# ***CAPITOLO 1: STATO DELL’ARTE***

## ***1.1 BUSINESS INTELLIGENCE***

Con il termine Business Intelligence (BI) ci si riferisce ad una serie di processi aziendali che ruotano attorno ai dati, con operazioni di raccolta, elaborazione, analisi, cui scopo è quello di produrre informazioni al servizio del management strategico e tattico, che trova supporto analitico, storico e previsionale alle proprie decisioni. La BI è stata collocata altresì nel sottoinsieme operativo, poiché sta assumendo un ruolo sempre più importante anche nelle normali attività giornaliere delle aziende.

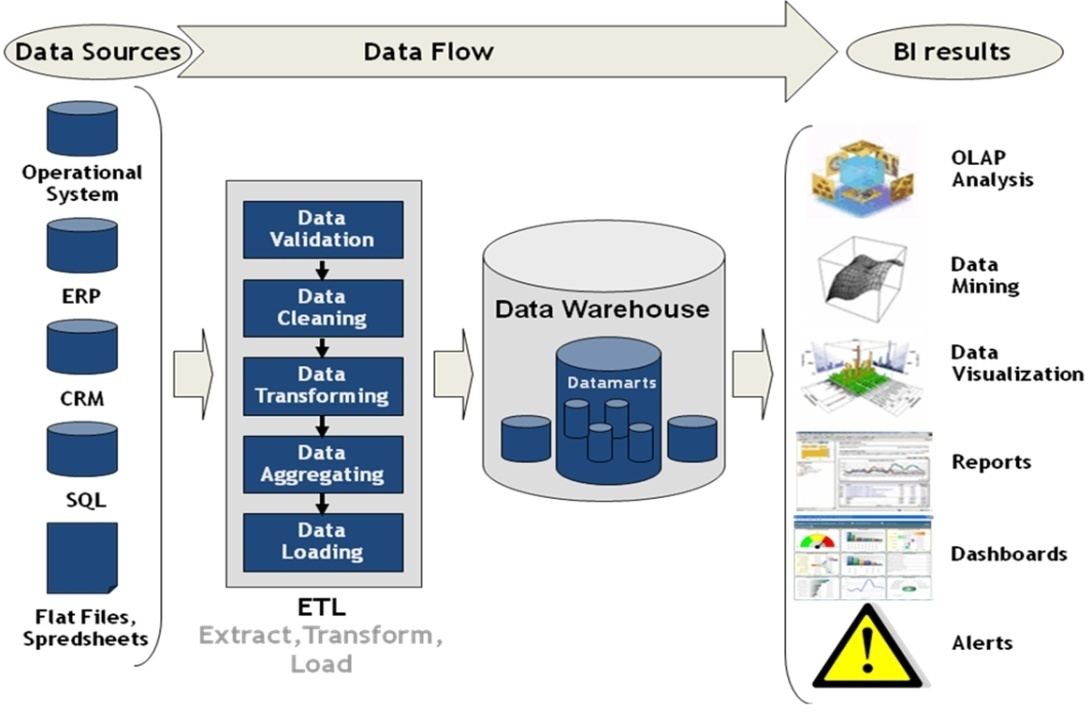


Figura 3: Business Intelligence Framework

Nell’ambito lavorativo moderno, il cui scopo principale è fare business e diventare leader di mercato, le aziende si trovano sempre più frequentemente a confrontarsi con realtà differenti dalla propria. Ciò avviene tramite l’analisi del comportamento dei competitors facenti parte dello stesso settore e lo studio del mercato in cui si trovano.

L’adozione della BI da parte delle imprese permette una conoscenza più approfondita non solo di loro stesse ma anche del mercato di riferimento.

Nel periodo attuale, il “cambiamento” è all’ordine del giorno, pertanto saper leggere in anticipo le tendenze dei mercati è un fattore competitivo a cui non si può e non si deve rinunciare.

Data l’elevata mole di dati generata ogni giorno, diventa necessario trovare un metodo che:

* Permetta di raccogliere e processare dati ad alta velocità (sempre più spesso si parla di processi Real-Time).
* Fornisce un servizio di pulizia del dato stesso, eliminando dati sporchi, ridondanti o errati tramite i processi di ETL “Extraction, Transformation & Loading” che prelevano i dati dai sistemi alimentanti (ERP, fogli Excel etc.) e li portano nel DWH certificandoli attraverso processi di Data Quality. Questo processo sarà spiegato in maniera specifica nel cap.2;
* Definisce un sistema consolidato e stabile di memorizzazione per i dati certificati (Data Warehouse);
* Trasforma l’informazione in fonte di conoscenza attraverso analisi di business sui dati stessi, determinando nuovi KPI.

I Big Data provengono da diverse fonti, sia interne che esterne, spesso sono in formati differenti e risiedono in posizioni multiple in numerosi sistemi legacy e altre applicazioni. I dati possono essere strutturati (dati conservati in Database relazionali, organizzati secondo schemi e tabelle rigide), non strutturati (dati conservati senza alcuno schema come forme libere di testo tra cui articoli e parti di e-mail, audio senza tag, immagini e video) o semi-strutturati (dati che presentano caratteristiche sia di quelli strutturati che di quelli non strutturati; un esempio è rappresentato dai file compilati con sintassi XML per i quali non ci sono ci sono limiti strutturali all’inserimento dei dati, ma le informazioni vengono organizzate secondo logiche strutturate e interoperabili). Dopo che i dati sono stati uniti, questi hanno bisogno di essere processati o trasformati, essendo in uno stato grezzo.

Il passo successivo consiste nella scelta della piattaforma e della tecnologia da utilizzare per le applicazioni di Big Data Analytics che includono queries, reports, OLAP e Data mining e alla visualizzazione, compresa in tutte queste applicazioni [22].

Un ruolo centrale in quest’ambito viene svolto dai Big Data Analytics, tecnologie di Business Intelligence & Analytics basate su come:

* *CRM & Customer Analytics*: soluzioni e tecnologie che raccolgono, organizzano e sintetizzano i dati dei clienti per aiutare le organizzazioni a risolvere i problemi di business riguardanti i consumatori attraverso tool, dashboard, portali e altri metodi negli ambiti di Marketing, Sales e Customer Service; i consumatori vengono poi segmentarli in gruppi sulla base dei comportamenti adottati, implementare azioni di Marketing personalizzate e determinare trend generali;
* *Predictive Analytics*: Analytics avanzati che implementano tecniche quali la regressione, i modelli predittivi e la statistica per analizzare i dati e i contenuti e rispondere alle domande “Cosa succederà” o “Cosa accadrà molto probabilmente?”;
* Social Analytics: tools che estraggono, analizzano e sintetizzano automaticamente i contenuti generati dagli utenti online. Questa tecnologia verrà descritto in modo approfondito nel successivo capitolo;
* *Text Analytics*: processo di estrazione delle informazioni dai testi, utilizzato per diversi scopi, tra cui il *riepilogo*, ovvero il tentativo di trovare i contenuti chiave in un grande insieme di informazioni, la *sentiment analysis*, già spiegate o per determinare cosa ha guidato un determinato commento di una persona e quindi per un fine esplicativo;
* *Web Analytics*: applicazioni analitiche utilizzate per capire e migliorare l’esperienza online del consumatore, l’acquisizione di utenti e l’ottimizzazione del digital Marketing e delle campagne pubblicitarie. Questi offrono reporting, segmentazione, gestione delle campagne e integrazione con altre fonti dati e processi.

## ***1.2 DATAWAREHOUSE***

I Data Warehouse (DWH) sono il principale strumento a supporto della Business Intelligence. Essi permettono di collezionare dati integrati, consistenti e certificati, afferenti a tutti i processi di business dell’azienda e provenienti dalle fonti operazionali. Questi dati vengono in seguito opportunatamente trasformati attraverso procedure ETL e controllati attraverso il sistema di Data Quality.

La Qualità dei dati è un requisito fondamentale per l’intero sistema informativo, in quanto, se i dati risultano sporchi, possono oltre che causare un peggioramento delle performance aziendali, portare a prendere decisioni inopportune, comportando costi aggiuntivi e perdita di opportunità.

L’obiettivo di un DWH è pertanto quello di supportare il “*knowledge Worker*” (dirigente, amministratore, gestore, analista) per aiutarlo a condurre analisi finalizzate all’attuazione di processi decisionali e al miglioramento del patrimonio informativo, e fornire un unico punto di accesso per tutti i dati dell’azienda resi consistenti e affidabili attraverso i processi di ETL. Il datawarehouse garantisce inoltre una profondità storica completa dei dati, poiché in esso viene persistito anche lo stato passato delle informazioni permettendo così un’analisi temporale.

Dovranno quindi essere attentamente progettati per gestire in maniera efficiente ed efficace le caratteristiche dei Big Data.

I Datawarehouse sono realizzati come principale base per il Decision Support System (DSS), cioè un *sistema* di supporto alle decisioni è un sistema in grado di fornire chiare informazioni agli utenti, in modo che essi possano analizzare dettagliatamente una situazione e prendere le opportune decisioni sulle azioni da intraprendere in modo facile e veloce [12]. Il DSS si appoggia su dati di uno o più database, spesso organizzati in strutture diverse con dati non omogenei.

In altre parole, un sistema di questo tipo deve supportare le attività di analisi e controllo manageriale di routine, le attività di ricerca delle cause di un problema (*focused search*) e le attività di gestione manageriale complessa (*decision making*), permettendo inoltre un facile utilizzo ad un’utenza con un tempo disponibile ridotto e riluttante verso nuove tecnologie (soprattutto nei casi in cui non riesce a percepire in breve tempo i benefici).

Immagine che contiene screenshot

Descrizione generata con affidabilità molto elevata

Figura 4: Datawarehouse

Andiamo a descriverne in dettaglio le caratteristiche:

* *Orientato al soggetto*: nel data warehouse i dati sono organizzati per soggetti rilevanti come, per esempio, i prodotti, i clienti, i fornitori e il periodo di tempo, al fine di offrire tutte le informazioni inerenti una specifica area.
* *Integrato*: il data warehouse deve essere in grado di integrarsi perfettamente con la moltitudine di standard utilizzati nelle diverse applicazioni. I dati devono essere ricodificati, per risultare omogenei dal punto di vista semantico, e devono utilizzare le stesse unità di misura.
* *Variabile nel tempo*: a differenza dei dati operazionali, quelli di un data warehouse hanno un orizzonte temporale molto ampio (anche 5-10 anni), risultando riutilizzabili in diversi istanti temporali.
* *Non volatile*: i dati operazionali sono aggiornati in modo continuo; nel data warehouse i dati sono caricati inizialmente con processi integrali e successivamente aggiornati con caricamenti parziali; i dati, una volta caricati, non vengono modificati e mantengono la loro integrità nel tempo.

È possibile che un Datawarehouse sia suddiviso in diversi Data Mart, ognuno dei quali specifico per un solo processo di business fra quelli presenti all’interno dell’azienda (ordini, vendite, clienti, marketing, etc.). Nel capitolo 2 vedremo, per l’appunto, come una Data Mart relativa ad un fashion retailer.

### ***1.2.1 Architettura Di Un Data Warehouse***

In fase di progettazione risulta fondamentale stabilire quali tipologie di architettura adottare. Chiaramente, da quando sono stati idealizzati, i modelli (descritti successivamente) si sono evoluti e, di conseguenza, un DWH deve essere costruito secondo i principi moderni [10].

I pattern descritti in questo paragrafo rimangono comunque delle basi da cui partire.

