interpretare, gestiscono le features categoriali, non richiedono il ridimensionamento delle features e sono in grado di lavorare con i dati non lineari (Decision tree classifier Spark MLlib, 2023).
Nel lavoro di ricerca è stato utilizzato come parametro il "max_depth" pari a 10 che indica la massima profondità dell'albero consentita durante il processo di addestramento. L'albero decisionale può crescere in modo fino a quando viene raggiunta la profondità massima specificata da questo parametro. Limitare la profondità massima può aiutare a evitare l'overfitting, ovvero un modello che si adatta eccessivamente ai dati di addestramento, perdendo la capacità di generalizzare su nuovi dati. Ridurre la profondità massima può migliorare la capacità di generalizzazione del modello a scapito della complessità.
Poi è stato utilizzato l'iperparametro "p_grid", ovvero p_grid = {"max_depth": [2, 3, 4, 5]} che indica che si desidera eseguire una ricerca dei migliori iperparametri per il parametro max_depth, che rappresenta la massima profondità consentita dell'albero decisionale.
In altre parole, ciò significa che la ricerca dei migliori iperparametri esplorerà diversi valori per max_depth, nello specifico 2, 3, 4 e 5. In sostanza, p_grid = {"max_depth": [2, 3, 4, 5]} indica che si desidera esaminare quale valore di max_depth (2, 3, 4 o 5) ottiene le migliori prestazioni per il modello di albero decisionale.
KFold invece è una tecnica di validazione incrociata (cross-validation) utilizzata per valutare le prestazioni di un modello di machine learning. Nella specifica notazione "KFold (n_splits=5)", KFold indica che il dataset verrà diviso in 5 fold o partizioni. Durante il processo di addestramento e validazione, il modello verrà allenato su 4 fold e testato sul fold rimanente. Questo processo viene ripetuto 5 volte, in modo che ogni fold sia utilizzato come set di test una volta. La tecnica di KFold aiuta a ottenere una stima più affidabile delle prestazioni del modello rispetto a una singola suddivisione dei dati in set di addestramento e test.
GridSearchCV (Cross-Validation Grid Search) è un metodo utilizzato per selezionare i migliori iperparametri di un modello di machine learning. È un approccio sistematico che esamina tutte le possibili combinazioni di

valori specificati in una griglia di ricerca (grid search) per i diversi iperparametri del modello. Durante il processo di ricerca della griglia, vengono eseguite iterazioni di addestramento e validazione utilizzando la tecnica di validazione incrociata (come KFold) per valutare le prestazioni di ogni combinazione di iperparametri.
In sintesi, il "max_depth" limita la profondità massima di un albero decisionale durante l'addestramento, "P_grid" rappresenta una griglia di valori di iperparametri da esplorare, KFold è una tecnica di validazione incrociata con 5 fold e il GridSearchCV è un metodo per selezionare i migliori iperparametri di un modello utilizzando la validazione incrociata

- **Algoritmo Logistic Regression**

La Logistic Regression è un algoritmo di apprendimento automatico comunemente utilizzato nei problemi di classificazione. Analogamente a una Regressione regolare, un modello logit fornisce coefficienti per ogni variabile regressore. La differenza principale tra i due modelli risiede nel modo in cui viene generato l'output, poiché una funzione logistica restituisce un output con valore reale all'interno dell'intervallo [0;1]. Nei problemi di classificazione, l'output di una regressione logistica può essere interpretato come la probabilità che una data osservazione appartenga alla classe da prevedere (Hosmer e Lemeshow, 2005).
I due principali vantaggi di questo algoritmo di Machine Learning sono che non è così impegnativo dal punto di vista computazionale come gli altri menzionati in precedenza, e inoltre il modello è interpretabile grazie al peso/tributo di ogni variabile di input che può essere compreso attraverso il suo rispettivo coefficiente, consentendo così una più chiara comprensione della relazione tra i predittori e le variabili di previsione (Park, 2013).
In questo lavoro di tesi, è stato adottato l'algoritmo LogisticRegression(random_state=1) che utilizza la funzione logistica per stimare la probabilità che una determinata istanza di dati appartenga a una classe specifica. Il parametro random_state=1, si assicura che l'algoritmo generi gli stessi risultati ogni volta che viene eseguito con gli stessi dati di addestramento. Questo è utile per ottenere risultati riproducibili e facilita il confronto tra modelli e la condivisione dei risultati.
Dopo aver importato i dati nel modello, viene addestrato su un sottoinsieme del set di dati completo. Il modello imparerà a identificare quale delle variabili o caratteristiche indipendenti è correlata con la variabile o la classe target e itererà sui dati, diventando progressivamente più accurato nel fare previsioni. Una volta addestrato, il modello di classificazione può essere valutato per valutarne l'accuratezza e utilizzato per fare previsioni su dati non etichettati.

- **Algoritmo Naive Bayes**

Gli algoritmi di classificazione Naive Bayes sono comunemente usati per problemi di classificazione del testo di Machine Learning, come prevedere il sentimento di un tweet, identificare la lingua di un pezzo di testo o classificare un ticket di supporto. Esistono in realtà diversi algoritmi di classificazione di Naive Bayes, che

utilizzano tutti una tecnica nota come Teorema di Bayes. Il concetto di base di questo, e la parte "ingenua" del nome, è che ogni coppia di caratteristiche è considerata indipendente l'una dall'altra ed è uguale (Matt Clarke, 2022, Sunday, May 08).

Ovviamente, questo presupposto è spesso sbagliato nelle situazioni del mondo reale, motivo per cui viene chiamato "ingenuo" Bayes. Tuttavia, ciò non impedisce che sia estremamente efficace.

In questo elaborato di tesi, per creare un modello di classificazione di base è stato utilizzato l'algoritmo Multinomial Naive Bayes tramite il modulo MultinomialNB in scikit-learn. Per poi utilizzare la funzione fit() per passare i dati balanced_trainX per il bilanciamento del train o trainX per il bilancviamento della Y e balanced_train_Y nel caso del bilanciamento della Y o trainY nel caso del bilanciamento della Y e addestrare il modello (Matt Clarke, 2022, Sunday, May 08).

Nel prossimo capitolo verranno commentati e mostrati i risultati di tutti gli algoritmi.

Capitolo 4. Risultati e discussioni

In questo capitolo, verranno presentati e commentati i risultati ottenuti per entrambi i bilanciamenti. Le metriche utilizzate per effettuare le analisi sono state l'Accuracy, la Precision, la Recall e l'F1- score. Per poi approfondire anche lo studio della matrice di confusione e l'AUC per la curva ROC e mostrare le features più rilevanti per ogni classificatore

4.1 Metriche utilizzate per confrontare i risultati

Nella valutazione delle prestazioni di un modello di classificazione multiclasse, l'Accuracy, la Recall, la Precision e l'F1-score sono metriche comunemente utilizzate per misurare diversi aspetti della qualità delle previsioni del modello. E' importante considerare tutte queste metriche insieme per valutare in modo completo le prestazioni di un modello di classificazione, poiché forniscono informazioni diverse sulla sua capacità di classificare correttamente le istanze di dati in diverse classi.

- Accuracy (Accuratezza):

L'Accuracy è una metrica che misura la percentuale di previsioni corrette rispetto al totale delle previsioni effettuate dal modello. È calcolata come il rapporto tra il numero di previsioni corrette e il numero totale di previsioni. L'Accuracy è una metrica generale che indica quanto il modello sia accurato nel classificare correttamente le istanze in tutte le classi. Più il valore di questa metrica si avvicina a 1.00, più il modello di classificazione è preciso e accurato.

$$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

Fonte: Hasty, May 17, 2023,https://hasty.ai/docs/mp-wiki/metrics/accuracy

- Precision (Precisione)

La Precision è una metrica di valutazione che misura la capacità di un modello di classificazione di identificare correttamente gli esempi positivi rispetto al numero totale di esempi classificati come positivi dal modello. È calcolata come il rapporto tra il numero di veri positivi e la somma dei veri positivi e falsi positivi. Il suo calore è compreso tra 0 e 1, dove un valore di Precision pari a 1 indica che tutti gli esempi classificati come positivi dal modello sono effettivamente positivi, mentre un valore di Precision pari a 0 indica che il modello non è in grado di identificare correttamente alcun esempio positivo. La Precision è una metrica utile quando è importante minimizzare i falsi positivi, ovvero quando si vuole essere sicuri che le istanze classificate come positive dal modello siano effettivamente corrette.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Fonte: Koo Ping Shung,(2018, Mar 15). Accuracy, Precision, Recall or F1?.Towards Data Science. https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

- Recall (Sensibilità):

La Recall misura la percentuale di istanze positive che sono state correttamente identificate dal modello rispetto al totale delle istanze positive presenti nel set di dati. È calcolata come il rapporto tra il numero di veri positivi e la somma dei veri positivi e falsi negativi. Restituisce un valore compreso tra 0 e 1, dove un valore pari a 1 indica che il modello è in grado di identificare tutti gli esempi positivi correttamente, mentre un valore di Recall pari a 0 indica che il modello non è in grado di identificare correttamente alcun esempio positivo. Questa metrica è utile quando è importante identificare correttamente tutte le istanze positive, minimizzando i falsi negativi.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Fonte: Koo Ping Shung, (2018, Mar 15). Accuracy, Precision, Recall or F1?.Towards Data Science. https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

- F1-score:

L'F1-score è una misura di bilanciamento tra Recall e Precision ed è calcolata come il rapporto tra il prodotto tra Precision e Recall e la loro somma. La metrica è un valore compreso tra 0 e 1, dove un valore più alto indica prestazioni migliori del modello. Un valore di F1- score pari a 1 indica una perfetta armonia tra Precisione e Recall, indicando che il modello è in grado di identificare correttamente tutti gli esempi positivi senza generare falsi positivi. L'F1-score è una metrica importante da utilizzare quando si desidera avere una valutazione complessiva delle prestazioni di un modello di classificazione, considerando sia la capacità di identificare correttamente gli esempi positivi (Recall) che di evitare falsi positivi (Precision).

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Fonte: Koo Ping Shung, (2018, Mar 15). Accuracy, Precision, Recall or F1?.Towards Data Science. https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

Matrice di confusione

La matrice di confusione è uno strumento utilizzato per valutare le prestazioni di un modello in un problema di classificazione multiclasse. Si compone di una tabella quadrata in cui le righe rappresentano le classi effettive e le colonne rappresentano le classi previste dal modello. Ogni cella contiene il numero di istanze che appartengono alla classe effettiva e sono state previste correttamente o erroneamente dal modello. Può aiutare a identificare le classi in cui il modello ha prestazioni migliori o peggiori e a individuare eventuali problemi di classificazione, come confusione tra classi simili o sbilanciamento delle previsioni tra le classi. Nella matrice di confusione è importante identificare i True Positives (TP), i True Negatives (TN), i False Positives (FP) e i False Negatives (FN):

- True Positives (TP) = Sono gli elementi presenti sulla diagonale principale della matrice, che rappresentano il numero di istanze correttamente classificate per ciascuna delle 9 classi, in altre parole dove il valore effettivo e il valore previsto sono gli stessi.
- True Negatives (TN)= Questi sono gli esempi che appartengono effettivamente alla classe negativa e sono stati correttamente riconosciuti come tali dal modello. Sono la somma di tutti gli elementi non presenti nella riga e nella colonna corrispondenti a una specifica classe.
- False Negatives (FN): Sono il numero di casi erroneamente classificati come negativi dal modello e sono i valori che appartengono effettivamente alla classe positiva, ma sono stati erroneamente classificati come negativi dal modello. Per individuarli essi sono gli elementi nella riga corrispondente a una specifica classe, esclusi gli elementi sulla diagonale principale.
- False Positives (FP): Rappresentano il numero di casi erroneamente classificati come positivi dal modello. Questi sono i valori che appartengono effettivamente alla classe negativa, ma sono stati erroneamente classificati come positivi dal modello. Essi sono gli elementi nella colonna corrispondente a una specifica classe, esclusi gli elementi sulla diagonale principale.

Curva ROC

Nella curva Roc viene calcolata un'altra metrica importante che misura le prestazioni complessive di un classificatore che è il valore "Area Under the Curve", o anche meglio conosciuto come AUC (Rohit Kundu, 2022, September 13). L'AUC rappresenta l'area sottesa alla curva ROC e fornisce una misura della capacità discriminante del modello per la classe specifica. I valori che può assumere variano tra 0 e 1, dove il valore 1 indica una discriminazione perfetta e un valore di 0.5 indica una discriminazione casuale. La curva ROC può avere varie forme, che riflettono le prestazioni del modello nel distinguere le classi. L'AUC per ciascuna classe può variare da 0.5 (discriminazione casuale) a 1 (discriminazione perfetta), con valori intermedi che indicano diversi gradi di capacità discriminante del modello per ciascuna classe rispetto alle altre classi. Naturalmente, un valore più alto di AUC rappresenta un classificatore migliore e dotato di una buona capacità di discriminare la classe in esame rispetto alle altre classi. Il grafico presenta sull'asse delle X i False Positive Rate, ovvero la proporzione di esempi negativi erroneamente classificati come positivi rispetto al numero totale di esempi

negativi. Mentre sull'asse delle Y abbiamo i True Positive Rate, ovvero la proporzione di esempi positivi correttamente classificati rispetto al numero totale di esempi positivi.

4.2 Risultati dei Classificatori

Per ogni algoritmo, si partirà col commentare i risultati ottenuti dal Bilanciamento "teorico", ovvero il Bilanciamento della Y, per poi confrontarli con i risultati ottenuti dal Bilanciamento del Train basato su dati reali, ovvero i risultati corretti da considerare per questo lavoro di tesi.

- **Random Forest**

I risultati ottenuti per il classificatore Random Forest sono riassunti nella seguente tabella che fornisce una panoramica dettagliata delle prestazioni del modello (in termini di Accuracy, Precision, Recall e F1-score) di classificazione multiclasse per ognuna delle 9 classi, nonché una valutazione complessiva delle prestazioni del modello attraverso le medie. Le medie macro e ponderate (weighted) rappresentano le medie delle metriche per tutte le classi, assegnando loro un peso uguale o in base al support. Si intende specificare che i risultati corretti sono quelli indicati dalle medie ponderate (ultima riga della tabella) in quanto indicano un buon livello di performance complessivo del modello, considerando l'equilibrio delle classi nel dataset di test.
Nella parte bassa della tabella, vengono forniti l'Accuracy totale e le medie non pesate e pesate delle metriche descritte in precedenza.

Bilanciamento Y

	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	201
1	1.00	1.00	1.00	166
2	0.93	0.99	0.96	115
3	0.99	1.00	1.00	175
4	0.87	0.95	0.91	105
5	0.95	0.84	0.89	121
6	0.96	0.92	0.94	110
7	1.00	1.00	1.00	173
8	1.00	1.00	1.00	153
accuracy			0.97	1319
macro avg	0.97	0.97	0.97	1319
weighted avg	0.98	0.97	0.97	1319

Bilanciamento Train

	Precision	Recall	F1-score	Support
0	0.83	0.77	0.80	13
1	1.00	0.15	0.27	13
2	0.82	0.56	0.67	59
3	0.76	0.86	0.80	43
4	0.82	0.92	0.87	486
5	0.73	0.46	0.57	97
6	0.81	0.82	0.81	170
7	0.70	0.73	0.71	41
8	0.50	0.30	0.37	10
accuracy			0.80	932
macro avg	0.77	0.62	0.65	932
weighted avg	0.80	0.80	0.79	932

I risultati ottenuti in termini di Accuracy, Precision, Recall e F1-score per entrambi i bilanciamenti indicano un buon rendimento del modello Random Forest in questo problema di classificazione multiclasse in quanto sono tutti valori superiori all'80%.

Andando più nel dettaglio possiamo notare un'Accuracy molto buona per entrambi i bilanciamenti: per il bilanciamento della Y molto alta pari al 97%, il che significa che il 97% delle istanze nel dataset è stata classificata correttamente dal modello e che quindi è in grado di classificare la maggior parte delle istanze correttamente. Mentre per il bilanciamento del Train, l'Accuracy è pari all'80% (un risultato anch'esso buono), che significa che l'80% delle istanze nel dataset è stata classificata correttamente dal modello.La Precision, che indica quanto il modello è preciso nel classificare le istanze positive, possiamo osservare che varia tra 87% e 100% per tutte le classi nel caso del bilanciamento della Y (indicando un'alta precisione per tutte le classi), mentre varia dal 50% al 100% nel caso del bilanciamento del Train.

Per il bilanciamento della Y, il valore di Precision pesato è 98% e indica che il 98% delle istanze classificate come positive dal modello sono effettivamente positive. Mentre per quanto riguarda il bilanciamento del Train il valore della Precision pesata è dell'80%. Possiamo dire che entrambi i valori sono molto buoni, difatti un valore elevato di Precision è particolarmente importante quando si vogliono evitare i falsi positivi.

La Recall, anche chiamata sensitivity o true positive rate, misura la capacità del modello di individuare correttamente le istanze positive ed è preferibile averla alta perché più alta è e meno falsi negativi ci sono.

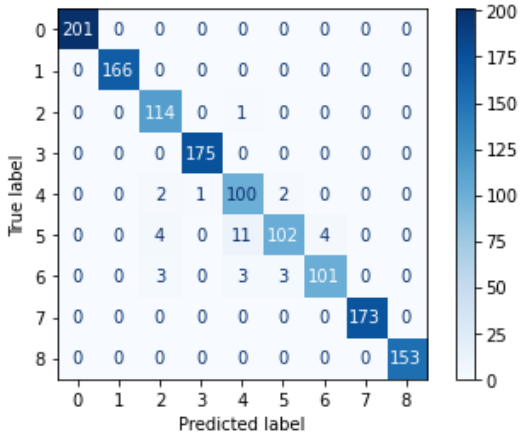
Nel caso del bilanciamento della Y, possiamo osservare che varia tra l'84% e il 100% per le classi, mentre nel caso del bilanciamento del Train il suo valore varia dal 15% al 92%. Il valore di Recall pesato è del 97% (per la figura di sinistra), quindi possiamo affermare che il 97% delle istanze positive nel dataset è stato correttamente identificato dal modello. Mentre otteniamo un valore di Recall dell'80% per quanto riguarda il bilanciamento del Train, il che significa che l'80% delle istanze positive nel dataset è stato correttamente identificato dal modello. L'F1-score è utile quando si desidera avere una valutazione complessiva delle

prestazioni del modello, tenendo conto sia delle metriche Precision e Recall e difatti questa metrica è una combinazione tra le due.

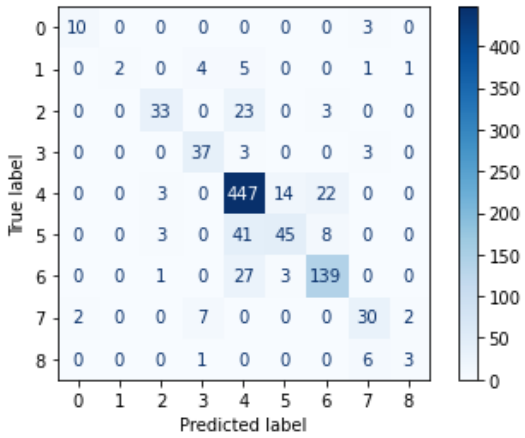
Per quanto concerne il bilanciamento della Y, possiamo constatare che l'F1-score varia tra 89% e 100% per le diverse classi, mentre per quanto riguarda il bilanciamento del Train il suo valore varia tra il 27% e l'87%. Il valore dell'F1 score pesato è del 97% nel caso del bilanciamento della Y e indica un equilibrio tra Precision e Recall suggerendo che il modello è in grado di ottenere buoni risultati su entrambe le metriche e di performance. Un buon valore di F1-score lo possiamo riportare anche per il bilanciamento del Train pari al 79%. Osserviamo che i risultati ottenuti dai due bilanciamenti sono diversi, difatti ci si aspettava che i risultati teorici fossero un po' più alti e vicino all'1 rispetto a quelli reali. Possiamo concludere che però siamo soddisfatti in quanto sono dei valori molto alti.

Matrice di confusione

Bilanciamento Y



Bilanciamento Train



Analizzando le matrici per entrambi i bilanciamenti, possiamo trarre le seguenti osservazioni:

Per la classe 0, per quanto riguarda il bilanciamento della Y sono presenti 201 istanze e tutte sono classificate correttamente come Classe 0. Non ci sono falsi positivi o falsi negativi per questa classe.

Per il bilanciamento del Train invece sono stati correttamente classificati 10 casi. Tuttavia, sono presenti 3 falsi negativi, indicando che alcuni casi che appartengono alla classe 0 sono stati erroneamente classificati come classe 7 e 2 falsi positivi.

Per la classe 1, per quanto riguarda il bilanciamento della Y, sono presenti 166 istanze e tutte sono classificate correttamente come Classe 1. Non ci sono falsi positivi o falsi negativi per questa classe.

Mentre per quanto riguarda il bilanciamento del Train sono stati correttamente classificati come classe 1 solo 2 casi, mentre ci sono stati 11 falsi negativi e 0 falsi positivi. In particolare, sono stati confusi con la classe 3, la classe 4, la classe 7 e la classe 8.

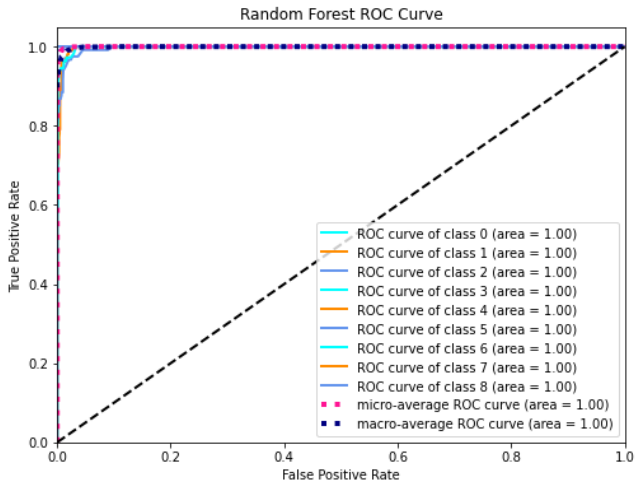
Per quanto concerne la classe 2 per il bilanciamento della Y sono presenti 115 istanze, di cui 114 sono classificate correttamente come Classe 2 e 1 viene erroneamente classificata come Classe 4. Quindi, abbiamo 1 falso negativo per la Classe 2 e 9 falsi positivi.
Mentre per quanto riguarda il bilanciamento del Train sono state correttamente predette 33 istanze come classe 2, mentre sono state erroneamente predette 23 istanze come classe 4, 3 istanze come classe 6. Ci sono inoltre 7 falsi positivi.
Per la classe 3 sono presenti 175 istanze e tutte sono classificate correttamente come Classe 3. Non ci sono falsi negativi ma 1 solo falso positivo per questa classe. Mentre per il bilanciamento del train sono state correttamente classificate come classe 3 solo 37 casi, e sono presenti 12 falsi positivi con altre classi. Tuttavia, sono presenti 6 falsi negativi, indicando che alcuni casi che appartengono alla classe 3 e alla classe 7 sono stati erroneamente classificati come altre classi.
Per quanto riguarda il bilanciamento della Y nella classe 4 sono presenti 105 istanze, di cui 100 sono classificate correttamente come Classe 4, 2 vengono erroneamente classificate come Classe 5, 2 vengono erroneamente classificate come Classe 2 e 1 viene erroneamente classificato come classe 3. Quindi, abbiamo 15 falsi positivi per la Classe 4 e 5 falsi negativi. Mentre nel bilanciamento del Train la classe 4 mostra una buona capacità di distinzione, con 447 casi correttamente classificati come classe 4. Tuttavia, sono presenti 99 falsi positivi e 39 falsi negativi.
Per la classe 5, Sono presenti 121 istanze, di cui 102 sono classificate correttamente come Classe 5, 11 vengono erroneamente classificate come Classe 4, 4 vengono erroneamente classificate come Classe 2 e 4 come classe 6. Quindi, abbiamo 5 falsi positivi e 19 falsi negativi. Mentre per quanto riguarda il bilanciamento del Train, sono stati correttamente classificati come classe 5 solo 45 casi. Tuttavia, sono presenti 17 falsi positivi e 52 falsi negativi per la classe 6.
Per la classe 6 nella matrice di confusione riferita al bilanciamento del Train sono presenti 110 istanze, di cui 101 sono classificate correttamente come Classe 6, 3 vengono erroneamente classificate come Classe 2, 3 vengono erroneamente classificate come Classe 4 e 3 vengono erroneamente classificati come classe 5. Quindi, abbiamo 9 falsi negativi e 4 falsi positivi. Mentre per quanto riguarda il bilanciamento del Train la classe 6 sono stati correttamente classificati come classe 6 ben 139 casi, ma ci sono 33 falsi positivi e 31 falsi negativi.
Mentre per la classe 7 nel grafico a sinistra sono presenti 173 istanze e tutte sono classificate correttamente come Classe 7. Non ci sono falsi positivi o falsi negativi per questa classe. Nell'altro bilanciamento sono stati correttamente classificati 30 casi. Tuttavia, sono presenti 11 falsi negativi e 13 falsi positivi.
Infine per l'ultima classe, ovvero la classe 8 sono presenti 153 istanze e tutte sono classificate correttamente. Non ci sono falsi positivi o falsi negativi per questa classe. Mentre per il bilanciamento del Train sono stati correttamente classificati come classe 8 solo 3 casi. Tuttavia, sono presenti 7 falsi negativi e 3 falsi positivi

In conclusione, la matrice di confusione del bilanciamento della Y mostra buone prestazioni del modello di classificazione, con la maggior parte delle istanze correttamente classificate per ciascuna classe. Purtroppo,

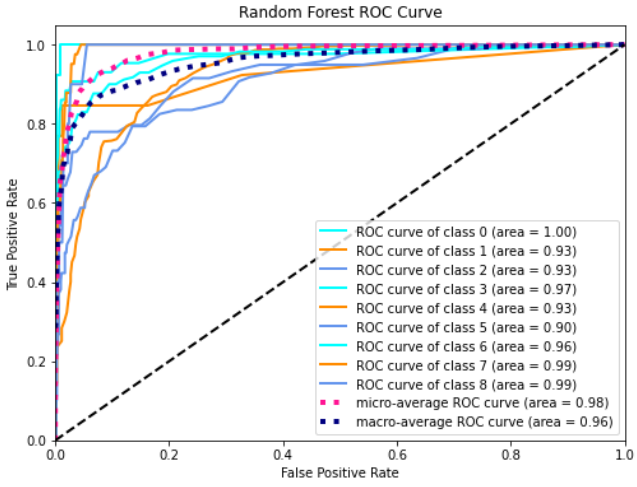
però non si può dire lo stesso per il bilanciamento del Train inquanto questa matrice di confusione indica una certa difficoltà nel distinguere alcune classi. Certamente alcune classi mostrano una buona capacità di distinzione, mentre altre mostrano una confusione significativa con diverse classi. Sarebbe utile analizzare ulteriormente le cause di questa confusione e cercare di migliorare le prestazioni del modello

Curva ROC

Bilanciamento Y



Bilanciamento Train

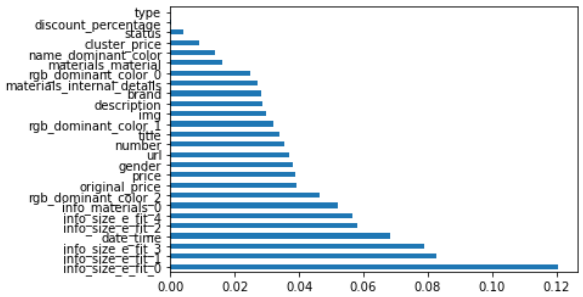


Dopo aver commentato i risultati della matrice di confusione passiamo al confronto dei risultati dei due bilanciamenti per la curva ROC andando ad analizzare i valori AUC per ciascuna classe. Possiamo constatare, già da un primo sguardo, che i risultati sono molto buoni.

