POLITECNICO DI TORINO

Corso di Laurea Magistrale in Ingegneria Gestionale

Tesi di Laurea Magistrale

STUDIO E SPERIMENTAZIONE DELLE METRICHE DI GESTIONE WEB ANALYTICS DEI SITI CORPORATE DI UN'AZIENDA BROADCASTER TELEVISIVO

Relatore

Chir.ma Prof.ssa
Tania CERQUITELLI

Candidato

Alessandro BIANCHI

Anno Accademico 2020/2021

Indice

INTRODUZIONE

"The story of how data scientists became sexy is mostly the story of the coupling of the mature discipline of statistics with a very young one-computer science." Questa citazione di Gil Press (2013), è utile per introdurre gli obiettivi perseguiti nella presente tesi sperimentale: descrivere un percorso tecnico-sperimentale di applicazione di Data Science ad una utenza industriale, basato proprio sulla integrazione di dati statistici e informatici.

Il termine "Data Science" è recente e specifica una branca della scienza che dovrebbe dare un senso agli sconfinati archivi di Big Data; procedura studiata da anni, sia in letteratura scientifica, sia in applicazioni industriali.

L'evoluzione del termine "Data Science" e la contestualizzazione delle sue branche è fondamentale per comprendere quali siano le possibili applicazioni nella realtà produttiva. Infatti, uno degli obiettivi della ricerca informatica è quello di fornire risultati che, dal rigore della sperimentazione scientifica, consentano di passare alle applicazioni in scala più ampia e in settori più speculativi, in cui si utilizzino tali risultati scientifici per avviare e supportare il Business; è proprio nel settore del Business Intelligence che si inseriscono le società di consulenza, le quali si propongono di sfruttare le conoscenze scientifiche, al fine di migliorare o supportare la transizione tecnologica che si posta dietro l'ondata del progresso. In questa realtà di transizione, ad oggi, migliaia di imprese sono portate ad effettuare aggiornamenti degli strumenti di analisi, talvolta per seguire il progresso, talvolta per migliorare le strategie di mercato e di produzione.

Nel presente lavoro di tesi sono analizzate le fasi di un progetto in tema di Analytics, reso necessario da un prodotto in procinto di andare in "End-of-Support", termine tecnico che indica la fase di inizio della sua scomparsa dal mercato. La sperimentazione ha riguardato il miglioramento del processo di utilizzo del nuovo applicativo e illustra quanto sia possibile integrare le varie aree di Data Science, applicando tecniche di Machine Learning in ambito Web Analytics.

CAPITOLO I

Evoluzione scientifica e tecnica degli analytics

1.1 Data Science & Data Analytics

Sono disponibili interi archivi di dati per qualsiasi campo, sia strutturati che non ma, visti singolarmente, non possono essere significativi. Nella moltitudine dei dati presenti in questi archivi, risultato della digitalizzazione di tutte quelle notazioni che in precedenza sfruttavano metodi analogici come registri cartacei e documenti, era inevitabile che la società moderna trovasse un modo per trarne valore e sfruttarli al massimo delle loro capacità. È con questa copiosa quantità di dati che si è cominciato a vedere gli strumenti di storage come effettiva fonte di valore, dando così i natali all'epoca dei Big Data. Questo strumento a disposizione del Business necessitava di essere supportato da una scienza che studiasse il dato e imparasse a sfruttarne a pieno il valore a livello speculativo, come strumento di previsione, come fonte di integrazione di informazione.

La Data Science è il processo di espansione della conoscenza applicata a campi come il Business, la finanza, la medicina e l'istruzione, che organizza i dati in modo che possano essere compresi e utilizzati; essa mira a migliorare i processi di sviluppo del prodotto, i processi decisionali e i processi di analisi delle tendenze e di previsione, sfruttando i vari campi dell'analisi dei dati. Di base, quindi, la Data Science copre numerose aree della conoscenza. Sebbene il termine "scienza dei dati" sia apparso per la prima volta all'inizio degli anni '60, le prime conferenze scientifiche su questo argomento si sono svolte alla fine degli anni '90.

Secondo Rachel Schutt e Cathy O'Neil (2013), un Data Scientist è "qualcuno che sa come estrarre significato e interpretare i dati, attività che richiede sia strumenti che metodi di statistica e machine learning, oltre ad essere umano". Al centro della scienza dei dati c'è una profonda connessione con la statistica, sebbene non debba essere ridotta solo ad algoritmi statistici e metodi di elaborazione delle informazioni. Dall'unione di più ambiti della conoscenza, dunque, nasce questa nuova area ibrida che

nel tempo è stata ramificata in diverse branche e discipline, una delle più famose divisioni (Fig. 2).

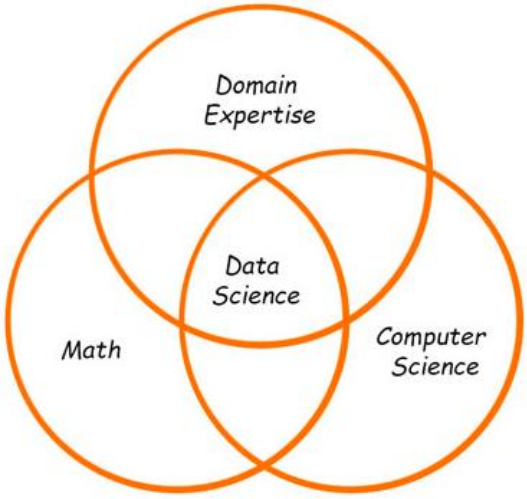

Figura 1. Integrazioni scientifiche che hanno portato alla nascita di Data Science.

Figura 2. Branche e discipline della Data Science.

Le divisioni sopra citate cominciano ad avere barriere sempre più sfocate e capita, molto frequentemente, che le varie sezioni, in sede di progetto, vengano applicate insieme; anche se una non prescinde dall'altra, è la cooperazione delle varie branche che fa della Data Science una scienza molto flessibile dove alle regole prevalgono le "Best Practices". In ogni caso, in concordanza con letteratura scientifica che tende a

differenziare i concetti, di seguito si farà riferimento prevalentemente alla branca dell'Analytics.

L'Analytics è il processo scientifico utile per scoprire e comunicare i modelli significativi che possono caratterizzare i dati. La descrizione più completa della data analytics è la seguente: "L'analisi dei dati è il processo di applicazione sistematica di tecniche statistiche e/o logiche per descrivere, illustrare, condensare, ricapitolare e valutare i dati" (Savenye and Robinson, 2004) Questa area si occupa di trasformare i dati grezzi in informazioni, con il fine di prendere decisioni strategiche; si basa sull'applicazione di statistiche, programmazione informatica e ricerca operativa per quantificare e acquisire informazioni sul significato dei dati. L'Analytics fornisce informazioni significative che potrebbero essere nascoste all'interno di grandi quantità di dati; quindi, è uno strumento che qualsiasi leader, manager o addetto ai lavori può utilizzare, soprattutto in maniera "data-driven".

L'analisi dei dati può fare la differenza, non solo nel mondo del business ma anche nello sport, nell'assistenza sanitaria e in quasi tutti i campi in cui vengono raccolte grandi quantità di dati. Questo strumento porta a realizzare modelli nel mondo che ci circonda, dai comportamenti dei consumatori, alle prestazioni di atleti e squadre, alla ricerca di connessioni tra attività e malattie. Diversamente alla Data Science, che ha un approccio più generale, la Data Analytics è più orientata alle attività ed è questa una visione più ampia dei moderni problemi scientifici e pratici (Tab.1).

Tabella 1. Confronto fra gli aspetti principali di differenziazione di Data Science e Data analytics.

	Data Science	Data Analytics
Type of analysis	Descriptive Analytics + Predictive Analytics	Descriptive Analytics
Scope of analysis	Macro	Micro
Goals of research	Find the right questions to confirm with available math methods	Find valuable data and its business meaning
Fields of application	Machine learning, AI, computer vision, automatic language translation, corporate business analysis	Businesses with on-the-spot data needs, transportation, policy and security, fraud and risk detection, healthcare, energy management, internet search, digital marketing

1.2 Applicazioni degli Analytics

Le applicazioni degli Analytics sono numerose ma il processo che ne identifica l'operato tra i vari ambiti non differisce particolarmente. Tale processo inizia a partire dalle informazioni esistenti, che sono già presenti nell'organizzazione o nel progetto. Viene effettuata una ampia descrizione delle informazioni relative all'area del dominio (contesto). Ciò richiede una profonda conoscenza dell'area disciplinare e la capacità di interpretare i risultati sotto forma di cifre in indicatori specifici del settore in cui opera l'analista. A tal proposito, spesso agli analisti si affiancano figure denominate "esperti di dominio", cioè personaggi con grande esperienza e conoscenza dell'area di applicazione, che spesso hanno il compito che validare il senso scientifico delle soluzioni proposte dagli analyst. Inoltre, è necessario avere la capacità di rappresentare e visualizzare correttamente i dati, in modo che siano comprensibili agli utenti del contesto aziendale.

Gli obiettivi degli Analytics sono: supportare le attività decisionali mediante la pulizia, l'analisi, la trasformazione e la formazione dei dati. Il processo di analisi dei dati consiste essenzialmente in una sequenza di passaggi; ognuno di questi può essere complicato e dispendioso in termini di tempo, a seconda delle problematiche di contesto:

a) si definiscono le domande a cui devono rispondere con i dati;
b) si definisce cosa e come misurare nei dati di input e output;
c) si raccolgono, si preparano e, eventualmente, si normalizzano i dati nel set;
d) si analizzano i dati;
e) si cerca la spiegazione per rende i dati di valore e comunicarli con efficacia.

Qualsiasi tipo di analisi si identifica in uno dei seguenti quattro tipi.

- ✓ *Descriptive analytics*, per descrivere cosa è successo in un determinato periodo di tempo.
- ✓ *Diagnostic analytics*, per scoprire perché si è verificato un evento specifico.
- ✓ *Predictive analytics*, per prevedere qualcosa che è probabile che accada.
- ✓ *Prescriptive analytics*, per suggerire soluzioni data-driven.

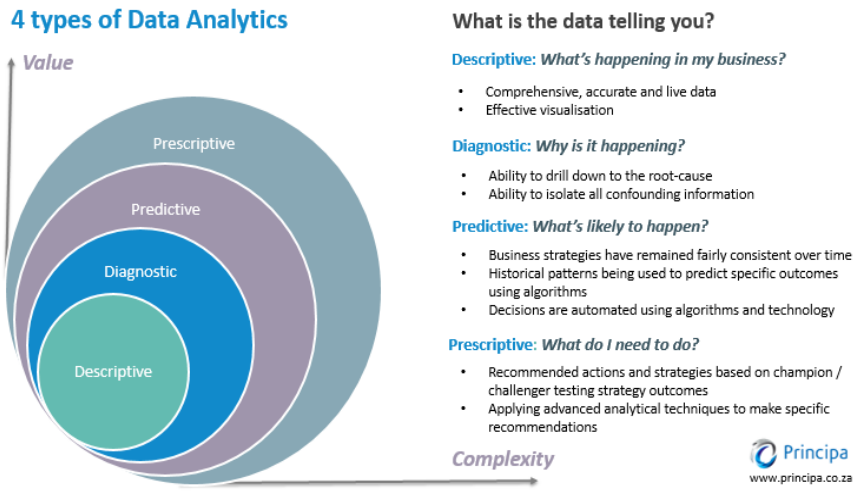

Figura 3. Tipi di Data Analytics.

In maniera più specifica, a seconda del tipo analisi che si effettua, possono essere applicate le seguenti tecniche.

Classification. Nella classificazione, proprietà caratteristiche vengono assegnate a gruppi di oggetti nel set di dati studiato, definite classi; secondo queste proprietà, un nuovo oggetto può essere assegnato a questa o quella classe.

Clustreing. La continuazione logica dell'idea di classificazione; questa attività tende ad associare gruppi di oggetti in base a caratteristiche meno intuitive, come la distanza nello spazio del dato. Tra questi metodi ci sono algoritmi del tipo K-means che studia la distanza tra i centroidi del dato oppure Dbscan che solitamente serve per filtrare gli outliers.

Association rules. Ricercano regole associative, vengono sviluppati modelli tra eventi correlati in un set di dati; esistono delle specifiche soglie che esprimono la significatività delle regole.

Forecasting. Come risultato della risoluzione del problema di previsione, i valori mancanti o futuri degli indici numerici target vengono stimati in base ai dati storici. Spesso utilizzato in logistica e produzione, prevede anche l'utilizzo di Euristiche.

Deviation detection. Rilevare e analizzare i dati più diversi dall'insieme dei dati, individuando i cosiddetti pattern non caratteristici.

Visualizzazione. Come risultato della visualizzazione, viene riportata rappresentazione grafica dei dati analizzati (rappresentazione dei dati in misurazioni 2D e 3D). Tra queste tecniche ricordiamo per esempio i grafici scatter o le heatmaps.
All'interno della Data Analyitics esiste un'ulteriore divisione che identifica le vaie aree funzionali in cui trovano applicazione di Business:

- web analytics;
- fraud analysis;
- risk analysis;
- advertisement and marketing;
- enterprise decision management;
- market optimization;
- market modeling.

Questa tesi si concentrerà sulla Web Analytics, tema che sarà analizzato, non solo dal punto di vista del cliente, che ha si interfaccia con la tematica Web, ma anche del consulente che ha il compito di progettare il sistema per costruirlo; è interessante immaginare che la conoscenza del consulente, affinché sia efficace, deve estendersi ad aree ancora più vaste che, teoricamente, non dovrebbero entrare in questi temi. Componenti Architetturali, Sistemi e Software devono essere temi ben chiari a chi propone il suo intervento in termini di consulenza, questo perché saranno gli attori che costruiranno i sistemi per realizzare gli Anlaystics.

1.3 Web Analytics

La Web Analytics (WA, anche Digital Analytics) è oramai fondamentale per il business digitale e il successo dei servizi online di aziende sia pubbliche che private. Con questo termine, si intende l'analisi di tutte quelle informazioni che riguardano gli utenti che visitano il sito web. Precisamente, c'è una definizione coniata dalla WAA (la Web Analytics Association, poi diventata Digital Analytics Association) che considera i Web Analytics come «la misurazione, la collezione, l'analisi e il reporting di dati internet allo scopo di capire e ottimizzare l'utilizzo del Web».
Web Analytics è la scienza e l'arte di migliorare i siti web aumentare la propria redditività migliorando l'esperienza del cliente sul sito web. È una scienza perché usa le statistiche, tecniche di data mining e un processo metodologico ed è un'arte perché,

come un brillante pittore, l'analista o il marketer può attingere da una gamma di colori diversificata (fonti di dati) per trovare il mix perfetto che produrrà intuizioni fruibili. È anche un'arte perché migliorare i siti web richiede un profondo livello di creatività, bilanciamento del design incentrato sull'utente, promozioni, contenuti, immagini, e molto altro. Inoltre, l'analista cammina sempre sulla linea sottile tra progettista di siti web, personale IT, esperto di marketing. Ormai, i gestori di siti web sono consapevoli che l'acquisizione dei visitatori è uno sforzo multiforme, che si avvale di tecniche del tipo: e-mail, posta, marketing di affiliazione e, naturalmente, ricerca. Un passo alla volta hanno acquisito know-how a trovare il visitatore giusto da portare ai loro siti web. Ad esempio, ogni il sito web ora ha una strategia di ottimizzazione dei motori di ricerca (SEO) che li aiuta a posizionarsi in alto nei risultati dei motori di ricerca. Allo stesso modo sono anche consapevoli che Pay-Per-Click (PPC) le campagne possono essere efficaci nell'indirizzare visitatori. Nel corso del tempo, gli strumenti di Web Analytics sono diventati sempre più importanti e sofisticati per l'analisi del comportamento degli utenti. Non a caso si parla di Big Data, proprio per sottolineare che i dati sono diventati sempre più numerosi e di difficile interpretazione, che possono trasformarsi in informazioni da cui indirizzare preziose strategie di marketing. «Tramite web analytics si possono analizzare con un ottimo grado di dettaglio le proprie piattaforme digitali, effettuando benchmark con i dati provenienti dal mercato o con altre organizzazioni e siti del settore [Federico della Bella, Associated Partner P4I – Digital Customer Experience Practice -]. Fare web analytics significa identificare le dimensioni chiave di performance delle proprie piattaforme web, del proprio eCommerce, delle proprie web-app e analizzarne i risultati, effettuando confronti e scoprendo come si comportano e come fruiscono dei servizi digitali i propri utenti. Il paragone con i "best in class", anche tra-settore, costituisce uno stimolo potente all'innovazione e al miglioramento». Web Analytics non è solo una tecnologia per produrre report; è un processo che propone un circolo virtuoso per il sito web ottimizzazione. Esiste infatti una differenza sostanziale tra statistiche del sito web e Web Analytics, le prime misurano il traffico che passa sul sito web; le seconde, dovrebbero spiegare quello che succede sul sito web. Quindi, le statistiche dovrebbero cogliere l'aspetto quantitativo, mentre gli analytics dovrebbero aiutare a fornire un'analisi qualitativa del dato. Dopo l'analisi digital Analytics, per le aziende con un grande numero di transazioni, può

essere molto utile dotarsi di programmi di Business Intelligence che permettono un'analisi grafica e analitica molto dettagliata del business, e facilitano diverse tipologie di analisi.

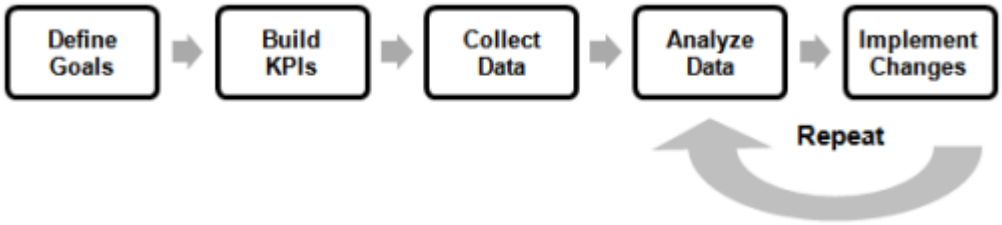

Figura 4. Web Analytics Framework.

Sulla base delle Best Practices, il processo per analizzare le prestazioni del sito web dovrebbe includere i seguenti passaggi (Fig. 4).

Definizione degli obiettivi. In questa fase si risponde all'interrogativo del perché il sito in questione sia sul Web. Ogni proprietario del sito web deve definire il successo secondo i propri obiettivi e rivederli. Gli obiettivi del sito web sono input critici che aiuteranno a identificare le metriche utili a misurare il successo del canale (per molte aziende, un sito web è visto come tale). Il sito web dovrebbe essere contabilizzato allo stesso modo delle altre spese aziendali; l'investimento deve essere misurato rispetto al rendimento.

