LUISS Guido Carli
LIBERA UNIVERSITÀ INTERNAZIONALE DEGLI STUDI SOCIALI

Dipartimento di Impresa e Management

Cattedra Economia Dei Mercati E Degli Intermediari Finanziari

La Data Science nell'intermediazione finanziaria: il caso Modefinance

Prof. Daniele Previtali

RELATORE

Daniele Fiorucci Matr. 250061

CANDIDATO

Anno Accademico 2021/2022

INDICE DEI CONTENUTI

INTRODUZIONE

Al giorno d'oggi, la nostra società è attraversata da quella che può essere definita come una vera e propria rivoluzione, non meno importante della rivoluzione industriale avvenuta nella seconda metà del '700 o della rivoluzione informatica avvenuta negli anni '70 del '900. Quella odierna è strettamente legata all'avvento dei Big Data.

In una comunità sempre più digitalizzata e connessa, ogni azione compiuta da un individuo genera un flusso continuo di dati. I movimenti in rete vengono tracciati, per poter così registrare informazioni che possono andare dagli interessi personali, passando per i gusti culinari, fino ad arrivare alle preferenze riguardo le prossime elezioni politiche.

In quella che può essere definita come "Era dei dati" sono profondi i cambiamenti che già al giorno d'oggi stanno avvenendo nella società, sia sotto il profilo sociale che economico, in cui le GAFAM (Google, Apple. Facebook, Amazon) si pongono in una posizione di spicco data l'enorme quantità di dati che circolano all'interno dei propri database.

È fondamentale per le aziende cercare di anticipare i trend emergenti in una comunità che si trova in uno scenario di continuo cambiamento, cercando di cogliere i benefici che la Data Science, e quindi le tecniche di estrazione di valore dai dati, possono apportare al mondo del business.

Tra i settori che utilizzano maggiormente i dati come materia prima, e che quindi possono beneficiare dell'evoluzione sistemica appena descritta, spiccano sicuramente gli intermediari finanziari.

I dati possono essere, in particolare, utilizzati dalle banche non solo come uno strumento per indirizzare la propria strategia, ma anche come mezzo per verificare l'affidabilità degli agenti economici con cui interagiscono quotidianamente. L'applicazione che più di tutte può trarre beneficio da una grande quantità di informazioni è, infatti, il Credit Scoring.

Il Credit Scoring è il sistema utilizzato dalle banche per svolgere la valutazione del merito creditizio dei clienti, questo processo per essere implementato necessita l'estrazione di informazioni utili da enormi quantità di dati disordinate e apparentemente non collegate tra loro, proprio questo è l'obiettivo principale della Data Science.

Se da un lato la diffusione e lo studio dei Big Data può generare un vantaggio competitivo per le aziende che ne intuiscono il valore, dall'altro queste portano con sé diversi rischi dal punto di vista della privacy dei cittadini e delle truffe dei clienti degli intermediari finanziari.

La scienza dei dati, inoltre, necessita grandi investimenti di capitale, per potersi munire dei strumenti tecnologicamente adeguati e per formare i data scientist.

Nel primo capitolo dell'elaborato, viene definito il concetto di Big Data e descritte le cause che hanno portato nel corso degli ultimi anni ad un aumento così grande dei dati. Dopodiché, dopo aver

introdotto e descritto la Data Science e l'interdisciplinarietà di questa scienza, verranno descritti in breve alcuni metodi di elaborazione dei dati e alcuni campi di applicazione degli stessi. Per concludere, sono riportate argomentazione a favore della Data Science come disciplina aziendale ed è descritto come le problematiche legate alla stessa scienza possono essere superate adottando una visione a lungo termine.

Nel secondo capitolo, si approfondiscono: le modalità di applicazione della Data Science in un campo ricco di dati come è quello degli intermediari creditizi, le implementazioni delle tecniche di credit scoring tramite il Data Mining, i fattori di successo e i rischi a cui si può andare incontro.

Nel terzo capitolo viene descritta la società Modefinance, società relativamente giovane che da anni si occupa di applicare i meccanismi di Data Science e AI ai classici modelli di valutazione del merito creditizio. Per concludere, sarà riportata la testimonianza del CEO e co-founder di Modefinance, Valentino Pediroda.

1. BIG DATA E AMBITI DI APPLICAZIONE

1.1. Definizione e caratteristiche di Big Data e Data Science

I Big Data sono enormi quantità di dati che negli ultimi anni hanno visto una crescita esponenziale, arrivando addirittura a poter essere misurati in Zettabyte ZB[1] (raggiunto per la prima volta nel 2011).

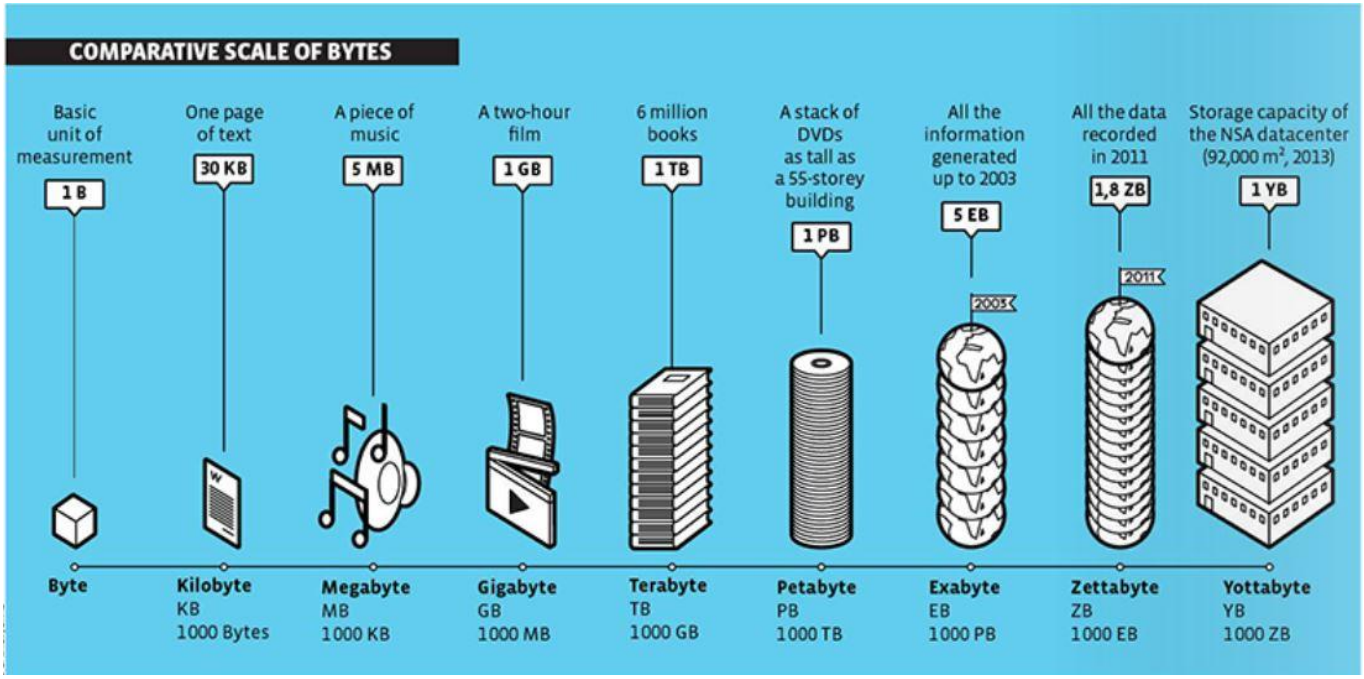

Figura 1.1 Scala comparative di byte

Fonte: Fisher, Tim. "Terabytes, Gigabytes, & Petabytes: How Big Are They?" Accessed from (2017).

Non è difficile immaginare che tale quantità di dati oggi possa essere generata in molto meno di un anno. Nell'ultimo decennio, infatti, stanno vedendo la luce unità di misura sempre maggiori, con la conseguente nascita di infrastrutture che hanno l'obiettivo di permettere ad una quasi illimitata quantità di dati di circolare .

Questi dati sono per la maggior parte generati dall'uomo in formato digitale. Sono una quantità notevole e provengono da numerosi fonti, come satelliti spaziali o più banalmente foto caricate su Instagram. Il seguente grafico a bolle mostra quelli che sono i maggiori produttori di dati. Tra parentesi sono riportati i dati generati per unità, in corsivo nel caso in cui non ci siano fonti riguardo le stime, mentre la grandezza delle bolle rappresenta la quantità degli stessi. In grassetto la tipologia e quantità degli oggetti utilizzati per redigere le stime.

[1]Zettabyte (ZB), multiplo del byte tale che 1 ZB = 10^21 byte.

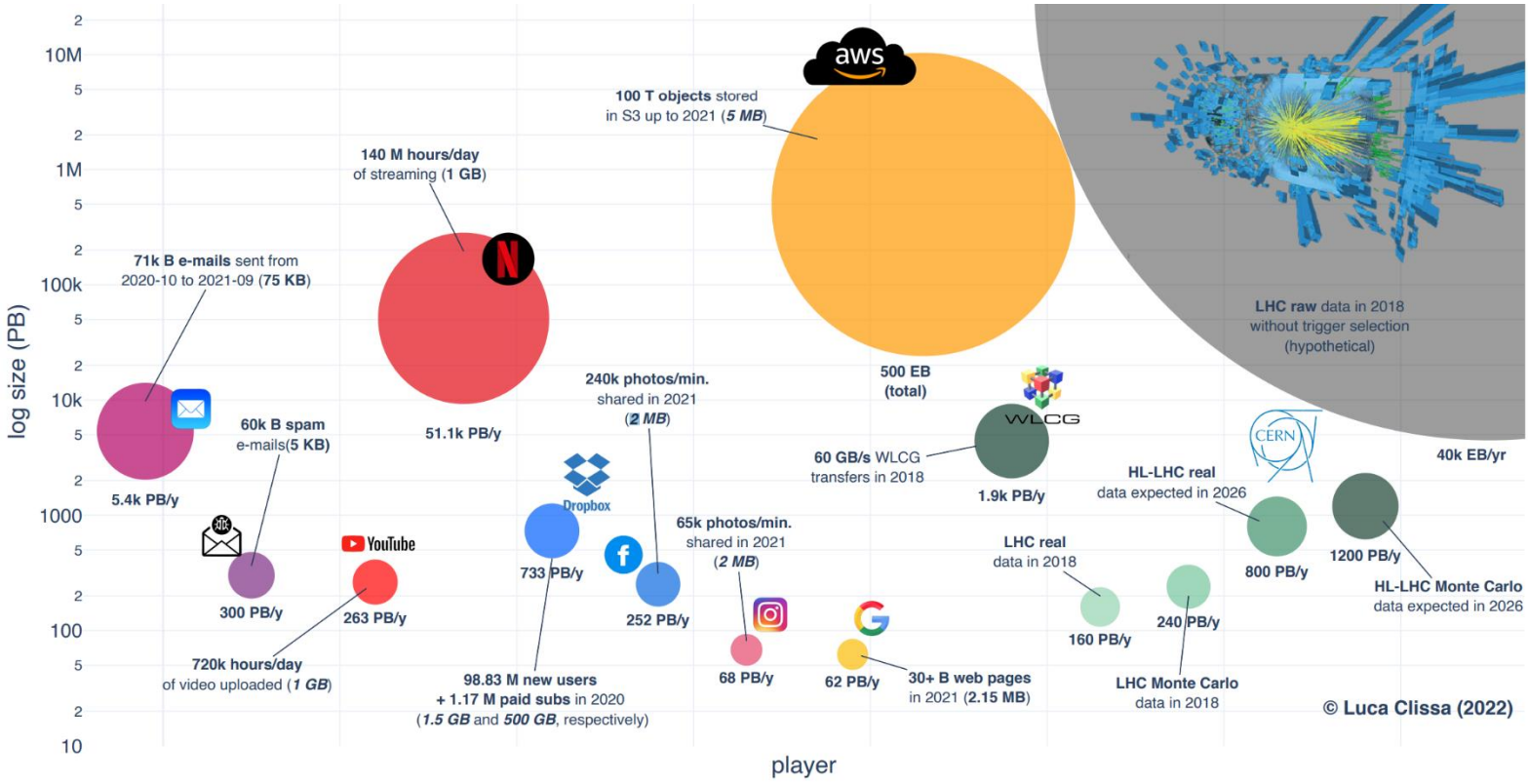

Figura 1.2 Maggiori produttori di dati

Fonte: Clissa, Luca. "Survey of Big Data sizes in 2021." arXiv preprint arXiv:2202.07659 (2022).

La rivoluzione dei Big Data, che sta attraversando il mondo da più di un decennio, è dovuta a due fattori principali. Il primo, riguarda la crescita senza precedenti di dati generati e raccolti mediante spazi di archiviazione sempre maggiori, il secondo, consiste nelle innovazioni tecnologiche e digitali recenti.

Le moderne tecnologie permettono di analizzare ed elaborare quantità di informazioni, inimmaginabili fino a pochi anni fa, in tempi decisamente ridotti. Il binomio tra Big Data e innovazioni tecnologiche può permettere di trarre vantaggio dalle grandi quantità di dati, estraendone informazioni che possono essere sfruttate dalle aziende per indirizzare la propria strategia.

Questo cambiamento non riguarda solo il mondo del business, sono, infatti, molteplici i campi d'applicazione che possono trarre benefici dall'analisi dei dati, o meglio dei Big Data.

Al riguardo giova ricordare il Global Legal Hackathon del 2019. Nel 2019 un gruppo di 20 avvocati ha sfidato un sistema di intelligenza artificiale (d'ora in poi IA) nell'analisi dei rischi contenuti in cinque documenti contenenti accordi di non divulgazione (d'ora in poi NDA). Il sistema di IA ha risposto con un'accuratezza del 94%, contro lo 85% del gruppo di specialisti. La risposta del computer è arrivata dopo 26 secondi, mentre gli avvocati hanno impiegato 92 minuti nell'analisi dei 5 documenti.