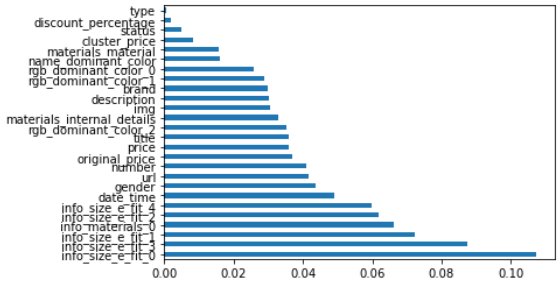
Per quanto riguarda la curva Roc del bilanciamento della Y possiamo notare un caso di discriminazione perfetta in quanto tutti i valori AUC per ciascuna classe sono pari a 1 e quindi il modello è in grado di distinguere completamente tutte le classi senza alcun errore di classificazione. In questo caso, la curva ROC per ciascuna classe sarà una linea verticale che raggiunge il punto (0,1) (TPR = 1, FPR = 0) e rimane a quel valore per tutti i livelli di soglia di decisione. L'AUC sarà pari a 1 per ogni classe, indicando una perfetta separazione tra le classi. Anche la curva Roc del bilanciamento del Train dà dei risultati molto buoni, anche se non perfetti come per la figura sulla sinistra, ma ci possiamo ritenere soddisfatti in quanto l'AUC per ogni classe varia da 0.93 a 1.00 quindi possiamo affermare che anche in questo caso il modello è in grado di distinguere molto bene tutte le classi senza alcun errore di classificazione.

Features più rilevanti

Bilanciamento Y



Bilanciamento Train



Come ultima analisi, si riportano i grafici per entrambi i bilanciamenti delle features più rilevanti per l'algoritmo Random Forest che, misurate tramite valore numerico (in ordine crescente), permettono di vedere quelle che hanno contribuito di più alla previsione del modello. Attraverso l'analisi accurata delle caratteristiche, è possibile identificare gli attributi più importanti e che quindi ci danno più informazioni, ci permettono di migliorare le prestazioni del modello e di ottenere una migliore comprensione del problema in esame.

Per entrambi i grafici la feature più rilevante è stata info_size_e_fit_0 con un valore pari a 0,12, per poi a seguire le altre features che appartengono alla stessa categoria della prima, ovvero info_size_e_fit_3 e info_size_e_fit_1 (riportando gli attributi più importanti del Bilanciamento del Train). Tutte e tre queste features danno informazioni precise sulle taglie e le misure relative alle borse e quindi questo sta ad indicare che l'algoritmo ha riconosciuto le informazioni riguardanti taglie e misure come importanti per discriminare le diverse tipologie di borse e le utilizza per prendere decisioni di classificazione più accurate.

- ## Classificatore XG-Boost

I risultati ottenuti in termini di Accuracy, Precision, Recall e F1 score per entrambi i bilanciamenti indicano un buon rendimento del modello XG-Boost in questo problema di classificazione multiclasse in quanto sono tutti valori superiori all'80%. Possiamo notare che i risultati sono molto simili al classificatore Random Forest. Andando più nel dettaglio possiamo notare un'Accuracy molto buona per entrambi i bilanciamenti: per il bilanciamento della Y molto alta pari al 98%, il che significa che il 98% delle istanze nel dataset è stata classificata correttamente dal modello e che quindi è in grado di classificare la maggior parte delle istanze correttamente. Mentre per il bilanciamento del Train l'Accuracy è pari all'80%, un risultato anch'esso buono che significa che l'80% delle istanze nel dataset è stata classificata correttamente dal modello. La Precision, che indica quanto il modello è preciso nel classificare le istanze positive, possiamo osservare che per tutte le classi varia tra 93% e 100% nel caso del bilanciamento della Y (indicando un'alta precisione per tutte le classi), mentre varia dal 33% al 88% nel caso del bilanciamento del Train.

Per il bilanciamento della Y, il valore di Precision pesato è 98% e indica che il 98% delle istanze classificate come positive dal modello sono effettivamente positive. Mentre per quanto riguarda il bilanciamento del Train

il valore della Precision pesata è dell'80%. Possiamo dire che entrambi i valori sono molto buoni, difatti un valore elevato di Precision è particolarmente importante quando si vogliono evitare i falsi positivi.
La Recall, che misura la capacità del modello di individuare correttamente le istanze positive, è preferibile averla alta perché più alta è e meno falsi negativi ci saranno. Nel caso del bilanciamento della Y, possiamo osservare che varia tra l'88% e il 100% per le classi, mentre nel caso del bilanciamento del Train il suo valore varia dal 15% all'88%. Il valore di Recall pesato è del 98% (per la figura di sinistra), quindi possiamo affermare che il 98% delle istanze positive nel dataset è stato correttamente identificato dal modello. Mentre otteniamo un valore di Recall dell'80% per quanto riguarda il bilanciamento del Train, il che significa che l'80% delle istanze positive nel dataset è stato correttamente identificato dal modello. Anche per questa metrica i risultati di entrambi i bilanciamenti sono molto buoni. L'F1-score è utile quando si desidera avere una valutazione complessiva delle prestazioni del modello, tenendo conto sia delle metriche Precision e Recall e difatti questa metrica è una combinazione tra le due.
Per quanto concerne il bilanciamento della Y, possiamo constatare che l'F1-score varia tra 92% e 100% per le diverse classi, mentre per quanto riguarda il bilanciamento del Train il suo valore varia tra lo 25% e l'87%. Il valore dell'F1-score pesato è del 98% nel caso del bilanciamento della Y e indica un equilibrio tra Precision e Recall suggerendo che il modello è in grado di ottenere buoni risultati su entrambe le metriche e di performance. Un buon valore di F1-score lo possiamo riportare anche per il bilanciamento del Train pari al 80%. Osserviamo che i risultati ottenuti dai due bilanciamenti sono diversi, difatti ci si aspettava che i risultati teorici fossero un po' più alti e vicino all'1 rispetto a quelli reali. Possiamo concludere che però siamo ugualmente soddisfatti in quanto sono dei valori molto alti, superiori all'80

Bilanciamento Y

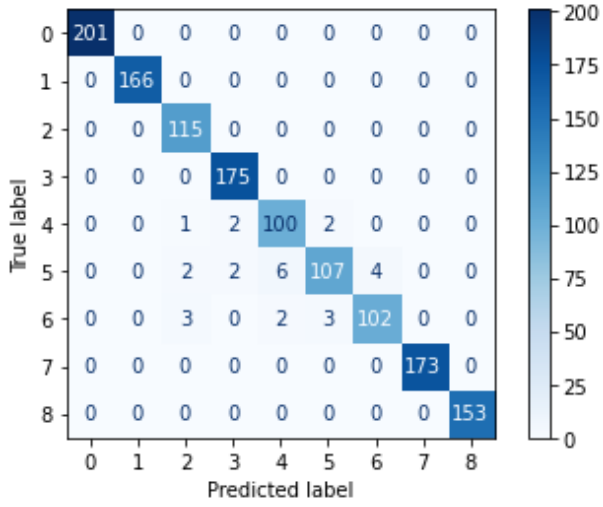
	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	201
1	1.00	1.00	1.00	166
2	0.95	1.00	0.97	115
3	0.98	1.00	0.99	175
4	0.93	0.95	0.94	105
5	0.96	0.88	0.92	121
6	0.96	0.93	0.94	110
7	1.00	1.00	1.00	173
8	1.00	1.00	1.00	153
accuracy			0.98	1319
macro avg	0.97	0.97	0.97	1319
weighted avg	0.98	0.98	0.98	131

Bilanciamento Train

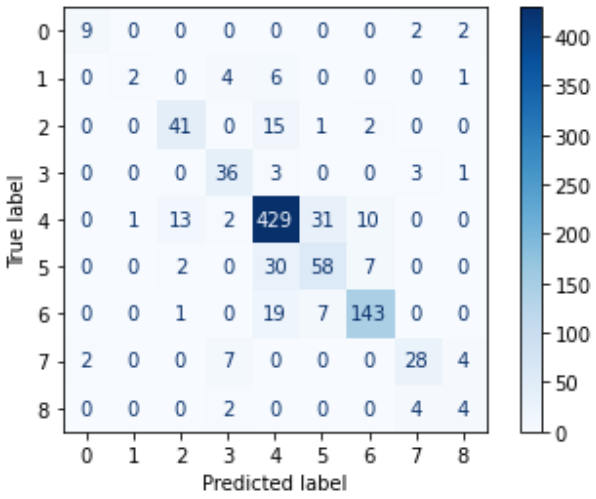
	Precision	Recall	F1-score	Support
0	0.82	0.69	0.75	13
1	0.67	0.15	0.25	13
2	0.72	0.69	0.71	59
3	0.71	0.84	0.77	43
4	0.85	0.88	0.87	486
5	0.60	0.60	0.60	97
6	0.88	0.84	0.86	170
7	0.76	0.68	0.72	41
8	0.33	0.40	0.36	10
accuracy			0.80	932
macro avg	0.70	0.64	0.65	932
weighted avg	0.80	0.80	0.80	932

Matrice di Confusione

Bilanciamento Y



Bilanciamento Train



Analizzando le matrici per entrambi i bilanciamenti, possiamo trarre le seguenti osservazioni:

Per la classe 0, per quanto riguarda il bilanciamento della Y sono presenti 201 istanze e tutte sono classificate correttamente come Classe 0. Non ci sono falsi positivi o falsi negativi per questa classe.

Per il bilanciamento del Train invece sono stati correttamente classificati 9 casi. Tuttavia, sono presenti 4 falsi negativi, indicando che alcuni casi che appartengono alla classe 0 sono stati erroneamente classificati come classe 7 e 8, mentre ci sono 2 falsi positivi.

Per la classe 1, per quanto riguarda il bilanciamento della Y sono presenti 166 istanze e tutte sono classificate correttamente come Classe 1. Non ci sono falsi positivi o falsi negativi per questa classe.

Mentre per quanto riguarda il bilanciamento del Train sono stati correttamente classificati come classe 1 solo 2 casi, mentre ci sono stati 11 falsi negativi e 1 falso positivo. In particolare, sono stati confusi con la classe 3, la classe 4 e la classe 8.

Per quanto concerne la classe 2 per il bilanciamento della Y sono presenti 115 istanze. Non ci sono falsi negativi ma ci sono 6 falsi negativi per questa classe.

Mentre per quanto riguarda il bilanciamento del Train sono state correttamente predette 41 istanze come classe 2, mentre sono state erroneamente predette 15 istanze come classe 4, 1 istanza come classe 5 e 2 istanze come classe 6 (falsi negativi). Per poi aggiungere che ci sono 16 falsi positivi.

Per la classe 3 sono presenti 175 istanze e tutte sono classificate correttamente come Classe 3. Non ci sono falsi negativi per questa classe ma ci sono 4 falsi positivi. Mentre per il bilanciamento del Train sono state correttamente classificate come classe 3 solo 36 casi, con 15 falsi positivi. Sono presenti 7 falsi negativi, indicando che alcuni casi che appartengono alla classe 4, 7 e 8 sono stati erroneamente classificati come altre classi.

Per quanto riguarda il bilanciamento della Y nella classe 4 sono presenti 105 istanze, di cui 100 sono classificate correttamente come Classe 4, 2 vengono erroneamente classificate come Classe 5, 2 vengono erroneamente classificate come Classe 2 e 1 viene erroneamente classificato come classe 3. Quindi, abbiamo 8 falsi positivi per la Classe 4 e 5 falsi negativi. Mentre nel bilanciamento del Train la classe 4 mostra una buona capacità di distinzione, con 429 casi correttamente classificati come classe 4. Tuttavia, sono presenti 73 falsi positivi e 57 falsi negativi.

Per la classe 5, Sono presenti 121 istanze, di cui 107 sono classificate correttamente come Classe 5, 6 vengono erroneamente classificate come Classe 4, 2 vengono erroneamente classificate come Classe 2, 2 vengono erroneamente classificati come classe 3 e 4 come classe 6. Quindi, abbiamo 14 falsi negativi e 8 falsi positivi. Mentre per quanto riguarda il bilanciamento del Train, sono stati correttamente classificati come classe 5 solo 58 casi. Tuttavia, sono presenti 39 falsi positivi e 39 falsi negativi.

Per la classe 6 nella matrice di confusione riferita al bilanciamento della Y sono presenti 110 istanze, di cui 102 sono classificate correttamente come Classe 6, 3 vengono erroneamente classificate come Classe 2, 2 vengono erroneamente classificate come Classe 4 e 3 vengono erroneamente classificati come classe 5. Quindi, abbiamo 8 falsi negativi e 4 falsi positivi. Mentre per quanto riguarda il bilanciamento del Train nella classe 6 sono stati correttamente classificati come classe 6 ben 143 casi, ma ci sono 19 falsi positivi e 27 falsi negativi.

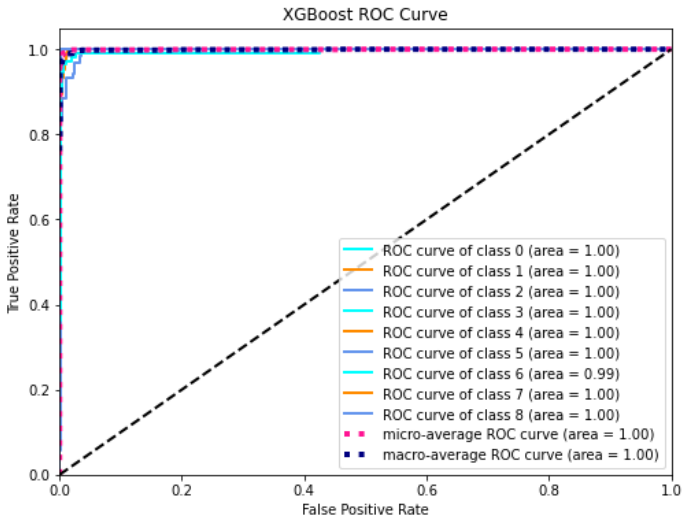
Mentre per la classe 7 nel grafico a sinistra sono presenti 173 istanze e tutte sono classificate correttamente come Classe 7. Non ci sono falsi positivi o falsi negativi per questa classe. Nell'altro bilanciamento sono stati correttamente classificati 28 casi. Tuttavia, sono presenti 13 falsi negativi e 9 falsi positivi.

Infine per l'ultima classe, ovvero la classe 8, sono presenti 153 istanze e tutte sono classificate correttamente per il bilanciamento della Y. Non ci sono falsi positivi o falsi negativi per questa classe. Mentre per il bilanciamento del Train sono stati correttamente classificati come classe 8 solo 4 casi. Tuttavia, sono presenti 6 falsi negativi e 8 falsi positivi.

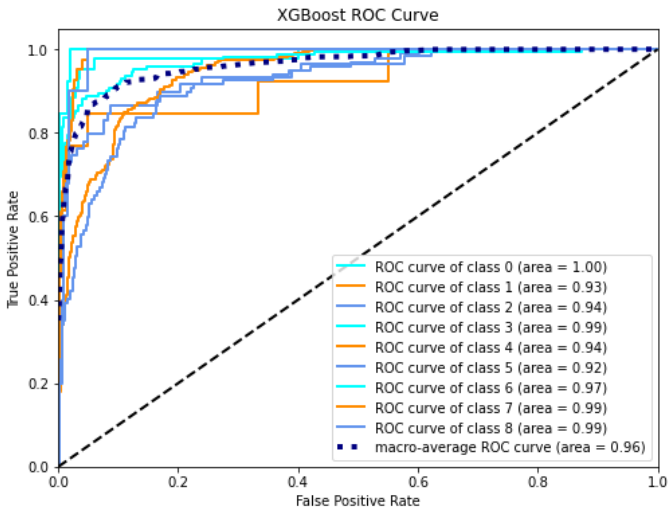
In generale, la matrice di confusione del bilanciamento della Y mostra migliori prestazioni del modello di classificazione rispetto alla matrice del bilancimaneto del Train, in quanto presenta un maggior numero di istanze correttamente classificate per ciascuna classe. Alcune classi mostrano una buona capacità di distinzione, mentre altre mostrano una confusione significativa con diverse classi. Sarebbe utile analizzare ulteriormente le cause di questa confusione e cercare di migliorare le prestazioni del modello.

Curva Roc

Bilanciamento Y



Bilanciamento Train

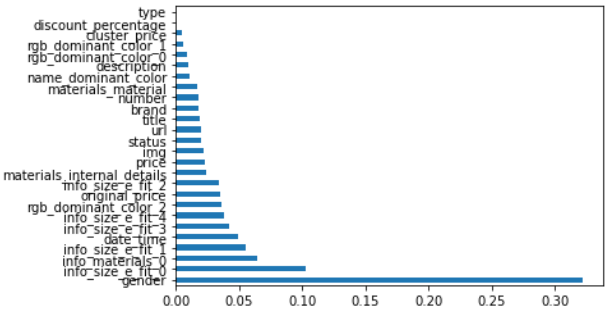


Dopo aver commentato i risultati della matrice di confusione, passiamo al confronto dei risultati dei due bilanciamenti per la curva ROC andando ad analizzare i valori AUC per ciascuna classe. Come per il precedente classificatore, ovvero il Random Forest, anche qui i risultati sono molto buoni.

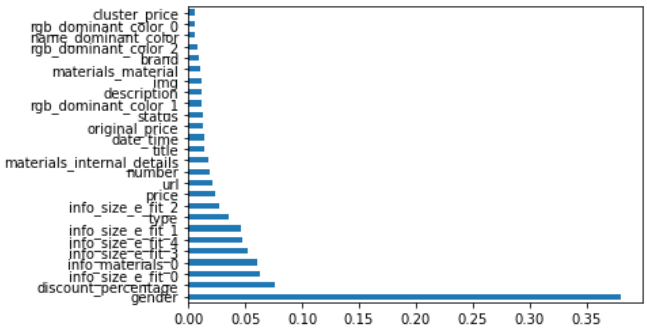
Per quanto riguarda la curva Roc del bilanciamento della Y possiamo notare un caso di discriminazione praticamente perfetta in quanto tutti i valori AUC per ciascuna classe (tranne che per la classe 6) sono pari a 1 e quindi possiamo affermare che il modello XG-Boost è in grado di distinguere completamente tutte le classi senza alcun errore di classificazione. Anche la curva Roc del bilancimaneto del Train dà dei risultati molto soddisfacenti, anche se non perfetti come per la figura di sinistra, ma ci possiamo ritenere soddisfatti in quanto l'AUC per ogni classe varia da 0.92 a 1.00 quindi possiamo affermare che anche in questo caso il modello è in grado di distinguere molto bene tutte le classi con pochi errori di classificazione.

Features più rilevanti

Bilanciamento Y



Bilanciamento Train



Per l'algoritmo XG-Boost, si può notare che, a differenza dell'algoritmo Random Forest, la feature più rilevante non è info_size_e_fit_0, ma bensì gender. Per il bilanciamento del Train, gender presenta un valore un po' più alto pari quasi a 0,40 rispetto al bilanciamento della Y che presenta un valore pari a 0,35. L'algoritmo XG-Boost ha identificato il genere come una caratteristica che fornisce informazioni discriminanti e influenti per la classificazione delle borse. Pertanto, il genere può essere utilizzato come la feature più rilevante per distinguere le diverse tipologie di borse all'interno di questo problema di classificazione multiclasse.
La feature info_size_e_fit_0, si trova sempre tra le prime con un valore simile a quello del precedente algoritmo ovvero 0,7 per il Bilanciamento del Train e 0,10 per il Bilanciamento della Y sottolineando che anche per questo algoritmo le informazioni riguardanti taglie e misure sono rilevanti per categorizzare le diverse tipologie di borse.

- **Classificatore KNN**

I risultati ottenuti in termini di Accuracy, Precision, Recall e F1 score indicano un buon rendimento del modello KNN in questo problema di classificazione multiclasse solo per il bilanciamento della Y in quanto sono tutti valori superiori all'80%, purtroppo però i risultati ottenuti dal bilanciamento del Train sono molto bassi e questo sicuramente ci sta ad indicare che questo modello non performa bene nel caso di dati reali.
Andando più nel dettaglio possiamo notare un buon valore di Accuracy per il bilanciamento della Y pari al 87%, il che significa che l'87% delle istanze nel dataset è stata classificata correttamente dal modello e che quindi è in grado di classificare la maggior parte delle istanze correttamente. Mentre per il bilanciamento del Train l'Accuracy è bassa pari al 40%, un risultato non buono che significa che solo il 40% delle istanze nel dataset è stata classificata correttamente dal modello.
La Precision, che indica quanto il modello è preciso nel classificare le istanze positive, possiamo osservare che per tutte le classi varia tra 70% e 97% nel caso del bilanciamento della Y (indicando un'alta precisione per tutte le classi), mentre varia dallo 0,06% al 66% nel caso del bilanciamento del Train.

Per il bilanciamento della Y, il valore di Precision pesato è dell'86% e indica che l'86% delle istanze classificate come positive dal modello sono effettivamente positive. Mentre per quanto riguarda il bilanciamento del Train il valore della Precision pesata è dell'48%. Possiamo dire che solo nel caso del bilanciamento della Y il valore di Precision sia buono, difatti un valore elevato di Precision è particolarmente importante quando si vogliono evitare i falsi positivi. Non possiamo dire lo stesso per il bilanciamento del Train.

La Recall, nel caso del bilanciamento della Y, possiamo osservare che varia tra il 47% e il 100% per le classi, mentre nel caso del bilanciamento del Train il suo valore varia dallo 8% al 47%. Il valore della Recall pesata è dell'87% (per la figura di sinistra), quindi possiamo affermare che l'87% delle istanze positive nel dataset è stato correttamente identificato dal modello ed è positivo. Mentre otteniamo un valore di Recall dell'40% per quanto riguarda il bilanciamento del Train, il che significa che l'40% delle istanze positive nel dataset è stato correttamente identificato dal modello.

L'F1- score è utile quando si desidera avere una valutazione complessiva delle prestazioni del modello e per quanto concerne il bilanciamento della Y, possiamo constatare che l'F1-score varia tra il 56% e 98% per le diverse classi, mentre per quanto riguarda il bilanciamento del Train il suo valore varia tra il 6% e il 50%. Il valore dell'F1- score pesato è dell'86% nel caso del bilanciamento della Y e indica un equilibrio tra Precision e Recall suggerendo che il modello è in grado di ottenere buoni risultati su entrambe le metriche e di performance. Mentre per il bilanciamento del Train anche il valore dell'F1 score è molto basso, pari al 42%.

Bilanciamento Y

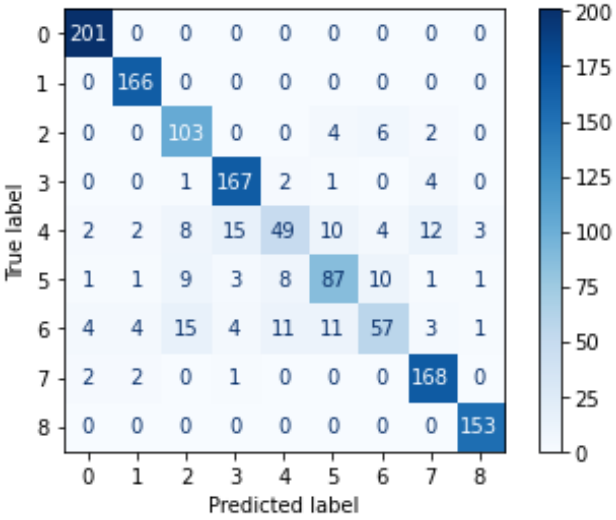
	Precision	Recall	F1-score	Support
0	0.96	1.00	0.98	201
1	0.95	1.00	0.97	166
2	0.76	0.90	0.82	115
3	0.88	0.95	0.92	175
4	0.70	0.47	0.56	105
5	0.77	0.72	0.74	121
6	0.74	0.52	0.61	110
7	0.88	0.97	0.93	173
8	0.97	1.00	0.98	153
accuracy			0.87	1319
macro avg	0.84	0.84	0.83	1319
weighted avg	0.86	0.87	0.86	1319

Bilanciamento Train

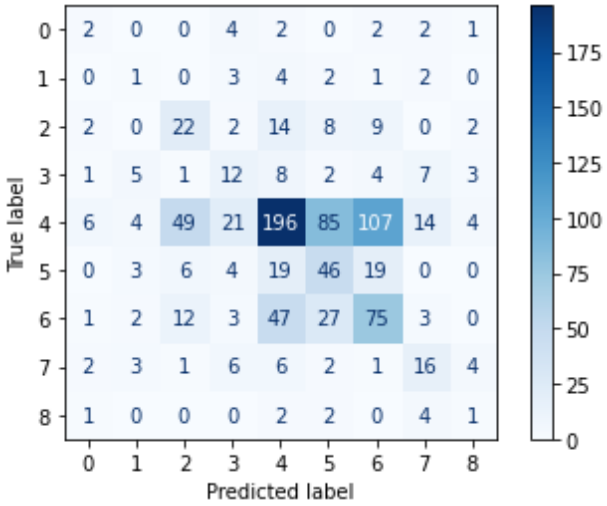
	Precision	Recall	F1-score	Support
0	0.13	0.15	0.14	13
1	0.06	0.08	0.06	13
2	0.24	0.37	0.29	59
3	0.22	0.28	0.24	43
4	0.66	0.40	0.50	486
5	0.26	0.47	0.34	97
6	0.34	0.44	0.39	170
7	0.33	0.39	0.36	41
8	0.07	0.10	0.08	10
accuracy			0.40	932
macro avg	0.26	0.30	0.27	932
weighted avg	0.48	0.40	0.42	932

Matrice di confusione

Bilanciamento Y



Bilanciamento Train



Analizzando in dettaglio le matrici per entrambi i bilanciamenti, possiamo trarre le seguenti osservazioni:

Per la classe 0, per quanto riguarda il bilanciamento della Y sono presenti 201 istanze e tutte sono classificate correttamente come Classe 0. Non ci sono falsi negativi per questa classe ma ci sono 9 falsi positivi.

Per il bilanciamento del Train invece sono stati correttamente classificati solo 2 casi. Tuttavia, sono presenti 11 falsi negativi, mentre ci sono 13 falsi positivi.

Per la classe 1, per quanto riguarda il bilanciamento della Y sono presenti 166 istanze che sono state classificate correttamente come Classe 1. Non ci sono falsi negativi per questa classe ma ci sono 9 falsi positivi. Mentre per quanto riguarda il bilanciamento del Train è stato correttamente classificato come classe 1 solo 1 caso, mentre ci sono stati 12 falsi negativi e 17 falsi positivi. In particolare, possiamo notare che per questa classe c'è una certa difficoltà nel classificare i valori.

Per quanto concerne la classe 2 per il bilanciamento della Y sono presenti 103 istanze correttamente classificate come classe 2. Ci sono 33 falsi positivi mentre ci sono 12 falsi negativi per questa classe.

Mentre per quanto riguarda il bilanciamento del Train sono state correttamente predette 22 istanze come classe 2, mentre ci sono 37 falsi negativi e 69 falsi positivi.

