Università Politecnica delle Marche

Facoltà di Ingegneria

Dipartimento di Ingegneria dell'Informazione

Corso di Laurea Magistrale in Ingegneria Informatica e dell'Automazione

---

Tesi di Laurea

Applicazione di tecniche di Data Science per l'analisi dell'efficacia delle promozioni relative ad un'azienda produttrice di articoli per la cura della persona

Application of Data Science techniques for the analysis of the effectiveness of the promotions for a company producing personal care items

Relatore

Prof. Domenico Ursino

Candidato

Bogdan Petru Borc

Anno Accademico 2020-2021

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

Negli ultimi decenni i dati hanno assunto un'importanza sempre maggiore nella vita di un'azienda e, soprattutto, rappresentano un vantaggio competitivo, se sono ben utilizzati. Per fare ciò, essi devono essere analizzati e studiati a fondo. L'analisi dei dati è un processo di ispezione, pulizia, trasformazione e modellazione di dati con il fine di evidenziare informazioni utili che supportino le decisioni strategiche aziendali. Se le decisioni aziendali vengono prese senza l'utilizzo della conoscenza dei dati, non parliamo di una *decisione strategica* ma di una semplice *ipotesi*.

Le aziende sono in possesso di una grande mole di dati; esse, spesso, non riescono ad interpretarli e sfruttare l'enorme potenziale che essi hanno. Per capire questo concetto, riportiamo una frase del matematico francese, Henri Poincaré, ovvero: "la scienza è fatta di dati come una casa è fatta di pietre. Ma un ammasso di dati non è scienza più di quanto un mucchio di pietre sia una vera casa". Questa frase ribadisce il fatto che, nonostante un'azienda sia in possesso di tantissimi dati, non necessariamente questi portano informazioni utili all'azienda. Tuttavia, attraverso una fase accurata di analisi, possono portare l'azienda stessa al successo.

I dati hanno un potenziale informativo enorme, che può aiutare le imprese sia a conoscere meglio se stesse (utilizzando i dati interni) sia il proprio mercato (utilizzando dati esterni), ma soprattutto i propri clienti. Secondo Peter Drucker, considerato il padre del management moderno, lo scopo di un'azienda è quello di creare e mantenere stretto un cliente attraverso il marketing e l'innovazione. È molto importante soddisfare il cliente con i propri prodotti/servizi; però, la maggior parte delle volte questo è molto difficile a causa della concorrenza. Per questo motivo, nell'ultimo periodo, sono tante le imprese che hanno iniziato a considerare di fondamentale importanza la *customer analytics*, per conoscere i propri clienti ed ottenere risultati migliori. Grazie a questa attività si riescono a capire i comportamenti dei propri clienti e le preferenze di questi ultimi al fine di prendere decisioni commerciali strategiche e tattiche.

In generale si può dire che tutti i dati sono importanti per il business, perché la maggior parte aiuta a scoprire e analizzare le aspettative dei clienti e le opportunità di crescita più efficaci. I dati utilizzati per eseguire questo tipo di analisi, la maggior parte delle volte, possono provenire da diversi reparti di un'azienda e riguardano le vendite, gli acquisti, i prodotti, ma anche le promozioni realizzate dall'azienda stessa.

Queste ultime svolgono un ruolo fondamentale nel soddisfare e attirare nuovi clienti, in quanto stimolano la richiesta del prodotto/servizio che, all'occhio del cliente, deve sembrare più attraente e innovativo. In base agli obiettivi dell'azienda, le strategie promozionali possono essere di tipo *pull* o *push*. Le prime cercano di far sì che i clienti "tirino" i prodotti dall'azienda, attraverso, per esempio, l'utilizzo di sconti. Le seconde cercano di spingere il prodotto dall'azienda verso distributori e rivenditori. Le prime permettono di costruire un legame più forte con il cliente, in quanto è a stretto contatto con questo.

In questo elaborato sono stati analizzati i dati relativi alle promozioni e alle vendite dell'azienda Fater. In una prima fase viene effettuata una pulizia dei dati; successivamente, viene eseguita un'esplorazione di questi per comprenderli meglio. Ci sarà, anche, un'analisi descrittiva dei contenuti più interessanti all'interno dei dataset. L'obiettivo finale è quello di capire quanto le promozioni, realizzate dall'azienda, hanno influenzato le vendite; per fare ciò, viene eseguita una fase di clustering; successivamente, per riuscire a esprimere quanto un nuovo prodotto in promozione influenzerà la vendita, viene svolta una fase di classificazione.

La presente tesi è composta da sette capitoli strutturati come di seguito specificato:

- Nel Capitolo 1 saranno introdotti il mondo della customer analytics e, successivamente, il ruolo delle promozioni nelle campagne di marketing.
- Nel Capitolo 2 si presenteranno le attività relative alle campagne promozionali in Fater. In particolare, si introdurranno le due tipologie di sconti utilizzate e la gerarchia dei clienti.
- Nel Capitolo 3 verrà introdotto il sistema dell'azienda riguardante le promozioni; successivamente, saranno illustrati i set di dati relativi alle promozioni e alle vendite dell'azienda.
- Nel Capitolo 4 verranno analizzate le fasi di ETL e di EDA sui dataset delle promozioni e delle vendite e sarà effettuata un'analisi descrittiva dei dati.
- Nel Capitolo 5 verranno analizzate le attività di clustering e classificazione dei dataset.
- Nel Capitolo 6 saranno riportati i risultati ottenuti.
- Nel Capitolo 7 verranno tratte le conclusioni e verranno delineati alcuni possibili sviluppi futuri.

1

La customer analytics e la gestione delle promozioni

In questo primo capitolo sarà introdotto il mondo della customer analytics e, successivamente, il ruolo delle promozioni nelle campagne di marketing.

1.1 Introduzione alla customer analytics

Negli ultimi anni si parla molto di "customer analytics"; però, per capire come funziona, si deve introdurre il concetto generale di data analytics. Fare data analytics vuol dire identificare, procurare e preparare grandi quantità di dati, che possono essere grezzi e non strutturati, successivamente analizzarli per estrarre valore da essi e per dare supporto a chi prende le decisioni. Fino a poco tempo fa chi prendeva le decisioni si basava soltanto sull'esperienza; oggi, invece, sono tanti coloro che adoperano la data analytics per ottenere migliori risultati. Questi ultimi sono ottenuti perché le decisioni vengono prese in base ad uno studio approfondito dei dati a disposizione.

In particolare, esistono quattro tipologie di analisi, che si differenziano in base ai risultati che esse producono:

- *analisi descrittiva*: riesce a sintetizzare il contenuto dei dati e risponde a domande ai eventi già accaduti;
- *analisi diagnostica*: analizzando i dati spiega perché qualcosa sta accadendo e quale è la causa;
- *analisi predittiva*: prevede cosa accadrà in futuro, basandosi sullo studio dei dati del passato;
- *analisi prescrittiva*: prevede cosa accadrà, spiegando anche perché e fornendo raccomandazioni su ciò che dovrà essere fatto per favorire o contrastare la previsione.

Ovviamente, come mostra la Figura 1.1, si può capire che l'ultima tipologia di analisi è la migliore, grazie al valore che riesce a realizzare, ma anche quella più difficile da realizzare, perché richiede una grandissima quantità di dati ed ha bisogno di persone altamente qualificate per la sua realizzazione.

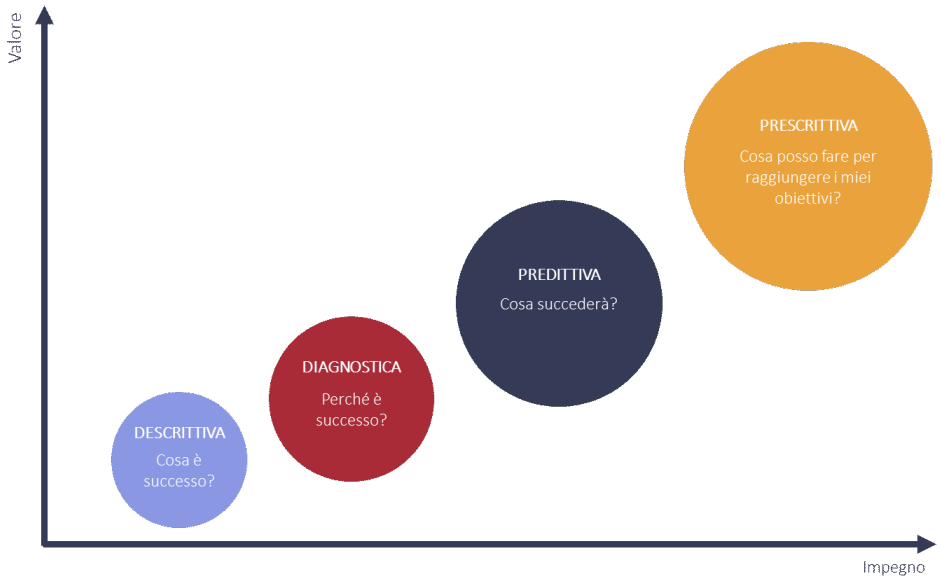

Figura 1.1. Grafico che mostra il valore che pare apportato dai vari tipi di analisi in funzione dell'impegno/ difficoltà nella sua realizzazione

1.1.1 Che cos'è la customer analytics

Secondo Peter Drucker, considerato il padre del management moderno, lo scopo di un'azienda è quello di creare e mantenere stretto un cliente attraverso il marketing e l'innovazione. È molto importante soddisfare il cliente con i propri prodotti/servizi; però, la maggior parte delle volte questo è molto difficile a causa della concorrenza. Per questo motivo sono tante le aziende che, nell'ultimo periodo, hanno iniziato a considerare di fondamentale importanza la customer analytics, per conoscere i propri clienti e ottenere ottimi risultati.

La customer analytics è un processo di raccolta e analisi dei dati dei clienti per imparare il comportamento e le preferenze di questi ultimi al fine di prendere decisioni commerciali strategiche e tattiche. Essa si basa su una serie di azioni e attività che devono essere eseguite al fine di giungere allo scopo finale, che è quello di comprendere i propri clienti. Le diverse attività da eseguire sono le seguenti:

- *Raccolta dei dati*: questa operazione è una delle più complicate, in quanto i dati riguardanti i clienti provengono da più fonti, e si deve riuscire a integrarli tutti insieme.
- *Individuare pattern attraverso l'utilizzo di metodi matematici*: grazie a determinati modelli si cerca di trasformare i dati, che possono essere sporchi, in informazioni più utili.
- *Individuare le intuizioni*: analizzando i dati, si possono capire le cause dei diversi comportamenti dei clienti.
- *Supporto alle decisioni*: sapendo il comportamento del cliente si possono cambiare le decisioni in modo opportuno.

- *Soddisfare il cliente*: si possono individuare i problemi dei clienti e si possono provare a risolverli in modo tale da rendere soddisfatti i clienti.

Anche se, a prima vista, queste attività possono sembrare abbastanza facili, sono molto complicate, in quanto si devono fare a livello del singolo cliente, e questo richiede l'accesso al livello più basso dei dati. Inoltre, si deve analizzare il comportamento del cliente nel tempo, e questo comporta la richiesta di tantissimi dati su di esso. Uno dei fattori che rende questa analisi molto complicata è dato proprio dal fatto che si devono analizzare gli atteggiamenti dei clienti, i quali influenzano le loro decisioni, e questo è molto complicato.

Per fare customer analytics le aziende posso utilizzare diverse tipologie di dati; questi si dividono principalmente in 4 categorie:

1. *Dati transazionali*: sono dati che descrivono eventi avvenuti all'interno di un'azienda. Si dividono in dati *finanziari, di lavoro e di logistica*. Quelli finanziari comprendono ordini, fatture e pagamenti, mentre quelli logistici contengono consegne, viaggi, etc. I dati relativi alle transazioni di lavoro contengono i diversi piani di lavoro all'interno dell'azienda. Ovviamente grazie a questa tipologia di dati, si analizzano le vendite e si riesce a stimare il potenziale di crescita e, soprattutto, ad ottimizzare le attività di marketing.
2. *Dati sull'utilizzo del prodotto/servizio*: analizzando i dati provenienti direttamente dai propri prodotti/servizi si può migliorare l'esperienza del cliente e, magari, apportare innovazione al prodotto.
3. *Dati del web*:vengono analizzati dati provenienti, per esempio, dal sito web di un'azienda. Analizzando ogni movimento del visitatore nel sito è possibile creare contenuti in evidenza in base alle preferenze del cliente.
4. *Dati dei clienti*: i clienti possono condividere il loro pensiero su un determinato prodotto/servizio sotto forma di recensione, o attraverso l'utilizzo dei social media. Analizzando tali contenuti si possono apportare miglioramenti per soddisfare i clienti.

Oltre a questi dati, le aziende hanno iniziato a integrare anche altri, provenienti dall'esterno, per avere una visione più completa del comportamento dei clienti.

1.1.2 Come fare customer analytics

Per fare customer analytics è necessario realizzare un sistema in grado di analizzare costantemente i dati, riuscire a ottenere dei pattern sul comportamento dei clienti e mostrare possibili modelli. Il sistema realizzato deve avere una struttura solida ma, allo stesso tempo, flessibile, cioè in grado di trasformarsi in base ai dati analizzati, in quanto i clienti possono cambiare le proprie abitudini. Per realizzare un sistema efficace dal punto di vista degli obiettivi dell'azienda si devono seguire i seguenti passaggi:

1. *Identificazione dell'obiettivo*: si deve comprendere a fondo l'obiettivo che si vuole raggiungere e, in base a questo, capire quali dati sono utili a tale scopo.
2. *Scelta delle metriche*: l'analisi dei clienti necessita di più tipologie di dati; per questo, nessuna metrica da sola è in grado di modellare l'analisi dei clienti. Ci sono tantissime metriche; le più utilizzate sono i ricavi, le transazioni, le visite

dei siti e l'usabilità del prodotto/servizio. In questa fase è molto importante riuscire a rappresentare tutte le metriche, magari utilizzando apposite dashboard e vedendo quali sono quelle ritenute più importanti per l'analisi.
3. *Analisi dei dati*: si ricercano dei pattern utili nei dati e, una volta trovati, ci si concentra su di essi, analizzandoli più in dettaglio.
4. *Elaborare modelli descrittivi o predittivi*: una volta che si hanno dei dati puliti e coerenti, si costruisce il modello, che può essere descrittivo o predittivo.
5. *Ottimizzare i modelli*: una volta realizzato il modello si deve verificare la sua solidità e, magari, capire se può essere migliorato in qualche modo.
6. *Mettere in pratica il modello*: durante questa fase si osserva se il funzionamento del modello è corretto. È molto importante monitorare il sistema in questa fase, per capire se l'output del modello è quello desiderato.

Grazie alla customer analytics vengono prese decisioni migliori in base ai dati, e questo, come si vedrà nella prossima sezione, può portare tantissimi vantaggi.

1.1.3 Benefici della customer analytics

In seguito saranno indicati i possibili benefici che possono essere ottenuti grazie all'utilizzo della customer analytics:

- *Campagne dirette*: è possibile raggiungere maggiore redditività, grazie a sforzi di marketing, che fanno in modo di ridurre i costi.
- *Prezzi competitivi*: è possibile stabilire i prezzi dei prodotti in base alla domanda e alle preferenze dei clienti.
- *Personalizzazione*: i clienti hanno la possibilità di scegliere il prodotto/servizio che li soddisfa di più.
- *Riduzione degli scarti*: anticipando le richieste dei clienti è possibile produrre il giusto e scartare di meno.
- *Consegna più veloce*: sapere in anticipo che prodotto si venderà, dove e quando, fa in modo che si conoscano i tempi giusti di produzione.
- *Clienti fedeli*: fornire il prodotto/servizio giusto al prezzo giusto aumenta la soddisfazione dei clienti, facendo in modo che diventino clienti fedeli, essenziali per la crescita a lungo termine.

1.2 Panoramica sulle promozioni

Una promozione svolge un ruolo fondamentale nel soddisfare e attirare nuovi clienti, in quanto stimola la richiesta del prodotto/servizio che, all'occhio del cliente, deve sembrare più attraente e innovativo. In seguito si vedrà cos'è una strategia promozionale, e i vari tipi di strategie che si possono condurre.

1.2.1 Che cos'è una strategia promozionale

Una strategia promozionale è definita dalle tattiche adottate nel piano di marketing per aumentare la domanda del prodotto/servizio all'interno di un'azienda. Quindi,

una promozione non è altro che una strategia in cui il prodotto/servizio viene promosso, usando iniziative attraenti a breve termine, per stimolarne la domanda ed aumentarne le vendite.

Una strategia promozionale è un modo per soddisfare gli obiettivi di vendita a breve termine di un'azienda, stimolando l'acquisto del prodotto. In base agli obiettivi, le promozioni possono essere indirizzate a diversi utenti. Più specificatamente è possibile distinguere le seguenti classi di promozioni:

- *Consumer promotion*: si tratta di una promozione rivolta direttamente al consumatore finale. L'obiettivo di tale promozione è quello di modificare il comportamento dell'acquirente sulle scelte, nella quantità e nella tempistica nel comprare il prodotto. Entrano in questo campo tutte quelle promozioni che fanno si che un prodotto sia scontato, in omaggio, ma anche, in determinati casi, il periodo di prova, che spinge il cliente ad avvicinarsi al prodotto ed acquistarlo.
- *Trade promotion*: sono quelle promozioni rivolte agli intermediari commerciali. L'obiettivo è quello di modificare il comportamento del distributore/rivenditore in modo che questo sostenga in modo attivo il prodotto. Un esempio è la possibilità di ottenere uno spazio all'interno di un supermercato dove poter esporre il proprio prodotto.
- *Sales force promotion*: sono le promozioni rivolte alla forza vendita. Attraverso premi e corsi di formazione il personale di vendita è contento e spinto a svolgere al meglio il proprio lavoro, aumentando i benefici.