Non esiste una definizione univoca di Big Data, si preferisce quindi procedere espletando quelle che sono le caratteristiche più importanti di queste enormi moli di dati. Gli attributi principali possono essere racchiusi in quelle che sono le cosiddette "5V", ossia volume, velocità, varietà, valore e veridicità. Analizziamole più nel dettaglio:

- *Volume*: in precedenza abbiamo accennato alla grande mole di dati generati e in costante aumento. Nel 2013 lo spazio occupato nel web dai dati era quantificabile nell'ordine dei Petabyte (10^15 byte)[2], al giorno d'oggi si stima, invece, che siano stati raggiunti i 44 Zettabyte[3] (10^21 byte). L'aspetto negativo di questa caratteristica riguarda la qualità e la comprensibilità delle informazioni immagazzinate, non sempre ottima.
- *Velocità*: Ogni giorno vengono acquisiti molto rapidamente dati strutturati e non. Ciò rende sempre più complessa e dispendiosa in termini di energia e denaro l'analisi di ogni mole nuova di dati. Basti pensare a quanto è immediata la ricezione di informazioni provenienti dai social network, da chiamate ad un'assistenza tecnica, da risultati di test e sondaggi o risultanti dalle localizzazioni abilitate sui nostri telefonini. Per questo motivo è necessario elaborare nuovi strumenti che possano aumentare la velocità di analisi e scomporre i dati ricevuti, in modo da rilevare elementi utili e profittevoli per il business, come quelli geografici, demografici, comportamentali ed economici.
- *Varietà*: i dati provengono da fonti diverse e sono eterogenei anche per quanto riguarda il formato (dati strutturati, semi-strutturati e non strutturati, approfonditi al paragrafo 1.3)
  Ciò avviene poiché sempre nuove tecnologie vengono introdotte e la platea, di conseguenza, sviluppa abitudini differenziate. Questa caratteristica se da un lato è sinonimo di ricchezza, dall'altro può portare a complicazioni nell'associare dati qualitativamente differenti tra loro.
- *Valore*: possedere grandi quantità di dati risulta inutile nel momento in cui non si è in grado di estrarre valore dalle stesse. Non sono i dati in quanto tali ad apportare un vantaggio competitivo alle aziende, quanto le informazioni che vengono ricavate da essi. L'obiettivo principale della scienza dei dati è proprio tradurre informazioni quantitative in qualitative, così che possano essere sfruttate dalle aziende per ottenere benefici.
- *Veridicità*: per veridicità si intende l'accuratezza del dato e il suo significato all'interno del giusto contesto. Dati "inquinati" possono contaminare i risultati delle analisi e portare a risultati distorti. Riuscire ad affinare i filtri in modo tale da ottenere dati veritieri

[2] Katal, Avita, Mohammad Wazid, and Rayan H. Goudar. "Big data: issues, challenges, tools and good practices." In *2013 Sixth international conference on contemporary computing (IC3)*, pp. 404-409. IEEE, 2013.
[3] Gutierrez, Felipe. "Cloud and big data." In *Spring Cloud Data Flow*, pp. 3-8. Apress, Berkeley, CA, 2021.

permetterebbe di prevedere processi capaci di portare a decisioni istantanee, che, altrimenti, sarebbero impraticabili senza una selezione precedente delle informazioni.

Un esempio importante, capace di dimostrare quanto essenziale sia la necessità di dati veritieri e quanto, invece, disastrose le conseguenze dell'utilizzo di dati inquinati è dato dal caso Loomis e dal software C.O.M.P.A.S. (*Correctional Offender Management Profiling for Alternative Sanctions*). Il Signor Loomis fu condannato a scontare 6 anni di reclusione poiché ritenuto socialmente pericoloso da C.O.M.P.A.S, un software di supporto alla decisione del giudice creato dalla Northpointe (ora Equivant) e usato nel sistema giudiziario USA per calcolare la probabilità di recidiva. Loomis impugnò la sentenza contestando la non conoscibilità dell'algoritmo e il fatto che questo conteneva dei pregiudizi. Se i precedenti penali e l'età hanno valore determinante, l'ipotesi di maggiori controlli eseguiti a carico di comunità etniche (o religiose) considerate a maggior rischio determina un maggior numero di condanne a carico dei componenti di quella comunità.

Ciò non significa, però, che tutti coloro che appartengono a quella comunità siano più pericolosi degli altri.

I dati inseriti nella macchina in questo caso riflettevano i pregiudizi della società, erano quindi dati inquinati, al punto di portare a decisioni errate con conseguenze gravi.

Una delle principali cause che hanno portato alla nascita del fenomeno dei Big Data, è sicuramente l'Internet of Things[4] (d'ora in poi IoT). L'IoT rappresenta la possibilità di poter collegare oggetti ad Internet, in modo tale che la condivisione dei dati tra le aziende produttrici dei software e gli utenti possa implementare e personalizzare l'esperienza di quest'ultimi.

In un momento storico in cui prolifera la quantità di dispositivi costantemente connessi alla rete, i dati rappresentano la "traccia digitale" del passaggio dell'essere umano sul pianeta Terra.

Per far sì che questa grande materia prima, possa essere sfruttata, è necessario che i Big Data vengano analizzati, classificati e manipolati nel modo corretto per ottenere nuove conoscenze. È a tal fine che nasce la Data Science.

È chiamata Data Science la scienza che, sfruttando metodi statistici, matematici ed informatici, si pone come obiettivo quello di estrarre informazioni e valore dai Big Data.

[4] (Internet delle cose) *loc. s.le m.* Rete di oggetti collegati tra loro, dotati di tecnologie di identificazione, in grado di comunicare sia reciprocamente, sia verso punti nodali del sistema, ma, in particolare, in grado di costituire un enorme network di cose, ciascuna delle quali è rintracciabile per nome e in riferimento alla posizione che occupa, Treccani, <https://www.treccani.it/vocabolario/internet-delle-cose_%28Neologismi%29/>.

Activities	Examples
Data gathering, preparation, and exploration	Survey data, experimental data, genomic data, textual data, administrative data, image data, web data, and sensor data Data cleaning and exploratory data analysis methods for checking on outliers and data quality
Data representation and transformation	Relational and nonrelational databases Networks and graphs Other mathematical structures for data
Computing with data	R and Python Programming packages, text manipulation languages Cluster and cloud computing Reproducible workflows
Data modeling	Determining or hypothesizing data generating probability functions, structural and predictive modeling
Data visualization and presentation	Types of visualizations and graphs Rules for labeling and presenting data Psychological impacts of various displays
Data archiving, indexing, and search and data governance	Standards for open data and reproducibility Determining rules for access and privacy protection where necessary
Science about data science	How people do data science Impacts of data science and big data on society

Figura 1.3 Principali attività della Data Science

Fonte: Brady, Henry E. "The challenge of big data and data science." Annual Review of Political Science 22 (2019): 297-323.

Tra i fondatori di questa scienza dobbiamo sicuramente includere lo statistico William S. Cleveland, egli già nel 2001 si rendeva conto che la disciplina della statistica dovesse essere ampliata per includere al suo interno un campo specializzato nel calcolo dei dati.

Egli presentò un piano per poter raggiungere questo scopo e, probabilmente per primo, coniò il termine "Data science".

Questo concetto è stato successivamente ripresentato dall'informatico Jim Gray nel 2007 al "*National Research Council's Computer Science and Telecommunications Board*" arrivando fino ad una descrizione della nuova scienza fornita nel 2015 dal NIST[5].

L'interdisciplinarietà, caratteristica alla base di questa recentissima scienza, fa sì che sia ancora incerto il suo confine d'azione. La Data Science è, infatti, legata ad altre discipline quali:

- *Artificial intelligence*: scienza che cerca di traslare i ragionamenti del pensiero umano all'interno delle macchine. La sfida principale sta nello sviluppare algoritmi che si aggiornino automaticamente in base alle esperienze vissute dalla macchina stessa. I Big Data sono il carburante che permette all'IA di esistere e di perfezionarsi. Raccogliere grandi quantità di

[5] Ufficio del Dipartimento del commercio del governo degli USA, con l'iniziale compito di conservare i vari campioni di unità di misura, attualmente articolato in sezioni che si occupano di ricerche in vari campi scientifici e tecnici, Treccani, https://www.treccani.it/enciclopedia/national-institute-of-standards-and-technology/.

dati è, oggi, un'operazione generalmente facile poiché sono di molto aumentate le tecnologie che ne producono.

- *Data mining*: si pone come obiettivo quello di estrarre valore da enormi quantità di dati disorganizzati cercando schemi frequenti e correlazioni tra essi. È strettamente legato all'artificial intelligence (AI), e, nello specifico, al machine learning.
- *Deep learning*: a differenza degli altri processi, in questo caso, è l'algoritmo stesso a stabilire le varie interconnessioni e classificazioni possibili di tutti i dati raccolti.
- *Machine learning*: è una branca dell'AI ed è forse la più datata tra queste discipline. Essa si pone come obiettivo quello di sviluppare algoritmi che si aggiornino automaticamente in base alle esperienze vissute dalla macchina, senza bisogno di aggiornamenti implementati da interventi esterni.

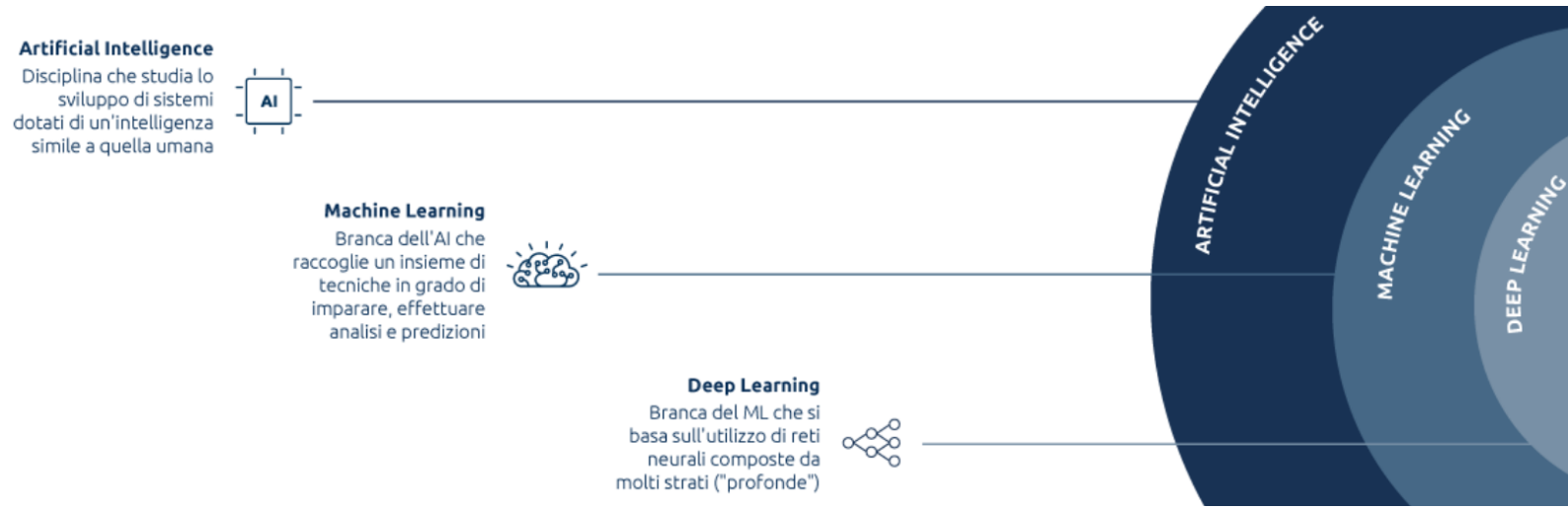

Figura 1.4 Tipologie di Learning

Fonte: Cirese, Claudio. "L'evoluzione del digital marketing nell'era dei big data: pro e contro del consumer profiling." (2021).

Il seguente grafico illustra quella che è la crescita delle ricerche Google tra il 2015 e il 2021 di temi quali la data mining, l'AI, la deep learning e la machine learning.
Riguardo gli ultimi tre, possiamo vedere come l'attenzione sia notevolmente cresciuta.

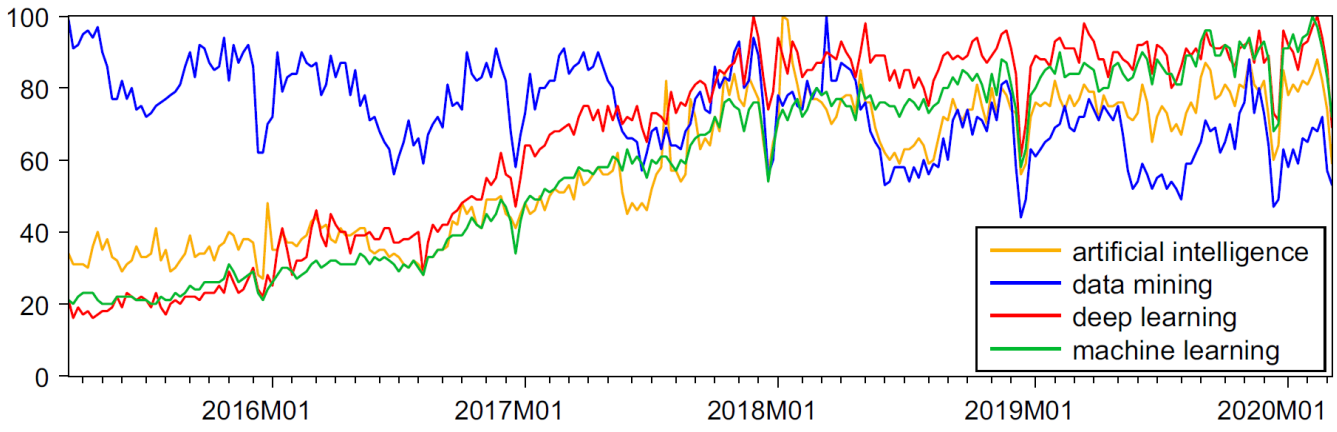

Fig.1.5 Frequenza di ricerca globale dei termini

Fonte: Google Trends

I processi produttivi sono fonte di dati, come lo sono anche alcuni prodotti. Un esempio di quest'ultimi possono essere i Google Glass, l'Apple Watch o gli assistenti vocali come Alexa e Siri. Inoltre, una recente fonte di dati è il riconoscimento facciale, che rinveniamo ormai su quasi tutti i telefoni e spesso anche implementato in app (es. App Store in cui il riconoscimento facciale permette di autorizzazione al download delle app).

Determinate categorie di dati prodotti, se non ormai quasi tutti, sono di grande valore per le aziende ed essere elaborati così da produrre reddito (approfondimento al paragrafo 1.4).

1.2. Fonti di generazione dei dati e ambiti di applicazione della Data Science

Sono molti i cambiamenti nello stile di vita, nelle abitudini e, quindi, nella quotidianità che la rivoluzione digitale ha apportato.