Per la classe 3 sono presenti 167 istanze classificate correttamente come Classe 3. Ci sono 8 falsi negativi per questa classe, mentre ci sono 23 falsi positivi. Mentre per il bilanciamento del Train sono state correttamente classificate come classe 3 solo 12 casi, con 43 falsi positivi con altre classi. Sono presenti 31 falsi negativi.

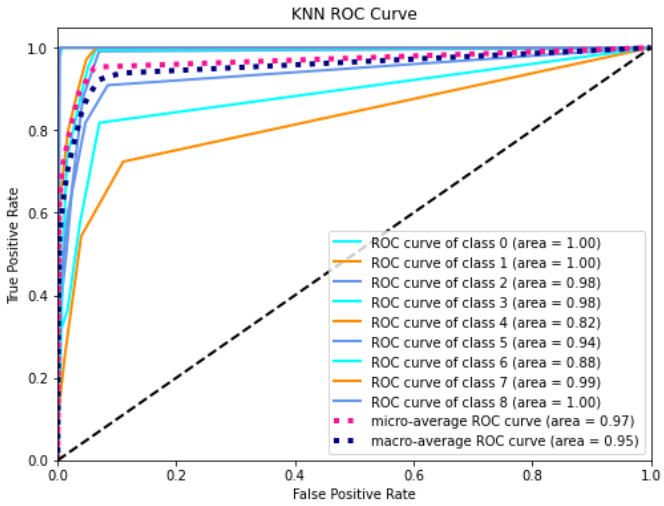
Per quanto riguarda il bilanciamento della Y nella classe 4 sono presenti 49 istanze classificate correttamente. Quindi, abbiamo 21 falsi positivi per la Classe 4 e 56 falsi negativi. Mentre nel bilanciamento del Train la classe 4 mostra una buona capacità di distinzione, con 196 casi correttamente classificati come classe 4. Tuttavia sono presenti 102 falsi positivi e 290 falsi negativi.

Per la classe 5, sono presenti 87 istanze classificate correttamente. Quindi, abbiamo 26 falsi positivi e 34 falsi negativi. Mentre per quanto riguarda il bilanciamento del Train, sono stati correttamente classificati come classe 5 solo 46 casi. Tuttavia, sono presenti 51 falsi negativi e 128 falsi positivi.
Per la classe 6 nella matrice di confusione riferita al bilanciamento del train sono presenti 57 istanze classificate correttamente come Classe 6. Quindi, abbiamo 53 falsi negativi e 20 falsi positivi. Mentre per quanto riguarda il bilanciamento del train nella classe 6 sono stati correttamente classificati come classe 6 ben 75 casi, ma ci sono 143 falsi positivi e 95 falsi negativi.
Mentre per la classe 7 nel grafico a sinistra sono presenti 168 istanze e tutte sono classificate correttamente come Classe 7. Ci sono 22 falsi positivi e 5 falsi negativi per questa classe. Nell'altro bilanciamento sono stati correttamente classificati 16 casi. Tuttavia, sono presenti 25 falsi negativi e 32 falsi positivi.
Infine per l'ultima classe, ovvero la classe 8 sono presenti 153 istanze e tutte sono classificate correttamente. Non ci sono falsi negativi per questa classe, ma ci sono 5 falsi positivi. Mentre per il bilanciamento del Train sono stati correttamente classificati come classe 8 solo 1 caso. Sono presenti 14 falsi positivi e 9 falsi negativi.

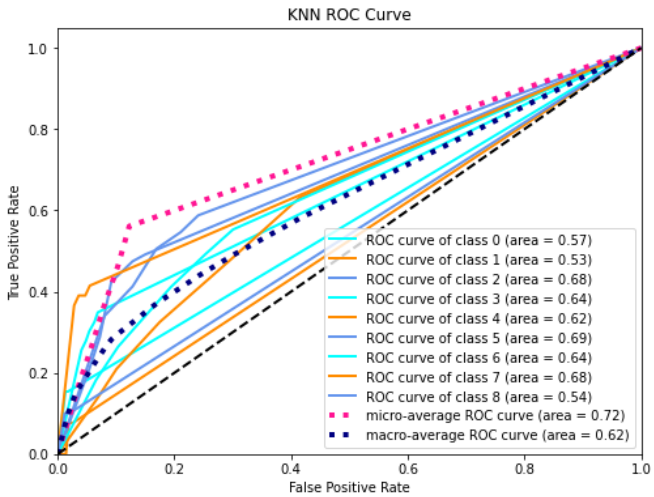
Possiamo concludere che la matrice di confusione del bilanciamento della Y mostra migliori prestazioni del modello di classificazione rispetto alla matrice del bilancimaneto del Train, in quanto presenta un maggior numero di istanze correttamente classificate per ciascuna classe. Alcune classi mostrano una buona capacità di distinzione, mentre altre mostrano una confusione significativa con diverse classi. Sarebbe utile analizzare ulteriormente le cause di questa confusione e cercare di migliorare le prestazioni del modello.

Curva ROC

Bilanciamento Y



Bilanciamento Train



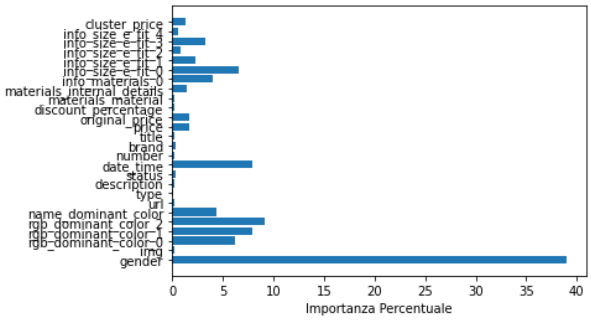
Dopo aver commentato i risultati della matrice di confusione passiamo al confronto dei risultati dei due bilanciamenti per la curva ROC e possiamo affermare che i risultati sono buoni solamente per il bilanciamento della Y, purtroppo i valori risultanti dal bilanciamento del train sono più vicino allo 0.5 che all'1.00.

Per quanto riguarda la curva Roc del bilanciamento della Y possiamo notare una buona discriminazione in quanto l'AUC per ogni classe varia da 0.82 a 1.00, certamente non è alta come il Random Forest e l'XG-Boost ma siamo comunque soddisfatti e quindi anche in questo caso il modello è in grado di distinguere molto bene tutte le classi senza grandi errori di classificazione.

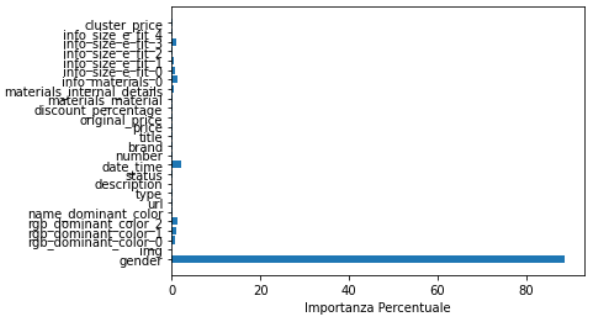
Purtroppo però la curva Roc del bilanciamento del Train dà dei risultati non molto soddisfacenti, in quanto l'AUC per ogni classe varia da 0.53 a 0.69, il che significa che il modello non è in grado di fare buone. In questo caso, la curva ROC per ciascuna classe sarà simile ad una linea diagonale, che va dal punto (0,0) al punto (1,1) in modo lineare. In questo caso abbiamo una discriminazione casuale, cioè che il modello KNN non ha buone capacità discriminante per distinguere le classi.

Features più rilevanti

Bilanciamento Y



Bilanciamento Train



Riportando i grafici per entrambi i bilanciamenti delle features più rilevanti per l'algoritmo KNN, si può notare che come per l'Algoritmo XG-Boost, la caratteristica più rilevante è anche qui gender. Per il bilanciamento del Train, gender presenta un valore (percentuale) più alto quasi al 90% rispetto al bilanciamento della Y che presenta un valore pari al 38%.

La feature info_size_e_fit_0 non si trova tra le prime ma questa volta tra le ultime. Gli altri attributi rilevanti insieme a gender sono stati img (ovvero l'immagine della borsa) e rgb_dominant_color_0 (ovvero i colori dei materili) anche se presentano valori di gran lunga più bassi rispetto a gender.

Anche l'algoritmo KNN ha identificato il genere come una caratteristica che fornisce informazioni discriminanti e influenti per la classificazione delle borse. Pertanto, il genere può essere utilizzato quasi come come unico indicatore importante per distinguere le diverse tipologie di borse.

- **Classificatore Decision Tree**

I risultati ottenuti in termini di Accuracy, Precision, Recall e F1-score indicano un basso rendimento del modello Decision Tree in questo problema di classificazione multiclasse per entrambi i bilanciamenti in quanto i risultati sono inferiori all'80%.

Andando più nel dettaglio possiamo notare un basso valore di Accuracy per il bilanciamento della Y pari al 66%, il che significa che il 66% delle istanze nel dataset è stata classificata correttamente dal modello e che

quindi è in grado di classificare la maggior parte delle istanze correttamente. Mentre per il bilanciamento del Train, l'Accuracy è pari al 47%, un risultato non buono che significa che solo il 47% delle istanze nel dataset è stata classificata correttamente dal modello.

La Precision, che indica quanto il modello è preciso nel classificare le istanze positive, possiamo osservare che per tutte le classi varia tra 48% e 91% nel caso del bilanciamento della Y (indicando un'alta precisione per tutte le classi), mentre varia dallo 0,08% all'88% nel caso del bilanciamento del Train.

Per il bilanciamento della Y, il valore di Precision pesato è del 70% e indica che il 70% delle istanze classificate come positive dal modello sono effettivamente positive. Mentre per quanto riguarda il bilanciamento del Train il valore della Precision pesata è dell'68%. Entrambi i valori seppur vicini all'80% purtroppo non possiamo affermare che siano buoni.

La recall, misura la capacità del modello di individuare correttamente le istanze positive ed è preferibile averla alta perché più alta è e meno falsi negativi ci sono, purtroppo però in questo caso non è così.

Nel caso del bilanciamento della Y, possiamo osservare che varia tra il 45% e il 92% per le classi, mentre nel caso del bilanciamento del Train il suo valore varia dallo 13% all'80%. Il valore di Recall pesato è dell'66% (per la figura di sinistra), quindi possiamo affermare che il 66% delle istanze positive nel dataset è stato correttamente identificato dal modello. Mentre otteniamo un valore di Recall del 47% per quanto riguarda il bilanciamento del Train, il che significa che l'47% delle istanze positive nel dataset è stato correttamente identificato dal modello.

L'F1-score, per quanto concerne il bilanciamento della Y, possiamo constatare che varia tra il 47% e il 77% per le diverse classi, mentre per quanto riguarda il bilanciamento del train il suo valore varia tra il 14% e il 70%. Il valore dell'F1-score pesato è dell'66% nel caso del bilanciamento della Y e indica un basso equilibrio tra Precision e Recall suggerendo che il modello non è molto in grado di ottenere buoni risultati su entrambe le metriche e di performance. Mentre per il bilanciamento del Train anche il valore dell'F1 score è ancora più basso, ovvero pari al 52%.

Bilanciamento Y

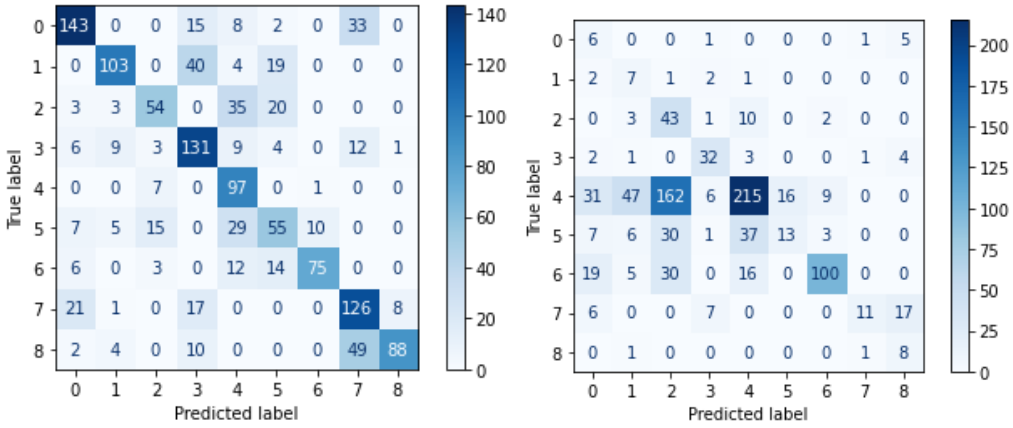
	Precision	Recall	F1-score	Support
0	0.76	0.71	0.74	201
1	0.82	0.62	0.71	166
2	0.66	0.47	0.55	115
3	0.62	0.75	0.68	175
4	0.50	0.92	0.65	105
5	0.48	0.45	0.47	121
6	0.87	0.68	0.77	110
7	0.57	0.73	0.64	173
8	0.91	0.58	0.70	153
accuracy			0.66	1319
macro avg	0.69	0.66	0.65	1319
weighted avg	0.70	0.66	0.66	1319

Bilanciamento Train

	Precision	Recall	F1-score	Support
0	0.08	0.46	0.14	13
1	0.10	0.54	0.17	13
2	0.16	0.73	0.26	59
3	0.64	0.74	0.69	43
4	0.76	0.44	0.56	486
5	0.45	0.13	0.21	97
6	0.88	0.59	0.70	170
7	0.79	0.27	0.40	41
8	0.24	0.80	0.36	10
accuracy			0.47	932
macro avg	0.45	0.52	0.39	932
weighted avg	0.68	0.47	0.52	932

Matrice di Confusione

Bilanciamento Y **Bilanciamento Train**



Analizzando in dettaglio le matrici per entrambi i bilanciamenti, possiamo trarre le seguenti osservazioni:

Per la classe 0, per quanto riguarda il bilanciamento della Y sono presenti 201 istanze di cui 143 classificate correttamente come Classe 0. I falsi negativi per questa classe sono 58 mentre i falsi positivi sono 45.

Per il bilanciamento del Train invece sono stati correttamente classificati solo 6 casi. Tuttavia, sono presenti 7 falsi negativi, mentre ci sono 67 falsi positivi.

Per la classe 1, per quanto riguarda il bilanciamento della Y sono presenti 103 casi classificati correttamente come Classe 1. Ci sono 63 falsi negativi per questa classe mentre i falsi positivi sono 22. Mentre per quanto riguarda il bilanciamento del Train sono stati correttamente classificati come Classe 1 solo 7 casi, mentre ci sono stati 6 falsi negativi e 63 falsi positivi. Possiamo notare che anche per questa classe c'è una certa difficoltà nel classificare i valori.

Per quanto concerne la classe 2, per il bilanciamento della Y sono presenti 54 istanze. Ci sono 28 falsi positivi mentre ci sono 61 falsi negativi per questa classe. Mentre per quanto riguarda il bilanciamento del Train sono state correttamente predette 43 casi come classe 2, mentre ci sono 16 falsi negativi e ben 266 falsi positivi.

Per la classe 3 sono presenti 131 casi classificati correttamente come Classe 3. Ci sono 44 falsi negativi per questa classe, mentre ci sono 82 falsi positivi. Mentre per il bilanciamento del Train sono state correttamente classificate come classe 3 appena 32 casi, con 18 falsi positivi con altre classi e 11 falsi negativi.

Per quanto riguarda il bilanciamento della Y nella classe 4 sono presenti 105 istanze, di cui 97 classificate correttamente. Poi ci sono 97 falsi positivi per la Classe 4 e 8 falsi negativi. Mentre nel bilanciamento del Train la classe 4 presenterebbe una buona capacità di distinzione, con 215 casi correttamente classificati come classe 4. Tuttavia, sono presenti un numero di falsi negativi molto alto pari a 271 (più alto dei casi classificati correttamente) e 67 falsi positivi.

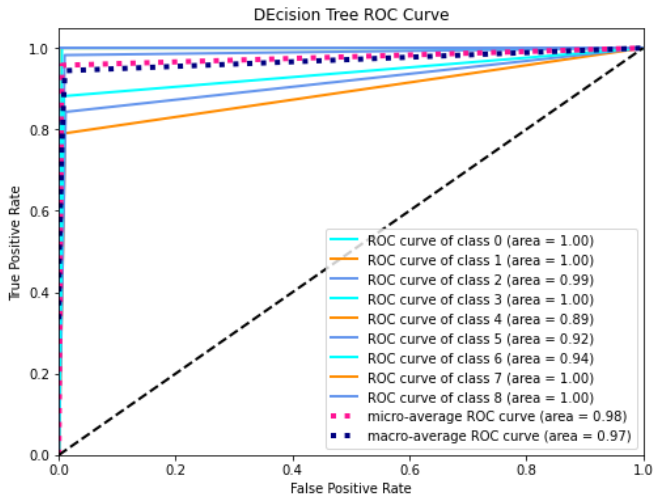
Per la classe 5, sono presenti 55 classi classificate correttamente. Quindi, abbiamo 59 falsi positivi e 66 falsi negativi. Mentre per quanto riguarda il bilanciamento del Train, sono stati correttamente classificati come classe 5 solo 13 casi. Sono presenti 84 falsi negativi e 16 falsi positivi.

Per la classe 6 nella matrice di confusione riferita al bilanciamento della Y sono presenti 75 istanze classificate correttamente come Classe 6. Quindi, abbiamo 35 falsi negativi e 11 falsi positivi. Mentre per quanto riguarda il bilanciamento del train nella classe 6 sono stati correttamente classificati come classe 6 ben 100 casi, anche se ci sono 14 falsi positivi e 70 falsi negativi.
Mentre per la classe 7 nel grafico a sinistra sono presenti 126 classi classificate correttamente come Classe 7. Ci sono 94 falsi positivi e 46 falsi negativi per questa classe. Nell'altro bilanciamento sono stati correttamente classificati solo 11 casi. Tuttavia, sono presenti 3 falsi positivi e 30 falsi negativi.
Infine per l'ultima classe, ovvero la classe 8 sono presenti 88 classi e tutte sono classificate correttamente. Ci sono 65 falsi negativi per questa classe, ma ci sono 9 falsi positivi. Mentre per il bilanciamento del Train sono stati correttamente classificati come classe 8 solo 8 casi. Sono presenti 2 falsi negativi e 26 falsi positivi.

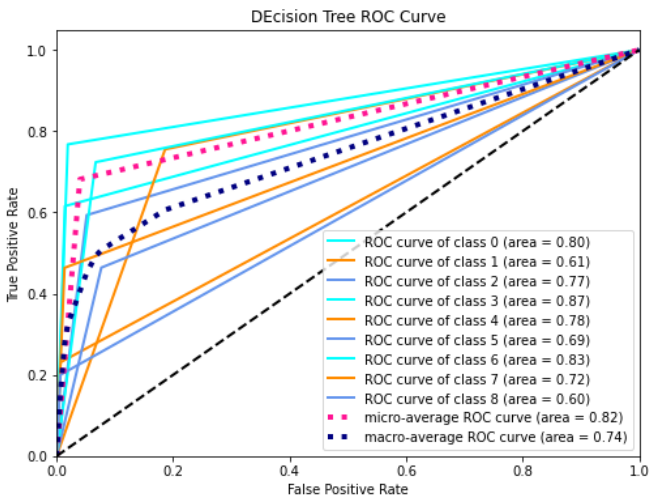
In conclusione, la matrice di confusione del bilanciamento della Y mostra migliori prestazioni del modello di classificazione rispetto alla matrice del bilancimaneto del Train, in quanto presenta un maggior numero di istanze correttamente classificate per ciascuna classe. Alcune classi mostrano una buona capacità di distinzione, mentre altre mostrano una confusione significativa con diverse classi. Sarebbe utile analizzare ulteriormente le cause di questa confusione e cercare di migliorare le prestazioni del modello.

Curva ROC

Bilanciamento Y



Bilanciamento Train



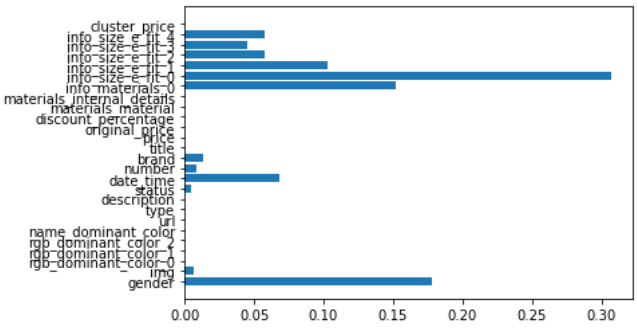
Dopo aver commentato i risultati della matrice di confusione passiamo al confronto dei risultati dei due bilanciamenti per la curva ROC e possiamo affermare che per il Decision Tree i risultati sono buoni solamente per il bilanciamento della Y, purtroppo i valori risultanti dal bilanciamento del Train sono più vicini allo 0.5 che all'1.00 anche se sono un po' più alti dei risultati ottenuti dal KNN.

Per quanto riguarda la curva Roc del bilanciamento della Y possiamo notare una buona discriminazione in quanto l'AUC per ogni classe varia da 0.89 a 1.00, certamente non è alta come il Random Forest o l'XG-Boost ma siamo comunque soddisfatti e quindi anche in questo caso il modello è in grado di distinguere molto bene tutte le classi senza grossolani errori di classificazione.

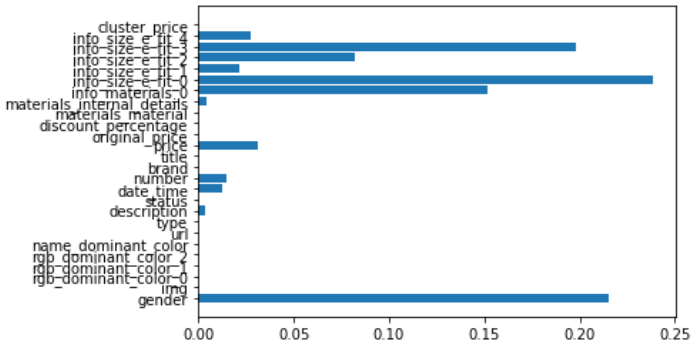
La curva Roc del bilanciamento del Train dà dei risultati un po' meno soddisfacenti, in quanto l'AUC per ogni classe varia da 0.60 a 0.80, il che significa che il modello non è in grado di fare buonissime previsioni. In questo caso, la curva ROC per ciascuna classe sarà simile ad una linea diagonale, che va dal punto (0,0) al punto (1,1) in modo lineare. In questo caso abbiamo una discriminazione casuale, cioè che il modello Decision Tree non ha grandi capacità discriminanti per distinguere le classi.

Features più rilevanti

Bilanciamento Y



Bilanciamento Train



Riportando i grafici per entrambi i bilanciamenti delle features più rilevanti per l'algoritmo Decision Tree, si può notare che, come per l'Algoritmo Random Forest, la caratteristica più rilevante è info_size_e_fit_0 con un valore vicino a 0,25 per quanto riguarda il Bilanciamento del Train, mentre di 0,30 per il Bilanciamento della Y. Si può notare che la features gender è al secondo posto per importanza quindi vediamo che si posiziona anche qui nelle più alte posizioni.

- **Classificatore Logistic Regression**

I risultati ottenuti in termini di Accuracy, Precision, Recall e F1-score indicano un basso rendimento del modello Logistic Regression in questo problema di classificazione multiclasse per entrambi i bilanciamenti in quanto i risultati sono inferiori all'80%.

Andando più nel dettaglio possiamo notare un basso valore di Accuracy per il bilanciamento della Y pari al 54% il che significa che il 54% delle istanze nel dataset è stata classificata correttamente dal modello e che quindi è in grado di classificare la maggior parte delle istanze correttamente. Mentre per il bilanciamento del Train l'Accuracy è pari al 33%, un risultato molto basso che significa che solo il 33% delle istanze nel dataset è stata classificata correttamente dal modello.

La Precision, che indica quanto il modello è preciso nel classificare le istanze positive, possiamo osservare che per tutte le classi varia tra 44% e 62% nel caso del bilanciamento della Y (indicando un'alta precisione per tutte le classi), mentre varia dallo 0,08% all'84% nel caso del bilanciamento del Train.
Per il bilanciamento della Y, il valore di Precision pesato è del 53% e indica che il 53% delle istanze classificate come positive dal modello sono effettivamente positive. Mentre per quanto riguarda il bilanciamento del Train il valore della Precision pesata è dell'59%. Entrambi i valori purtroppo non possiamo affermare che siano buoni in quanto sono molto vicini al 50%.
La Recall, nel caso del bilanciamento della Y, possiamo osservare che varia tra il 26% e il 68% per le classi, mentre nel caso del bilanciamento del Train il suo valore varia dallo 23% all'60%. Il valore di Recall pesato è del 54% (per la figura di sinistra), quindi possiamo affermare che il 54% delle istanze positive nel dataset è stato correttamente identificato dal modello. Mentre otteniamo un valore di Recall dell'33% per quanto riguarda il bilanciamento del Train, il che significa che l'33% delle istanze positive nel dataset è stato correttamente identificato dal modello. Entrambi i risultati del modello sono molto bassi.
L'F1-score per quanto concerne il bilanciamento della Y, possiamo constatare che varia tra il 31% e il 65% per le diverse classi, mentre per quanto riguarda il bilanciamento del Train il suo valore varia tra il 13% e il 52%. Il valore dell'F1-score pesato è dell'53% nel caso del bilanciamento della Y, mentre per il bilanciamento del Train anche il valore dell'F1 score è molto basso, pari al 36%. Entrambi i risultati suggeriscono che il modello è in grado di ottenere bassi risultati di performance.

Bilanciamento Y

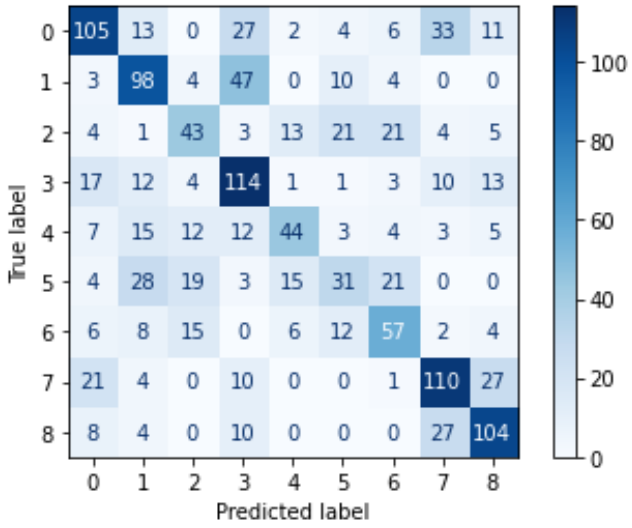
	Precision	Recall	F1-score	Support
0	0.60	0.52	0.56	201
1	0.54	0.59	0.56	166
2	0.44	0.37	0.41	115
3	0.50	0.65	0.57	175
4	0.54	0.42	0.47	105
5	0.38	0.26	0.31	121
6	0.49	0.52	0.50	110
7	0.58	0.64	0.61	173
8	0.62	0.68	0.65	153
accuracy			0.54	1319
macro avg	0.52	0.52	0.51	1319
weighted avg	0.53	0.54	0.53	1319

Bilanciamento Train

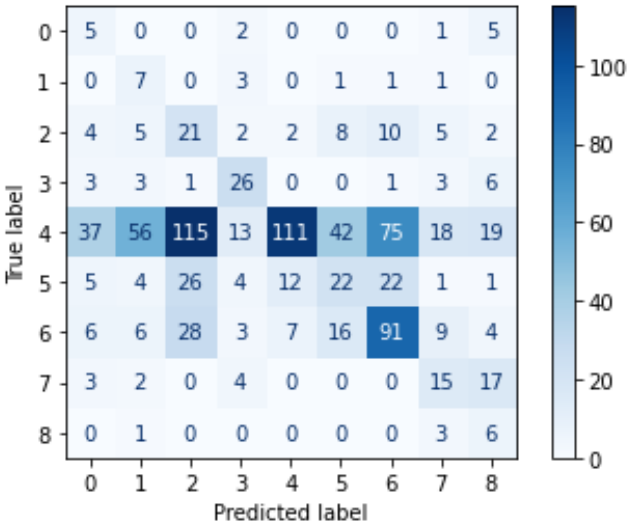
	Precision	Recall	F1-score	Support
0	0.08	0.38	0.13	13
1	0.08	0.54	0.14	13
2	0.11	0.36	0.17	59
3	0.46	0.60	0.52	43
4	0.84	0.23	0.36	486
5	0.25	0.23	0.24	97
6	0.46	0.54	0.49	170
7	0.27	0.37	0.31	41
8	0.10	0.60	0.17	10
accuracy			0.33	932
macro avg	0.29	0.43	0.28	932
weighted avg	0.59	0.33	0.36	932

Matrice di confusione

Bilanciamento Y



Bilanciamento Train



Per la classe 0, per quanto riguarda il bilanciamento della Y sono presenti 201 istanze di cui 105 classificate correttamente come Classe 0. I falsi negativi per questa classe sono 96 mentre i falsi positivi sono 70.