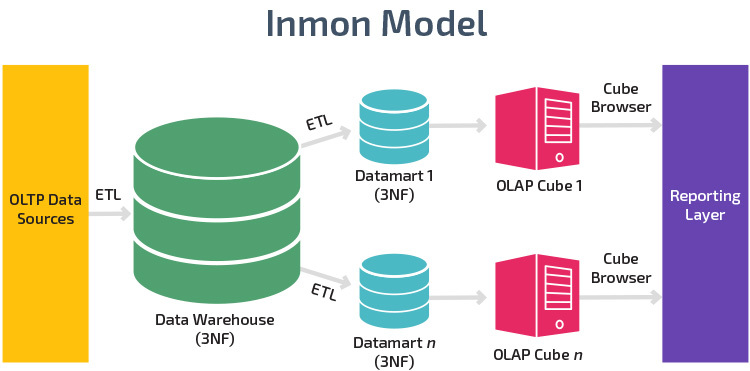
***Modello di Inmon - Corporate Information Factory****:* I Datawarehouse si costruiscono nella loro totalità fin dal principio come un unico blocco monolitico; non è possibile vederli come la composizione dei DM. Viene adottata una visione Top-Down.

Figura 5: Modello di Inmon

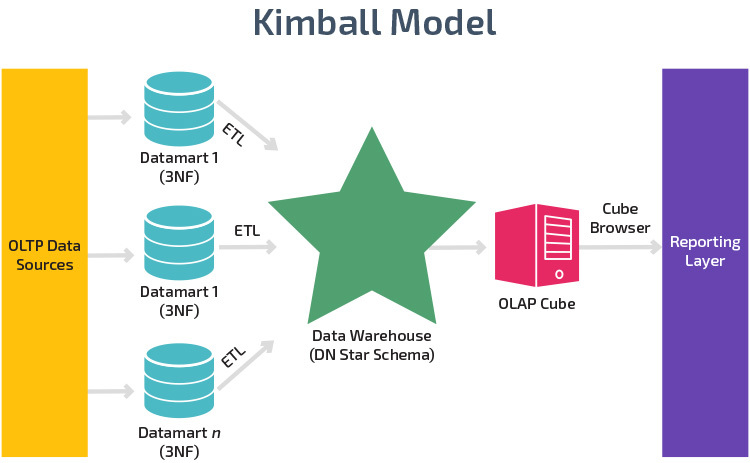
***Modello di Kimball - Dimensional Model****:* adotta un approccio Bottom-up in cui il Datawarehouse nasce dall’unione dei vari Data Mart che riferiscono ognuno ad una specifica area di business.

Figura 6: Modello di Kimball

È stato dimostrato che gli approcci di Inmon e Kimball funzionano per consegnare con successo i data warehouse. Esistono persino organizzazioni in cui è stata implementata una combinazione di entrambi, In un modello ibrido: il data warehouse viene creato utilizzando il modello Inmon e, oltre al data warehouse integrato, i Data Mart orientati ai processi aziendali vengono creati utilizzando lo schema a stella per la creazione di report. Non possiamo generalizzare e affermare che un approccio è migliore dell'altro; entrambi hanno i loro vantaggi e svantaggi, ed entrambi funzionano bene in diversi scenari. L'architetto deve selezionare un approccio per il data warehouse in base ai diversi fattori; Infine, affinché qualsiasi approccio abbia successo, deve essere attentamente studiato, discusso in dettaglio e progettato per soddisfare le esigenze di reporting della BI dell'organizzazione e dovrebbe anche integrarsi con la cultura dell'organizzazione.

### ***1.2.2 Extraction, Transformation And Loading (ETL)***

Il ruolo degli strumenti di ETL è quello di alimentare una sorgente dati singola, dettagliata, esauriente e di alta qualità che possa a sua volta alimentare il Data Warehouse. Le operazioni da essi svolte vengono spesso indicate con il termine *riconciliazione* che, durante il processo di alimentazione del Data Warehouse avviene in due occasioni: quando il DW viene popolato per la prima volta e periodicamente quando viene aggiornato. La riconciliazione consiste di quattro distinti processi detti rispettivamente:

* Extraction o Capture.
* Cleaning o Scrubbing.
* Trasformation.
* Loading.

In linea generale il confine tra pulitura e trasformazione è abbastanza nebuloso quindi per semplicità si assume che l’operazione di pulitura sia essenzialmente mirata alla correzione dei valori dei dati, mentre la trasformazione si occupa più propriamente del loro formato.

#### **1.2.2.1 Extraction**

La Data Integration è composta da due sottofasi chiamate estrazione e pulitura.

Durante la prima sottofase i dati rilevanti vengono estratti dalle sorgenti e questa operazione può essere di tipo:

* *Statico:* viene effettuata quando il DW deve essere popolato per la prima volta e consiste concettualmente in una fotografia dei dati operazionali.
* *Incrementale:* viene usata per l’aggiornamento periodico del DW, e cattura solamente i cambiamenti avvenuti nelle sorgenti dall’ultima estrazione. L’idea alla base è quella di utilizzare i cambiamenti registrati a livello dei dati per aggiornare il DWH. I benefici derivabili sono il volume molto piccolo dei dati coinvolti di volta in volta nell’operazione rispetto all’estrazione statica, e che la maggior parte dei dati nel Datawarehouse restano invariati e vengono analizzati solo i dati che hanno subito modifiche. Vengono usate tecniche CDC (Change Data Capture) che permettono di monitorare le sorgenti dati con l’obiettivo di individuare i cambiamenti avvenuti a livello dei dati. Queste tecniche sono particolarmente importanti per la Data Warehouse maintenance grazie alla propagazione dei cambiamenti rilevati a livello della sorgente.

La Pulitura, invece, è la sottofase che si occupa di migliorare la qualità dei dati andando ad eliminare dati “sporchi” dovuti a duplicazioni, inconsistenze, dati mancanti, valori errati etc.

Le principali funzionalità di pulitura dei dati riscontrabili negli strumenti ETL sono la correzione e l’omogeneizzazione, che utilizzano dizionari appositi per correggere errori di scrittura e riconoscere sinonimi, e la pulitura basata su regole, che applica regole proprie del dominio per stabilire le corrette corrispondenze tra i valori.

#### **1.2.2.2 Trasformation**

È la fase centrale del processo di riconciliazione e ha l’obiettivo di convertire i dati dal formato operazionale sorgente a quello del DW. Tra le funzionalità di questo livello per l’alimentazione del livello dei dati riconciliati si hanno:

* *Conversione e normalizzazione*: operano sia a livello di formato di memorizzazione sia a livello di unità di misura al fine di uniformare i dati.
* *Matching*: che stabilisce corrispondenze tra campi equivalenti in sorgenti diverse.
* *Selezione*: che riduce, se necessario, il numero di campi e record rispetto alle sorgenti.

Nella fase di alimentazione del DW si hanno invece due sostanziali differenze: La normalizzazione viene sostituita dalla denormalizzazione e si introduce l’aggregazione che realizza le opportune sintesi dei dati.

#### **1.2.2.3 Loading**

In questa fase avviene il caricamento dei dati sul Datawarehouse attraverso due modalità alternative:

* *refresh:* i dati vengono riscritti integralmente sostituendo completamente quelli precedenti. In generale questa tecnica viene utilizzata solo durante la fase iniziale di popolamento.
* *update:* vengono aggiunti al Datawarehouse solo i cambiamenti avvenuti sui dati senza sovrascrivere ad ogni iterazione tutti i dati. Questa tecnica viene utilizzata, in abbinamento all’estrazione incrementale per l’aggiornamento periodico.

Un modo per ridurre il tempo di caricamento è quello di parallelizzare il processo ETL. Questo può

si verificano in due modi: più passaggi eseguiti in parallelo e un singolo passaggio in esecuzione

in parallelo.

* *Passi di carico multipli*. Il flusso di lavoro ETL è diviso in più indipendenti lavori presentati insieme. È necessario riflettere attentamente su ciò che accade ogni lavoro; l'obiettivo principale è creare posti di lavoro indipendenti. Molto più sicuro per la gestione di eventuali errori.
* *Pipeline*. Il database stesso può anche identificare determinati compiti che può eseguire in parallelo. Ad esempio, la creazione di un indice può essere in genere parallela attraverso tutti i processori disponibili sulla macchina.

#### **1.2.2.4 Possibili Problemi Nell’ ETL E Come Risolverli**

Dopo che il sistema ETL è in produzione, i guasti possono verificarsi per innumerevoli motivi.

Le Cause comuni di guasti alla produzione di ETL includere:

* Errore di rete
* Errore del database
* Errore del disco
* Errore di memoria
* Errore nella qualità dei dati
* Aggiornamento di sistema senza preavviso

Per proteggersi da questi guasti, è necessario un solido sistema di backup e un sistema compagno di ripristino e riavvio. Devi pianificare per errori irreversibili durante il caricamento perché accadrà. Il sistema dovrebbe anticipare questo e fornire funzionalità di recupero, arresto e riavvio di arresto anomalo.

Ad esempio, Per un processo di caricamento dovrebbe impegnare serie relativamente piccole di record alla volta e tenere traccia di ciò che è stato commesso. La dimensione del set dovrebbe essere regolabile perché le dimensioni della transazione hanno implicazioni di prestazioni su diversi DBMS.

Il sistema di ripristino e riavvio viene utilizzato, ovviamente, per riprendere un lavoro che è entrato in errore e si è fermato o per far riportate l'intero lavoro indietro tramite backup e riavviarlo. Questo sistema è significativo dipende dalle capacità del sistema di backup. Quando si verifica un errore, la reazione iniziale istintiva è tentare di salvare qualsiasi cosa sia stata elaborata e riavviare il processo da quel punto. Ciò richiede uno strumento ETL solido e affidabile funzionalità di checkpoint, in modo che possa determinare perfettamente cosa ha elaborato e cosa non deve riavviare il lavoro esattamente nel punto giusto. In molti casi, potrebbe essere meglio uscire da tutte le righe che sono state caricate come parte del processo e riavviare dall'inizio.

Per questo motivo è consigliato di progettare tabelle dei fatti con un surrogato primario a singola colonna chiave. Questa chiave surrogata è un numero intero semplice che viene assegnato in sequenza come le righe vengono create per essere aggiunte alla tabella dei fatti. Con la chiave surrogata della tabella dei fatti, puoi facilmente riprendere un carico che viene fermato o estrarre tutte le righe nel carico limitando un intervallo di chiavi surrogate.

Quanto più un processo ETL è lungo, tanto più devi essere consapevole delle vulnerabilità a causa di un errore. La progettazione di un sistema ETL modulare composto da processi efficienti, resistenti agli arresti anomali e alle interruzioni impreviste, può ridurre il rischio di un guasto con conseguente notevole recupero. Un'attenta considerazione di quando mettere fisicamente i dati scrivendoli su disco, insieme a punti di recupero accuratamente predisposti e caricamento di data / ora o di tabelle sequenziali dei fatti consente di specificare la logica di riavvio appropriata.

## ***1.3 OLTP vs OLAP***

**On-Line Transaction Processing (OLTP)**

A livello di database, gli On-Line Transaction Processing si basano su query multi-access veloci ed efficaci. Le principali operazioni svolte sono INSERT, DELETE e UPDATE in quanto modificano direttamente i dati. Questi ultimi vengono costantemente aggiornati e, di conseguenza, richiedono un efficiente supporto alle operazioni di riscrittura. Una caratteristica fondamentale di questi sistemi è la normalizzazione, la quale fornisce un modo rapido ed efficace per effettuare scrittura nel database.

**On-Line Analytical Processing (OLAP)**

L’On-Line Analytical Processing è un insieme di tecniche software per l'analisi accelerata e interattiva di grandi moli di dati, con la possibilità di farlo da punti di vista differenti. Questi sistemi si riveleranno molto utili per l’ottenimento di informazioni di sintesi, che avranno lo scopo di supportare e migliorare i processi decisionali aziendale. Esempi di strumenti OLAP sono i Data Warehouse, i Cubi Multidimensionali.