Da un punto di vista informatico, alla base del cambiamento c'è stata la traduzione e la possibilità di codificare testi, immagini e suoni in sequenze di codici binari. Da un punto di vista storico, la rivoluzione digitale ha fatto sì che nascesse un'intersezione, un'area di studio comune, tra discipline che fino a quel momento erano considerate individualmente.

Quest'ultimo cambiamento, insieme all'avvento di Internet, ha mutato irreversibilmente le abitudini degli esseri umani, il cui risultato può essere riassunto nella locuzione "digital life style"[6].

[6] Sciotto, Fortunata, Giovanna Zucchi, Daniel Guzzetti, and Giorgio Corbucci. "La rivoluzione digitale." *GIAC* 6, no. 2 (2003).

La “collaborazione” tra settori quali l’informatica e l’elettronica ha portato alla nascita di quella che è probabilmente una delle maggiori “cause” dell’aumento delle fonti di generazione dei dati del 21esimo secolo: l’IoT.

L’IoT è il concetto secondo il quale oggetti di uso comune possono essere costantemente connessi alla rete, alla quale vengono trasmessi i dati raccolti tramite sensori. Questo processo permette ai software che gestiscono le funzionalità degli oggetti di essere costantemente aggiornati e di poter, tramite le informazioni raccolte e salvate nel cloud, personalizzare l’esperienza dell’utente.

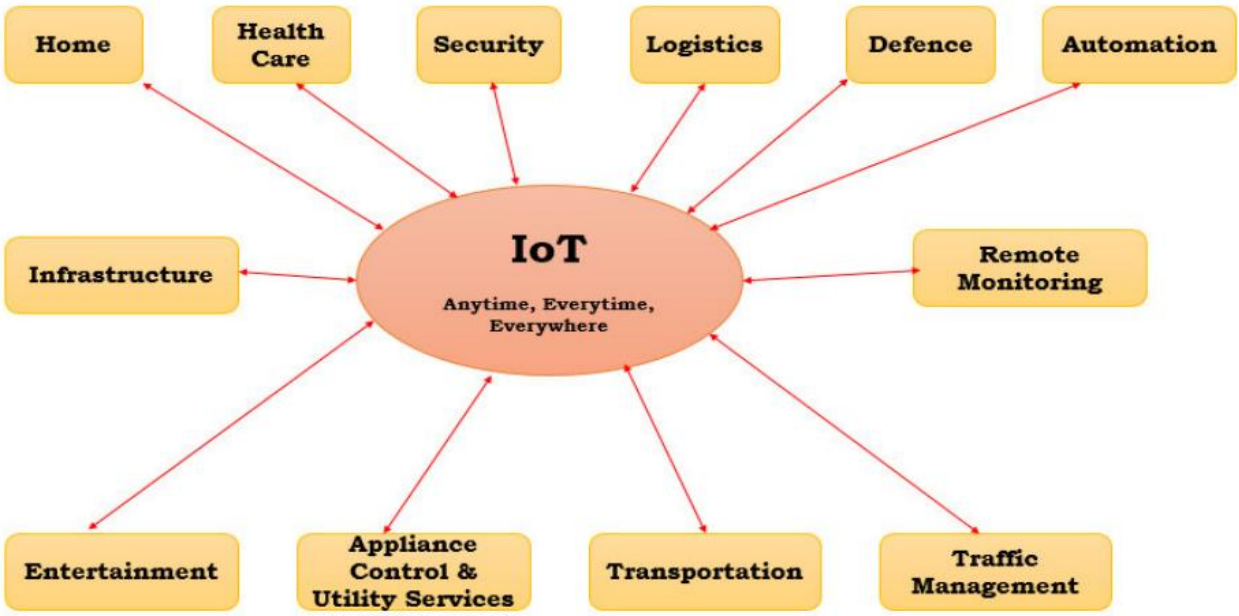

Figura 1.5 Aree applicazione IoT

Fonte: Vyas, Daiwat A., Dvijesh Bhatt, and Dhaval Jha. "IoT: trends, challenges and future scope." IJCSC 7, no. 1 (2015): 186-197.

Assimilabile al boom di Internet negli anni ’90, guidato da nuovi calcolatori di dimensioni sempre più contenute, lo studio dei dati è stato, inoltre, favorito da computer sempre più potenti che permettono di migliorare non solo la quantità dei dati memorizzabili, con costi sempre minori, ma soprattutto la qualità degli stessi.

Mentre prima la registrazione di una spedizione poteva racchiudere al suo interno solo le informazioni essenziali, come, ad esempio, luogo e avvenuta consegna, ora un’operazione come questa potrà generare dati che comprendo anche i tempi di spedizione, il peso dei pacchi e il carburante speso per la spedizione.

Come anticipato, una delle 5V che definiscono le caratteristiche fondamentali dei Big Data è quella della Varietà.

I dati possono essere, oggi, generati da una moltitudine di sorgenti che differiscono, anche grandemente, tra loro.

Lo studio dei dati ha un impatto rilevante in diversi campi, tra cui:

- *Social network*: fino a pochi anni, era impensabile fosse possibile di studiare su larga scala e a costi relativamente ridotti i gusti e lo stato d'animo del consumatore prima, durante e dopo aver fruito di un servizio. Il compito di raccogliere questi dati è oggi svolto dai social network che registrano una vastissima quantità di informazioni tra cui, ad esempio, le modalità con cui un utente si rapporta ad un determinato brand, come lo valuta in termini di qualità, assistenza, prezzo.
- *Agricoltura*: è frequente l'utilizzo di sensori per massimizzare la resa delle colture. Essi, registrano, in particolare, i cambiamenti delle condizioni ambientali dovuti a stimoli esterni così da verificare quale siano i processi che impattano negativamente i livelli ideali[7].
- *Healthcare:* in un mercato con pochi margini di manovra come quello che riguarda le forniture per l'assistenza sanitaria pubblica, la Data Science sta intervenendo per ottimizzare i tempi di reazione delle strutture sanitarie pubbliche alle variabili ambientali e sociali, ma anche al fine di realizzare una riduzione dei costi per i venditori e i compratori. Sono state, inoltre, dimostrate le applicazioni e l'utilità imprescindibile della scienza dello studio dei dati nella prevenzione delle malattie e nell'identificazione delle modalità di intervento migliori per la risoluzione delle varie possibili disfunzioni[8].
- *Smartphone:* lo studio dei Big Data ha permesso di analizzare i parametri fisici registrati tramite la fotocamera degli smartphone per realizzare uno strumento che fino ad un decennio fa sarebbe stato ipotizzabile solo in un film di fantascienza, cioè il riconoscimento facciale. La sua introduzione e il successivo sviluppo di questa funzionalità, ormai ampiamente utilizzata, ha permesso un aumento della sicurezza dei dispositivi elettronici e, vantaggio non di poco conto, una riduzione ai costi dell'assistenza che riguardava il recupero delle password e l'accesso a dei servizi particolarmente protetti (es. conto bancario).
- *Business*: come verrà approfondito in seguito, a causa della sua intrinseca caratteristica di generare enormi quantità di dati, il business è sicuramente la macroarea che può sfruttare nel migliore dei modi le mille sfaccettature della Data Science.

[7] Sravanthi, Kuchipudi, and Tatireddy Subba Reddy. "Applications of big data in various fields." *International Journal of Computer Science and Information Technologies* 6, no. 5 (2015): 4629-4632.
[8] Hansen, M. M., T. Miron-Shatz, A. Y. S. Lau, and C. Paton. "Big data in science and healthcare: a review of recent literature and perspectives." *Yearbook of medical informatics* 23, no. 01 (2014): 21-26.

1.3. Metodi di elaborazione dei dati e principali sfide della Data Science

È chiaro che, se i dati rappresentano la più grande ricchezza del 21esimo secolo, per renderli utilizzabili è necessario estrarne valore.

Lo sviluppo tecnologico ha fornito gli strumenti che permettono di fare ciò, ad esempio, tramite la costruzione di moderni algoritmi capaci di realizzare analisi precedentemente non possibili con i sistemi di elaborazione dati tradizionali.

Nell'ampio ventaglio di modelli e metodi dedicati allo studio dei dati, quelli applicati più frequentemente al campo economico possono essere ricondotti a quattro classi principali[9]:

- *Deep learning models*: la loro caratteristica principale, che li rende utili in diversi campi, è la capacità di riconoscere schemi ricorrenti in conglomerati di dati grezzi. A differenza dei modelli di machine learning è comunque necessaria una supervisione umana esterna. Riescono a gerarchizzare in modo efficiente campioni di dati diversi tra loro e disorganizzati.
- *Hybrid deep learning models*: è una combinazione tra processi di machine learning e deep learning. Gli hybrid deep learning models, nonostante in termini di precisione siano migliori i deep learning models, sono la classe di modelli più applicata nell'economia. Essi applicano sia processi generativi, che hanno come obiettivo quello di prevedere cosa accadrà in un tempo immediatamente successivo, sia processi discriminativi, utilizzati per la classificazione.
- *Hybrid machine learning*: sono il risultato della combinazione tra modelli di machine learning. Questa tipologia di modelli ibridi permette di implementare l'efficienza di previsione di eventi futuri ed è spesso utilizzata per formulare ipotesi riguardo il futuro prezzo delle azioni o la valutazione del rischio di credito (c.d. Credit Scoring).
- *Ensemble models*: è combinazione di più algoritmi di Machine learning che portano alla generazione di un modello unico migliore in termini di apprendimento dei dati e precisione.

Dalla definizione di queste quattro classi di modelli è ancor più evidente, come già sottolineato in precedenza, la diretta connessione tra i processi di Data Science, Deep learning e Machine learning.

Se un sistema è progettato mediante modelli di Machine learning, ed è quindi in grado di registrare le proprie esperienze così da correggere i propri errori, "apprendendo" dagli input che riceve, allora sarà ancora più evidente la distinzione rispetto ad un sistema predeterminato.

[9] Waller, Matthew A., and Stanley E. Fawcett. "Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management." *Journal of Business Logistics* 34, no. 2 (2013): 77-84.

I dati necessitano di essere processati per poter essere effettivamente utili, non basta che vengano conservati. Dopo essere stati raccolti devono essere contestualizzati tramite l'utilizzo di ulteriori informazioni.

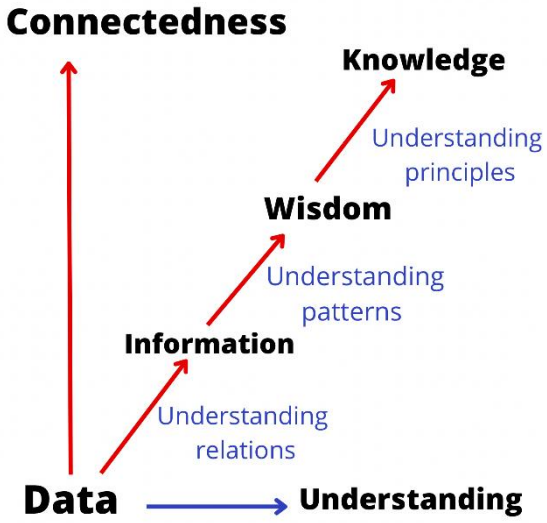

Figura 1.6 Dai dati alla conoscenza
Fonte: Cirese, Claudio. "L'evoluzione del digital marketing nell'era dei big data: pro e contro del consumer profiling." (2021).

Le esperienze degli utenti generano informazioni che sono da considerare come il carburante dell'IA, ciò che permette a questa di elaborare meccanismi di risposta sempre più precisi e rapidi.
Parallelamente all'innovazione e ai benefici che l'applicazione della Data Science è in grado di apportare a campi di studio diversi tra loro, essa porta con sé anche diversi rischi e sfide, nodi il cui districamento è di fondamentale rilievo per lo sviluppo della scienza stessa.
Tra i problemi riscontrati vi è che, proiettando nel tempo la crescita dei dati generati avvenuta negli ultimi anni in seguito all'avvento dei big data, presto la quantità di spazio destinato all'archiviazione non sarà sufficiente per immagazzinare l'enorme quantità di informazioni.
Questa sfida tecnologica sta quindi nel riuscire ad allineare la velocità di produzione dei dati quantomeno a quella con cui crescono gli spazi per l'archiviazione degli stessi.
Per far sì che lo studio proposto in seguito sul problema dell'analisi della qualità dei dati possa essere compreso in modo proficuo, deve essere prima introdotta la seguente classificazione di dati:

- *Dati strutturati*: dati che seguono un determinato schema e rispettano una struttura standard e possono essere immagazzinati secondo modalità predefinite. Esempi di dati strutturati sono le tabelle, i numeri o le stringhe.

- *Dati semi-strutturati*: dati che seppur non rispettando una struttura specifica, possiedono indicatori che permettono la scomposizione e classificazione.
  Esempi di dati semi-strutturati sono i codici HTML.
- *Dati non strutturati*: insiemi di dati diversi tra loro per tipologia e caratteristiche, non sono riconducibili a strutture standard e non possono essere immagazzinati in modo funzionale all'analisi previa precedente elaborazione.
  Esempi di dati non strutturati sono i file audio o le immagini.

