UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO

DIPARTIMENTO DI INFORMATICA

Corso di laurea in informatica

Tesi di laurea

L'utilizzo dell'intelligenza artificiale per analizzare il mercato delle startup:
la variabilità del successo di un'azienda in seguito alle caratteristiche che presenta

Relatore:
Prof. Donato Impedovo

Laureando:
Giovanni Pio Amato

Correlatore:
Vincenzo Dentamaro

ANNO ACADEMICO 2023/2024

ABSTRACT

Le startup giocano un ruolo fondamentale per la ricerca e per il progresso tecnologico di tutta l'umanità. Sono queste aziende che, introducendo prodotti innovativi sul mercato, tendono a rendere la vita dell'uomo migliore. Infatti, i loro prodotti sono solitamente tesi a creare qualcosa di innovativo, non presente sul mercato, a cui nessuno prima aveva pensato.
Solitamente il successo di un'impresa di questo genere è una combinazione di più fattori. A incidere come fattori esterni sono essenzialmente: lo stato in cui nasce l'azienda, determinante per le leggi esistenti sul territorio; i competitor cioè coloro che propongono soluzioni simili sul mercato e il numero di investitori, che dipendono dal prodotto che si vuole realizzare. Oltre ai fattori esterni, ciò che maggiormente incide per la loro affermazione, sono le scelte aziendali e ciò che la startup andrà a creare.
Questo nuovo modo di fare impresa è stato introdotto da poco. Di conseguenza numerosi sono gli interrogativi relativi ai fattori che più incidono per l'affermazione di una startup sul mercato. In tal senso, il mio elaborato mira proprio a far luce su questi elementi e attraverso l'uso di modelli di intelligenza artificiale stabilisce, in base ai dati restituiti dallo studio e dalla ricerca, cosa realmente viene considerato importante per la crescita ed il consolidamento di una startup sul mercato.

INTRODUZIONE

Negli ultimi anni le startup si sono affermate sempre di più nel panorama mondiale, costituendo talvolta (come nel caso di Space X) delle valide alternative ad aziende governative che hanno un budget decisamente più elevato. Queste realtà stanno prendendo sempre più piede e, ad oggi, grazie anche a diversi sgravi fiscali operati dagli stati in cui vengono fondate, sono, a tutti gli effetti, una parte cruciale dell'ecosistema economico. Il mercato dell'innovazione tecnologica, ad esempio, è letteralmente esploso in questi anni grazie a chatgpt, il modello linguistico sviluppato da OpenAi, capace di comprendere e generare il testo in linguaggio naturale. La tecnologia completamente nuova per i più, i fondi investiti e i relativi soci che si sono prodigati affinché il progetto andasse in porto (tra i tanti investitori Bill Gates e Elon Musk), sicuramente hanno contribuito all'enorme successo dell'azienda. Certamente non possono essere gli unici fattori da prendere in considerazione per l'esplosione e la conseguente affermazione della stessa.
La mia tesi approfondisce la realtà delle startup e, attraverso l'addestramento di modelli di intelligenza artificiale, prova a stabilire quali sono i fattori che più incidono per il successo della stessa, in base ad un dataset limitato creato con l'uso del web. Inoltre, crea un modello di Digital Twin che simula l'andamento di una startup. Queste simulazioni prevedono l'andamento aziendale in risposta a diversi scenari di change management acquisiti dall'utente.
Prima di affrontare il tema legato all'intelligenza artificiale e alla digital twin, si farà luce sul tema della startup, analizzando in prima battuta cosa essa sia e

come è costituita. Il primo capitolo, infatti, si concentrerà sulla startup, percorrendo tutte le definizioni attribuitele dalla letteratura, esaminando tutte le tipologie esistenti, approfondendone il ciclo di vita, il ciclo di finanziamento e i motivi dietro al fallimento delle stesse. Nel secondo capitolo, invece, ci si addentrerà più sulla parte pratica ovvero si andrà ad esaminare più nel dettaglio come il dataset è stato costruito, i modelli di intelligenza artificiale utilizzati, valutando i pro e i contro per ogni modello e le relative differenze, l'interfaccia grafica che è stata utilizzata per acquisire i dati relativi alla digital twin. In particolare si andrà a vedere cosa è il change management e come esso possa influire sulla vita di una startup e sulle persone che la costituiscono.