Le maggiori differenze fra i due sistemi sono riportati in tabella [10]:

Tabella 1: OLTP VS OLAP

|  |  |  |
| --- | --- | --- |
|  | **OLTP** | **OLAP** |
| **Finalità** | Supporto all’operatività | Supporto al processo decisionale |
| **Modalità di utilizzo** | Guidata, per processi e stati successivi | Interrogazione ad hoc |
| **Quantità di dati per operazione elementare** | Bassa: centinaia di record per ogni transazione | Alta: milioni di record per ogni query |
| **Qualità** | In termini di integrità | In termini di consistenza |
| **Orientamento** | Per processo/applicazione | Per Soggetto |
| **Frequenza di aggiornamento** | Continua, tramite azioni | Sporadica, tramite funzioni esplicite |
| **Copertura temporale** | Dati correnti | Storica |
| **Ottimizzazione** | Per accessi in lettura e scrittura su una porzione di dati | Per accessi in sola lettura su tutta la base di dati |

In base alla memorizzazione dei dati, si avranno diverse architetture OLAP, ognuna delle quali con i propri pro e contro [10]:

* *Relational OLAP (ROLAP):*i dati vengono memorizzati in un database relazionale come supporto al motore OLAP. Le analisi multidimensionali vengono tradotte in query, restituendo risultati in forma multidimensionale.
* *Multidimensional OLAP (MOLAP):*si ha sia il database che il motore multidimensionale. Per le operazioni di Drill-Down non è il sistema ideale, in quanto, può generare errori.
* *Hybryd OLAP (HOLAP):*unisce i vantaggi dei due sistemi precedenti. In particolare, pre-aggrega i dati in sistemi multidimensionali per un’analisi efficiente e veloce, mentre vengono ricercate in un data base relazionale in caso di Drill-Down.
* *Desktop OLAP (DOLAP):*i dati vengono caricati in un sistema client e vengono calcolati dal motore in locale.

## ***1.4 COSA SI INTENDE CON IL TERMINE BIG DATA***

Il termine Big Data indica una raccolta di dati estesa in termini di *volume*, *velocità* e *varietà* che richiede tecnologie e metodi analitici specifici per l'estrazione di valore o conoscenza. Il termine è utilizzato in riferimento alla capacità di analizzare ovvero estrapolare e mettere in relazione un'enorme mole di dati eterogenei, strutturati e non strutturati, allo scopo di scoprire i legami tra fenomeni diversi e prevedere quelli futuri.

Le dimensioni variano nei diversi settori, da dozzine di terabyte a centinaia di petabyte (1000 terabyte), in base anche agli svariati strumenti software a disposizione. Esse aumenteranno sicuramente nel tempo grazie i continui avanzamenti tecnologici.

In questa definizione emergono le cosiddette 5V che caratterizzano i Big Data, ovvero il volume, velocità, varietà, veridicità e valore [2].

Il *volume* fa appunto riferimento all’enorme massa di dati generata attraverso numerosi canali.

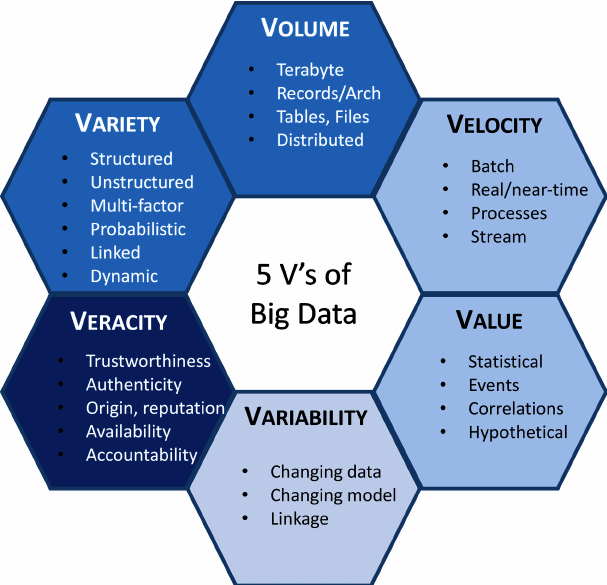
La *velocità* si riferisce alla rapidità con cui i dati vengono acquisiti e utilizzati grazie a transazioni sempre più frequenti e veloci: le aziende non solo raccolgono i dati più velocemente, ma cercano di sfruttarli il prima possibile, spesso in real-time.

Figura 7: Le 5V dei Big Data

La *varietà* è legata alle differenti tipologie di dati disponibili provenienti da un numero crescente di fonti di dati sia strutturati sia non strutturati; in particolare è possibile identificare cinque categorie di informazioni che costituiscono i Big Data:

* Dati generati da smartphone e altri dispositivi mobile relativi a persone, attività e localizzazione, tra cui dati RFID (radio-frequency identification), dispositivi che tracciano il prodotto, e dati da dispositivi di controllo come i contatori per il monitoraggio dell’acqua o del gas.
* Dati di vendita e pricing, dati generati dall’attività delle carte fedeltà e degli eventi promozionali.
* Computer log Data, come i click streams dai siti web.
* Informazioni dai social media come Twitter e Facebook.
* Social multimediali e altre informazioni da YouTube e siti simili.

La *veridicità* riguarda la questione relativa alla qualità dei dati e al loro livello di sicurezza, la cui garanzia rappresenta una sfida molto importante. Per poter sfruttare i Big Data è necessario saper agire per poter estrarne il *valore* e, quindi, incrementare la produttività e la competitività delle aziende e creando un surplus economico per i consumatori.

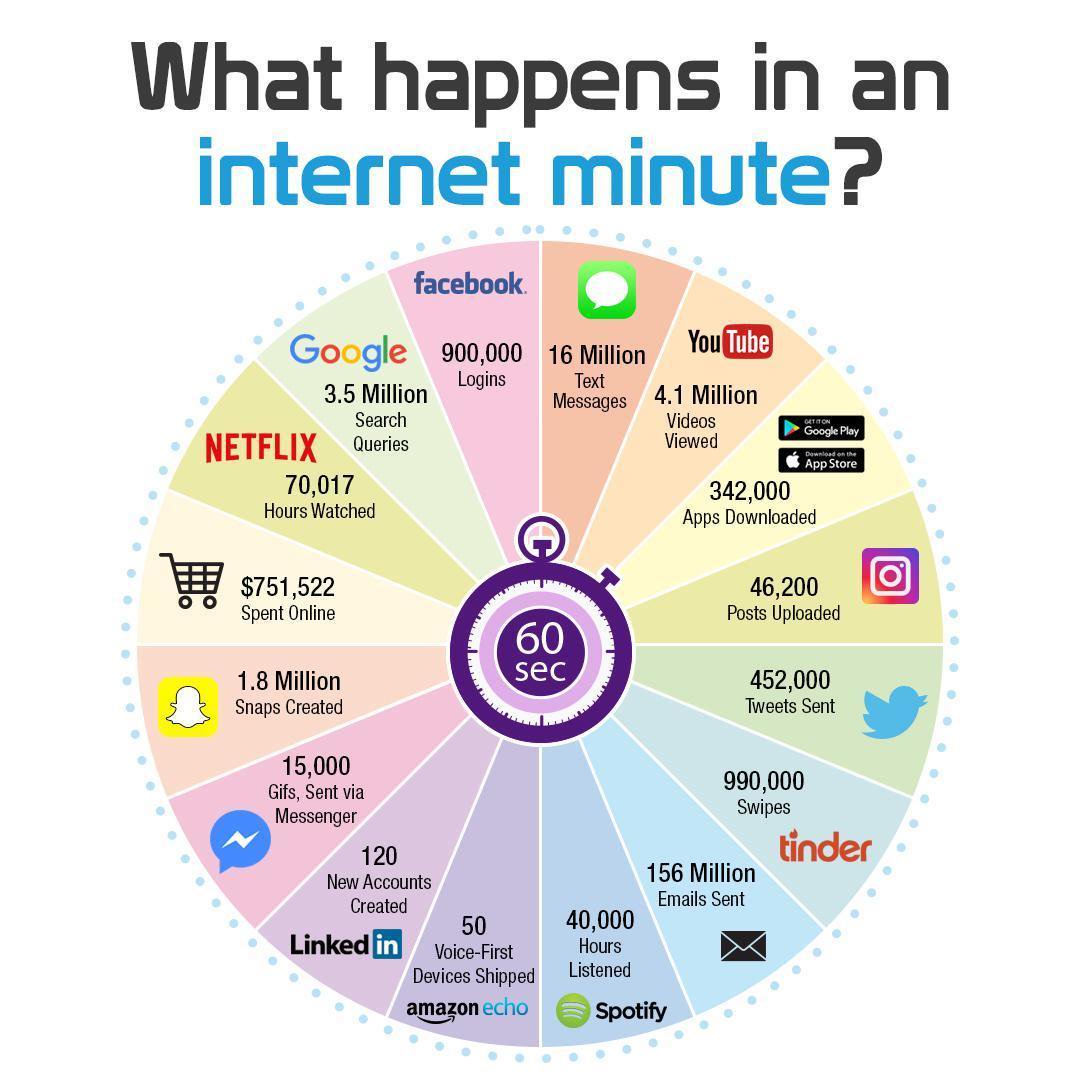
La capacità di memorizzare, aggregare i dati e di utilizzare i risultati per svolgere analisi di business profonde, è in continuo miglioramento grazie alla disponibilità di strumenti software e tecniche sempre più sofisticate combinate a una crescente potenza di calcolo. Stiamo assistendo ad un enorme cambiamento della capacità di generare, comunicare, condividere e accedere ai dati dovuto all’aumento del numero di persone, strumenti e sensori ora connessi da reti digitali. Per capire la grandezza del fenomeno, basta osservare la figura sottostante che mostra quanti dati vengono generati in un minuto.

Figura 8: Dati generati in un minuto nei social

### ***1.4.1 Benifici e Barriere Sull’utilizzo Dei Big Data***

I Big Data rappresentano una grande opportunità per le aziende e per le economie nazionali in quanto consentono di ottenere diversi benefici significativi:

* *Rivelare le variabilità delle performance e migliorare le prestazioni*: La creazione e la memorizzazione di dati transazionali in forma digitale consente alle aziende di avere dati più accurati e dettagliati su svariate performance, dallo stato dei magazzini ai giorni di malattia del personale, tutto in tempo reale o quasi. Inoltre, esse utilizzando i dati per analizzare la variabilità delle prestazioni e per capirne le cause più profonde.
* *Personalizzare le azioni:* I Big Data consentono di creare specifici segmenti di clienti chiamati cluster e di personalizzare prodotti e servizi sulla base delle loro esigenze per realizzare promozioni e pubblicità adatte ad esse.
* *Migliorare le previsioni e Supportare le persone nel processo di decision making:* Utilizzando Analytics sofisticati su interi Dataset è possibile automatizzare e migliorare i processi decisionali tramite le predizioni dei Key Performance Indicators (KPI), minimizzare i rischi e scoprire preziosi insight; Questi benefici naturalmente non possono essere perseguiti con l’analisi e la gestione di piccoli campioni di dati tramite i fogli di calcolo. I rivenditori per esempio possono utilizzare algoritmi che consentono la messa a punto automatica e l’ottimizzazione degli inventari e dei prezzi a partire dai dati in tempo reale relativi alle vendite nei negozi e a quelle online.
* *Creare trasparenza:* Un accesso facile e tempestivo ai Big Data rende disponibile una maggiore quantità di informazione e facilita la condivisione dei dati tra le diverse unità organizzative di un’impresa.
* *Profiling dei consumatori:* La disponibilità quasi in real-time di dati da smartphone fornisce caratteristiche dettagliate sui clienti e sul loro complesso processo decisionale quando fanno acquisti: i Big Data permettono infatti di identificare i modelli comportamentali dei consumatori e far luce sulle loro intenzioni.
* *Creare nuovi prodotti e servizi, nuove tipologie di aziende e innovativi modelli di business.* Le società possono sfruttare i Big Data per migliorare lo sviluppo di modelli futuri e per creare servizi post-vendita innovativi.
* *Incrementare la produttività e la profittabilità delle aziende.* Lo sfruttamento dei Big Data può portare ad un aumento dell’efficacia e dell’efficienza delle imprese, le quali potranno realizzare più output utilizzando meno input e migliorare il livello di qualità dell’output stesso.