Definizione delle metriche (KPI). È possibile misurare il raggiungimento degli obiettivi creando Key Performance Indicators (KPI), che mostrano se il sito web si sta avvicinando o meno ai suoi obiettivi. È conoscenza comune nella comunità di Web Analytics che le informazioni che non generano insight non andrebbero raccolte, ci dovrebbe essere un'azione legata a ciascun KPI proposto per ogni sito web. Ad esempio, se viene misurato il costo di marketing per visitatore di un sito, ci dovrebbero essere due azioni ad esso correlate: una per un calo di questo numero e una per un aumento di esso. Steve Bennett, ex CEO di Intuit, è noto per essere sostenitore dell'identificazione dei "few critics": priorità, obiettivi, metriche, KPI (Schroeder and Taylor, 2003). Tutti probabilmente hanno al massimo tre metriche "critiche" che definiscono la priorità di analisi. Una caratteristica importante di un KPI è la flessibilità: ogni azienda, dipartimento o persona dovrebbe avere i suoi KPI definiti in base all'azienda o ai propri obiettivi e interessi. Una visione comune è quella di avere KPI

trasversali nell'industria divisi in maniera gerarchica: il Senior-management riceve rapporti sul raggiungimento complessivo degli obiettivi del sito web; il middle-management riceve rapporti sulla campagna e sul "sito" e sui risultati di ottimizzazione; gli analisti ricevono informazioni dettagliate e tecniche sulle prestazioni del sito web. Inoltre, ci dovrebbe essere un chiaro allineamento tra gli obiettivi aziendali e ciò per cui ogni livello dell'organizzazione lavorando. Affinché un KPI si possa ritenere rappresentativo dovrebbe contenere quattro attributi:

1) *non complesso*, le decisioni nelle aziende sono prese da persone in diversi reparti con differenti contesti. Se solo l'analista web capisce il KPI, è improbabile che il senior management lo utilizzerà;
2) *rilevante*, ogni azienda è unica, anche le aziende che sembrano essere nella stessa attività. È importante individuare quelli più importanti per il proprio modello di Business;
3) *tempestivo*, le migliori metriche devono essere fornite tempestivamente, in modo che i decisori possano prendere decisioni tempestive: persino ottimi KPI sono inutili se ci vuole un mese per ottenere informazioni in presenza di alta mobilità di settore;
4) *immediatamente utile*, è fondamentale capire rapidamente cosa il KPI è, in modo che si possa trovare il primo rossore di intuizioni non appena lo si guarda.

Un buon esempio di un ottimo KPI che soddisfa tutti i criteri precedenti è la bounce rate (percentuale di single visite di pagine visualizzate). Non è complesso perché è facile da comprendere, spiegare e condividere. È rilevante perché identifica dove si stanno sprecando risorse/liquidità di marketing e quali pagine siano poco performanti. È tempestivo perché è uno standard in tutto Strumenti di Web Analytics ed è subito utile, perché il proprietario del sito web può guardarlo e sapere di cosa ha bisogno di attenzione: avere una pagina con un rimbalzo del 50% significa che il sito ha qualche problema, avere una pagina campagna o parola chiave con una frequenza di rimbalzo del 70% anche.

Raccolta dei dati. È fondamentale che i dati vengano raccolti in modo accurato e salvati su in locale o in cloud per ulteriori analisi. La raccolta dei dati è cruciale per i risultati dell'analisi.

Analisi dei dati. Per comprendere il comportamento del cliente, il l'analista dovrebbe seguire alcuni passaggi iniziali.

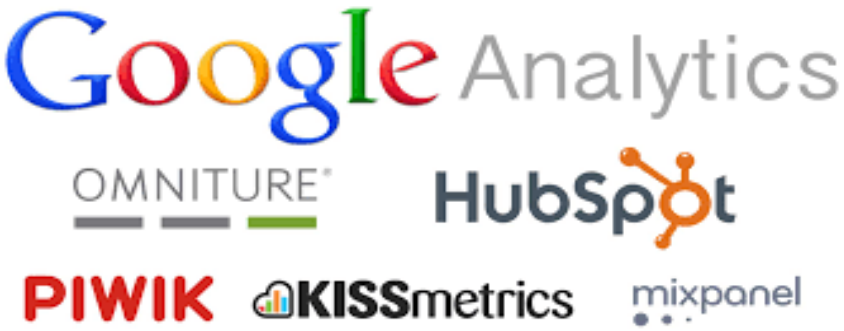

Figura 5. Player nella Web Analytics.

Sia i più esperti che i principianti utilizzano agilmente tool di analytics come Google Analytics, Facebook Analytics e Matomo Analytics, i quali forniscono diversi indicatori che aiutano nel comprendere la parte quantitativa, ovvero cosa sta succedendo. In base al mercato di riferimento e alle esigenze aziendali, le metriche di interesse possono cambiare o è possibile crearne di nuove che inquadrino meglio le caratteristiche del sito che si vuole monitorare.

CAPITOLO II

Prestazioni e Modalità di estrazione e applicazioni dei web analitics

2.1 Tracciamento delle informazioni

Esistono varie modalità di tracciamento dei dati con cui è possibile acquisire le informazioni relative alle visite degli utenti nei vari siti. Tra queste esistono modalità asincrone (log analytics) in cui si acquisiscono i dati e si cariano in sessioni, le quali possono essere separate a livello temporale, oppure sincrone (JavaScript) e danno la possibilità di effettuare tracciamenti in "real time". La prima modalità sarà affrontata è quella dell'analisi dei log.

2.1.1 Analisi Log

Ogni volta che un visitatore di un sito web richiede informazioni, il server del sito registra questa richiesta in un file di log. Il log può avere diversi formati; ELF (Extended Log File Format), è il più comune e salva le seguenti informazioni: indirizzo IP del client che ha richiesto informazioni, data/ora della richiesta, pagina richiesta, codice HTTP, byte serviti, user agent e referrer (Waisberg D. and Kaushik A., 2009). Tuttavia, i log del server in genere non raccolgono informazioni specifiche dell'utente. Vantaggi di questo metodo sono:

- il proprietario del sito Web possiede i dati (al contrario di JavaScript Tagging sotto), ciò significa che il proprietario ha il pieno controllo sulla riservatezza delle informazioni;
- nei registri Web sono disponibili gli storici, questo consente ai gestori del sito web per rianalizzare le campagne passate e rielaborare i dati;
- sono registrati i comportamenti degli Spider (Robot del browser che visitano il sito per indicizzarli e mostrarli nei risultati della ricerca).

I Web I log sono raccolti seguendo la seguente successione operativa (Fig.6):

1. l'utente digita l'URL in un browser;
2. la richiesta arriva a uno dei server web;
3. il server web crea una voce nel file di registro;
4. la pagina torna al cliente.

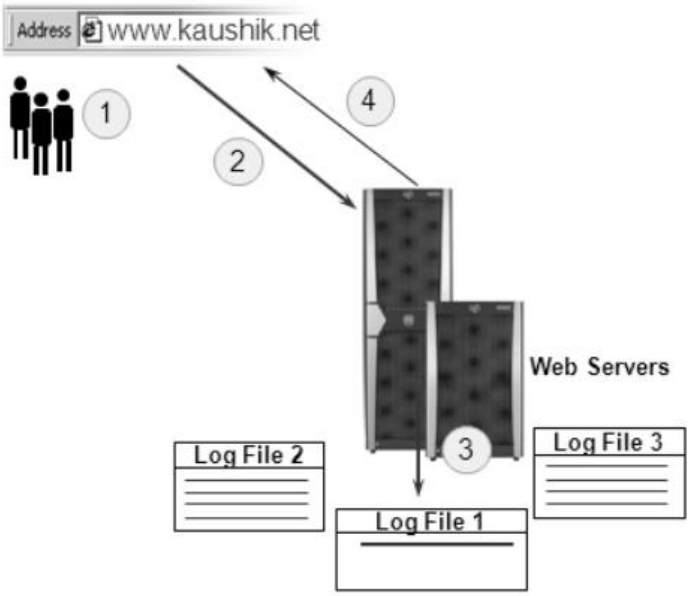

Figura 6. Funzionamento del Log-Tracking.

Alla fine della sessione viene registrato e salvato il log nella sezione predisposta e, in un momento successivo, i file saranno caricati in maniera automatica (con schedulatori) oppure manualmente seguendo le procedure del prodotto utilizzato.

2.1.2 Tag JavaScript

Questa tecnologia consiste nell'inserire piccoli snippet di JavaScript (che non può essere memorizzato nella cache) in ogni pagina di un sito web. Quindi, ogni volta che un visitatore apre una pagina, questo JavaScript è attivato in modo che le informazioni e le azioni del visitatore vengono salvate in un file separato oppure spedite direttamente al database tramite chiamate "push ()" (metodo di trasformazione oggetto in linguaggio JavaScript). I vantaggi di questo metodo sono:

- conta ogni visita su un sito Web (a meno che il cliente non chiuda il pagina prima che lo script venga caricato) mentre i file di log possono essere sviati dalle pagine memorizzate nella cache dal Proxy (il provider di connessione di rete); inoltre il browser, che può inviare una pagina a un visitatore senza registrazione di un file di registro nel server; le informazioni memorizzate in cache vengono perse ogni volta che si analizzano i files log, riducendo l'accuratezza delle informazioni del cliente.

- Il JavaScript non viene letto dagli Spider, che generano elevate quantità di traffico e non sono rappresentativi del comportamento dei clienti. I robot possono essere esclusi dall'analisi; tuttavia, è dispendioso a livello di tempo.
- L'oggetto di analisi è esterno all'azienda, ovvero l'azienda non deve elaborare e salvare i dati internamente.

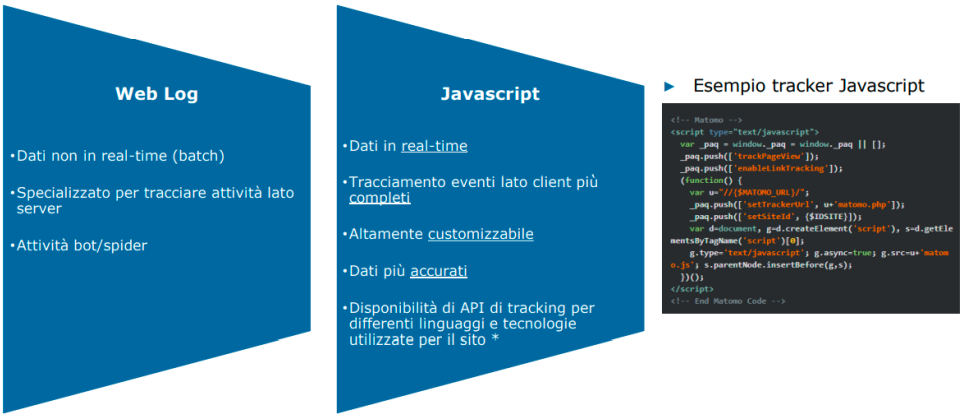

Figura 7. JavaScript e Weblog a confronto.

Di seguito è riportata una successione operativa di come JavaScript (Kaushik, 2007):

1. l'utente digita l'URL in un browser;
2. la richiesta arriva a uno dei server web;
3. il server Web restituisce la pagina insieme ad uno snippet di codice JavaScript aggiunto;
4. quando la pagina viene caricata, esegue il codice JavaScript, che cattura i dettagli sulla sessione del visitatore, i cookies, e li rinvia al server di raccolta;
5. in alcuni casi, al ricevimento della prima serie di dati, il server restituisce il codice aggiuntivo al browser per impostare cookie aggiuntivi o raccoglierne altri dati.

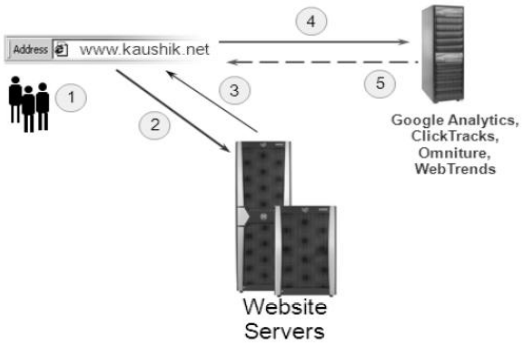

Figura 8. Modalità di tracciamento JavaScript.

2.1.3 Web beacon

Questa tecnologia viene utilizzata per misurare le impressioni dei banner e le sezioni adibite al click degli utenti. Sebbene non vengano utilizzati spesso, può succedere che i web beacon ci siano ancora sul web. Un grande vantaggio dei web beacon è che possono monitorare il comportamento dei clienti in diversi siti web. Di seguito è riportata una successione operativa di come Web beacon vengono raccolti (Fig. 9) (Kaushik, 2007):

1. l'utente digita l'URL in un browser;
2. la richiesta arriva ad uno dei server web;
3. il server Web restituisce la pagina insieme a un "get request" per un'immagine di 1x1 pixel da un server di terza parte;
4. man mano che la pagina viene caricata esegue la chiamata per l'immagine di 1x1 pixel inviando così i dati sulla visione della pagina ad al server terzo.
5. il server terzo, re-invia l'immagine al browser insieme al codice in grado di leggere i cookie e acquisire dati anonimi sul comportamento dei visitatori.

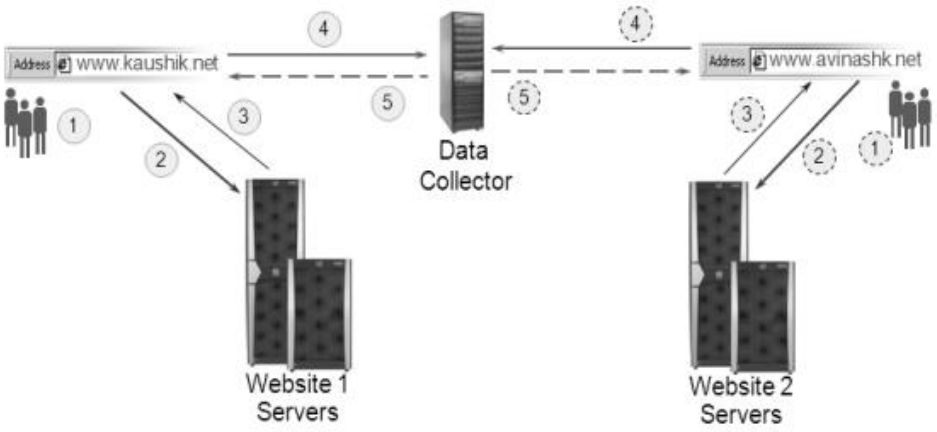

Figura 9. Modalità di tracciamento con Web beacon.

2.1.4 Packet Sniffing

Sebbene lo sniffing dei pacchetti sia molto avanzato in termini di tecnologia, viene utilizzato principalmente per i test multivariati. Suo il più grande vantaggio è che non ha bisogno di taggare le pagine; tutte le informazioni passano attraverso il Packet

sniffer (hardware). Di seguito è riportata una successione operativa di come opera il Packet Sniffing (Fig. 10) (Kaushik, 2007):

1. l'utente digita l'URL in un browser;
2. la richiesta viene spostata al server web passando tramite pacchetti basati su software o hardware sniffer che raccoglie gli attributi della richiesta;
3. il Packet sniffer invia la richiesta al web server;
4. la richiesta viene rinviata all'utente ma prima passa dal Packet sniffer;
5. il Packet sniffer acquisisce informazioni sulla pagina che torna indietro, memorizza i dati e invia la pagina sul browser del visitatore. Qualche venditore applica alle soluzioni di Packet Sniffing un tag JavaScript che può inviare più dati del visitatore al Packet Sniffing.

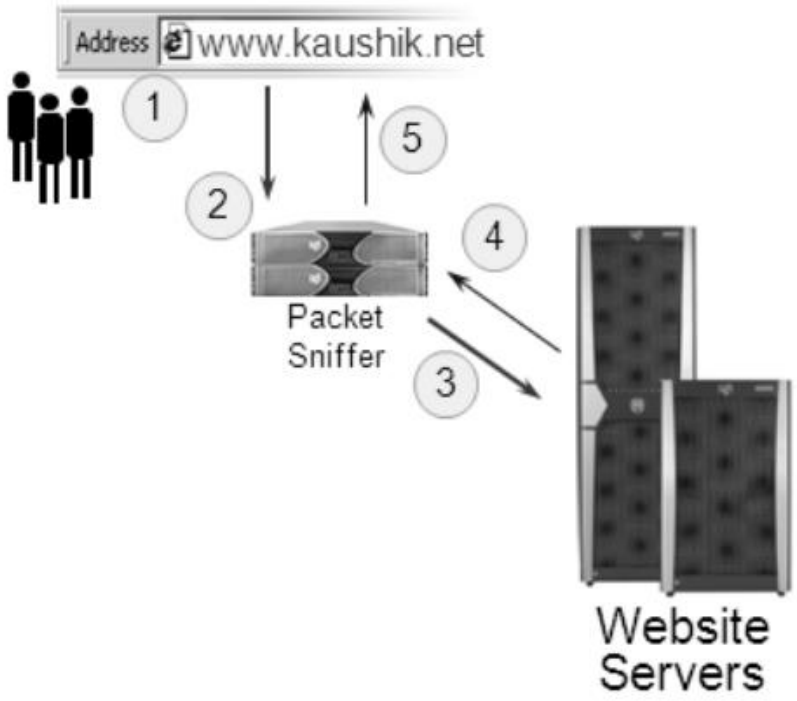

Figura 10. Modalità di tracciamento con Packet sniffer.

Questa ultima soluzione, generalmente, è meno commerciale ed utilizzata in ambiti diversi dall'Analytics. Bisogna precisare che le soluzioni maggiormente in uso sono le prime due ed è bene tenere presente che i maggiori utilizzatori dei Web Analytics erano e sono ancora legati agli E-commerce, cui interessa il comportamento e la reazione dei

potenziali acquirenti rispetto alle campagne (Pubblicitarie) in corso; pertanto, è sempre stato avvertito il bisogno di strumenti sempre più "User-friendly", che sostituissero i linguaggi di programmazione, in quanto questi presupponevano conoscenze informatiche non tipiche di chi si occupa di Marketing. Le imprese, quindi, erano costrette ad avvalersi di sviluppatori, spesso esterni. Da questo bisogno sono nate soluzioni "Low-code" che hanno permesso di abbattere i "Programming boundaries" che attanagliavano questo settore. Nascono dunque i Tag Management System (TMS) ovvero sistemi utili a gestire il ciclo di vita dei tag di e-marketing (a volte indicati come pixel di tracciamento o web beacon). Tale funzionalità può includere analisi dei dati web, analisi delle campagne, misurazione del pubblico, personalizzazione, test A/B, retargeting comportamentale e monitoraggio delle conversioni. I principali vantaggi di un sistema di gestione dei tag è quello di consentire, a utenti non sviluppatori, di eseguire diverse attività di tracking, migliorando le prestazioni con la riduzione del codice scritto. Si accede quindi separatamente al sistema di gestione dei tag (normalmente tramite un sito Web) per assegnare la priorità e "attivare" i singoli tag, in base alle regole aziendali, agli eventi di navigazione e ai dati noti. In concreto quindi il TMS è molto semplice da utilizzare, lo si imposta direttamente da interfaccia client ed il collegamento con il sito è reso possibile dal codice JavaScript che viene generato, generalmente, in automatico dalla piattaforma di analyitics.

In Matomo nel codice è presente il collegamento a Matomo tramite la presenza di "Containers": veri e propri contenitori che contengono gli strumenti del tracciamento. Tutto ciò che riguarda il tracciamento viene convogliato nel container che presenta, al suo interno, dei parametri customizzabili direttamente che servono per tracciare:

- *Tags*, cioè gli oggetti di tracciamento, possono essere eventi, punti specifici o variabili all'interno delle pagine;
- *Triggers*, l'evento che mette in moto dall'identificazione del tag;
- *Variabile,* cioè Variabili che possono essere utilizzate sia come oggetto del Tag che come strumento per il trigger;
- *Versions*, per validare il funzionamento di un Tag quando si utilizza la modalità Debug; una volta validato si può salvare la versione ufficiale.

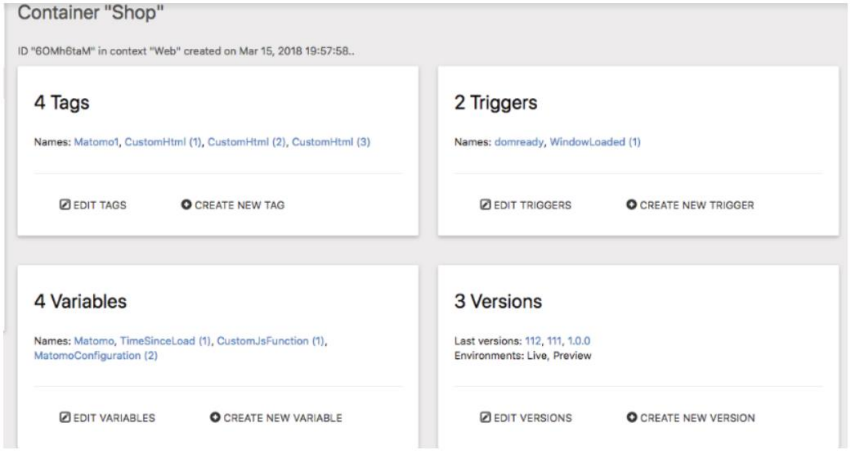

Figura 11. Interfaccia Tag Manager di Matomo.