1.2.2 Tipi di strategie promozionali

In base agli obiettivi e al tipo di marketing, le strategie promozionali possono essere divise in tre grandi categorie:

- *Strategie pull*, alla base dell'*inbound marketing*; cercano di far sì che i clienti "tirino" i prodotti dall'azienda. Implica l'uso di diverse strategie e iniziative verso i clienti, ad esempio gli sconti stagionali.
- *Strategie push*, alla base dell'*outbound marketing*; cercano di spingere il prodotto dall'azienda verso distributori e rivenditori. Questa strategia implica l'uso di tattiche sviluppate specialmente per commercianti, rivenditori, distributori ed agenti.
- *Strategia ibrida*: usa entrambe le strategie menzionate in precedenza per vendere il prodotto con la minor resistenza possibile. Si tratta di attirare i clienti utilizzando coupon speciali e anche di fornire incentivi ai commercianti per vendere i propri prodotti.

L'outbound marketing non permette ad un'azienda di costruire un legame forte con il consumatore finale in quanto non è a stretto contatto con quest'ultimo. Questa tipologia di marketing è fortemente guidata dalle vendite. Al contrario, come anticipato in precedenza, l'inbound marketing guarda direttamente il consumatore finale e fa in modo di trasformarlo in un cliente fedele.

In seguito sono indicati alcuni esempi di strategie dell'inbound marketing più utilizzate:

- *Content marketing*: è la branca del marketing che include l'insieme delle tecniche per creare e condividere contenuti digitali intorno ad un prodotto/servizio, in modo tale da orientare il cliente all'acquisto. Questi contenuti possono essere di diversi tipi, partendo da semplici testi, fino a immagini e video. In questo ramo è molto importante ricordare che i social media giocano un ruolo fondamentale per diffondere i propri contenuti, in quanto oltre 3,5 miliardi di persone ne utilizzano almeno uno. I più gettonati sono Facebook, Instagram e YouTube.
- *Email marketing*: si concentra sulla costruzione di un legame con il ricevitore delle e-mail, e di non "spammare" messaggi ad utenti se si sa che questi difficilmente ne sarebbero interessati. Una volta creata la rete di indirizzi e-mail delle persone che possono essere interessate al prodotto/servizio, basta fornire a queste ultime contenuti di qualità, preferibilmente mirati attraverso la segmentazione.
- *Referral marketing*: si basa sul fatto che le persone preferiscono fidarsi di un amico/conoscenza sulla qualità di un prodotto/servizio; essa è detta, anche, pubblicità del passaparola. Secondo gli esperti è una delle tecniche più forti per far conoscere il proprio prodotto, ma anche una delle più difficili da realizzare, soprattutto in fase iniziale; successivamente i clienti felici inizieranno a pubblicizzare il marchio senza aggiungere alcuno sforzo extra al processo.
- *Campioni gratuiti e coupon di vendita*: si basa sul fatto che il marchio è sicuro che chi prova il proprio prodotto, anche in maniera gratuita, successivamente lo acquisterà.
- *Offrire rimborsi*: i rimborsi rappresentano un modo di garantire ai clienti la qualità del prodotto. Questa tattica di marketing promozionale funziona come la psicologia inversa; cioè, sapendo che si possono avere indietro i soldi se non si è soddisfatti del prodotto, fa venire voglia al cliente di provarlo.

1.2.3 Compito delle promozioni

Qualsiasi campagna promozionale deve cercare di raggiungere uno o più dei seguenti obiettivi:

- *Creare consapevolezza*: la maggior parte delle volte le imprese falliscono perché le persone non le conoscono e non sanno neanche cosa fanno.
- *Far provare i prodotti ai consumatori*: la promozione è quasi sempre usata per far provare un nuovo prodotto o uno già esistente a chi non lo usa.
- *Fornire informazioni*: la promozione è volta a spiegare perché un determinato prodotto è migliore rispetto ad un altro della concorrenza, ad informare il cliente di un nuovo prezzo più basso, a spiegare dove poter acquistare il prodotto, etc.
- *Mantenere i clienti fedeli*: le promozioni vengono usate per evitare che i clienti cambino idea sul prodotto e ne acquistino uno di un'altra marca.
- *Aumentare la quantità e la frequenza d'uso del prodotto*: la promozione viene spesso usata per far sì che la gente usi di più un prodotto.
- *Identificare i clienti target*: la promozione aiuta a trovare nuovi clienti.

2

La gestione delle promozioni in Fater

Nel capitolo corrente si presenteranno le attività relative alle campagne promozionali in Fater. In particolare, si introdurranno le due tipologie di sconto utilizzate e la gerarchia dei clienti.

2.1 Introduzione al mondo Fater

In questo capitolo verrà presentato il contenuto relativo a Fater, in particolare le campagne promozionali da essa condotte e sulle quali, successivamente, concentreremo le nostre attività di data analytics.

Le origini di Fater risalgono al 1867, quando i fratelli Bucco fondano a Pescara l'azienda chimico-farmaceutica dall'omonimo cognome "F.lli Bucco", che verrà rilevata da Francesco Angelini nel 1958. Quest'ultimo la ribattezza Fater, acronimo di "Farmaceutici aterni", che ricorda il tipo di prodotto di cui si occupa la società e la sua origine (dal fiume Aterno, che si trova nella città di Pescara).

Dagli anni 60 ci sarà un progressivo sviluppo dell'azienda, dovuta alle fiorenti condizioni del mercato italiano ed all'ingegno di Iginio Angelini.

Dal 1992, anno in cui la Product & Gamble acquisisce il 50% delle quote della Fater, dando vita a una joint venture paritetica, l'azienda iniziò, grazie ai prodotti per bambini, ad avere un grande successo, soprattutto con i famosissimi Pampers.

Andando avanti con gli anni, nuovi prodotti iniziarono a far parte delle offerte di quest'azienda. Oggi la Fater produce e distribuisce in 39 paesi e le collaborazioni con imprese abruzzesi sono 180, con le quali collabora nella produzione e vendita dei prodotti dei marchi Pampers, Lines, ACE, Tampax, Linidor, Comet, Dignity e NEOBLANC, tutti sono in continua crescita.

Gli stabilimenti dell'azienda sono quattro, due in Italia (Pescara e Campochiaro), uno in Portogallo e l'ultimo in Turchia.

2.2 La campagna promozionale di Fater

L'azienda utilizza un sistema ERP (Enterprise Resource Planning), un tipo di software indispensabile per gestire un'azienda con queste dimensioni.

I sistemi ERP si basano su un'unica struttura di dati che condivide, in genere, ad un database comune. In questo modo si fa sì che le informazioni utilizzate in tutta l'azienda siano normalizzate e basate su definizioni comuni. Tali costrutti principali vengono, quindi, interconnessi con i processi aziendali guidati dai flussi di lavoro tra i vari reparti aziendali, i sistemi di connessione e le persone che li usano. In poche parole, l'ERP è il veicolo per l'integrazione di persone, processi e tecnologie in un'azienda moderna.

Per quanto riguarda le promozioni in Fater, queste sono gestite dal *sistema POP*, un sottosistema dell'ERP aziendale. All'interno del sistema POP si trovano tutti i dati che riguardano le campagne promozionali realizzate negli anni passati.

Il sistema è fatto in modo tale da permettere ai clienti, che gestiscono un *calendario promozionale*, di introdurre un evento promozionale, che raggruppa tutte le attività che saranno svolte nei loro punti vendita per incentivare le vendite. L'informazione riguardo tale evento può essere inserita con mesi di anticipo e il sistema permette che essa non sia completa; tuttavia fornisce la possibilità di completarla in seguito.

Le informazioni basilari che vengono riportate alla creazione di un evento promozionale sono:

- *Codice calendario*: è l'id del calendario dove si sta inserendo l'evento promozionale. Un calendario promozionale viene gestito da un responsabile e può avere una struttura gerarchica, anche se ciò non è indispensabile. In particolare, tale evento può essere nazionale, intermedio e locale. La gerarchia è data dal fatto che l'informazione presente nel calendario nazionale può essere ereditata da un calendario figlio intermedio (esiste un attributo che sottolinea la presenza di un calendario padre) e sia presente anche in questo (magari con più informazioni). Se un calendario locale ha un padre intermedio, eredita le promozioni di quest'ultimo alle quali può, comunque, applicare ulteriori promozioni.
- *Descrizione del tema*: spiega l'evento promozionale.
- *Id tema*: è un numero progressivo.
- *Data di inizio e fine sell in*: si intende il periodo in cui il cliente farà gli ordini all'azienda per usufruire degli sconti.
- *Data di inizio e fine sell out*: si intende il periodo in cui la merce sarà effettivamente esposta nei vari punti vendita.

Le altre informazioni riguardanti l'evento potranno essere completate in seguito, mano a mano che si hanno più dettagli riguardanti i clienti, l'occasione d'uso, il target, il tipo di promozione, fino ad arrivare ai prodotti effettivamente in promozione.

Tipi di promozione

Una volta definita l'occasione d'uso dal cliente responsabile, il sistema permette di scegliere tutti i particolari delle promozioni, soprattutto di indicare il tipo di promozione che può essere a Volantino, TPR (Taglio prezzo) o Promo web. Fatto ciò si può indicare se nel punto vendita ci sarà la presenza di hostess, isole, comarketing o casse display, e scegliere le referenze che parteciperanno alla promozione. La *referenza è un numero di 18 cifre* che identifica univocamente un prodotto; in tale

numero i primi due caratteri indicano il target, il terzo e il quarto carattere indicano l'occasione d'uso, il quinto, il sesto e il settimo denotano il segmento. Le restanti cifre indicano il prodotto vero e proprio.

Di recente, per non inserire manualmente tutte le referenze in un evento promozionale, sono stati realizzati dei *nodi promo* all'interno dei quali sono state inserite delle referenze che, di solito, vanno insieme in promozione. In questo modo tutt'al più vengono deselezionate le referenze che non sono d'interesse. i nodi promo, inoltre, consentono di velocizzare il processo di selezione delle referenze alla creazione di un evento. Ovviamente si ha ancora la possibilità di scegliere a mano la referenza voluta, senza l'utilizzo del nodo promo.

Campagna Promozionale

Ciò che è stato spiegato fino ad adesso è come un responsabile di un calendario è in grado di realizzare un evento promozionale scegliendo le referenze che andranno in promozione. Ovviamente dal lato di Fater si deve decidere quali referenze in un determinato periodo sono in sconto. Per questo, ogni semestre, il reparto vendite prepara un *menù promozionale*, che comprende più *campagne promozionali*, al fine d'incentivare le vendite dei prodotti dell'azienda presso i clienti. La campagne promozionali contengono gli sconti associati ai vari prodotti in un determinato periodo temporale.

Una volta che il cliente sceglie le referenze che saranno in promozione, questo può vedere quali di queste sono presenti nelle varie campagne promozionali attive e usufruire dello sconto, che è valido solo se l'ordine viene fatto dall'inizio del periodo di *sell in* fine alla fine di questo.

Tipologia di sconti

In Fater gli sconti possono essere di due tipi:

- *In fattura*: al momento del pagamento dell'ordine si procede ad applicare lo sconto.
- *Fuori fattura*: sono degli accordi extra concordati con il cliente, in cui l'azienda fornisce un pagamento con una *nota di credito* oppure con una *fattura promozionale*.

La prima tipologia di sconto è la più conosciuta, mentre la seconda riguarda il mondo Fater. Per comprendere lo sconto fuori fattura si deve introdurre il concetto dei due budget che l'azienda ha a disposizione per scopi diversi; questi sono:

- *Budget CMA*: è un budget messo disposizione per spingere le vendite su determinati clienti.
- *Budget CMO*: è un budget messo a disposizione per spingere la vendita dei prodotti a prescindere dai clienti.

Selezionato il budget, per decidere lo sconto fuori fattura ci sono delle *regole di calcolo* che devono essere applicate; queste sono:

- *Percentuale sul fatturato*: per esempio una piccola percentuale su tutto il fatturato (esempio 2%).

- *Valore per quantità*: si dà un premio, ad esempio per ogni cartone venduto.
- *Importo a carico* : 100 euro per ogni carico completo.
- *Importo manuale*: a prescindere da tutto, si dà una quantità in denaro fissa al cliente.

La differenza tra nota di credito e fattura promozionale sta nel fatto che quest'ultima tiene conto, nel calcolo del bonus, anche del tipo di servizio che si sta chiedendo al cliente, ovvero se è un evento a volantino, sottocosto, oppure, per esempio, se c'è la presenza anche di isole promozionali (spazio all'interno della punto vendita dedicato solo a prodotti Fater), hostess (persona fisica che organizza e pubblicizza l'evento), comarketing o casse display (articolo che contiene più prodotti anche diversi tra loro).

Gerarchia clienti

All'interno del sistema POP i clienti hanno una gerarchia; questa è riportata nella Tabella 2.1. Ciò vuol dire che, partendo dal basso (cliente) verso l'alto (canale), un *cliente* può appartenere ad un solo *Bussines partner*, ma questo può contenere più clienti, e cosi via fino al *canale*. È utile capire questa gerarchia, in quanto, quando si selezionano, per esempio, i clienti parteciperanno ad un evento promozionale, se si sceglie solo un determinato gruppo regionale, saranno automaticamente selezionati tutti i Business Partner insieme a tutti i clienti che ne fanno parte. Ovviamente, se lo si desidera, alcuni clienti possono essere deselezionati a scelta dell'organizzatore dell'evento.

Livello	Nome
1	Canale
2	Centrale Internazionale
3	Super Centrale
4	Centrale Nazionale
5	Gruppo Regionale
6	Business Partner
7	Cliente

Tabella 2.1. Gerarchia dei clienti

3

Dataset di riferimento

Nel capitolo corrente verrà introdotto il sistema POP e, successivamente, saranno illustrati i set di dati relativi alle promozioni e alle vendite dell'azienda.

3.1 Introduzione

In questo capitolo verranno introdotti i dati utilizzati nell'attività di data analytics.

Prima di avere direttamente i dati necessari per il lavoro, sono state necessarie alcune ore di interlocuzione con dei responsabili della Fater, che ci hanno spiegato in dettaglio il *sistema POP* dell'azienda. Ci hanno fornito un file *Excel* dove erano riportate le tabelle estratte dal sistema. Per avere un quadro migliore sul contenuto di questo file, è stata realizzata una rappresentazione grafica delle tabelle (mostrata in Figura 3.1), per riuscire a comprendere meglio i collegamenti tra i dati. Una volta conclusa la fase in cui ci hanno spiegato il funzionamento del *sistema POP*, ci sono stati forniti due set di dati in formato *.csv* che saranno analizzati successivamente. Il primo file conteneva i dati relativi al *sistema POP* dal 2020 in poi, mentre il secondo riguardava le vendite dell'azienda nello stesso periodo. Il secondo set di dati è stato necessario per le fasi successive di analisi, per vedere, se possibile, come le promozioni hanno influenzato le vendite. Nelle prossime sezioni analizziamo più in dettaglio il contenuto della Figura 3.1, spiegando quali sono i campi più interessanti e utili per le nostre analisi.

3.2 Sistema POP

Osservando la Figura 3.1 si può notare il numero elevato di tabelle; per comprendere meglio il contenuto, in Figura 3.2 sono state evidenziate le macro regioni che verranno analizzate in seguito più in dettaglio e che permettono di capire i vari collegamenti all'interno del sistema. Le regioni di interesse sono le seguenti:

- In blu è rappresentato tutto ciò che riguarda il cliente: oltre ai dati anagrafici, vengono indicate anche le insegne a lui associate.