Questo elenco di benefici mette in evidenza come l’investimento nei Big Data porti alla creazione di valore per le aziende e quindi all’ottenimento di vantaggio competitivo nel lungo termine. Risulta quindi fondamentale per loro sviluppare competenze in questo ambito.

Nonostante le opportunità offerte dai Big Data siano enormi, c’è ancora un certo scetticismo all’interno delle aziende sui reali benefici apportati a causa degli scarsi risultati ottenuti in pratica [3].

Esistono quindi una serie di barriere da considerare, che possono essere classificate in 6 categorie:

* *Barriere tecniche*: Difficoltà di integrazione dei dati, Basso grado di influenza del business, Scarsa qualità dei dati.
* *Barriere legate alle competenze*: Difficoltà di comprensione degli strumenti analitici e di quantificazione dei benefici, Carenza di talenti, Difficoltà nella scelta del tool adatto.
* *Barriere organizzative/gestionali*: Mancanza di commitment dei top manager che non sono coinvolti nelle iniziative di Big Data, verso le quali mostrano poco interesse, risultando inefficaci.
* *Barriere culturali*: La maggior parte delle aziende non è ancora pronta e del tutto aperta alle innovazioni che i Big Data potrebbero portare, in quanto il loro sfruttamento richiederebbe significativi cambiamenti culturali e organizzativi: Inerzia.
* *Barriere economiche*: Le iniziative Big Data richiedono ingenti spese in termini di tecnologie implementate e di nuove figure professionali da assumere.
* *Barriere legate alle privacy:* I consumatori non vogliono che le loro informazioni personali, come i personal location Data e i dati elettronici generati dal loro uso di Internet, vengano utilizzate dalle aziende, soprattutto perchè non sanno dove e come queste verranno sfruttate dalle organizzazioni, le quali devono considerare anche le leggi relative alle privacy dei diversi Paesi. Tools che consentono di tracciare ogni movimento dei dipendenti e di misurare continuamente le loro performance fanno gli interessi delle organizzazioni e non dei singoli individui, che vedono minacciata la loro privacy.

### ***1.4.2 Tecniche Per L’analisi Dei Big Data***

Fino ad ora abbiamo parlato della ideologia e del valore aggiunto ch possono portare i Big Data ad una azienda. Di seguito, invece, saranno elencate le principali tecniche e le tecnologie utilizzate per aggregare, manipolare, gestire e analizzare i dati.

* *A/B testing*: tecnica in cui un gruppo di controllo viene confrontato con gruppi di test al fine di determinare quali modifiche e azioni miglioreranno una data variabile obiettivo, come il tasso di risposta a una campagna di Marketing.
* *Crowdsourcing:* tecnica utilizzata per raccogliere dati, sottoposta a un grande gruppo di persone o a una comunità, attraverso, per esempio, il Web.
* *Data integration*: insieme di tecniche che integrano e analizzano dati provenienti da diverse fonti al fine di sviluppare insight più efficienti e accurati rispetto a quelli ottenuti esaminando una singola fonte.
* *Modelli predittivi*: tecniche in cui viene creato o scelto un modello matematico per prevedere la probabilità di un risultato.
* *Data mining*: insieme di tecniche di classificazione, cluster analysis, regole associative e regressione, che permette di estrarre modelli da grandi dataset combinando metodi statistici e di machine learning con la gestione dei database.
* *Machine Learning*: parte della computer science riguardante la progettazione e lo sviluppo di algoritmi che consentono ai computer di identificare i comportamenti basandosi su dati empirici e, in particolare, di riconoscere schemi complessi e prevedere decisioni per mezzo della intelligenza artificiale.
* Natural language processing (NLP): insieme di tecniche di computer science e linguistica che si ricorrono ai computer per analizzare il linguaggio umano.
* *Regressione*: set di tecniche che permettono di determinare come il valore di una variabile dipendente cambia quando una o più variabili indipendenti vengono modificate.
* *Ottimizzazione:* insieme di tecniche numeriche utilizzate per riprogettare sistemi e processi complessi al fine di migliorare le performance relativamente a uno o più aspetti, tra cui costi, velocità e affidabilità.
* *Sentiment Analysis*: applicazione del processing natural language e di altre tecniche analitiche per identificare ed estrarre informazioni soggettive dai testi, per esempio la “polarità” (positiva, negativa o neutra) delle caratteristiche o dei prodotti su cui le persone hanno espresso un’opinione e il grado e la forza dell’opinione stessa.
* *Statistica*: scienza della raccolta, organizzazione e interpretazione dei dati, utilizzata per esprimere giudizi sulle relazioni tra variabili che potrebbero essersi verificate per caso (ipotesi nulla) e su quelle causali (statisticamente significative).
* *Data Visualization*: tecniche di creazione di immagini, diagrammi o animazioni che consentono di comunicare, capire e migliorare i risultati dell’analisi dei Big Data.

## ***1.5 PROGETTI DI BIG DATA NEL MARKETING***

Lo sfruttamento dei Big Data in ambito Marketing rappresenta un enorme potenziale, tanto che le aziende si stanno dedicando e hanno un grande interesse verso progetti che prevedono il loro utilizzo in quest’area. Oltre ai Social Analytics che affronteremo in seguito, possiamo classificarne altri 6: il Direct e il Digital Marketing, la Customer Micro-Segmentation, il Location-based Marketing, Price Optimization, l’In-store Analysis e il Cross-Selling/Up-Selling.

### ***1.5.1 Direct E Digital Marketing***

Il Direct Marketing comprende tutte le tecniche di Marketing che consentono alle aziende di comunicare in modo mirato e personalizzato direttamente con il cliente o l’utente finale. La continua e significativa crescita di internet e della sua importanza ha comportato il rapido sviluppo del Digital Marketing, che assume la forma di display advertising, contenuti su Facebook, video clip su Youtube, e-mail personalizzate e molto altro. Le aziende per fare Digital Marketing oggi possono contare sull’enorme ammontare di informazioni degli utenti, che trascorrono ore e ore al giorno su Internet, relative ai loro interessi, ai contenuti delle loro comunicazioni, agli acquisti che fanno e molto altro [4].

Il Direct Marketing si serve di molte tecniche di Big Data, oltre che per identificare i clienti più profittevoli e quelli che risponderanno con maggiore probabilità, soprattutto per profilare i clienti, in modo da prevedere anche il comportamento di quelli sconosciuti. Vengono utilizzate sia tecniche di apprendimento supervisionato, come i modelli di ottimizzazione, le reti neurali bayesiane e gli alberi decisionali sia quelle non supervisionate, tra cui il clustering. Per ottenere risultati migliori l’ideale è combinare diverse tecniche [5].

I vantaggi apportati dai Big Data al Direct Marketing sono, oltre alla personalizzazione del messaggio, la visione a 360° del cliente, l’identificazione dei contenuti, del timing e del canale più appropriato per inviare il messaggio e la possibilità di fare questo in real time.

Da ciò deriva un incremento del tasso di conversione, ovvero del numero di visitatori che decidono di cliccare su un certo contenuto casuale o di visitare un sito web come risultato di un’azione guidata, e quindi la massimizzazione del Digital ROI, l’acquisizione di nuovi clienti e la fidelizzazione di quelli che già si rivolgono all’azienda.

### ***1.5.2 Customer Micro-Segmentation***

La molteplicità di nuove tipologie di dati e lo sviluppo di Analytics avanzati permette di avere dettagli granulari e un numero maggiore di informazioni sui consumatori e, quindi, di generare micro-segmenti molto precisi, costituti da un piccolo numero di persone. Molti Retailer affermano addirittura di essere impegnati nella personalizzazione e non più nella semplice segmentazione [1]. I tradizionali segmenti B2C (Business to Customers) e B2B (Business to Business) basati rispettivamente su dati demografici, psicografici e comportamentali e sulle dimensioni delle aziende o sui criteri di acquisto adottati sono ormai superati.

Sfruttando quindi:

• *Activity-Based Data*: click-stream data dal web, le storie degli acquisti, i dati dei call center, i dati mobile.

• *Profili dei social network*: la storia lavorativa e l’appartenenza a gruppi.

• *Sentiment Data*: associazioni a prodotti e aziende (like o follows) e commenti online.

• *Dati tradizionali*: dati delle ricerche di mercato e quelli transazionali.

È possibile costruire segmenti molto più stretti. Gli uomini di Marketing possono quindi creare offerte, prodotti e servizi personalizzati e su misura per ciascun cluster, con ovvi benefici sui ritorni. Questi dati possono inoltre essere aggiornati in real-time, riuscendo quindi ad identificare i cambiamenti dei clienti e delle loro preferenze.

### ***1.5.3 Price Optimization***

Le aziende possono sfruttare la crescente granularità dei dati sulle vendite e i potenti Analytics per ottimizzare i prezzi. L’ammontare di informazioni a loro disposizione è enorme, dalle serie storiche della domanda, ai dati relativi alle scorte, a quelli riguardanti i competitor, fino al livello delle vendite attuali. Questa base di dati è in continuo aumento considerata l’esplosione di nuovi canali di vendita online dove i consumatori possono confrontare i prezzi, incrementando la competizione tra le varie firme presenti nel mercato, andando incontro alle esigenze del cliente [6].

Da queste ingenti quantità di dati, attraverso opportuni tools, i pricing manager sono in grado di estrarre insight per definire quasi in real-time il prezzo ottimale che un consumatore è disposto a pagare per ciascun prodotto, basandosi sulle sue caratteristiche.

La price optimization può considerare, per esempio, l’elasticità della domanda al prezzo, con specifici modelli che analizzano i dati delle vendite storiche per ricavare insight sul pricing di ciascuna unità, che possono poi essere utilizzati per fare promozioni o per ridurre i prezzi, valutando i costi conseguenti. I benefici che le aziende riescono a conseguire in questo modo sono un aumento dei ricavi, dei margini e della quota di mercato.

Tuttavia, per riuscire a sfruttare i Big Data in quest’area è necessario costruire una fiducia verso il cliente, Identificare le opportunità più promettenti, che comprendono determinare quanto il consumatore vuole pagare esattamente per un dato prodotto attraverso la customer segmentation e le promozioni personalizzate, e Ascoltare i dati su cui le organizzazioni devono saper far leva. Particolare attenzione è focalizzata nel corretto utilizzo di adeguati Analytics per identificare elementi che spesso vengono trascurati e per determinare i fattori guida per ciascun cliente e prodotto che porterà alla scelta finale di prezzo [7].

### ***1.5.4 Location-Based Marketing***

Il Location-based Marketing si basa sull’adozione crescente di smartphone e di altri mobile device che generano i personal location Data, i quali permettono di conoscere posizione e comportamento delle persone in real-time sfruttando il GPS o il WI-Fi, favorendo lo sviluppo di una strategia di Marketing che considera le abitudini lavorative e di divertimento e non solo le preferenze dei consumatori. Altre fonti utilizzate sono i segnali delle torri di triangolazione cellulare e i pagamenti tramite carte di credito e di debito, le quali, attraverso il terminale del punto di vendita, rendono disponibili i dati di identificazione personale.