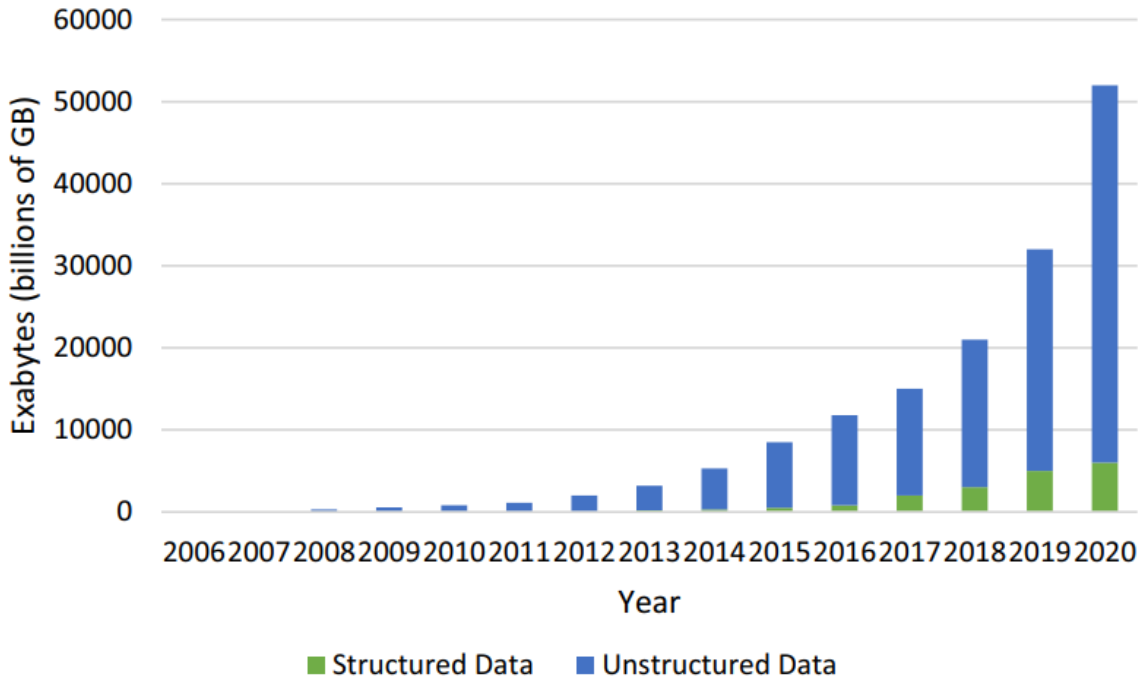

Figura 1.7 Crescita negli anni di dati strutturati e non strutturati
Fonte: Azad, Poopak, Nima Jafari Navimipour, Amir Masoud Rahmani, and Arash Sharifi. "The role of structured and unstructured data managing mechanisms in the Internet of things." Cluster computing 23, no. 2 (2020): 1185-1198.

La crescita esponenziale di dati non strutturati rispetto ai dati strutturati è dovuta soprattutto alla diffusione dell'IoT. Gli "oggetti" connessi al network sono col passare del tempo sempre più variegati, e i flussi di dati che riescono a produrre sempre maggiori sia per portata che per velocità. Alla luce di questi fatti, diventa fondamentale l'implementazione di database[10] che dovranno essere configurati nell'ottica di un aumento di velocità d'interrogazione, scalabilità e flessibilità, riducendo allo stesso tempo i costi di archiviazione richiesti dagli stessi.

[10] Archivio elettronico di dati correlati, registrati nella memoria di un computer e organizzati in modo da poter essere facilmente, rapidamente e selettivamente rintracciabili uno per uno, oppure per gruppi determinati, mediante appositi programmi di gestione e di ricerca (chiamati anch'essi *data base*, ma più propr. denominati *data base management system*, in sigla *DBMS*), Treccani, < https://www.treccani.it/vocabolario/data-base/>

Il futuro sviluppo della Data Science dovrà inoltre concentrarsi sulla qualità dei dati[11].

Questo perché riuscire ad acquisire grandi quantità di dati, espandere gli spazi di archiviazione e sviluppare nuove metodologie e algoritmi per lo studio dei Big Data risulta del tutto inutile se poi gli stessi non rispecchiano la realtà, anzi una bassa accuratezza dei dati potrebbe portare a studi con risultati fuorvianti e quindi a processi decisionali sbagliati.

Il problema della qualità dei dati è sorto nel momento in cui le aziende hanno iniziato a pianificare la propria strategia non solo sui dati prodotti dai propri sistemi ma, con l'avvento del fenomeno dei Big Data, anche su quelli generati e raccolti da fonti quali per esempio i social network, l'IoT e sensori di ultima generazione. Questa varietà di fonti porta ad una difficoltà, organizzativa e tecnologica prima ancora che in termini di costi, nel riuscire a classificare e a stabilire delle relazioni tra un gran numero di tipologie di dati.

Quando i dati utilizzati dalle imprese erano, se non interamente, generati internamente alle stesse, l'azienda poteva pensare di svolgere il compito di trasformare i dati non strutturati in dati strutturati manualmente senza impiegare un tempo eccessivo.

Con l'avvento dei Big Data, non essendo questo più possibile, la priorità dei reparti di Data Science delle aziende, e non solo, sta nel riuscire a progettare algoritmi che riescano in automatico, mediante meccanismi di machine learning, ad immagazzinare, filtrare, classificare ed elaborare dati , nel minor tempo possibile.

Un ulteriore problematica riguarda la tutela dei dati personali, e le relative lacune normative, in un momento storico in cui essi rappresentano il nuovo oro nero per le aziende.

I Big Data rappresentano, infatti, il veicolo attraverso il quale le imprese hanno la possibilità di personalizzare, adattando i propri prodotti, l'esperienza del cliente. Per far ciò è necessario utilizzare come materia prima i dati degli utenti, in particolare quelli personali.

Queste informazioni sono, o almeno dovrebbero, essere tutelate ai sensi del Regolamento (UE) 2016/679 che ha come uno dei principali obiettivi quello di limitare lo sfruttamento di questi dati da parte delle aziende.

La criticità principale sta nel fatto che la velocità a cui si muove il progresso tecnologico, non va di pari passo con l'adeguamento della normativa riguardante la tutela dei dati personali. Questo disallineamento ha come diretta conseguenza il fatto che le aziende cerchino di eludere le norme riguardanti la privacy degli utenti, così da ottenere quello che può essere considerato sul mercato come un vero e proprio vantaggio competitivo[12].

[11] Cai, Li, and Yangyong Zhu. "The challenges of data quality and data quality assessment in the big data era." *Data science journal* 14 (2015).

[12] Mandarà, Elena. "Tutela dei dati personali e antitrust: il caso dei sistemi IoT alla luce della sentenza Facebook/Germany (Bundeskartellamt B6-22/16)." (2021).

Per sovvertire il trend, è necessaria una stretta collaborazione tra il legislatore e gli esperti di Data Science. Questa cooperazione porterebbe ad un doppio vantaggio che si manifesterebbe da un lato in una maggior tutela della concorrenza, dall'altro nella tutela dei dati personali e della privacy degli utenti.

1.4. Data science nel business: una grande potenzialità per le aziende

I dati, o meglio i Big Data, possono essere sfruttati dalle aziende per migliorare i processi produttivi, per ottimizzare l'allocazione delle risorse e soprattutto, come già espletato nel paragrafo precedente, per orientare la propria offerta sulla base di quelle che sono le preferenze dei clienti target.

Sono diversi i settori correlati al mondo aziendale per cui la Data Science può operare come uno strumento capace di creare valore aggiunto[13].

Come possiamo osservare dalla figura sottostante, le pubblicazioni scientifiche riguardanti applicazioni della Data Science all'economia, crescono anno dopo anno, questo perché maggiore è la crescita della produzione dei dati generati maggiore è il rilievo che lo studio di questi può assumere.

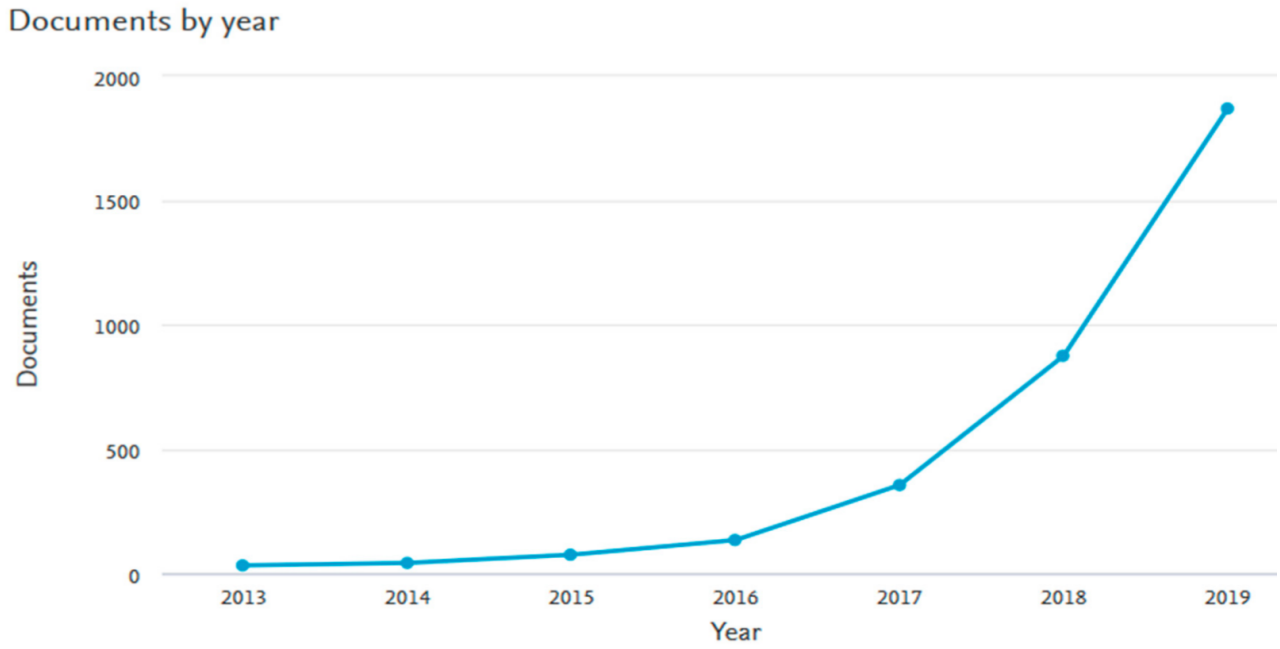

Fig.1.8 Applicazioni annuali Data Science in economia

Fonte: Nosratabadi, Saeed, Amirhosein Mosavi, Puhong Duan, Pedram Ghamisi, Ferdinand Filip, Shahab S. Band, Uwe Reuter, Joao Gama, and Amir H. Gandomi. "Data Science in Economics: Comprehensive Review of Advanced Machine Learning and Deep Learning Methods." Mathematics (Basel) 8, no. 10 (2020): 1-25.

I principali campi d'applicazione risultato essere:

[13] Nosratabadi, Saeed, Amirhosein Mosavi, Puhong Duan, Pedram Ghamisi, Ferdinand Filip, Shahab S. Band, Uwe Reuter, Joao Gama, and Amir H. Gandomi. "Data science in economics: comprehensive review of advanced machine learning and deep learning methods." *Mathematics* 8, no. 10 (2020): 1799.

- *Management*: la Data Science, estraendo informazioni innovative dalle enormi quantità di dati registrati nei database aziendali, permette alla direzione dell'azienda di prendere decisioni sempre più consapevoli sulle iniziative della stessa. La scienza dei dati rappresenta, quindi, il mezzo che permette ai manager di conoscere in modo più approfondito i risultati della propria azienda così da pianificare una strategia aziendale fondata su dati avanzati, favorendo il raggiungimento del vantaggio competitivo.
- *Supply chain management*: come già detto in precedenza, un grande numero di funzioni interdipendenti tra loro in un determinato processo fa sì che siano presenti una moltitudine di fonti di generazione dei dati, sarà quindi maggiore il contributo che la Data Science potrà fornire, in termini qualitativi, al processo stesso.

  Con "*supply chain management*" si definisce la gestione del flusso di valore dell'azienda incorporato nell'insieme delle attività fondamentali dell'impresa, dal reperimento delle materie prime, passando per la distribuzione del prodotto, fino ad arrivare ad attività di supporto come la gestione dei servizi post-vendita.

  In questo caso, in cui le singole attività devono essere viste in un'ottica di sistema oltre che individuale, l'analisi dei dati generati da ciascuna delle stesse lungo la supply chain può permettere di poter implementare migliorie organizzative e comunicative così da ridurre non solo i costi ma anche i tempi e, allo stesso tempo, migliorare la qualità del prodotto[14].

  Il primo step per l'applicazione della scienza dei dati è quello di massimizzare il valore generato da ogni singolo segmento della catena. Questo può avvenire, per esempio, nella ricerca delle materie prime, nella scelta dello stock di risorse nei magazzini, nella selezione della localizzazione degli impianti, nella manutenzione preventiva dei macchinari o ancora nelle decisioni che riguardano le modalità di distribuzione finali.

  Dopodiché, è fondamentale l'ottimizzazione dell'organizzazione e della comunicazione all'interno della supply chain in un'ottica sistemica.
- *Marketing*: come già enunciato nel paragrafo 1.2, un'altra fonte di generazione dei dati che ha acquisito sempre più valore e importanza negli ultimi due decenni è sicuramente quella dei social network. Questi dati rappresentano informazioni preziose in forma grezza da analizzare tramite metodi di Data Science al fine di studiare i comportamenti e i sentimenti dei clienti. Queste sono materie che fino a pochi anni fa non erano esaminabili su larga scala.

  Il marketing è un insieme di più processi, anche la scienza dei dati applicata a questo settore, di conseguenza, spazia molto trovando diverse tipologie di applicazione. Uno degli scopi

[14] Hung, Jui-Long, Wu He, and Jiancheng Shen. "Big data analytics for supply chain relationship in banking." *Industrial Marketing Management* 86 (2020): 144-153.

principale dell'analisi dei dati in questa area è sicuramente quello di anticipare la variazione, in termini di sentimenti, dei consumatori di fronte a cambiamenti nelle caratteristiche di determinati prodotti. Un'altra applicazione è quella di sistemi di riconoscimento facciale in grado di rilevare gli stimoli d'interesse di consumatori messi di fronte al lancio simulato di attività promozionali. Infine, un'ulteriore dimostrazione dell'utilità della Data Science nel market riguarda, per esempio, mappe di calore degli scaffali per il rilevamento dello scaffale vuoto o previsioni sull'abbandono dei clienti.

- *Fallimento aziendale*: prevedere il fallimento aziendale è una delle applicazioni della Data Science che, in un ambiente aziendale, potrebbero avere maggior peso.

  Per studiare un fenomeno come questo devono essere analizzati dati provenienti da fonti molto diverse tra loro poiché il fallimento di un'azienda può essere provocato da eventi che, in un primo momento, possono sembrare non correlati tra loro. Fondamentali a questo fine sono, quindi, i moderni algoritmi che riescono nell'impresa di riuscire non solo ad analizzare dati diversi tra loro, ma anche a trovare connessioni e relazioni tra essi, per stabilire soglie oltre le quali il recupero dell'azienda non è più possibile.
- *Mercato azionario e criptovalute*: l'obiettivo principale della Data Science applicata al mercato azionario è, analogamente a quello degli analisti del settore, quello di stimare i prezzi futuri delle azioni, al fine di ottenere un rendimento.