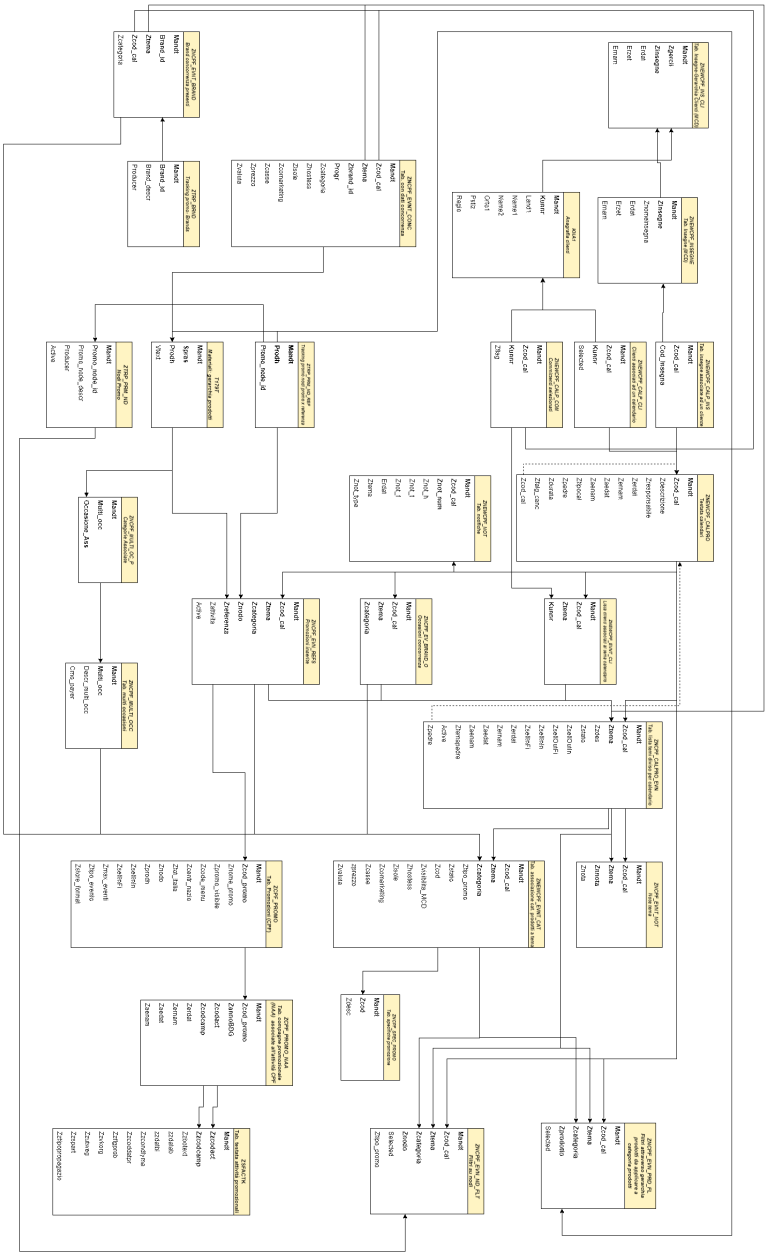

Figura 3.1. Collegamenti delle tabelle del sistema POP

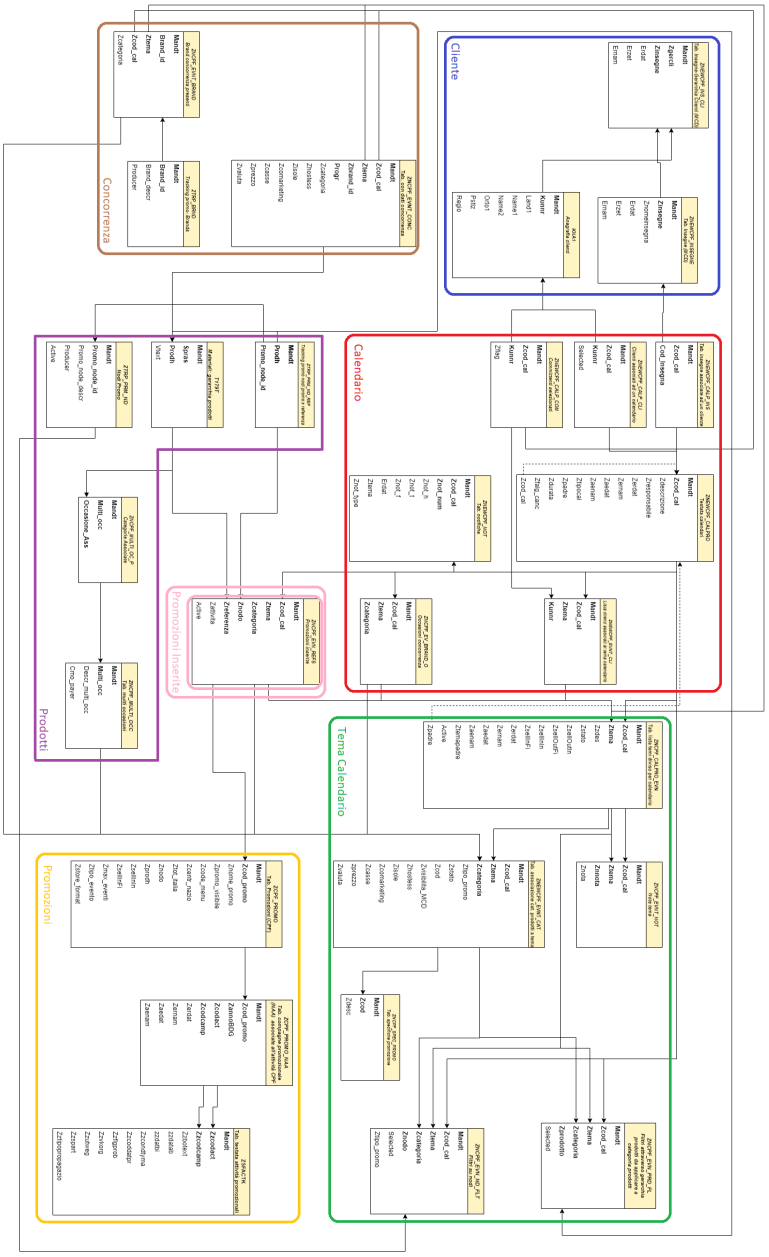

Figura 3.2. Regioni d'interesse del sistema POP

- In rosso viene indicato tutto ciò che riguarda un calendario; in particolare si può vedere quali sono i clienti ad esso associati e i temi. La tabella delle notifiche indica tutte le notifiche apportate al sistema.
- In verde è rappresentato tutto ciò che riguarda un tema di un calendario; in particolare, oltre all'informazione del tema stesso, vengono indicate le categorie di prodotti associate a questo.
- In giallo viene rappresentato tutto ciò che riguarda le promozioni dal lato del reparto vendite, ovvero quali sono i prodotti effettivamente in sconto in un determinato periodo.
- In viola è rappresentato tutto ciò che riguarda i prodotti, insieme ai nodi promo introdotti nel capitolo precedente.
- In marrone sono rappresentate le tabelle della concorrenza; esse, però, non saranno analizzate in quanto questo è un punto ancora in via di sviluppo all'interno del *sistema POP*.
- La tabella cerchiata in rosa, come si può notare, è stata evidenziata due volte, in quanto è il punto chiave di estrazione dei dati del sistema POP che saranno analizzati in seguito. In questa tabella ci sono le promozioni inserite: il responsabile del calendario indica quali sono le referenze che faranno parte di un determinato tema e, se queste appartengono ad una promozione interna, usufruiranno del corrispettivo sconto, altrimenti sono senza sconto.

Calendario

In Figura 3.3 è rappresentato tutto quello che riguarda direttamente un calendario promozionale. La tabella centrale in figura, i cui dettagli sono anche presentati nella Tabella 3.1, indica i campi creati all'inserimento di un nuovo calendario. Si può notare che la chiave è `MANDT e ZCOD_CAL`; il primo campo è un numero fisso che viene rappresentato in tutte le tabelle ed è dovuto al sistema ERP dell'azienda (in caso di aziende molto grandi, tale campo può cambiare nei vari reparti), mentre il secondo è il codice del calendario. Vengono riportate, anche, informazioni in merito al responsabile, al tipo e a quando è stato inserito il calendario. Tralasciando gli altri campi spiegati nella Tabella 3.1 è molto importante segnalare il campo `ZPADRE` in quanto questo permette di realizzare la gerarchia, spiegata nel capitolo precedente, tra i calendari.

La tabella delle notifiche, mostrata in figura 3.3, viene riportata nella Tabella 3.2 ed è importante perché riporta tutti gli aggiornamenti eseguiti all'interno del sistema. In particolare, preso un determinato calendario e un tema ad esso associato, grazie al campo `ZNOTTYPE` si può specificare il tipo di notifica.

Le altre tabelle nella figura indicano quali clienti, insieme alla insegne, vengono selezionati per un determinato tema all'interno di un calendario. Come è stato detto in precedenza, non analizziamo le tabelle riguardanti la concorrenza.

Cliente

In Figura 3.4 sono riportate le tabelle che riguardano il cliente; in particolare ci soffermiamo sulla tabella *Anagrafica clienti* in quanto le altre due indicano solo le insegne associate ai clienti. Nella Tabella 3.3 sono riportati tutti i campi di un cliente con una breve descrizione per ciascuno di essi.

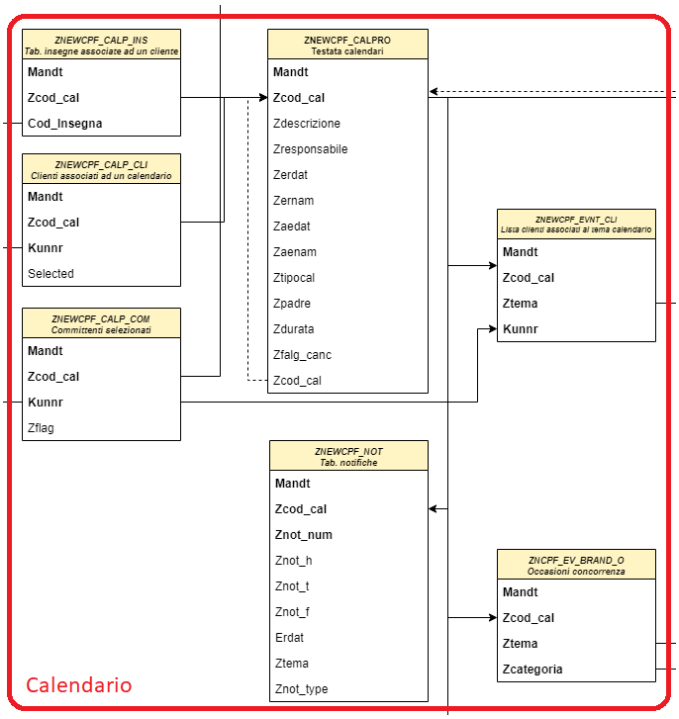

Figura 3.3. Regioni d'interesse: Calendario

None Campo	*Tipo Dati*	*Lunghezza*	*Descrizione breve*
MANDT	CLNT	3	Mandante
ZCOD_CAL	NUMC	8	Codice calendario
ZDESCRIZIONE	CHAR	50	Descrizione calendario
ZRESPONSABILE	CHAR	12	Codice dell'utente responsabile del calendario
ZERDAT	DATS	8	Data di inserimento del calendario
ZERNAM	CHAR	12	Nome dell'utente che ha inserito l'oggetto
ZAEDAT	DATS	8	Data dell'ultima modifica
ZAENAM	CHAR	12	Nome dell'utente autore della modifica
ZTIPOCAL	CHAR	1	Tipo Calendario: 1-Nazionale, 2-Intermedio, 3-Locale
ZPADRE	NUMC	8	Codice calendario padre (se esiste)
ZDURATA	NUMC	4	Parametri contatore
ZFLAG_CANC	CHAR	1	Flag Inclusione (X = Calendario cancellato)

Tabella 3.1. Struttura della Tabella Testata Calendari

Tema calendario

Come già detto, ogni calendario può avere diversi temi, per esempio il tema *Natale* durante il periodo invernale, il tema *Pasqua* nel periodo primaverile, e così via. Per ogni tema possono essere selezionati clienti diversi e, soprattutto, prodotti diversi. In Figura 3.5 si può notare la tabella principale *Tab. lista temi diviso per calendario*,la cui struttura è riportata nella Tabella 3.4, che riporta i dati principali di un tema.

None Campo	*Tipo Dati*	*Lunghezza*	*Descrizione breve*
MANDT	CLNT	3	Mandante
ZCOD_CAL	NUMC	8	Codice del calendario
ZNOT_NUM	CHAR	8	Numero della notifica
ZNOT_H	CHAR	40	Titolo della notifica
ZNOT_T	CHAR	255	Testo della notifica
ZNOT_F	CHAR	1	Flag lettura
ERDAT	DATS	8	Data di inserimento del record
ZTEMA	NUMC	8	Tema dell'operazione del cliente
ZNOT_TYPE	NUMC	1	Tipo notifiche: 1-Committente rimosso dal calendario; 2-Committente aggiunto al calendario; 3-Manca il sell-in; 4-Manca la promozione alla referenza; 5-Promozione non attiva

Tabella 3.2. Struttura della Tabella Notifiche

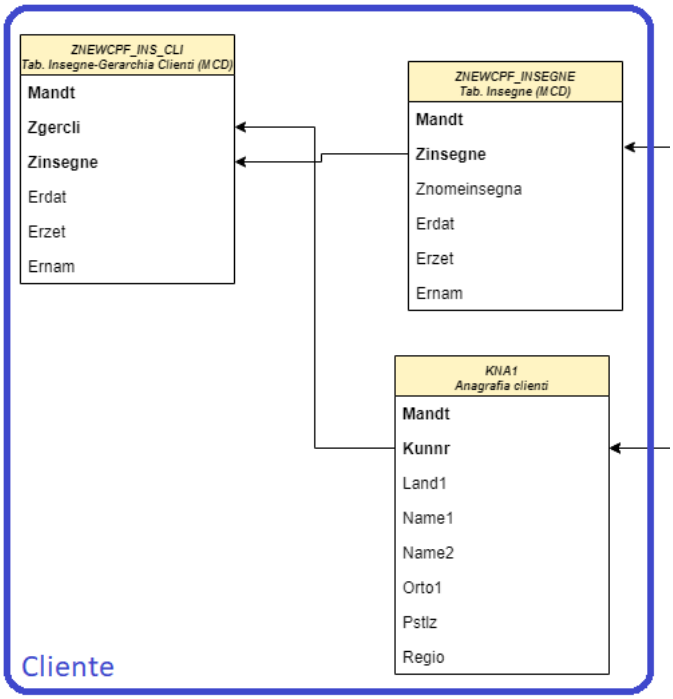

Figura 3.4. Regioni d'interesse: Cliente

None Campo	*Tipo Dati*	*Lunghezza*	*Descrizione breve*
MANDT	CLNT	3	Mandante
KUNNR	CHAR	10	Codice del cliente
LAND1	CHAR	3	Codice del paese
NAME1	CHAR	35	Primo nome
NAME2	CHAR	35	Secondo nome
ORT01	CHAR	35	Località
PSTLZ	CHAR	10	Codice di avviamento postale
REGIO	CHAR	3	Regione (stato federale, stato federato, provincia, contea)

Tabella 3.3. Tabella Anagrafica Clienti

Il campo `ZCOD_Cal` indica il calendario a cui è associato il tema, `ZTEMA` è un numero che incrementa ogni volta che viene creato un nuovo tema. Il campo `ZSTATO` indica lo stato del tema, ovvero se esso, alla sua creazione, è stato ereditato da altri eventi oppure è completamente nuovo. I campi più importanti sono le quattro date di inizio e fine di *sell-in e sell-out*, che sono state spiegate nel capitolo precedente. Esse sono fondamentali in quanto, se il cliente acquista durante il periodo di sell-in, riceve la tipologia di sconto per quel determinato prodotto. Il periodo di sell-out indica quando il prodotto è effettivamente in esposizione nei centri vendita. Il campo `ZAEDAT` è fondamentale per il concetto del tema incrementale, ovvero il tema può essere inserito in un determinato momento e modificato successivamente, quando si hanno informazioni utili su di esso, per esempio i prodotti che saranno in promozione.

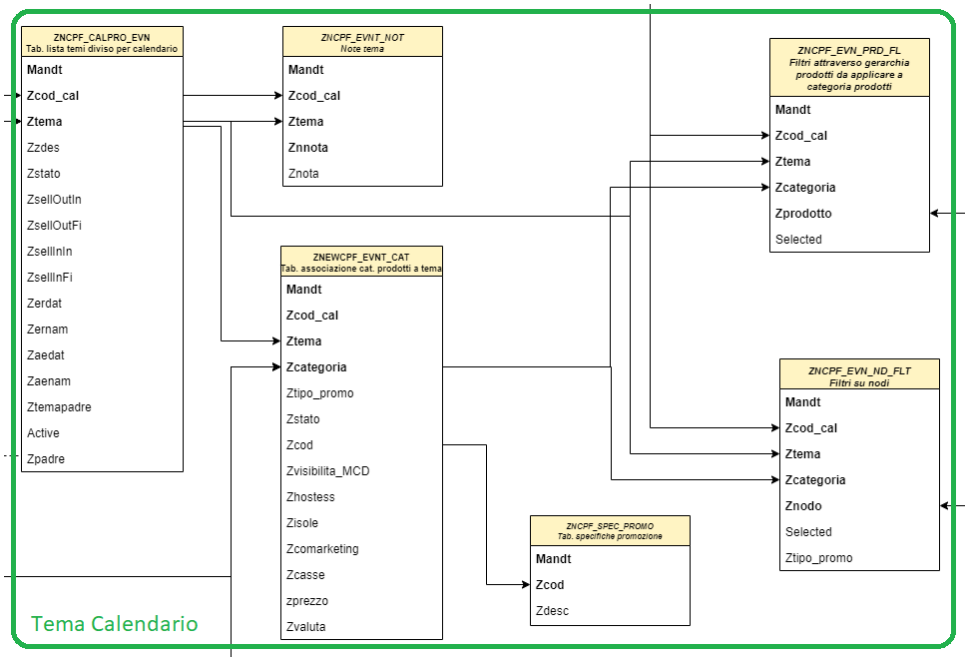

Figura 3.5. Regioni d'interesse: Tema calendario

La tabella *Note tema* in Figura 3.5 indica alcune annotazioni sul tema in analisi, mentre le altre tabelle riguardano i prodotti selezionati per un determinato tema. Ci concentriamo sulla tabella *Tab.associazione cat. prodotti a tema*, la cui struttura è riportata nella Tabella 3.5, perché le altre sono solo filtri sui prodotti o sui nodi promo. In questa tabella, oltre al codice calendario e a quello del tema, viene introdotta la categoria del prodotto; in particolare questo campo, grazie a filtri successivi, arriva ad indicare la referenza che sarà effettivamente in promozione. Altri campi all'interno della tabella indicano come sarà esposto il prodotto in negozio; la disposizione sarà indicata sulla base di quale campo tra `ZHOSTESS, ZISOLE, ZCOMARKETING e ZCASSE` è `TRUE` (vuol dire che nel centro vendita ci sarà

None Campo	*Tipo Dati*	*Lunghezza*	*Descrizione breve*
MANDT	CLNT	3	Mandante
ZCOD_CAL	NUMC	8	Codice calendario
ZTEMA	NUMC	8	Tema dell'operazione del cliente
ZZDES	CHAR	50	Descrizione menù
ZSTATO	CHAR	2	Stato completamento evento (11-Evento nazionale Propagato; 12-Evento nazionale non Propagato; 21-Evento intermedio Ereditato; 22-Evento intermedio non Ereditato; 32-Evento locale On Top).
ZSELLOUTIN	DATS	8	Sell-Out Inizio
ZSELLOUTFI	DATS	8	Sell-Out Fine
ZSELLININ	DATS	8	Sell-In Inizio
ZSELLINFI	DATS	8	Sell-In Fine
ZERDAT	DATS	8	Data di inserimento del record
ZERNAM	CHAR	12	Nome dell'utente che ha inserito l'oggetto
ZAEDAT	DATS	8	Data dell'ultima modifica
ZAENAM	CHAR	12	Nome dell'utente autore della modifica
ZTEMA_PADRE	NUMC	8	tema padre
ACTIVE	CHAR	1	Flag inclusione
ZPADRE	NUMC	8	Codice calendario padre (se esiste)

Tabella 3.4. Tabella lista temi diviso per calendari

una hostess, un'isola, etc).