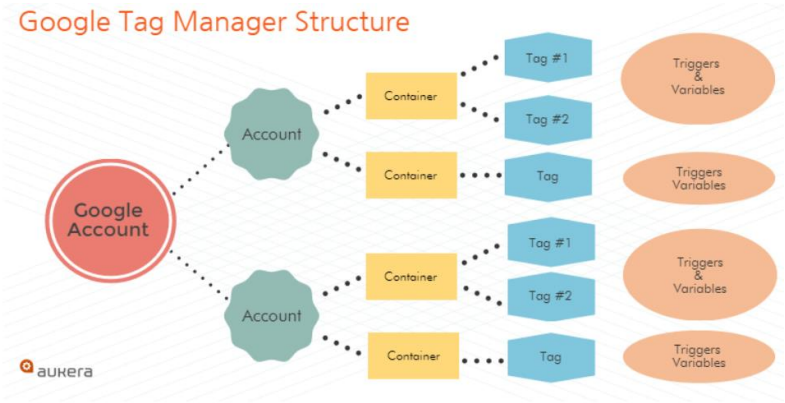

Figura 12. Schema funzionamento di Google Tag Manager.

I vantaggi dei sistemi di gestione dei tag includono:

- *agilità*, poiché c'è ridotta dipendenza dalle risorse tecniche e dai cicli IT conferisce maggiore agilità agli utenti aziendali;
- *prestazioni*, poiché i tempi di caricamento della pagina sono ridotti grazie al caricamento asincrono dei tag, al caricamento dei tag condizionale e alla funzionalità di time-out dei tag;
- *risparmio sui costi*, poiché non c'è bisogno di uno sviluppatore per gestire lo strumento;
- *controllo dei dati*, poiché hanno capacità di controllare la fuga di dati a terzi e di rispettare la legislazione sulla privacy dei dati (consenso ai cookie); TMS forniscono anche

un altro livello di astrazione per la gestione della complessità dei siti Web di grandi dimensioni;

- *anteprima sicura*, in particolare, alcuni gestori di tag, come Google Tag Manager, Ensighten e Matomo, includono una modalità di anteprima che consente di verificare la formattazione e i problemi di sicurezza, prima di distribuire i tag in produzione.

Figura 13. JavaScript & TMS codici di tracciamento.

2.2 Web Performance

Una volta estratto il dato, la fase successiva è quella di aggregarlo secondo le metriche di interesse. Qualsiasi strumento di analisi web presenta un report di sintesi, cioè un gruppo di metriche di base che sono disponibili immediatamente, dopo aver effettuato l'accesso (Fig. 14).

Figura 14. Videata del report sintesi del tag Matomo.

Le statistiche in figura 14 sono le seguenti.

- Visits: numero di sessioni sul sito e numero di volte in cui l'utente interagisce con il sito.
- Bounce Rate (Visit have bounced): percentuale di singole visioni di pagina (metrica che può essere differenziata inserendo il tracciamento con visite che sono durate meno di 5 s)
- Page Views: numero di pagine (del sito) richieste in tutte le visite.
- Average Time on Site: tempo medio di permanenza sul sito.

Le metriche precedenti variano a seconda del settore e, per questo motivo, non esiste un valore di riferimento assoluto con cui confrontarsi. Il metodo migliore è quello di seguire il trend temporale, registrando più dati possibili, per capire se il sito web sta migliorando. Un alto numero di pagine viste sul sito web è un buon segno nella maggior parte dei casi, tranne, ad esempio, per i siti web di supporto, in cui il cliente vuole trovare le informazioni velocemente. Inoltre, il tempo di visita su un sito web potrebbe essere un buon indicatore del coinvolgimento dei visitatori. Il Bounce Rate misura la qualità di traffico che si sta acquisendo, e aiuta a capire se i visitatori tendono ad uscire dal sito per qualche problema.

Di seguito è riportato un elenco delle metriche di maggiore importanza.

- *Direct Traffic* rappresenta i visitatori che si presentano sul sito web inserendo l'URL del sito o da un segnalibro. Nel caso in cui i visitatori vengono reindirizzati, anche se tramite codice maligno, questo traffico è sempre considerato come "diretto". Dal Direct traffic si può capire quanto traffico proviene da utenti che conoscono abbastanza bene il sito, avendo acquisito o inserito tra i segnalibri l'URL.
- *URL* di riferimento (Referrer) sono altri siti web collegati al sito web analizzato. Questi potrebbero essere il risultato di banner pubblicitari, di campagne, o provenire da blog coinvolti al sito web. Gli URL di riferimento aiutano a identificare le fonti che non si sa se generano traffico nel sito web analizzato.
- *Pagine di uscita*. Conoscere quali pagine sono le ultime viste può aiutare a comprendere cosa gli utenti cercano sul sito. Analizzando le pagine di uscita, si può

anche monitorare il tempo dedicato dai visitatori a rimanere su ogni pagina del sito.

- *Tasso di conversione*. Il sito web è una preziosa opportunità per generare contatti e vendite. Il tasso di conversione valuta il sito in termini di convincimento degli utenti a effettuare gli acquisti desiderati.
- *Impressioni*. Sono viste e sono comunemente usate in pubblicità online; è facile finire su una pagina web di un sito ma è molto più difficile ottenere un utente che visualizzi la pagina per esplorare il resto del sito web.
- *Posizionamento*. Indica la posizione nella quale il sito Web site compare nelle pagine di risultati di Search Engine. L'obiettivo che si dovrebbe avere per il sito web è quello di essere classificato, o posizionato, più in alto possibile dai motori di ricerca. Le tattiche di ottimizzazione di Search Engine sono fondamentali nell'amplificazione del posizionamento sito (SEO).
- *Backlinks*. I links possono essere interni o esterni e sono considerati vitali nell'aiutare il sito a guadagnare traffico e ricerca organici (traffico derivante dai motori di ricerca). Collegamento a un sito web esterno irrilevante può influenzare negativamente SEO del sito web. La SEO è il processo che aiuta ad incrementare la qualità o quantità del traffico in un sito web migliorane sensibilmente la probabilità di essere trovato (da chi o cerca) sui motori di ricerca i quali, utilizzano proprio la qualità e la quantità di backlinks per determinare la posizione di un sito web nei risultati di ricerca.

Oltre all'analisi delle attività dell'utente sulle pagine del sito ed il monitoraggio del suo comportamento per fini di Marketing, la Web analytics è utilizzata anche per il monitoraggio della qualità e del buon funzionamento del sito. Misurare le prestazioni delle pagine è fondamentale per diagnosticare e migliorare le prestazioni nel tempo. Senza i dati, è impossibile sapere con certezza se le modifiche poste in essere sul sito raggiungono effettivamente i risultati desiderati. Molti applicativi di Analytics forniscono la Real User Monitoring (RUM) e supportano le metriche "Core Web Vitals" nei loro strumenti.

La Real User Monitoring (RUM) è una tecnologia di monitoraggio passivo che registra tutte le interazioni degli utenti con un sito web o un client; questi interagiscono con un

server o un'applicazione basata su cloud. Il monitoraggio dell'effettiva interazione con un sito web o un'applicazione è importante per determinare se gli utenti sono "serviti" in modo rapido e senza errori; si può altresì valutare quale parte di un processo aziendale sta fallendo. Il *software as a service* (Saas) e gli *application service providers* (ASP) utilizzano la RUM per monitorare e gestire la qualità del servizio fornito ai loro clienti. I dati di monitoraggio degli utenti reali sono utilizzati per determinare l'effettiva qualità del servizio fornito agli utenti finali e per rilevare errori o rallentamenti sui siti web.
Il monitoraggio degli utenti reali è tipicamente "monitoraggio passivo", cioè il dispositivo RUM raccoglie il traffico web senza avere alcun effetto sul funzionamento del sito. Nella maggior parte dei casi, una pozione di JavaScript viene iniettata nella pagina o codice sorgente all'interno dell'applicazione, per fornire un riscontro dal browser o client. Questi dati vengono poi raccolti e consolidati da vari operatori (Oyama et al., 2011).
L'ottimizzazione per la qualità dell'esperienza dell'utente è la chiave per il successo a lungo termine di qualsiasi sito web. Google ha fornito una serie di strumenti per misurare e riferire le prestazioni sul web; i *Web Vitals*, derivano da un'iniziativa di Google per fornire una guida unificata dei segnali di qualità, che sono essenziali per fornire una migliore esperienza utente sul web. L'iniziativa *Web Vitals* mira a semplificare il paesaggio e aiutare i siti a concentrarsi sulle metriche che contano di più, i *Core Web Vitals*.
Questi ultimi sono un sottoinsieme di Web Vitals e si applicano a tutte le pagine web e dovrebbero essere tenuti in considerazione su tutti i siti. Ciascuno dei Core Web Vitals rappresenta un aspetto distinto dell'esperienza utente, è misurabile sul campo e riflette l'esperienza del mondo reale di un risultato critico incentrato sull'utente. Le metriche che compongono i Core Web Vitals si sono evolute e si evolveranno nel tempo. Il programma per il 2020 è stato concentrato su tre aspetti: il loading, l'interattività e la stabilità visiva; esso comprende le seguenti metriche e le rispettive soglie (Fig. 15).

- *Largest Contentful Paint* (LCP): misura le prestazioni di caricamento. Per fornire una buona esperienza utente, LCP dovrebbe verificarsi entro 2,5 s da quando la pagina inizia il caricamento.

- *First Input Delay* (FID): misura l'interattività. Per fornire una buona esperienza utente, le pagine dovrebbero avere un FID di 100 ms o meno.
- *Cumulative Layout Shift* (CLS): misura la stabilità visiva, ovvero la stabilità del layout di pagina. Per fornire una buona esperienza utente, le pagine dovrebbero mantenere un CLS di 0,1 o inferiore.

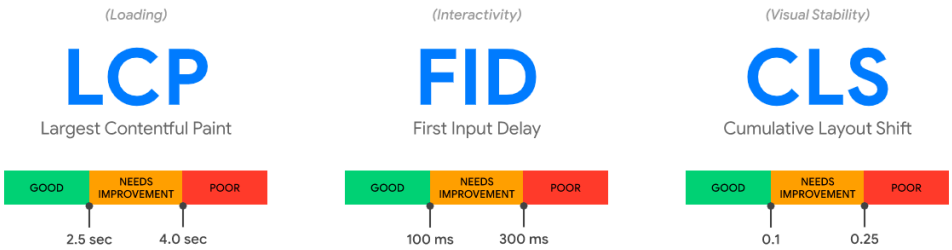

Figura 15. Metriche "Core Web Vitals".

Per ciascuna delle metriche di sopra citate, al fine di garantire che si sta colpendo il target consigliato per la maggior parte dei vostri utenti, una buona soglia per misurare è il 75° percentile di caricamento di pagine, segmentato tra dispositivi mobili e desktop. Gli strumenti che valutano la conformità ai Core Web Vitals dovrebbero considerare il passaggio di una pagina se soddisfa gli obiettivi raccomandati al 75° percentile per tutte e tre le metriche.

2.3 Ciclo di vita dei Web Analytics Services

Ciascuna metrica da monitorare nel corso della vita di un sito internet tende non solo a variare di importanza a seconda dell'ambito di applicazione, ma varia anche, riferendosi in particolare alle performance richieste dal mercato, con il passare del tempo. Questo continuo processo di sostituzione e cambiamento è tipico del mercato IT e si rispecchia nei prodotti che si utilizzano per fare Analytics, Web Analytics e che servono per l'acquisizione di statistiche in generale, a supporto della Business intelligence. I vendors, infatti, si sono muniti di politiche ben delineate di aggiornamento, gestione e svecchiamento che sono tenuti a fornire non appena rilasciano il prodotto sul mercato. Questo tipo di politiche deve essere inteso come uno strumento fondamentale per coloro che acquistano il prodotto, ai quali di solito

viene garantito un periodo minimo di aggiornamenti in cui il prodotto dovrebbe rimanere sul mercato.

Le suddette politiche sono una pratica estremamene diffusa in questo settore e si è diffusa una standardizzazione delle poclicy da parte di tutti i vendors. Nelle figure 16, 17 e 18 sono riportate le politiche di gestione del ciclo di vita del prodotto di alcuni tra i più importanti fornitori di prodotti IT: Dell, IDM e VidyoCloud; si nota che, a parte alcune variazioni di terminologie, tutte le politiche prese in esame si assomigliano per Stages e tempistiche. In generale, la vita segue un processo che inizia con la "BOL": Beginning Of Life, procede con una fase di maturità "MOL": Middle Of Life e termina con il declino del prodotto fino al raggiungimento della "EOL": End Of Life.

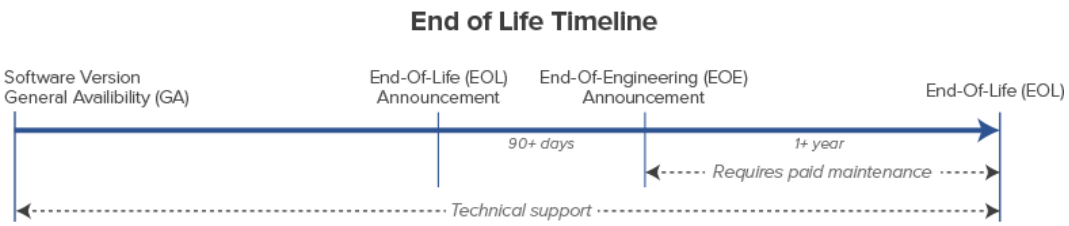

Figura 16. Release policy di Cycle IDM.

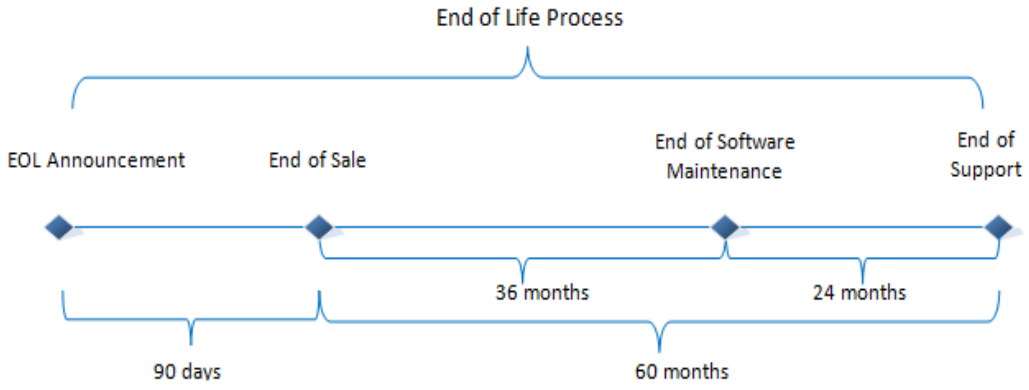

Figura 17. Release policy di Vidyo Cloud.

In linea con il ciclo generale (Fig. 19) adottato dalle policy di gestione della "Software release", il primo periodo di interesse, che corrisponde alla General Availability (GA), è la fase di commercializzazione, in cui tutte le attività per la vendita necessarie sono state completate e il prodotto software è disponibile per l'acquisto. Le attività di

commercializzazione potrebbero includere test di sicurezza e conformità, nonché localizzazione e disponibilità a livello mondiale. Il tempo tra RTM (Release to manufacturing) e GA può variare da settimane a mesi, prima che il rilascio possa essere annunciato a causa del tempo necessario per completare tutte le attività di commercializzazione richieste da GA.

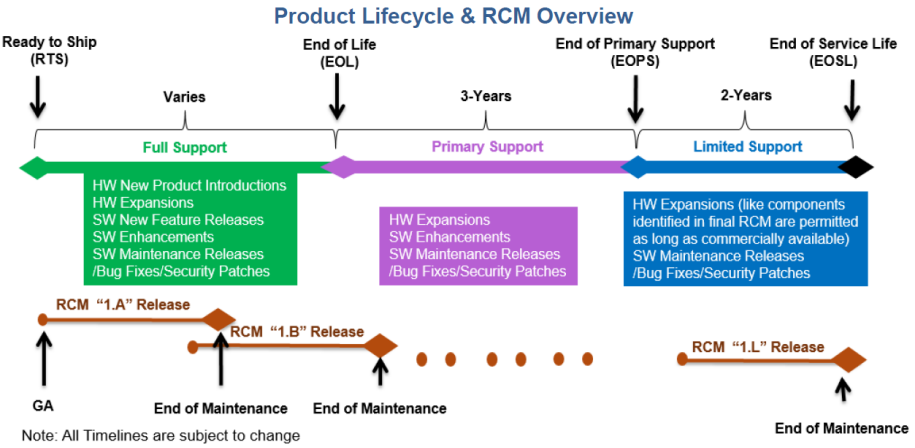

Figura 18. Release policy di DELL.

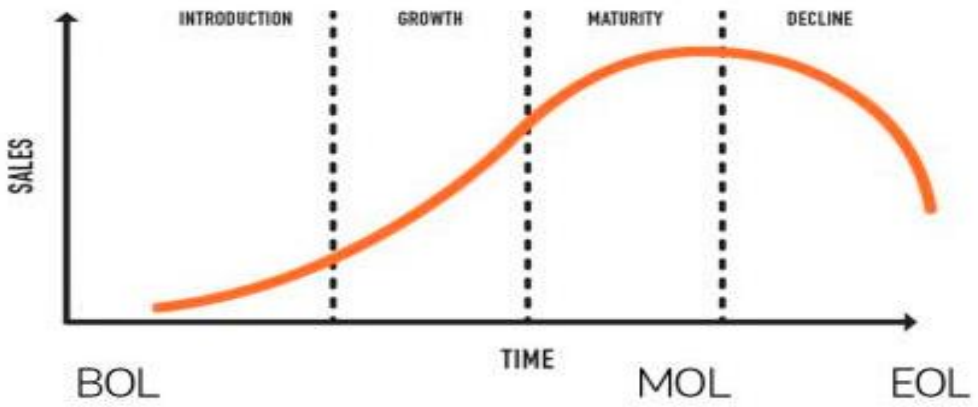

Figure 19. General Product life Cycle.