  La difficoltà non sta tanto nell'analizzare il rischio specifico, possibile anche attraverso i sistemi di analisi tradizionali, ma nello studiare il rischio sistematico, cioè associabile al mercato in generale.

  Per cercare di stimare questo rischio lo studio delle serie temporali finanziarie non è più sufficiente. È fondamentale che a questo venga associato lo studio del sentimento dei diversi soggetti e l'analisi dei dati provenienti da tutti quei fattori esterni che possono influire sui mercati. Altri obiettivi della Data Science in questo campo riguardano la gestione automatica e personalizzata del portafoglio, lo studio e la previsione dell'andamento degli indici di mercato, in primis dell'S&P 500, e dei fondi d'investimento.

  Come accade per la previsione del prezzo delle azioni, è possibile prevedere le tendenze riguardo l'andamento futuro delle criptovalute.
- *Banking*: se è vero che l'economia è la scienza che meglio tra tutte riesce ad integrarsi con la scienza dei dati, in forza della vastissima quantità di informazioni utili che possono essere ricavate tramite l'analisi degli stessi, a maggior ragione il settore degli intermediari creditizi, che per loro natura generare una mole enorme di dati, può essere fonte di valore.

D’altro canto però, l’utilizzo dei dati dei clienti per personalizzare i prodotti finanziari risulta rischioso in termini di sicurezza, visto che, ad esempio, una violazione degli archivi di una banca potrebbe far diventare di pubblico dominio delle informazioni personali dei clienti stessi.

2. LA DATA SCIENCE NELL’INTERMEDIAZIONE CREDITIZIA

2.1. Intermediazione creditizia e tecniche di Data Mining: il Credit Scoring

Per Credit Scoring si intende il processo attraverso il quale viene stabilita una valutazione quantitativa del rischio di credito associato ad un determinato soggetto, ossia il rischio associato al fatto che il debitore possa non essere in grado di assolvere ai suoi obblighi di pagamento. Viene stimata la c.d. “Probability of Default (d’ora in poi PD)”. Anderson per dare una definizione efficace di Credit Scoring, nel suo libro “*The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*”[15], ha scomposto il termine per poi studiare in modo indipendente i due elementi.

Analizzando “*credit*” è possibile recepire il fatto che si stia parlando di un obbligo in capo al beneficiario di una determinata prestazione tramite un pagamento che avverrà in un diverso momento. Il componente “*scoring*” fa riferimento ad un punteggio associato alla persona che potrà poi permettere di stilare una classifica o ancor meglio di suddividere i casi in categorie omogenee tra loro, le cosiddette “Classi di rating”.

Una definizione, questa, che a differenza della precedente non fa riferimento all’etimologia della parola, quanto più alla statistica come scienza alla base dei processi di credit scoring. “*Il credit scoring aiuta gli istituti di credito a valutare il rischio potenziale dei nuovi clienti e a valutare il comportamento futuro dei clienti esistenti, utilizzando modelli statistici per trasformare i dati rilevanti in misure numeriche che guidano il credito*.” (Abdou, H.A. and Pointon, J., 2011)[16].

Volgendo per un attimo lo sguardo al passato, i concetti di debito e di credito sono nati poco dopo, se non nello stesso momento, rispetto alla nascita del sistema finanziario cosiddetto “primitivo”.

[15] Anderson, R. (2007). *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation*. Oxford University Press.

[16] Abdou, Hussein A., and John Pointon. "Credit scoring, statistical techniques and evaluation criteria: a review of the literature." *Intelligent systems in accounting, finance and management* 18, no. 2-3 (2011): 59-88.

Questo prevedeva per la prima volta nella storia l'introduzione della moneta per velocizzare e semplificare gli scambi commerciali.

Per quanto riguarda il Credit Scoring, esso, così come tutte tecniche che si basano su una modellizzazione della realtà, non potrà che sorgere in concomitanza ai primi computer intorno agli anni '60 del 900.

Un altro balzo in avanti, in termini di diffusione di articoli a questo tema, è avvenuto agli inizi del ventunesimo secolo, sempre parallelamente alle innovazioni tecnologiche che l'inizio del nuovo millennio ha portato con sé.

In quest'ultimo periodo si è ampliata notevolmente anche la numerosità e la diversità dei campi in cui il credit scoring può essere applicato. Nonostante ciò, gli intermediari creditizi, e in particolare le banche, sono i protagonisti della sperimentazione di nuove tecniche, che, tramite l'applicazione delle innovative tecnologie di Data Science e machine learning hanno il potenziale per ottenere profitti smisurati.

Questo è possibile perché, attraverso la messa in pratica delle scienze per lo studio dei Big Data, si raggiungerà una rivoluzione dei modelli classici di credit scoring che si basavano unicamente sull'analisi della storia finanziaria del prenditore.

La Data Science prevede, infatti, l'inclusione nei modelli anche di variabili studiate in base ai dati provenienti da sorgenti innovative, come ad esempio i social network (approfondito più avanti).

Dal punto di vista degli intermediari finanziari, la possibilità di stimare, nelle modalità più precise possibili, questa tipologia di rischi rappresenta un mezzo fondamentale per stabilire quella che è la propria strategia di concessione del credito.

Il comitato di Basilea[17] ha introdotto e regolamentato la disciplina prudenziale per quanto riguarda gli intermediari creditizi. Questo processo è avvenuto con l'introduzione dei gruppi di regole che prendono il nome del comitato stesso (BIS I, BIS II, BIS III e BIS IV).

La disciplina prudenziale prevede che, in contrapposizione a rischi di credito assunti a seguito della concessione del credito stesso, ci debba essere un requisito minimo di capitale quantificato in base alla all'importo dell'erogazione ponderato per la probabilità di insolvenza della controparte.

Il requisito minimo di capitale previsto dalla normativa, rappresenta, quindi, il "prezzo" che la banca è costretta a pagare nel momento in cui stabilisce una determinata concessione di credito.

[17] Comitato per la vigilanza bancaria, spesso denomitato CB. Fu creato nel 1974 dalle banche centrali appartenenti al Gruppo dei 10 (G 10; ➔ G 20), a seguito della crisi determinata dal fallimento della banca tedesca Herstatt. Tale episodio, che metteva in evidenza i rischi di contagio derivanti dai rapporti di credito e debito fra banche attive a livello internazionale, portò alla creazione del CB come forum permanente di coordinamento delle politiche di vigilanza prudenziale sulle banche, Treccani, https://www.treccani.it/enciclopedia/comitato-di-basilea_%28Dizionario-di-Economia-e-Finanza%29/

Questo fattore lega inevitabilmente la strategia di credito degli intermediari alla disciplina prudenziale.

Nel momento in cui si verifica un peggioramento del merito di credito, la banca dovrà non solo aumentare il capitale per fronteggiare il rischio, assimilabile ad un aumento dei "costi" dovuti a tenere fermo una maggior quantità di capitale, ma anche sostenere maggiori spese per aumentare la frequenza di monitoraggio del credito stesso.

Una stima più efficiente permette di evitare di essere presi alla sprovvista da costi inaspettati.

Un adeguato modello di Credit Scoring può rappresentare un importante mezzo di analisi e valutazione che, utilizzato in modo efficace dal risk management, può permettere una corretta stima del rischio di credito e quindi l'elaborazione della migliore strategia possibile. Garg et al. (2017) espongono quelli che sono i 3 fattori determinanti che spianano la strada ad un'inclusione della Data Science tra le discipline aziendale, in particolar modo delle banche.

Il primo fattore, già sottolineato in precedenza, è il progresso tecnologico recente. Questo ha fatto sì che la quantità di dati generati, oltre che le fonti dei dati stessi, aumentassero in modo esponenziale negli ultimi anni. In parallelo a questa crescita, anche le aziende hanno investito una porzione di budget sempre maggiore per raggiungere la migliore potenza di calcolo possibile, così da essere in grado di sviluppare modelli e algoritmi più efficaci.

Secondo elemento, è il fatto che le aziende operano in una situazione di forte pressione economica, il 54% dei primi istituti al mondo ha un prezzo inferiore al valore di mercato[18]. Questo significa che ogni possibilità di innalzare il profitto, o migliorare le proprie strategie può essere un'enorme stimolo d'investimento per questi istituti.

Terzo, e ultimo, fattore riguarda gli sforzi di digitalizzazione delle banche effettuate negli ultimi anni per poter ampliare e migliorare i propri database. Questo trend può essere sfruttato per introdurre sistemi di machine learning all'interno degli stessi intermediari creditizi.

In sintesi, le banche possono cogliere i frutti di una rivoluzione tecnologica ancora incompleta, investendo nelle infrastrutture e nelle sperimentazioni, rendendo la Data Science una vera e propria disciplina aziendale.

La più grande obiezione delle banche che ancora non hanno investito nella scienza dei dati è che essa è ancora troppo teorica. Secondo questi dirigenti, con le tecnologie attuali ottenere profitti considerevoli è ancora qualcosa di utopico.

[18] Garg, Amit, Davide Grande, Gloria Macias-Lizaso Miranda, Christoph Sporleder, and Eckart Windhagen. "Analytics in banking: Time to realize the value." *Режим достуna: http://www. mckinsey. com/industries/financial-services/our-insights/analytics-in-banking-time-to-realize-the-value/(дата обращения: 18.03. 2018)* (2017).

Questa è, però, una visione poco lungimirante. Non possono essere considerati solo i benefici che gli investimenti garantiscono oggi in questo settore, ma vanno valutati anche quelli che potrebbero essere apportati in un futuro prossimo. Basti ricordare la rivoluzione e la ricchezza che il marketing ha generato per le aziende nel momento in cui è stato identificato ufficialmente come disciplina aziendale.

Per raggiungere il suo scopo, la Data Science fa riferimento alla Data Mining, ossia al processo che permette di ottenere informazioni utili da modelli che si alimentano tramite grandi cumuli di dati grezzi.

Il procedimento più in generale, e che quindi incorpora al suo interno le tecniche di Data Mining, è detto "Knowledge Discovery from Data" (d'ora in poi KDD)[19].

Il KDD prevede 5 fasi che sono:

- *Selection*: creazione di un dataset sulla base delle conoscenze che si possiedono e sulla tipologia di informazioni che si vogliono ottenere. È, inoltre, fondamentale la selezione delle variabili che si andranno a considerare.
- *Pre-processing*: Riduzione della quantità di dati a disposizione ed eliminazione dei dati considerati scorretti o fuorvianti.

  L'aumento della velocità di produzione di dati negli ultimi anni, se da un lato ha permesso agli analisti di avere più materia prima, dall'altro ha fatto sì che gli stessi incontrassero diverse difficoltà ad analizzare i dati in modo efficace, perché troppo numerosi.

  La soluzione è quella di affidare questa fase all'automazione. Tra le caratteristiche che rendono questo step idoneo ad essere automaticamente eseguito vi è il fatto che le operazioni da svolgere sono molto meccaniche e che i problemi possono essere scomposti in sotto-problemi.
- *Feature selection*: spesso i dataset possiedono un gran numero di informazioni non funzionali al tipo di modello che si sta elaborando. È fondamentale quindi stabilire quali attributi sono necessari per raggiungere l'obiettivo che ci si è posti in principio e quali sono i dati che possono essere omessi.
- *Data mining*: utilizzo di algoritmi (quali associazione, clustering, predizione e classificazione che verranno analizzate nello specifico in seguito) con il compito di individuare le informazioni target dai dataset.
- *Interpretation/Evaluation*: studio delle informazioni ottenute come risultato dalla fase di Data Mining. A differenza degli altri step, che possono essere automatizzati e svolti dalle macchine,

[19] Fayyad, Usama M., David Haussler, and Paul E. Stolorz. "KDD for Science Data Analysis: Issues and Examples." In *KDD*, pp. 50-56. 1996.

in questo processo sono fondamentali le capacità di sintesi, analisi critica e formulazione di ipotesi. Se il risultato non soddisfa le richieste, sarà necessario svolgere nuovamente il KDD dall'inizio.

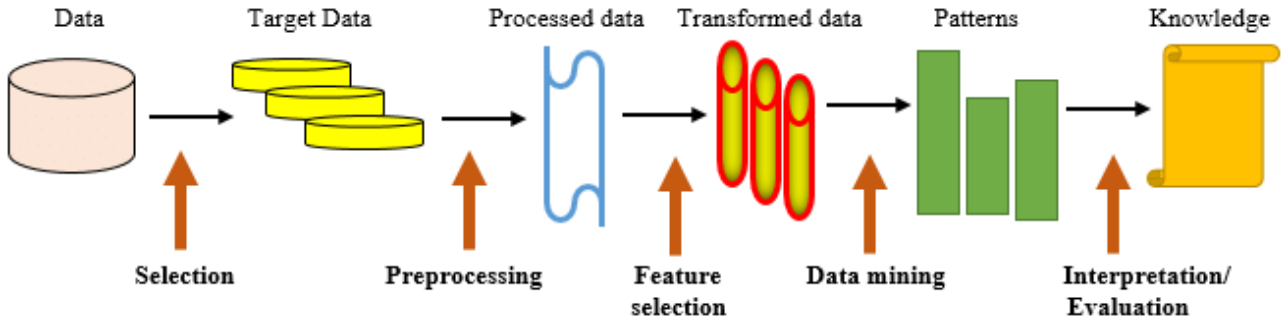

Figura 2.1: I 5 processi del KDD

Fonte: Fayyad, Usama M., David Haussler, and Paul E. Stolorz. "KDD for Science Data Analysis: Issues and Examples." In KDD, pp. 50-56. 1996.

2.2. Le principali applicazioni della Data Science nell'intermediazione finanziaria: fattori critici di successo

La Data Science, per ottenere risultati qualitativamente significativi, necessita di una grande quantità di materia prima, cioè di dati.

È per questo motivo che la scienza dello studio dei dati trova nell'intermediazione finanziaria terreno fertile, vista l'enorme quantità di dati che questa genera.

Tra i temi e le discussioni che riguardano l'applicazione della Data Science, in ambito bancario e non solo, andremo ad analizzare quelli più degni di nota: la sicurezza e il rilevamento delle frodi, la regolamentazione della tutela dei dati in ambito di Credit Scoring e i modelli di Credit Scoring che sfruttano i dati provenienti dai social network.

L'implementazione delle nuove tecnologie ha fatto sì che in parallelo ai benefici, descritti precedentemente, si evolvessero anche comportamenti negativi, quali le frodi.