Per il bilanciamento del Train invece sono stati correttamente classificati solo 5 casi. Tuttavia, sono presenti 8 falsi negativi, mentre ci sono 58 falsi positivi.

Per la classe 1, per quanto riguarda il bilanciamento della Y sono presenti 98 casi classificati correttamente come Classe 1. Ci sono 68 falsi negativi per questa classe, mentre i falsi positivi sono 85. Mentre per quanto riguarda il bilanciamento del Train è stato correttamente classificato come classe 1 solo 7 casi, mentre ci sono stati 6 falsi negativi e 77 falsi positivi. In particolare, possiamo notare che per questa classe c'è una certa difficoltà nel classificare i valori.

Per quanto concerne la classe 2 per il bilanciamento della Y sono presenti 43 casi classificati correttamente. Ci sono 54 falsi positivi mentre ci sono 72 falsi negativi per questa classe.

Mentre per quanto riguarda il bilanciamento del Train sono state correttamente predette 21 casi come classe 2, mentre ci sono 38 falsi negativi e 170 falsi positivi.

Per la classe 3 sono presenti 114 casi classificati correttamente come Classe 3. Ci sono 61 falsi negativi per questa classe, mentre ci sono 112 falsi positivi. Mentre per il bilanciamento del Train sono state correttamente classificate come classe 3 appena 26 casi, con 31 falsi positivi e 17 falsi negativi.

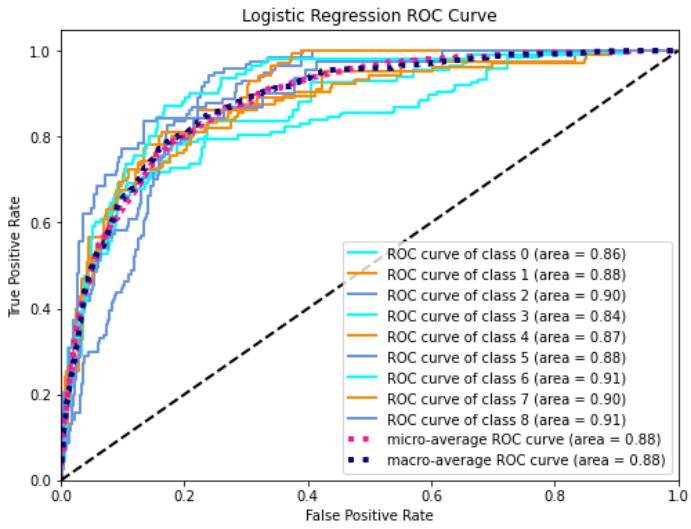
Per quanto riguarda il bilanciamento della Y nella classe 4 sono presenti 105 istanze di cui 44 classificate correttamente. Poi ci sono 37 falsi positivi per la Classe 4 e 61 falsi negativi. Mentre nel bilanciamento del Train la classe 4 mostra una buona capacità di distinzione, con 111 casi correttamente classificati come classe 4. Anche se tuttavia sono presenti un numero di falsi negativi molto alto pari a 375 (più alto dei casi classificati correttamente) e 21 falsi positivi.

Per la classe 5, sono presenti 31 classi classificate correttamente. Quindi, abbiamo 51 falsi positivi e 90 falsi negativi. Mentre per quanto riguarda il bilanciamento del Train, sono stati correttamente classificati come classe 5 solo 22 casi. Sono presenti 75 falsi negativi e 67 falsi positivi.
Per la classe 6 nella matrice di confusione riferita al bilanciamento della Y sono presenti 57 istanze classificate correttamente come Classe 6. Quindi, abbiamo 53 falsi negativi e 60 falsi positivi. Mentre per quanto riguarda il bilanciamento del Train nella classe 6 sono stati correttamente classificati come classe 6 solo 91 casi, anche se ci sono 109 falsi positivi e 79 falsi negativi.
Mentre per la classe 7 nel grafico a sinistra sono presenti 110 classi classificate correttamente come Classe 7. Ci sono 79 falsi positivi e 63 falsi negativi per questa classe. Nell'altro bilanciamento sono stati correttamente classificati solo 15 casi. Tuttavia, sono presenti 26 falsi negativi e 41 falsi positivi.
Infine, per l'ultima classe, ovvero la classe 8 sono presenti 104 classi classificate correttamente. Ci sono 49 falsi negativi per questa classe e 65 falsi positivi. Mentre per il bilanciamento del Train sono stati correttamente classificati come classe 8 solo 6 casi. Sono presenti 4 falsi negativi e 54 falsi positivi.

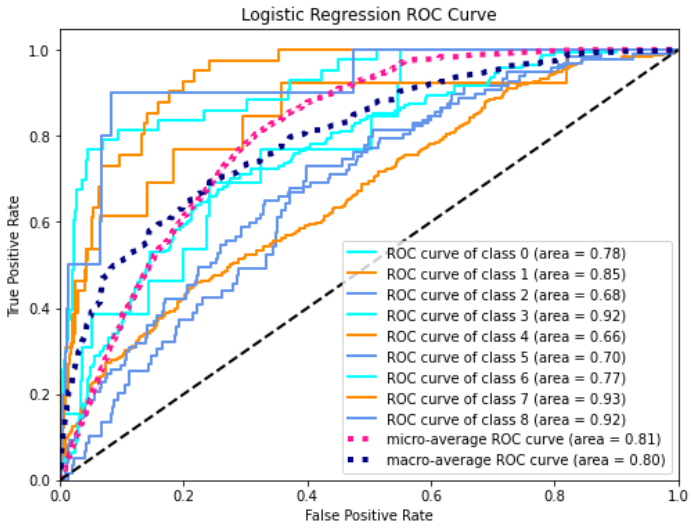
In definitiva, la matrice di confusione del bilanciamento della Y mostra migliori prestazioni del modello di classificazione rispetto alla matrice del bilancimaneto del Train, in quanto presenta un maggior numero di istanze correttamente classificate per ciascuna classe. Alcune classi mostrano una buona capacità di distinzione, mentre altre mostrano una confusione significativa con diverse classi. Sarebbe utile analizzare ulteriormente le cause di questa confusione e cercare di migliorare le prestazioni del modello.

Curva ROC

Bilanciamento Y



Bilanciamento Train

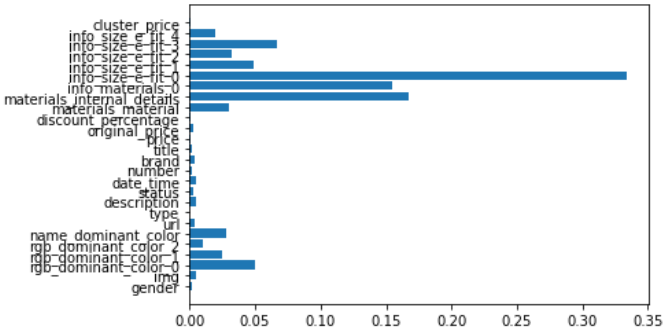


Dopo aver commentato i risultati della matrice di confusione passiamo al confronto dei risultati dei due bilanciamenti per la curva ROC e possiamo affermare che la Logistic Regression presenta dei risultati che sono discretamente buoni per entrambi i bilancimaneti.

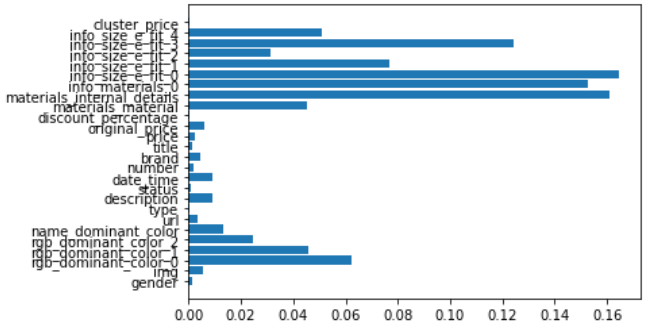
Per quanto riguarda la curva Roc del bilanciamento della Y possiamo notare una buona discriminazione in quanto l'AUC per ogni classe varia da 0.84 a 0.91, certamente non è alta come il Decision Tree ma siamo comunque soddisfatti perché i valori sono buoni e vicini all'1.00 e quindi anche in questo caso il modello è in grado di distinguere molto bene tutte le classi senza errori di classificazione.
La curva Roc del bilanciamento del Train dà dei risultati un po' meno soddisfacenti, in quanto l'AUC per ogni classe varia da 0.66 a 0.93, il che significa che il modello è in grado di fare previsioni discrete, quindi il modello Logistic Regression presenta buone ma non ottime capacità di distinguere le classi.

Features più rilevanti

Bilanciamento Y



Bilanciamento Train



Riportando i grafici per entrambi i bilanciamenti delle features più rilevanti per l'algoritmo Logistic Regression si può osservare che, come per l'Algoritmo Random Forest e Decision Tree, la caratteristica più rilevante è info_size_e_fit_0 con un valore molto vicino allo 0,16 per quanto riguarda il Bilanciamento del Train, mentre di 0,32 per quanto riguarda il bilanciamento della Y. Incredibilmente, in questo caso, si può osservare che Gender è la feature con il valore più basso di tutte e quindi questo algoritmo non identifica il genere come una caratteristica rilevante che fornisce informazioni per la categorizzazione delle borse.

- **Classificatore Naïve Bayas**

I risultati ottenuti in termini di Accuracy, Precision, Recall e F1 score indicano un basso rendimento del modello Naïve Bayas in questo problema di classificazione multiclasse per entrambi i bilanciamenti in quanto i risultati sono inferiori all'80% e possiamo affermare che questo è il classificatore che presenta i risultati più bassi di tutti.
Andando più nel dettaglio possiamo notare un basso valore di Accuracy per il bilanciamento della Y pari al 36% il che significa che il 36% delle istanze nel dataset è stata classificata correttamente dal modello e che quindi è in grado di classificare solo poche istanze correttamente. Mentre per il bilanciamento del Train l'Accuracy è pari al 19%, un risultato non buono che significa che solo il 19% delle istanze nel dataset è stata classificata correttamente dal modello.

La Precision, che indica quanto il modello è preciso nel classificare le istanze positive, possiamo osservare che per tutte le classi varia tra 28% e 54% nel caso del bilanciamento della Y (indicando un'alta precisione per tutte le classi), mentre varia dallo 0,04% all'75% nel caso del bilanciamento del Train.
Per il bilanciamento della Y, il valore di Precision pesato è del 38% e indica che il 38% delle istanze classificate come positive dal modello sono effettivamente positive. Mentre per quanto riguarda il bilanciamento del Train il valore della Precision pesata è dell'49%. Entrambi i valori purtroppo non possiamo affermare che siano buoni in quanto sono addirittura più bassi del 50%.
La Recall, nel caso del bilanciamento della Y, possiamo osservare che varia tra il 18% e il 70% per le classi, mentre nel caso del bilanciamento del Train il suo valore varia dal 12% al 70%. Il valore di Recall pesato è del 36% (per la figura di sinistra), quindi possiamo affermare che l'36% delle istanze positive nel dataset è stato correttamente identificato dal modello. Mentre otteniamo un valore di Recall del 19% per quanto riguarda il bilanciamento del Train, il che significa che il 19% delle istanze positive nel dataset è stato correttamente identificato dal modello. I risultati ottenuti sono veramente molto bassi.
L'F1-score per quanto concerne il bilanciamento della Y, possiamo constatare che l'F1-score varia tra il 24% e il 50% per le diverse classi, mentre per quanto riguarda il bilanciamento del Train il suo valore varia tra il 7% e il 35%. Il valore dell'F1-score pesato è dell'35% nel caso del bilanciamento della Y, mentre per il bilanciamento del Train anche il valore dell'F1 score è molto basso, pari al 21%. Ciò suggerisce che il modello non è in grado di ottenere buoni risultati di performance per entrambi i bilanciamenti.

Bilanciamento Y

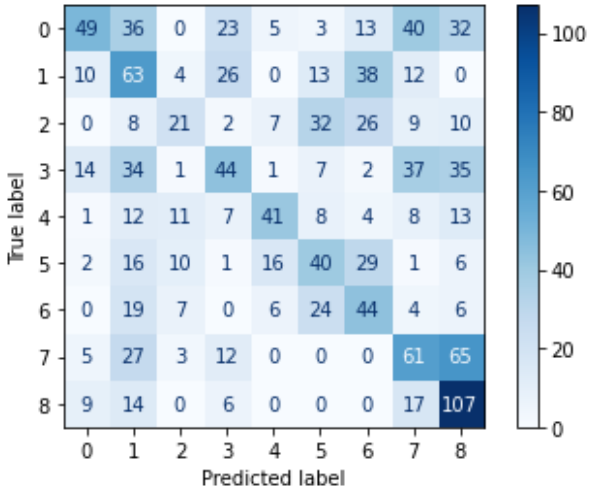
	Precision	Recall	F1-score	Support
0	0.54	0.24	0.34	201
1	0.28	0.38	0.32	166
2	0.37	0.18	0.24	115
3	0.36	0.25	0.30	175
4	0.54	0.39	0.45	105
5	0.31	0.33	0.32	121
6	0.28	0.40	0.33	110
7	0.32	0.35	0.34	173
8	0.39	0.70	0.50	153
accuracy			0.36	1319
macro avg	0.38	0.36	0.35	1319
weighted avg	0.38	0.36	0.35	1319

Bilanciamento Train

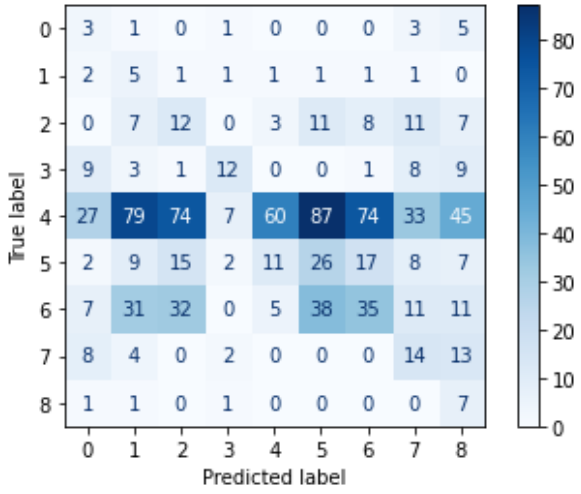
	Precision	Recall	F1-score	Support
0	0.05	0.23	0.08	13
1	0.04	0.38	0.07	13
2	0.09	0.20	0.12	59
3	0.46	0.28	0.35	43
4	0.75	0.12	0.21	486
5	0.16	0.27	0.20	97
6	0.26	0.21	0.23	170
7	0.16	0.34	0.22	41
8	0.07	0.70	0.12	10
accuracy			0.19	932
macro avg	0.23	0.30	0.18	932
weighted avg	0.49	0.19	0.21	932

Matrice di confusione

Bilanciamento Y



Bilanciamento Train



Per la classe 0, per quanto riguarda il bilanciamento della Y sono presenti 201 istanze di cui 49 classificate correttamente come Classe 0. I falsi negativi per questa classe sono 152 mentre i falsi positivi sono 41. Per il bilanciamento del Train invece sono stati correttamente classificati solo 3 casi. Tuttavia, sono presenti 10 falsi negativi, mentre ci sono 56 falsi positivi.

Per la classe 1, per quanto riguarda il bilanciamento della Y sono presenti 63 casi classificati correttamente come Classe 1. Ci sono 103 falsi negativi per questa classe mentre i falsi positivi sono 166. Mentre per quanto riguarda il bilanciamento del Train sono stati correttamente classificati come classe 1 solo 5 casi, mentre ci sono stati 8 falsi negativi e 135 falsi positivi.

Per quanto concerne la classe 2 per il bilanciamento della Y sono presenti 21 casi classificati correttamente. Ci sono 36 falsi positivi, mentre ci sono 36 falsi negativi per questa classe.

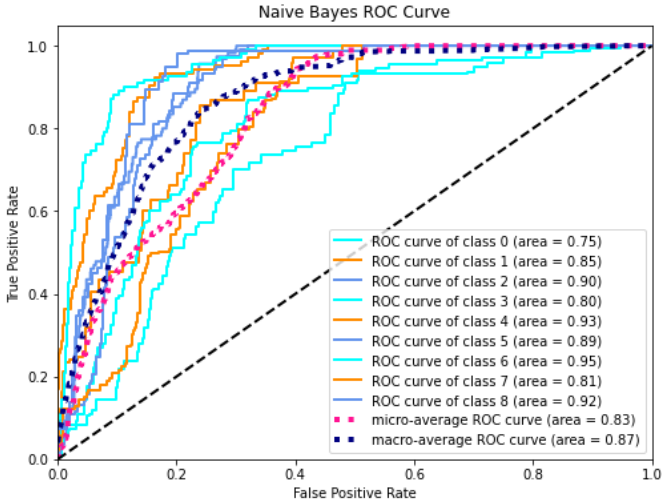
Mentre per quanto riguarda il bilanciamento del Train sono state correttamente predette 12 casi come classe 2, mentre ci sono 47 falsi negativi e 123 falsi positivi.

Per la classe 3 sono presenti 44 casi classificati correttamente come Classe 3. Ci sono 131 falsi negativi per questa classe, mentre ci sono 77 falsi positivi. Mentre per il bilanciamento del Train sono state correttamente classificate come classe 3 appena 12 casi, con 14 falsi positivi e 31 falsi negativi.
Per quanto riguarda il bilanciamento della Y nella classe 4 sono presenti 41 casi classificati correttamente. Poi ci sono 35 falsi positivi per la Classe 4 e 64 falsi negativi. Mentre nel bilanciamento del Train la classe 4 mostra una buona capacità di distinzione, con 60 casi correttamente classificati come classe 4. Tuttavia, sono presenti un numero di falsi negativi molto alto pari a 426 (più alto dei casi classificati correttamente) e 20 falsi positivi.
Per la classe 5, sono presenti 40 classi classificate correttamente. Quindi, abbiamo 87 falsi positivi e 81 falsi negativi. Mentre per quanto riguarda il bilanciamento del Train, sono stati correttamente classificati come classe 5 solo 26 casi. Sono presenti 71 falsi negativi e 137 falsi positivi.
Per la classe 6 nella matrice di confusione riferita al bilanciamento della Y sono presenti 44 casi classificate correttamente come Classe 6. Quindi, abbiamo 66 falsi negativi e 112 falsi positivi. Mentre per quanto riguarda il bilanciamento del Train nella classe 6 sono stati correttamente classificati come classe 6 sono 35 casi, anche se ci sono 101 falsi positivi e 135 falsi negativi.
Mentre per la classe 7 nel grafico a sinistra sono presenti 61 classi classificate correttamente come Classe 7. Ci sono 128 falsi positivi e 112 falsi negativi per questa classe. Nell'altro bilanciamento sono stati correttamente classificati solo 14 casi. Tuttavia, sono presenti 27 falsi negativi e 75 falsi positivi.
Infine per l'ultima classe, ovvero la classe 8 sono presenti 107 classi classificate correttamente. Ci sono 46 falsi negativi per questa classe e 167 falsi positivi. Mentre per il bilanciamento del Train sono stati correttamente classificati come classe 8 solo 7 casi. Sono presenti 3 falsi negativi e 97 falsi positivi.

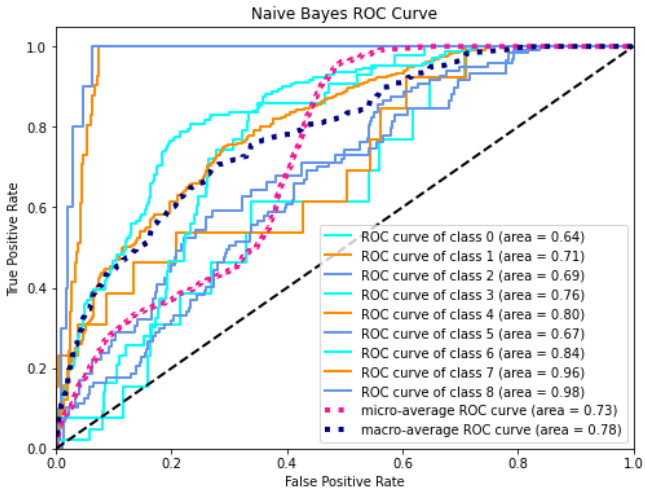
Per concludere, la matrice di confusione del bilanciamento della Y mostra migliori prestazioni del modello di classificazione rispetto alla matrice del bilancimaneto del Train, in quanto presenta un maggior numero di istanze correttamente classificate per ciascuna classe che se entrambe le matrici presentano un alto livello di confusione. Alcune classi mostrano una buona capacità di distinzione, mentre altre mostrano una confusione significativa con diverse classi. Sarebbe utile analizzare ulteriormente le cause di questa confusione e cercare di migliorare le prestazioni del modello.

Curva ROC

Bilanciamento Y



Bilanciamento Train



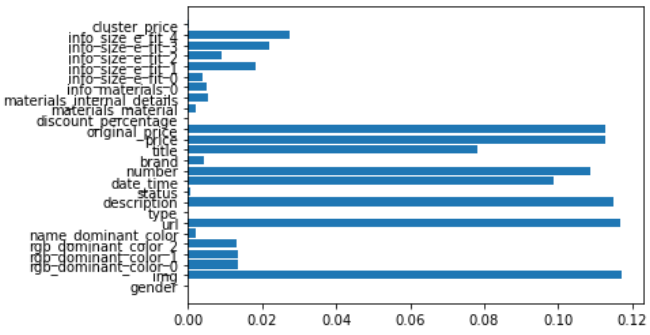
Dopo aver commentato i risultati della matrice di confusione passiamo al confronto dei risultati dei due bilanciamenti per la curva ROC e possiamo affermare che i risultati sono discretamente buoni per entrambi i bilanciamenti (anche se sono sempre migliori per il bilanciamento della Y).

Per quanto riguarda la curva Roc del bilanciamento della Y possiamo notare una buona discriminazione in quanto l'AUC per ogni classe varia da 0.75 a 0.95, certamente non è alta come la Logistic Regression o il Decision Tree ma siamo comunque soddisfatti perché i valori sono buoni e più vicini all'1.00 che allo 0.50 e quindi anche in questo caso il modello è in grado di distinguere abbastanza bene tutte le classi senza errori di classificazione.

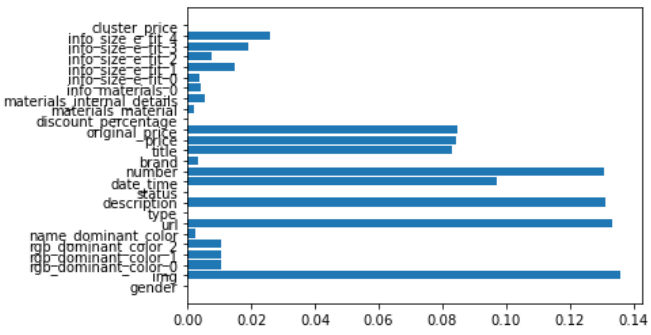
La curva Roc del bilanciamento del train dà dei risultati un po' meno soddisfacenti, in quanto l'AUC per ogni classe varia da 0.64 a 0.98, il che significa che il modello è in grado di fare previsioni discrete; quindi, il modello Naïve Bayas presenta buone ma non ottime capacità di distinguere le classi.

Features più rilevanti

Bilanciamento Y



Bilanciamento Train



Infine, riportando i grafici per entrambi i bilanciamenti delle features più rilevanti per l'algoritmo Naïve Bayas si può osservare che, come per l'Algoritmo XG-Boost e KNN, la caratteristica più rilevante è gender con un valore quasi vicino allo 0,14 per quanto riguarda il Bilanciamento del Train, mentre di uno 0,12 per il Bilanciamento della Y. Altre features rilevanti che presentano valori alti, vicini alla feature gender, sono url (url del prodotto sul sito), description (descrizione del prodotto) e number (codice del prodotto).

4.3 Applicazioni nel Marketing

In un mercato virtuale sempre più affollato, diventa sempre più difficile per gli E-tailers creare un'esperienza di qualità per cercare di soddisfare e superare le aspettative dei clienti (Rose Sebastianelli & Nabil Tamimi, 2013). I consumatori, negli ultimi anni, sempre più frequentemente si affidano alle piattaforme di shopping online ed E-commerce per effettuare i loro acquisti proprio per il fatto che sono più convenienti, permettono un maggior risparmio di tempo e una grande varietà di scelta rispetto ai negozi fisici. Di conseguenza la quantità di informazioni disponibili online è aumentata drasticamente e i clienti sono spesso sopraffatti dalle scelte nel loro processo decisionale (Chong 2013). L'enorme quantità di dati a disposizione degli E-tailer possono riguardare, nel contesto specifico dei prodotti di moda, le preferenze dei consumatori sul colore, il tessuto, lo stile, la vestibilità e molti altri. Ciò consente agli E-tailer di personalizzare e ottimizzare la presentazione dei prodotti, i messaggi pubblicitari, le strategie di pricing in base a tali attributi e non solo.
Questa enorme quantità di informazioni raccolte sono fondamentali per comprendere meglio i clienti e offrire un'esperienza di acquisto online personalizzata. Se i dati vengono utilizzati in modo efficace per offrire prodotti pertinenti, raccomandazioni personalizzate e un'interfaccia utente-intuitiva, ciò può portare a una Customer Satisfaction positiva da parte dei clienti. Al contrario, se i dati non vengono sfruttati adeguatamente o in modo errato, ciò può comportare una Customer Satisfaction negativa nell'esperienza di acquisto online.
Utilizzando algoritmi di Intelligenza Artificiale, è possibile analizzare questi dati per identificare modelli, correlazioni e insight che possono aiutare gli E-tailer a migliorare le loro strategie di Marketing e personalizzare l'esperienza di acquisto online per i clienti in modo da indirizzarne il comportamento d'acquisto.