A questo punto, il software è "Gone live"; fase che coincide con quella che per Dell è definita "Ready to Ship" (RTS), in cui tutte le case madri, si impegnano a fornire il massimo supporto (Fig. 18). La fase che segue segna l'inizio del declino del prodotto con "End-of-Life Announcement" come primo stadio, che dà il via alla "End-of-Life" (EOL). Quando il software non è più venduto o supportato, si dice che ha raggiunto

la fine della vita; è quindi diventato obsoleto, ma è possibile che il forte legame dei clienti con questo prodotto possa portare a lunghe proroghe la sia esistenza.
Dopo la data di fine vita, il produttore di solito non implementa nuove funzionalità, non fornisce alcun supporto per il prodotto e non corregge difetti esistenti, bug o vulnerabilità; se il produttore lo desidera, può rilasciare il codice sorgente, in modo che la piattaforma sopravviva e sia mantenuta in vita da volontari. Si creano spesso delle community che tendono a mantenere in vita alcuni prodotti ormai diventati obsoleti, come se si creasse un sottomercato. Inevitabilmente, "End-of-Sale", periodo in cui il prodotto non viene più venduto dalla casa madre, costituisce l'ultimo stadio del Life Cycle (Figg. 16-17-18). Ovviamente, questo non vieta che rivenditori terzi possano ancora commercializzare il prodotto o servire assistenza. Questa milestone viene inglobata da Dell nella fase di "End-of-Life Date", data in cui la Dell Technologies interrompe la vendita di un sistema convergente (Fig. 18). Questo segna l'inizio della fase di Supporto Primario, durante la quale nuove introduzioni di componenti seguiranno il loro corso: miglioramenti di funzionalità, manutenzione del software, correzioni debug, patch di sicurezza, etc.
Le milestones successive, caratterizzate da una varietà di nomi quali "End-Of-Engineering" (EOE), "End-of-Maintenance" e "End-of-Primary Support Date" (EOPS) (nel caso di Dell), hanno generalmente gli stessi sviluppi (Figg. 16-17-18) e corrispondono all'ultima data in cui il produttore rilascia aggiornamenti di manutenzione o correzioni di bug per un prodotto. Dopo questa data, il produttore non svilupperà, riparerà, manterrà o testerà più nuove versioni software per il Prodotto. Per IDM (Fig. 18) la data di fine vendita segna la cessazione dell'ingegneria relativa al Prodotto Software e l'azienda cessa di fornire supporto tecnico relativo alla versione applicabile, riservandosi il diritto di interrompere il supporto tecnico per qualsiasi Release del Software, resa generalmente disponibile dopo due Release successive. Quando una Release raggiunge la data di Fine Engineering, non sarà più attivamente supportata dall'ingegneria. Il Supporto Tecnico è l'assistenza tecnica fornita ai clienti IDM in manutenzione attiva e supporto e include una guida generale sui prodotti IDM ed eventuali soluzioni, correzioni di bug, ecc. che sono stati progettati prima della fine dell'ingegneria; questo sarà disponibile fino alla data di fine

vita, corrispondente a un anno dopo la data di fine di ingegneria, fino a quando il cliente è sotto manutenzione o pagamento.

Dell, invece, non offre più sviluppo di componenti o di funzionalità software per un sistema convergente dopo l'EOPS. Questo segnerà anche l'inizio della fase che per Dell è definita "Supporto limitato", durante la quale i prodotti continueranno a ricevere manutenzione del software, correzioni di bug e solo patch di sicurezza critiche. La data ultima da ricordare in questo ciclo è la data che Dell definisce:" End-of-Service Life" (EOSL; fig. 18) o più comunemente la data di "End-of-Support" (EOS). Una volta che un prodotto ha raggiunto la fine del supporto, il prodotto sarà considerato a fine vita e non verrà riparato, sostituito o altrimenti supportato. Dopo questa data la casa madre non produrrà alcun servizio di manutenzione e neanche patch critiche.

CAPITOLO III

Studio e sperimentazione delle metriche di gestione web analytics dei siti corporate di un'azienda broadcaster televisivo

3.1 Stato dell'arte

Il mondo della web analytics, come visto nei capitoli introduttivi, trova applicazione principalmente nel mondo degli E-commerce così come in quello dei Media, dove ormai la componente web ha definitivamente preso il sopravvento nel core business di tutte le imprese di distribuzione televisive (Géczy et al, 2009). In ogni caso, è buona norma per le imprese conservare e monitorare siti non solo internet ma anche intranet, lì dove solamente i dipendenti, o chi possiede un accesso previlegiato, nella VPN della corporate può navigare all'interno (Géczy et al, 2009). D'altra parte, quando si parla di siti corporate, le metriche di interesse variano totalmente, dando priorità alle prestazioni delle pagine, così come la loro funzionalità, ai fini di monitorare se e come gli utenti le stiano impiegando (Géczy et al, 2009).

Nella letteratura scientifica, numerosi studi sono stati effettuati per proporre valutazioni e sperimentazioni in questo ambito. Schneider et al. (2021) hanno fornito una panoramica sugli strumenti di analisi web, identificando le opportunità che derivano dall'applicazione di questi potenti strumenti agganciati ad un software; il loro studio evidenzia che diverse categorie, utenti e sviluppatori possono trarre vantaggio da questi prodotti: gli utenti ne trarrebbero beneficio con risultati più precisi di ricerca e gli sviluppatori potrebbero facilmente monitorare la qualità di funzionamento del software. Le prove e le valutazioni sono state validate con una prototipazione basata su una piattaforma commerciale integrata al software di analytics Piwik, oggi conosciuto come Matomo.

Relativamente al tasso di rimbalzo è stato dimostrato che, tra le metriche di analisi web, il Bounce Rate fornisce una effettiva misura della soddisfazione dell'utente, proponendosi anche la questione di come effettuarne delle previsioni (Sculley et al., 2009). Una soluzione proposta consiste nell'utilizzo di un metodo di learninig su larga scala, che pesi le caratteristiche dei disegni creativi pubblicitari in aggiunta alle loro

parole chiave e alle loro pagine di destinazione; i risultati di tale approfondimento hanno portato a definire metodi di stima del Bounce rate in base al comportamento degli utenti, dimostrando che, anche in assenza di dati relativi a banner di click sulle pagine web, il bounce rate può essere stimato con il machine learning se applicato a specifici attributi estratti dalle ricerche pubblicitarie e i relativi tempi di caricamento (Sculley et al., 2009). Inoltre, risulta che i miglioramenti apportati nella stima della metrica sono significativi e sufficienti per portare gli ad-provider a ritorni di guadagno sugli investimenti consistenti in corrispondenza di un miglioramento del tasso di conversione, riscontrando anche che se il bounce rate può essere utile per identificare problemi di qualità, da solo, non suggerisce immediatamente le azioni che i provider pubblicitari possano prendere per veicolare i problemi (Sculley et al., 2009).

Parwez et al. (2017) hanno introdotto un approccio di rilevamento delle anomalie nella rete mobile, servendosi di algoritmi di clustering, impiegando informazioni spaziali e temporali contenute nelle reti mobili per analizzare le attività degli utenti. I dati utilizzati contengono ID, timestamp, attività di chiamate in entrata, attività di chiamate in uscita e attività di SMS in arrivo e in uscita; le attività insolite dell'utente sono state classificate come anomalie, in quanto causano elevate esigenze di traffico. In questo studio, utilizzando algoritmi come il K-means ed il clustering gerarchico, le attività anomale degli utenti sono state raggruppate e identificate, inoltre sono state identificate le regioni di interesse ed effettuate azioni correttive nell'assegnazione del traffico (Parwez et al., 2017).

Un nuovo approccio è stato introdotto per individuare utenti che presentano pattern simili identificando modelli di comportamento tramite dati raccolti dai cellulari degli utenti, inclusi registri di profili, ID e registri di interazione come i giochi elettronici: dopo una fase di normalizzazione, utile per dare una rappresentazione più generale dei dati, è stato utilizzato un modello di fattorizzazione della matrice bayesiana basato su vincoli per estrarre abitudini comuni dai comportamenti generali degli utenti, che sono stati successivamente trasformati in vettori di abitudini comuni denominati modelli Iper-comportamentali in uno spazio più esteso (Ma et al., 2012). Per confrontare il grado di similarità ogni due record, e generare cluster di comportamenti simili è stata calcolata la distanza coseno e i risultati dell'esperimento hanno mostrato che

l'approccio può ridurre la dispersione del comportamento in forma vettoriale ed individuare utenti simili in base alle loro abitudini in modo efficace (Ma et al., 2012).
Rivera et Brian (2021) hanno individuato i prodotti più popolari nel mondo web analytics, successivamente servendosi del software R, per focalizzarsi sui report raccolti in Google Analytics; gli autori hanno effettuato la clusterizzazione sugli utenti tramite l'algoritmo K-means, al fine di indirizzare la vendita di prodotti e segmentare il mercato inesplorato delle pubblicità online tramite le informazioni derivanti dai cluster, questo ha consentito un significativo risparmio in termini di campagne pubblicitarie (Rivera et Brian, 2021).
Mettendo a confronto gli antipodi della web analytics, open source e software proprietario, è stata effettuata un'analisi approfondita, senza passare per l'implementazione strutturale del prodotto, dell'applicazione del Machine Learning ed AI ad un tema che viene definito particolarmente critico (Sujo et Ruano, 2019): la stampa digitale intesa non solo come il nuovo giornale ma anche come prodotto online, definito "prodotto che integra diversi media (testo, immagine, audio e video) e consente a colui che riceve il servizio di ottenere uno strumento di azione" (Armañanzas et al., 1996). Questo studio, si è occupato di predire la popolarità di nuove notizie in base ai fattori che le influenzano, individuandoli e trovando modelli in grado di identificare il comportamento di una notizia e di definire il suo sentimento prima di essere pubblicata (Sujo et Ruano, 2019); tramite l'analisi dei dati per mezzo di vari algoritmi di machine learning, è stato introdotto un algoritmo in grado di prevedere se una notizia sarà popolare o meno (Sujo et Ruano, 2019).
Un'altra ricerca ha avuto come obbiettivo quello di riconoscere pattern comuni nell'utilizzo di una app, sempre servendosi di tecniche di machine learning, ed ha portato a identificare, con il clustering, un numero definito di pattern compreso di relativa qualità del modello (He Yu, 2018); con questo procedimento sono state individuate 6 diverse determinanti, dell'utilizzo dell'App in questione: baseline, attività, istruzione, esercizio, flusso utenti e sessioni. Ognuna di queste determinanti è stata analizzata utilizzando metriche differenti presenti nel software Matomo (He Yu, 2018).
Da quanto innanzi si evince che nella letteratura scientifica, sono stati analizzati i temi del flusso dei clienti e delle attività di marketing sotto vari aspetti, senza però trattare le interazioni nel campo dell'intranet. Pertanto, il presente studio sperimentale si è

concentrato sulla definizione di un modello che possa identificare diverse classi di pagine, in base alla concezione di funzionalità che ha l'utente corporate; la ricerca è stata svolta partendo dall'analisi di mercato, fino a definire l'architettura progettuale di un applicativo da impiegare nel settore dei media televisivi. Inoltre, è stato identificato un modello rappresentante due importanti variabili analizzate nel mondo della Web Analytics.

3.2 Materiali e Metodi

La ricerca effettuata è stata richiesta da un noto broadcaster televisivo operante nel settore dei media, con la necessità di sostituire un prodotto per l'analisi web delle statistiche intranet ed internet di alcuni profili.

La procedura sperimentale si è articolata nelle seguenti fasi: analisi di mercato, implementazione dell'applicativo, analisi delle metriche.

3.2.1 Analisi di mercato

I requisiti che il nuovo applicativo avrebbe dovuto avere da Richiesta di Offerta RdO sono stati divisi in n. 3 macrocategorie: requisiti funzionali, requisiti di integrazione, requisiti tecnici e di esercizio.

Relativamente ai requisiti funzionali, il sistema doveva garantire la disponibilità dei principali strumenti di analytics, a partire dalla definizione dei diversi siti\contesti per arrivare alla definizione delle principali strategie di raccolta dei dati, in modo da poter generare report avanzati e configurabili relativi a:

- accessi alle pagine,
- conteggio dei visitatori loggati e non,
- clusterizzazione dei visitatori,
- tracking degli eventi generati dall'utente sulle pagine, durante la fruizione di contenuti multimediali, o interagendo con contesti social.

Un altro elemento di valutazione è stato la disponibilità di API per il completamento dei dati raccolti, tramite tag o dati terzi, così come di API per l'esportazione dei dati raccolti; inoltre, è stata valutata la disponibilità di strumenti di analisi più avanzati quali la gestione tag, disponibilità di heatmap, catalogazione pagine, tracciamenti cross site e gestione form.

Relativamente ai requisiti di integrazione, un fondamentale elemento di valutazione della soluzione è stato la disponibilità e la semplicità di soluzioni di integrazione facilmente adottabili dai diversi siti del richiedente. Particolare rilievo, da questo punto di vista, è stato la disponibilità di ampia documentazione e di numerose esemplificazioni dell'utilizzo degli script da integrare nelle pagine web e nelle eventuali applicazioni mobile. In particolare, un parametro di valutazione è stato la disponibilità di kit ed esempi di integrazione compatibili con i framework di sviluppo più utilizzati (es. Angular).

Relativamente ai requisiti tecnici e di esercizio, oltre alla verifica dell'integrazione con i web server IIS e Apache, nonché col supporto al CMS Wordpress, è stato verificato il potenziale della soluzione in termini di esercibilità. In particolare, nel caso di soluzioni cloud, sono state verificate la scalabilità della soluzione, la disponibilità di supporto di tipo premium in linea con gli SLA di servizio richiesti, la collocazione geografica del datacenter del servizio e la verifica della compatibilità con gli standard richiesti dalle normative GDPR.

Nel caso di soluzione on premise è stato necessario verificare:

- la compatibilità dell'intero stack operativo con gli standard dei data center del richiedente, privilegiando soluzioni compatibili per OS, database con l'ecosistema del richiedente, nonché la possibilità di adottare sistemi virtuali;
- la scalabilità;
- la possibilità di disegnare architetture in alta affidabilità;
- l'integrazione con i sistemi di autenticazione aziendali.

L'ultimo requisito considerato in questa fase della ricerca è stato la preferenza verso soluzioni Open-Source. Uno dei fattori fondamentali nell'adozione di una soluzione open source è verificare che il prodotto sia adeguatamente supportato in modo da garantirne costanti aggiornamenti per il monitoraggio di problematiche funzionali e di sicurezza, per garantirne il funzionamento a fronte di aggiornamento di sistemi operativi, database, web server, etc. A tal fine è stato fondamentale verificare le dimensioni e l'attività della community degli sviluppatori e delle aziende che supportano la soluzione. Gli indicatori principali di questi fattori sono stati la verifica dei repository dei sorgenti del software, verificando il numero di contributori al progetto e la frequenza di aggiornamento dei sorgenti, l'adozione e il rilascio di

aggiornamenti per la compatibilità ai sistemi operativi\database\web server aggiornati. Nel caso specifico delle soluzioni di web analytics, un interessante elemento di valutazione è stato la verifica della diffusione della soluzione nell'ambito delle più diffuse community di supporto alla realizzazione di siti web come quella di Wordpress. Altro fattore di valutazione è stato la disponibilità di piattaforme cloud che offrano l'utilizzo della soluzione con l'eventuale disponibilità di un supporto premium.
Sulla base dei requisiti analizzati, è stata effettuata un'analisi di mercato che guardasse alle soluzioni disponibili ad oggi, il cui deliverable ultimo è stato una matrice di valutazione che riassume tutte risposte dei prodotti sul mercato ai requisiti funzionali, di integrazione e tecnici. In figura 20 sono riportati gli applicativi partecipanti alla valutazione.

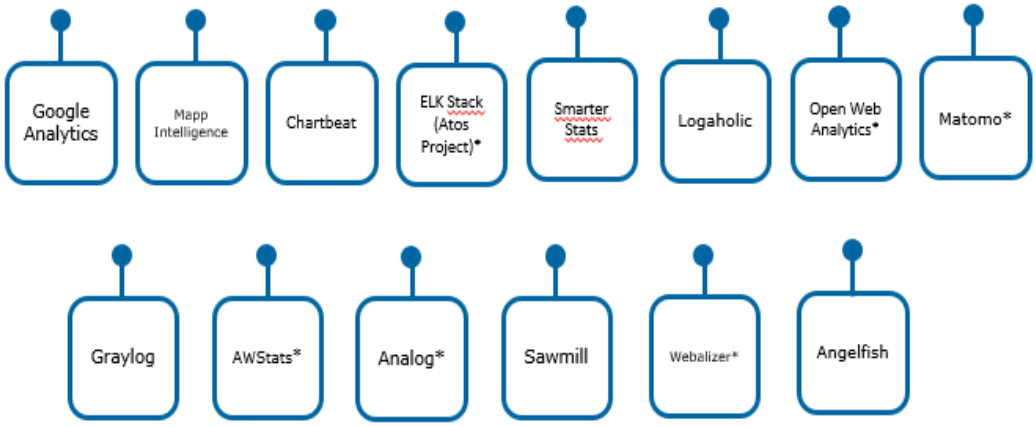

Figura 20. Prodotti sul mercato studiati.

Dall'analisi delle criticità individuate, sono poi stati scartati alcuni dei prodotti studiati e la scelta del prodotto è stata effettuata congiuntamente ai titolari della ricerca, mediante n. 3 matrici estratte: matrice dei requisiti tecnici, matrice dei requisiti funzionali, matrice dei requisiti di integrazione (Fig. 21).
Il processo di valutazione si è basato su una scala qualitativa che indicasse in un range a tre livelli (FIT, PARTIAL FIT, NOT FIT) il grado di compatibilità con ciascuno dei requisiti citati e riassunti nelle matrici. (Duraisamy and Atan, 2013)

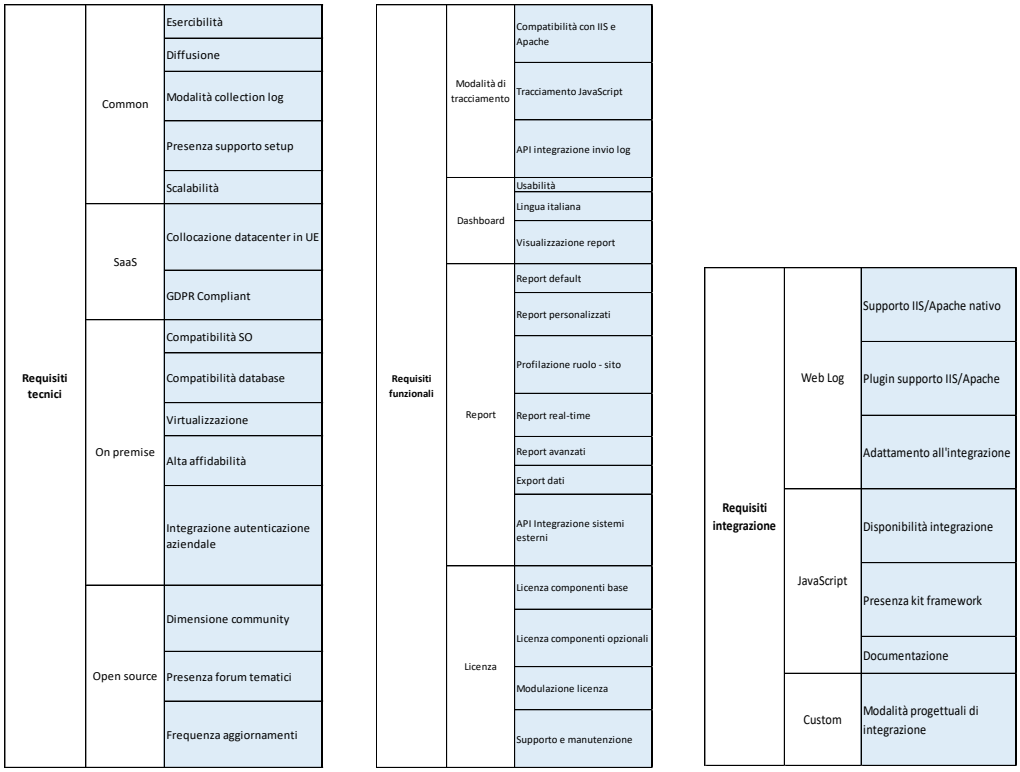

Figura 21. Matrici dei requisiti tecnici, funzionali e di integrazione.

Come sarà meglio illustrato nel successivo paragrafo, i risultati della presente fase della ricerca hanno consentito di selezionare il prodotto *Matomo*, che ha riportato il "Grado di fit" più elevato.

3.2.2 Implementazione dell'applicativo

Scelto il prodotto, si è proceduto all'implementazione della soluzione applicativa, definendo i seguenti requisiti da soddisfare con i componenti architetturali.

- Numero nodi Frontend: ≥ 2.
- Hardware:
    - 8 Virtual CPU;
    - 16 GB RAM;
    - 50 GB disco per Sistema Operativo, applicazione e log applicativi.
- Software:
    - Linux RedHat 7.9 OS;
    - Apache HTTPD 2.4;
    - PHP 7.x (7.2.5 o maggiore);

    - Python 3.x (3.5 o maggiore);
    - Matomo 4.x;
    - Keytab-URL alias Matomo e hostname nodi.