Secondo lo studio Javelin (2007)[20], nello stesso anno sono stati ben 6,8 milioni i cittadini truffati tramite carte di credito, a cui si aggiungono le frodi contabili, assicurative e molte altre.

L'individuazione di attività illecite nell'enorme flusso di dati a disposizione delle banche è sempre più difficile e richiede lo sviluppo di nuovi modelli per tutelare i clienti. Questo è dovuto all'evoluzione digitale che ha permesso un miglioramento delle tipologie di attacchi, non più fronteggiabili con gli strumenti antifrode comuni, divenuti ormai obsoleti.

[20] Strategy, Javelin. "Research; 2007 Identity Fraud Survey Report—Consumer Version How Consumers Can Protect Themselves."

Le nuove attività illecite possono, invece, essere contrastate tramite tecniche di Data Science e Data Mining che si pongono come scopo quello di analizzare schemi insoliti che potrebbero condurre ad azioni illegali.
Le tipologie di modelli di Data Mining sono classificati in[21]:

- Analisi esplorativa: analisi globale di un dataset il cui obiettivo è individuarne i caratteri principali.
- Modellazione descrittiva: schematizza e stabilisce relazione tra i dati.
- Modellazione predittiva: sfrutta le conoscenze estratte dai dati per generare previsioni future

Tra le tecniche di Data Science e Data Mining, le principali e più ampiamente utilizzate in diversi campi sono: le procedure di classificazione, clustering (raggruppamento), associazione e predizione.
La classificazione ha come obiettivo quello di etichettare e suddividere i dati raccolti in classi, talvolta stabilendo anche con che percentuale il dato stesso è associabile ad una determinata categoria.
Con il processo definito clustering (raggruppamento) i dati vengono raggruppati in insiemi in base a caratteristiche che, a differenza della classificazione, non si conoscono a priori. Questa tecnica è ampiamente diffusa in tutti i settori per effettuare il cosiddetto cross-selling, una strategia di vendita che si basa sul fatto che ad ogni gruppo di clienti verrà somministrata un'offerta diversa, che rispecchia i gusti degli stessi.
La regola di associazione è il metodo secondo cui si cercano di stabilire relazioni e connessioni tra dati diversi, tra loro secondo una logica consequenziale del tipo "se x allora y".
Il processo di predizione permette di raggiungere il risultato che stiamo perseguendo fin dalla prima fase, ossia la possibilità di ottenere una stima di un valore futuro partendo dalle variabili analizzate precedentemente.
Con l'implementazione e l'applicazione di quest'ultima tecnica nel settore bancario si raggiunge lo scopo di sviluppare un modello che, utilizzando tutti i processi analizzati in precedenza, può permettere la prevenzione di truffe e frodi.

[21] John, Samuel Ndueso, C. Anele, O. Okokpujie Kennedy, F. Olajide, and Chinyere Grace Kennedy. "Realtime fraud detection in the banking sector using data mining techniques/algorithm." In *2016 international conference on computational science and computational intelligence (CSCI)*, pp. 1186-1191. IEEE, 2016.

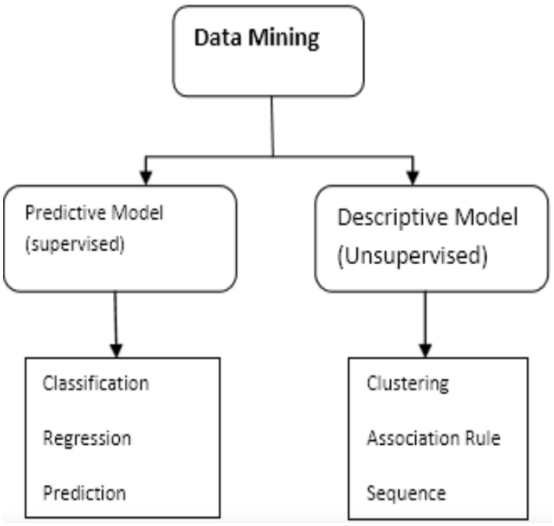

Figura 2.2 Tecniche e modelli di Data Mining

Fonte: John, Samuel Ndueso, C. Anele, O. Okokpujie Kennedy, F. Olajide, and Chinyere Grace Kennedy. "Realtime fraud detection in the banking sector using data mining techniques/algorithm." In 2016 international conference on computational science and computation.

Modelli di Data Mining e processi di riconoscimento dovranno essere bilanciati, perché se da un lato permetteranno al cliente di svolgere operazioni sui propri risparmi con tutta la sicurezza del caso, allo stesso tempo non devono eccedere con richieste di autenticazione troppo lunghe e/o complesse, anche se in molti casi queste sono direttamente proporzionali alla sicurezza del cliente stesso.

Una soluzione potrebbe essere quella di integrare gli schemi sviluppati con le tecniche di Data Mining con un aumento delle fasi di login, in numero e difficoltà, solo nel momento in cui si rilevino dati sospetti o pericolosi.

L'inclusione finanziaria, definita come il diritto di ciascun cittadino a poter accedere ai servizi bancari, è uno dei principali obiettivi che la sinergia tra startup Fintech e legislazione stanno tentando di raggiungere.

Recentemente la disciplina con l'inserimento di leggi antidiscriminatorie ha stabilito che alcune variabili e informazioni riguardanti il possibile prenditore, non potessero essere inserite nei modelli di Credit Scoring perché potenzialmente lesive agli interessi delle categorie protette.

La principale conseguenza di una decisione come questa è stata un peggioramento della precisione e dell'efficienza dei modelli di scoring stessi, causato da una diminuzione delle informazioni analizzabili. Questo è dovuto al fatto che i modelli di credit scoring, basati su tecniche statistiche e matematiche, si nutrono delle informazioni raccolte. Maggiore è la quantità delle informazioni che vengono prese in considerazione e migliore sarà l'analisi sotto il punto di vista di affidabilità creditizia della controparte.

L'attività legislativa, inoltre, portare danni enormi alle banche. Con la recente normativa del Comitato di Basilea, che prevede un capitale minimo in contrapposizione alla concessione di prestiti, una cattiva analisi dei rischi di insolvenza del prenditore può provocare un aumento dei rischi e quindi una maggior quantità di capitale paralizzato per far fronte agli stessi.

Alla base di questi provvedimenti, esiste il concetto che rimuovere le variabili che potevano andare a "riconoscere" l'appartenenza dell'individuo ad una determinata minoranza, poteva essere la soluzione migliore per preservare i gruppi protetti di cittadini, come definito dal documento "Guide to Credit" redatto nel Regno Unito.

L'inefficacia di queste leggi antidiscriminatorie è stata, però, accertata da uno studio statunitense realizzato da Munnell et al. (1996)[22], questo ha dimostrato che anche nel momento in cui venivano omesse le variabili correlate a caratteristiche che potevano ricondurre ad una determinata minoranza, gli individui appartenenti alle stesse avevano meno della metà delle possibilità di ottenere un finanziamento.

Come definito da Capon (1982)[23], questo era dovuto alle cosiddette variabili "surrogate". Esse erano, dal punto di vista legale, accettate singolarmente, ma combinate tra loro producevano come risultato quello di una "discriminazione indiretta".

È ora necessario esaminare tre fattori che mirano a sottolineare che l'includere tutte le variabili nei modelli di Credit Scoring non rappresenta un problema per la tutela dell'inclusione finanziaria di minoranze[24].

Una delle caratteristiche alla base dei modelli di scoring è il loro essere associativi, perciò è impossibile prevedere la connessione tra l'inclusione di una determinata variabile e lo score finale, dato che il modello non può affermare nessi di causalità.

Il secondo fattore riguarda il fatto che i modelli di Credit Scoring permettono di ottenere una stima, che per definizione non potrà mai raggiungere una precisione totale.

In ultimo, è necessario evidenziare che i modelli di Credit Scoring non rappresentano la regola a cui gli operatori finanziari devono obbligatoriamente attenersi, ma solo uno strumento che possono utilizzare durante la redazione della strategia di credito della banca.

Come affermato ripetutamente, la quantità di dati prodotti e salvati in rete in questo momento storico e tecnologico è enorme, così come si è ampliata la quantità di sorgenti che possono generare i dati stessi. Tra queste fonti, quella che sicuramente raccoglie informazioni, per di più fortemente eterogenee, è costituita dai social network.

---

[22]Munnell, Alicia H., Geoffrey MB Tootell, Lynn E. Browne, and James McEneaney. "Mortgage lending in Boston: Interpreting HMDA data." *The American Economic Review* (1996): 25-53.

[23] Capon, Noel. "Credit scoring systems: A critical analysis." *Journal of Marketing* 46, no. 2 (1982): 82-91.

[24] Chan, Wen Li, and Hsin-Vonn Seow. "Legally scored." *Journal of Financial Regulation and Compliance* (2013).

Se uno dei problemi principali a cui la Fintech si trova a far fronte è quello dell'inclusione finanziaria, la Data Science è lo strumento tramite il quale essa può realizzare quest'obiettivo.

I modelli di Credit Scoring classici raccolgono al loro interno le variabili che riguardano sia le informazioni anagrafiche sia quelle riguardanti la storia finanziaria del possibile prenditore. In base allo "score", ossia al punteggio stabilito dall'analisi, verrà deciso non solo se emettere o meno il credito, ma anche a quale tasso (TAEG).

Nel momento in cui una qualsiasi persona, che alle sue spalle non ha una storia finanziaria affidabile, faccia richiesta per il credito vedrà molto probabilmente negata la sua richiesta.

Difficoltà nel ricevere dei finanziamenti, per esempio per poter acquistare casa o aprire un'attività, rappresenta una mancata occasione di inclusione nella società.

È proprio a questo scopo che diverse società hanno elaborato modelli di Credit Scoring che basano le proprie analisi unicamente dalle variabili ottenute dai dati raccolti sui profili dei social network per valutare il merito creditizio dei consumatori[25].

Questi nuovi meccanismi rappresentano una svolta in tema di inclusione finanziaria, molte più persone, anche con redditi bassi, avranno l'opportunità di ricevere finanziamenti che con i modelli di Credit Scoring classici non avrebbero mai potuto ottenere.

3. Il caso Modefinance

Modefinance è una società italiana che ha fatto dell'applicazione di tecnologie di analisi dei dati in ambito di intermediazione creditizia il suo pilastro fondante.

Figura 3.1 Logo Modefinance

Fonte https://www.modefinance.com/it

Modefinance è un'azienda Fintech fondata nel 2009 a Trieste, da Mattia Ciprian e Valentino Pediroda, con l'obiettivo di fornire valutazioni di rating oggettive tramite l'utilizzo e la ricerca di soluzioni di IA.

Dal 2015 è autorizzata dall'European Securities and Markets Authority ad operare, per prima nel mondo accademico, come credit rating agency.

[25] Wei, Yanhao, Pinar Yildirim, Christophe Van den Bulte, and Chrysanthos Dellarocas. "Credit scoring with social network data." *Marketing Science* 35, no. 2 (2016): 234-258.

Un anno dopo diventa l'unica Agenzia di Rating italiana certificata per la valutazione delle banche e poi, grazie alla certificazione ECAI (External Credit Assessment Institution), la prima Fintech Rating Agency certificata a livello europeo.

Tra i temi più controversi della crisi dei mutui subprime, che ha influenzato negativamente l'economia mondiale, ricorre la responsabilità delle agenzie di rating per le alte valutazioni dei titoli tossici in quegli anni. Modefinance nasce, infatti, proprio a seguito della crisi finanziaria ponendosi come obiettivo quello di fornire delle valutazioni del rischio di credito attraverso algoritmi chiari e verificabili, cosicché gli investitori e l'intero mondo finanziario possano ritrovare la fiducia nelle agenzie di rating persa negli ultimi decenni.

Quest'obiettivo viene perseguito tramite la coniugazione tra metodologie di Data Science, AI e processi di credit scoring.

L'approccio interdisciplinare, che rappresenta una coraggiosa innovazione nel panorama italiano, è il punto di forza di questa società che le ha permesso non solo di diventare una tra le più innovative e importanti aziende Fintech in Italia, ma anche di farsi conoscere oltre i confini del Paese.

Figura 3.2 Integrazione tra Data Science, AI e tecnologia cloud

Fonte https://www.modefinance.com/en/solutions/credit-rating-agency-modefinance

L'indipendenza e il rigore con il quale vengono elaborate le valutazioni del credito, sono anche dimostrate dal fatto che Crowfundme, società che si occupa di predisporre raccolte fondi tramite il

pubblico, ha scelto proprio Modefinance per la valutazione delle start-up Fintech che richiedono questa fonte di finanziamento.

Oltre al caso appena esplicato, sono molti e profondamente differenti tra loro i clienti che fanno affidamento su Modefinance, dalle banche, passando per le società di gestione del risparmio (SGR), fino ad arrivare alle imprese piccole, medie e alle multinazionali.

Tra i propri clienti conta, infatti, oltre 65 mila banche e 250 milioni di imprese tra cui: Burgo, Piaggio, Levoni, Finanziaria Internazionale, Samsung, Banca Valsabbina, Banca Progetto, ELITE, Azimut Direct, Fabrick, Anthilia SGR, Lindt, CrowdFundMe, Metsa Group, e molti altri.

In un contesto storico di ampio sfruttamento dei dati sia pubblici che privati, il mercato si è orientato nell'ottica dell'Explainable AI. Con Explainable AI si intendono tutti quegli schemi e strumenti che permettono all'essere umano di comprendere i processi svolti da tecnologie di intelligenza artificiale, così da capirne non solo il funzionamento, ma anche di prevederne, seppur in modo approssimativo, il risultato. Questo concetto insieme ad una regolamentazione riguardante la privacy (come sottolineato più volte in precedenza) aggiornata ed efficiente, permette di elevare la fiducia dei cittadini nei confronti di tecnologie innovative e in continua evoluzione.

In quest'ottica, si posiziona perfettamente Modefinance che proprio tra i principi fondanti prevede chiarezza, trasparenza, indipendenza, rispetto della regolamentazione ed un'adeguata informazione.

Il principale processo di valutazione del rischio di credito di Modefinance è MORE (Multi Objective Rating Evaluation).

Il modello è stato sviluppato nel 2011 tramite l'utilizzo della Data Science, Machine Learning e IA. Esso si avvale di reti neurali artificiali[26] ed è specializzato nell'analisi delle capacità di una società di far fronte ai propri impegni finanziari tramite le risorse generate.