Quello che le aziende fanno solitamente prende il nome di il *Geo-Targeted Advertising*, ovvero effettuare azioni di advertising in tempo reale in base alla localizzazione dei propri clienti. Infatti, per ottenere enormi vantaggi, le aziende ricorrono alle push notifications, cioè delle offerte customizzate e aggiornate per un determinato cliente mentre, per esempio, cammina con in mano lo smartphone all’interno del negozio. Pertanto, lo sfruttamento dei dati di geolocalizzazione può portare ad un aumento delle vendite, ad un incremento dei profitti e ad un miglioramento della customer experience e perciò alla fidelizzazione della clientela.

Tuttavia, relativamente a questo progetto le aziende si trovano a dover affrontare due sfide: La privacy ed un trade-off, cioè se gli utenti desiderano ricevere offerte mobile quando si trovano in prossimità dello store stesso.

### ***1.5.5 In-Store Analysis***

L’in-store analysis prevede l’analisi dei dati real-time relativi al comportamento dei consumatori tramite la posizione e il percorso dei clienti all’interno dello store vengono tracciati attraverso svariate tecnologie: video camere, Wi-Fi, strumenti Bluetooth, sistemi dei punti di vendita, carte di pagamento, trasponder dei carrelli, applicazioni degli smartphone, Path Intelligence e tag RFID sulle carte d’acquisto.

Tools e Analytics, quali web dashboard, app mobile, real-time alert e strumenti di Data mining, vengono utilizzati per organizzare, analizzare e visualizzare questo grande ammontare di dati, identificare trend e confrontare le prestazioni dei diversi periodi. Così facendo vengono estratti insight relativi ai comportamenti dei consumatori all’interno dello store, con l’obiettivo ultimo di migliorare la customer experience.

In particolare, gli insight ottenuti sono relativi a quanti clienti entrano nel negozio, a come si comportano gli shopper all’interno dello store e per conoscere il consumatore tramite gli attributi sesso, età, se è la prima volta che entra nel negozio, se ritorna spesso, da dove viene e quali sono i suoi interessi

Le aziende si servono di questi insight per migliorare efficacemente l’organizzazione, ovvero per ottimizzare il layout dello store, le sue caratteristiche, il posizionamento sugli scaffali e il mix di prodotti offerti per trasformare i clienti una tantum in clienti abituali, per incrementare la frequenza delle loro visite e delle loro spese migliorando la store experience, per aumentare la dimensione media della transazione e attirare un numero sempre maggiore di consumatori. Le organizzazioni mettono in atto aggiustamenti in tempo reale per ottimizzare l’intero processo d’acquisto.

### ***1.5.6 Cross-Selling / Up-Selling***

I Big Data offrono grandi opportunità per aumentare la dimensione media dell’acquisto di un consumatore sia mettendo a disposizione prodotti o servizi collegati con la scelta d’acquisto iniziale sia offrendo qualcosa di maggior valore rispetto a questa, ovvero per migliorare le azioni di Cross-Selling e di Up-Selling. Dati quali le caratteristiche demografiche dei clienti, la posizione real-time, le preferenze, la storia degli acquisti passati vengono utilizzati a tal fine. Gli algoritmi come le regole di associazione, si basano su questi dati per prevedere il comportamento dei consumatori in vari scenari di vendita ed estrarre insight per capire molto prima cosa vogliono e determinare quindi il miglior approccio [9].

I benefici che le aziende traggono sono un aumento delle vendite e dei profitti e la fidelizzazione dei clienti.

Caso esemplare è quello di Amazon che raccoglie i dati da tutti gli utenti, riconoscendo i trend nelle persone che fanno acquisti simili attraverso tools di Analytics, in modo da cogliere potenziali opportunità e in base a ciascun prodotto o servizio visitato dall’utente sul sito suggerisce “potresti anche volere” ed è proprio in questo modo che riesce ad incrementare significativamente le vendite.

## ***1.6 KNOWLEDGE DISCOVERY IN DATABASE (KDD)***

Immagine che contiene interni

Descrizione generata con affidabilità molto elevata

Figura 9: KDD Process

Il KDD è una procedura interattiva e iterativa, che cerca di estrarre dai dati informazioni implicite, sconosciute a priori e potenzialmente utili.

Andiamo ad analizzare ora le singole fasi:

* *Identificazione degli obiettivi*: l'oggetto di questa fase è l'individuazione dell'ambito di applicazione in cui deve essere considerato il KDD, individuando gli obiettivi da perseguire. Si tratta forse della fase più difficile sia in termini di allocazione risorse sia perché devono essere determinate, in modo preciso, le misure del successo e i criteri per misurare successi e fallimenti. Si può fare una lista solo parziale dei molteplici aspetti che vanno presi in considerazione, alcuni sono il costo stimato del progetto e la scelta degli strumenti di data mining da utilizzare.
* *Selezione*: In questa fase deve essere selezionato l'insieme iniziale dei dati, da sottoporre all’ analisi. I dati grezzi vengono segmentati e selezionati secondo alcuni criteri al fine di pervenire ad un sottoinsieme di dati, che rappresentano il nostro target. Se i dati originali sono collocati in un flat file, la creazione del target risulta molto semplice. I sistemi di gestione dei database immagazzinano e manipolano dati transazionali, ciò consente ai sistemi informatici, relativi a tali sistemi, di fare aggiornamenti e di estrarre informazioni in modo rapido. Ciò è dovuto alla strutturazione dei dati tramite modelli relazionali, il cui scopo è ridurre la ridondanza dei dati, tramite la decomposizione di singole tabelle in più strutture relazionali, ed accelerare l'accesso alle informazioni. Del resto lo scopo del DM è proprio utilizzare la ridondanza dei dati per reperire “conoscenza", ecco perché è necessario ricomporre le strutture relazionali. Si intuisce quindi che è stretto il legame tra DM e Data Warehouse, il cui scopo è proprio quello di mettere insieme i dati, e non scomporli, al fine di sfruttarne la ridondanza. Spesso è anche necessario mettere insieme informazioni estratte da più fonti, cosa che può rendere complessa la fase di selezione in quanto bisogna trasformare i dati in modo da assicurare l'omogeneità in quanto, ad esempio, la codifica dei dati deve essere uguale per tutti i record dei dati target, altrimenti l'analisi risulta di scarsa utilità.
* *Preelaborazione*: Generalmente il target data disponibile non deve essere analizzato interamente ma basta estrarne un campione opportuno, eseguendo poi un'analisi su base campionaria. Inoltre i dati devono essere preprocessati, cioè “puliti", trattando in maniera opportuna i dati anomali e mancanti. Vanno individuati i valori errati delle variabili; trovare gli errori nei dati categorici diventa un problema quando si analizzano dataset molto grandi. I dati vanno anche semplificati; queste tecniche di data smoothing sono mirate alla riduzione del numero di valori per una variabile numerica. Alcuni classificatori, come le reti neurali, utilizzano funzioni che effettuano la semplificazione durante il processo di classificazione, eseguendo così un data smoothing interno. Due semplici tecniche di semplificazione sono il calcolo e l'arrotondamento dei valori medi.
* *Trasformazione*: I dati, per essere utilizzati, spesso devono essere trasformati; questa fase può assumere varie forme e può essere necessaria per varie ragioni. Si possono convertire tipi di dati in altri o definirne di nuovi, ottenuti attraverso l'uso di operazioni matematiche e logiche sulle variabili, eseguire delle normalizzazioni (scalamento decimale, normalizzazione min-max o con lo z-score) o addirittura eliminare delle variabili. In genere infatti gli algoritmi di DM non lavorano in modo efficiente se i dati contengono una grande quantità di variabili che non sono in grado di prevedere la classe di appartenenza. Si rende quindi utile una ricerca ed una successiva eliminazione delle variabili ridondanti e “inutili" per il problema in questione. A volte le variabili con poco potere previsivo possono essere combinate con altre per formare nuove variabili con un alto grado di capacità previsiva.
* *Data mining*: Ai dati trasformati vengono applicate una serie di tecniche in modo da poterne ricavare dell'informazione non banale o scontata. Sono gli obiettivi che si vogliono raggiungere a dare un'indicazione sul tipo di tecnica/algoritmo che deve essere applicata.
* *Interpretazione e valutazione*: Scopo di questa fase è determinare la validità del modello ottenuto con il DM; in sintesi non basta interpretare i risultati ma bisogna capire in che misura questo modello o risultato possa essere utile. Questo può essere fatto in vari modi sia attraverso un'analisi statistica che euristica o sperimentale.
* *Data Visualization*: L'ultimo obiettivo consiste nell'utilizzare ciò che è stato appreso, creando un report o un rapporto tecnico su ciò che è stato scoperto, cercando di capire in che modo sfruttare ciò che è stato scoperto.

Si capisce bene quindi che il processo di estrazione della conoscenza è lungo e piuttosto articolato, perciò, sono fondamentali le scelte che si fanno per il trattamento di anomalie o errori nei dati e l'identificazione chiara degli obiettivi che si vogliono perseguire.

### ***1.6.1 Data Mining vs Machine Learning***

Il data mining si riferisce all'estrazione di conoscenza da una grande quantità di dati ed è il processo per scoprire vari tipi di pattern che sono ereditati nei dati e che sono accurati, nuovi e utili. È un processo iterativo di creazione di un modello predittivo e descrittivo, attraverso la scoperta di tendenze e pattern precedentemente sconosciuti con grandi quantità di dati per supportare il processo decisionale. Può essere definito anche come il sottoinsieme dell'analisi aziendale, simile alla ricerca sperimentale. Le fonti del data mining sono i database e i metodi statistici.

Il Machine Learning indica un ambito di ricerca all’interno dell’Intelligenza Artificiale e, grazie all'esperienza basata sui dati, implica lo studio di algoritmi che sono in grado di estrarre informazioni automaticamente. Sono necessarie due fonti di dati: dati di addestramento e dati di test. Di solito, il Machine Learning utilizza tecniche di data mining e un altro algoritmo di apprendimento per costruire modelli di ciò che sta accadendo dietro alcuni dati in modo che possa prevedere i risultati futuri.

Ma vediamo in tabella le varie differenze:

Tabella 2: Data Mining vs Machine Learning [13]

|  |  |  |
| --- | --- | --- |
|  | **Data mining** | **Machine learning** |
| Definizione | Estrarre Knowledge da una grande quantità di dati | Introdurre un nuovo algoritmo da dati e esperienza passata |
| Storia | Introdotto nel 1930 | Introdotto nel 1950 |
| Responsabilità | Il data mining viene utilizzato per ottenere le regole dai dati esistenti. | L'apprendimento automatico insegna al computer a imparare e comprendere le regole date. |
| Origine | Database tradizionali con dati non strutturati | Dati esistenti e algoritmi. |
| Implementazione | We can develop our own models where we can use data mining techniques for | Possiamo usare l'algoritmo di machine learning nell'albero decisionale, nelle reti neurali e in qualche altra area di intelligenza artificiale. |
| Natura | Manuale | Automatico |
| Applicationi | Usato nella cluster analysis | Usato in web search, spam filter, credit scoring, fraud detection, computer design |
| Tecniche | Il data mining è più di una ricerca che utilizza metodi come l'apprendimento automatico | Auto apprendimento e insegnamento fatto da task intelligenti. |
| Scopo | Area limitata | Vasta area. |

## ***1.7 ALGORITMI DI DATA MINING***

L'obiettivo del data mining consiste nell'estrarre nuove informazioni dai dati esistenti. Come vedremo, esistono due approcci per raggiungerlo: l’apprendimento supervisionato e l’apprendimento non supervisionato [14].