None Campo	*Tipo Dati*	*Lunghezza*	*Descrizione breve*
MANDT	CLNT	3	Mandante
ZCOD_CAL	NUMC	8	Codice Calendario
ZTEMA	NUMC	8	Tema dell'operazione del cliente
ZCATEGORIA	CHAR	18	Categoria prodotto
ZTIPO_PROMO	CHAR	2	Tipo promo (01-Volantino; 02-TPR (taglio prezzo); 03-Promo Web)
ZSTATO	CHAR	2	Origine tema e obbligatorietà: (11-Attività Nazionale Obbligatoria; 12-Attività Nazionale Facoltativa; 21-Attività Intermedia Ereditata Obbligatoria; 22-Attività Intermedia Ereditata Facoltativa Aderita; 23-Attività Intermedia Ereditata Facoltativa Non Aderita; 24-Attività Intermedia Propria; 34-Attività Locale Propria).
ZCOD	CHAR	3	Codice della specifica
ZVISIBILITA_MCD	CHAR	1	S ->Visibile; Blank ->Non visibile
ZHOSTESS	CHAR	1	Presenza di hostess
ZISOLE	CHAR	1	Presenza di isole
ZCOMARKETING	CHAR	1	Presenza di comarketing
ZCASSE	CHAR	1	Presenza di casse display
ZPREZZO	CURR	15	Valore netto
ZVALUTA	CUKY	5	Chiave divisa

Tabella 3.5. Tabella di associazione categoria prodotti a tema

Promozioni

Come già anticipato, le promozioni (Figura 3.6) vengono realizzate dal reparto vendita che, periodicamente, sceglie quali prodotti devono andare in promozione. Nella Tabella 3.6 vengono indicati i campi di una promozione con una breve descrizione.

Si può notare che ci sono campi per indicare dov'è valida la promozione e i prodotti o i nodi promo interessati da essa.

Una promozione su un prodotto può essere creata in un qualsiasi momento e deve essere associata ad un'attività e a una campagna promozionale, in quanto, come si può notare nella Tabella 3.7, è qui che si determinano il bonus e i tipi di sconto che avranno i vari prodotti grazie al campo `ZZBOTEXT`.

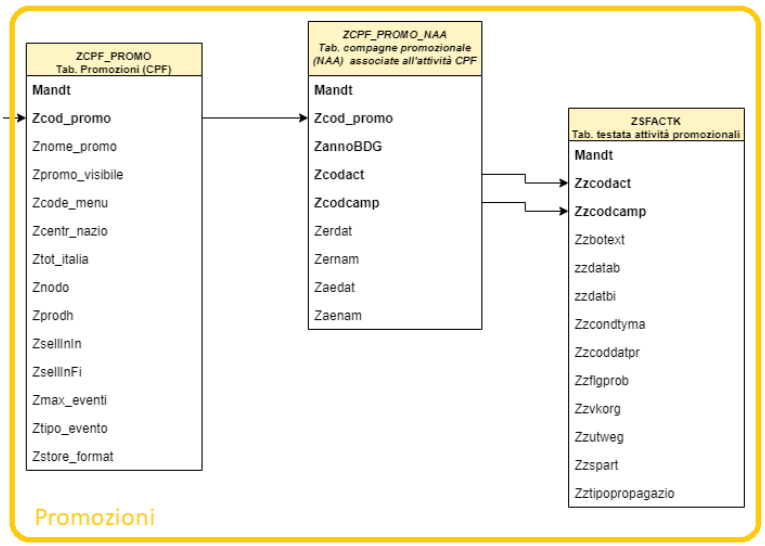

Figura 3.6. Regioni d'interesse: Promozioni

None Campo	*Tipo Dati*	*Lunghezza*	*Descrizione breve*
`MANDT`	CLNT	3	Mandante
`ZCOD_PROMO`	NUMC	8	Codice promozione
`ZNOME_PROMO`	CHAR	50	Descrizione Promozione
`ZPROMO_VISIBILE`	CHAR	1	Visibilità promozione
`ZCOD_MENU`	NUMC	5	Codice menù Promozionale
`ZCENTR_NAZIO`	CHAR	10	Promozione valida per il canale Centrale Nazionale
`ZTOT_ITALIA`	CHAR	1	Promozione valida per tutta l'Italia
`ZNODO`	CHAR	2	Gruppo materiali
`ZPRODH`	CHAR	18	Gerarchia prodotti
`ZSELLININ`	DATS	8	Sell-In Inizio
`ZSELLINFI`	DATS	8	Sell-In Fine
`ZMAX_EVENTI`	CHAR	2	Numero Max eventi
`ZTIPO_EVENTO`	NUMC	1	Tipologia evento
`ZSTORE_FORMAT`	NUMC	1	Store Format

Tabella 3.6. Tabella Promozioni

None Campo	*Tipo Dati*	*Lunghezza*	*Descrizione breve*
MANDT	CLNT	3	Mandante
ZZCODACT	NUMC	5	Codice attività
ZZCODCAMP	NUMC	10	Codice campagna
ZZBOTEXT	CHAR	40	Definizione di un accordo bonus
ZZDATAB	DATS	8	Data di inizio validità
ZZDATBI	DATS	8	Data di fine validità
ZZCONDTYMA	CHAR	4	Tipo condizione manuale
ZZCODDATPR	CHAR	1	Criterio per la determinazione della data del prezzo
ZZFLGPROB	CHAR	1	Flag prenotazione obbligatori
ZZVKORG	CHAR	4	Organizzazione commerciale
ZZVTWEG	CHAR	2	Canale di distribuzione
ZZSPART	CHAR	2	Settore merceologico
ZZTIPOPROPAGAZIO	CHAR	1	Tipo propagazione

Tabella 3.7. Tabella testata attività promozionali

Prodotti e Promozioni Inserite

I prodotti sono gestiti da un altro sottosistema del sistema ERP aziendale. Per quanto riguarda tale tematica, per ciò che concerne le nostre analisi, ci interessa soltanto in concetto di *referenza* e di *nodo promo* presenti nelle tabelle della Figura 3.7. La *referenza è un numero di 18 cifre* che identifica univocamente un prodotto; in tale numero i primi due caratteri indicano il target, il terzo e il quarto carattere indicano l'occasione d'uso, il quinto, il sesto e il settimo denotano il segmento. Le restanti cifre indicano il prodotto vero e proprio. Un *nodo promo* è costituito da un insieme di referenze che, di solito, vanno insieme in promozione.

La tabella *Promozioni Inserite* in Figura 3.7 la cui struttura è riportata nella Tabella 3.8 è fondamentale in quanto indica i prodotti che i responsabili di un calendario hanno selezionato per un determinato tema. La tabella mostra che si può selezionare sia la singola referenza (grazie al campo ZREFERENZA), sia il nodo promo che ha più referenze all'interno. Un concetto fondamentale è che una referenza inserita in un tema di un calendario non ha necessariamente uno sconto; lo ha solo se il campo ZATTIVITA contiene il codice di una promozione associata a una campagna promozionale.

None Campo	*Tipo Dati*	*Lunghezza*	*Descrizione breve*
MANDT	CLNT	3	Mandante
ZCOD_CAL	NUMC	8	Codice del Calendario
ZTEMA	NUMC	8	Tema dell'operazione del cliente
ZCATEGORIA	CHAR	18	Categoria del prodotto
ZNODO	CHAR	10	Nodo Promozionale: ID
ZREFERENZA	CHAR	18	Gerarchia prodotti
ZATTIVITA	NUMC	8	Codice Promo.
ACTIVE	CHAR	1	Attivo (X ->Prenotazione generata)

Tabella 3.8. Tabella Promozioni inserite

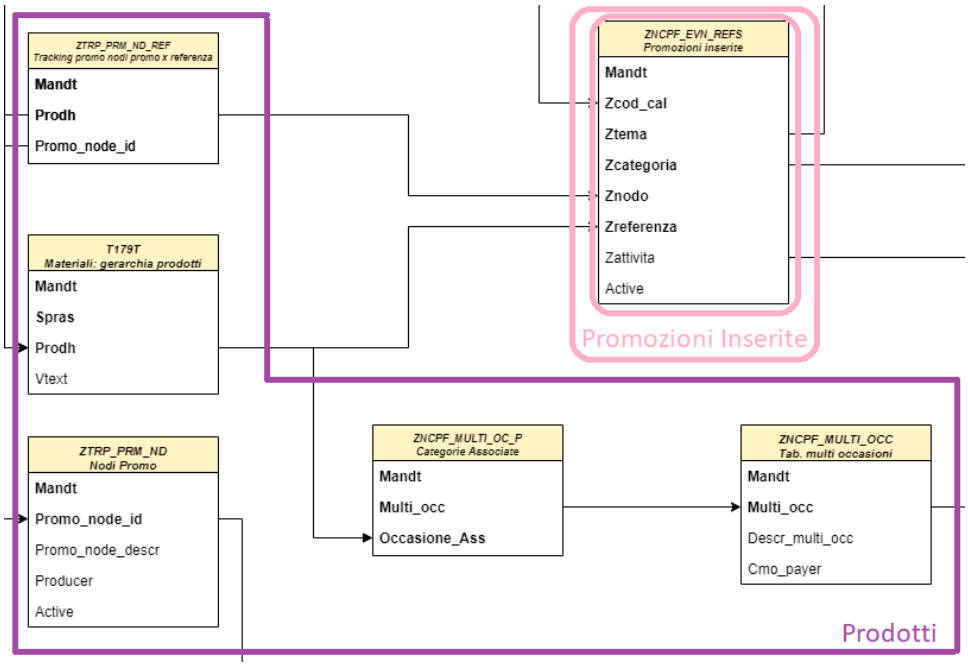

Figura 3.7. Regioni d'interesse: Prodotti e Promozioni Inserite

3.3 Dataset Promozioni

Il set di dati sulle promozioni è costituito da due file in formato *CSV*. Il primo ha una dimensione di circa 230 Mb per un totale di 473417 righe e 26 colonne, e copre i dati relativi alle promozioni inserite nel 2020. Il secondo ha una dimensione pari a 133 Mb e 295634 righe con lo stesso numero di colonne del precedente, ma i dati riguardano l'anno 2021. Il numero di colonne è più elevato rispetto alla Tabella 3.8 perché sono state aggiunte più informazioni provenienti da altre tabelle. La granularità dei dati è per singolo committente; quindi, per ogni tema di un calendario, viene riportata una riga in base alle combinazioni tra i committenti e le referenze del tema stesso. I nomi delle colonne sono riportati in seguito con una piccola descrizione:

- Calendario (Full Name): stringa che contiene l'Id del calendario insieme al suo nome.
- Tipo Calendario: stringa che contiene il tipo del calendario. Può assumere i valori seguenti:
  - "1 - Nazionale": indica un calendario nazionale;
  - "2 - Intermedio": indica un calendario intermedio, che non ha validità nazionale;
  - "3 - Locale": indica un calendario che ha validità solo in un determinato posto.
- Tema Cliente: stringa che contiene l'Id del tema insieme al suo nome.
- Tipo Promozione: indica se la promozione è a volantino oppure un taglio prezzo.
- Insegna: indica l'insegna del cliente.

- Sell-Out Inizio: data di inizio del sell-out, ovvero quando i prodotti iniziano ad essere in esposizione nei vari centri.
- Sell-Out Fine: indica la fine del periodo di sell-out.
- Sell-In Inizio: data a partire dalla quale, se si effettua un acquisto, si ottiene il bonus/sconto rispettivo della promozione.
- Sell-In Fine: indica la fine del periodo di sell-in.
- Gerarchia clienti: indica, per ogni client, tutta la gerarchia:
  - Canale:
  - Super Centrale;
  - Centrale Nazionale;
  - Gruppo Regionale;
  - Business Partner;
  - Committente (FullName): contiene sia l'Id del committente che il suo nome.
- Occasione d'Uso: indica l'occasione d'uso della referenza (prodotto) di ogni riga.
- Segmento: indica il segmento della referenza.
- Formato: indica il formato della referenza.
- Target Consumo: indica il target consumo della referenza.
- Nodo Promo: indica se la referenza inserita appartiene ad un nodo promo.
- Referenza: è la referenza inserita nella promozione.
- Stato Promozione: può essere "completa", nel qualcaso la referenza ha degli sconti associati, oppure "senza sconto", che indica l'assenza di sconto.
- Promozione (FullName): è una stringa che contiene l'id della promozione insieme al nome.
- Menu Promo: codice del menù promozionale.
- Naa Promo: stringa che contiene al suo interno sia l'Id dell'attività promozionale, quello della campagna promozionale, nonchè una piccola descrizione del bonus.

3.4 Dataset Vendite

Il set di dati sulle vendite è composto da tre file in formato *CSV*. Due di questi riguardano le vendite nel 2020. Le le dimensioni cambiano rispetto al dataset delle promozioni, in quanto un file occupa 356 Mb per un totale di 1048566 righe, mentre il secondo 228 Mb per un totale di 674425 righe. Il terzo file è composto dai dati del 2021 ed è più piccolo in dimensione (284 Mb) rispetto all'anno precedente, con un numero di righe pari a 939953, a causa della mancanza dei dati della seconda parte dell'anno. La granularità è la stessa del dataset delle promozioni. In seguito vengo mostrate le colonne del dataset; come si può osservare, molti campi sono identici a quello delle promozioni; per tale ragione, vengono spiegati solo i nuovi campi:

- Data Rif: data in cui, per una determinata referenza, è stata eseguita un'azione di fatturazione, accredito o addebito in base al tipo di documento;
- Referenza: referenza che è impegnata nell'atto di vendita;
- Nodo Promo;
- Formato;
- Segmento;
- Occasione Uso;

- Target Consumo;
- N. Conf in Cart: numero che indica quante confezioni ci sono in un cartone;
- SU Factor (CartToSU): è un numero che rappresenta un coefficiente che viene utilizzato dall'azienda per mettere tutti i prodotti sulla stessa scala. Una confezione di pannolini è diversa da una confezione di assorbenti; utilizzando tale coefficiente viene messo tutto alla pari. Ogni referenza ha il proprio coefficiente.
- Canale;
- Centrale Internazionale;
- Super Centrale;
- Centrale Nazionale;
- Gruppo Regionale;
- Business Partner;
- Pagatore: identifica chi fa il pagamento; spesso coincide con il committente, ma non sempre ciò accade a causa della gerarchia clienti.
- Committente;
- Tipo Documento: i tipi di documenti possibili sono rappresentati nella Tabella 3.9.
- Qtà Cartoni: indica quanti cartoni si considerano nella fattura per la referenza.
- Qtà CONF: indica quante confezioni si considerano nella fattura per la referenza.
- Qtà SU: quantità che si basa sul "SU Factor".
- Fatturato Netto: indica il fatturato della referenza.

Tipo Documento	*Descrizione*
Fattura YF1	Fattura prodotti finiti
Fattura YF5-fattura Gare Divos	Prodotti venduti alla Pubblica Amministrazione (ASL) tramite gara
Fattura YFD-fattura Cons. Domic.	Fattura di prodotti che si vendono alla Pubblica Amministrazione (ASL) tramite gara che prevede la consegna domiciliare
Fattura YFR-Fatt.Val. PF (no UM)	Rifatturazione di prodotti che non prevede il movimento di merce (viene utilizzato per correggere fatture già effettuate)
Storno fattura ZSFT-Storno	Documento di annullamento di una fattura
Accredito YG2-NC Prezzi / Sconti	Documento per accreditare al cliente valori derivanti da un errata attribuzione di prezzo e/o sconti
Accredito YG3-NC Campagne	Documento per accreditare al cliente valori derivanti da campagne a premio
Accredito YGC-NC Varie	Documento per accreditare al cliente valori senza uno specifico motivo
Accredito YRE-NC Reso	Documento per accreditare al cliente valori derivanti da restituzione merce
Accredito YRV-NC Reso a Valore	Documento per accreditare al cliente valori derivanti da restituzione valori
Nota di debito YL2-ND Prezzi / Sconti	Documento per addebitare al cliente valori derivanti da un errata attribuzione di prezzo e/o sconti
Nota di debito YLC-ND Sconto Cassa	Documento per addebitare al cliente valori derivanti da un errata attribuzione dello sconto cassa puntualità di pagamento
Nota di debito YLV-ND Varie	Documento per addebitare al cliente valori senza uno specifico motivo
Nota di debito YLZ-ND Cons. in eccesso	Documento per addebitare al cliente valori derivanti da una consegna di merce superiore a quella riportata nella fattura originale
Storno nota cred. ZSNC	Documento di annullamento di una nota di accredito

Tabella 3.9. Tabella con i tipi di documento

4

ETL e analisi descrittiva dei dati

In questo capitolo verranno analizzate le fasi di ETL e di EDA sui dataset delle promozioni e delle vendite e sarà effettuata un'analisi descrittiva dei dati.

4.1 Introduzione

In questo capitolo verranno analizzati più in dettaglio i dati contenuti nei dataset relativi alle promozioni e alle vendite. Prima di analizzare i nostri dati, verrà introdotto il concetto di ETL e di EDA. Per motivi di riservatezza, in tali analisi, saranno oscurati i dati sensibili dell'azienda come, ad esempio, il fatturato, in ottemperanza agli obblighi che sono stati assunti in seguito alla firma di un NDA (Non Disclosure Agreement) con l'azienda.