L'accuratezza dell'algoritmo è garantita da diversi fattori, come il fatto che le informazioni vengono prese in considerazione solo se corrette e aggiornate. La quantità di variabili contenute all'interno dell'algoritmo sono collocate nell'ordine delle decine di migliaia, e, inoltre, la combinazione tra standardizzazione e flessibilità fa sì che possa essere applicato ad aziende e banche di tutto il mondo.

[26] Modello matematico delle funzioni cerebrali, costruito a partire da semplici unità elementari di calcolo, che sono a volte chiamate neuroni, per analogia con le loro controparti biologiche, Treccani, <https://www.treccani.it/enciclopedia/rete-neurale-artificiale_%28Enciclopedia-della-Scienza-e-della-Tecnica%29/>

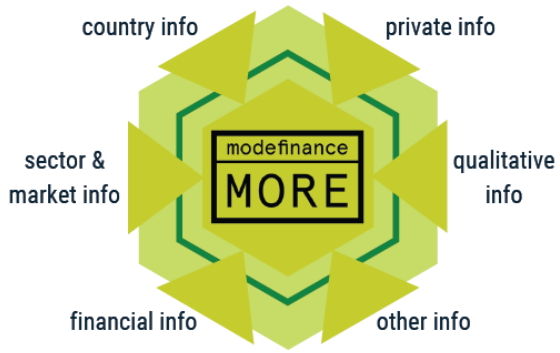

Figura 3.3 Informazioni considerate da MORE

Fonte https://www.modefinance.com/it/tecnologie/more

La stabilità dell'azienda/banca analizzata è direttamente proporzionale all'equilibrio tra le seguenti aree, che caratterizzano la stessa sia sotto il profilo economico che finanziario.

Figura 3.4 Aree di interesse

Fonte https://www.modefinance.com/it/tecnologie/more

Il processo che viene seguito, prima di arrivare all'output finale del modello, si articola in 4 fasi principali:

- Definizione e scelta degli indici: la scelta degli indici deve avvenire tenendo conto del fatto che devono essere indicativi del rischio di default e rappresentativi della situazione economico-finanziaria dell'azienda/banca.

- Estrazione informazioni qualitative dai dati: tramite la Fuzzy logic[27], vengono trasformati i valori provenienti dagli indici in informazioni di tipo qualitativo.
- Equilibrio economico-finanziario: vengono preferite le aziende che possiedono indici più in equilibrio tra loro, questi rifletteranno situazioni più stabili sotto il profilo economico-finanziario.
- MORE score: associazione di una classe di rischio alla società, rappresenta un giudizio sintetico sulla salute della stessa.

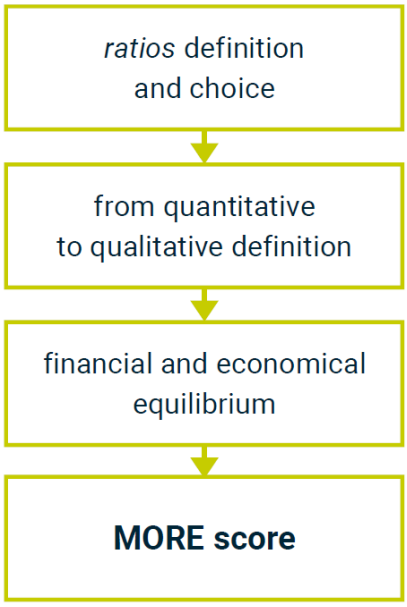

Figura 3.5

Fonte https://www.modefinance.com/it/tecnologie/more

Il modello MORE, e quindi la classe di rischio associata all'azienda, rappresenta lo strumento principale tramite il quale gli analisti di Modefinance potranno redigere le valutazioni di rating complete.

[27] fuzzy logic (ingl., letteralmente: «logica sfumata» o «logica sfocata») tipo di logica polivalente, cioè che, a differenza di quella classica (aristotelica o booleana), è in grado di trattare contesti ambigui, imprecisi, non esattamente definiti, Treccani, < https://www.treccani.it/enciclopedia/fuzzy-logic_%28Enciclopedia-della-Matematica%29/#:~:text=fuzzy%20logic%20(ingl.%2C%20letteralmente,%2C%20imprecisi%2C%20non%20esattamente%20definiti.>

MORE score	Macro classi di score	Giudizio
AAA	Sana	La solvibiltà dell'impresa è ritenuta massima.
AA		L'impresa ha una solvibilità molto alta
A		L'impresa ha una solviblità alta
BBB	Equilibrata	L'equilibrio patrimoniale, finanziario ed eocnomico è considerato adeguato
BB		La performance dell'impresa è adeguata considerando il settore e il paese nei quali opera
B	Vulnerabile	L'impresa manifesta segnali di elevata vulnerabilità
CCC		L'impresa mpostra squilibri nella sua struttura patrimoniale, finanziaria ed economica
CC	Rischiosa	L'impresa mostra dei segnali di elevata vulnerabilità
C		L'impresa manifesta situaizoni patologiche considerevoli
D		L'impresa manifesta situazioni patologiche gravi

Figura 3.6 Classi di rischio

Fonte https://www.modefinance.com/it/tecnologie/more

Tramite la testimonianza di Valentino Pediroda, CEO e co-founder di Modefinance, nonché docente di Metodi Numerici all'Università di Trieste, è possibile aggiungere valore alle considerazioni di cui si è discusso fino a questo momento.

1) Al momento della fondazione della società, nel 2009, lo studio dei dati era ancora un campo poco esplorato, è stato difficile, quindi, ottenere la fiducia dei primi finanziatori?

"Effettivamente sì. Essere, sotto alcuni punti di vista, i precursori, non è mai semplice. Il fatto che Modefinance abbia indicato la strada che, col tempo si è rivelata vincente, ma che di fronte alla quale inizialmente non tutti erano pronti, si è rivelata una missione davvero complessa."

2) Quali erano i bisogni che mirava a soddisfare Modefinance nel momento in cui è stata fondata? Sono cambiati al giorno d'oggi?

"Nel momento in cui siamo entrati in questo business, l'obiettivo era quello di creare uno scoring automatizzato che fosse il più possibile globalizzato per tutte le aziende. In realtà, lo schema iniziale si è rivelato col tempo il mattoncino su cui abbiamo costruito quelle che oggi sono le nostre tecnologie. Quel mattoncino ha permesso alla nostra società di creare il prodotto Tigran, la piattaforma in cui le aziende e gli istituti finanziari possono automatizzare i propri processi di valutazione. Tutto è quindi iniziato con un'idea, anche se poi la sfida maggiore è stata realizzare l'infrastruttura e rendere il progetto funzionante e funzionale ai clienti."

3) Per quanto riguarda il processo MORE, tra le varie informazioni che vengono acquisite per elaborare quello che poi è il rating, ci sono le "private info". Le chiedo, quindi, quanto è importante per la vostra società l'attenzione alla privacy?

"È fondamentale. Lavorando anche con istituti finanziari è chiaro che il concetto di privacy e la salvaguardia del dato, e quindi del cliente, siano imprescindibili."

4) Se da un lato la produzione di dati in continua crescita è sinonimo di ricchezza di materie prime, dall'altro può essere un problema per i database che devono essere man mano adeguati alle nuove esigenze, la società come si sta muovendo sotto questo punto di vista?

"Modefinance è consumatore di dati più che creatore. La direzione verso cui si sta andando è una massificazione del concetto di API[28], partendo dai dati singoli per poi integrarli a proprio favore. La difficoltà sta più nell'organizzazione del tutto che nell'adeguamento delle tecnologie. Il fatto che ci si trovi di fronte ad API molto modulari, fa sì che si possa aumentare l'informazione della singola azienda con dati innovativi. Ricapitolando, quello dei database non è, quindi, un problema così complesso, perché ci si approccia a modelli modulari di API."

5) In relazione al tema della fiducia, quanto è importante che tutti i processi svolti dall'intelligenza artificiale avvengano alla luce del sole e siano ben spiegati?

"Fondamentale, il tema dell'Explainable Artificial Intelligence soprattutto nel campo della valutazione del rischio è imprescindibile."

6) Per quanto riguarda gli indici ESG (Environmental Social Governance), la società si è adeguata fornendo anche questa tipologia di valutazioni. Come sappiamo, il problema nel calcolo di questo fattore è legato ai criteri da prendere in considerazione, poiché possono variare da un'agenzia di rating all'altra. L'uso della data science e dell'intelligenza artificiale può aiutare nel rendere questo indice più oggettivo?

"Modefinance se ne sta occupando all'interno del progetto dell'Unione Europea "TranspArEEnS", su cui si lavora per standardizzare le modalità con cui vengono calcolati gli indici ESG.
Il problema non riguarda tanto il benchmarking, quanto più le difficoltà nel recupero dei dati.
Ad oggi per redigere in modo efficace il rating ESG di un'azienda, è indispensabile essere a stretto contatto con la stessa.
La società Modefinance sta lavorando per automatizzare il processo di redazione del rating ESG. Al giorno d'oggi una stima può essere realizzata utilizzando gli indici pubblici, come quelli associati al

[28]API (Application programming interface). – In informatica, regole e specifiche per la comunicazione tra software. Tali regole fungono da interfaccia tra i vari software e ne facilitano l'interazione, allo stesso modo in cui l'interfaccia utente facilita l'interazione tra uomo e computer, Treccani, <https://www.treccani.it/enciclopedia/api_%28Lessico-del-XXI-Secolo%29/#:~:text=s.%20f.%20invar.,interazione%20tra%20uomo%20e%20computer.>

rischio climatico, ma sul rating ESG vero e proprio il coinvolgimento dell'azienda è ancora fondamentale."

7) Essendo la data science una disciplina molto giovane, sono molti i vantaggi che devono essere ancora sfruttati. Quali di questi ci si aspetta che possano essere scoperti nel breve termine?

"Nel breve termine si possono ottenere vantaggi che riguardano principalmente i dati non supervisionati.
Ad oggi, nel nostro mondo, si sfrutta l'intelligenza artificiale per recuperare un target, per esempio la probabilità di default PD oppure la probabilità che un'azienda possa essere acquisita, partendo da una grande mole di dati. Quello che può essere il prossimo passo della Data Science riguarda l'auto creazione dei target, stabilendo automaticamente delle relazioni che possano spiegare determinate sinergie.
Il problema riguardo queste nuove implementazioni dell'analisi dei dati non è tanto la potenzialità numerica e scientifica, che è concreta, ma nel cercare di capire se il mercato è pronto per determinate scoperte.
Questo perché il mercato è data driven, si muove in base ai dati raccolti. Bisogna quindi capire se esso, e le applicazioni di questi innovativi processi, siano pronti per una tale innovazione mentale, prima che pratica."

8) Perché un cliente dovrebbe rivolgersi ad un'agenzia di rating Fintech come Modefinance piuttosto che ad una che utilizzi algoritmi di rating classici?

"Mi accorgo che sempre di più non è fondamentale avere solo gli algoritmi e i mezzi tecnici, ma è necessaria la messa in opera di queste tecnologie, e questo è molto complesso. Un'azienda come Modefinance possiede tre anime: da una parte è una rating agency, dall'altra una fintech e dall'altra ancora sviluppa in prima persona le tecnologie. È l'unione tra queste tre identità che permette di ottenere il vantaggio competitivo che possiede oggi. Questo è evidente dal fatto che Modefinance riesce non solo a dare una visione globale delle aziende, ma farlo integrando piattaforme IT, riuscendo così ad adattarsi a quelle che sono le esigenze procedurali del cliente."

9) Secondo lei quale può essere il contributo pratico che la figura del data scientist può apportare all'interno di una qualsiasi società che opera nel mondo del business?

"I data scientist sono fondamentali all'interno delle società perché senza l'aiuto dell'analisi dei dati nel 2022 non è più possibile essere sul mercato.

D’altro canto quello che è fondamentale è avere a che fare con data scientist preparati, il grande vantaggio si ha quando queste figure possiedono una profonda conoscenza tecnica della materia, non solo degli algoritmi di base.”

Dalla testimonianza di Pediroda, è possibile evincere l’importanza di alcuni temi già trattati nel corso dell’elaborato.

La società Modefinance ha avuto difficoltà iniziali nel reperire finanziatori essendo una delle prime realtà ad investire realmente nella Data Science, ma questo “rischio” è stato ripagato sul piano economico da un costante aumento del fatturato nel corso degli anni. Modefinance è la dimostrazione di come l’investimento nel medio-lungo termine nella scienza dell’analisi dei dati apporterà senz’altro benefici alle società che ne sapranno cogliere il valore.

Il CEO di Modefinance ha, inoltre, confermato l’importanza del tema della privacy, in un momento storico in cui la diffusione dei dati, personali e non, avviene con molta frequenza e facilità. Egli ha ribadito che, operando con tecnologie innovative e per niente banali, sia necessario prestare molta attenzione ad un tema fondamentale come quello della fiducia. A tal fine, sta assumendo sempre di più un ruolo centrale l’Explainable Artificial Intelligence, ossia i modelli che permettono di spiegare agli utenti i processi svolti dall’AI e i risultati ottenuti dalla stessa.

Un altro argomento trattato da Pediroda è stato quello del collegamento tra Data Science e un tema molto diffuso al giorno d’oggi tra le società Fintech e non solo: il calcolo degli indici ESG. L’azienda triestina oltre che collaborare nel progetto dell’Unione Europea “TranspArEEnS”, sta investendo risorse e tempo per raggiungere l’automatizzazione del calcolo di questo indice e l’indipendenza dalle aziende, oggi ancora indispensabile per l’ottenimento dei dati necessari a fare ciò.

Il co-fondatore di Modefinance, con l’ultima risposta all’intervista, ha evidenziato il fatto che, in questo periodo storico per poter essere competitivi sul mercato ogni società debba acquisire la figura del data scientist. L’investimento nella scienza dei dati di cui si è parlato in precedenza deve avvenire anche, se non soprattutto, nella selezione del personale più preparato e nella formazione dello stesso.

Pediroda, infine, anticipa quello che potrebbe essere uno sviluppo della Data Science nel prossimo futuro: l’auto creazione dei target all’interno di dataset, stabilendo senza la necessità di interventi umani le relazioni tra dati anche molto diversi tra loro. Questa evoluzione potrà diventare realtà solo dopo aver verificato se un mercato data driven, come quello odierno, sia pronto ad un’innovazione di tanto dirompente.