* ***Supervised learning***: metodologia di apprendimento automatico in cui vengono passati alla macchina degli esempi composti da una coppia di dati contenenti il dato originale e il risultato atteso. Compito della macchina è quello di trovare la regola (funzione o modello) con cui creare una relazione tra i due in modo tale che, al presentarsi di un esempio sconosciuto in precedenza, possa ottenere il risultato corretto. I dati sono precedentemente etichettati, ovvero assegnati ad una certa categoria. L’apprendimento supervisionato è utilizzato principalmente per i problemi di *classificazione*, come, ad esempio, si usa nel marketing per classificare i clienti potenziali e proporre i prodotti a cui potrebbero essere interessati sulla base del profilo e della storia degli acquisti. Un altro esempio sono i sistemi anti-spam delle email che, all’arrivo di un messaggio, riescono a decidere se una determinata email debba essere etichettata come spam o meno.
* *Unsupervised learning:* a differenza del precedente, non utilizza dati classificati e etichettati in precedenza; non sappiamo, quindi, a quale categoria essi appartengano. Alla macchina viene chiesto, quindi, di estrarre una regola che raggruppi i casi presentati secondo caratteristiche che ricava dai dati stessi. Per questo è anche definito apprendimento di caratteristiche (feature learning). Gli algoritmi in questo caso cercano una relazione tra i dati per capire se e come essi siano collegati tra di loro. Non contenendo alcuna informazione preimpostata, l’algoritmo è chiamato a creare “nuova conoscenza” (knowledge discovery). Una delle applicazioni principali è il *clustering*, ovvero il raggruppamento dei dati in gruppi omogenei definiti *cluster*. L’apprendimento non supervisionato, quindi, serve generalmente ad estrarre informazioni non ancora note.

Immagine che contiene screenshot

Descrizione generata con affidabilità molto elevata

Figura 10: Data Mining Algoritms

Utilizzando alcune delle tecniche sopra citate possiamo creare modelli predittivi. Qualunque sia la loro applicazione, i modelli predittivi usano l'esperienza per assegnare punteggi e livelli di confidenza, ad alcuni risultati rilevanti in futuro. Per far ciò, bisogna dividere il processo in due fasi:

La prima fase è la formazione, in cui il modello viene creato utilizzando i dati del passato, mentre la seconda è il punteggio, in cui il modello creato viene testato con dati non visibili per vedere come ha segnato.

Non bisogna mai dimenticare che il più importante è quello di ottenere buoni risultati nei dati invisibili e non nei dati di allenamento. L'*overfitting* è la situazione che si verifica quando il modello spiega i dati dell'allenamento ma non può generalizzare per testare i dati.

Le innovazioni che utilizzano l’intelligenza artificiale e il Machine Learning sono tra le principali tendenze tecnologiche nel mondo del retail. Stanno avendo un grande impatto sul settore, in particolare nelle aziende di e-commerce che si affidano alle vendite online, dove l’uso di una qualche forma di Machine Learning è oggi molto comune, sopratutto nei retail.

Grandi retailer online come eBay, Amazon o Alibaba hanno integrato con successo le tecnologie AI nell’intero ciclo di vendita, dalla logistica di stoccaggio al servizio clienti post- vendita.

Le aziende che utilizzano i sistemi di raccomandazione ottengono aumenti delle vendite a seguito di offerte personalizzate e di una migliore esperienza del cliente. Le raccomandazioni, in genere, accelerano le ricerche e rendono più facile acquisire e fidelizzare i clienti inviando e-mail con collegamenti a nuove offerte che soddisfano gli interessi dei destinatari e si adattano ai loro profili.

Quando l’utente inizia a sentirsi compreso, è più propenso ad acquistare prodotti aggiuntivi. Conoscendo ciò che un cliente vuole e mostrandoglielo subito, è meno probabile che egli lasci la piattaforma. Ciò si traduce in una maggiore possibilità di acquisto e in una diminuzione della minaccia di perdere un cliente a favore di un concorrente.

Includendo l’offerta, la stagionalità, gli eventi esterni relativi alla tua attività (ad esempio un concerto, una partita, un festival), la domanda e l’offerta del mercato, un sistema automatico di prezzi può adeguare in modo efficiente i prezzi.

Vediamo nel dettaglio i più comuni algoritmi usati dal Machine Learning per andare in contro al cliente.

### ***1.7.1 Clustering***

L’obiettivo della clusterizzazione è di organizzare gli oggetti esaminati in gruppi, che condividono proprietà simili. Il Clustering si può considerare uno dei più importanti metodi di apprendimento non supervisionato e, come ogni metodo appartenente a questa categoria, non fa uso di identificatori determinati a priori per intuire la possibile struttura dei dati.

Esistono varie forme di clustering [15]:

1. *Clustering Esclusivo*: Ogni elemento può appartenere solamente ad un cluster, ossia le intersezioni tra i clusters sono sempre insiemi vuoti; questa procedura prende anche il nome di *Hard Clustering*.
2. *Clustering Inclusivo*: Ogni elemento può appartenere a più cluster contemporaneamente, con un indice che decreta il grado di appartenenza ad ogni cluster, procedura che prende il nome di *Soft* o *Fuzzy Clustering*.
3. *Clustering Partizionale*: Si utilizza il concetto di distanza tra gli elementi, i quali appartengono ad un particolare gruppo in base alla loro relazione con un punto significativo del dataset.
4. *Clustering Gerarchico*: Si costruisce una gerarchia di partizioni, costruita sia per aggregazione che per divisione, mediante una rappresentazione ad albero che prende il nome di *Dendogramma*. Esistono altre suddivisioni per quanto riguarda il Clustering Partizionale, più dettagliate, le quali si differenziano per la valutazione della distanza tra gli elementi e la relativa creazione dei cluster[19]. Questa tecnica si suddivide in due approcci:

* *Agglomerativo*: Il processo inizia considerando ogni punto come un cluster, ad ogni step si unificano i punti secondo una particolare funzione di similitudine arbitraria, fino ad ottenere un cluster unico ed il relativo dendogramma. Questo approccio si basa sullo sviluppo di una *Matrice di Prossimità* tra i cluster e risulta di fondamentale importanza la funzione per il calcolo della similitudine tra due cluster
* *Divisivo:* Caso complementare in cui si parte da un unico cluster e si suddivide ad ogni iterazione, fino ad ottenere un numero di cluster pari al numero di punti che costituiscono la base dati.

La complessità è nell’ordine di *O*(*N*3), e come in K-Means, la presenza di outliers condiziona negativamente questo approccio.

D seguito, sono descritte le principali strategie di Clustering ed algoritmi utilizzati nell’ambito Fashion.

#### **1.7.1.1 Clustering Basato su Centroidi (K-Means)**

Il clustering basato su centroidi è di tipo partizionale e ogni cluster è rappresentato da un prototipo chiamato *centroide* che tipicamente è la media tra le distanze dei punti del cluster. Uno dei più famosi algoritmi di clustering appartenenti a questa categoria è il K-Means che richiede di specificare il numero *K*di cluster che si vogliono ottenere. L’algoritmo iterativamente elegge i *K* centroidi del cluster, ed ogni elemento viene associato al centroide più vicino. L’algoritmo è il seguente:

|  |
| --- |
| **K-MEANS ALGORITHM** |
| 1: **function** K-Means(*clusters K*) |
| 2: *Elezione K Centroidi* |
| 3: **repeat** |
| 4: *Assegnamento di ogni elemento al punto K piu*´ *vicino* |
| 5: *Ricalcolo dei K Centroidi* |
| 6: **until** *I Centroidi non variano* |

Inizialmente, i centroidi vengono scelti randomicamente mentre, nelle iterazioni successive dell’algoritmo, essi consistono tipicamente nella media tra le distanze dei punti del cluster. Esistono differenti metodologie per calcolare tale distanza: *Distanza Euclidea*, *Cosine Similarity*, *Correlazione*. L’algoritmo converge per le misure di similitudine elencate. Tale convergenza si manifesta principalmente nelle prime iterazioni, seguite da una fase di assestamento. In essa, infatti, spesso la condizione di stop viene rilassata, ammettendo una soglia minima di cambiamento tra i centroidi.

La scelta dei centroidi è una fase molto sensibile, infatti vengono applicate le seguenti tecniche per risolvere, anche se non completamente, il problema:

* Si eseguono molteplici esecuzioni, stimando i centroidi in modi differenti oppure semplicemente randomicamente, in seguito si valuta la qualità del risultato ottenuto, per mezzo degli strumenti di validazione che saranno descritti in seguito.
* Si utilizza la procedura di *Clustering Gerarchico per eseguire K* suddivisioni e si calcolano i centroidi dei cluster ottenuti, questi saranno i punti di partenza per l’algoritmo K- Means.
* Si stima un numero di centroidi *N > K*, e vengono considerati solamente i K migliori, tramite
* tecniche di postprocessing, come eliminazione di piccoli clusters, unione di cluster molto simili tra di loro e suddivisione di cluster troppo grandi.
* Immagine che contiene testo, mappa

  Descrizione generata con affidabilità molto elevataImmagine che contiene testo

  Descrizione generata con affidabilità molto elevataSi utilizza l’algoritmo *Bisecting K-Means*, esso consiste in un approccio gerarchico attraverso il quale partendo da un unico cluster, si suddivide tramite algoritmo 2-Meansun numero arbitrario di volte. Si prende l’iterazione che ha prodotto i migliori cluster e si applica ricorsivamente l’algoritmo fino a che non si ottengono i Kcluster desiderati.

Figura 11: K-Means Algorithm

La complessità dell’algoritmo è *O*(*n\* K \* I \* d*) dove *n* è il numero di punti, *K* il numero di cluster, *I* il numero di iterazioni e *d* il numero di attributi su cui si basa la funzione per il calcolo della distanza utilizzata.

In conclusione, l’algoritmo K-Means presenta difficoltà nella gestione di dati la cui presenza di outliers è troppo elevata, infatti sono spesso eseguite procedure di *Preprocessing* per attenuare la problematica. Inoltre, come detto in precedenza, la scelta dei centroidi è spesso difficoltosa, soprattutto quando si ha a che fare con dati ad elevata densità. Tuttavia, K-Means risulta uno degli algoritmi più utilizzati soprattutto per quanto riguarda il problema della Customer Segmentation.

#### **1.7.1.2 Density-Based Clustering**

Il density-based clustering si basa sul concetto di *Densità.* L’idea di base è trovare clusters definiti implicitamente da regioni ad alta densità separate da regioni a bassa densità. Uno degli algoritmi più famosi di questa categoria è il DBSCAN che usa due parametri per identificare aree dense: un raggio *ε*, che serve a identificare un’area attorno ad un determinato punto, e un numero minimo di punti *MinPts* che devono essere presenti all’interno del raggio *ε*.

Ogni punto viene etichettato secondo 3 differenti categorie:

* *Core Point*: tutti i punti che superano la soglia *MinPts* all’interno del raggio *ε*.
* *Border Point*: tutti i punti che non superano la soglia *MinPts* ma nel loro raggio *ε* hanno almeno un Core Point.
* *Noise Point*: tutti i punti che non sono Border o Core Point.

L’algoritmo parte da un punto casuale. Sono calcolati tutti i punti compresi nel raggio *ε* e se contiene un numero *MinPts* di punti, viene creato un nuovo cluster altrimenti viene etichettato come Noise-Point. Il punto potrebbe essere successivamente ritrovato in quanto incluso nel raggio *ε* di un vicino e di conseguenza essere inserito in un cluster.