ETL sta ad indicare i processi di *Extract, Transform e Load*, ovvero estrazione, trasformazione e caricamento dei dati. I dati vengono raccolti da più sorgenti, trasformati in base a ciò che si vuole fare con essi e caricati in un sistema comune dove, successivamente, verranno utilizzati. La parte più importante è quella della trasformazione che, di solito, include diverse operazioni, tra cui quelle di filtraggio, ordinamento, aggregazione, join e pulizia generale, eseguite da un motore specializzato.

Spesso le fasi di ETL vengono eseguite in parallelo, al fine di ridurre i tempi di esecuzione. Per esempio, durante l'estrazione dei dati, può già essere avviata un'operazione di trasformazione; inoltre, non si aspetta la fine delle operazioni per eseguire la fase di caricamento dei dati, ma essi si caricano volta per volta.

Si parla anche di ELT, ovvero *Extract, Load e Transform*, e la differenza rispetto al caso precedente sta solo nel punto in cui avvengono le trasformazioni sui dati, che vengono realizzate direttamente nel sistema di destinazione, non più da un motore specializzato. Tale approccio richiede che il sistema finale sia abbastanza potente per permettere, in maniera efficiente, le varie operazioni sui dati.

EDA, invece, sta per *Exploratory Data Analysis*, ed è una tecnica che viene usata nel campo della *Data Science* per approfondire la conoscenza del dataset e poter, successivamente, lavorare con esso. Questa tecnica permette di capire i tipi di dati con cui si lavorerà, nonché di avere delle statistiche immediate su di essi.

Nelle prossime sezioni analizzeremo la fase di trasformazione dei due set di dati introdotti nel capitolo precedente, capiremo con che dati dovremo lavorare e, infine, proporremo un'analisi descrittiva dei contenuti dei dataset.

4.2 Trasformazione ed esplorazione dei dati

Nel nostro caso non è possibile parlare di una vera e propria attività di ETL, in quanto i dati ci sono stati forniti direttamente in formato *CSV*, saltando, dunque, la fase di *estrazione*, che è stata precedentemente eseguita dai responsabile della Fater prima di consegnarci i dati.

Lo strumento utilizzato per lavorare con i file *CSV* è *Pandas*, una libreria software scritta per il linguaggio di programmazione *Python*, usata per la manipolazione e l'analisi dei dati. Tale libreria permette di lavorare con dati salvati in diversi formati, tra cui il *CSV*, grazie all'apposito metodo `read()` che, nel nostro caso, è `read_csv()`. Tale metodo trasforma i formati diversi in un formato specifico che prende il nome di *DataFrame*, una tabella strutturata su colonne dove i dati sono distribuiti per righe. Sia le colonne che le righe sono indicizzate al fine di facilitare l'accesso ai dati da esse contenute. In particolare, per rappresentare una riga o una colonna di un *DataFrame*, si utilizza la *Series*, un array monodimensionale etichettato, in grado di contenere qualsiasi tipo di dati.

Una volta caricati i dati nel nostro ambiente di lavoro, grazie ai tanti metodi offerti dalla libreria, è stata eseguita una fase di pulizia all'interno delle colonne dei due dataset, in quanto l'informazione contenuta poteva essere scomposta in più informazioni utili. Principalmente è stato utilizzato il metodo `Series.str.split()`, per dividere il contenuto di una colonna in più colonne, ciascuna con un significato ben preciso; per capire meglio il concetto, prendiamo come esempio la colonna del dataset delle promozioni *Calendario (Full Name)*. Inizialmente aveva due informazioni all'interno, separate dal simbolo " − ". La prima informazione era l'Id del calendario mentre la seconda era rappresentata dal nome. Utilizzando il comando `df["Calendario (Full Name)"].str.split(" - ", expand = True)`, dove `df["Calendario (Full Name)"]` è la serie dei dati su cui attuare il metodo di `split()`, si divideranno in due colonne i dati, in base al parametro " − " passato ad esso. Se il parametro "*expand*" è `True` vuol dire che saranno aggiunte al DataFrame le nuove colonne. Per indicare il nome di queste colonne si può utilizzare il metodo `rename()`, che permette di sceglierlo.

Una volta che ad ogni colonna è stata associata una sola informazione utile, si è passati all'espolorazione dei dati per vedere che informazioni si possono ricavare da essi. Una buona pratica per iniziare ad esplorare i dati è quella di capire con che tipi e con quanti dati abbiamo a che fare. Per fare ciò sono stati utilizzati diversi metodi indicati in seguito:

- `DataFrame.shape()`: questo metodo restituisce la dimensione del DataFrame su cui si applica, in particolare il numero di righe e il numero di colonne.
- `DataFrame.info()`: questo metodo restituisce le informazioni riguardanti il DataFrame in analisi; in particolare, per ogni colonna, indicando se può contenere valori nulli e quanta memoria essa occupa.

- `DataFrame.describe()`: genera delle statistiche descrittive dei dati contenuti nel DataFrame. Esse includono quelle che riassumono la tendenza centrale, la dispersione e la forma della distribuzione di una serie di dati, esclusi i valori `NaN`. Analizza sia le serie numeriche che quelle di oggetti, ma su queste ultime fornisce meno informazioni, in quanto ci dice soltanto, per ogni colonna, il numero di elementi `unique` all'interno, l'oggetto più utilizzato e la sua frequenza.

Utilizzando questi metodi ci possiamo fare un'idea sui dati. Nel nostro caso, per quanto riguarda i dati delle promozioni, si è notato che ci sono soltanto dati categorici, che sono i vari ID delle colonne, o altri dati, che sono stringhe che rappresentano il nome o una descrizione di un campo. Solo quattro colonne indicavano delle date (*sell in e sell-out*). Quindi, per quanto riguarda il primo set di dati, come vedremo in seguito, le analisi eseguite su di esso sono di tipo descrittivo, attraverso l'utilizzo di grafici, per vedere la distribuzione dei campi più importanti.

I dataset sulle vendite, come detto nel capitolo precedente, hanno tanti campi in comune con il primo dataset, ma quelli in più ci hanno permesso di avere anche dei campi numerici. In particolare, nella Tabella 4.1, sono riportati i dati che riguardano le statistiche descrittive delle vendite del 2020 (i due *CSV* del 2020 sono stati concatenati grazie al metodo `concat()`, arrivando ad un numero di righe pari a 1.722.991). Una cosa importante da vedere è il valore del `count` nella tabella, ovvero 785189.0, che indica il numero totale di righe del dataset delle vendite del 2020; questo è diverso dal numero precedente, perché tante righe sono state cancellate. Questa cancellazione è stata necessaria nei casi in cui i campi `Qta Cartoni`, `Qta CONF`, `Qta SU` e `Fatturato Netto` erano tutti uguali a zero contemporaneamente; chiedendo ai responsabile dell'azienda, abbiamo scoperto che quelle righe erano un refuso che generava il modulo BI da cui sono stati estratti i dati e che erano inutili per l'analisi.

Metrica	*Qta Cartoni*	*Qta CONF*	*Qta SU*	*Fatturato Netto*
`count`	785189.0	785189.0	785189.0	****
`mean`	62.676756	387.025897	43.657603	****
`std`	246.902023	1333.000927	147.986979	****
`min`	-2580.0	-31590.0	-2207.0	****
`25%`	2.0	10.0	1.0	****
`50%`	6.0	28.0	5.0	****
`75%`	42.0	224.0	33.0	****
`max`	20520.0	116736.0	11080.0	****

Tabella 4.1. Tabella con le statistiche descrittive del dataset delle vendite 2020

4.3 Analisi descrittiva

In questa sezione vengono riportati alcuni grafici utili per la comprensione del contenuto dei dati. In particolare, verranno analizzati per prima i dati delle promozioni, per poi vedere quelli delle vendite. I vari grafici riguardano soltanto i dati dell'anno 2020, in quanto era interessante vedere i dati all'interno di un anno, e, come già detto, i dati riguardanti il 2021 non erano completi.

Dati promozioni

Per prima cosa, nella Tabella 4.2, viene mostrato, per ogni campo (senza considerare i vari Id, che sono categorici, e le quattro date) il numero totale di valori distinti. Per visualizzare il contenuto del dataset delle promozioni, a livello di codice, ovvero utilizzando i vari metodi offerti da *Pandas*, i dati sono stati analizzati molto più in dettaglio. A causa di troppi valori categorici, di stringhe e date, la loro rappresentazione è stata difficile, in quanto, per rappresentarli, è stato sempre necessario l'utilizzo di un `count` e i grafici ottenuti possono essere molto simili tra loro anche se l'informazione che rappresentano è diversa.

In Figura 4.1 viene rappresentato l'andamento del numero dei calendari promozionali e il numero delle promozioni nei mesi dell'anno. Si può notare che, all'incirca, l'andamento del numero del calendario è lo stesso nei vari mesi, mentre, a partire da Giugno, il numero delle promozioni incrementa. L'andamento dei calendari ha senso perché, come già detto, al loro interno, ci sono i diversi temi che possono richiamare promozioni diverse. Quindi lo stesso calendario può essere, o meno, presente in più mesi, e le promozioni possono essere richiamate in più calendari. Le promozioni vengono gestite dal reparto vendite e il loro aumento nella seconda parte dell'anno è sicuramente dovuto alla maggior richiesta sul mercato nel periodo estivo e pre-natalizio.

Per quanto riguarda il numero di temi associati ai vari calendari nei mesi, l'andamento è mostrato in Figura 4.2. Dall'analisi della figura si può notare che Gennaio e Luglio hanno un picco. Tutto sommato l'andamento dei temi è molto simile a quello dei calendari; in particolare si può dedurre che, mediamente, ogni calendario ha due o tre temi associati al mese.

Campo	*Valori distinti*
`Nome Calendario`	134
`Tipo Calendario`	2
`Tema Cliente`	1672
`Tipo Promozione`	2
`Insegna`	143
`Canale`	6
`Super Centrale`	21
`Centrale Nazionale`	26
`Gruppo Regionale`	86
`Business Partner`	118
`Committente`	340
`Occasione d'Uso`	20
`Segmento`	46
`Formato`	133
`Nodo Promo`	207
`Target Consumo`	7
`Referenza`	671
`Stato Promozione`	2
`Nome Promozione`	103
`Menu Promo`	22
`Naa Promo`	99
`Anno Fatturazione`	2
`Sconto`	29

Tabella 4.2. Tabella con i valori distinti per i vari campi

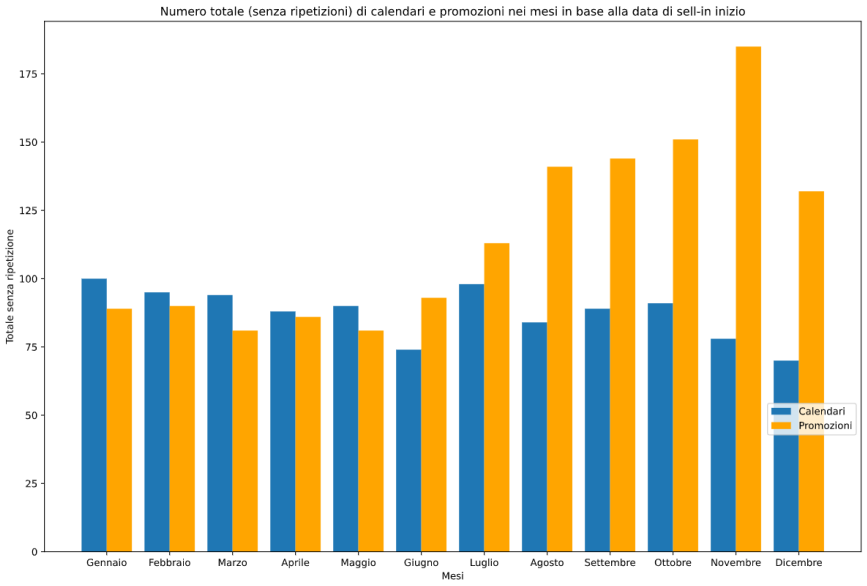

Figura 4.1. Andamento dei calendari e delle promozioni nell'anno

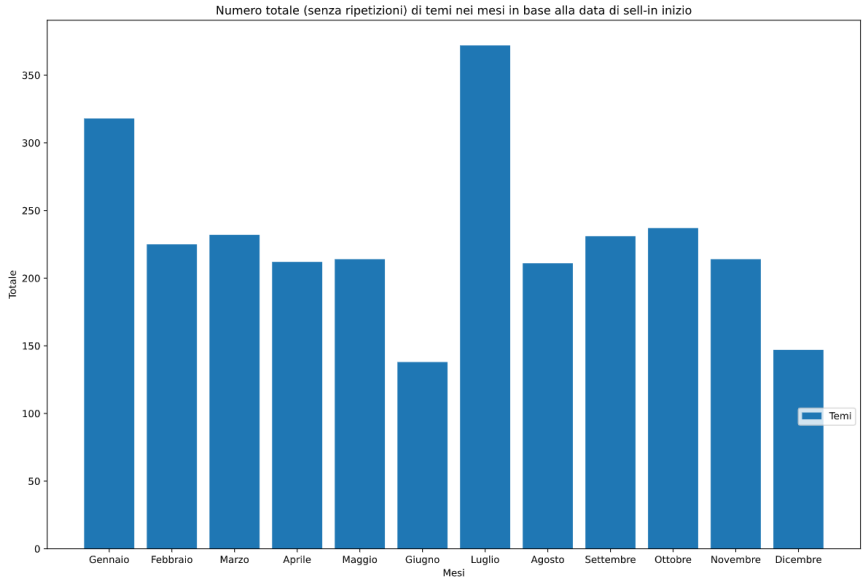

Figura 4.2. Andamento dei temi nell'anno

Successivamente, si è visto quanti committenti sono associati ai vari calendari. Come mostra la Figura 4.3, attraverso il passaggio dell'indicatore del mouse sul grafico, questo mostra il nome del calendario insieme all'Id e al numero di committenti totali di quel calendario. La grandezza delle bolle nel grafico indica la presenza di più committenti; anche se ci sono tanti punti con pochi committenti, il grafico è fatto in modo tale da poter ingrandire la zona desiderata e osservare le informazioni utili.

Lo stesso approccio è stato applicato anche in Figura 4.4 per vedere il numero di insegne associate ad un calendario. In particolare, si può notare che la maggior parte dei calendari hanno da 1 a 5 insegne associate, mentre sono solo due i calendari che hanno, rispettivamente, 17 e 22 insegne.

Per ogni insegna sono stati calcolati il numero di committenti ad essa associati; il risultato è mostrato in Figura 4.5.

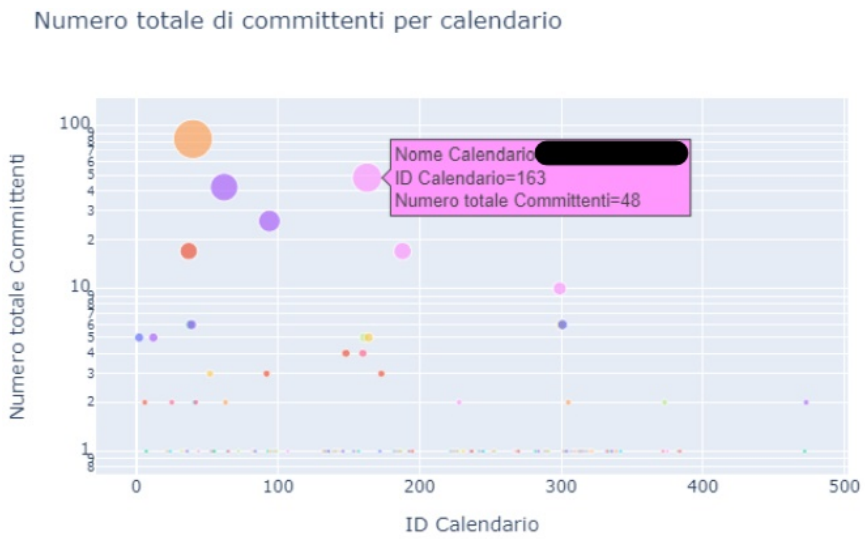

Figura 4.3. Rappresentazione del numero di committenti per calendario

Per quanto riguarda le *referenze*, oltre a vedere quante di queste sono presenti all'interno del dataset, nella Tabella 4.2 vengono riportate le informazioni utili riguardo ad esse. In particolare si può vedere il numero di *occasioni d'uso*, di *segmenti*, di *formati* e di *target* presenti. In Figura 4.6 viene mostrato l'andamento dei 6 tipi di target nei mesi; in particolare, a differenza del numero rappresentato in tabella, vengono mostrati solo 6 target e non 7, in quanto uno di essi era rappresentato dalla stringa "`Non Definito`", e il numero di righe che conteneva tale stringa erano solo 48 e sono state cancellate. Anche il numero di *Altri (Cosmetici)* è molto basso e in figura non si nota. Anche l'andamento dei target è molto simile a quello delle promozioni e si può notare che i target più quotati sono *Adult, Fem e Baby Care*.