Nel terzo capitolo verrà approfondito il tema dell'intelligenza artificiale sia da un punto di vista prettamente tecnico, sia da un punto di vista etico e sociale attraverso le mie personali considerazioni relative all'utilizzo di questa risorsa.

Il quarto capitolo, infine, analizzerà i risultati e presenterà pareri soggettivi tesi a migliorare il lavoro fatto, con lo scopo di ampliarlo e renderlo effettivamente fruibile da chiunque voglia fondare una nuova azienda o da coloro che vogliano dedicarsi allo studio delle previsioni dell'andamento di una startup.

CAPITOLO 1: LE STARTUP

1.1 Definizione

Il concetto di startup nasce nel 1902 in California, più precisamente nella Silicon Valley, la culla tecnologica di molte grandi aziende che, ancora oggi, dominano il panorama mondiale. Il termine startup nell'immaginario comune viene associato sia a business innovativi creati da ragazzi visionari sia a grandi imprese tecnologiche. In realtà il termine non ha ancora dei limiti di confine definiti e viene spesso confuso con la fase di inizio di una impresa. Molti esperti in questo campo hanno provato a definire il significato di startup esprimendo spesso dei concetti simili con accezioni diverse, risultando a volte troppo generici e evidenziando sempre piccole differenze.

Secondo Eric Ries, autore del libro, The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful, "Una startup è un organizzazione umana progettata per creare un nuovo prodotto o servizio in condizioni di estrema incertezza". Per Neil Blumenthal, cofondatore e co-CEO di Warby Parker "Una startup è un'azienda che lavora per risolvere un problema in cui la soluzione non è ovvia e il successo non è garantito". Adora Cheung la definisce così "È quando le persone si uniscono alla tua azienda prendendo la decisione esplicita di rinunciare alla stabilità in cambio della promessa di una crescita straordinaria".

Secondo Merriam-Webster, startup significa "un atto o un'istanza di messa in funzione o movimento" oppure "un'impresa commerciale alle prime armi". L'American Heritage Dictionary suggerisce che si tratta di "un'attività o impresa che ha recentemente iniziato a operare"

Per Paul Graham, capo dell'acceleratore di Y Combinator, invece "la sola caratteristica essenziale di una startup è la crescita". "Un'azienda

di cinque anni può ancora essere una startup", "Dieci anni inizierebbero a essere una forzatura".
Continua dicendo “Mi metterò in disparte e dirò categoricamente che dopo circa tre anni di attività, la maggior parte delle startup cessa di essere startup. Questo spesso coincide con altri fattori: acquisizione da parte di un'azienda più grande, ricavi superiori a $ 20 milioni, più di 80 dipendenti, oltre cinque persone nel consiglio e fondatori che hanno venduto personalmente azioni. Ironia della sorte, quando una startup diventa redditizia, è probabile che si allontani dalla startup.”
Il problema con questa definizione è che ogni negozio a conduzione familiare nel tuo quartiere che è stato aperto per generazioni sarebbe quindi considerato una startup. Allo stesso modo, se hai appena creato una piccola impresa a scopo di lucro e sei intenzionato a diventare abbastanza grande da conquistare il mondo, anche se stai ancora lavorando dalla tua camera da letto, probabilmente sei un fondatore di startup.
Si può riconoscere, da questa moltitudine di definizioni, che l’attributo chiave di una startup è la capacità di crescere. Lo scopo, come dice Graham, è la crescita rapida e di conseguenza scalare rapidamente. È questa propensione alla crescita che differenzia una startup da una piccola impresa.
Alyson Shontell, Editor-in-Chief di Business Insider US, definisce “Una startup è un roller coaster emotivo che può portare a enormi fallimenti o successi, dopo di che il totale del proprio conto in banca può aumentare o diminuire drasticamente. La persona dietro una startup è un fondatore, una persona spesso molto brillante, un po' pazza che trova noioso un normale lavoro nine-to-five1 e si illude di credere di poter cambiare il mondo lavorando instancabilmente davanti allo

schermo di un computer. Il lavoro implacabile è noto per consumare alcuni anni di vita di un fondatore aggiungendo prematuri capelli grigi, ma può essere molto gratificante sia emotivamente che finanziariamente per coloro che lo perseguono"