Se un punto è associato ad un cluster, sono inseriti in esso anche i punti presenti all’interno del suo raggio *ε*, e di conseguenza anche i loro vicini all’interno sempre del raggio stabilito. Questo processo continua fino a quando non sono stati inseriti tutti i vicini. Ogni punto a cui è associato un cluster viene marcato come visitato e l’algoritmo prosegue eseguendo la stessa procedura per un punto successivo che non è ancora stato visitato.

L’algoritmo ha complessità *O*(*n*2) che tuttavia può essere ridotta a *O*(*n* log *n*) tramite utilizzo di strutture indicizzate per l’interrogazione del vicinato.

Il punto di forza di questo approccio è dato dalla buona gestione di outliers e dalla conseguente capacità di riuscire a gestire cluster di forme e dimensioni molto differenti. Tuttavia, risulta inefficiente quando si ha a che fare con dati che son caratterizzati da densità troppo variabili. È molto usato per clusterizzare tramite la Geo-localizzazione.

### ***1.7.2 Classification And Regression Trees (CART)***

CART è una procedura non parametrica dove non è necessario pre-testare la normalità o altre assunzioni che riguardano la distribuzione statistica dei dati. L’albero finale include solo le variabili indipendenti che risultano essere predittive della variabile dipendente; le altre variabili indipendenti non predittive non hanno effetto sul risultato finale; anche sotto questo aspetto CART si differenzia dalle altre procedure statistiche tradizionali. Con il termine classificazione si intende il processo che data una collezione di record, denominata *Training Set,* cerca di costruire un modello in grado di attribuire una caratteristica, denominata *attributo Classe,* basandosi sulla combinazione delle altre proprietà che caratterizzano il singolo individuo della popolazione. Una volta ottenuto il modello, esso può essere usato per predire la classe di nuove istanze di record per cui la classe è sconosciuta.

La struttura di un classification tree include i nodi non terminali (*parent nodes*), i quali hanno due discendenti diretti (*child nodes*), ed i nodi terminali che non subiscono ulteriori bipartizioni (*terminal nodes*). Il primo nodo radice (*root node*) contiene tutte le osservazioni. Dal nodo radice discendono due “nodi figli”. Ogni child node, che indichiamo con la lettera t contiene un sottocampione del campione originale, in cui i membri condividono le stesse caratteristiche, che influenzano la variabile dipendente di interesse. Ogni t, a sua volta, costituisce un potenziale parent node che può essere ancora suddiviso in due child node. Il processo continua fino a che l’albero non termina la sua crescita. I nodi terminali sono i nodi finali dell’albero decisionale e contengono insiemi di osservazioni che vanno a formare classi molto omogenee al loro interno e il più possibile eterogenee tra loro.

Vi sono alcuni step importanti da seguire quando si costruisce un albero decisionale con la procedura CART; gli step includono: adottare un criterio di bontà della tecnica con i cui i nodi vengono suddivisi da parent nodes a child nodes (split criterion); stabilire una regola di arresto di crescita dell’albero (stopping rule).

Per scegliere le split criterion si utilizza generalmente una tecnica di *Recursive Binary Splitting.*

Il metodo è binario e ricorsivo: binario, poiché ogni parent node si divide in due discendenti diretti, e ricorsivo, poiché i nodi (non terminali) nati dalla suddivisione del parent node in due discendenti diretti possono diventare, a loro volta, parent node e suddivisi in due nodi successivi.

Alberi decisionali con molti nodi e numero di divisioni possono portare a un sovradattamento dei dati (definito più propriamente dal termine *overfitting*). Ciò significa che il modello risulta di difficile interpretazione in quanto diventa inaccurato per previsioni successive ed ha bisogno delle stopping rule. I metodi per evitare questo problema sono impostare un numero minimo di dati di allenamento da utilizzare su ciascun nodo foglia o impostare la profondità massima del modello, che si riferisce alla lunghezza del percorso più lungo dal nodo radice al nodo foglia.

Immagine che contiene testo

Descrizione generata con affidabilità molto elevata

Figura 12: Decision Tree

I differenti algoritmi esistenti si differenziano in base alla strategia impiegata sui singoli nodi, per la valutazione dello Split. Esistono infatti differenti indici per la validazione di una classificazione:

* Immagine che contiene oggetto, orologio

  Descrizione generata con affidabilità molto elevata*GINI INDEX* identifica la qualità dello split. Considerando *p(i)* la frequenza relativa della classe *i* al nodo:
* Immagine che contiene oggetto, orologio

  Descrizione generata con affidabilità molto elevata*GAIN INDEX*: si basa sul concetto di *entropia*, indice relativo alla omogeneità del nodo, ottenuta eseguendo un particolare split sul nodo *pi*:

#### **1.7.2.2 Altri Tipi Di Classificatori**

* *Basata su Istanze*: Consiste in una famiglia di algoritmi i quali, anzi che eseguire generalizzazioni esplicite, confrontano nuove istanze direttamente con i record analizzati e opportunamente memorizzati dal training set. Degna di nota è la procedura Nearest-Neighbor che utilizza una particolare ed arbitraria metrica per il calcolo della distanza ed un parametro k rappresentante il numero minimo di vicini da estrarre [15]. Per ogni record che deve essere classificato, si calcola la distanza dal training set identificando i k record ritenuti più vicini e si usano i valori assunti dai loro attributi per classificare il record in esame.
* *Classificatore Byesiano:* Consiste in un framework probabilistico per risolvere il problema della classificazione. Si considerano gli attributi e la classe come variabili casuali Si basandoci fortemente sul concetto di Probabilità Condizionata. Dato un record con attributi (A1, A2, ..., An), l’obiettivo è quello di prevedere la classe C, ossia vogliamo trovare il valore di C che massimizza la probabilità P (C A1, A2, ..., An). Ne segue il Teorema di Bayes:

*P (C|A) = P(A|C)\* P(C)/ P(A)*

Grazie al teorema di Bayes, si ottiene un problema di ottimizzazione equivalente che consiste nel trovare C che massimizza: P (A|C) = P(A,C) . Esistono differenti modi per la stima di tale probabilità basandosi sui dati, come distribuzione normale, stima di densità, m-estimate, Laplace [15].

* *Support Vector Machine (SVM):* La classificazione viene eseguita trovando l’iperpiano che massimizza il margine tra due classi. I vettori (possibili attributi della classe) che definiscono l’iperpiano, sono definiti vettori di supporto. Il vantaggio di questo metodo consiste nel fatto che se i dati sono linearmente separabili, allora esiste un minimo globale unico. Una SVM ideale dovrebbe produrre un iperpiano che separa completamente i vettori di due classi non sovrapposte. In genere, la completa separazione non è sempre possibile, ma spesso si arriva ad ottenere un modello con troppi possibili casi che comporta una classificazione non corretta [18].

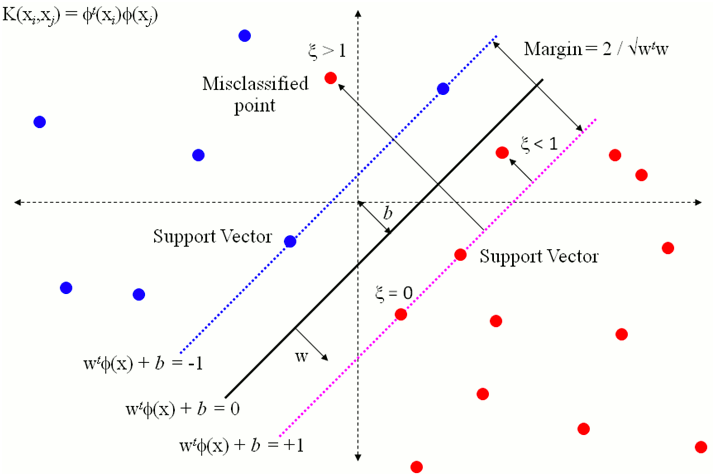


Figura 13. SVM

La *Validazione* di questi processo è di fondamentale importanza, in quanto permette di valutare le prestazioni del modello costruito e di poterlo confrontare con altre possibili modellazioni. Le misure di valutazione si basano sul Test-Set, partizione dei dati su cui applicare il modello predittivo.

L’applicazione del modello sul Test-Set produce la Matrice di Confusione, ossia una matrice indicante l’incidenza tra le classi predette e il loro valore reale dei record nel Test-Set. Si possono quindi determinare le seguenti tipologie di previsione:

* *True Positive*: Predizioni Positive Corrette
* *False Positive*: Predizioni Negative Corrette
* *True Negative*: Predizioni Positive Errate
* *False Negative*: Predizioni Negative Errate

Queste possono essere applicate a qualsiasi tipologia di attributo, non solamente alle classi binarie. Le metriche più utilizzate sono: Accuracy, Precision, Recall, F-Measure.

**Immagine che contiene screenshot

Descrizione generata con affidabilità molto elevata**

Figura 14: Confusion Matrix

### ***1.7.3 Predizione: Association Rules***

La base di partenza di un algoritmo per l’estrazione di regole associative è costituita da un insieme di Transazioni. Ogni transazione consiste in un insieme di *item*. Estrarre le *Regole di Associazione* consiste nel prevedere l’occorrenza di un item in base all’occorrenza di altri item compresi anch’essi nelle transazioni a disposizione.

Risulta importante definire alcuni concetti alla base di questa tecnica:

* *Itemset*: Collezione di uno o più elementi generalmente definiti per mezzo del parametro *k*, indicativo della sua dimensione nella forma *k*-*Itemset*.
* *Supporto Itemset*: Dato un itemset *I*, il supporto è la frazione delle transazioni che contengono *I* e si indica con *supp*(*I*).
* *Itemset Frequente*: Tutti gli itemset che superano un’arbitraria soglia minima di supporto.
* Una *Regola di Associazione* è un’implicazione espressa nella forma: *X* → *Y* con *X, Y itemset* dove *X* prende il nome di *Premessa* ed *Y Conseguenza* della regola.

Oltre al supporto, visto precedentemente, esiste un’altra forma di validazione della regole che tiene conto sia della premessa che della conseguenza: la *Confidenza*. Essa indica quanto spesso una particolare regola è verificata, consiste nella proporzione tra il numero delle transazioni che contengono l’intera regole e le transazioni che contengono la premessa:

*conf* (*X* → *Y*) = *supp*(*X* ∪ *Y* ) */ supp*(*X*)

Formalmente il supporto *supp*(*X* ∪*Y* ) può essere riscritto come la probabilità congiunta *P*(*EX ∩ EY* ), dove *EX* e *EY* sono tutte le transazioni che contengono *X* o *Y* rispettivamente. Quindi, possiamo esprimere la confidenza come la probabilità condizionata *P* (*EY* | *EX*).

Dato un set di transazioni, l’obiettivo consiste nell’estrazione di tutte le regole che rispettano le soglie arbitrarie di supporto e confidenza. La loro estrazione non può essere eseguita con un approccio Brute-Force, a causa dell’elevato numero di regole che possono essere generate. Per ridurre il numero di possibili regole, si sfrutta il Principio Apriori.

#### **1.7.3.1 Principio Apriori**

Questo principio si basa sulla proprietà *anti-monotona* del supporto, che ci permette di stabilire con certezza che se un itemset non risulta frequente, allora nemmeno tutti gli itemset che lo contengono risulteranno frequenti. Tale proprietà è cosi formalizzata, con *X* e *Y* itemset:

*∀X, Y : (X ⊆ Y ) ⇒ supp(X) ≥ supp(Y )*

Questa proprietà è alla base dell’algoritmo *Apriori*, dove, partendo da tutti i possibili item con cardinalità 1, si costruisce tutti i gli itemset di dimensione *n* + 1 con *n* la dimensione dell’itemset di partenza e ad ogni iterazione verifica se l’itemset generato è frequente o meno.

La proprietà anti-monotona permette di escludere itemset non frequenti e di conseguenza tutti possibili itemset derivanti da essi.