In Figura 4.7 invece, vengono rappresentati i *nodi promo*, con il rispettivo numero di referenze ad essi associato. In questo caso, visto il numero abbastanza elevato di nodi promo, l'istogramma è fatto in modo tale da riuscire ad ingrandire una zona

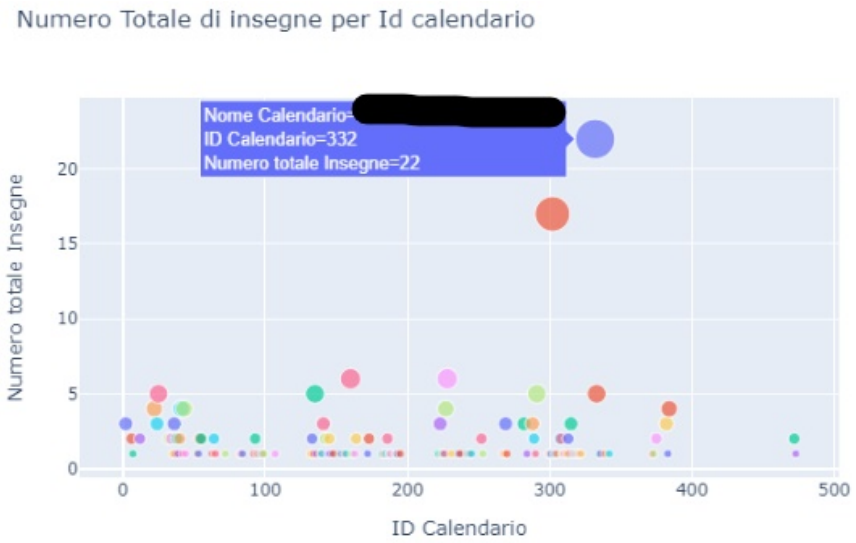

Figura 4.4. Rappresentazione del numero di insegne per calendario

Figura 4.5. Rappresentazione del numero di committenti per insegna

specifica, attraverso la selezione di un'area nel grafico per vedere da vicino i nodi interessati.

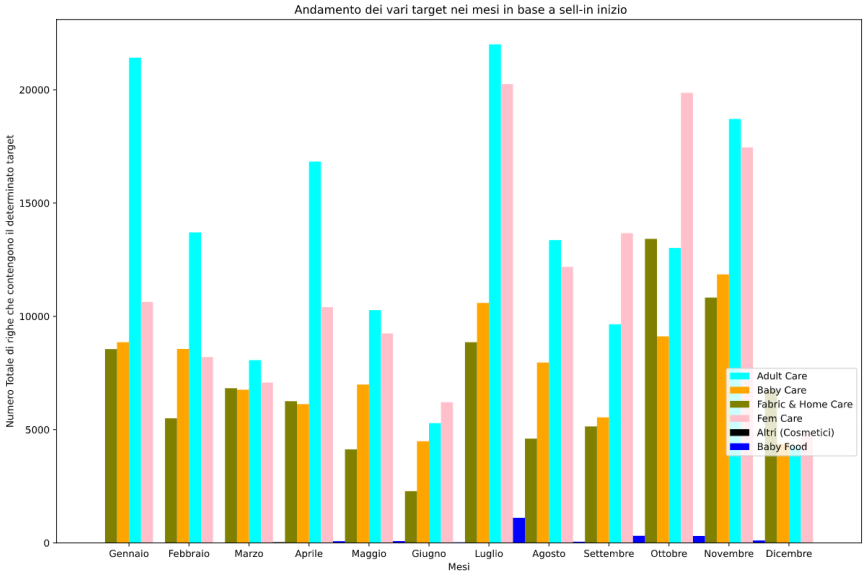

Figura 4.6. Andamento dei target nei mesi

Nella Figura 4.8 è stata analizzata la durata media nei mesi del periodo di *sell in* per vedere se ci fosse qualche mese con valori diversi e capire se la durata poteva avere qualche collegamento con gli altri grafici realizzati in precedenza. In figura si può osservare che l'andamento è sempre intorno alla media (33.17 giorni), tranne, per il mese di Giugno che, come negli altri grafici, è quello con valori più bassi.

La stessa cosa è stata fatta per il periodo di *sell out*, per rappresentare la sua durata media e per vedere se ci fosse qualcosa di interessante. Come si può notare, a differenza del grafico precedente, la media è molto più bassa, in quanto tale periodo rappresenta il tempo in cui il prodotto si trova nel centro vendita, e questo è più breve del precedente.

Dati vendite

Anche nel caso dei dati delle vendite si considerano solo quelli dell'anno 2020. Nella Tabella 4.3 sono riportati i campi del dataset delle vendite, con il rispettivo numero di valori distinti che contengono (anche qui non considerando le date, i valori categorici e quelli numerici già considerati nella Tabella 4.2). Come si può facilmente notare, i numeri dei campi di questa tabella sono diversi, più alti rispetto ai valori del dataset delle promozioni (Tabella 4.2). Ciò è dovuto al fatto che chi compra, può usufruire o meno di una promozione.

In seguito sarà analizzato il contenuto del dataset delle vendite, soffermandosi sui nuovi campi rispetto a quelli delle promozioni, considerando in particolare, l'andamento del fatturato e delle quantità presenti. In Figura 4.10 viene mostrato

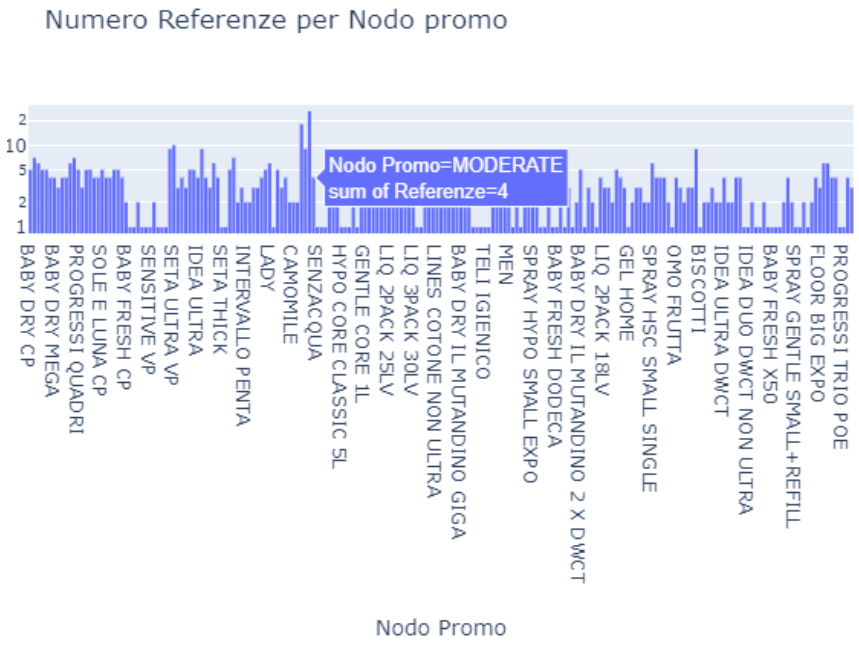

Figura 4.7. Rappresentazione del numero di referenze per nodo promo

Figura 4.8. Durata del periodo di sell in

Figura 4.9. Durata del periodo di sell out

l'andamento del fatturato in base al campo `Data Rif`. Per essere realizzato, per ogni giorno presente nel dataset delle vendite, è stato realizzato un `groupby('Data Rif').agg('sum')` che sommava tutti i fatturati delle diverse referenze di quel determinato giorno. Si può notare che, ad ogni fine mese, sono presenti valori molto più elevati del fatturato, e questo porta a pensare che gli acquisti da parte dei committenti si concentrano in questa parte del mese. La linea tratteggiata nella figura indica la media, che è pari a ****.

Si è notato che all'interno della colonna del fatturato a volte comparivano valori negativi; per questo, nelle Figure 4.11 e 4.12, sono stati riportati gli andamenti dei fatturati positivi e negativi, con la rispettiva media. L'andamento del fatturato positivo è simile a quello totale; cambiano soltanto alcuni valori, come la media, pari a ****. I valori negativi, invece, sono dovuti al fatto che è l'azienda che paga il committente in base al tipo di documento che corrisponde al fatturato. I tipi di documenti sono già stati spiegati in precedenza. Per questo, nella Figura 4.13, vengono riportati solo l'andamento dei fatturati di questi ultimi senza spiegarli in dettaglio. Da notare che i documenti di tipo fattura sono quelli che hanno valori più alti.

In Figura 4.14 viene mostrato l'andamento dei target nei mesi. Tale grafico è uguale a quello delle promozioni per quanto riguarda la forma, ma cambia il contenuto, in quanto si possono osservare valori diversi dei target. In questo caso si nota che la presenza del target *Baby Care* ha valori molto più elevati rispetto al set dei dati precedenti, e questo porta a pensare che in tanti acquistano tali prodotti, anche se essi non sono effettivamente in una promozione, oppure il committente stesso non appartiene ai diversi temi che permettono i benefici.

L'ultimo grafico, in Figura 4.15, rappresenta l'andamento delle quantità presenti nel dataset. Si può osservare che, come è normale che sia, le tre quantità hanno andamento simile ma cambia il numero, in quanto, per ogni riga, vengono riportate

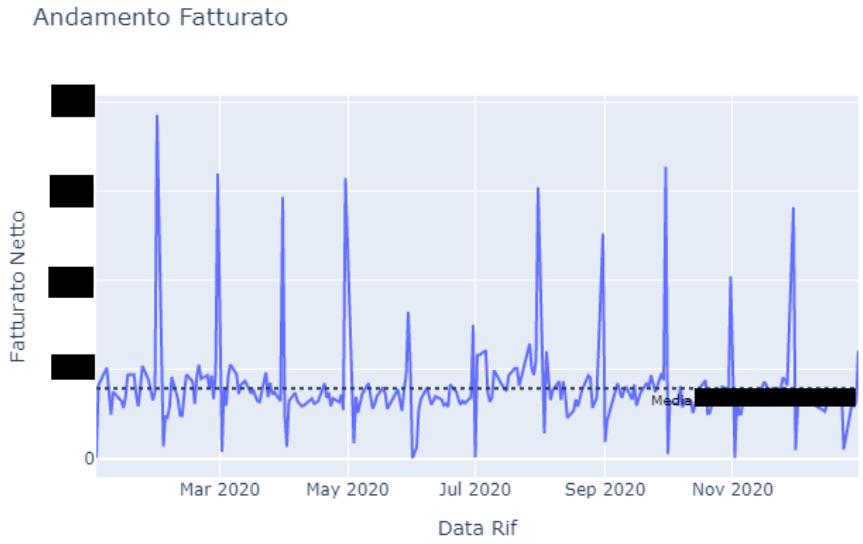

Figura 4.10. Andamento del fatturato

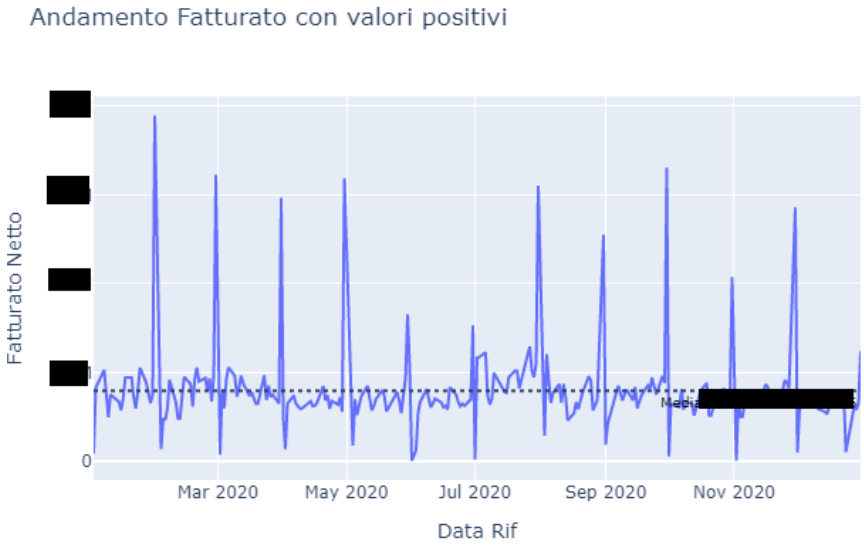

Figura 4.11. Andamento del fatturato con valore positivo

Campo	*Valori distinti*
Referenza	865
Nodo Promo	239
Formato	160
Segmento	55
Occasione d'Uso	25
Target	7
Canale	15
Centrale Internazionale	35
Super Centrale	57
Centrale Nazionale	89
Gruppo Regionale	338
Business Partner	1119
Committente	5899
Pagatore	4795
Tipo documento	14

Tabella 4.3. Tabella con i valori valori distinti per i vari campi del dataset delle vendite

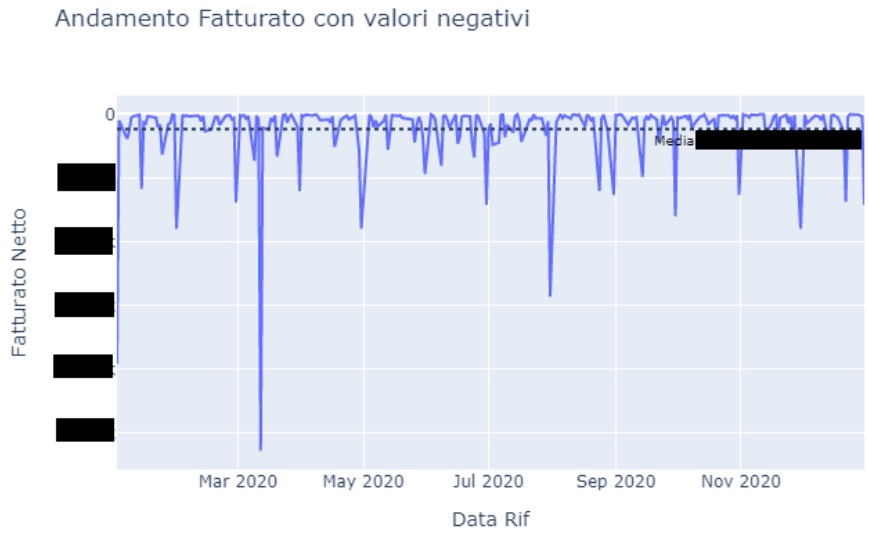

Figura 4.12. Andamento del fatturato con valore negativo

tutte e tre, e queste sono tra loro correlate: più *confezioni* possono appartenere ad un *cartone* e, la *quantità SU* dipende da un fattore interno dell'azienda che mette tutti i prodotti alla pari; quindi, sicuramente, in tale fattore, vengono considerati anche i dati dei cartoni e delle confezioni.

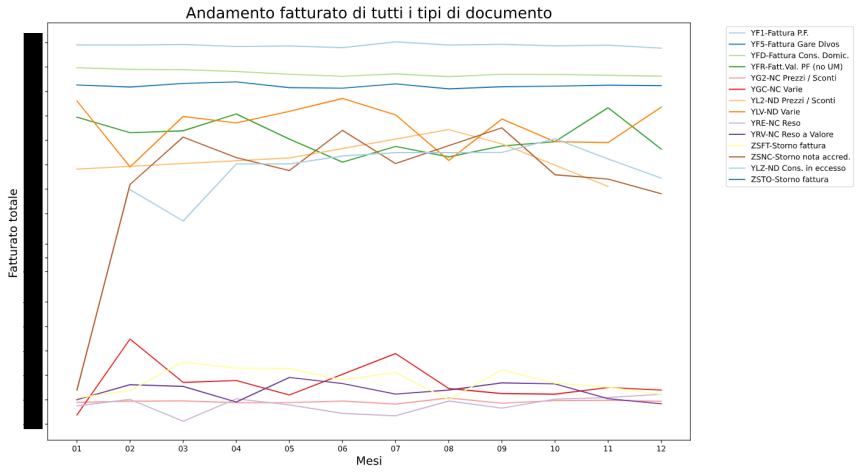

Figura 4.13. Andamento del fatturato di tutti i tipi di documento

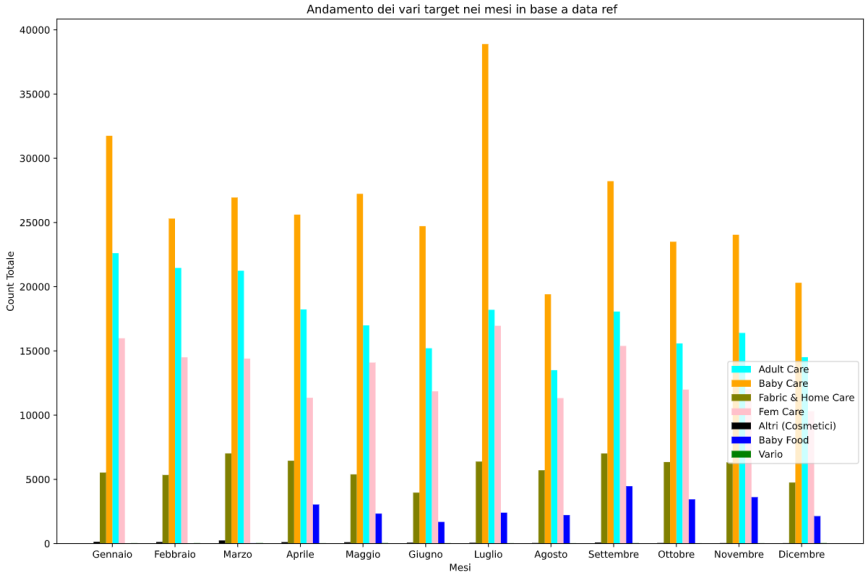

Figura 4.14. Andamento dei tipi di target nei mesi

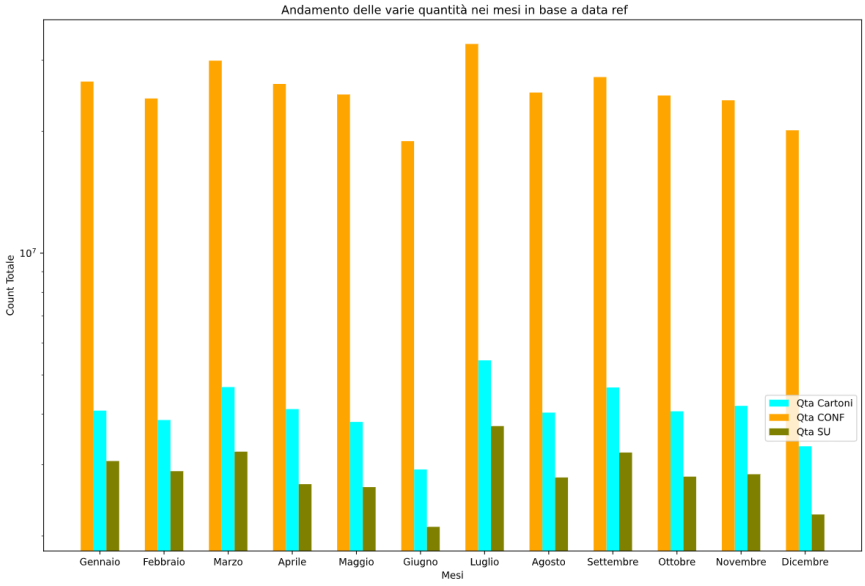

Figura 4.15. Andamento delle quantità di confezioni, di cartoni e SU nei mesi

5

Implementaione della campagna di Data Analytics

In questo capitolo verranno analizzate le fasi di clustering e classificazione dei dataset.