Ma tra tutte le definizioni la più popolare e universalmente accettata è quella data da Steve Blank nel suo libro "The Startup Owner's Manual" in cui afferma che "Una startup è un'organizzazione temporanea progettata per cercare un business model ripetibile e scalabile"

Questa definizione evidenzia i tre requisiti fondamentali che, secondo Blank, la startup deve avere:

-Temporaneità: Secondo Blank, uno degli elementi che caratterizza una startup è la temporaneità. A differenza di un'impresa di piccole dimensioni, una startup company nasce già con l'obiettivo di crescere rapidamente e trasformarsi in un'impresa di grandi dimensioni (scaleup company).

-Execution vs Ricerca: Per soddisfare i clienti, le imprese tradizionali mettono in atto un business model già sviluppato e validato (execution). Le startup al contrario nella fase iniziale del loro sviluppo sono ancora alla "ricerca" di un modello di business innovativo che permetta loro di creare valore e che soddisfi al meglio i loro clienti. L'alto rischio al quale questo tipo di imprese è associato si lega al fatto che si debba "testare" sul mercato una formula imprenditoriale ancora non definita.

-Model Business ripetibile e scalabile: Questo modello di business deve essere poi ripetibile nei suoi processi (ingegneria, marketing, ecc.) e su vari Paesi oltre ad essere scalabile. Esso deve cioè permettere una crescita esponenziale in termini di dimensioni, fatturato e investimenti senza un proporzionale aumento dei costi (esempi possono essere i

libri in formato ebook e la stampa 3D).

1.2 Tipi di Startup

Le startup vengono solitamente categorizzate:
- In base alla qualità della loro idea di prodotto o servizio vengono classificate come startup visionarie, startup innovative o startup ordinarie
-A seconda del mercato o al segmento al quale si rivolgono possono essere classificate in startup Tech suddivise poi in: MedTech (medico), startup FinTech (finanziario), startup FoodTech (food & beverage), BioTech(biotecnologie). Startup viaggi (Turismo), startup fashion o startup digitale. Nel nostro caso di studio si è scelto di considerare solo aziende catalogate come tecnologiche (Tech), considerando anche poche startup che operano o operavano nel mercato finanziario(Fintech) e solo alcune riguardanti le biotecnologie(BioTech).
-Secondo le tipologie di startup fornite da Steve Blank in un articolo comparso nel 2013 nella sezione Business del sito del Wall Street Journal:

- Lifestyle
- Small business
- Scalabile
- Acquisibile
- Large company
- Sociale

Lifestyle Startup: Work to Live Their Passion

Gli imprenditori lifestyle sono paragonati ai surfisti californiani, che danno lezioni di surf per pagarsi le bollette facendo della propria passione un lavoro. Queste persone vivono la vita che amano, non lavorando per nessuno. Quello che Blank indica come “l’equivalente in Silicon Valley” è il programmatore o web designer dipendente, che ama la tecnologia e accetta lavori di coding e U/I, per poter perseguire con tali incarichi la sua passione.

Small business Startup: Work to Feed the Family

Si tratta della maggioranza delle startup presenti attualmente negli Stati Uniti, nelle quali l’imprenditore è colui che gestisce direttamente l’attività. Si tratta di persone che investono il proprio capitale nel business (o quello preso in prestito da familiari e amici, o dalle banche). Spesso queste attività sono a malapena redditizie, ma nella maggior parte dei casi questi imprenditori sono quelli che Blank definisce più rappresentativi del concetto di “imprenditorialità”, in quanto lavorano con passione e dedizione creando nuovi posti di lavoro a livello locale. L'imprenditorialità delle piccole imprese non è progettata per scalare.