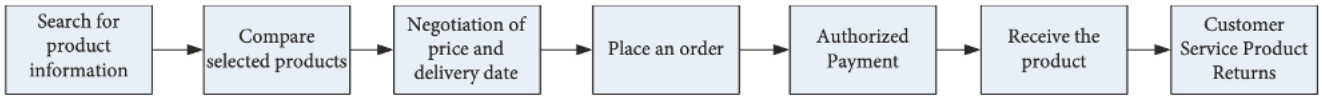


FIGURE : Basic flow chart of online shopping.

Fonte: Xiong, Y. (2022)

Il processo di acquisto di beni online, dal punto di vista del consumatore, passa solitamente attraverso diverse fasi: 1) Fase di consapevolezza di avere bisogno di un prodotto. L'acquirente riconosce un bisogno che può essere suscitato da uno stimolo interno o da uno esterno. I Marketing manager devono determinare i fattori e le situazioni che suscitano nel consumatore la percezione del problema e quindi come è possibile spingere il cliente a scegliere un determinato prodotto (Redazione, 2022). 2) Fase di ricerca delle informazioni. La

quantità delle informazioni ricercate dal consumatore dipende dall'entità dello stimolo, dalla quantità delle informazioni iniziali, dalla semplicità di ottenere nuove informazioni. In questa seconda fase, l'acquirente può ottenere informazioni da diverse fonti, come i motori di ricerca, le comunità online, articoli di blog, siti web, E-commerce, post e sponsorizzazioni sui social, recensioni, email ecc. (Redazione, 2022). Quindi è molto importante da parte dei Marketers capire quali sono i canali che gli utenti utilizzano in quanto il web è una fonte ricca di informazioni anche se molto diversi e variegati tra di loro. 3) Fase di valutazione delle alternative. Il consumatore utilizza le informazioni per giungere a restringere la propria scelta tra un numero limitato di marche per poi passare alla 4) Fase di acquisto del prodotto. Ciò comprende la circolazione delle informazioni e dei prodotti correlati, la negoziazione del prezzo e delle modalità di consegna tra acquirenti e venditori, la scelta del metodo e del termine di pagamento. In questa fase il consumatore acquisterà la marca preferita in assoluto e l'interazione principale, in questo momento del processo di acquisto, che ha il consumatore è con i siti web di E-commerce (Redazione, 2022). 5) Fase di esperienza post-acquisto. Questa comprende la fornitura di servizi post-vendita, la gestione dei reclami e la restituzione dei prodotti. Proprio qui comincia l'opera di fidelizzazione che comprende la richiesta di feedback, recensioni o l'invio di sondaggi per valutare la Customer Satisfaction (Redazione, 2022). La soddisfazione del cliente è uno dei fattori cruciali nella ricerca del comportamento dei consumatori negli ambienti commerciali tradizionali e online ed è generalmente considerata in diversi contesti comportamentali a causa del confronto tra le aspettative e l'esperienza del consumatore; questa, si raggiunge quando la consegna finale soddisfa o meglio supera le aspettative del cliente (Izyan, 2014). Nel loro studio, Sathiya et all. (2016), hanno rilevato che la soddisfazione degli acquirenti online era correlata alla zona di residenza, all'età, al sesso, all'occupazione e al costo dell'acquisto recente. Inoltre, in un ambiente online, la soddisfazione dei clienti è uno dei fattori chiave che portano a una maggiore fidelizzazione dei clienti e alla crescita a lungo termine dei negozi online (Urvashi, 2017). La maggior parte dei clienti è molto soddisfatta del prezzo del prodotto, seguito dallo sconto offerto e dal tempo impiegato per l'acquisto online (V. Mathan e R. Velmurugan, 2017; Nebojša, 2019).

Offerte e programmi di fidelizzazione sono ottime strategie di Marketing per rafforzare la Brand Loyalty, ovvero il rapporto di fedeltà alla marca da parte del cliente dopo la vendita. Se ogni fase del processo di acquisto è stata fatta correttamente, si genererà un processo di passaparola positivo sia online che offline capace di aumentare i clienti, implementare le vendite e rafforzare la Brand Awareness (Redazione, 2022).

Come illustrato nella Figura 2, le informazioni esterne sono necessarie per supportare le informazioni sul marchio o sul prodotto sia nella fase di pre-consumo che in quella di post-consumo (Xiong, Y., 2022). Ogni sito web fornisce informazioni diverse sul marchio o sul prodotto. Alcuni promuovono direttamente il marchio, altri vendono direttamente. Anche se il messaggio del marchio o del prodotto trasmesso in ogni fase varia, ogni fase richiede una pubblicità accurata e la giusta forza del messaggio per ottenere un mix mediatico perfetto (Xiong, Y., 2022).

Quando i consumatori sceglievano i metodi di acquisto tradizionali, la scelta dei rivenditori o dei negozi teneva solitamente conto di fattori quali la posizione geografica, la localizzazione del traffico del negozio, la

circolazione della rete di vendita, il passaparola, la pubblicità del prodotto e quindi si preoccupavano principalmente della qualità del servizio clienti e del servizio post-vendita al momento dell'acquisto, della comodità dell'ambiente di acquisto e dell'esplorazione del prodotto durante il processo di acquisto (Xiong, Y., 2022). La scelta degli acquirenti per lo shopping online, invece, si riflette principalmente nella scelta dei siti web commerciali quindi i fattori chiave da considerare sono le dimensioni del sito, l'adeguatezza delle informazioni sui prodotti fornite e la disponibilità di siti simili (Xiong, Y., 2022).

Un altro aspetto cruciale dell'applicazione dell'IA nel Marketing dei prodotti di moda è la capacità di personalizzare l'esperienza di acquisto per ogni singolo cliente. Utilizzando algoritmi di apprendimento automatico, l'IA può analizzare i dati sugli acquisti precedenti, le preferenze dichiarate dai consumatori e altri fattori per creare profili individuali dei clienti. Questi profili possono essere utilizzati per offrire raccomandazioni di prodotti personalizzati, suggerimenti di stile e promozioni mirate, aumentando così le probabilità di conversione e fedeltà del cliente. Le aziende che riescono ad adottare in modo efficace l'IA per l'analisi dei dati di Marketing avranno un vantaggio significativo nel mercato altamente competitivo dell'E-tailer soprattutto per migliorare e personalizzare l'esperienza d'acquisto dei consumatori.

Un aspetto importante, perciò, su cui gli E-tailers dovrebbero concentrarsi maggiormente è sicuramente la comprensione di come i clienti percepiscono e valutano la qualità delle esperienze di acquisto online, al fine di aumentare la fedeltà dei clienti. Zeithaml et all. (2002, p. 1) sostengono che "le aziende devono spostare l'attenzione dell'e-business dall'e-commerce (le transazioni) all'e-service (tutti gli spunti e gli incontri che avvengono prima, durante e dopo le transazioni)".

A questo punto possiamo definire l'esperienza d'acquisto per i clienti come la somma totale di sensazioni, emozioni e percezioni che un cliente sperimenta durante l'intero processo di acquisto. Questa esperienza può essere influenzata da diversi fattori, come l'interfaccia e l'usabilità del sito web o dell'applicazione di acquisto, la qualità dei prodotti o dei servizi offerti, il livello di assistenza e supporto al cliente, nonché la facilità e la convenienza del processo di acquisto. L'aumento della concorrenza esercita anche pressioni sulle aziende affinché attraggano i clienti e assicurino le vendite più rapidamente. Pertanto, l'analisi dei fattori che potrebbero influenzare le vendite dei prodotti sulle piattaforme online è cruciale per il successo aziendale e, di conseguenza, molte di queste variabili sono state ampiamente studiate in letteratura. Molti lavori di ricerca hanno approfondito lo studio sulla qualità nell'area del commercio elettronico (e-quality) cercando di identificare e valutare gli attributi (o dimensioni) che influenzano la percezione dei consumatori online sulla qualità dell'E-tailer (e-quality) e della soddisfazione dell'esperienza di acquisto per lo shopping online (Rose Sebastianelli & Nabil Tamimi, 2013).

Alcuni studi hanno generalmente enfatizzato la concettualizzazione della e-quality utilizzando quadri multidimensionali, e un certo numero di ricercatori ha affrontato la questione empiricamente sviluppando scale multi-item per misurare i costrutti della e-quality (Rose Sebastianelli & Nabil Tamimi, 2013).

Rose Sebastianelli & Nabil Tamimi (2013), nel loro lavoro di ricerca dal titolo "An Examination of Attributes Affecting Consumers' Perceptions of E-tailer Quality", hanno condotto una Conjoint Analysis, una

metodologia utilizzata da diverso tempo negli studi di Marketing (ad esempio, Green e Rao 1971), ma relativamente nuova per il commercio elettronico (ad esempio, Schaupp e Belanger 2005; Chen, Hsu e Lin 2010), per ricavare l'importanza relativa degli attributi nelle preferenze dei consumatori per prodotti e/o servizi, fornendo i mezzi per determinare il valore relativo di specifici attributi dell'e-tailing per le percezioni dei clienti online sulla qualità dell'e-tailer (Tamimi, N., & Sebastianelli, R., 2016).
In questo modo, l'importanza relativa di ciascun attributo può essere individuata in uno scenario decisionale più realistico (poiché quando viene chiesto di valutare direttamente l'importanza degli attributi, gli intervistati tendono a valutarli tutti come importanti). I ricercatori presentano i risultati di uno studio sperimentale in cui i partecipanti forniscono giudizi complessivi (classifiche) sulla qualità del rivenditore elettronico descritta in termini di cinque attributi (reputazione del rivenditore, usabilità del sito, sicurezza, consegna e assistenza clienti) che si sono rivelati salienti in ricerche precedenti (Rose Sebastianelli & Nabil Tamimi, 2013). Queste classifiche servono come base per la stima dei modelli Conjoint a livello individuale, da cui si ricava l'importanza relativa degli attributi. I ricercatori utilizzano l'analisi dei cluster, basata sull'importanza relativa di questi attributi, per raggruppare gli individui in diversi segmenti (Tamimi, N., & Sebastianelli, R., 2016). Forniscono un profilo di ciascun segmento, non solo in termini di percezione di come questi attributi influenzano la qualità dell'E-tailer, ma anche in termini di caratteristiche demografiche e comportamentali del cliente (Rose Sebastianelli & Nabil Tamimi, 2013). Coerentemente con i risultati della Conjoint Analysis, la voce "sicurezza delle transazioni" ha ricevuto la valutazione media più alta. Tuttavia, anche altri tre elementi, ossia la reputazione del rivenditore online, la puntualità nella consegna dell'ordine e la facilità d'uso dell'interfaccia web, hanno ricevuto valutazioni molto alte (Tamimi, N., & Sebastianelli, R., 2016). Ciò rafforza l'idea che quando agli intervistati viene chiesto di valutare direttamente l'importanza dei singoli attributi del rivenditore online essi tendono a considerarli tutti importanti. Inoltre, ciò supporta l'uso della Conjoint Analysis per accertare il valore relativo degli attributi dell'E-tailing, ossia per rappresentare una visione più realistica del modo in cui i clienti online effettuano il trade-off tra tali attributi quando valutano la qualità dell'E-tailer (Tamimi, N., & Sebastianelli, R., 2016). Anche Keen et all. (2004) hanno utilizzato questo approccio per comprendere le decisioni di acquisto dei consumatori di fronte a formati di vendita al dettaglio alternativi, ovvero negozio, catalogo e Internet. Hanno scoperto che i due attributi più importanti che guidano l'esperienza d'acquisto erano il formato di vendita al dettaglio e il prezzo del prodotto.

La qualità dell'esperienza d'acquisto dei clienti online, come abbiamo precedentemente visto, è un fattore molto importante nel processo di acquisto e questo può influenzare anche l'intenzione di rivisitazione di un determinato sito Web e quindi di conseguenza l'intenzione di riacquisto da parte di un acquirente sul sito online. Molti studi precedenti hanno riportato l'impatto dell'analisi dei Big Data sui valori e sulle sfide aziendali (Akter, 2016; Barton e Court, 2012). Tuttavia, non sono state condotte ricerche sufficienti sul punto di vista dei clienti per esaminare come questi reagiscono all'applicazione dei Big Data Analytics sull'intenzione di riacquisto online. Pertanto, la ricerca sull'intenzione di riacquisto online dei clienti basata sull'applicazione

dei Big Data Analytics, sta diventando una tendenza avanzata nella strategia di Marketing (Le e Liaw, 2017). Inoltre, anche l'effetto di mediazione tra la soddisfazione del cliente basata sui fattori di Big Data Analytics e l'intenzione di riacquisto è stato studiato in modo limitato (Hui, S. C. S., Dastane, O., Johari, Z., & Roslee, M., 2021).

Molti studiosi come Lin (2014) hanno affermato che l'intenzione di riacquisto si ha quando un comportamento ripetuto diventa abituale come risultato di un processo cognitivo automatizzato. I clienti con personalità distinte possono reagire in modo diverso a benefici simili, determinando quindi differenze nel valore percepito, che possono in ultima istanza mostrare intenzioni di riacquisto diverse (Fang, 2016). Altri come Jobo (2016) sostengono che l'intenzione di riacquisto coinvolge l'esperienza generale dei clienti fino ad oggi, la loro soddisfazione per i prodotti o i servizi e la capacità di mantenere la soddisfazione dei clienti per incoraggiare la loro intenzione di riacquisto. Ed è per questo che, come afferma anche Quoc TP (2018), l'intenzione di riacquisto è uno degli obiettivi comportamentali di Marketing più significativi che assicurano che i clienti siano disposti ad acquistare nuovamente presso lo stesso negozio o venditore online. In effetti se un cliente ha avuto un'esperienza d'acquisto di alta qualità, che supera le aspettative e soddisfa le sue esigenze, è probabile che sviluppi un'intenzione di riacquisto positiva (Hui, S. C. S., Dastane, O., Johari, Z., & Roslee, M., 2021). C'è anche da dire che piattaforme di shopping online incontrano maggiori difficoltà rispetto a un contesto off-line a causa dell'assenza di interazioni faccia a faccia con i distributori, dell'inaffidabilità dei dati di configurazione di Internet, della diffidenza, dei bassi costi di commutazione, dell'incertezza e della rapida diffusione attraverso il passaparola. (Donni, Dastane, Haba e Selvaraj, 2018). Ma se la qualità dell'esperienza d'acquisto è curata nei minimi dettagli, sicuramente questo creerà fiducia, fedeltà e una connessione emotiva con il marchio o il venditore, che a sua volta aumenterà l'intenzione di riacquisto.

In conclusione, sulla base dei risultati empirici identificati dalla letteratura sul Marketing su Internet, è emerso che anche la costruzione dell'immagine può influenzare positivamente la soddisfazione e l'intenzione di riacquisto (Arons 1961; Rich e Portis 1964; Higie. Feick e Price 1987; Dodd 1991). Come evidenziato da Alba et al. (1997), è il web design che fornirà all'E-tailer un vantaggio comparativo nell'ambiente di vendita al dettaglio. Se il design promuove la frequenza di navigazione e rivisitazione di un sito Web, sicuramente questi comportamenti contribuiranno ad incentivare l'intenzione di acquistare e riacquistare un prodotto (Hopkins, C. D. (2001).

Capitolo 5. Conclusioni e sviluppi futuri

5.1 Conclusioni

Questo lavoro di tesi ha permesso la realizzazione di un modello predittivo che affronta un problema di classificazione multiclasse, utilizzando diversi algoritmi di Machine Learning, per poter assegnare correttamente le etichette delle diverse classi di borse di un E-tailer in base alle loro caratteristiche.

Come si è visto nel Capitolo 4, i risultati migliori sono dati dall'algoritmo XG-Boost per entrambi i bilanciamenti, in quanto ha ottenuto risultati molto alti su tutte le metriche di valutazione delle prestazioni pari all'80% per quanto riguarda Accuracy, Precision, Recall e F1-score nel Bilanciamento del Train. Anche se possiamo dire che i risultati ottenuti dal Random Forest in tutte le metriche utilizzate per valutare le performance sono stati molto vicini a quelli ottenuti dall' XG-Boost, si può affermare che entrambi gli algoritmi sono risultati i migliori. Da entrambe le tabelle riportate qui sotto, si può notare che i valori ottenuti per entrambi i bilanciamenti non sono uguali, anzi nel caso del bilanciamento del Train sono più bassi. Questo perché durante il processo di addestramento, il modello ha ottimizzato i suoi parametri per minimizzare l'errore tra le sue previsioni e le etichette corrette presenti nel Train set, poiché è necessario che contenga dati correttamente etichettati o annotati, in modo che il modello possa apprendere la corretta relazione tra le caratteristiche dei dati e le risposte desiderate. Sono stati riportati i risultati del Bilanciamento della Y perché mostrano le performance del modello a livello ideale e teorico, ovvero come sarebbero stati se si fosse trattato di un caso concettuale. Però i risultati corretti e quelli su cui dobbiamo basare la nostra analisi sono naturalmente quelli derivanti dal Bilanciamento del Train, perché consente di costruire un modello predittivo rappresentativo del problema reale che si vuole risolvere riflettendo la natura dei dati che si incontreranno nel mondo reale. L'obiettivo principale dei dati di Train è quello di fornire al modello di Machine Learning esempi realistici che gli consentano di imparare e generalizzare correttamente su nuovi dati.

Procedendo con gli altri algoritmi, il KNN ha avuto dei buoni risultati solo nel caso del bilanciamento della Y in cui ha ottenuto una Precision e una Recall leggermente inferiori rispetto ai primi due algoritmi, pari all'86% e all'87%, ma comunque buoni. Purtroppo non si può dire lo stesso per il bilanciamento del Train che presenta risultati molto bassi per tutte le metriche, cioè inferiori al 50%. Gli algoritmi Decision Tree, Logistic Regression e Naïve Bayes hanno ottenuto risultati significativamente inferiori rispetto ai classificatori descritti in precedenza, soprattutto Logistic Regression e Naïve Bayes. Ciò fa presupporre che questi algoritmi non siano molto performanti per questo problema di classificazione multiclasse.

In conclusione, le features più rilevanti, ovvero quelle che hanno contribuito di più alla previsione del modello sono (per gli algoritmi che hanno avuto i migliori risultati): la feature gender (ovvero genere uomo/donna) per l'algoritmo XG-Boost e la feature info_size_e_fit_0 (ovvero informazioni riguardanti taglie e misure delle borse) per l'algoritmo Random Forest. Attraverso l'analisi accurata delle caratteristiche, è possibile identificare gli attributi più importanti e che quindi ci danno più informazioni e che ci permettono di migliorare le prestazioni del modello al fine di ottenere una migliore comprensione del problema in esame. L'algoritmo XG-

Boost ha identificato il genere (uomo/donna) come la feature più rilevante per distinguere le diverse tipologie di borse all'interno di questo problema di classificazione multiclasse, mentre per il Random Forest la feature più rilevante è quella che dà informazioni precise sulle taglie e le misure relative alle borse. Questo sta ad indicare che l'algoritmo ha riconosciuto le informazioni riguardanti taglie e misure come importanti per categorizzare le diverse tipologie di borse e le utilizza per classificazione più accurate.

Bilanciamento Train

Algoritmo	Accuracy	Precision (Weighted Average)	Recall (Weighted Average)	F-1 Score (Weighted Average)
Random Forest	0.80	0.80	0.80	0.79
XG-Boost	0.80	0.80	0.80	0.80
KNN	0.40	0.48	0.40	0.42
Decision Tree	0.47	0.68	0.47	0.52
Logistic Regression	0.33	0.59	0.33	0.36
Naïve Bayas	0.19	0.49	0.19	0.21

Bilanciamento Y

Algoritmo	Accuracy	Precision (Weighted Average)	Recall (Weighted Average)	F-1 Score (Weighted Average)
Random Forest	0.97	0.98	0.97	0.97
XG-Boost	0.98	0.98	0.98	0.98
KNN	0.87	0.86	0.87	0.86
Decision Tree	0.66	0.70	0.66	0.66
Logistic Regression	0.54	0.53	0.54	0.53
Naïve Bayas	0.36	0.38	0.36	0.35

5.2 Implicazioni manageriali e sviluppi futuri

L'obiettivo principale di questo lavoro di tesi consiste nel possibile utilizzo da parte degli E-tailer di un modello predittivo che affronta un problema di classificazione multiclasse utilizzando alcuni algoritmi di Machine Learning come il Random Forest e l'XG-Boost (ovvero gli algoritmi che hanno avuto i migliori risultati in questo problema di classificazione), che possa assegnare correttamente le etichette delle diverse classi di borse di un E-tailers in base alle loro caratteristiche. L'impiego dell'Intelligenza Artificiale permetterà alle aziende di accorciare il tempo impiegato e le risorse spese per fare questo lavoro manualmente così da poter utilizzare questo tempo e le informazioni raccolte in altre attività aziendali come:

1. Personalizzazione dei prodotti: le aziende possono utilizzare queste informazioni per personalizzare l'offerta di prodotti e servizi offrendo suggerimenti e raccomandazioni personalizzate in base ai gusti e alle preferenze individuali dei clienti. L'utilizzo dell'Intelligenza Artificiale potrebbe consentire agli E-tailer di analizzare anche come gli attributi dei prodotti di moda influenzano i comportamenti e le preferenze dei clienti che potrebbero essere utilizzati per creare modelli di raccomandazione personalizzati basati sui gusti e sulle preferenze degli acquirenti. Ciò potrebbe consentire alle aziende di offrire suggerimenti di borse più pertinenti e adatte ai singoli clienti.

2. Migliorare l'esperienza dell'utente: attraverso l'analisi dell'Intelligenza Artificiale, è possibile identificare quali attributi (ad esempio, colore, taglia, stile, materiale) sono più rilevanti per i clienti e ottimizzare l'organizzazione e la presentazione dei prodotti sulla piattaforma di E-commerce in modo da poter migliorare l'esperienza di shopping online.

3. Ottimizzazione delle strategie di Marketing: l'intelligenza artificiale può aiutare le aziende a identificare quali attributi dei prodotti di moda sono più attraenti per i clienti e come posizionarli efficacemente nelle strategie di Marketing. Queste informazioni consentono di sviluppare campagne pubblicitarie mirate e creare contenuti ad hoc per i consumatori target che sono più inclini ad essere interessati a determinati tipi di borse e quindi presentare loro annunci pertinenti in modo tale da aumentare il loro coinvolgimento, ovvero la Customer Engagement.

4. Previsione della domanda e gestione dell'inventario: con una migliore comprensione di come gli attributi per specifici prodotti di moda influenzano la domanda dei clienti, le aziende possono ottimizzare la gestione dell'inventario regolando i livelli di stock, evitando sovra stock o scorte insufficienti e migliorare così l'efficienza nella gestione dell'inventario.

5. Innovazione del prodotto: l'analisi degli attributi dei prodotti di moda può rivelare nuove tendenze o combinazioni di attributi che sono particolarmente attraenti per i clienti. Queste informazioni possono

guidare le aziende nella progettazione e nello sviluppo di nuovi prodotti innovativi, rispondendo alle esigenze e alle preferenze dei clienti in modo più accurato e tempestivo.

6. Analisi di tendenze e stili: gli algoritmi di Machine Learning potrebbero aiutare le aziende a identificare nuove tendenze di stile, comprendere meglio i gusti dei consumatori e sviluppare borse che soddisfino le esigenze del mercato in evoluzione tramite la raccolta di grandi quantità di dati relativi alle borse come i dati di vendita, recensioni dei clienti, influenze dei social media e tendenze di moda.

7. Analisi della concorrenza: le aziende possono monitorare e analizzare i dati relativi alle borse dei concorrenti come tendenze di mercato, strategie di prezzo e caratteristiche di prodotto di successo della concorrenza. Queste informazioni consentono loro di adattare le proprie strategie di marketing e differenziarsi sul mercato.

Questo modello predittivo può essere implementato per categorizzare non solo le borse, ma anche altre categorie di prodotti moda come abbigliamento, scarpe e accessori.
L'applicazione di tali algoritmi potrebbe essere estesa anche a prodotti commercializzati da E-tailer che operano in settori diversi da quello della moda e questo permetterebbe alle aziende di sfruttare le potenzialità dell'apprendimento automatico per una varietà più ampia di prodotti, migliorando ulteriormente l'efficienza e la velocità delle operazioni aziendali.
E-tailers che operano in settori differenti sicuramente potrebbero utilizzare altri algoritmi che non sono stati implementati in questo lavoro di tesi come ad esempio l'SVM o Reti neurali artificiali in quanto non esiste un algoritmo unico che funzioni meglio su tutti i problemi di apprendimento supervisionato ma ogni problema deve essere trattato in maniera differente, quindi sicuramente gli algoritmi più performanti in questo lavoro di tesi sicuramente non saranno gli stessi che hanno avuto i migliori risultati per altri problemi di classificazione.
Un'ultima implicazione manageriale potrebbe essere cercare di colmare l'enorme divario tra le varie catene del valore dell'industria della moda come il design, la produzione e il Marketing. Questo perché dalla letteratura è emerso che i dataset di moda disponibili o sono troppo piccoli, o provengono da un'unica fonte di dati, o sono stati creati su misura per un compito specifico, o coprono un breve periodo di tempo. Manca un buon set di dati di riferimento per addestrare, testare, valutare e confrontare le prestazioni dei diversi algoritmi per l'analisi della moda. Pertanto, sarebbe estremamente utile per i ricercatori se esistesse un dataset di moda unificato su larga scala, che contenga dati di diverse modalità e che copra un lungo periodo di tempo (Fashion analysis and understanding with artificial intelligence) per raffinare la loro capacità di categorizzazione delle borse.

Come prospettiva futura a questo lavoro di tesi, si potrebbero migliorare i risultati ottenuti modificando alcuni parametri utilizzati per i diversi algoritmi di classificazione o implementando modelli più sofisticati e complessi che offrano risultati di classificazione ancora più accurati, riducendo sempre di più la necessità di un intervento manuale.

Si potrebbero anche utilizzare dati non strutturati in quanto attualmente l'impiego di algoritmi di Machine Learning per categorizzare le borse si basa principalmente su attributi strutturati come dimensioni, materiali e colori. In un futuro si potrebbe, quindi, mirare a integrare algoritmi di apprendimento automatico in grado di elaborare dati non strutturati come immagini o descrizioni testuali delle borse, consentendo una categorizzazione ancora più dettagliata e precisa.

Una prospettiva futura potrebbe anche essere quella di combinare questo modello di Machine Learning con l'utilizzo di modelli di Deep Learning come reti neurali convoluzionali (CNN) e reti neurali ricorrenti (RNN) in quanto offrono potenzialità avanzate nella comprensione di immagini e testi per poter migliorare ulteriormente la precisione e l'efficienza della categorizzazione delle handbags.

In conclusione si potrebbe utilizzare un dataset più ricco di attributi e informazioni, più dettagliato rispetto a quello che è stato esaminato in questo lavoro di tesi in modo da poter migliorare ancora di più la categorizzazione delle borse e si potrebbe estendere il lavoro di ricerca non limitandolo alla sola Nazione Italia ma anche a tutte le altre nazioni in cui opera l'Etailer Mytheresa ovvero Giappone, Korea del Sud e Stati Uniti d'America.

Bibliografia

Alba, J., Lynch, J., Weitz, B., Janiszewki, C,, Lutz, R. Sawyer. A., and Wood S. (1997). "Interactive Home Shopping: Consumer. Retailer, and Manufacturer Incentives to Participate in Electronic Market Places. Journal of Marketing. 61. 38-53.

Adeline, C. P. (2006). E-Commerce: A Study on Online Shopping in Malaysia. J. Soc. Sci., 13(3), 231–242.