5.1 Introduzione

Classificazione e Clustering sono i due tipi di metodi di apprendimento che organizzano gli oggetti in gruppi in base a una o più caratteristiche. Questi processi sembrano essere simili, ma presentano differenze nel contesto del data mining. La differenza principale tra la classificazione e il clustering è che la prima è una tecnica di apprendimento supervisionato, mentre la seconda è usata nell'apprendimento non supervisionato. In seguito, viene spiegata la differenza tra i due tipi di apprendimento:

- *Apprendimento supervisionato*: vengono presentati al computer degli input di esempio, ed i relativi output desiderati, con lo scopo di apprendere una regola generale in grado di "mappare" gli input negli output.
- *Apprendimento non supervisionato*: al computer vengono forniti solo dei dati in input, senza alcun output atteso, con lo scopo di individuare una qualche struttura nei dati d'ingresso.

La classificazione è il processo di apprendimento di un modello che opera con diverse classi presenti nei dati. E' composta da due fasi: la prima, di apprendimento, in cui viene costruito un modello, e la seconda, di classificazione, in cui il modello costruito viene usato per assegnare le etichette di classe ai nuovi dati passati ad esso.

Il clustering è una tecnica di organizzazione di un gruppo di dati in classi, o *cluster* dove gli oggetti che risiedono all'interno di uno stesso cluster avranno un'alta somiglianza fra loro, mentre gli oggetti di due cluster differenti saranno dissimili. L'obiettivo principale del clustering è quello di dividere l'insieme dei dati in più cluster. A differenza del processo di classificazione, le etichette di classe degli oggetti non sono note a priori.

Nel clustering, la somiglianza tra due oggetti è misurata dalla *funzione di somiglianza*, che misura la loro distanza. Più sarà breve tale distanza e più sarà alta la somiglianza, e viceversa.

Classificazione e clustering sono i metodi usati nel data mining per analizzare i set di dati e dividerli, sulla base di opportune regole. La classificazione categorizza i dati con il supporto di dati di training. D'altra parte, il clustering utilizza diverse misure di somiglianza per categorizzare i dati.

Esistono tanti algoritmi per realizzare i due tipi di apprendimento sopra introdotti, ma nelle prossime sezioni verranno descritti solo quelli da noi utilizzati.

Prima di introdurre il clustering e la classificazione introduciamo il set di dati utilizzato in queste fasi. In particolare, è stato realizzato l'`append()` dei due set di dati riguardanti le promozioni (anno 2020 e 2021) e quelli delle vendite, in modo tale da avere solo due dataset. Nel dataset delle vendite sono state considerate solo le righe che, come documento, avevano una fattura, tralasciando addebiti e accrediti (per vedere come le promozioni hanno incrementato le vendite ci interessano, infatti, le fatture). Successivamente, è stato realizzato un `merge` tra i dati delle vendite e quelli delle promozioni. Lo scopo è quello di avere l'informazione utile del fatturato di una determinata riga, insieme all'informazione della possibile promozione attiva, per il determinato committente, alla data dell'acquisto.

Per realizzare il merge è stata utilizzata la funzione `merge(df1,df2,on ='campi in comune', how='tipo join')` di *Pandas* che, oltre ai dataset passati come parametri e ai campi, richiedeva, anche, di specificare il tipo di join; nel nostro caso, è stato realizzato un *left join* tra le vendite e le promozioni.

Il merge è stato possibile grazie alla presenza, in entrambi i dataset, dei campi `ID Referenza` e `IDCommittente`; in particolare, il campo `Data Rif` del dataset delle vendite doveva ricadere all'interno del periodo di *sell-in* presente nel dataset delle promozioni. Il risultato è un DataFrame di 153948 righe.

Su questo dataset verranno applicate le tecniche di clustering e classificazione.

5.2 Clustering

Come introdotto in precedenza, il processo di clustering viene utilizzato quando non ci sono le etichette e si vogliono dividere i dati in più cluster. Nel nostro caso, questo processo è stato necessario in quanto il nostro dataset era privo di etichette e, per riuscire a realizzare una successiva classificazione, è stata necessaria questa fase.

L'algoritmo utilizzato è *K-means*, che è uno degli algoritmi di clustering più diffuso e più "performante". Tale algoritmo si basa sui seguenti passi:

1. Non conoscendo le classi presenti nel dataset di ingresso, la prima cosa da fare è decidere il numero di cluster in cui si vuole suddividere il dataset. Questo numero è detto k, da cui il nome del metodo *K-means*. Il termine *means* sottintende l'uso dei centroidi, cioè i punti medi.
2. Si scelgono in modo casuale K centroidi appartenenti allo spazio delle feature passate all'algoritmo. L'unica condizione è che questi non siano coincidenti, in quanto, l'algoritmo potrebbe avere problemi a convergere.
3. Si calcola la distanza di ogni punto del dataset rispetto ad ogni centroide.

4. Ogni punto del dataset viene assegnato al cluster del centroide più vicino.
5. Si calcola la nuova posizione dei centroidi facendo la media delle posizioni di tutti i punti associati ai vari centroidi.
6. Si itera ricorsivamente dal punto 3 fino a quando non cambia più la posizione finale dei centroidi.

Per facilitare la scelta del numero K esiste un metodo che prende il nome di *elbow method*, o metodo del gomito. Questo itera l'algoritmo *K-means* per diversi valori di k ed, ogni volta, calcola la somma delle distanze al quadrato tra ogni centroide ed i punti del proprio cluster. Successivamente, crea un grafico con i risultati e sulla base del quale si può scegliere facilmente il numero da assegnare a k.

La sfida più difficile è la scelta di che feature passare all'algoritmo per effettuare il clustering. Nel nostro caso il dataset era ricco di stringhe e valori categorici. Gli unici dati a disposizione erano le varie quantità ed il fatturato. L'idea di partenza era quella di andare a vedere quanto una promozione ha influenzato le vendite; per fare questo la colonna del fatturato ha giocato un ruolo fondamentale.

Per facilitare la comprensione, quando viene nominato il dataset "merge", facciamo riferimento al dataset spiegato precedentemente, mentre dataset "vendite" si riferisce al dataset delle vendite in cui ci sono solo documenti di tipo fattura.

Per effettuare il clustering è stata aggiunta un colonna al dataset, che rappresenta un indice da noi calcolato. Tale indice indica quanto una determinata referenza è stata venduta grazie alle promozioni. A tal fine, sono state analizzate tre possibili strade, ovvero:

1. Calcolare il fattore come la divisione tra tutto il fatturato di una determinata referenza all'interno del dataset merge e il fatturato della stessa referenza all'interno del dataset delle vendite. Così facendo non consideriamo il calendario e il tema. Più il numero è elevato e più le promozioni hanno influenzato le vendite.
2. Considerare anche il calendario. Quindi, quando si trovava il fatturato totale di una determinata referenza all'interno del dataset merge, questo dipendeva anche dal calendario in cui si trovava la referenza. Così facendo consideriamo quanto i vari calendari hanno contribuito alla vendita.
3. Considerare anche i temi all'interno dei calendari.

Una volta realizzato tale indice è stata passata all'algoritmo solo la colonna contenente quest'ultimo e scegliendo come k il numero 5. Questa scelta è stata presa sulla base del metodo del gomito, che ci indicava di scegliere un numero compreso tra quattro e sei.

Il codice per eseguire l'algoritmo *K-means* è molto semplice: basta importarlo e utilizzare la funzione apposita di `fit()` dell'algoritmo, che riceve i dati come parametro e restituisce un array, all'interno del quale ci sono le etichette calcolate. Tali etichette sono state, successivamente, aggiunte al dataset merge. Così facendo, abbiamo ottenuto le etichette di cui avevamo bisogno per poter realizzare la classificazione.

5.3 Classificazione

Grazie alle etichette calcolate durante il clustering, possiamo implementare i metodi di classificazione. Prima di fare questo si devono decidere quali campi devono essere utilizzati per effettuare tale attività. Nel nostro caso, i campi del dataset di merge che sono stati utilizzati per la classificazione sono riportati in Tabella 5.1. Si può notare che il numero di campi è inferiore rispetto a quello totale; i campi scelti sono quelli che sono principalmente inseriti prima di inserire la referenza finale quando viene inserito un nuovo tema.

Campo	*Definizione*
`Formato`	Formato della referenza.
`Segmento`	Segmento della referenza.
`Occasione d'Uso`	Occasione d'uso della referenza.
`Target`	Target della referenza.
`Tipo Promozione`	Tipo della promozione (Taglio prezzo, Volantino o Promo Web).
`Durata`	Numero intero che indica la durata del *sell in* in giorni.
`Quadrimestre`	Indica a quale quadrimestre appartiene la determinata riga del dataset. Tale colonna è stata ricavata dalla colonna `Data Rif` e serviva per portarli dietro anche l'informazione del periodo dell'anno nella classificazione.
`Cluster`	Indica l'etichetta della riga calcolata con l'algoritmo *K-means*.

Tabella 5.1. Tabella dei campi utilizzati per la classificazione

Successivamente, come detto in precedenza, la classificazione utilizza un apprendimento supervisionato; per questo è stato necessario dividere i dati in *train e test set*. Il test set serve per verificare l'accuratezza del modello, mentre il train set serve per allenare il modello. Per realizzare la divisione è stato utilizzato il metodo `train_test_split("dataset", test_size = "percentuale")` che, grazie alla percentuale passata al parametro `test_size`, permette la divisione del dataset.

Un altro problema da affrontare riguarda il fatto che le colonne considerate sono anche stringhe, e gli algoritmi si aspettano numeri; per questo è stata utilizzata la funzione `sklearn.preprocessing.LabelEncoder().fit("Campo")`, che codifica i dati dei campi passati in numeri. Si ha, però, sempre la possibilità di tornare ai valori originali grazie al metodo `inverse_transform()`.

Una volta preparato il dataset, sia quello di train che di test, per effettuare la classificare sono stati utilizzati sia gli *alberi di decisione* semplici, sia il *Random Forest*, e sia, infine, il metodo *K-Nearest Neighbors*.

Gli alberi di decisione sono un modo di classificare i dati in un numero finito di classi, in maniera molto semplice e intuitiva. Tali alberi vengono costruiti dividendo (grazie a condizioni di *split*) ricorsivamente il training set in modo da formare sottoinsiemi omogenei rispetto alla variabile di classificazione (nel nostro caso è il valore della colonna `Cluster`) per creare un modello. All'interno dell'albero, i nodi sono etichettati con i possibili nomi degli attributi, gli archi indicano i valori

possibili di questi, mentre le foglie denotano la classe di appartenenza. Un nuovo oggetto è classificato, grazie al modello creato, in base alla posizione che prende nell'albero grazie ai valori dei suoi attributi.

Random Forest, invece, sfrutta più alberi di decisione per ottenere un risultato finale migliore rispetto ai singoli alberi. In particolare, la predizione della classe di un nuovo dato, viene effettuata prima su tutti gli alberi per poi decidere la classe finale in base a tutti i risultati. Si può capire che è molto più complesso computazionalmente, ma, in generale, fornisce risultati migliori.

Infine, l'algoritmo *K-Nearest Neighbors*, considera la somiglianza tra il dato che si vuole classificare e i k elementi ad esso più vicini. In particolare, viene assegnata la classe dominante nei k elementi (per questo k deve essere preferibilmente dispari, per facilitare la decisione).

I parametri migliori da passare agli algoritmi sono stati scelti grazie alla funzione `GridSearchCV()` che aiuta a scegliere quelli che portano ad avere l'accuratezza migliore (un esempio è il k dell'algoritmo Nearest Neighbors, la profondità massima negli alberi di decisione, etc). Nel capitolo successivo verranno mostrati i risultati dei classificatori utilizzati, basati sui tre risultati del clustering.

6

Risultati della campagna di Data Analytics

In questo capitolo vengono riportati i risultati ottenuti durante la fase di classificazione.

6.1 Risultati

Nel capitolo precedente abbiamo descritto come è stato impostato il processo di classificazione. In questo capitolo ci concentriamo sui risultati ottenuti. In particolare prima di vederli, è stato riportato, nelle Tabelle 6.1-6.3, come l'algoritmo *K-means* abbia assegnato i dati del dataset ai cluster durante la fase di clustering nei tre casi analizzati. Viene riportata la media dell'indice creato, e, in base a questa, si è scelto di assegnare le etichette ai vari cluster per comprendere meglio il risultato. In particolare, come detto, l'indice denota quanto le promozioni abbiano influenzato le vendite di una determinata referenza; per questo le etichette vengono assegnate in base a tale contributo. Tali etichette sono: *Molto basso, Basso, Normale, Alto e Molto Alto*. Da notare che, nelle tre tabelle, i valori delle medie cambiano, e questo è normale perché abbiamo una granularità dei dati diversa nel calcolo dell'indice. Anche il numero di elementi assegnati ad ogni cluster cambia, e ciò è dovuto al fatto che cambiano i valori l'output dell'algoritmo è diverso.

Cluster	*Numero elementi*	*Media indice cluster*
`0-Molto basso`	16778	0.152029
`1-Basso`	50494	0.328277
`2-Normale`	30972	0.448316
`3-Alto`	43361	0.583000
`4-Molto Alto`	12343	0.778863

Tabella 6.1. Tabella dei valori dei cluster dove non vengono considerati i calendari e i temi

Nelle Figure 6.1-6.3 viene illustrato il *report di classificazione* del metodo *K-Nearest Neighbors* per i tre casi in analisi. In tutti e tre i casi il k usato dall'algoritmo

Cluster	*Numero elementi*	*Media indice cluster*
0-Molto basso	89702	0.011079
1-Basso	36429	0.049032
2-Normale	13727	0.110725
3-Alto	9577	0.234351
4-Molto Alto	4513	0.592500

Tabella 6.2. Tabella dei valori dei cluster dove vengono considerati i calendari, ma non i temi

Cluster	*Numero elementi*	*Media indice cluster*
0-Molto basso	134172	0.003960
1-Basso	14229	0.032466
2-Normale	4439	0.093334
3-Alto	790	0.198446
4-Molto Alto	318	0.423061

Tabella 6.3. Tabella dei valori dei cluster dove vengono considerati anche i calendari che i temi

è 19, suggerito dalla funzione `GridSearchCV()`. Nel report sono riportate seguenti metriche:

- *Precision*: è il rapporto tra le osservazioni positive previste correttamente e il totale delle osservazioni positive previste;
- *Recall*: è il rapporto tra le osservazioni positive correttamente previste e tutte le osservazioni nella classe effettiva;
- *F1 score*: è la media ponderata di Precisione e Recall;
- *Accuratezza*: è il rapporto tra le osservazioni previste correttamente e il totale delle osservazioni.

Si può notare che, se analizziamo l'accuratezza, il caso migliore è quello in Figura 6.3, in cui consideriamo sia il calendario che il tema per realizzare il calcolo dell'indice. L'accuratezza, in questo caso, è pari a 0.8874, contro 0.7482 della Figura 6.1, e 0.6937 della Figura 6.2. In Figura 6.3 si nota che i risultati ottimi sono ottenuti per la classe *Molto basso*, ciò è dovuto al fatto che le righe del dataset con tale etichetta sono molte (134172 su 153948 totali), e ciò può influenzare la classificazione. Se consideriamo tutte le metriche insieme, il caso migliore è mostrato in Figura 6.1, dove la Precision, il Recall e l' F1 score hanno valori abbastanza elevati.

Per quanto riguarda gli alberi di decisione sono stati riportati sia quelli realizzati attraverso l'utilizzo dell'albero semplice di decisione (Figure 6.4-6.6), sia quelli realizzati attraverso l'utilizzo di Random Forest (Figure 6.7-6.9) per i tre casi in analisi. Da notare che la profondità dell'albero in tutti i casi è pari a 3. La funzione `GridSearchCV()` ha fornito un valore di profondità pari a 30, ma per motivi di visualizzazione gli alberi sono stati rappresentati con tre livelli. La condizione di split di ogni nodo è binaria e il colore associato a questo indica la classe assegnata al dato in base ai valori degli attributi dei dati presenti in esso.