Startup scalabili: Born to Be Big

Fin dalla creazione, i founder credono che cambieranno il mondo. Google, Uber, Facebook, Twitter sono solo gli ultimi esempi di startup scalabili. Tali startup assumono i migliori e i più brillanti. Cercano sempre un modello di business ripetibile e scalabile. Quando lo trovano, iniziano a cercare più capitale di rischio per incrementare le

loro attività. Spesso le startup scalabili si raggruppano in cluster di innovazione (Silicon Valley, Shanghai, New York, Boston, Israele, ecc.)

Startup acquistabili: Acquisition Targets

Negli ultimi cinque anni, le startup che offrono soluzioni Web e app mobili sono state vendute ad aziende più grandi. Questa tendenza diventa sempre più popolare. Il loro obiettivo non è costruire una società da miliardi di dollari, ma essere venduti a una società più grande per $5-$50 M.

Large Company Startup: Innovate or Evaporate

Le grandi aziende hanno una durata di vita finita. Cambiamenti nelle preferenze dei clienti, nuove tecnologie, problemi legislativi, nuovi concorrenti creano pressione, costringendo le grandi aziende a creare nuovi prodotti innovativi per nuovi clienti in nuovi mercati (ad esempio Google e Android)

Startup Sociali: Driven to Make a Difference

Sono appassionate e spinte ad avere un impatto. Tuttavia, a differenza delle startup scalabili, la loro missione è rendere il mondo un posto migliore. Di conseguenza non sempre hanno come scopo il profitto ma possono essere organizzazioni che lavorano per lo sviluppo del bene comune senza trarne un vero guadagno.
I fondatori di queste imprese sono tutti imprenditori ma tra vari tipi di startup vi sono significative differenze riguardo ad obiettivi finanziari, motivazioni dei team e strategie finanziarie da attuare. Se non si tiene conto di queste differenze, afferma Blank, si riducono drasticamente le probabilità di successo.

Proprio l’imprenditori, che decidono di avventurarsi in questo mondo, costituiscono un ulteriore differenza tra le startup. Possiamo distinguere tre tipi di configurazioni aziendali dipendenti dal tipo di imprenditore che ne è a capo:

-Cluster C1: Imprenditori nascenti contro la loro volontà

Questi imprenditori sono caratterizzati da un basso bisogno di realizzazione, basso locus of control interno e bassa iniziativa personale. Mostrano una ridotta motivazione alla sicurezza e risorse personali sfavorevoli. La loro situazione è aggravata dalla mancanza di supporto sociale e dalla scarsa percezione dell'importanza delle reti di contatti. Durante il processo di avvio, tendono a sottovalutare gli sforzi organizzativi e a fare scarso uso delle informazioni.

-Cluster C2: "Imprenditori nascenti potenziali"

Questo gruppo è motivato dalla realizzazione personale e ha una forte percezione di modelli positivi. Mostrano un migliorato locus of control interno, ma affrontano una situazione finanziaria sfavorevole e una maggiore motivazione alla sicurezza. Di conseguenza, percepiscono maggiori sforzi organizzativi nel processo di avvio, dovuti alle attività necessarie per stabilire una base finanziaria. Questo modello ambivalente riflette le sfide che incontrano nel bilanciare le loro aspirazioni personali con le realtà finanziarie

-Cluster C3: Imprenditori nascenti con rete e modelli di evitamento del rischio

La caratteristica principale di questo gruppo è la ridotta propensione al rischio, che potrebbe spiegare la loro alta considerazione del fallimento. Tuttavia, percepiscono un ambiente fortemente favorevole, indicato sia da alti valori di supporto sia dall'importanza delle reti di contatti. Utilizzano intensamente le informazioni,

affrontano pochi problemi e richiedono pochi sforzi organizzativi. Godono di una situazione di risorse sopra la media, il che li pone in una posizione di sicurezza durante il processo di avvio. La loro cautela è vista come una valutazione attenta piuttosto che procrastinazione. Detto ciò si può chiaramente concludere che l'imprenditore e le scelte manageriali sono certamente le più influenti sul successo di un'azienda. Certo è che l'effetto varia a seconda del contesto istituzionale e del livello di sviluppo economico dello stato in cui la startup viene fondata. Il processo di avvio delle imprese è influenzato dalle istituzioni e dalle politiche governative, che possono facilitare o ostacolare l'imprenditorialità. Nei paesi sviluppati, l'imprenditorialità tende ad aumentare nella fase di innovazione, mentre nei paesi in via di sviluppo è più prevalente nella fase di efficienza, dove l'economia cerca di aumentare la produttività e la qualità della forza lavoro. Il contesto istituzionale e le politiche pubbliche sono cruciali per promuovere un ambiente favorevole all'imprenditorialità, specialmente nei paesi in via di sviluppo.