Al-Halah, Z., Stiefelhagen, R., & Grauman, K. (2017, October 22–29). Fashion forward: Forecasting visual style in fashion. In 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy (pp. 388–397). https://doi.org/10.1109/ICCV.2017.50

Akter, S. F. (2016). Big Data Analytics in ECommerce:A Systematic Review and Agenda for Future Research. Electronic Markets-The International Journal.

Aggarwal, C.C., Zhai, C. (2012). A survey of text classification algorithms. In: Aggarwal, C.C., Zhai, C. (eds.) Mining Text Data, pp. 163–222. Springer, Boston (2012). https://doi.org/10.1007/978-1-4614-3223-4 6

Agrawal R. and Srikant R. (2001, May). On Integrating Catalogs. In Proceedings of the Tenth International World Wide Web Conference (WWW2001), Hong Kong.

Agrawal, A., Gans, J. S., & Goldfarb, A. (2018). Prediction machines: The simple economics of artificial intelligence. Harvard Business School Press.

Andrew, O. L. (2014). Factors Influencing Consumers Repurchase Intention of Groupon. Faculty Of Accountancy And Management Deparment Of International Business.

Ankush, S. (2017). Dynamic Recommendation System Using Enhanced K-means Clustering Algorithm for E-commerce. International Journal of Advanced Research in Computer Science, 8(5), 159–163.

Anwuri, P. N., & Eke. (2020). H.O. Content Marketing strategies and customer patronage of E-tailers in Port Harcourt.

Amin, M., Rezaei, S., Valaei, N., Wan Ismail, W.K., (2015). Gender differences and consumers' repurchase intention: the impact of trust propensity usefulness and ease of use for implication of innovative online retail. Int. J. Innovat. Learn. 17 (2), 217–233.

Arabasadi Z. et al. (2013, October). Prediction and optimization offire pro offing properties of intumescent flame retardant coatings using artificial intelligence techniques," Fire Saf. J., vol. 61, pp. 193–199.

Arons, L., (1961). Does Television Viewing Influence Store Image and Shopping Frequency? Journal o f Retailing, 3 7 .1-13.

Asdecker, B., & Karl, D. (2018, September). Big data analytics in returns management–Are complex techniques necessary to forecast consumer returns properly?. In 2nd International Conference on Advanced Reserach Methods and Analytics (CARMA 2018) (pp. 39-46). Editorial Universitat Politècnica de València.

Avinash, B.M., A. B. (2017). Big Data Analytics for E-Commerce – Its Impact on Value Creation. International Journal of Advanced Research in Computer and Communication Engineering, 6(12), 181–188.

Baird, N. (2018, March). What Digital Transformation Actually Means for Retail. Retrieved from https://www.fo
rbes.com/sites/nikkibaird/2018/03/13/what-digital-transformation-actually-means-for-retail/#35f7acaf7038

Başak, D. (2013). Revenue Management in E-Commerce: A Case Study. Proceedings of the International MultiConference of Engineers and Computer Scientists.

Brad, A., Delemare, A., Hurley, N., Lenikus, V., Mulrenan, R., Nemes, N., Trunk, U., Urbancic, N. (2018). The False Promise of Certification; Technical Report; Changing Markets Foundation: Utrecht, The Netherlands.

Breiman, L. (2001). Random forests. Machine Learning, 45(1):5–32.

Brosnan T. and D.-W. Sun (2004). Improving quality inspection of food products by computer vision—A review. J. Food Eng., vol. 61, no. 1, pp. 3–16.

Besson M., Faita F., Peretz I., A.-M.Bonnel, and J. Requin (1998). Singing in the brain: Independence of lyrics and tunes. Psychological Science, 9.

Carpenter, J. (2015). IBM's Virginia Rometty tells NU grads: Technology will enhance us. Retrieved February 11, 2019 from https://www. chicagotribune.com/bluesky/originals/ct-northwestern-virginia-rometty-ibm-bsi-20150619-story.html.

Castro, W., Oblitas, J., De-La-Torre, M., Cotrina, C., Bazán, K., & Avila-George, H. (2019). Classification of cape gooseberry fruit according to its level of ripeness using machine learning techniques and different color spaces. IEEE access, 7, 27389-27400.

Chaitanya, B. (2018). Opportunistic Returns and Dynamic Pricing: Empirical Evidence from Online Retailing in Emerging Markets. Academic Press.

Chakrabarti, S. et al. (1997). Using taxonomy, discriminants, and signatures for navigating in text databases. In Proceedings of the 23rd VLDB Conference, Greece.

Chen, Y., I. Hsu, and C. Lin. 2010. Website attributes that increase consumer purchase intention: A conjoint analysis. Journal of Business Research 63 (9=10): 1007–1014.

Chong, A. Y., Chan, F. T. S., & Ooi, K.-B. (2012). Predicting consumer decisions to adopt mobile com- merce: Cross country empirical examination between China and Malaysia. Decision Support Systems, 53(1), 34–43. doi:10.1016/j.dss.2011.12.001

Chong AYL (2013) Previsione dei determinanti dell'adozione del m-commerce: un approccio di rete neurale. Esperto Sist Appl 40(2):523–530

Chou, S.-W., & Hsu, C.-S. (2016). Understanding online repurchase intention: Social exchange theory and shopping habit. Information Systems and e-Business Management, 45(1), 14–19. doi:10.100710257- 015-0272-9

Chui, M., Manyika, J., Miremadi, M., Henke, N., Chung, R., Nel, P., & Malhotra, S. (2018). Notes from the AI frontier: Applications and value of deep learning. McKinsey global institute discussion paper, April 2018. Retrieved June 12, 2019 from https://www.mckinsey. com/featured-insights/artificial-intelligence/notes-from-the-ai- frontier-applications-and-value-of-deep-learning.

Ciecierski, K. A., & Kamola, M. (2020). Comparison of text classification methods for government documents. In Artificial Intelligence and Soft Computing: 19th International Conference, ICAISC 2020, Zakopane, Poland, October 12-14, 2020, Proceedings, Part I 19 (pp. 39-49). Springer International Publishing.

Coleman, S., Göb, R., Manco, G., Pievatolo, A., Tort-Martorelle, X., & Reis, M. S. (2016). How Can SMEs Benefit from Big Data? Challenges and a Path Forward. Quality and Reliability Engineering International 32(6), 2151–2164.

Columbus, L. (2019). 10 charts that will change your perspective of AI in marketing. Forbes, July 07. Retrieved on July 09, 2019 from https:// www.forbes.com/sites/louiscolumbus/2019/07/07/10-charts-that-will-change-your-perspective-of-ai-in-marketing/amp/

C. Cortes and V. Vapnik (1995). Support-Vector Networks, Machine Learning, 20:273-297.

C. J. C. Burges (1998). A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, 2(2):955-974.

Dai, B., Liu, N., & Jian, Z. (2021). Providing Data Insights to Suppliers for Product Development: Incentive Analysis of an E-Tailer. Discrete Dynamics in Nature and Society, 2021, 1-15.

da Silva, D. V. (2018). Modeling the pricing strategy of an e-commerce luxury fashion retailer: A machine learning approach.

Dacko, S.G. (2017), "Enabling smart retail settings via mobile augmented reality shopping apps", Technological Forecasting and Social Change, Vol. 124, pp. 243-256.

Davenport, T. (2006). Competing on Analytics. Harvard Business, 84, 98–107. PMID:16447373

Davenport, T. H., Dalle Mule, L., & Lucker, J. (2011). Know what your customers want before they do. Harvard Business Review, 89(12), 84–92.

Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. Harvard Business Review, 96(1), 108–116.

Davenport, T. H. (2018). The AI advantage: How to put the artificial intelligence revolution to work. MIT Press.

Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. Journal of the Academy of Marketing Science, 48, 24-42.

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. Management Information Systems Quarterly, 13(3), 319–340. doi:10.2307/249008

Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dodds. W.B. (1991). In Search of Value: How Price and Store Information Influence Buyers' Product Perceptions. Journal of Consumer Marketing, 2. 15-24.

Donni, R., Dastane, O., Haba, H. F., & Selvaraj, K. (2018). Consumer Perception Factors for Fashion M-Commerce and its Impact on Loyalty among Working Adults. Business and Economic Review, 8(2), 168–192. doi:10.5296/ber.v8i2.13044

Domingos, P. 2012. A few useful things to know about machine learning. Communications of the ACM 55(10): 78–87.

Ding, Y., Korotkiy, M., Omelayenko, B., Kartseva, V., Zykov, V., Klein, M., ... & Fensel, D. (2002, April). Goldenbullet: Automated classification of product data in e-commerce. In Proceedings of the 5th international conference on business information systems (Vol. 5, No. 7).

Du C.-J and D.-W. Sun (2008), "Multi-classification of pizza using computer vision and support vector machine," J. Food Eng., vol. 86, no. 2, pp. 234–242.
Dumais S. and Chen H. (2000, August). Hierarchical Classification of Web Content. In Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval, pp. 256-263, Athens, Greece.

Erevelles, S., Fukawa, N. and Swayne, L. (2015), Big data consumer analytics and the transformation of marketing. Journal of Business Research, in press.

Fagerstrøm, A., Eriksson, N., Sigurdsson, V., (2020). Investigating the impact of Internet of Things services from a smartphone app on grocery shopping. J. Retailing Consum. Serv. 52, 101927.

Fan, X., Ning, N., & Deng, N. (2020). The impact of the quality of intelligent experience on smart retail engagement. Marketing Intelligence & Planning.

Fang, J. (2016). Consumer Heterogeneity, Precieved Value and Repurchase Decision Makring In Onle Shopping: The Role of Genederm Agem abd Shopping Motives. Journal of Electronic Commerce Re- search, 17(2), 116–133.

Fawzy, M. (2018). E- Commerce Adoption and An Analysis of the Popular E- Commerce Business Sites in Malaysia. Journal of Internet Banking and Commerce, 23(1), 1–10.

Farhang, S. (2012). The Impact of Website Information Convenience On E-commerce Success Of Companies. Procedia: Social and Behavioral Sciences, 57, 381–387. doi:10.1016/j.sbspro.2012.09.1201

Fensel, D.,Ding, Y., Schulten, E., Omelayenko, B., Botquin, G., Brown, M. and Flett, A. (2001). Product Data Integration in B2B E-commerce, IEEE Intelligent System , 16(3).

Fiona, F.-H. N. (2005). Information Search Patterns in E-Commerce Product Comparison Services. SIGHCI 2005 Proceedings, 3.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics 7(2), 179–188.

Fisher, M., R. Vaidyanathan. 2012. Which Products Should You Stock? Harvard Business Review, November 2012.
Fisher, M., R. Vaidyanathan. 2014. An algorithm and demand esti- mation procedure for retail assortment optimization. Manage- ment Sci. 60(10): 2401–2415.

Fisher, M., S. Gallino, J. Li. 2017. Competition-based dynamic pricing in online retailing: A methodology validated with field experiments. Management Sci. Available at https://doi. org/10.1287/mnsc.2017.2753 (accessed date January 4, 2017).

Fisher, M., & Raman, A. (2018). Using data and big data in retailing. Production and Operations Management, 27(9), 1665-1669.

Gaasbeek, A., Golsteijn, L., Vieira, M. (2015). LCA in the Labelling Industry—Understand the LCA Landscape of the Production of Self- Adhesive Labels; Technical Report; PRé Consultants BV: Amersfoort, The Netherlands

Gans, J., Agrawal, A., & Goldfarb, A. (2017). How AI will change strategy: A thought experiment. Harvard business review online. Retrieved February 11, 2019 from https://hbr.org/product/how-ai- will-change-strategy-a-thought-experiment/H03XDI-PDF-ENG.

Gaudin, S. (2016). At stitch fix, data scientists and a.I. become personal stylists. Retrieved February 11, 2019 from https://www. computerworld.com/article/3067264/artificial-intelligence/at-stitch- fix-data-scientists-and-ai-become-personal-stylists.html.

Gers, F.A., Schmidhuber, J., Cummins ,F. (1999). Learningtoforget:Continualprediction with LSTM

Getman, R. R., Green, D. N., Bala, K., Mall, U., Rawat, N., Appasamy, S., & Hariharan, B. (2021). Machine learning (ML) for tracking fashion trends: Documenting the frequency of the baseball cap on social media and the runway. Clothing and Textiles Research Journal, 39(4), 281-296.

Green, P., and V. Rao. 1971. Conjoint measurement for quantifying judgmental data. Journal of Marketing Research 8:355–363.

G. Salton, A. Wong, and C. S. Yang: (1975). A vector space model for automatic indexing, Communications of the ACM, 18(7):613-620, 1975.

Gu, X., Gao, F., Tan, M., & Peng, P. (2020). Fashion analysis and understanding with artificial intelligence. Information Processing & Management, 57(5), 102276.

Gupta, A. R. (2013). Online Shopping: A Shining Future. International Journal of Techno-Management Research, 1(1), 2321–3744.

Gunasekaran, A., Marri, H. B., McGaughey, R. E., & Nebhwani, M. D. (2002). E-commerce and its impact on operations management. International Journal of Production Economics, 1(75), 185–197.

Harding, K. (2017). AI and machine learning for predictive data scoring. Retrieved February 11, 2019 from https://www.objectiveit.com/ blog/use-ai-and-machine-learning-for-predictive-lead-scoring on 13 February 2019.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). The elements of statistical learning: data mining, inference and prediction. Springer, 2 edition.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Data Mining, Inference, and Prediction (2nd ed.). New York: Springer.

Higie, R. A.. Feick. L. F.. & Price, L. L. (1987). Types and Amount of Word-of-Mouth Communications About Retailers. Journal of Retailing 63.260-278.

Hofacker, C. F., Malthouse, E. C., & Sultan, F. (2016). Big data and consumer behavior: Imminent opportunities. Journal of consumer marketing, 33(2), 89-97.

Hopkins, C. D. (2001). Analysis of affective and functional dimensions of e-tailer image as antecedents to online search motives, involvement, and purchase intentions: An empirical test and explanation. Mississippi State University.

Hosmer, D. W. and Lemeshow, S. (2005). Applied Logistic Regression. Wiley-Blackwell.

Huang, H., & He, S. (2011, August). The study on E-tailing. In 2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC) (pp. 1771-1774). IEEE., doi: 10.1109/AIMSEC.2011.6010538).

Huang, M. H., & Rust, R. T. (2018). Artificial intelligence in service. Journal of Service Research, 21(2), 155–172.

Hui, S. C. S., Dastane, O., Johari, Z., & Roslee, M. (2021). Enhancing Online Repurchase Intention via Application of Big Data Analytics in E-Commerce. In Handbook of Research on Innovation and Development of E-Commerce and E-Business in ASEAN (pp. 395-434). IGI Global.

Jansen, B. and McNeese M. (2005) "Evaluating the Effectiveness of And Patterns of Interactions with Automated Searching Assistance," Journal of the American Society for Information Science and Technology, (56:14), 2005, 1480-1503.

Jiang, S., Shao, M., Jia, C., & Fu, Y. (2016). Consensus style centralizing auto-encoder for weak style classification. AAAI1223–1229.

Jiwat, R. H. (2017). Examing Impacts of Big Data Analytics on Consumer Finance: A Case of China. International Journal of Managing Information Technology, 9(3), 13–22. doi:10.5121/ijmit.2017.9302

J. M. Gomez-Hidalgo and M. B. Rodriguez (1997, July). Integrating a lexical database and a training collection for text categorization. In the Proceedings of ACL/EACL (the Association for Computational Linguistics/European Association for Computational Linguistics: Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications), Madrid, Spain.

Joachims T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In C. Nedellec and C. Rouveirol (eds.), Machine Learning: ECML-98.

Jobo, D. (2016). South African generation-X online shopper satisfaction and their repurchase intentions. Investment Management and Financial Innovations, 13(3), 371–382. doi:10.21511/imfi.13(3-2).2016.09

Jurafsky, D., Martin, J.H. (2009). Speech and Language Processing, 2nd edn. Prentice- Hall Inc, Upper Saddle River.

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implica- tions of artificial intelligence. Business Horizons, 62(1), 15–25.

Kautish, P., (2011). Consumer satisfaction and internet shopping: a literature review. International Journal of Marketing and Management Research 2 (5), 40–61.

Kautish, P., Paul, J., & Sharma, R. (2021). The effect of assortment and fulfillment on shopping assistance and efficiency: An e-tail servicescape perspective. Journal of Retailing and Consumer Services, 59, 102393.

Keen, C., M. Wetzels, K. Ruyter, and R. Feinberg. 2004. E-tailers versus retailers: Which factors determine consumer preferences. Journal of Business Research 57 (7): 685–695.

Kedzia, P., Piasecki, M., Orlin´ska, M. (2016). WoSeDon. http://hdl.handle.net/ 11321/290. CLARIN-PL digital repository

Khare, A., Kautish, P., (2021). Cosmopolitanism, self-identity, online communities, and green apparel perception. Market. Intell. Plann. https://doi.org/10.1108/MIP-11- 2019-0556 (in press).

Kiapour, M. H., Yamaguchi, K., Berg, A. C., & Berg, T. L. (2014). Hipster wars: Discovering elements of fashion styles. Eccv (1)8689. Eccv (1) 472–488.

Kiong, S. C., Lee, L. Y., Chong, S. H., Azlan, M. A., & Muhd Nor, N. H. (2013). Decision making with the analytical hierarchy process (AHP) for material selection in screw manufacturing for minimizing environmental impacts. In Applied Mechanics and Materials (Vol. 315, pp. 57-62). Trans Tech Publications Ltd.

Koehrsen, W. (2017). Random forest simple explanation. Medium.

Koirala. (2012). What is Big Data Analytics and its Application in E-Commerce? Retrieved from www.venturecity.com

Koller D. and Sahami M. (1997, July). Hierarchically classifying documents using very few words. In Proceedings of the International Conference on Machine Learning, vol 14, Morgan-Kaufmann.

InternetWorldStats. (2014). Retrieved April 2017, from www.internetworldstats.com: www.internetworldstats.com/emarketing.htm

Izyan, H. b. (2014). Factors Influencing Customer Satisfaction and E-loyalty:Online Shopping Environ- ment among the Young Adults. Management Dynamics in the Knowledge Economy, 2(3), 462–471.

Lake, K. (2018). Stitch Fix's CEO on selling personal style to the mass market. Harvard Business Review, 96(3), 35-40. Retrieved from https://hbr.org/2018/05/ stitch-fixs-ceo-on-selling-person- al-style-to-the-mass-market

Larson, K. (2019). Data privacy and AI ethics stepped to the fore in 2018. Retrieved February 11 from https://medium.com/@Smalltofeds/data-privacy-and-ai-ethics-stepped-to-the-fore-in-2018- 4e0207f28210.

Laurier, C., Grivolla, J., & Herrera, P. (2008, December). Multimodal music mood classification using audio and lyrics. In 2008 seventh international conference on machine learning and applications (pp. 688-693). IEEE.

Le, T., & Liaw, S. Y. (2017). Effects of pros and cons of applying big data analytics to consumers' re- sponses in an E-commerce context. *Sustainability*, *9*(5), 798. doi:10.3390u9050798

Li, Q. C., J. X. (2017). The Impact of Big Data Analytics on Customers Online Behaviour. Proceedings of the International MultiConference of Engineers and Computer Scientists.

Luce, L. (2019). Artificial intelligence for fashion: How AI is revolu- tionizing the fashion industry. New York: Apress.

Lin, C., & Lekhawipat, W. (2014). Factors affecting online repurchase intention. Industrial Management & Data Systems, 114(4), 597–611. doi:10.1108/IMDS-10-2013-0432

Luchese C.L., P. D. Gurak, and L. D. F. Marczak (2015). Osmotic dehydration of physalis (physalis peruviana L.): Evaluation of water loss and sucrose incorporation and the quantification of carotenoids. LWT-Food Sci. Tech- nol., vol. 63, no. 2, pp. 1128–1136.

Ma, Y., Jia, J., Zhou, S., Fu, J., Liu, Y., & Tong, Z. (2017). Towards better understanding the clothing fashion styles: A multimodal deep learning approach. AAAI38–44.

Malone, T. W. (2018). How human-computer "superminds" are redefining the future ofwork. MIT Sloan Management Review, 59(4), 34-41. Retrieved from https://sloanreview.mit.edu/ article/how-human-computer- superminds-are-redefining-the- future-of-work/

Manyika, J. C. (2011). Big data: The next frontier for innovation,competition, and productivity. Aca- demic Press.

Mathan, V., K. a. (2017). Customer Satisfaction Towards Online Shopping in Coimbatore District. International Journal of Pure and Applied Mathematics, 117(15), 41–49.

Mathan, V., & Velmurugan, R. (2017). Customer Satisfaction Towards Online Shopping in Coimbatore District. International Journal of Pure and Applied Mathematics, 117(15), 41–49.

Matzen, K., Bala, K., & Snavely, N. (2017). StreetStyle: Exploring world-wide clothing styles from millions of photos. ArXiv. http://arxiv.org/abs/1706.01869

Mehta, N., Detroja, P., & Agashe, A. (2018). Amazon changes prices on its products about every 10 minutes — here's how and why they do it. Retrieved February 11, 2019 from https://www.businessinsider. com/amazon-price-changes-2018-8?international=true&r=US&IR= T.

Meta Group (2001), 3D Data Management: Controlling Data Volume, Velocity and Variety, Gartner.

Mittal, A. (2013). E-commerce: It's Impact on consumer Behavior. Global Journal of Management and Business Studies., 3(2), 131–138.

Modi, D., & Zhao, L. (2019). Trunk Club: Revolutionizing the retail model in fashion. Process innovation in the global fashion industry, 99-121.

Moon, B.-J. (2004). Consumer adoption of the internet as an information search and product purchase channel: Some research hypotheses. International Journal of Internet Marketing and Advertising, 1(1), 104–116. doi:10.1504/IJIMA.2004.003692
Modi, D., & Zhao, L. (2019). Trunk Club: Revolutionizing the retail model in fashion. In B. Jin, & E. Cedrola (Eds.), Process innovation in the global fashion industry (pp. 99-121). New York: Palgrave Pivot.

Nadikattu, R. R. (2020). Research on data science, data analytics and big data. International Journal of Engineering, Science And, 9(5), 99-105.

Nebojša, V. (2019). The Influence of Online Shopping Determinants on Customer Satisfaction in the Serbian Market. Journal of Theoretical and Applied Electronic Commerce Research, 14(2), 70–89.

Ngo, M. A. (2015). Content marketing for Small and Medium Online Retailers. 70(9), 1-46.

Oliver, H. (2011). Price Discrimination in E-Commerce? An Examination of Dynamic Pricing in Name- Your-Own-Price Markets. Management Information Systems Quarterly, 35(1).

Orapin, L. (2009). Factors influencing Internet Shopping Behavior: A Survey of Consumers in Thailand. Journal of Fashion Marketing and Management, 13(4), 501-513

Pandya, R. (2015). C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning. International Journal of Computer Applications 117(16), 18– 21.

Parekh, J. (2018). Why Programmatic provides a better digital marketing landscape. Retrieved February 13, 2019 from https://www.adweek. com/programmatic/why-programmatic-provides-a-better-digital- marketing-landscape/.

Park, H.-A. (2013). An introduction to logistic regression: from basic concepts to interpreta- tion with particular attention to nursing domain. College of Nursing and System Biomedical Informatics National Core Research Center, Seoul National University.

Pawłowski, M. (2022). Machine learning based product classification for ecommerce. Journal of Computer Information Systems, 62(4), 730-739.

Pee, L. G., Jiang, J., & Klein, G. (2018). Signaling Effect of Website Usability on Repurchase Intention. International Journal of Information Management, 39, 228–241. doi:10.1016/j.ijinfomgt.2017.12.010

Pedregosa, F., et al. (2011). Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830

Polikar, R. (2006). Ensemble based systems in decision making. IEEE Circuits and Systems Magazine 6(3), 21–45.

Proceedings of the 10th European Conference on Machine Learning, Lecture Notes in Artificial Intelligence, LNAI 1398, pages 137-142.

Radu, L. (2015). Personalization in E-commerce using profiles similarity. Economic Sciences, 8(57), 275–282.

Reid, K. (2012, July). Chicago Startup Trunk Club Is Disrupting the Men' s Fashion Industry. Retrieved from https://www.forbes.com/sites/kellyreid/ 2012/07/13/trunk-club/#38305b5b5323

Renko, S. and Druzijanic, M. (2014), "Perceived usefulness of innovative technology in retailing: consumers' and retailers' point of view", Journal of Retailing and Consumer Services, Vol. 21 No. 5, pp. 836-843.

Rich, S. U. & Portis, B. D. (1964). The Imageries of Department Stores. Journal of Marketing. 28, 10-15

Roberto Grandinetti (2020). How artificial intelligence can change the core of marketing theory. Innovative Marketing , 16(2), 91-103. doi:10.21511/im.16(2).2020.08

Robert M., W. (2001). Online Dynamic Pricing: Efficiency, Equity and the Future of E-commerce. Virginal Journal of Law and Technology, 16(2).

Romika, Y. M. (2015). Usage of Big Data Analytics for Customer Relationship Management. Interna- tional Journal of Advanced Research in Computer Science, 6(2)

Rose Sebastianelli & Nabil Tamimi (2013). An Examination of Attributes Affecting Consumers' Perceptions of E-tailer Quality, Journal of Internet Commerce, 12:3, 268-283, DOI: 10.1080/15332861.2013.859039

Sayad, S. (2018). Logistic regression

Sathiya, B. a. (2016). Satisfaction on Online Shopping A Study with Special Reference to Pollachi Taluk. International Journal of Multidisciplinary Research and Modern Education, 2(1), 90–94.

Satinet, C., & Fouss, F. (2022). A Supervised Machine Learning Classification Framework for Clothing Products' Sustainability. Sustainability, 14(3), 1334.

Schaupp, L., and F. Belanger. 2005. A conjoint analysis of online consumer satisfaction. Journal of Electronic Commerce Research 6 (2): 95–111.

Segment. (2018). How segment helps Trunk Club deliver personal style. Retrieved from https://segment.com/customers/trunk-club/story/

Shankar, V. (2018). How artificial intelligence (AI) is reshaping retailing. Journal of Retailing, 94(4), vi–xi.

Shen, A. (2014). Recommendations as personalized marketing: Insights from customer experiences. Journal of Services Marketing, 28(5), 414-427. Retrieved from https://www.semanticscholar.

org/paper/Recommendations-as-personalized-marketing%3A-insights-Shen/25999e994e588410a bcb22758f590b7c185b780c

Shi, S., Wang, Y., Chen, X., Zhang, Q., 2020. Conceptualization of omnichannel customer experience and its impact on shopping intention: a mixed-method approach. Int. J. Inf. Manag. 50, 325–336.

Shin, D. (2017). Adaptive Algorithms for Ordinal Optimization and Dynamic Pricing in E-commerce. ProQuest.

Syam, N., & Sharma, A. (2018). Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelli- gence in sales research and practice. Industrial Marketing Management, 69, 135–146.

Singh, R., Söderlund, M., 2020. Extending the experience construct: an examination of online grocery shopping. Eur. J. Market. 54 (10), 2419–2446.

Simo-Serra, E., & Ishikawa, H. (2016). Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction. CVPR298–307.

Shvaiko P., Euzenat J., (2005). A Survey of Schema-Based Matching Approaches, Journal on Data Semantics IV, 146–171.

Statista, 2020. Fashion e-commerce report 2020. https://www.statista.com/study/38 340/ecommerce-report-fashion/. (Accessed 29 August 2020).

Tamimi, N., & Sebastianelli, R. (2016). How e-tailing attributes affect perceived quality: The potential impact of customer demographics and online behaviors. The TQM Journal.