Gli step da cui è costituita la procedura sono i seguenti:

|  |  |
| --- | --- |
| **ALGORITHM APRIORI** | **DESCRIPTION** |
| 1: **function** Apriori(*T, s*) | (*set transazioni T* , *minSupport*) |
| 2: *L*1 ← {*large* 1 − *itemsets*} | *k* = 1 e *Generazione itemset con cardinalit*à 1 |
| 3: k=2 |
| 4: **while** *Lk*−1 ƒ= ∅ **do** | *Generazione itemset di cardinalità* *k* + 1. |
| 5: *Ck* ← *Generate*(*Lk*−1) |
| 6: **for** *transaction t* ∈ *T* **do** | *Eliminazione itemset contenenti non frequenti.* |
| 7: *Ct* ← *Subset*(*Ck*1*, t*) |
| 8: **for** *candidates c* ∈ *Ct* **do** | *Calcolo support itemset generati.* |
| 9: *count*[*c*] ← *count*[*c*] + 1 |
| 10: *Lk* ← {*c* ∈ *Ck*|*count*[*c*] ≥ *s*} | Eliminazione itemset non frequenti |
| 11: *k* ← *k* + 1 |
| 12: **return** U*k Li* |  |

Al termine di questa procedura, otteniamo tutti gli itemset che hanno superato la soglia supporto. Bisogna procedere con l’estrazione delle regole di associazione dagli itemset ottenuti. Le regole generate saranno valutate in base alla loro Confidenza (soglia arbitraria), e quest’ultima generalmente non gode della proprietà *anti-monotona*.

In questo caso specifico, invece, la confidenza delle regole generate dal solito itemset è a*nti-monotona* rispetto al numero di item che compongono la premessa della stessa regola possiede la seguente proprietà.

Indicando con *Conf* (*X* ⇒ *Y*) la confidenza della regola *X* ⇒ *Y* si otterrà:

*Conf* (*ABC* ⇒ *D*) ≥ *Conf* (*AB* ⇒ *CD*) ≥ *Conf* (*A* ⇒ *BCD*)

Si procede quindi generando le regole che possiedono solo un item nella conseguenza, eliminando tutte le regole che non superano la soglia minima di confidenza. Sulla base delle regole rimaste, si procede generando e valutando le regole con un item addizionale nella conseguenza, procedendo fino a che non sono state generate tutte le possibili regole.

Le regole estratte sono sottoposte ad un’ulteriore fase di post-processing, in quanto la confidenza a volte può essere fuorviante come indice di validità per una regola. Questo aspetto emerge per itemset che fanno parte della premessa di una regola, caratterizzati da alto supporto.

Un itemset molto frequente tende ad alzare l’indice di confidenza delle regole di cui esso costituisce la premessa, indipendentemente dal fatto che la regola sia contestualmente valida.

Per avere un’ottima validazione ci si basa sui seguenti indici:

Immagine che contiene oggetto

Descrizione generata con affidabilità molto elevata

Figura 15: Indici Di Validazione Di Una Regola Di Associazione

### ***1.7.4 Artificial Neural Networks & Deep Learning***

Il Deep Learning è un metodo specifico di apprendimento automatico che incorpora un numero elevato di reti neurali in vari strati successivi tra loro per apprendere dai dati in modo iterativo.

Le reti neurali e l'apprendimento profondo sono spesso usati nelle applicazioni di riconoscimento immagini, parlato e computer vision.

Una rete neurale è particolarmente utile quando si cerca di studiare i pattern da dati non strutturati e sono progettati per emulare come, attraverso l’intelligenza artificiale, i computer possano essere addestrati per trattare problemi che non sono ben definiti [21].

Essa consiste in tre o più livelli: uno strato di input, uno o più livelli nascosti e un livello di output. I dati sono ingeriti attraverso il livello di input. Quindi, i dati vengono modificati ed elaborati nel livello nascosto, ottenendo diversi livelli di output in base ai pesi applicati ai singoli nodi nascosti.

La tipica rete neurale può essere composta da migliaia o anche milioni di nodi di elaborazione semplici che sono densamente interconnessi.

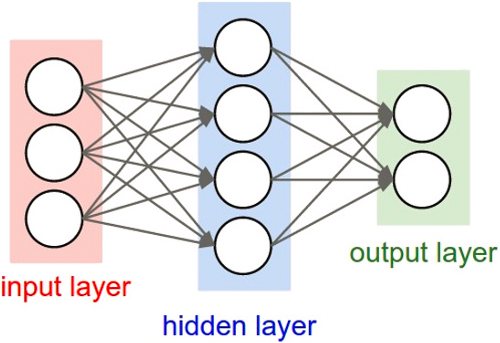


Figura 16: Artificial Neural Network Schema

In una rete neurale, gli input inviati ai *input layer* sono costituiti dal valore degli attributi dell’istanza che deve essere analizzata. L’output di questo primo livello della rete rimane invariato, poiché in uscita dai nodi di input sono presenti gli stessi valori che vengono forniti per l’analisi. In ogni nodo appartenente ai livelli successivi al primo, *hidden layer* e *output layer*, avviene l’effettiva computazione. Infatti, gli input di questi livelli corrispondono agli output dei livelli precedenti in cui, però, bisogna considerare il peso associato al collegamento tra i due nodi ed un valore caratteristico del nodo, l’*offset*. Considerando un nodo 𝑛 tra i nodi nascosti o tra quelli di output, il suo input 𝐼𝑛 è dato dalla seguente relazione:

Immagine che contiene oggetto, orologio

Descrizione generata con affidabilità elevata

dove 𝑤𝑖,𝑛 è il peso del collegamento tra il nodo 𝑖 del livello precedente e il nodo 𝑛 preso in considerazione, 𝑂𝑖 è l’output del nodo 𝑖 del livello precedente e 𝑜𝑓𝑓𝑠𝑒𝑡𝑛 è l’offset associato al nodo 𝑛 considerato.

Ogni nodo, inoltre, applica poi una funzione di attivazionesul valore che riceve in input ed invia il suo output al livello successivo. Infine, quando viene generato l’output dai nodi di output, se durante la fase di apprendimento si verifica un errore tra il valore della classe calcolato e quello previsto per un’istanza, viene calcolato l’errore da ogni nodo di output e viene propagato ai livelli precedenti, dove vengono sistemati i valori dei pesi e degli offset di tutti i nodi di tutti i livelli che costituiscono la rete neurale.

Il termine Deep Learning viene utilizzato quando ci sono più livelli nascosti all'interno di una rete neurale. Usando un iterativo approccio, una rete neurale si adatta e fa continuamente inferenze fino al raggiungimento di un punto di arresto specifico. Praticamente, è una tecnica di apprendimento automatico che utilizza la gerarchia delle reti neurali per imparare da dati non etichettati e non strutturati tramite una combinazione tra algoritmi non supervisionati e algoritmi supervisiona. Spesso viene chiamato Deep Learning una sotto-disciplina del Machine Learning.

Il Deep Learning viene usato nelle applicazioni dell’Internet of Things (IoT) o per prevedere quando una macchina funzionerà male.

### ***1.7.5 Regressione Lineare***

L'analisi di regressione è una tecnica statistica utilizzata per determinare una relazione tra una variabile dipendente e un insieme di fattori esplicativi. La variabile dipendente, indicata come variabile Y, è il valore che stiamo cercando di determinare in base ai fattori esplicativi.

I fattori esplicativi, indicati come variabili X, vengono anche chiamati fattori indipendenti, variabili predittive o semplicemente fattori modello. L'analisi di regressione aiuta gli analisti a scoprire la sensibilità della variabile dipendente ai cambiamenti nei fattori esplicativi. Queste sensibilità sono essenziali per una corretta gestione del rischio.

Esistono tre tipi di dati comunemente utilizzati nell'analisi di regressione: serie temporali, sezioni trasversali e dati raggruppati.

* Serie temporali: dati raccolti in un periodo di tempo. Nelle serie economiche e finanziarie questi dati si riferiscono spesso a rendimenti di mercato, rendimenti dell'indice, prezzi e valori delle attività, PIL, disoccupazione, tassi di interesse, ecc. Questi dati vengono raccolti a intervalli di tempo uguali come giornaliero, mensile, trimestrale, ecc.
* Sezione trasversale: dati raccolti per una famiglia di variabili nello stesso momento. Ad esempio, nell'analisi fondamentale raccogliamo spesso informazioni specifiche dell'azienda come il rapporto prezzo / utili, il rapporto prezzo / valore contabile, il rapporto debito / capitale netto o il volume medio giornaliero degli scambi.
* Dati raggruppati: dati che sono una combinazione di serie temporali e dati cross section.

Nel caso in cui abbiamo un singolo fattore esplicativo, l'analisi è chiamata un semplice modello di regressione e ha la forma:

Y = b 0+ b 1X + ε

dove Y è la variabile dipendente (quello che stiamo cercando di prevedere), X è il fattore esplicativo (quello che stiamo usando per predire), e ε è il rumore casuale(errore). Inoltre, la variabile dipendente Y, i fattori esplicativi x e il termine di errore ε sono vettori di colonne di valori.

Nell'equazione precedente, b 0 e b 1 sono i parametri del modello attuale che definiscono l'esatta sensibilità della variabile dipendente ai fattori esplicativi, e ε è la quantità di variabilità che non è spiegata dal modello.

In pratica, questi valori esatti non sono noti con certezza e devono essere stimati dai dati. Per fare ciò si utilizzano l’Expected Value e la Varianza:

Expected Value= E [b 0] = b 0\* E [b 1] = b 1 \* E [ε] = 0

Varianza= Var [ε] = σε2

Immagine che contiene testo, mappa

Descrizione generata con affidabilità molto elevata

Figura 17: Linear Regession

Nel caso in cui abbiamo più fattori esplicativi, l'analisi è denominata modello di regressione multipla e ha la forma:

Y = b 0+ b 1X+ b 2X2 + ⋯ b kXk + ε

L'obiettivo dell'analisi di regressione è determinare l'insieme di fattori esplicativi e sensibilità corrispondenti che spieghino il più possibile i valori dipendenti osservati.

**Metriche di valutazione e Presupposti del modello.**

Nell'effettuare l'analisi di regressione, le metriche importanti per analizzare l'analisi sono:

* b k = Parametro del modello - sensibilità stimata di Y al fattore k
* SE (b k) = errore standard del parametro stimato b k
* σ y = errore standard del modello di regressione
* R2 = bontà di adattamento (la percentuale della varianza complessiva spiegata dal modello)
* T-stat = valore critico per il parametro stimato
* F-stat = valore critico per l'intero modello

Esistono sei ipotesi o proprietà principali di un modello di regressione lineare. Se qualcuno di questi presupposti viene violato, i risultati dell'analisi potrebbero essere sospetti e potenzialmente fornire informazioni errate sulla reale relazione tra una variabile dipendente e i suoi fattori. In questi casi, gli analisti devono apportare modifiche ai dati. I principali presupposti del modello di regressione lineare sono:

* Relazione lineare tra variabile dipendente e fattori esplicativi.
* Termine di errore medio zero: il valore atteso del termine di errore è zero. E (ε) = 0.
* Variazione costante: ogni termine di errore ha la stessa varianza, ad esempio, nessuna eteroschedasticità. Var (ε k) = σ2 per tutti i k.
* Termini di errore non correlati, ad es. nessuna autocorrelazione. E (ε k \*ε k - t) = 0 per tutti i periodi di tempo ritardati t.
* Gli errori sono indipendenti da fattori esplicativi. Cov (ε, b k) = 0 per tutti i fattori k.
* I fattori esplicativi sono indipendenti. Cov (Xj, Xk) = 0 per tutti i fattori j e k.