L'accuratezza di questi alberi segue molto quella dell'algoritmo *K-Nearest Neighbors*, come si può notare nella Tabella 6.4. L'accuratezza del Random Forest nei

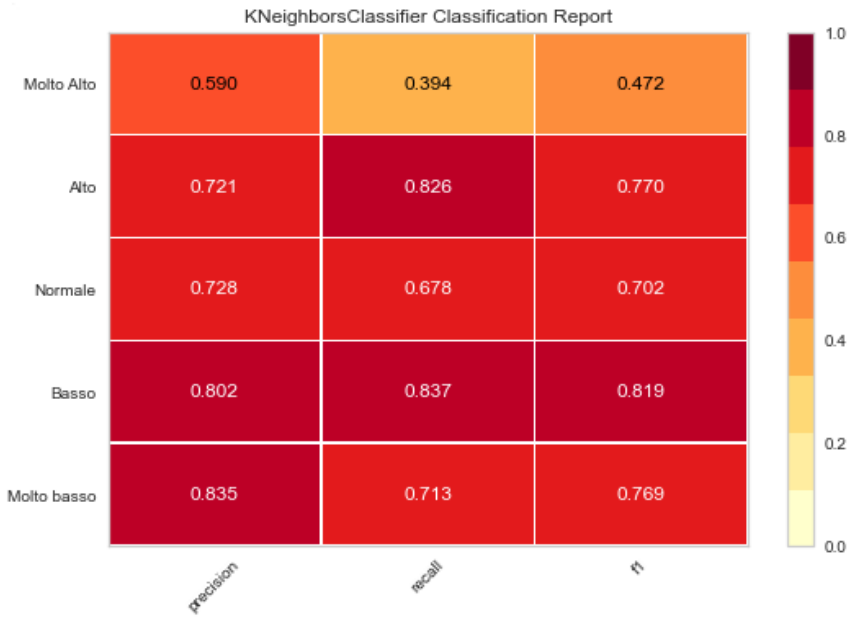

Figura 6.1. Report di classificazione del *K-Nearest Neighbors* con l'accuratezza pari a 0.7482

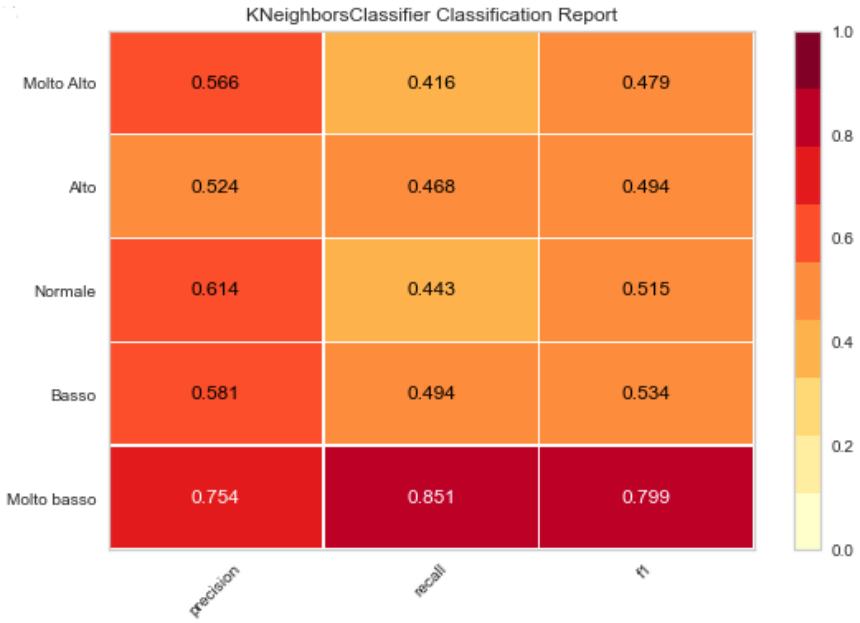

Figura 6.2. Report di classificazione del *K-Nearest Neighbors* con l'accuratezza pari a 0.6937

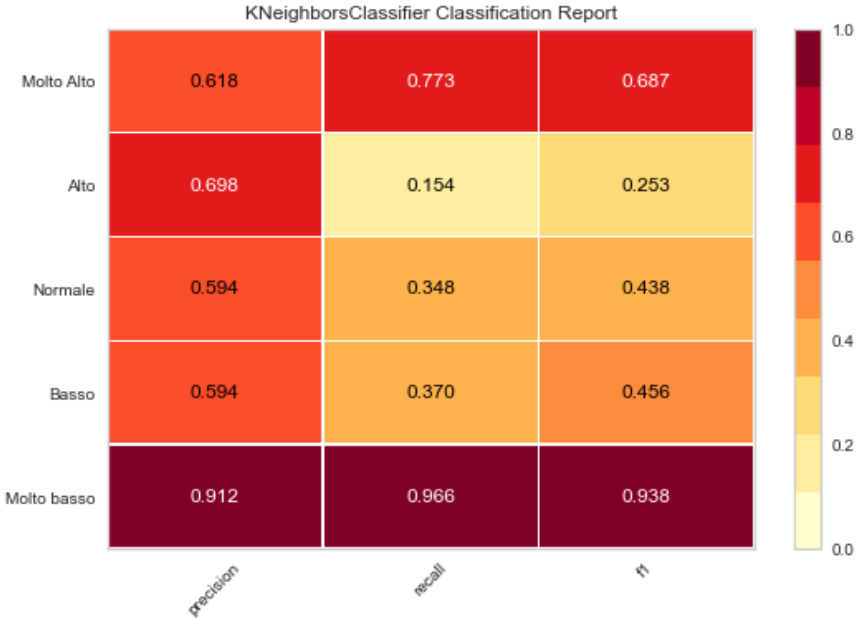

Figura 6.3. Report di classificazione del *K-Nearest Neighbors* con l'accuratezza pari a 0.8874

tre casi è lievemente migliore rispetto all'albero semplice di decisione. Nelle Figure 6.4 6.7, si può notare che, sebbene si utilizzino solo tre livelli, sono possibili più classi. Invece, nelle altre figure (Figure 6.5,6.6,6.8 e 6.9), domina una sola classe a causa dei pochi livelli.

Metodo	*Caso 1*	*Caso 2*	*Caso 3*
`K-Nearest Neighbors`	0.7482	0.6937	0.8874
`Alberi decisione semplici`	0.7563	0.7037	0.8897
`Random Forest`	0.7570	0.7040	0.8900

Tabella 6.4. Tabella delle accuratezze dei metodi di classificazione

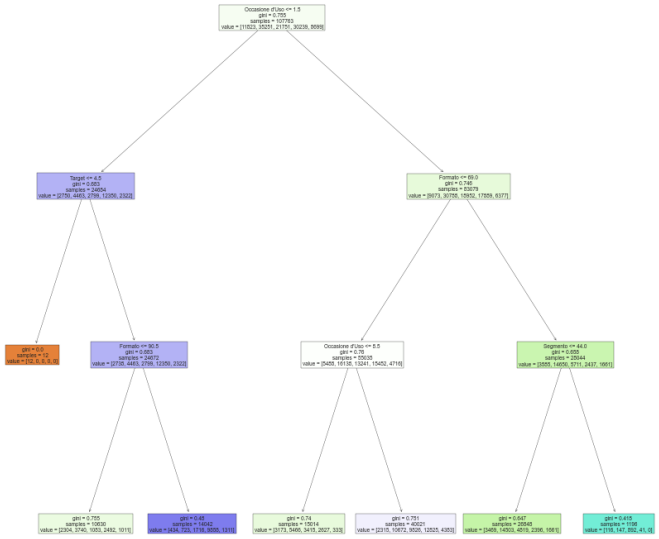

Figura 6.4. Caso 1: Albero di decisione con l'accuratezza pari a 0.7564

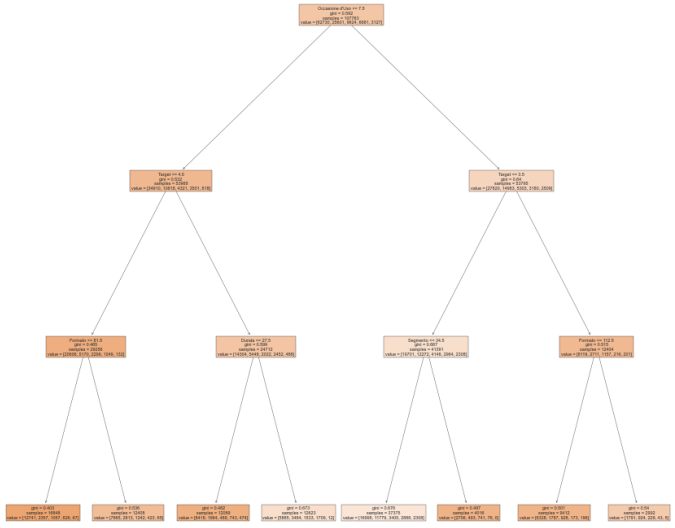

Figura 6.5. Caso 2: Albero di decisione con l'accuratezza pari a 0.7037

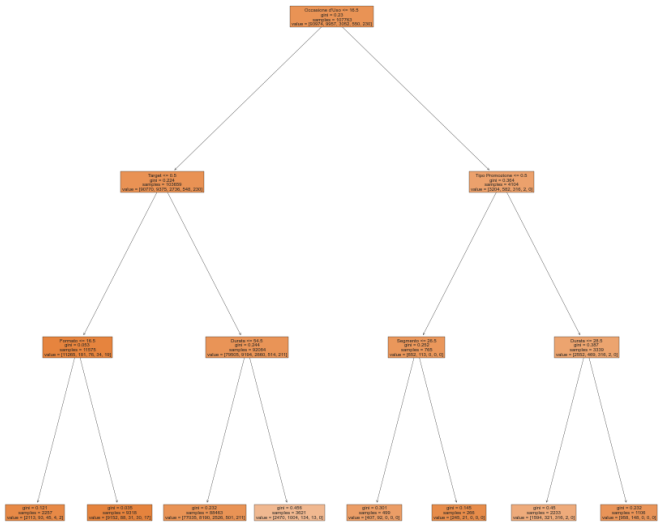

Figura 6.6. Caso3: Albero di decisione con l'accuratezza pari a 0.8897

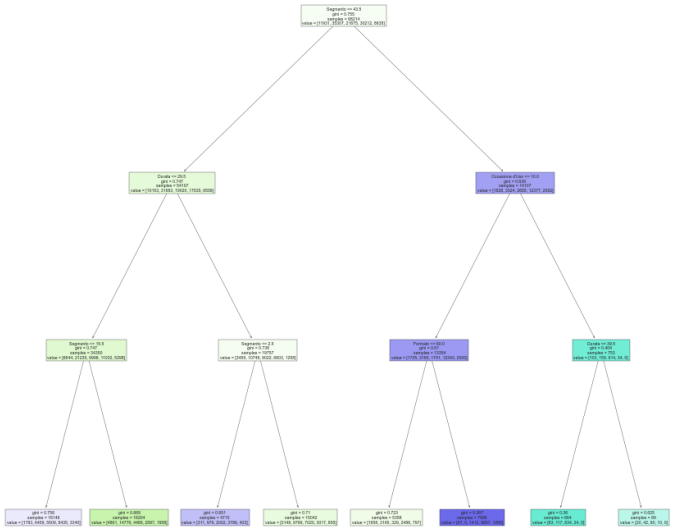

Figura 6.7. Caso 1: Albero di decisione realizzato tramite Random Forest con l'accuratezza pari a 0.7570

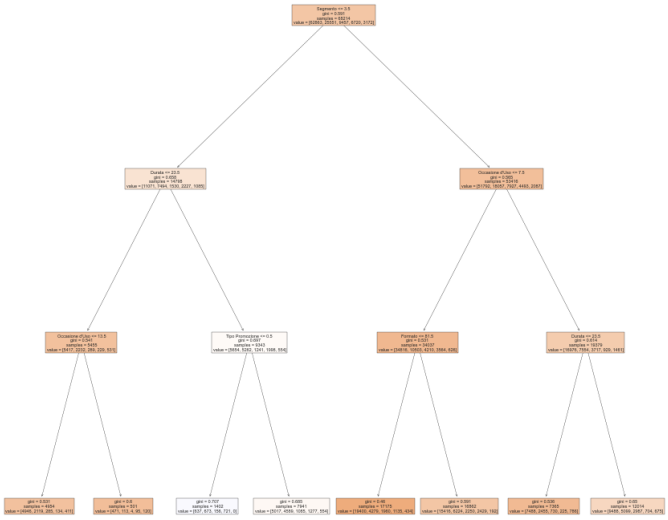

Figura 6.8. Caso 2: Albero di decisione realizzato tramite Random Forest con l'accuratezza pari a 0.7039

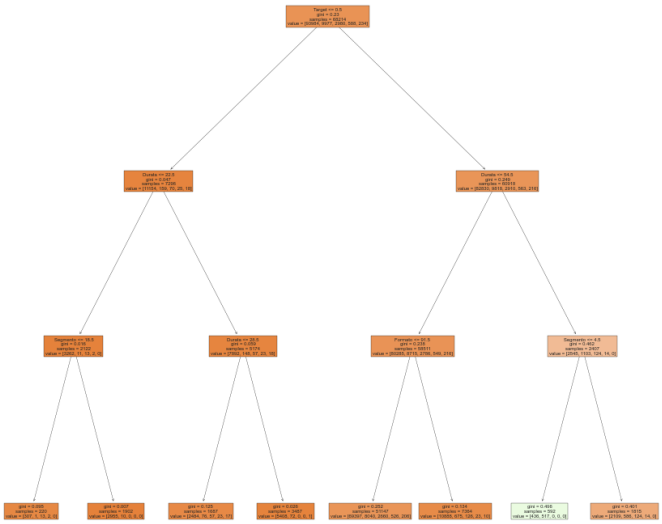

Figura 6.9. Caso 3: Albero di decisione realizzato tramite Random Forest con l'accuratezza pari a 0.8900

7

Conclusioni

In questo lavoro di tesi sono stati analizzati i dati relativi alle promozioni e alle vendite dell'azienda Fater. In una prima fase sono state necessarie delle call con dei responsabili dell'azienda, che ci hanno descritto come vengono gestite le promozioni presso la loro azienda. Una volta conclusa questa parte ed ottenuti i dati dall'azienda, si è svolta una fase di pulizia di questi ultimi per renderli adatti alle analisi successive.

In seguito, è stata eseguita una fase di esplorazione dei dati, per comprenderli più a fondo, aiutandoci anche con l'utilizzo di alcuni grafici realizzati che riportavano il contenuto più importante. L'obiettivo finale è stato quello di capire quanto le promozioni hanno influenzato le vendite; per fare questo, è stata eseguita un'attività di clustering seguita da un'attività di classificazione. Per effettuare il clustering sono state considerate le vendite dei prodotti effettuate, grazie a promozioni attive, in funzione delle vendite totali di quei determinati prodotti, tramite l'utilizzo di un indice. Sono state analizzate diverse strade, in base a come veniva realizzato tale indice. Una volta individuate le caratteristiche di ciascun cluster ottenuto, la successiva attività di classificazione è stata più semplice.

I risultati ottenuti sono abbastanza buoni, anche se migliorabili. Il lavoro descritto sulla tesi non può essere considerato un punto di arrivo; infatti ci si sta sforzando ancora a trovare soluzioni migliori nella scelta dell'indice utilizzato per realizzare il clustering, essendo questa la fase più delicata di tutto l'approccio. Sono state provate diverse strade per confrontare i risultati, ma, come sappiamo, ogni volta che si prova una nuova strada e si ottengono dei cluster, l'operazione più difficile è quella di riuscire a comprendere il significato di questi, e ciò non è sempre possibile.

Riferimenti bibliografici

1. Christopher Cheng. *Python*. Walker Books Australia, 2021.
2. Kenneth E. Clow and Donald Baack. *Integrated advertising, promotion, and marketing communications*. Pearson, 2014.
3. Douglas J. Dalrymple and Leonard J. Parsons. *Basic marketing management*. Wiley, 2000.
4. Michael Heydt. *Learning pandas: high perfomance data manipulation and analysis using Python*. Packt, 2017.
5. Ihab F. Ilyas and Xu Chu. *Data Cleaning*. Association for Computing Machinery, 2019.
6. John D. Kelleher and Brendan Tierney. *Data science*. The MIT Press, 2018.
7. Ron Kenett and Thomas C. Redman. *The real work of data science: turning data into information, better decisions, and stronger organizations*. John Wiley & Sons, Inc., 2019.
8. Stephen Klosterman. *Data science projects with Python: a case study approach to successful data science projects using Python, pandas, and scikit-learn*. Packt, 2019.
9. Mark Lutz. *Programming Python*. O'Reilly, 2011.
10. Mark Lutz. *Python*. Ed. O'Reilly, 2014.
11. Sandya Mannarswamy. *Data science: learn the what, where, and how of data science*. Apress, 2014.
12. Wes McKinney. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media, Inc., 2nd edition, 2017.
13. Wes Mckinney. *Python for data analysis: data wrangling with pandas, numpy, and ipython*. O'Reilly Media, 2017.
14. Ryszard S. Michalski, Jaime G. Carbonell, and Tom M. Mitchell. *Machine learning: an artificial intelligence approach*. Tioga Pub. Co., 1983.
15. Tom M. Mitchell. *Machine learning*. McGraw Hill, 2017.
16. Fabio Nelli. *Python Data Analytics: With Pandas, NumPy, and Matplotlib*. Apress, 2018.
17. Sonya Newland. *Machine learning*. Wayland, 2021.
18. Garreta Raul, Guillermo Moncecchi, Trent Hauck, and Gavin Hackeling. *Scikit-learn: machine learning simplified*. Packt Publishing, 2017.
19. Jesus Rogel-Salazar. *Data Science and Analytics with Python*. Chapman & Hall/CRC, 1st edition, 2017.
20. William Vance. *Data science*. William Vance, 2020.

Ringraziamenti

Il primo ringraziamento va a mia mamma, che mi ha dato la possibilità, per nulla scontata, di intraprendere e portare a termine questo percorso universitario, sostenendomi nei momenti di difficoltà.

Un altro importantissimo ringraziamento va ad Eliana, la mia ragazza, che mi è stata sempre vicina e mi ha supportato ed aiutato.

Vorrei ringraziare il Prof. Ursino, che mi ha accompagnato nel lavoro di tesi seguendomi costantemente con estrema disponibilità, e il dottorando, Gianluca Bonifazi, che mi ha aiutato nella parte implementativa del lavoro.

Ringrazio di cuore i miei compagni di studio: Michele, Albo, Gatto, Kevin, Lorenzo e Chiara, per il supporto che mi hanno fornito in questi anni. Soprattutto ringrazio Michele, che, nonostante la pandemia, è stato sempre disponibile per studiare e superare i vari ostacoli insieme. Un ulteriore ringraziamento va ad Albo, perché grazie a lui sono riuscito a finire il percorso universitario in forma, con gli allenamenti eseguiti insieme anche a distanza.