Thi & Shu. (2017). Effects of Pros and Cons of Applying Big Data Analytics to Consumers' Responses in an E-Commerce Context. Sustainability, 9(798), 1–19.

Tomas, R. (2017). Business Intelligence for Big Data Analytics. International Journal of Computer Applications Technology and Research, 6(1), 1–8. doi:10.7753/IJCATR0601.1001

Urbanke, P., Kranz, J., & Kolbe, L. (2015). Predicting Product Returns in E-Commerce: The Contribution of Mahalanobis Feature Extraction. Proceedings of the 36th International Conference on Information Systems (ICIS), 1–12.

Urvashi, T. (2017). Analyzing customer satisfaction: Users perspective towards online shopping. Nankai Business Review International, 8(3), 266–288. doi:10.1108/NBRI-04-2016-0012

Vazquez, D., Dennis, C. and Zhang, Y. (2017), "Understanding the effect of smart retail brand - consumer communications via mobile instant messaging (MIM) - an empirical study in the Chinese context", Machines in Human Behavior, Vol. 77, pp. 425-436.

Verganti, R., Vendraminelli, L., & Iansiti, M. (2020). Innovation and design in the age of artificial intelligence. Journal of Product Innovation Management, 37(3), 212-227.

Vijay, V. (2018). Factors Influencing Consumer Behavior and Prospective Purchase Decisions in a Dy- namic Pricing Environment—An Exploratory Factor Analysis Approach. Social Sciences, 7(153), 1–14.

Wajeeha, A. (2018). Influencing Factor of Brand Perception on Consumers on repurchase intention: An Examination of Online Apparel Shopping. Journal of Contemporary Management Issues, 23(2), 87–101.

Wim, d. (2014). Dynamic Pricing as a Service for E-Commerce Applications. 20th Twente Student Conference on IT.

Xiong, Y. (2022). The impact of artificial intelligence and digital economy consumer online shopping behavior on market changes. Discrete Dynamics in Nature and Society, 2022.

Xin, K L, J. L. (2018). Facotrs Influencing Repurchase Intention in Online Shopping Context: The Me- diating Role of Satisfaction. Journal of Applied Structural Equation Modeling, 2(1), 29–43.

Yan, G. (2018). Mobile e-Commerce Recommendation System Based on Multi-Source Information Fusion for Sustainable e-Business. Sustainability, 10(147), 1–13.

Yusepaldo, P. (2018). The Effect of Online Customer Experience towards Repurchase Intention. Int. J Sup. Chain. Mgt, 7(5), 548–558.

Zakaluk R. and R. S. Ranjan (2006), ‘‘Artificial neural network modelling of leaf water potential for potatoes using RGB digital images: A greenhouse study,’’ Potato Res., vol. 49, no. 4, pp. 255–272.

Zhang, Y. S.-L.-N.-L., Guo, S.-L., Han, L.-N., & Li, T.-L. (2016). Application and Exploration of Big Data Mining in Clinical Medicine. Chinese Medical Journal, 129(6), 731–739. doi:10.4103/0366- 6999.178019 PMID:26960378

Zhang, Z., Xu, G., & Zhang, P. (2016). Research on E-Commerce Platform-Based Personalized Recommendation Algorithm. Applied Computational Intelligence and Soft Computing, 2016, 1–7. doi:10.1155/2016/5160460

Zeithaml, V., Parasuraman, A. and Malhotra, A. (2000), “A conceptual framework for understanding e-service quality: implications for future research and managerial practice”, Report No. 00-115, Marketing Science Institute, Cambridge, MA.

Zeithaml, V., Parasuraman, A. and Malhorta, A. (2002), “Service quality delivery through web sites: a critical review of extant knowledge”, Journal of the Academy of Marketing Science, Vol. 30 No. 4, pp. 362-375.

Zeithaml, V.A. (2002), “Service excellence in electronic channels”, Managing Service Quality, Vol. 12 No. 3, pp. 135-139.

Sitografia

Adam Shafi (Feb 2023). K-Nearest Neighbors (KNN) Classification with scikit-learn. (https://www.datacamp.com/tutorial/k-nearest-neighbor-classification-scikit-learn?irclickid=Ui-WvOVx%3AxyNTOQ2vSxwRRtOUkAT6x3Xi16W000&irgwc=1&utm_medium=affiliate&utm_source=impact&utm_campaign=000000_1-1310690_2-mix_3-all_4-na_5-na_6-na_7-mp_8-affl-ip_9-na_10-bau_11-Admitad%20-%201310690&utm_content=TEXT_LINK&utm_term=44276).

Anaconda (2023). Anaconda.com.

Amazon (2023). Amazon.com

Data Science Team (2020, 15 May) . Impara la scienza dei dati online. https://datascience.eu/it/programmazione/xgboost/

Decision tree classifier Spark MLlib, https://spark.apache.org/docs/2.1.0/mllib-decision-tree.html)

emarketer.com (2013)

Hasty, (2023, May 17). https://hasty.ai/docs/mp-wiki/metrics/accuracy

Jason Brownlee (2020, July 24). Train-Test Split for Evaluating Machine Learning Algorithms. Python Machine Learning. https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/

Judy Howell (2021, July 25) . Differenza tra E Tailing ed E Commerce. https://it.strephonsays.com/e-tailing-and-vs-e-commerce-15337

KamilTaylan.blog, Enciclopedia finanziaria (2021). Vendita al dettaglio elettronica (E-tailing). https://it.kamiltaylan.blog/electronic-retailing-e-tailing/

Koo Ping Shung, (2018, Mar 15). Accuracy, Precision, Recall or F1?.Towards Data Science. https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

Nour Al-Rahman Al-Serw (2021, Feb 21). Undersampling and oversampling: An old and a new approach. Analytics Vidhya. https://medium.com/analytics-vidhya/undersampling-and-oversampling-an-old-and-a-new-approach-4f984a0e8392

Matt Clarke, (2021, Saturday, May 29), https://practicaldatascience.co.uk/machine-learning/how-to-create-a-classification-model-using-xgboost

Matt Clarke, (2022, Sunday, May 08). How to create a Naive Bayes text classification model using scikit-learn.https://practicaldatascience.co.uk/machine-learning/how-to-create-a-naive-bayes-text-classification-model-using-scikit-learn).

Mytheresa.com (2023). (https://www.mytheresa.com/it/it?log=geo&rdr_src=mag)

Raffaele Felaco (2020). UpBiz. https://up-biz.net/2020/05/27/cose-le-tailing/

Red hat (2019). Cos'è un ambiente di sviluppo integrato (IDE)?.

https://www.redhat.com/it/topics/middleware/what-is-ide

Redazione (2022). Il percorso d'acquisto del cliente: le 5 fasi della vendita. https://www.ictsviluppo.it/blog/percorso-acquisto-cliente-fasi-vendita

Rohit Kundu (2022, September 13). Confusion Matrix: How To Use It & Interpret Results [Examples]. V7lab. https://www.v7labs.com/blog/confusion-matrix-guide

Satyam Kumar (2021,Sep 19). 5 Techniques to work with Imbalanced Data in Machine Learning. Towards Data Science. https://towardsdatascience.com/5-techniques-to-work-with-imbalanced-data-in-machine-learning-80836d45d30c

Statista (2023). Fashion eCommerce report 2023. https://www.statista.com/study/38340/ecommerce-report-fashion/

Appendice

L'intero codice di Python si trova al presente link:

https://drive.google.com/file/d/1_zQ5SvEIKCkzdmiNIOQUlthNhcsXfxcS/view?usp=drive_link

Riassunto

Introduzione

Analisi del contesto di riferimento

Come l'introduzione di Internet ha dato vita allo shopping online

Negli ultimi anni il rapido sviluppo della tecnologia informatica e di Internet ha creato un mercato molto favorevole allo sviluppo dello shopping online facendo di Internet lo strumento principale per accedere a varie risorse e informazioni (Xiong, Y., 2022).

L'uso diffuso di tali tecnologie segna l'ingresso della società umana nella "economia di rete" permettendo alle aziende, attraverso il commercio elettronico, di creare una connessione più diretta tra venditori e acquirenti (Xiong, Y., 2022). Quindi, per innovarsi e sopravvivere all'interno del mercato, le aziende non si sono potute limitare solo a vendere tramite punti vendita fisici ma anche tramite Internet.

Per effettuare i loro acquisti online, i consumatori si affidano sempre di più agli E-tailers che negli ultimi tempi stanno diventando protagonisti del commercio elettronico, offrendo ai consumatori la possibilità di acquistare comodamente online e aprendo così nuove opportunità di business (Kautish, P., Paul, J., & Sharma, R., 2021). Gli analisti prevedono che il mercato globale dell'E-tailing raggiungerà i 4,11 trilioni di dollari nel 2023 (Statista, 2023), con il settore moda in particolare, il più grande segmento di E-commerce del mercato al dettaglio, che si stima crescerà dell'11,5% annuo (Statista 2023). Con i rapidi progressi tecnologici guidati da Internet (Singh e Söderlund, 2020), le aziende hanno avuto prospettive di business senza precedenti oltre ai confini fisici dei tradizionali modelli di vendita al dettaglio (Amin et al., 2015; Shi et al., 2020).

La nascita dell'E-tailer

l'E-tailer (o anche vendita al dettaglio elettronica) si riferisce ad un'azienda o ad un rivenditore che effettua la vendita di beni al dettaglio tramite transazioni Business-to-Consumer (B2C) su Internet (Kautish, 2011; FagerstrØm et al., 2020).

La vendita al dettaglio online è un formato (o dimensione) abbastanza nuovo che alcuni rivenditori hanno adottato come conseguenza dell'evoluzione dell'E-commerce.

Ma che cosa fa dell'E-tailer un modello così di successo? L'E-tailing può portare vantaggi sia ai rivenditori che ai consumatori, infatti richiede alle aziende di adattare i propri modelli di business per catturare le vendite su Internet che possono includere la creazione di canali di distribuzione come magazzini, pagine Web e centri di spedizione dei prodotti. In particolare, i canali di distribuzione sono fondamentali per la vendita al dettaglio elettronica poiché permettono di portare il prodotto a casa del cliente (KamilTaylan.blog, economia

finanziaria, 2021). Diventare un E-tailer offre quindi ai rivenditori l'opportunità di ampliare la propria base di clienti aiutandoli a raggiungerne di più rispetto a prima (Raffaele Felaco, 2020).
La vendita al dettaglio online comprende un'ampia gamma di aziende e settori (KamilTaylan.blog, economia finanziaria, 2021) come quello della moda che è analizzato in questo lavoro di tesi. Tuttavia, ci sono somiglianze tra la maggior parte delle società di E-tailing poiché tutte includono un sito Web accattivante, una strategia di Marketing online, una distribuzione efficiente di prodotti o servizi e un'analisi dei dati dei clienti (KamilTaylan.blog, economia finanziaria, 2021). Un altro fattore di successo dell'E-tailer è sicuramente la possibilità di ridurre i costi in quanto non hanno bisogno di costruire negozi fisici e solitamente hanno meno lavoro di ufficio rispetto ai rivenditori tradizionali (Huang, H., & He, S., 2011, August), mentre i vantaggi per i consumatori sono sicuramente la possibilità di avere più alternative e fornitori disponibili in quanto i clienti possono essere raggiunti dagli E-tailer di tutto il mondo. Inoltre il confronto sul prezzo e le caratteristiche dei prodotti per i clienti è molto più facile in quanto è velocissimo passare da un sito Web all'altro dei vari rivenditori scegliendo l'articolo più conveniente. Va considerato, infatti, che il fattore comodità e risparmio di tempo è al terzo posto tra i motivi per cui le persone acquistano online (Huang, H., & He, S., 2011, August).
E infine, un ulteriore vantaggio, consiste nell' avere accesso a maggiori informazioni da parte dei consumatori: una delle principali caratteristiche di Internet è l'empowerment delle informazioni. Internet può fornire molte più informazioni al pubblico su prodotti che non è probabile o non facile da ottenere rispetto a qualsiasi altro media. Tali informazioni possono aumentare il potere contrattuale quando i clienti stanno negoziando con i rivenditori (Huang, H., & He, S., 2011, August).

Obiettivi della ricerca e il Gap nello stato dell'arte

Per gli E-tailer di moda sta diventando sempre più importante riuscire a catalogare velocemente i nuovi prodotti presenti sul loro sito online in quanto ogni articolo è provvisto di caratteristiche diverse che lo contraddistingue.
Questo lavoro di tesi ha permesso la realizzazione di un modello predittivo che affronta un problema di classificazione multiclasse, utilizzando diversi algoritmi di Machine Learning, in grado di assegnare correttamente le etichette delle diverse classi di borse dell'E-tailer di moda MyTheresa in base alle loro caratteristiche. Durante il processo di addestramento del modello, vengono utilizzate diverse tipologie di borse con le rispettive etichette di classe così da far imparare al modello a riconoscere le svariate caratteristiche che distinguono le diverse categorie di prodotto.
Il presente lavoro di tesi si pone come obiettivo quello di permettere ad un E-tailer, più precisamente un E-tailer del settore fashion, di non dover categorizzare manualmente ogni prodotto presente nel catalogo online ma di poter velocizzare questa attività aziendale mediante l'Intelligenza Artificiale così da poter utilizzare questo tempo risparmiato e le informazioni raccolte in altre attività aziendali.
La rilevanza di questa ricerca è inoltre data dalla sua scalabilità, ovvero dalla possibilità di essere estesa non solo alle borse ma anche a qualsiasi altro prodotto di ogni settore di mercato.

Questo lavoro di tesi, ha evidenziato un Gap nella letteratura riguardante l'applicazione dell'Intelligenza Artificiale, e quindi del Machine Learning, in maniera approfondita, nel settore degli E-tailers di borse e accessori in base agli attributi. Le applicazioni più rilevanti, sviluppate in campo fashion riguardano per lo più la categoria abbigliamento per la quale sono state utilizzate tecnologie di AI per problemi di classificazione multiclasse. Si sono rilevate comunque carenze di studio e di applicazioni in questo ambito specifico.
Data la grande necessità di innovare e stare sempre di più al passo con i cambiamenti nel settore della moda, e quindi di borse e accessori, si propone con questo lavoro di superare il Gap creando un modello predittivo che consenta di categorizzare le borse di un E-tailer in base alle loro caratteristiche, evidenziandone quelle più rilevanti, per dare la possibilità alle aziende non solo di velocizzare questi processi aziendali, ma anche per arricchirle di dati per offrire ai consumatori prodotti sempre più personalizzati e su misura per loro.

Stato dell'arte

In questo capitolo, dedicato alla Literature Review passata, si è andati ad esplorare l'evoluzione dei Big Data e il loro utilizzo da parte degli E-tailers, come l'utilizzo dei Big Data da parte di un E-tailer di abbigliamento per la gestione dei resi, l'impiego dei Big Data per aggiornare gli assortimenti degli E-tailer, nonché l'utilizzo di Big Data per monitorare i prezzi dei concorrenti online.

Il secondo paragrafo si addentra più nel dettaglio nel mondo dell'Intelligenza Artificiale mostrando quali sono state le sue applicazioni negli E-tailers. Vengono presi in esame due E-tailer di moda: il primo è l'E-tailer Stitch Fix, il quale ha deciso di adottare un sistema di Intelligenza Artificiale che utilizza una grande quantità di informazioni per selezionare una serie di cinque articoli da inserire in più in ogni spedizione. Queste informazioni sono fornite in gran parte dai clienti che rispondono a un questionario dettagliato sulle loro preferenze di stile, taglia e prezzo, oltre a immagini o altri dati non numerici su di loro (dalle pagine Pinterest e dai like dei clienti) (Roberto Grandinetti, 2020). Katrina Lake, fondatrice dell'E-tailer, definisce il modello commerciale che ha inventato come semplice: "Noi ti inviamo capi di abbigliamento e accessori che pensiamo ti piacciano; tu tieni i capi che vuoi e rispedisci gli altri" (Lake, 2018, p. 35); il secondo è L'E-tailer Trunk Club che innova il processo di acquisto dei consumatori fornendo servizi di personal styling basati sulle loro esigenze, sulla taglia, sul budget e sulle preferenze di stile, con l'aiuto di un team di esperti di styling dedicati, attraverso il suo sito web o i suoi negozi in tutti gli Stati Uniti (Modi, D., & Zhao, L., 2019). Nella selezione dei prodotti per i suoi clienti, l'azienda utilizza l'Intelligenza Artificiale adottando l'apprendimento automatico e gli algoritmi di raccomandazione personalizzati insieme a stilisti personali, che costituiscono un buon esempio di innovazione di processo sconvolgendo così il tradizionale modello di vendita al dettaglio (Reid 2012). E per concludere questo paragrafo, non si potevano non riportare come esempi due E-tailers famosi in tutto il mondo come Amazon e Netflix.

Nel terzo paragrafo sono stati riportati alcuni lavori di ricerca che hanno utilizzato modelli di classificazione di Machine Learning nel mondo della moda e in altri ambiti di applicazione. Nel mondo della moda è stata riportata una collaborazione tra storici della moda e informatici per esaminare la tendenza di un particolare articolo di moda, ovvero il cappellino sportivo casual (il berretto da baseball), sulle passerelle di moda mai visto prima dal 2008. E' stato riportato poi un altro studio che utilizza tecniche di Machine Learning per sviluppare un modello predittivo, applicato alla categoria dei prodotti di abbigliamento, che possa valutare facilmente e rapidamente la sostenibilità ambientale dei prodotti durante il loro ciclo di vita in base alle caratteristiche. Poi ancora due studi che adottano diverse tecniche di apprendimento supervisionato per classificare differenti tipologie di prodotti presenti in un catalogo di un E-commerce.
In ambiti diversi dal settore moda, si è esplorato il mondo musicale in cui si è voluto riportare un lavoro di ricerca riguardante l'utilizzo dell'apprendimento automatico per categorizzare brani musicali in base all'umore (ovvero felice/ non felice e arrabbiato/non arrabbiato) combinando le informazioni del testo e quelle dell'audio in base a caratteristiche di diverso tipo come quelle timbriche, ritmiche, tonali e descrittori temporali. Anche in campo agroalimentare sono stati utilizzati modelli di classificazione di Machine Learning e difatti si è analizzato che l'elaborazione e la classificazione combinata di immagini ottenute dalla spettroscopia del dominio del tempo dei Terahertz e l'utilizzo di algoritmi di apprendimento automatico possono essere utilizzati per categorizzare i diversi stati di maturazione dell'uva spina. In fine, in ambito politico, grazie all'utilizzo di tecniche di apprendimento supervisionato, un'istituzione governativa può eseguire la scansione di tutta la corrispondenza in arrivo (documenti civili, legali e governativi in ambito ministeriale) e assegnarla automaticamente ai dipartimenti appropriati (Ciecierski, K. A., & Kamola, M., 2020) potendo rilevare spam, tentativi di phishing o messaggi falsi.

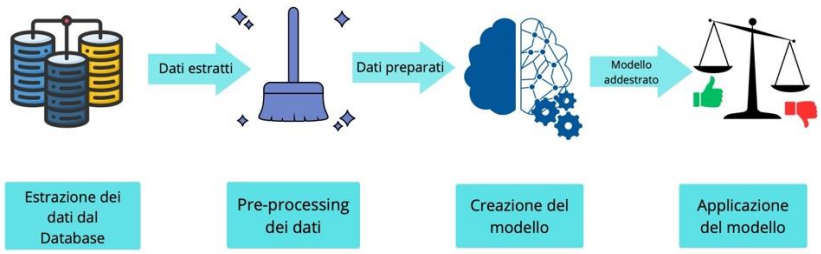
Materiali e metodi

Il lavoro di ricerca è stato svolto su Anaconda, una distribuzione dei linguaggi di programmazione Python e R per il calcolo scientifico (scienza dei dati, applicazioni di apprendimento automatico, elaborazione dati su larga scala, analisi predittiva, ecc.), che mira a semplificare l'installazione e la gestione delle librerie necessarie per l'analisi dei dati. Contiene un insieme di strumenti preinstallati, tra cui l'ambiente di sviluppo integrato (IDE) Spyder su cui sono state svolte le analisi del lavoro di ricerca.

Il Dataset utilizzato per questo lavoro di ricerca proviene dall'E-tailer di moda Mytheresa.
Mytheresa è un E-tailer di moda di lusso specializzato nella vendita di collezioni firmate di moda donna, uomo, bambino/a e articoli lifestyle. L'Etailer vanta una consolidata esperienza nel mondo della moda, frutto di oltre 30 anni di successi. Tutto è iniziato in una boutique nel centro di Monaco di Baviera specializzata in collezioni e designer internazionali che è diventata uno dei più innovativi e lussuosi E-tailer al mondo (Mytheresa.com,

2023). La loro accurata selezione permette di scegliere fra oltre 200 brand internazionali del lusso tra cui Bottega Veneta, Givenchy e Burberry e offre una piattaforma online semplice e intuitiva per un'esperienza di shopping proprio come in boutique che ti permette ogni giorno di essere sempre al passo con gli ultimi trend del mondo della moda (Mytheresa.com, 2023).

Descrizione del Modello Predittivo



5) Estrazione dei dati dal Database

In questa prima fase vengono individuati i dati di interesse, per poi effettuare l'estrazione dei dati da un database, file csv... e importare il dataset dell'E-tailer di Mytheresa su Spyder.

2) Pre-processing dei dati

Nella seconda fase i dati estratti vengono così puliti e pre-processati in quanto capita spesso che siano mancanti, incoerenti e rumorosi a causa della loro origine eterogenea.

I valori mancanti vengono sostituiti con la media nel caso di features numeriche, mentre con la moda (ovvero il valore più frequente) nel caso di features categoriche per poi procedere con l'eliminazione dei valori duplicati. Si andrà poi a fare il bilanciamento del dataset nel caso ci siano variabili sbilanciate.

In questo lavoro di tesi, al fine di effettuare un migliore confronto dei risultati ottenuti, sono stati attuati due bilanciamenti diversi del dataset, ovvero il Bilanciamento della variabile Y e il Bilanciamento del Train set.

È stato effettuato il Bilanciamento della variabile Y in quanto mostra le performance del modello a livello ideale e teorico, ovvero permette di dire quali sarebbero stati i risultati se fosse stato un caso puramente concettuale, mentre il Bilanciamento del Train set è stato fatto perché consente di costruire un modello predittivo rappresentativo del problema reale che si vuole risolvere, riflettendo la natura dei dati che effettivamente si incontreranno nel mondo reale. L'obiettivo principale dei dati di Train è quello di fornire al modello di Machine Learning esempi realistici che gli consentano di imparare e generalizzare correttamente su nuovi dati quindi tra i due diversi bilanciamenti i risultati più corretti e rilevanti. Per questo lavoro di tesi sono naturalmente quelli derivanti dal Bilanciamento del Train.

I due diversi bilanciamenti richiedono due diverse procedure per lo split del Dataset in Train set e Test Set. Nel Bilanciamento del Train si farà subito dopo aver convertito tutte le features da categoriche in numeriche così una volta splittato il dataset si potrà bilanciare solo la parte di Train, mentre nel caso del Bilanciamento

della Y lo split del dataset viene effettuato dopo che si è eseguito tutto il processo di Data-cleaning e quindi dopo che è stato effettuato il bilanciamento.
Il dataset in questo lavoro di tesi è stato diviso per l'80% in Train set e per il 20% in Test set:

c) Train Set: parte del dataset che è utilizzata per addestrare un algoritmo di Machine Learning
d) Test Set: parte del dataset utilizzata per verificare se il modello utilizzato può funzionare bene nel riconoscere nuovi dati, e quindi di conseguenza la precisione dell'algoritmo.

Sono state fatte analisi di statistica descrittiva che, non sono solo molto importanti perché permettono di vedere nel dettaglio come sono distribuite le variabili, ma soprattutto per capire se ci sono degli outlier, ovvero dei valori anomali, da dover rimuovere dalla variabile che si sta analizzando. In questo lavoro è stato presentato un confronto tra i grafici di distribuzione e del boxplot dei due diversi bilanciamenti effettuati per ogni variabile e quindi questo serve per vedere se da un bilanciamento all'altro ci sono stati dei risultati diversi per il grafico della distribuzione e del boxplot.
Successivamente è stata effettuata un'analisi della matrice di correlazione che ci permette di capire quali sono state le variabili più correlate tra loro e quelle maggiormente correlate con la variabile Y mostrando se ci sono stati miglioramenti o peggioramenti nei valori degli indici di correlazione tra le variabili prima e dopo aver effettuato il Data-cleaning. L'analisi della matrice di correlazione è molto importante per selezionare le features che verranno scelte per creare il modello predittivo. Vengono infine selezionati ed estratti gli attributi utili ai fini della costruzione del modello sulla base di tutte queste analisi utilizzando le tecniche di feature selection. L'output risultante, è un dataset pronto per la prossima fase, ovvero per essere utilizzato nella fase di creazione del modello.

3) Creazione del modello

Nella terza fase viene generato e addestrato il modello predittivo utilizzando i classificatori di Machine Learning. Per la creazione e l'addestramento del modello predittivo si sono utilizzati 6 algoritmi di Machine Learning, ovvero il Random Forest, l'XG-Boost, il KNN, il Decision Tree, la Logistic Regression e il Naïve Bayes.

4) Applicazione del modello

Nell'ultima fase, il modello addestrato viene testato sulle features sconosciute e per ciascuno di essi il modello fornisce la sua predizione.
Successivamente si valuta la precisione del modello attraverso il calcolo delle metriche di Accuracy, Precision, Recall e F1 score. Vengono, poi, rappresentati i grafici della Confusion Matrix e della curva ROC (in cui si andrà a valutare il valore AUC, ovvero "Area Under the Curve". E infine, saranno mostrate le features più rilevanti per ogni algoritmo.

Risultati e Conclusioni

Bilanciamento Train

Algoritmo	Accuracy	Precision (Weighted Average)	Recall (Weighted Average)	F-1 Score (Weighted Average)
Random Forest	0.80	0.80	0.80	0.79
XG-Boost	0.80	0.80	0.80	0.80
KNN	0.40	0.48	0.40	0.42
Decision Tree	0.47	0.68	0.47	0.52
Logistic Regression	0.33	0.59	0.33	0.36
Naïve Bayas	0.19	0.49	0.19	0.21

Bilanciamento Y

Algoritmo	Accuracy	Precision (Weighted Average)	Recall (Weighted Average)	F-1 Score (Weighted Average)
Random Forest	0.97	0.98	0.97	0.97
XG-Boost	0.98	0.98	0.98	0.98
KNN	0.87	0.86	0.87	0.86
Decision Tree	0.66	0.70	0.66	0.66
Logistic Regression	0.54	0.53	0.54	0.53
Naïve Bayas	0.36	0.38	0.36	0.35

Questo lavoro di tesi ha permesso la realizzazione di un modello predittivo che affronta un problema di classificazione multiclasse, utilizzando diversi algoritmi di Machine Learning, per poter assegnare correttamente le etichette delle diverse classi di borse dell'E-tailer Mytheresa in base alle loro caratteristiche.