Requisiti per la cartella dei web log:

- NAS share 500 GB disco.

Requisiti necessari per singola macchina di Back-end:

- Numero nodi Backend: ≥ 2;
- Hardware:
    - 8 Virtual CPU;
    - 16 GB RAM;
    - 30 GB disco per S.O;
    - 500 GB disco per MySQL.
- Software:
    - Linux RedHat 7.9 OS;
    - MySQL 8.x.

Requisiti necessari per la partizione di archiviazione e backup del database:

- NAS share 500 GB disco.

Le componenti architetturali e le configurazioni sono state replicate per entrambi gli ambienti di collaudo e produzione ed è stata definita una soluzione al fine di soddisfare i requisiti espressi dal richiedente della ricerca.

3.2.3 Analisi delle metriche

Una volta terminati i processi di selezione ed implementazione del prodotto, l'ultima fase dell'analisi svolta si è interessata del miglioramento del processo. È stato infatti migliorato nella sua precisione e nella profondità delle intuizioni dovute alle sole statistiche rese disponibili sull'applicativo. Successivamente alle fasi di collaudo del prodotto, sono stati raccolti ed esportati i risultati dei tracciamenti delle metriche identificate e sono stati analizzati secondo quella che è attualmente conosciuta nel mondo della data science come "Pipeline di KDD" (Fig.22). Il temine KDD sta a significare "Knowledge Discovery in Datataset" ed è un framework molto diffuso e abbastanza schematico che consiste in una successione ben delineata nell'analisi dei dataset affinché i risultati siano di buona qualità. (Bethaz P. et Al, 2021)

Figura 22. Pipeline di kdd.

Il framework di KDD consiste nelle seguenti fasi:

1. Estrazione dei dati ed eventuale costruzione del dataset;
2. Selezione delle metriche di interesse nel dataset;
3. Pre-processing: divisa in outlier detection e features selection;
4. Trasformazione dati ed eventuale generazione di indicatori utili per l'analisi;
5. Estrazione della conoscenza per mezzo di algoritmi o altri strumenti.
6. Visualizzazione ed interpretazione della conoscenza estratta in modo si possa trovare un'applicazione dei risultati dello studio effettuato;
7. Generalizzazione dello studio che si trasforma in conoscenza applicabile (Fig. 22).

Il dataset delle metriche di Matomo è stato realizzato esportando uno storico di 4 mesi di aggregazioni operate dal prodotto. Una volta messi aggregati i dati tramite funzioni di tabulazione di Excel il resto dello studio è stato svolto interamente su Jupiter Notebook, editor che supporta il linguaggio di programmazione Python con cui si è svolta l'analisi. In particolare, il dataset è frutto delle metriche registrate da Matomo rispetto a ciascuna pagina di un sito preso come campione per lo studio. La selezione dei dati componenti il dataset è stata effettuata aggregando i dati delle prime 150 pagine di questo sito ordinate in maniera decrescente rispetto al traffico sulla pagina. Il foglio Excel finale era quindi costituito da 600 righe e 21 colonne rappresentanti rispettivamente le pagine e metriche relative. Alcuni di questi attributi, tuttavia, non sono stati utilizzati ed eliminati successivamente nella fase di gestione delle colonne a causa della mancanza di informazioni ufficiali relative al significato della metrica. Queste sono state le seguenti: *total time spent by visitors (in seconds), nb_hits_with_time_server, total time spent by visitors (in seconds) after entering here.*

Prima di effettuare qualsiasi tipo di analisi sono state attentamente studiate le metriche di interesse raccolte il cui significato è riassunto nella seguente tabella (Tab. 2).

Tabella 2. Riepilogo delle metriche utilizzate nello studio.

Categoria	Metrica	Significato
BEHAVIOUR/PERFORMANCE	Bounce Rate(%)	% VISITE CHE SONO DURATE MENO DI 5 SEC SULLA PAGINA
	Exit rate(%)	% DI USCITE DAL SITO DOPO AVER VISTO QUELLA PAGINA
	LABEL	IDENTIFICATIVO UNICO DELLA PAGINA
BEHAVIOUR	Unique Pageviews	VOLTE IN CUI LA PAGINA E' STATA VISTA, NON TENENDO CONTO DELLE RIPETIZIONI
	Pageviews	VOLTE IN CUI LA PAGINA E' STATA VISTA, TENENDO CONTO DELLE RIPETIZIONI
	Entrances	NUMERO DI VISITE CHE SONO INIZIATE DA QUESTA PAGINA
	Actions after entering here	NUMERO DI AZIONI EFFETTUATE SULLA PAGINA DOPO ESSERCI ENTRATI
	Bounces	NUMERO DI VISITE CHE SONO INIZATE E FINITE SU QUESTA PAGINA, IN CUI IL VISITATORE è RIMASTO SOLO IN QUELLA PAGINA
	Exits	NUMERO DI VISITE CHE SONO TERMINATE SU QUESTA PAGINA
	Unique visitors	NUMERO DI VISITATORI DIVERSI CHE SONO ANDATI SULLA PAGINA
PERFORMANCE	Avg. server time	IL TEMPO MEDIO CHE IL SERVER IMPIEGA PER GENERARE LA PAGINA, DALLA RICEZIONE DELLA REQUEST ALLA EFFETTIVA PREPARAZIONED DELLA RISPOSTA.
	Avg. time on page	TEMPO MEDIO DI STAZIONAMENTO SULLA PAGINA(NON SUL SITO)
	min_time_server	MINIMO TEMPO DI RISPOSTA DEL SERVER (MENSILE)
	max_time_server	MASSIMO TEMPO DI RISPOSTA DEL SERVER (MENSILE)

Dopo aver effettuato lo studio delle metriche si è proceduto nella direzione di KDD effettuando le operazioni di pulizia del dataset passando su Jupiter Notebook utilizzando le seguenti librerie:

- Pandas
- Numpy
- Matplotlib
- Seaborn

Utilizzando un metodo di Pandas ".info ()" è stato possibile definire la percentuale di nulli e non nulli e, visto che le percentuali di valori nulli nel dataset si aggiravano tra il 2% e 4% (Fig. 23) si è pensato di gestirli tramite un riempimento del valore medio, utilizzando il metodo ".interpolate ()" della stessa libreria.

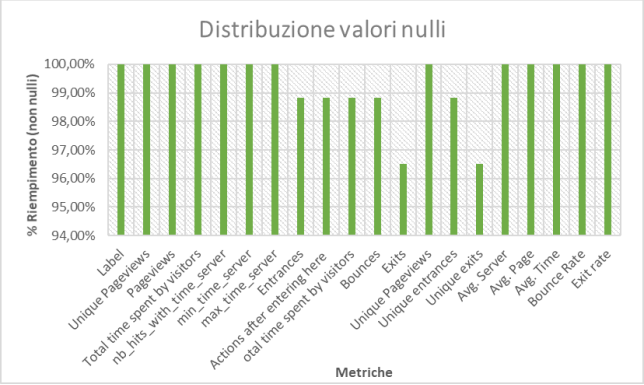

Figura 23. Percentuale non-nulli per ciascuna metrica.

Una volta raggiunta la totalità dei riempimenti si è passati alle fasi di Outlier detection e Feature selection.

Per quanto riguarda la selezione delle metriche, sono state utilizzate tutte quelle disponibili nel glossario fornito dalla piattaforma fatta eccezione dell'attributo attributo "Pageviews" che portava con sé lo stesso significato "Unique Pageviews". Diverso è stato l'approccio utilizzato per la selezione dei record, è ipotizzata e verificata la normalità delle distribuzioni anche data la elevata numerosità dei campioni, quindi, è stato deciso di considerare outlier tutti i valori che si trovassero ad una distanza $\mu \pm 3\sigma$. L'ultima fase di preparazione del dataset è stata la normalizzazione del tipo Min-Max così che non ci fosse troppa disparità tra scale di analisi, cosa che potrebbe penalizzare i modelli.

Una volta preparato il dataset sono stati ricercati all'interno del dataset "Pattern" comuni che potessero generare qualche informazione utile per integrare le metriche prese in esame. Gli algoritmi applicati sono stati i seguenti:

Supervised learning: *Regressione Lineare (OLS), Regressione Polinomiale;*

Unsupervised learning: *Clustering Kmeans, Clustering DBscan.*

Gli algoritmi sopra citati sono stati implementati entrambi utilizzando la libreria "Sckit-learn" di python e ad eccezione della regressione lineare che è stata implementata anche utilizzando un API verso il software "Stata" utile studiare la regressione OLS (Ordinary Least Squares) cioè una tecnica di ottimizzazione che permette di trovare una funzione, rappresentata da una curva ottima, che si avvicini il più possibile ad un insieme di dati. Al fine di migliorare la significatività dell'analisi e dei suoi risultati, inoltre, è stata analizzata la correlazione di ciascuna variabile presa in esame in modo da proporre nel modello di regressione variabili non correlate tra loro in una regressione multipla cercando di eludere il più possibile distorsioni da variabile omessa.

3.3 Risultati e discussione

I risultati dello studio effettuato hanno portato alla definizione dell'applicativo più adatto al caso specifico, nonché allo sviluppo della sua architettura e al consolidamento del processo di analisi delle metriche adottato, tramite l'utilizzo di tecniche di machine

learning; tuttavia, si possono individuare aspetti generali che è possibile estendere a tutte le fasi di analisi ed implementazione di un nuovo prodotto di analytics.

3.3.1 Risultati dell'analisi di mercato

L'analisi di mercato ha portato alla definizione dell'applicativo più adatto alla realtà economico-produttiva studiata.

In primo luogo, molti dei prodotti sottoposti ad analisi hanno evidenziato criticità rilevanti, che li hanno portati ad essere scartati e solo otto di essi sono risultati idonei (Fig. 24).

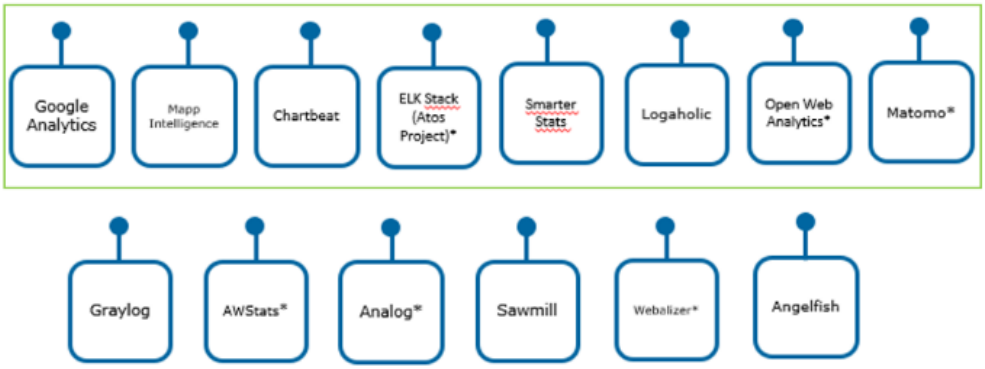

Figura 24. Risultati dello studio dei prodotti: nel riquadro sono raggruppati i prodotti selezionati.

Le criticità riscontrate nei prodotti scartati sono state le seguenti:

1. log collector e analizzatori di log a fini sistemistici;
2. scarso supporto, evoluzione del prodotto;
3. interfaccia statica e datata;
4. ecosistema proprietario, non standard.

La prima criticità è stata riscontrata prevalentemente nei prodotti: Graylog, AWStats, Analog, Sawmill e Webalizer. Infatti, è risultato che i log collector e gli analizzatori di log siano adatti per fini sistemistici diversi da quelli richiesti dalla presente ricerca. La seconda criticità è stata riscontrata nei prodotti: AWStats, Analog e Webalizer, per i quali è risultato che il supporto offerto sul prodotto sia poco affidabile, scarno e con una evoluzione poco flessibile; questo limite rende il prodotto poco applicabile alla realtà industriale studiata perché l'esigenza di una soluzione open-source porta, nei casi in cui la comunity non sia molto attiva, a difficoltà nella risoluzione dei problemi, scarsi aggiornamenti, release e patch. Inoltre, gli stessi prodotti presentano interfacce statiche

e datate. La terza criticità è stata riscontrata nel prodotto Angelfish; in questo caso è stato rilevato un sistema non standard in termini di database e web server, nonché scarsamente documentato sia dal punto di vista dell'istallazione proprietaria, sia per quanto riguarda la scalabilità del prodotto. In tutti questi casi, il prodotto non si addice alla dimensione economico-produttiva del fruitore specifico, perché lontano da molti requisiti di integrazione.

Nelle figure 25,26 e 27 sono riportati gli sviluppi delle matrici di valutazione per i prodotti che superato la prima selezione dovuta alle criticità; da esse si evincono le risposte dei prodotti sul mercato ai requisiti funzionali, di integrazione e tecnici.

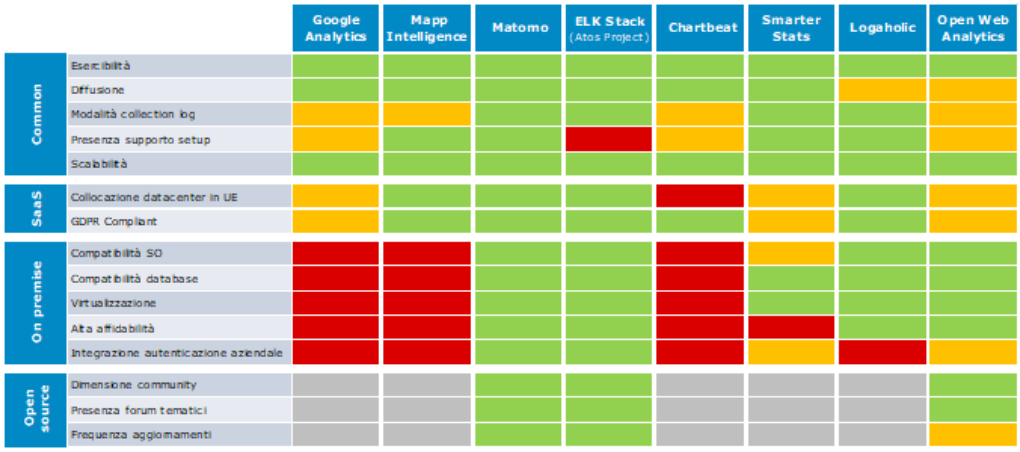

Figura 25. Risultati dello sviluppo della matrice dei requisiti tecnici.

Nella matrice dei requisiti tecnici (Fig. 25) è evidente che i prodotti Google Analytics, Mapp Intelligence e Chartbeat sono sviluppati per essere prevalentemente utilizzati sul cloud e quindi non possono soddisfare il requisito di struttura proprietaria ma unicamente quello di eternizzazione del servizio. Altre criticità si riscontrano in ELK Stack (Fig. 25) che non presenta un supporto per il setup adeguato; Logaholic invece presenta, nella forma cloud, problematiche di integrazione con i sistemi aziendali (integrazioni AD, LDAP etc.; Fig. 25). Smarter Stat risulta avere problemi di bassa affidabilità, a causa della difficoltà di clusterizzazione (Fig. 25). Il prodotto Open web Analytics evidenzia una situazione tecnica di parziale fit su una buona parte dei requisiti (Fig. 25), dovuta principalmente ad una scarsa diffusione del prodotto che, nei casi di open source, non è un aspetto positivo.

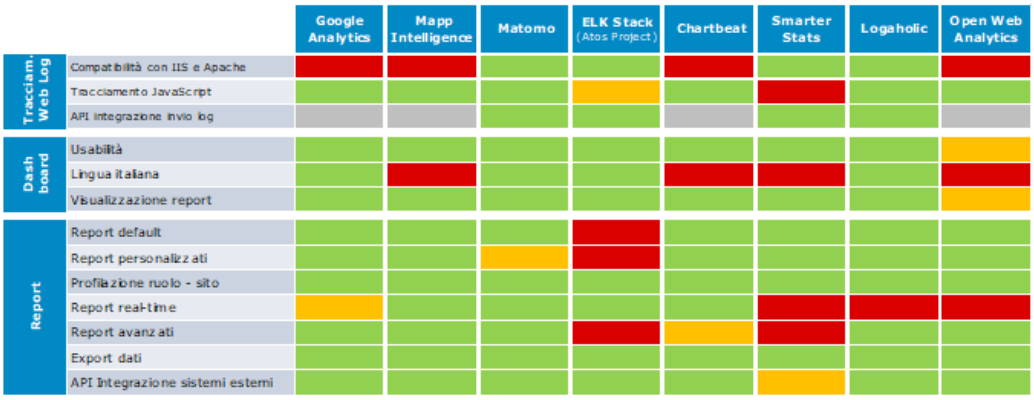

Figura 26. Risultati dello sviluppo della matrice dei requisiti funzionali.

Per quanto riguarda i requisiti funzionali (Fig. 26), così come per quelli tecnici, Google Analytics, Mapp Intelligence e Chartbeat, insieme in questo caso ad Open Web, non essendo prodotti on premise, non presentano la compatibilità con i web server interessati. Infatti, difficilmente prodotti sul cloud fanno anche log analytics, concentrandosi prevalentemente sul metodo JavaScript. Il prodotto ELK S presenta criticità funzionali legate al fatto che tutti i report devono essere creati scrivendo query specifiche in linguaggio DSL; dunque, è altamente customizzabile ma richiede una skill non comune. Smarter Stats è un prodotto che non presenta la modalità di tracciamento JavaScritpt (Fig. 26) e quindi non può garantire report particolarmente avanzati e in tempo reale, problema riscontrato anche in Logaholic.

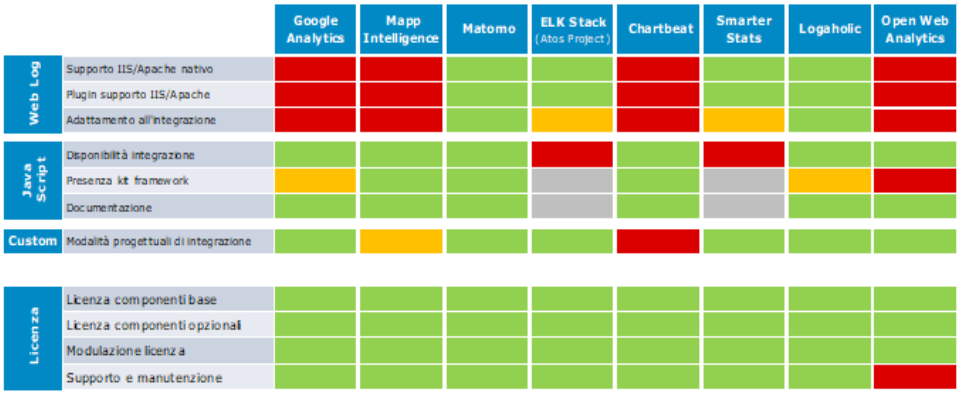

Figura 27. Risultati dello sviluppo della matrice di confronto requisiti di integrazione.

Analizzando infine la matrice dei requisiti di integrazione con le necessità del fruitore della ricerca (Fig. 27), si nota che, anche sotto questo aspetto, Google Analytics, Mapp

Intelligence, Chartbeat e Open Web presentano carenze, con particolare riferimento all'esigenza di effettuare Log Analytics. Le criticità emerse per gli altri prodotti sono dovute, invece, alla scarsa reperibilità, se non all'assenza di documentazione chiara e funzionale. Questo limite caratterizza tutti i restanti prodotti, ad eccezione di Matomo e Logaholic.

Dal confronto delle matrici (Figg. 25-26-27) risulta che il prodotto che rispetta in maniera più precisa e puntuale tutti i requisiti fissati nel presente studio è *Matomo*, per il quale l'unico "Partial-Fit" è dato dall'impossibilità di creare report customizzati, se non agendo direttamente sul codice o acquistando il plug-in.