CONCLUSIONI

Lo studio si è proposto come obiettivo quello di analizzare un argomento innovativo come la scienza dell'analisi dei dati, che può essere paragonata, per portata dell'impatto nella società, alla diffusione di Internet negli anni '90. La "rivoluzione dei dati" cambierà nel breve termine le modalità secondo cui funzionerà l'economia, l'organizzazione delle aziende e quindi i meccanismi intrinsechi della società nel suo complesso.

L'analisi, dopo aver enunciato il significato di Big Data e Data Science, ha cercato di spiegare, sia qualitativamente che quantitativamente, le cause del perché negli ultimi 15 anni questo fenomeno ha acquisito sempre più rilevanza. Questo è avvenuto principalmente per due fattori, entrambi conseguenza delle innovazioni tecnologiche digitali che hanno travolto la società: l'aumento dei dati prodotti nelle ultime due decadi, dovuto principalmente alla diffusione del fenomeno dell'IoT, e lo sviluppo di strumenti di calcolo sempre più potenti.

È stata, poi, descritta quella che è la principale caratteristica della Data Science: l'interdisciplinarietà. Interdisciplinarietà che ha una duplice interpretazione, sia dal punto di vista della composizione di questa scienza, ossia che si basa dalla collaborazione tra matematica, fisica e statistica, sia dal fatto che i processi di analisi dei dati hanno un numero infinito di applicazioni, in tutti i settori del mercato. La Data Science, tramite l'utilizzo di tecniche di Data Mining e dei processi di AI, permette alle aziende di poter osservare gli eventi sotto un nuovo punto di vista, del tutto oggettivo, che tiene conto di una quantità di dati mai visti prima, riuscendo a scoprire legami e relazioni tra informazioni apparentemente lontane tra loro. Come il marketing nel passato, il riconoscimento della Data Science come disciplina aziendale, e gli investimenti in ottica di vantaggi che si potranno ottenere nel futuro prossimo, non potranno che apportare benefici alle aziende che avranno il coraggio di ragionare in un'ottica di forward looking.

Sono stati, poi, illustrati solo degli esempi di quello che è il potenziale della Data Science nei vari business, anche se le aziende che possono ottenere il vantaggio competitivo maggiore, sono quelle che si occupano di intermediazione finanziaria. La quantità di dati che scorre nei database delle società che operano in questo settore è enorme e la tecnologia applicata in ambito finanziario è ancora lontana da raggiungere la sua maturità.

L'elaborato approfondisce, dunque, l'applicazione della Data Science in tema di credit scoring. Questo binomio è dovuto al fatto che se la mission principale dei modelli di scoring consiste nel prevedere eventi futuri sulla base di dati passati e odierni, l'analisi dei dati può ottenere risultati inimmaginabili sotto il profilo della velocità e della qualità di analisi di quantità di dati enormi.

Tra le aziende di spicco sia in Italia, che in Europa, che fin da subito hanno investito nella Data Science, deve essere sicuramente menzionata Modefinance. Modefinance è una realtà italiana fondata nel 2009 a Trieste da Valentino Pediroda e Mattia Ciprian, certificata come prima Fintech Rating Agency a livello europeo.

Nella parte conclusiva dell'elaborato è riportata un'intervista originale a Valentino Pediroda, CEO e co-founder di Modefinance. Questa, ha permesso di fissare quelli che sono i vantaggi già espletati in precedenza riguardo la Data Science in tema di credit scoring, la figura del data scientist in una qualsiasi azienda operante nel mercato, ma anche dei vantaggi dell'utilizzo della scienza dei dati in termini di automatizzazione del calcolo dell'indice ESG.

D'altro canto, i Big Data e la Data Science, così come tutte le grandi innovazioni, portano con sé oltre che grandi vantaggi, anche innumerevoli rischi e sfide. Pediroda ha, infatti, confermato che uno dei lati critici a cui le aziende che lavorano nel settore dell'intermediazione creditizia e finanziaria devono porre maggiore attenzione è il tema della privacy. La protezione dei dati sensibili dei clienti deve essere accompagnata non solo da una normativa adeguata, ma essa deve essere tutelata anche da chi utilizza quei dati, una perdita di fiducia da parte del cliente rappresenterebbe un'enorme perdita di capitale per tutto il settore. Direttamente collegato al tema della fiducia è quello dell'Explainable Artificial Intelligence, che deve essere implementato da tutte le aziende che utilizzano tecnologie di AI.

Inoltre, il transito di grandi quantità di dati nei database degli intermediari finanziari, amplifica quello che può essere il rischio di frodi. In questo la Data Science può fornire gli strumenti adeguati ad identificare ed eliminare le minacce.

Lo studio dei dati rappresenta, quindi, al giorno d'oggi un possibile propulsore che, se sviluppato nel modo corretto, può permettere alle aziende appartenenti ai settori più vari di ottenere importanti vantaggi competitivi sul mercato.

Per quanto riguarda il futuro, non ci resta che godere dei benefici apportati da quest'innovativa scienza in quello che è il mondo del business, quindi nella società, quindi nelle nostre vite.

BIBLIOGRAFIA

Abdou, Hussein A., and John Pointon. "Credit scoring, statistical techniques and evaluation criteria: a review of the literature." *Intelligent systems in accounting, finance and management* 18, no. 2-3 (2011): 59-88.

Athanassopoulos, Antreas D., and Dimitris Giokas. "The use of data envelopment analysis in banking institutions: evidence from the commercial bank of Greece." *Interfaces* 30, no. 2 (2000): 81-95.

Azad, Poopak, Nima Jafari Navimipour, Amir Masoud Rahmani, and Arash Sharifi. "The role of structured and unstructured data managing mechanisms in the Internet of things." *Cluster computing* 23, no. 2 (2020): 1185-1198.

Bae, Jae Kwon, and Jinhwa Kim. "A personal credit rating prediction model using data mining in smart ubiquitous environments." *International Journal of Distributed Sensor Networks* 11, no. 9 (2015): 179060.

Brady, Henry E. "The challenge of big data and data science." *Annual Review of Political Science* 22 (2019): 297-323.

Cai, Li, and Yangyong Zhu. "The challenges of data quality and data quality assessment in the big data era." *Data science journal* 14 (2015).

Cao, Longbing, Qiang Yang, and Philip S. Yu. "Data science and AI in FinTech: An overview." *International Journal of Data Science and Analytics* 12, no. 2 (2021): 81-99.

Cirese, Claudio. "L'evoluzione del digital marketing nell'era dei big data: pro e contro del consumer profiling." (2021).

Clissa, Luca. "Survey of Big Data sizes in 2021." *arXiv preprint arXiv:2202.07659* (2022).

Chakravaram, Venkamaraju, Jangirala Srinivas, and Sunitha Ratnakaram. "The role of big data, data science and data analytics in financial engineering." In *Proceedings of the 2019 international conference on big data engineering*, pp. 44-50. 2019.

Chan, Wen Li, and Hsin-Vonn Seow. "Legally scored." *Journal of Financial Regulation and Compliance* (2013).

Dhar, Vasant. "Data science and prediction." Communications of the ACM 56, no. 12 (2013): 64-73.

Doerr, Sebastian, Leonardo Gambacorta, and José María Serena Garralda. "Big data and machine learning in central banking." *BIS Working Papers* 930 (2021).

Dotto, Sara. "Big Data: nuove potenzialità per l'azienda."

Fayyad, Usama M., David Haussler, and Paul E. Stolorz. "KDD for Science Data Analysis: Issues and Examples." In *KDD*, pp. 50-56. 1996.

Fisher, Tim. "Terabytes, Gigabytes, & Petabytes: How Big Are They?." *Accessed from* (2017).

Garg, Amit, Davide Grande, Gloria Macias-Lizaso Miranda, Christoph Sporleder, and Eckart Windhagen. "Analytics in banking: Time to realize the value." *Режим доступа: http://www. mckinsey. com/industries/financial-services/our-insights/analytics-in-banking-time-to-realize-the-value/(дата обращения: 18.03. 2018)* (2017).

George, Gerard, Ernst C. Osinga, Dovev Lavie, and Brent A. Scott. "Big data and data science methods for management research." *Academy of Management Journal* 59, no. 5 (2016): 1493-1507.

Hansen, M. M., T. Miron-Shatz, A. Y. S. Lau, and C. Paton. "Big data in science and healthcare: a review of recent literature and perspectives." *Yearbook of medical informatics* 23, no. 01 (2014): 21-26.

Hassani, Hossein, Xu Huang, and Emmanuel Silva. "Digitalisation and big data mining in banking." *Big Data and Cognitive Computing* 2, no. 3 (2018): 18.

Hassani, Hossein, Xu Huang, Emmanuel Silva, and Mansi Ghodsi. "Deep learning and implementations in banking." *Annals of Data Science* 7, no. 3 (2020): 433-446.

Hung, Jui-Long, Wu He, and Jiancheng Shen. "Big data analytics for supply chain relationship in banking." *Industrial Marketing Management* 86 (2020): 144-153.

Huttunen, J. E. N. N. I. F. E. R., Jaana Jauhiainen, L. A. U. R. A. Lehti, A. N. N. I. N. A. Nylund, Minna Martikainen, and O. M. Lehner. "Big data, cloud computing and data science applications in finance and accounting." *ACRN Journal of Finance and Risk Perspectives* 8 (2019): 16-30.

John, Samuel Ndueso, C. Anele, O. Okokpujie Kennedy, F. Olajide, and Chinyere Grace Kennedy. "Realtime fraud detection in the banking sector using data mining techniques/algorithm." In *2016 international conference on computational science and computational intelligence (CSCI)*, pp. 1186-1191. IEEE, 2016.

Joloudari, Javad Hassannataj, Hamid Saadatfar, Abdollah Dehzangi, and Shahaboddin Shamshirband. "Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection." *Informatics in medicine unlocked* 17 (2019): 100255.

Katal, Avita, Mohammad Wazid, and Rayan H. Goudar. "Big data: issues, challenges, tools and good practices." In *2013 Sixth international conference on contemporary computing (IC3)*, pp. 404-409. IEEE, 2013.

Madyatmadja, Evaristus Didik, and Mediana Aryuni. "Comparative study of data mining model for credit card application scoring in bank." *Journal of Theoretical and Applied Information Technology* 59, no. 2 (2014): 269-274.

Man, Shen Yi. "Data driven banking: applying Big Data to accurately determine consumer creditworthiness." Master's thesis, University of Twente, 2016.

Mandarà, Elena. "Tutela dei dati personali e antitrust: il caso dei sistemi IoT alla luce della sentenza Facebook/Germany (Bundeskartellamt B6-22/16)." (2021).

Nosratabadi, Saeed, Amirhosein Mosavi, Puhong Duan, Pedram Ghamisi, Ferdinand Filip, Shahab S. Band, Uwe Reuter, Joao Gama, and Amir H. Gandomi. "Data science in economics: comprehensive review of advanced machine learning and deep learning methods." *Mathematics* 8, no. 10 (2020): 1799.

Onay, Ceylan, and Elif Öztürk. "A review of credit scoring research in the age of Big Data." *Journal of Financial Regulation and Compliance* (2018).

Provost, Foster, and Tom Fawcett. "Data science and its relationship to big data and data-driven decision making." *Big data* 1, no. 1 (2013): 51-59.

Sadatrasoul, Seyed Mahdi, Mohammadreza Gholamian, Mohammad Siami, and Zeynab Hajimohammadi. "Credit scoring in banks and financial institutions via data mining techniques: A literature review." *Journal of AI and Data Mining* 1, no. 2 (2013): 119-129.

Sciotto, Fortunata, Giovanna Zucchi, Daniel Guzzetti, and Giorgio Corbucci. "La rivoluzione digitale." *GIAC* 6, no. 2 (2003).

Sheng, Jie, Joseph Amankwah-Amoah, and Xiaojun Wang. "A multidisciplinary perspective of big data in management research." *International Journal of Production Economics* 191 (2017): 97-112.

Signore, Davide. "Analisi del fenomeno dei big data." (2016).

Sravanthi, Kuchipudi, and Tatireddy Subba Reddy. "Applications of big data in various fields." *International Journal of Computer Science and Information Technologies* 6, no. 5 (2015): 4629-4632.

Treccani, sd.

Varetto, Franco, and Emanuele Scoccia. "Credit Scoring mediante tecniche di machine learning." (2021).

Varetto, Franco, and Niccolò Mangione. "Credit Risk Scoring Model con metodologie di data science."

Vyas, Daiwat A., Dvijesh Bhatt, and Dhaval Jha. "IoT: trends, challenges and future scope." *IJCSC* 7, no. 1 (2015): 186-197.

Waller, Matthew A., and Stanley E. Fawcett. "Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management." *Journal of Business Logistics* 34, no. 2 (2013): 77-84.

Wei, Yanhao, Pinar Yildirim, Christophe Van den Bulte, and Chrysanthos Dellarocas. "Credit scoring with social network data." *Marketing Science* 35, no. 2 (2016): 234-258.

Zhang, Yaoxue, Ju Ren, Jiagang Liu, Chugui Xu, Hui Guo, and Yaping Liu. "A survey on emerging computing paradigms for big data." *Chinese Journal of Electronics* 26, no. 1 (2017): 1-12.

SITOGRAFIA

www.google.com

www.investopedia.com

www.modefinance.com

www.treccani.it

RINGRAZIAMENTI

Vorrei ora ringraziare tutte le persone senza le quali la stesura di quest'elaborato non sarebbe stata possibile.

Desidero innanzitutto ringraziare il prof. Previtali, mio relatore, nonché accompagnatore, per i suoi consigli, il supporto e la comprensione dimostratomi.

Una menzione speciale è per il Dott. Pediroda, resosi disponibile a contribuire in prima persona alla stesura dell'elaborato.

Desidero inoltre ringraziare l'AS Luiss, che ormai da anni permette a ragazzi come me di conciliare quella che è la propria passione, la pallacanestro nel mio caso, con l'impegno accademico in questa prestigiosa Università.

Un grazie di cuore alla mia famiglia, che fin da piccolo mi ha sostenuto permettendomi di portare avanti entrambi i percorsi intrapresi.

Ringrazio i miei nonni, in particolare Giancarlo, che starà festeggiando questo mio traguardo così come ha sempre fatto, fin da quando ero piccolo.

Infine, ci tengo a ringraziare tutte le persone che mi hanno accompagnato durante questo viaggio, i miei compagni di squadra, i colleghi, gli amici da una vita Giuseppe e Francesco, e Irene, mia più grande forza.