Come possiamo vedere dalle tabelle qui sopra, i risultati migliori sono dati dall'algoritmo XG-Boost per entrambi i bilanciamenti, in quanto ha ottenuto risultati molto buoni su tutte le metriche di valutazione delle prestazioni (Accuracy, Precision, Recall e F1-score) pari all'80% per quanto riguarda il Bilanciamento del Train. I risultati ottenuti dal Random Forest in tutte le metriche utilizzate per valutare le performance sono stati molto vicini a quelli ottenuti dall' XG-Boost (sempre intorno all'80%) e quindi si può affermare che entrambi gli algoritmi sono risultati i migliori. Da entrambe le tabelle riportate, si può notare che i valori ottenuti per entrambi i bilanciamenti non sono uguali, anzi nel caso del bilanciamento del Train sono più bassi. Questo perché durante il processo di addestramento, il modello ha ottimizzato i suoi parametri per minimizzare l'errore tra le sue previsioni e le etichette corrette presenti nel Train set, poiché è necessario che contenga dati correttamente etichettati o annotati. In questo modo il modello può apprendere la corretta relazione tra le caratteristiche dei dati e le risposte desiderate. Sono stati riportati i risultati del Bilanciamento della Y perché mostrano le performance del modello a livello ideale e teorico, ovvero come sarebbero stati se si fosse trattato di un caso concettuale. Nel caso nell'algoritmo XG-Boost i risultati per tutte le metriche di Accuracy, Precision e Recall sono veramente molto alti pari al 98%, mentre per il Random Forest i risultati sono leggermente più bassi ma comunque buoni intorno al 97%. Però i risultati corretti e quelli su cui dobbiamo basare la nostra analisi sono naturalmente quelli derivanti dal Bilanciamento del Train, perché consente di costruire un modello predittivo rappresentativo del problema reale che si vuole risolvere riflettendo la natura dei dati che si incontreranno nel mondo reale. L'obiettivo principale dei dati di Train è quello di fornire al modello di Machine Learning esempi realistici che gli consentano di imparare e generalizzare correttamente su nuovi dati.

Procedendo con i risultati degli altri algoritmi, il KNN ha avuto dei buoni risultati solo nel caso del bilanciamento della Y in cui ha ottenuto una Precision e una Recall leggermente inferiori rispetto ai primi due algoritmi, pari, cioè, all'86% e all'87%, ma comunque buoni. Purtroppo non si può dire lo stesso per il bilanciamento del Train che presenta risultati molto bassi per tutte le metriche, cioè inferiori al 50%. Gli algoritmi Decision Tree, Logistic Regression e Naïve Bayes hanno ottenuto risultati significativamente inferiori rispetto ai classificatori descritti in precedenza. Ciò fa presupporre che questi algoritmi non siano molto performanti per questo problema di classificazione multiclasse.

In conclusione, le features più rilevanti, ovvero quelle che hanno contribuito di più alla previsione del modello sono (per gli algoritmi che hanno avuto i migliori risultati): la feature gender (ovvero genere uomo/donna) per l'algoritmo XG-Boost e la feature info_size_e_fit_0 (ovvero informazioni riguardanti taglie e misure delle borse) per l'algoritmo Random Forest. Attraverso l'analisi accurata delle caratteristiche, è possibile identificare gli attributi più importanti e che quindi ci danno più informazioni e che ci permettono di migliorare le prestazioni del modello al fine di ottenere una migliore comprensione del problema in esame. L'algoritmo XG-Boost ha identificato il genere (uomo/donna) come la feature più rilevante per distinguere le diverse tipologie di borse all'interno di questo problema di classificazione multiclasse, mentre per il Random Forest la feature più rilevante è quella che dà informazioni precise sulle taglie e le misure relative alle borse. Questo sta ad

indicare che l'algoritmo ha riconosciuto le informazioni riguardanti taglie e misure come importanti per categorizzare le diverse tipologie di borse e le utilizza per classificazioni più accurate.

Applicazioni nel Marketing

In un mercato virtuale sempre più affollato, diventa sempre più difficile per gli E-tailers creare un'esperienza di qualità per cercare di soddisfare e superare le aspettative dei clienti (Rose Sebastianelli & Nabil Tamimi, 2013). I consumatori, negli ultimi anni, sempre più frequentemente si affidano alle piattaforme di shopping online ed E-commerce per effettuare i loro acquisti proprio per il fatto che sono più convenienti, permettono un maggior risparmio di tempo e una grande varietà di scelta rispetto ai negozi fisici. Di conseguenza la quantità di informazioni disponibili online è aumentata drasticamente e i clienti sono spesso sopraffatti dalle scelte nel loro processo decisionale (Chong 2013). L'enorme quantità di dati a disposizione degli E-tailer possono riguardare, nel contesto specifico dei prodotti di moda, le preferenze dei consumatori sul colore, il tessuto, lo stile, la vestibilità e molti altri. Ciò consente agli E-tailer di personalizzare e ottimizzare la presentazione dei prodotti, i messaggi pubblicitari, le strategie di pricing in base a tali attributi e non solo.
Questa enorme quantità di informazioni raccolte sono fondamentali per comprendere meglio i clienti e offrire un'esperienza di acquisto online personalizzata. Se i dati vengono utilizzati in modo efficace per offrire prodotti pertinenti, raccomandazioni personalizzate e un'interfaccia utente-intuitiva, ciò può portare a una Customer Satisfaction positiva da parte dei clienti. Al contrario, se i dati non vengono sfruttati adeguatamente o in modo errato, ciò può comportare una Customer Satisfaction negativa nell'esperienza di acquisto online.
Utilizzando algoritmi di Intelligenza Artificiale, è possibile analizzare questi dati per identificare modelli, correlazioni e insight che possono aiutare gli E-tailer a migliorare le loro strategie di Marketing e personalizzare l'esperienza di acquisto online per i clienti in modo da indirizzarne il comportamento d'acquisto.

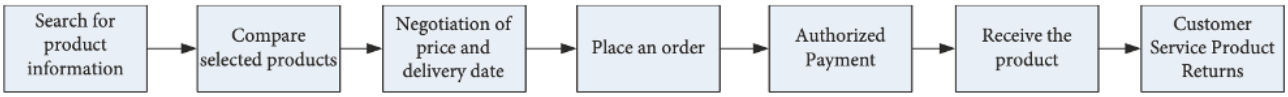


FIGURE : Basic flow chart of online shopping.

Fonte: Xiong, Y. (2022)

Il processo di acquisto di beni online, dal punto di vista del consumatore, passa solitamente attraverso diverse fasi: 1) Fase di consapevolezza di avere bisogno di un prodotto. L'acquirente riconosce un bisogno che può essere suscitato da uno stimolo interno o da uno esterno. I Marketing manager devono determinare i fattori e le situazioni che suscitano nel consumatore la percezione del problema e quindi come è possibile spingere il cliente a scegliere un determinato prodotto (Redazione, 2022). 2) Fase di ricerca delle informazioni. La quantità delle informazioni ricercate dal consumatore dipende dall'entità dello stimolo, dalla quantità delle informazioni iniziali, dalla semplicità di ottenere nuove informazioni. In questa seconda fase, l'acquirente può

ottenere informazioni da diverse fonti, come i motori di ricerca, le comunità online, articoli di blog, siti web, E-commerce, post e sponsorizzazioni sui social, recensioni, email ecc. (Redazione, 2022). Quindi è molto importante da parte dei Marketers capire quali sono i canali che gli utenti utilizzano in quanto il web è una fonte ricca di informazioni anche se molto diversi e variegati tra di loro. 3) Fase di valutazione delle alternative. Il consumatore utilizza le informazioni per giungere a restringere la propria scelta tra un numero limitato di marche per poi passare alla 4) Fase di acquisto del prodotto. Ciò comprende la circolazione delle informazioni e dei prodotti correlati, la negoziazione del prezzo e delle modalità di consegna tra acquirenti e venditori, la scelta del metodo e del termine di pagamento. In questa fase il consumatore acquisterà la marca preferita in assoluto e l'interazione principale, in questo momento del processo di acquisto, che ha il consumatore è con i siti web di E-commerce (Redazione, 2022). 5) Fase di esperienza post-acquisto. Questa comprende la fornitura di servizi post-vendita, la gestione dei reclami e la restituzione dei prodotti. Proprio qui comincia l'opera di fidelizzazione che comprende la richiesta di feedback, recensioni o l'invio di sondaggi per valutare la Customer Satisfaction (Redazione, 2022). La soddisfazione del cliente è uno dei fattori cruciali nella ricerca del comportamento dei consumatori negli ambienti commerciali tradizionali e online ed è generalmente considerata in diversi contesti comportamentali a causa del confronto tra le aspettative e l'esperienza del consumatore; questa, si raggiunge quando la consegna finale soddisfa o meglio supera le aspettative del cliente (Izyan, 2014). Nel loro studio, Sathiya et all. (2016), hanno rilevato che la soddisfazione degli acquirenti online era correlata alla zona di residenza, all'età, al sesso, all'occupazione e al costo dell'acquisto recente. Inoltre, in un ambiente online, la soddisfazione dei clienti è uno dei fattori chiave che portano a una maggiore fidelizzazione dei clienti e alla crescita a lungo termine dei negozi online (Urvashi, 2017). La maggior parte dei clienti è molto soddisfatta del prezzo del prodotto, seguito dallo sconto offerto e dal tempo impiegato per l'acquisto online (V. Mathan e R. Velmurugan, 2017; Nebojša, 2019).

Offerte e programmi di fidelizzazione sono ottime strategie di Marketing per rafforzare la Brand Loyalty, ovvero il rapporto di fedeltà alla marca da parte del cliente dopo la vendita. Se ogni fase del processo di acquisto è stata fatta correttamente, si genererà un processo di passaparola positivo sia online che offline capace di aumentare i clienti, implementare le vendite e rafforzare la Brand Awareness (Redazione, 2022).

Come illustrato nella Figura 2, le informazioni esterne sono necessarie per supportare le informazioni sul marchio o sul prodotto sia nella fase di pre-consumo che in quella di post-consumo (Xiong, Y., 2022). Ogni sito web fornisce informazioni diverse sul marchio o sul prodotto. Alcuni promuovono direttamente il marchio, altri vendono direttamente. Anche se il messaggio del marchio o del prodotto trasmesso in ogni fase varia, ogni fase richiede una pubblicità accurata e la giusta forza del messaggio per ottenere un mix mediatico perfetto (Xiong, Y., 2022).

Quando i consumatori sceglievano i metodi di acquisto tradizionali, la scelta dei rivenditori o dei negozi teneva solitamente conto di fattori quali la posizione geografica, la localizzazione del traffico del negozio, la circolazione della rete di vendita, il passaparola, la pubblicità del prodotto e quindi si preoccupavano principalmente della qualità del servizio clienti e del servizio post-vendita al momento dell'acquisto, della

comodità dell'ambiente di acquisto e dell'esplorazione del prodotto durante il processo di acquisto (Xiong, Y., 2022). La scelta degli acquirenti per lo shopping online, invece, si riflette principalmente nella scelta dei siti web commerciali quindi i fattori chiave da considerare sono le dimensioni del sito, l'adeguatezza delle informazioni sui prodotti fornite e la disponibilità di siti simili (Xiong, Y., 2022).

Un altro aspetto cruciale dell'applicazione dell'IA nel Marketing dei prodotti di moda è la capacità di personalizzare l'esperienza di acquisto per ogni singolo cliente. Utilizzando algoritmi di apprendimento automatico, l'IA può analizzare i dati sugli acquisti precedenti, le preferenze dichiarate dai consumatori e altri fattori per creare profili individuali dei clienti. Questi profili possono essere utilizzati per offrire raccomandazioni di prodotti personalizzati, suggerimenti di stile e promozioni mirate, aumentando così le probabilità di conversione e fedeltà del cliente. Le aziende che riescono ad adottare in modo efficace l'IA per l'analisi dei dati di Marketing avranno un vantaggio significativo nel mercato altamente competitivo dell'E-tailer soprattutto per migliorare e personalizzare l'esperienza d'acquisto dei consumatori.

Un aspetto importante, perciò, su cui gli E-tailers dovrebbero concentrarsi maggiormente è sicuramente la comprensione di come i clienti percepiscono e valutano la qualità delle esperienze di acquisto online, al fine di aumentare la fedeltà dei clienti. Zeithaml et all. (2002, p. 1) sostengono che "le aziende devono spostare l'attenzione dell'e-business dall'e-commerce (le transazioni) all'e-service (tutti gli spunti e gli incontri che avvengono prima, durante e dopo le transazioni)".

A questo punto possiamo definire l'esperienza d'acquisto per i clienti come la somma totale di sensazioni, emozioni e percezioni che un cliente sperimenta durante l'intero processo di acquisto. Questa esperienza può essere influenzata da diversi fattori, come l'interfaccia e l'usabilità del sito web o dell'applicazione di acquisto, la qualità dei prodotti o dei servizi offerti, il livello di assistenza e supporto al cliente, nonché la facilità e la convenienza del processo di acquisto. L'aumento della concorrenza esercita anche pressioni sulle aziende affinché attraggano i clienti e assicurino le vendite più rapidamente. Pertanto, l'analisi dei fattori che potrebbero influenzare le vendite dei prodotti sulle piattaforme online è cruciale per il successo aziendale e, di conseguenza, molte di queste variabili sono state ampiamente studiate in letteratura. Molti lavori di ricerca hanno approfondito lo studio sulla qualità nell'area del commercio elettronico (e-quality) cercando di identificare e valutare gli attributi (o dimensioni) che influenzano la percezione dei consumatori online sulla qualità dell'E-tailer (e-quality) e della soddisfazione dell'esperienza di acquisto per lo shopping online (Rose Sebastianelli & Nabil Tamimi, 2013).

Alcuni studi hanno generalmente enfatizzato la concettualizzazione della e-quality utilizzando quadri multidimensionali, e un certo numero di ricercatori ha affrontato la questione empiricamente sviluppando scale multi-item per misurare i costrutti della e-quality (Rose Sebastianelli & Nabil Tamimi, 2013).

Rose Sebastianelli & Nabil Tamimi (2013), nel loro lavoro di ricerca dal titolo "An Examination of Attributes Affecting Consumers' Perceptions of E-tailer Quality", hanno condotto una Conjoint Analysis, una metodologia utilizzata da diverso tempo negli studi di Marketing (ad esempio, Green e Rao 1971), ma relativamente nuova per il commercio elettronico (ad esempio, Schaupp e Belanger 2005; Chen, Hsu e Lin

2010), per ricavare l'importanza relativa degli attributi nelle preferenze dei consumatori per prodotti e/o servizi, fornendo i mezzi per determinare il valore relativo di specifici attributi dell'e-tailing per le percezioni dei clienti online sulla qualità dell'e-tailer (Tamimi, N., & Sebastianelli, R., 2016).
In questo modo, l'importanza relativa di ciascun attributo può essere individuata in uno scenario decisionale più realistico (poiché quando viene chiesto di valutare direttamente l'importanza degli attributi, gli intervistati tendono a valutarli tutti come importanti). I ricercatori presentano i risultati di uno studio sperimentale in cui i partecipanti forniscono giudizi complessivi (classifiche) sulla qualità del rivenditore elettronico descritta in termini di cinque attributi (reputazione del rivenditore, usabilità del sito, sicurezza, consegna e assistenza clienti) che si sono rivelati salienti in ricerche precedenti (Rose Sebastianelli & Nabil Tamimi, 2013). Queste classifiche servono come base per la stima dei modelli Conjoint a livello individuale, da cui si ricava l'importanza relativa degli attributi. I ricercatori utilizzano l'analisi dei cluster, basata sull'importanza relativa di questi attributi, per raggruppare gli individui in diversi segmenti (Tamimi, N., & Sebastianelli, R., 2016). Forniscono un profilo di ciascun segmento, non solo in termini di percezione di come questi attributi influenzano la qualità dell'E-tailer, ma anche in termini di caratteristiche demografiche e comportamentali del cliente (Rose Sebastianelli & Nabil Tamimi, 2013). Coerentemente con i risultati della Conjoint Analysis, la voce "sicurezza delle transazioni" ha ricevuto la valutazione media più alta. Tuttavia, anche altri tre elementi, ossia la reputazione del rivenditore online, la puntualità nella consegna dell'ordine e la facilità d'uso dell'interfaccia web, hanno ricevuto valutazioni molto alte (Tamimi, N., & Sebastianelli, R., 2016). Ciò rafforza l'idea che quando agli intervistati viene chiesto di valutare direttamente l'importanza dei singoli attributi del rivenditore online essi tendono a considerarli tutti importanti. Inoltre, ciò supporta l'uso della Conjoint Analysis per accertare il valore relativo degli attributi dell'E-tailing, ossia per rappresentare una visione più realistica del modo in cui i clienti online effettuano il trade-off tra tali attributi quando valutano la qualità dell'E-tailer (Tamimi, N., & Sebastianelli, R., 2016). Anche Keen et all. (2004) hanno utilizzato questo approccio per comprendere le decisioni di acquisto dei consumatori di fronte a formati di vendita al dettaglio alternativi, ovvero negozio, catalogo e Internet. Hanno scoperto che i due attributi più importanti che guidano l'esperienza d'acquisto erano il formato di vendita al dettaglio e il prezzo del prodotto.

La qualità dell'esperienza d'acquisto dei clienti online, come abbiamo precedentemente visto, è un fattore molto importante nel processo di acquisto e questo può influenzare anche l'intenzione di rivisitazione di un determinato sito Web e quindi di conseguenza l'intenzione di riacquisto da parte di un acquirente sul sito online. Molti studi precedenti hanno riportato l'impatto dell'analisi dei Big Data sui valori e sulle sfide aziendali (Akter, 2016; Barton e Court, 2012). Tuttavia, non sono state condotte ricerche sufficienti sul punto di vista dei clienti per esaminare come questi reagiscono all'applicazione dei Big Data Analytics sull'intenzione di riacquisto online. Pertanto, la ricerca sull'intenzione di riacquisto online dei clienti basata sull'applicazione dei Big Data Analytics, sta diventando una tendenza avanzata nella strategia di Marketing (Le e Liaw, 2017). Inoltre, anche l'effetto di mediazione tra la soddisfazione del cliente basata sui fattori di Big Data Analytics e

l'intenzione di riacquisto è stato studiato in modo limitato (Hui, S. C. S., Dastane, O., Johari, Z., & Roslee, M., 2021).

Molti studiosi come Lin (2014) hanno affermato che l'intenzione di riacquisto si ha quando un comportamento ripetuto diventa abituale come risultato di un processo cognitivo automatizzato. I clienti con personalità distinte possono reagire in modo diverso a benefici simili, determinando quindi differenze nel valore percepito, che possono in ultima istanza mostrare intenzioni di riacquisto diverse (Fang, 2016). Altri come Jobo (2016) sostengono che l'intenzione di riacquisto coinvolge l'esperienza generale dei clienti fino ad oggi, la loro soddisfazione per i prodotti o i servizi e la capacità di mantenere la soddisfazione dei clienti per incoraggiare la loro intenzione di riacquisto. Ed è per questo che, come afferma anche Quoc TP (2018), l'intenzione di riacquisto è uno degli obiettivi comportamentali di Marketing più significativi che assicurano che i clienti siano disposti ad acquistare nuovamente presso lo stesso negozio o venditore online. In effetti se un cliente ha avuto un'esperienza d'acquisto di alta qualità, che supera le aspettative e soddisfa le sue esigenze, è probabile che sviluppi un'intenzione di riacquisto positiva (Hui, S. C. S., Dastane, O., Johari, Z., & Roslee, M., 2021). C'è anche da dire che piattaforme di shopping online incontrano maggiori difficoltà rispetto a un contesto off-line a causa dell'assenza di interazioni faccia a faccia con i distributori, dell'inaffidabilità dei dati di configurazione di Internet, della diffidenza, dei bassi costi di commutazione, dell'incertezza e della rapida diffusione attraverso il passaparola. (Donni, Dastane, Haba e Selvaraj, 2018). Ma se la qualità dell'esperienza d'acquisto è curata nei minimi dettagli, sicuramente questo creerà fiducia, fedeltà e una connessione emotiva con il marchio o il venditore, che a sua volta aumenterà l'intenzione di riacquisto.

In conclusione, sulla base dei risultati empirici identificati dalla letteratura sul Marketing su Internet, è emerso che anche la costruzione dell'immagine può influenzare positivamente la soddisfazione e l'intenzione di riacquisto (Arons 1961; Rich e Portis 1964; Higie. Feick e Price 1987; Dodd 1991). Come evidenziato da Alba et al. (1997), è il web design che fornirà all'E-tailer un vantaggio comparativo nell'ambiente di vendita al dettaglio. Se il design promuove la frequenza di navigazione e rivisitazione di un sito Web, sicuramente questi comportamenti contribuiranno ad incentivare l'intenzione di acquistare e riacquistare un prodotto (Hopkins, C. D. (2001).

Implicazioni manageriali e sviluppi futuri

L'obiettivo principale di questo lavoro di tesi consiste nel possibile utilizzo da parte degli E-tailer di un modello predittivo che affronta un problema di classificazione multiclasse utilizzando alcuni algoritmi di Machine Learning come il Random Forest e l'XG-Boost (ovvero gli algoritmi che hanno avuto i migliori risultati in questo problema di classificazione), che possa assegnare correttamente le etichette delle diverse classi di borse di un E-tailers in base alle loro caratteristiche. L'impiego dell'Intelligenza Artificiale permetterà alle aziende di accorciare il tempo impiegato e le risorse spese per fare questo lavoro manualmente così da poter utilizzare questo tempo e le informazioni raccolte in altre attività aziendali come:

8. Personalizzazione dei prodotti: le aziende possono utilizzare queste informazioni per personalizzare l'offerta di prodotti e servizi offrendo suggerimenti e raccomandazioni personalizzate in base ai gusti e alle preferenze individuali dei clienti. L'utilizzo dell'Intelligenza Artificiale potrebbe consentire agli E-tailer di analizzare anche come gli attributi dei prodotti di moda influenzano i comportamenti e le preferenze dei clienti che potrebbero essere utilizzati per creare modelli di raccomandazione personalizzati basati sui gusti e sulle preferenze degli acquirenti. Ciò potrebbe consentire alle aziende di offrire suggerimenti di borse più pertinenti e adatte ai singoli clienti.

9. Migliorare l'esperienza dell'utente: attraverso l'analisi dell'Intelligenza Artificiale, è possibile identificare quali attributi (ad esempio, colore, taglia, stile, materiale) sono più rilevanti per i clienti e ottimizzare l'organizzazione e la presentazione dei prodotti sulla piattaforma di E-commerce in modo da poter migliorare l'esperienza di shopping online.

10. Ottimizzazione delle strategie di Marketing: l'intelligenza artificiale può aiutare le aziende a identificare quali attributi dei prodotti di moda sono più attraenti per i clienti e come posizionarli efficacemente nelle strategie di Marketing. Queste informazioni consentono di sviluppare campagne pubblicitarie mirate e creare contenuti ad hoc per i consumatori target che sono più inclini ad essere interessati a determinati tipi di borse e quindi presentare loro annunci pertinenti in modo tale da aumentare il loro coinvolgimento, ovvero la Customer Engagement.
11. Previsione della domanda e gestione dell'inventario: con una migliore comprensione di come gli attributi per specifici prodotti di moda influenzano la domanda dei clienti, le aziende possono ottimizzare la gestione dell'inventario regolando i livelli di stock, evitando sovra stock o scorte insufficienti e migliorare così l'efficienza nella gestione dell'inventario.

12. Innovazione del prodotto: l'analisi degli attributi dei prodotti di moda può rivelare nuove tendenze o combinazioni di attributi che sono particolarmente attraenti per i clienti. Queste informazioni possono guidare le aziende nella progettazione e nello sviluppo di nuovi prodotti innovativi, rispondendo alle esigenze e alle preferenze dei clienti in modo più accurato e tempestivo.

13. Analisi di tendenze e stili: gli algoritmi di Machine Learning potrebbero aiutare le aziende a identificare nuove tendenze di stile, comprendere meglio i gusti dei consumatori e sviluppare borse che soddisfino le esigenze del mercato in evoluzione tramite la raccolta di grandi quantità di dati relativi alle borse come i dati di vendita, recensioni dei clienti, influenze dei social media e tendenze di moda.

14. Analisi della concorrenza: le aziende possono monitorare e analizzare i dati relativi alle borse dei concorrenti come tendenze di mercato, strategie di prezzo e caratteristiche di prodotto di successo della concorrenza. Queste informazioni consentono loro di adattare le proprie strategie di marketing e differenziarsi sul mercato.

Questo modello predittivo può essere implementato per categorizzare non solo le borse, ma anche altre categorie di prodotti moda come abbigliamento, scarpe e accessori. L'applicazione di tali algoritmi potrebbe essere estesa anche a prodotti commercializzati da E-tailer che operano in settori diversi da quello della moda e questo permetterebbe alle aziende di sfruttare le potenzialità dell'apprendimento automatico per una varietà più ampia di prodotti, migliorando ulteriormente l'efficienza e la velocità delle operazioni aziendali.
E-tailers che operano in settori differenti sicuramente potrebbero utilizzare altri algoritmi che non sono stati implementati in questo lavoro di tesi come ad esempio l'SVM o Reti neurali artificiali in quanto non esiste un algoritmo unico che funzioni meglio su tutti i problemi di apprendimento supervisionato ma ogni problema deve essere trattato in maniera differente, quindi sicuramente gli algoritmi più performanti in questo lavoro di tesi sicuramente non saranno gli stessi che hanno avuto i migliori risultati per altri problemi di classificazione.
Un'ultima implicazione manageriale potrebbe essere cercare di colmare l'enorme divario tra le varie catene del valore dell'industria della moda come il design, la produzione e il Marketing. Questo perché dalla letteratura è emerso che i dataset di moda disponibili o sono troppo piccoli, o provengono da un'unica fonte di dati, o sono stati creati su misura per un compito specifico, o coprono un breve periodo di tempo. Manca un buon set di dati di riferimento per addestrare, testare, valutare e confrontare le prestazioni dei diversi algoritmi per l'analisi della moda. Pertanto, sarebbe estremamente utile per i ricercatori se esistesse un dataset di moda unificato su larga scala, che contenga dati di diverse modalità e che copra un lungo periodo di tempo (Fashion analysis and understanding with artificial intelligence) per raffinare la loro capacità di categorizzazione delle borse.

Come prospettiva futura a questo lavoro di tesi, si potrebbero migliorare i risultati ottenuti modificando alcuni parametri utilizzati per i diversi algoritmi di classificazione o implementando modelli più sofisticati e complessi che offrano risultati di classificazione ancora più accurati, riducendo sempre di più la necessità di un intervento manuale.
Si potrebbero anche utilizzare dati non strutturati in quanto attualmente l'impiego di algoritmi di Machine Learning per categorizzare le borse si basa principalmente su attributi strutturati come dimensioni, materiali e colori. In un futuro si potrebbe, quindi, mirare a integrare algoritmi di apprendimento automatico in grado di elaborare dati non strutturati come immagini o descrizioni testuali delle borse, consentendo una categorizzazione ancora più dettagliata e precisa.
Una prospettiva futura potrebbe anche essere quella di combinare questo modello di Machine Learning con l'utilizzo di modelli di Deep Learning come reti neurali convoluzionali (CNN) e reti neurali ricorrenti (RNN)

in quanto offrono potenzialità avanzate nella comprensione di immagini e testi per poter migliorare ulteriormente la precisione e l'efficienza della categorizzazione delle handbags.

In conclusione si potrebbe utilizzare un dataset più ricco di attributi e informazioni, più dettagliato rispetto a quello che è stato esaminato in questo lavoro di tesi in modo da poter migliorare ancora di più la categorizzazione delle borse e si potrebbe estendere il lavoro di ricerca non limitandolo alla sola Nazione Italia ma anche a tutte le altre nazioni in cui opera l'Etailer Mytheresa ovvero Giappone, Korea del Sud e Stati Uniti d'America.