Più precisamente, il prodotto *Matomo* risulta più idoneo all'applicazione specifica perché presenta le seguenti caratteristiche.

- *Diffusione*: top 10 prodotti web analytics con il più alto traffico web secondo l'analisi di w3techs (Prodotto scelto per il progetto Web Analytics in Beta della PA).
- *Esercibilità*:
    - alta frequenza di aggiornamenti, dell'ordine di 1/mese;
    - installabile su stack standard;
    - roadmap di sviluppo ben definita;
    - documentazione per procedure di manutenzione ben dettagliata.
- *Modalità di tracking tramite web log*: retrocompatibilità con i siti del richiedente (web log) e possibilità di migliorare il livello di tracking (JavaScript).
- *Compatibilità con log prodotti dai web server IIS e Apache*.
- *Presenza supporto*: servizi di supporto professionali opzionali anche per versione on premise.
- *Diverse modalità di installazione*: versione SaaS (in cloud) oppure On Premise ad alta affidabilità su macchine fisiche o virtuali.
- *Open source*: community diffusa e attiva con forum di discussione che conta oltre 20.000 utenti.
- *100% GDPR Compliant* (General Data Protection Regulation), con collocazione data center e server in UE.

- *Nessun campionamento sui dati,* statistiche effettuate sulla totalità delle collezioni e non solamente su una parte di questi.
- *Dashboard* smart, modulare, personalizzabile e localizzata in italiano.
- *Profilazione per ruoli*: autorizzazioni ruoli per sito (admin, standard).
- *Licenze aggiuntive ad integrazione*: modulabile con componenti base gratis e componenti opzionali e pacchetto di supporto a pagamento (SAML).
- *Modalità progettuali di integrazione possibili*: setup compatibile con architetture del cliente; tuning del sistema dedicato per ogni sito; alta integrabilità dell'applicazione.
- *Piwik Pro:* branch proprietario che garantisce supporto dedicato e funzionalità enterprise.

3.3.2 Risultati dell'analisi di implementazione applicativo

L'implementazione dell'applicativo è stata caratterizzata dalla definizione e progettazione dell'architettura, implementata per tutti i componenti necessari al funzionamento dell'applicativo. I requisiti sono stati replicati per entrambi gli ambienti di collaudo e produzione.

Requisiti di Frontend necessari per il corretto funzionamento.

- Numero nodi Frontend: 2. La presenza dei due nodi garantisce l'affidabilità del sistema in caso di guasti, per questo è stato deciso di implementare un'architettura di Front-End che permetta di gestire eventuali disfunzioni di uno dei due nodi, instradando gli eventuali flussi di utenti sull'altro nodo nel caso in cui verifichino. Inoltre, questa decisione è stata presa anche per evitare sovraccaricare i due nodi durante il caricamento periodico dei dati.
- Hardware: Virtual CPU; 16 GB RAM; 50 GB disco per S.O. e log applicativi. I parametri operativi definiti sono stati concordati insieme al richiedente della ricerca; al fine di mantenere una soglia di sicurezza rispetto alla memoria a disco, si è optato invece per una RAM di 16 GB in quanto considerata dimensione minima per una idonea velocità di lettura lato Client.
- Software: Linux RedHat 7.9 OS; Apache HTTPD 2.4; PHP 7.x (7.2.5 o maggiore); Python 3.x (3.5 o maggiore); Matomo 4.x; Keytab-URL alias Matomo e hostname nodi. Gli applicativi sopracitati sono stati necessari poichè

Matomo richiede la presenza di un LAMP (Linux, Apache, mySQL, PHP/Pearl/Python) una "Solution Stack" (serie di componenti usate in maniera aggregata) abbastanza diffusa per il supporto delle applicazioni web. Il Keytab invece è un certificato che autorizza il protocollo di connessione HTTPS più sicuro del http. Questa è stata una scelta per soddisfare l'esigenza di aumentare il livello di sicurezza del sistema.

Requisiti per la cartella dei web log.

- NAS share 500 GB disco. Si è ritenuto che, prima che entri in vigore una policy di svecchiamento dei log presenti in cartella, ci sia la capacità di contenere almeno uno storico pari 6 mesi di file per ciascun sito (30 profili) corrispondente circa a 200 GB. Tale soglia è stata ottenuta grazie ad una stima effettuata sul peso dei dati pregressi. Il resto dello spazio è stato ritenuto necessario per il caricamento di file di prova, per eventuali azioni future e per politiche di Back-up dei file stessi.

Requisiti di Backend necessari per singola macchina.

- Numero nodi Backend: 2. La presenza dei due nodi è stata definita in base alla volontà di gestire l'affidabilità del sistema in caso di guasti. È per questo che è stato deciso di implementare un'architettura di Back-End di tipo "Master-Slave" che sta ad indicare che i dati vengono replicati dal nodo "Maestro" al nodo "Servo" in modo che ci sia maggiore affidabilità in caso di disfazioni.
- Hardware: 8 Virtual CPU; 16 GB RAM; 30 GB disco per S.O.; 500 GB disco per MySQL. Le grandezze scelte sono state concordate insieme al richiedente del servizio al fine di mantenere una soglia di sicurezza rispetto alla memoria a disco, si è optato invece per una RAM di 16 GB in quanto considerata dimensione minima per una idonea velocità di lettura lato Server in particolare per non sovraccaricare in fase di caricamento dei log.
- Software: Linux RedHat 7.9 OS; MySQL 8.x. Software necessari affinché il database sia operativo.

L'architettura è stata implementata in modo che ci siano due nodi sia lato Front-end sia lato Back-end che rispettivamente mettano in sicurezza i sistemi da eventuali guasti

e che permettano di alleggerire il carico di dati importati, spartendo il flusso tra i due nodi (Figg. 28 e 29).

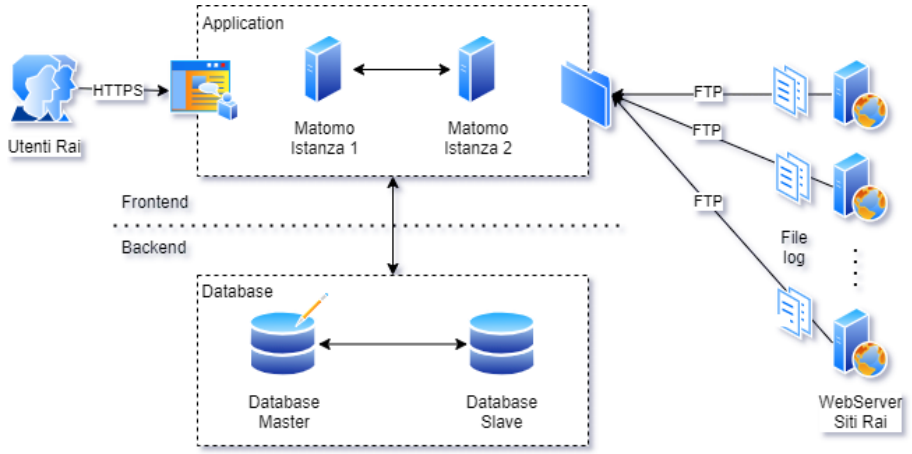

Figura 28. Architettura di ad alto livello designata.

Pertanto, sono stati definiti i seguenti layer per la soluzione proposta (Fig. 28).

Frontend: per il tracking, processing dei report e interfaccia utente e sul quale sono state installate le istanze applicative di Matomo.

Backend: che ospiterà il Database MySQL per l'archiviazione delle statistiche prodotte da Matomo.

Si è previsto, l'impiego di una locazione di memoria, condivisa tra i due nodi di Frontend, nella quale vanno a inserirsi tutti i file log delle attività prodotte dai siti, i quali sono inviati in batch al sistema tramite una procedura schedulata, con una frequenza di 24 h. L'architettura è stata studiata in modo che gli utenti di business si colleghino in VPN alla rete Intranet ed accedano in HTTP/HTTPS all'interfaccia web di Matomo tramite un URL alias (Fig. 29). Sul Frontend insistono due nodi virtuali collegati alla scheda di rete "VLAN FE" che comunicano attraverso un bilanciatore con logica Active/Active Round Robin con stickiness delle sessioni; il bilanciatore ha il compito quindi di inviare le nuove richieste esclusivamente al nodo attivo in caso di failure di un nodo e indirizza le richieste degli utenti ad uno dei due nodi Frontend, mentre le singole sessioni sono preservate su uno stesso nodo (Fig. 29).

Sui due nodi Frontend è montata una cartella NAS SHARE come file system che è la home folder che raccoglie i file log in arrivo dai WebServer (Fig. 29). Il flusso dello

schedulatore accede alla rete "VLAN FE" tramite il protocollo di rete SSH. I due nodi Frontend sono inoltre collegati alla scheda di rete "VLAN FE2BE" che è dedicata alla comunicazione con il Backend, livello in cui risiede il database MySQL (Fig. 29). Frontend e Backend sono, per ragioni di sicurezza, su due reti separate, rispettivamente su "VLAN FE2BE" e "VLAN BE" (Fig. 29). La comunicazione tra i due layer avviene mediante interfaccia di terzo livello (Routing). I due nodi di database, come precedentemente illustrato, sono configurati in modalità master/slave. Sul nodo slave è prevista una partizione NAS share per l'archiviazione di backup.

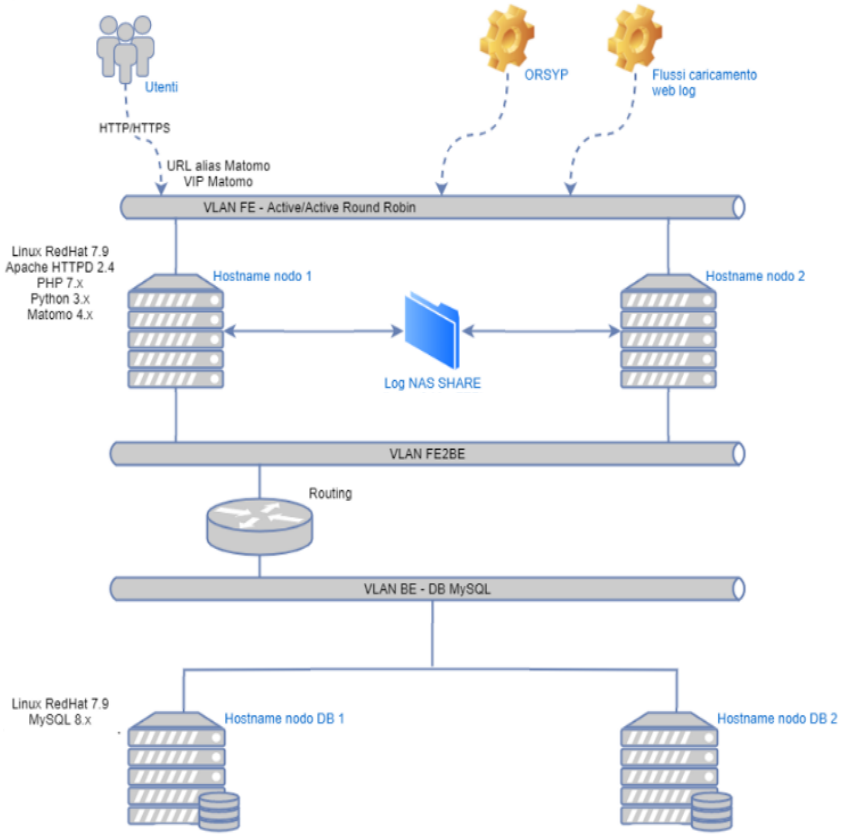

Figura 29. Architettura di dettaglio designata.

3.3.3 Risultati dell'analisi delle metriche

L'analisi delle metriche è stata effettuata tramite tecniche di regressione, mediante le quali è stato sviluppato un modello che possa prevedere con buona approssimazione la variabile "Bouce rate" in funzione dell'"Exit rate" (Fig. 30).

La prima tecnica utilizzata è stata la Regressione Lineare, che consiste nel minimizzare la somma residua dei quadrati tra la variabile target (Bounce rate) e le variabili di interesse. Il dataset è stato ripartito in un set di dati di test, corrispondente al 30% del dataset, mentre il restante è servito per generare il modello e sviluppare le predizioni. I risultati del modello generato mediante la regressione lineare e usando sia la libreria di Python che una API al software Stata, non evidenziano particolari differenze di prestazione tra i due strumenti. Quindi è stato possibile risalire alle seguenti equazioni della retta di regressione (Figg. 31-32).

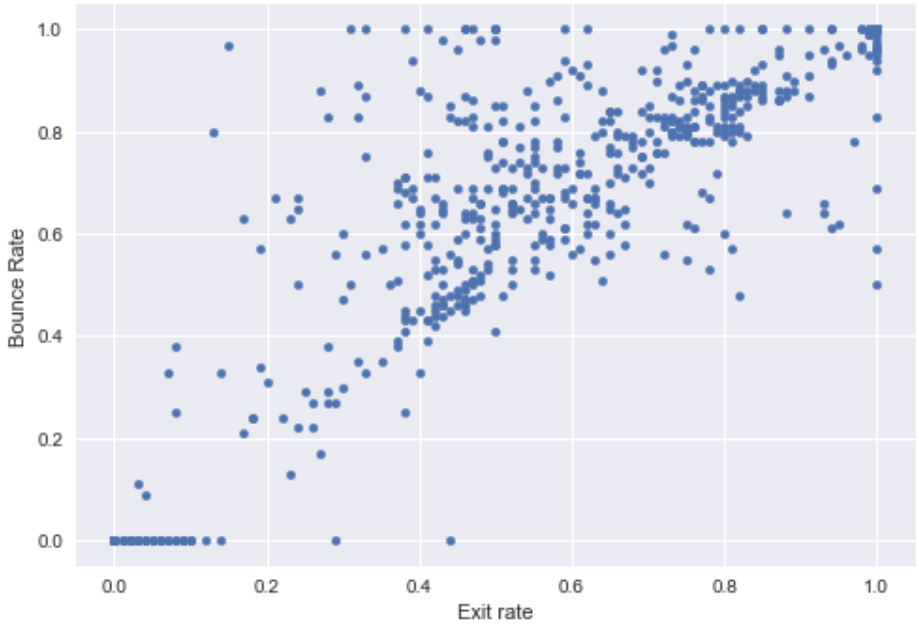

Figura 30. Grafico Scatter plot Exit rate in relazione a Bounce rate.

Risultati Sckit-learn:

$$Bouce\ rate = 0.16 + 0.886 \times Exit\ Rate$$

Risultati API stata:

$$Bouce\ rate = \ 0.154 + \ 0.885 \times Exit\ Rate$$

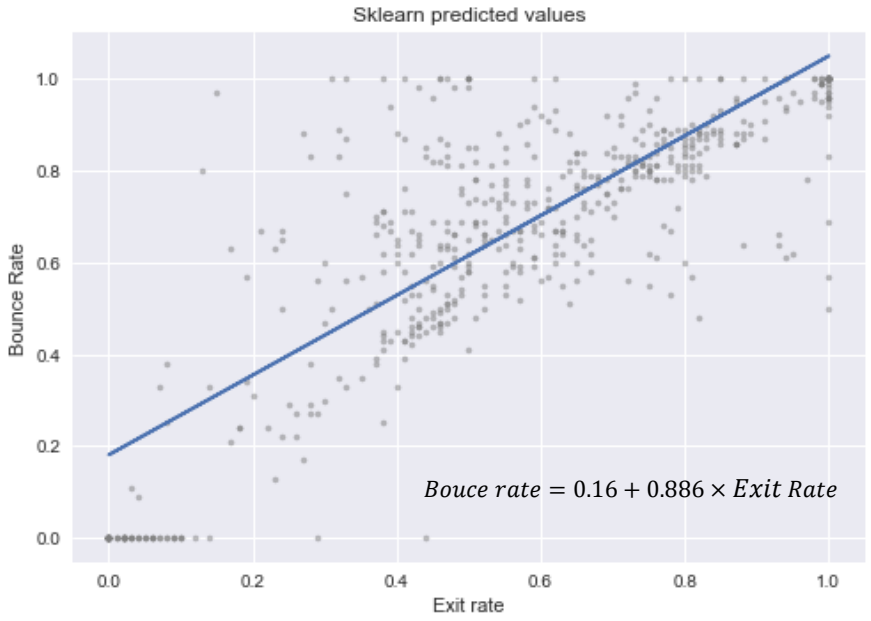

Figura 31. Regressione sckitlearn e relativa equazione.

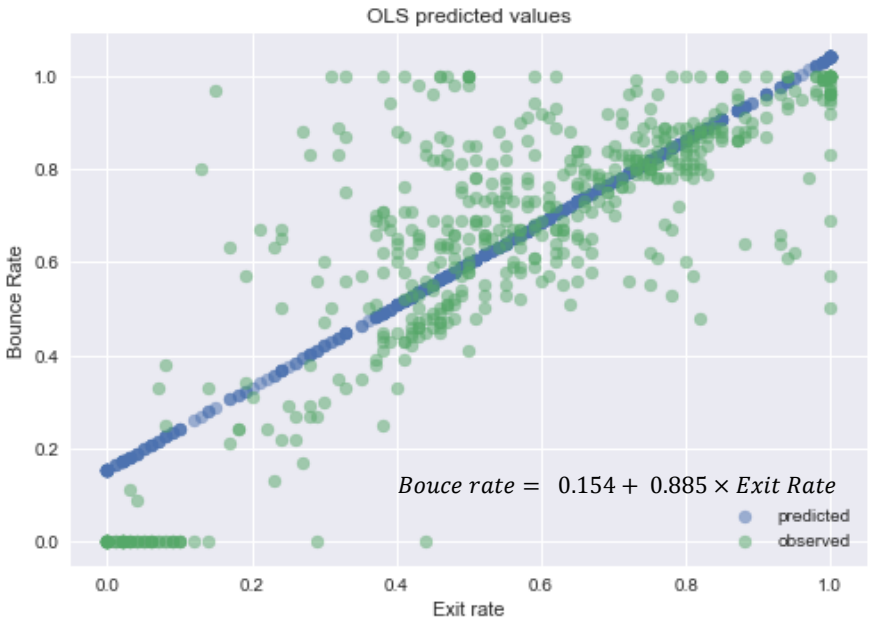

Figura 32. Regressione API a Stata e relativa equazione.

La valutazione della regressione implementata con Sckit-learn è stata effettuata utilizzando la "Cross validation", una tecnica che consiste nell'effettuare una valutazione incrociata di una specifica metrica misurata tra il dataset di train e quello di test. È stato utilizzato il metodo "cross_val_score ()", per definire una funzione che

valuti il punteggio della varianza spiegata in ambo i dataset (Baralis and Pasini, 2021), fornendo in output un valore con un margine di incertezza (Fig. 33).

```
#print the score
def print_score(reg, x, y):
    r2 = cross_val_score(reg, x[:, np.newaxis], y, cv=5, scoring='r2')
    print("R2: %0.2f (+/- %0.2f)" % (r2.mean(), r2.std() * 2))
```

Figura 33. Funzione di calcolo usata per la cross-validation di R^2.

Nel caso in esame, il punteggio di validità della regressione è valutato in maniera incrociata rispetto alla varianza spiegata (R^2) dalla regressione, cioè la percentuale di varianza della variabile dipendente che la variabile di interesse è in grado spiegare. I risultati sono l'intercetta della retta di regressione ed il suo coefficiente, definito "stimatore" (Figg. 31-32). L'equazione della R^2 ed i suoi elementi di composizione è:

$$R^2 = \frac{ESS}{TSS}$$

in cui:

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$

$$ESS = TSS - RSS.$$

- y_i sono i dati osservati;
- $\overline{y}$ è la loro media;
- $\hat{y}_i$ sono i dati stimati dal modello.

Come altro parametro di qualità del modello è stato calcolato il Root Mean Squared Error (RMSE), calcolato come deviazione standard degli scarti. La formula del RMSE, o ∈, è la seguente:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2}$$

dove *u* sono gli errori campionari e le *n* sono le osservazioni.

Il valore della varianza spiegata della regressione ottenuta usando Sckit-learn è 0.62, con una incertezza di (± 0.36). Il valore dell'errore residuo ∈ della regressione è 0.0235.

La seconda equazione è stata implementata calcolando lo stimatore OLS, che minimizza la somma dei quadrati residui e porta ad un'espressione in forma chiusa per il valore stimato del vettore di parametro sconosciuto β:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$$

dove *y* è l'i-esimo vettore dell'i-esima osservazione della variabile dipendente e *X* è la matrice in cui l'ij-esimo elemento rappresenta l'i-esima osservazione della j-esima variabile indipendente. In questo caso, non è stata effettuata una cross-validation ed è stata valutata direttamente tramite il comando richiamante l'API, ottenendo come risultati un $R^2 = 0.700$ ed un $\in = 0.039$.

Dato che la varianza spiegata varia tra 0 ed 1 e l'errore, in entrambi i casi, è abbastanza basso (varia tra 0 e +∞) si può dire che i modelli siano abbastanza attendibili (Stock et Watson, 2012).

Dopo la regressione lineare è stata valutata anche la possibilità di usare la regressione polinomiale (Fig. 34), nella quale si tende a tenere il grado del polinomio il più basso possibile, per non incorrere nel fenomeno dell'"Overfitting", che si verifica quando il modello è così accurato sul set di dati che tende a perdere la caratteristica di generalizzazione (Baralis and Cerquitelli, 2021). Tuttavia, anche alzando il grado del polinomio, la varianza spiegata non cresce in maniera significativa, rimane sempre sul 0.69-0.70 riscontrato anche nella regressione lineare.

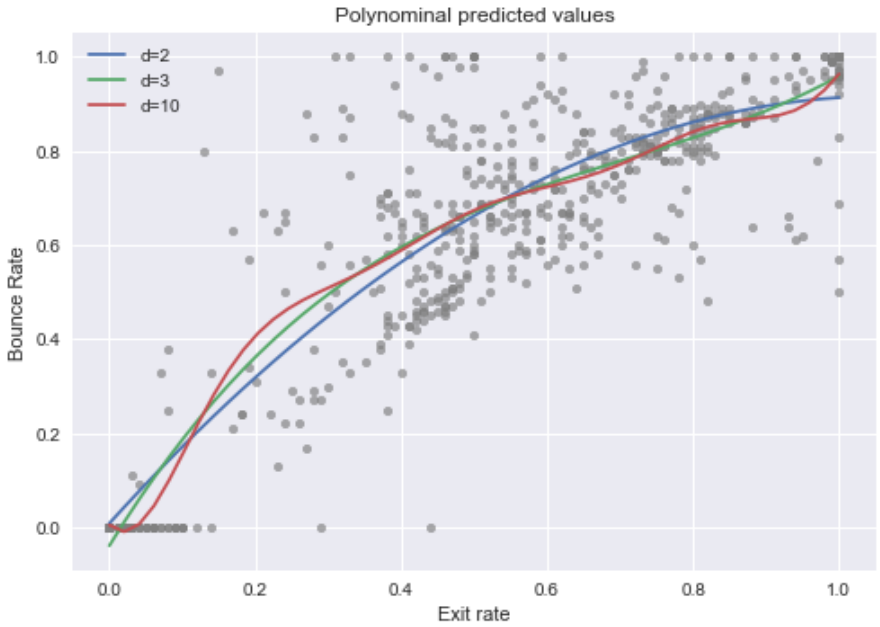

Figura 34. Regressione polinomiale di grado 2°,3° e 10°.

Infine, si riportano i risultati dell'analisi di correlazione, che è stata effettuata per identificare le variabili che siano correlate con l'attributo "Exit rate", in modo da non inserirle nelle regressioni successive, che sono regressioni "multiple". Questo tipo di regressioni sono utili per consolidare oppure smentire il modello proposto dalla regressione singola e consistono nell'aggiungere una serie di regressori che, generalmente, contribuiscono a migliorare il modello e ad evitare la distorsione da variabile omessa, ovvero l'errore di generare un modello non completo, il cui completamento è dato dall'aggiunta di quella variabile (Stock et Watson, 2012).
Pertanto, è stata elaborata una regressione che tenga conto, nel modello lineare precedentemente studiato, delle seguenti variabili: *Avg.time on page, Actions after entering here, Avg. After load time e min_time_server*.

Tabella 3. Risultati della regressione multipla con l'aggiunta delle variabili: Avg.time on page, Actions after entering here, Avg. After load time e min_time_server.

OLS Regressions

	Model 1	Model 2	Model 3	Model 4
const	0.14***	0.14***	0.11***	0.11***
	(0.02)	(0.02)	(0.02)	(0.02)
Exit rate	0.89***	0.89***	0.88***	0.88***
	(0.02)	(0.02)	(0.02)	(0.02)
Avg. time on page	0.08**	0.09***	0.15***	0.15***
	(0.03)	(0.03)	(0.03)	(0.03)
Actions after entering here		-0.11**	-0.08*	-0.07*
		(0.04)	(0.04)	(0.04)
Avg. page load time(s)			0.42***	0.35***
			(0.07)	(0.09)
min_time_server				0.07
				(0.05)
R-squared	0.70	0.71	0.72	0.72
R-squared Adj.	0.70	0.70	0.72	0.72
R-squared	0.70	0.71	0.72	0.72
No. observations	567	567	567	567

Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01

I risultati del nuovo modello dimostrano che, non è stata commessa alcuna distorsione, infatti viene unicamente spiegato il 2% in più della varianza della regressione singola (Tab. 3).
Successivamente allo sviluppo dei modelli è stato svolto un processo di *Unsupervised learing*, che ha permesso di ottenere risultati ancora più significativi. In particolare, sono stati utilizzati due algoritmi di clustering, ovvero strumenti in grado di raggruppare

oggetti in base ad alcune proprietà comuni relative alla distanza (Cerquitelli T et Al., 2021). Il primo dei due strumenti è stato *l'algoritmo Kmeans* appartenente alla famiglia dei "centroid-based", termine che indica gli algoritmi che sono basati sull'individuazione dei centroidi (media di tutti i punti del cluster) oppure dei medoidi (punti più rappresentativi del cluster), intesi come i punti in cui sono centrati i cluster di oggetti.

I suddetti algoritmi tendono a creare i gruppi simili, in base alla minimizzazione della distanza dai centroidi/medoidi Kmeans in particolare richiede in input la quantità di cluster (*K*) da individuare e seleziona in maniera randomica il centro, procedendo alla minimizzazione delle distanze in modo da creare il numero di cluster desiderato. Si procede quindi alla selezione empirica del numero di cluster da far identificare all'algoritmo, valutandone ad ogni tentativo alcune metriche di coesione intra ed inter-cluster. In maniera analoga si procede per il secondo algoritmo utilizzato, *l'algoritmo Dbscan* ma secondo una procedura differente. Questo secondo algoritmo è appartenente alla famiglia dei "density-based" che producono il cluster sulla base della densità della regione di punti analizzata; esso riceve in input il raggio massimo che la regione di interesse deve avere (Eps) ed il numero minimo di oggetti per cluster (MinPoints). Il funzionamento consiste nell'aggregare tutti i punti nella regione compresa dall'area coperta da Eps, aggregando un minimo di MinPoints punti e individuando "Core points", "Border points" e "Noise points" (Tan et al., 2006) (Fig. 35).

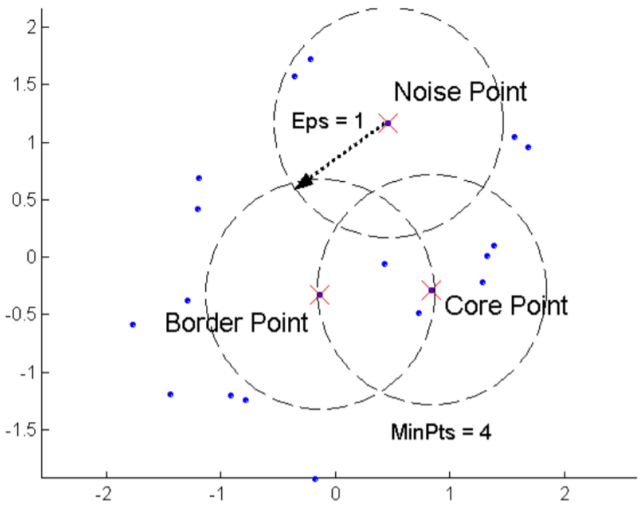

Figura 35. Scatter plot identificazione Core points", "Border points" e "Noise points".

Per questa tipologia di clustering è generalmente più complicato trovare la migliore combinazione di parametri, ma esistono delle soluzioni che aiutano sensibilmente nella scelta, in particolare della Eps; una di queste, è rappresentare le distanze ordinate, in maniera ascendente del k-esimo oggetto più vicino, in funzione della distanza del k-esimo oggetto più vicino, ottenendo così, nella maggior parte dei casi, un grafico che presenta un *gomito* in cui conviene scegliere il raggio, che rappresenta la distanza alla quale si trova la maggior parte dei k-esimi vicini.

Lo scopo di questa fase dell'elaborazione è stato quello di individuare gruppi di pagine che spieghino, con le dovute approssimazioni, il comportamento degli utenti, in modo da valutare l'utilità/performance percepita dai visitatori. Sono state quindi prese in esame le distribuzioni delle variabili "Bounces" e "Avg. time on page" a cui sono state integrate successivamente informazioni relative alla variabile "Actions after entering here", performando un clustering su tutte le 3 dimensioni. Le valutazioni sulla bontà dei cluster sono state poi effettuate secondo due metriche: la Silhouette e il Davies-Bouldin Index (DBI). La prima è calcolata usando la distanza media interna tra punti del cluster (a) e la distanza media del cluster più vicino (b). L'indice è poi calcolato come rapporto della differenza di *b* ed *a* con il massimo tra i due; questa misura viene calcolata per ogni cluster e poi ne viene fatta una "Silhouette media", che rappresenta appunto la metrica utilizzata; esso varia tra -1 ed 1 ed indica la bontà del cluster al suo crescere (BSD License, Scikit-learn Developers, 2007-2021).

Il DBI invece, è definito come la media della misura della similarità di ciascun cluster con il suo cluster più simile, dove la similarità rappresenta il rapporto tra le distanze interne del cluster e le distanze tra cluster. I cluster che sono più lontani e meno dispersi risulteranno i migliori. Questo indice varia tra 0 e $+\infty$ ed è tanto migliore quanto più si avvicina allo zero (BSD License, Scikit-learn Developers, 2007-2021).

Il primo step è stato l'applicazione di Kmeans mediante la stessa libreria di Python utilizzata per le regressioni, che ha permesso di trovare dei cluster ben definiti tra "Bounces" e "Avg. time on page" in cui spiccano, particolarmente per qualità delle metriche, i clustering con K=10, K=5 e K=3 (Figg. 37-38-39). I cluster individuati presentano le seguenti coppie di Silhouette e DBI: (0.55259, 0.547336), (0.548935, 0.604253) e (0.539412, 0.636785) (Fig. 36). Le coppie di valori appena ciatate, fanno ben

sperare in termini di affidabilità del modello in cui i cluster indicati sono abbastaza coesi e distanziati.

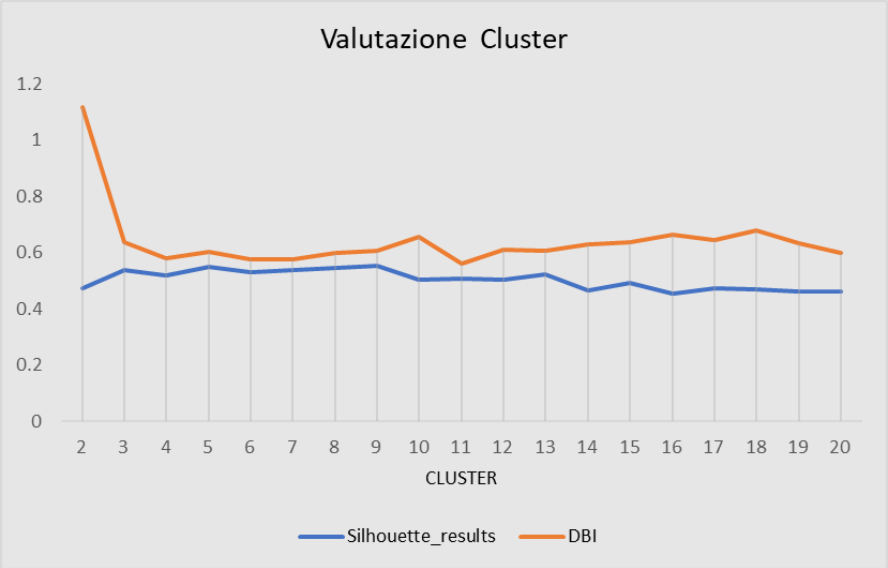

Figura 36. Silhouette e DBI kmeans 2D a confronto.

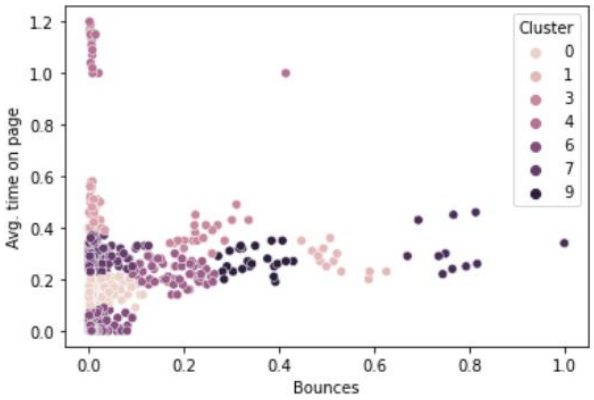

Figura 37. Grafico scatter k-means K=10.

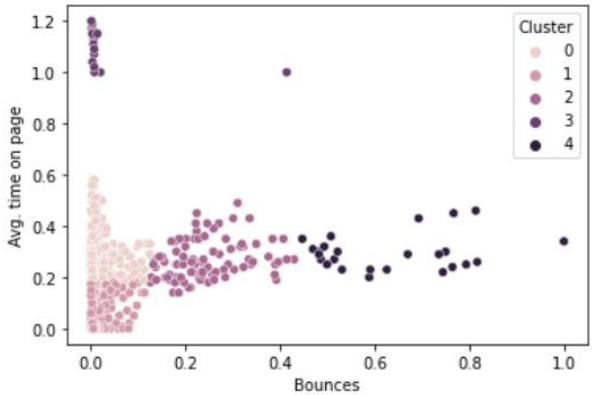

Figura 38. Grafico scatter k-means K=5.

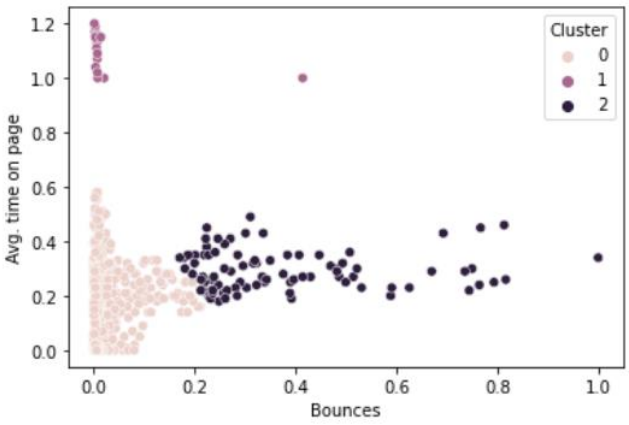

Figura 39. Grafico scatter k-means K=3.

Dai cluster ottenuti, dividendo le pagine per classi in cui gli utenti rimangono più o meno a lungo prima di uscire, sono stati ottenuti gruppi che identificano le pagine in categorie; tuttavia, è necessario acquisire informazioni più specifiche riferite a come gli utenti considerino o usino tali pagine, con lo scopo di limitare le possibili supposizioni corrette per l'interpretazione. Pertanto, è stato analizzato il clustering ed inserite in esso un'altra variabile che indichi il numero delle azioni svolte dagli utenti. Questa procedura ha permesso di incrementare sensibilmente lo spessore delle supposizioni riguardanti l'etichettatura delle classi identificate; è stato quindi utilizzato lo stesso algoritmo ma su centroidi nello spazio e non più in un'area.

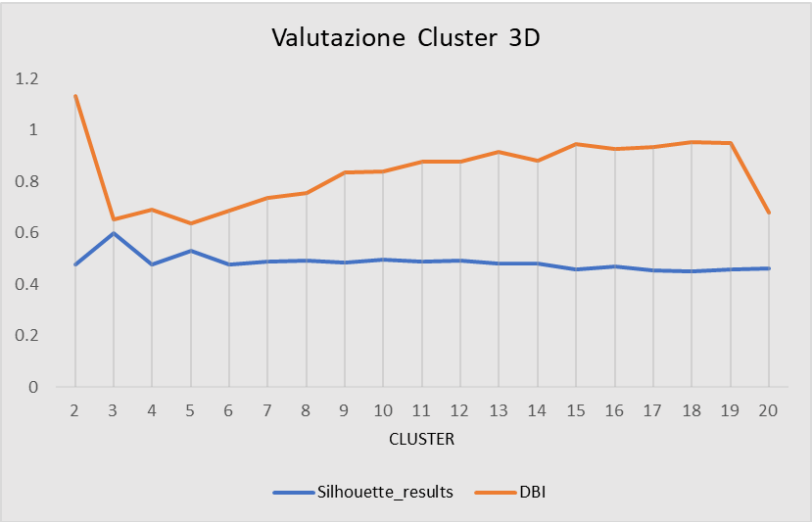

Figura 40. Silhouette e DBI kmeans 3D a confronto.

Le informazioni derivanti dal clustering su 3 variabili si dimostrano non solo chiarificatrici, poichè restringono la scelta tra K=3 e K=5 (Fig. 40), ma consentono anche di interpretare con più chiarezza i gruppi di pagine. Le metriche di valutazione risultano superiori al 2D plot e, dato la scelta di qualità risulta indifferente tra i due, si è scelto di analizzare la divisione per 5 cluster (Fig. 42), in quanto dà la possibilità di analizzare più scenari.

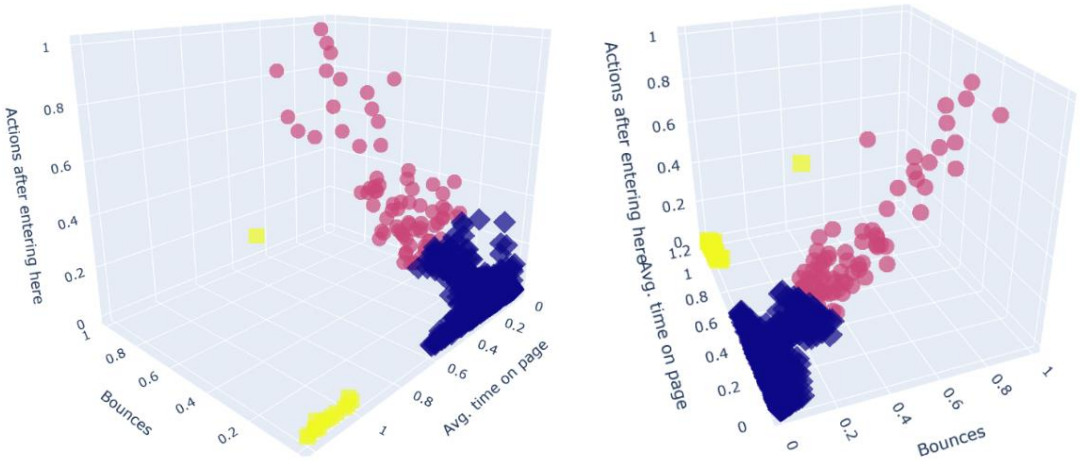

Figura 41. Grafici scatter Kmeans 3D per K=3.

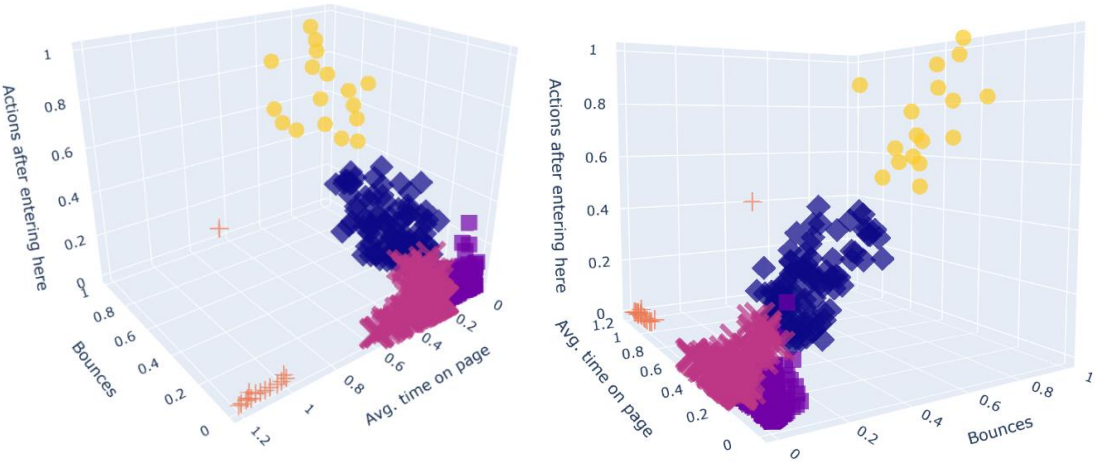

Figura 42. Grafici scatter Kmeans 3D per K=5.

In base ai risultati di Kmeans per K=5 (Fig. 44) sono state definite le classi di seguito riportate in tabella 4. Si nota che nel sito analizzato, in particolare nel cluster 4 corrispondente alla categoria "C5", è presente la pagina base delle FAQ, in cui si rileva il massimo tempo di stazionamento dell'utente di più in assoluto effettuando un basso numero di azioni, indice di un buon riconoscimento del cluster da parte dell'algoritmo (Fig. 43).

Tabella 4. Classi identificate dai cluster.

CLASSE	ATTRIBUTI	SPIEGAZIONE	AZIONE CORRETTIVA
C1	ALTO BOUNCES ALTO NUMERO AZIONI MEDIO TEMPO STAZIONAMENTO	Gli utenti sono su questa pagnia fino alla fine della navigazione, per tempi medi, facendo numerose operazioni su di questa prima di lasciarla (sono disposti a rimanere per un tempo medio al fine di comletare le azioni compiute su di essa)	Nessuna
C2	MEDIO BOUNCES MEDIO NUMERO AZIONI MEDIO TEMPO STAZIONAMENTO	Essendo tutto medio risulta difficile valuarne l'utilità o il buon funzionameno. essendo tutto nella media sembra che le pagine siano comprese dagli utenti	Nessuna
C3	BASSO BOUNCES BASSO NUMERO AZIONI BASSO TEMPO STAZIONAMENTO	Gli utenti non utilizzano la pagina fino alla fine della loro navigazione, non ci fanno molte azioni e rimangono al di sopra per poco. Qesto ci fa capire che molti utenti si posizionano su di esse per errore opure siano delle pagine di transizione (potrebbero essere sovrabbondanti).	Monitorare/Sostituire
C4	BASSO BOUNCES BASSO NUMERO AZIONI MEDIO TEMPO STAZIONAMENTO	Gli utenti stazionano sulla pagina poco e non rimangono su di essa fino alla fine, il tempo medio di stazionamento medio può portare a pensare che la pagine sia poco rilevante e si potrebbe integrare in altre o eliminarla.	Monitorare/Sostituire
C5	BASSO BOUNCES BASSO NUMERO AZIONI ALTO TEMPO STAZIONAMENTO	Gli utenti entrano e non rimangono solo su questa pagina facendo poche azioni questo può significare o che la pagina è solo informativa oppure che è fatta in maniera poco chiara e non viene compresa da chi la utilizza.	Monitorare/Sostituire

Tabella 5. Commenti di ciascuna classe.

CLASSE	COMMENTO
C1	la pagina è funzionale, nonostante ci siano numerose azioni da fare non costringe l'utente a rimanere tutto il tempo.
C2	Impossibile valutare.
C3	Valutazione singola pagina in base a metriche che possono indicare l'utilità della pagina e valutane il Bounce rate per capire se gli utenti ci vanno per sbaglio, in quel caso sarebbe bene fare un'analisi del traffico per valutarne le modalità di collegamento
C4	Valutazione singola pagina in base a metriche che possono indicare l'utilità della pagina .
C5	Due possibilità: - pagina informativa, l'utente non fa molte azioni e legge il contenuto (es. F.A.Q) in ogni caso data la bassa % di Bounces si immagina che l'utente non abbia risolto i suoi dubbi con questa pagina. - pagina progettata male, e gli utenti non la capiscono.

Allo stesso modo, ispezionando quali FAQ sono presenti nel cluster 0, relativo alla categoria "C1", possiamo ritrovare le FAQ che risolvono meglio i dubbi degli utenti (Tab. 5). Si nota anche che nelle classi più della metà delle FAQ o sezioni dedicate ad esse non sono risolutrici dei dubbi dell'utente (Fig. 44). Ciò porta a proporre la revisione delle FAQ situate nelle classi 4 e 3 (Fig. 44).

4	/[redacted]/faq.aspx
4	/[redacted]/faq.aspx

Figura 43. Ricerca parola "FAQ" nel cluster 4 (C5).

Un'altra soluzione migliorativa può essere una classificazione spalmata su più siti, così da poter arrivare a predirne la positività o negatività della FAQ e analizzarle nel tempo. L'ultimo algoritmo utilizzato è stato DBscan che non ha riscontrato risultati migliori del precedente, individuando soltanto 2 cluster, anche se con dei valori di qualità buoni (Silhouette: 0.5884); tuttavia, l'individuazione di così pochi cluster non è funzionale allo studio.

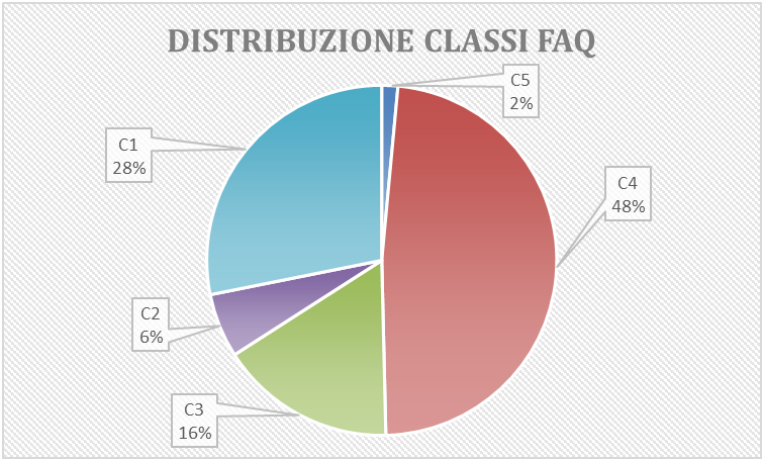

Figura 44. Distribuzione delle pagine "FAQ" nelle classi individuate.

Di seguito sono riportati il grafico scatter 3D (Fig. 46) ed il grafico dei K-esimi vicini utilizzato per la scelta dei parametri (Fig. 45). Uno dei motivi per cui questo algoritmo dà questi risultati è dovuto al suo impiego che, di solito, sfrutta il cluster 0 per la raccolta

del rumore durante outlier detection. Infatti, l'algoritmo tende a tenere i core point ed i punti di frontiera nei cluster ed il rumore nel cluster 0.

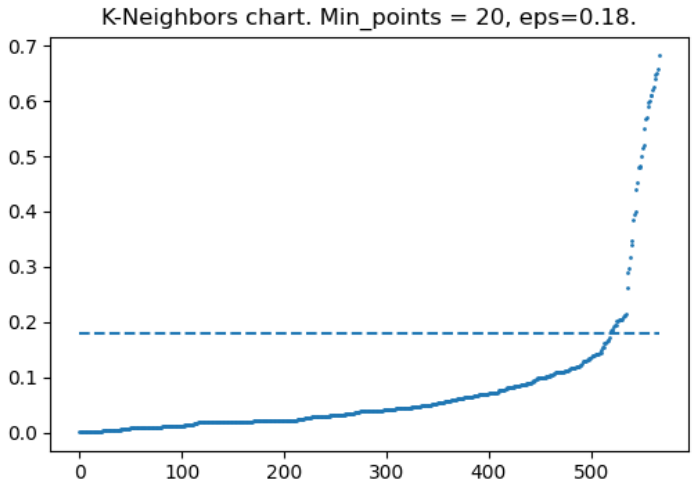

Figura 45. Grafico del 20° vicino dataset analizzato.

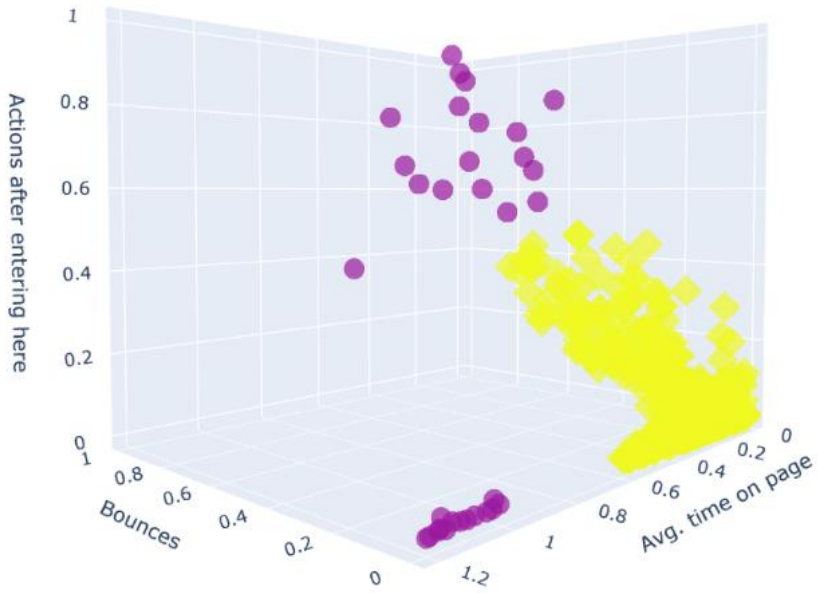

Figura 46. Grafico scatter DBscan 3D per min_points = 20, eps = 0.18.

CONCLUSIONI

Informatizzare un processo significa custodire e valorizzare un patrimonio di valore inestimabile, in modo da utilizzarlo efficacemente, per far sviluppare strategie aziendali che rendano le realtà imprenditoriali competitive, sia dal punto di vista qualitativo che della produttività, in un mercato globale sempre più esigente che impone garanzie ed evoluzione tecnologica continue. Il presente lavoro di tesi evidenzia che una corretta applicazione della Data Science all'utenza industriale consente di perseguire con successo questi obiettivi.

In quest'ottica, si inserisce la cooperazione fra ricerca e industria, che rappresenta un'opportunità imperdibile per le aziende, ma è anche un'occasione per le piccole e grandi imprese di settori afferenti al mondo dell'innovazione in generale: il modello collaborativo promuove il confronto fra ricercatori e mondo imprenditoriale, agevolando la sperimentazione in campo e consentendo di accedere ad un livello economico più elevato. La creazione di "Gruppi Operativi" in materia di produttività e sostenibilità, costituiti da diversi portatori d'interessi (imprese, ricercatori, consulenti, distretti produttivi e tecnologici, centri di sperimentazione, ecc.) consente di individuare le esigenze di innovazione e sviluppo prioritarie per l'imprenditoria.

Il lavoro di ricerca e sperimentazione effettuato, ha consentito di implementare un percorso di applicazione di Data Science ad un noto broadcaster televisivo operante nel settore dei media, con la necessità di sostituire un prodotto per l'analisi web delle statistiche intranet ed internet di alcuni profili.

Sono stati definiti criteri e soluzioni progettuali, di scelta e di miglioramento del processo di utilizzo del nuovo applicativo. La sperimentazione è stata svolta integrando varie aree di Data Science mediante tecniche di Machine Learning in ambito Web Analytics; con questo approccio scientifico, è stato possibile analizzare i temi del flusso degli utenti sulle pagine web, tenendo anche conto delle interazioni nel campo intranet. I risultati hanno consentito di individuare l'applicativo più indicato per l'ecosistema del fruitore della ricerca, di sviluppare una architettura idonea al funzionamento del prodotto, nonché di definire, mediante analisi supervisionata e non, le metriche principalmente monitorate sui siti corporate, proponendo un modello rappresentante due importanti variabili analizzate nel mondo della Web Analytics. È stato proposto

inoltre anche un metodo di identificazione delle diverse classi di pagine, in base alla concezione di funzionalità che ha l'utente corporate.

Da quanto innanzi, si può affermare anche che, i risultati della presente ricerca contribuiscono anche a valorizzare il coraggio imprenditoriale di cambiare e innovare le aziende, lo sconfinato potere dei benefici legati a un adeguato percorso di digitalizzazione, nonché a ridurre la reticenza di molte imprese a modificare processi radicati anche se, talvolta, è l'ambiente circostante che costringe al cambiamento.

Nonostante il presente studio abbia caratterizzato una specifica realtà economico-produttiva, molti risultati ottenuti e la metodologia adottata si possono considerare generalizzabili e applicabili in tutti i casi di analisi ed implementazione di prodotti di analytics, e non unicamente nell'ambito delle telecomunicazioni.

Bibliografia

Schutt R. and O'Neil C. (2013). Doing Data Science: Straight Talk from the Frontline. O'Reilly Media, ISBN: 9781449358655, 1449358659, 1005 Gravenstein Highway North, Sebastopol, CA. pp. 1-365.

Arikan A. (2008). Multichannel Marketing: Metrics and Methods for On and Offline Success. Wiley Publishing, Indiana.

Cutler M. and Sterne J. (2000), E-Metrics: Business Metrics for The New economy, Copyright 2000, NetGenesis Corp. Chicago, IL. pp. 1-60.

Kaushik A. (2007). Web Analytics: An Hour a Day. Wiley Publishing, Indiana. pp.1-418.

Phippen A., Sheppad L., Furnell S. (2004). "A practical evaluation of web analytics". Internet Research, 14, 4, pp. 284-293.

Sterne Jim (2002). Web Metrics: Proven Methods for Measuring Web Site Success. John Wiley & Sons, New York. pp. 201-238.

Duraisamy G, Atan R. (2013). Requirement traceability matrix through ocumentation for scrum methodology, Journal of Theoretical and Applied Information Technology. ISSN 1992-8645, ESSN: 1817-3195 Selangor, Malaysia pp. 154-159.

Waisberg D. and Kaushik A. (2009). Web Analytics 2.0: Empowering Customer Centricity. Copyright 2009 SEMJ.org, 2, Issue 1.

Oyama K., Takeuchi A, Ming H., Chang C. (2011). "A Concept Lattice for Recognition of User Problems in Real User Monitoring", Proceedings - Asia-Pacific Software Engineering Conference, APSEC, Ho Chi Minh, Vietnam. pp. 59-66.

Tan P., Steinbach M., Kumar V. (2006). Introduction to Data Mining. Copyright 2006 Pearson Addison-Wesley, Boston, US. pp. 1-145.

Bethaz P., Cavaglion S., Cricelli S., Liore E., Manfredi E., Salio S., Regalia A., Conicella F., Greco S., Cerquitelli T. (2021). Empowering Commercial Vehicles through Data-Driven Methodologies. *Electronics* 2021, 10, 2381.

Giordano D., Mellia M., Cerquitelli T. (2021). K-MDTSC: K-Multi-Dimensional Time-Series Clustering Algorithm. Electronics 2021 10, 1166.

Attanasio G., Giobergia F., Pasini A., Ventura F., Baralis E., Cagliero L., Garza P., Apiletti D., Cerquitelli T., Chiusano S. (2020). DSLE: A Smart Platform for Designing Data Science Competitions. pp. 133-142.

Stock J. H. and Watson M. W. (2012). Introducione all'econometria. Pearson Milano-Torino, Italia, pp. 101-123.

Clemente Rivera B. E. (2021). Segmentación de lectores digitales registrados de un sitio web informativo con el algoritmo de análisis Cluster k-means. Trabajo de suficiencia profesional para optar el título de ingeniero estadístico informático. Universidad nacional agraria, Lima, Perú, pp. 1-28.

Sujo J.M. and Ruano V.J. (2019), Data Science application for the news profitability study of a digital media, Expuso su Trabajo de Final de Máster, Escola Tècnica Superior d'Enginyeria en Electrònica i Informàtica La Salle, Barcellona, Spain. Pp. 1-78.

He Yu, Bach K. (2018). Data Analysis for the Mobile Application of the selfBACK Decision Support System. Norwegian University of Science and Technology, Trondheim, Norway. pp. 1-100.

Parwez M. S., Rawat D. B., Garuba M. (2017), Big data analytics for user-activity analysis and user-anomaly detection in mobile wireless network. IEEE Transactions On Industrial Informatics 13, Issue: 4. pp. 2058-2065.

Ma H., Cao H., Yang Q., Chen E., Tian. J. (2012). A habit mining approach for discovering similar mobile users. In: International Conference on World Wide Web, Association for Computing Machinery Lyon, France. pp 231–240.

Géczy P., Izumi N., Akaho S., Hasida K. (2009) Analytics and Management of Collaborative Intranets. In: Bertino E., Joshi J.B.D. (eds) Collaborative Computing: Networking, Applications and Worksharing. CollaborateCom 2008. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 10. Springer, Berlin, Heidelberg. pp. 623-631.

Schneider W., Steinhoff A. (2021), Applying Web Analytics Tools in the Context of Enterprise Social Software, Technische Universität München, Munich, Germany.

Sculley D., Malkin R. G., Sugato B., Bayardo J. R. (2009). Predicting bounce rates in sponsored search advertisements. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09). Association for Computing Machinery, New York, NY, USA, 1325–1334.

Sitografia

https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/?sh=1ca99ea055cf

https://www.delltechnologies.com/asset/en-ee/products/converged-infrastructure/legal-pricing/eosl-for-converged-systems.pdf

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_samples.html

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html

https://support.vidyocloud.com/hc/en-us/articles/235905908-Vidyo-Product-End-of-Life-EOL-Policy

https://www.panorama.com/blog/the-role-of-data-analytics-in-the-telecom-industry/#:~:text=Telecom%20Data%20Analytics%20Allows%20Better,their%20data%20across%20departmental%20lines

https://medium.com/@springboard_ind/data-science-vs-data-analytics-how-to-decide-which-one-is-right-for-you-41e7bdec080e consultato nel 2021.

https://www.slideshare.net/JamesChristopher2/essential-data-science-for-product-designers-and-nonscientists

https://svitla.com/blog/data-science-vs-data-analytics

https://www.ultraedit.com/end-of-life.html

https://www.upgrad.com/blog/data-science-vs-data-analytics/

https://www.techopedia.com/definition/30296/analytics

https://glossary.matomo.org/

https://www.trustradius.com/tag-management

https://ppcmasterminds.com/analytics-web/setup-google-tag-manager/

https://www.ensighten.com/training/manage-technical/

https://w3techs.com/technologies/overview/tag_manager

https://stackify.com/what-is-real-user-monitoring/

https://www.dynatrace.com/platform/real-user-monitoring/

https://www.digital4.biz/marketing/customer-experience-definizione-modello/

https://www.digital4.biz/marketing/trend-digital-marketing-2020/

https://matomo.org/intranet-analytics/