Molto determinante è anche l'apporto degli investitori e dei finanziamenti ottenuti appunto da questi ultimi. Distinguiamo due tipi di individui in un contesto aziendale gli stakeholder e gli shareholder. Gli shareholder sono individui che possiedono azioni di una società. Il loro unico interesse è legato al profitto della startup e quindi al suo successo in termini di dividendi.

Gli stakeholder sono individui portatori di interessi che influenzano le decisioni aziendali e allo stesso modo possono essere influenzati dall'azienda stessa.

Gli stakeholder possono essere azionisti oppure semplicemente dipendenti, clienti o comunità. I loro obiettivi sono vari ma non sono

focalizzati sul profitto dell'azienda. Tuttavia, possono influenzare con le loro azioni il funzionamento e i servizi forniti dall'azienda.

1.3 Ciclo di vita

Le startup, durante il loro ciclo di vita, attraversano una serie di fasi diverse in cui le

esigenze, gli obiettivi e le caratteristiche di ciascuna variano notevolmente. Vediamo nel

dettaglio quali sono le fasi più importanti:

- Pre-seed
- Seed
- Early stage
- Early growth
- Growth
- Expansion
- Exit

1.3.1 Pre-seed

Questa è la fase di creazione dell'idea. Bisogna capire se l'idea può funzionare, se risolve un bisogno e se c'è un potenziale mercato.
In questa fase non esiste ancora un prodotto minimo funzionale2 o un modello di business convalidato. Questo è il momento di cercare un co-founder e formare il team iniziale, gettare le basi legali e iniziare a pensare su come trasformare la propria idea in qualcosa di reale,

fattibile. Bisogna capire se l'idea può trasformarsi in un prodotto o servizio e se la gente sia disposta a pagare per questo prodotto o servizio. È una fase molto delicata che lo startupper dovrebbe condurre in maniera rigorosamente oggettiva. In questo momento è già possibile ottenere piccoli investimenti, chiamati FFF (Friends, Family and Fools). Spesso è molto difficile accedere ai capitali, quindi saranno gli ideatori di tale idea che dovranno investire sui costi iniziali. Una delle opzioni che esistono per aiutare in queste prime fasi è quella di partecipare a un acceleratore di startup, ovvero un programma, offerto da terze parti, progettato per supportare le startup emergenti (di solito in fase iniziale) nella loro crescita e sviluppo. Solitamente chi fornisce questo genere di servizi sono le grandi aziende in supporto a startup che hanno idee complementari alle loro. Altri possibili fornitori sono li enti governativi o le università, con lo scopo di commercializzare la propria ricerca.

1.3.2 Seed

La fase seed è forse una delle fasi più importanti del ciclo di vita di una startup. È la fase in cui il progetto diventa realtà e l'obiettivo principale è quello di sviluppare correttamente l'idea e validare il modello di business. La convalida di un'idea è il processo mediante il quale vengono raccolte prove, attraverso la sperimentazione, per prendere decisioni rapide, informate e prive di rischi. Vanno proposte alcune ipotesi e assunzioni iniziali, che attraverso un metodo di verifica saranno confermate o respinte. Il rifiuto, o la mancata convalida,

dell'ipotesi iniziale riflette la necessità di orientarsi verso una nuova ipotesi. Durante il seed si lavora per affinare il business model Canvas o Lean Canvas, si struttura il business plan e si fanno i primi passi verso la creazione di un prodotto minimo funzionante (Minimum Viable Product – MVP) che consenta di testare il proprio prodotto sul mercato, con clienti reali.
La cosa più importante è ottenere la convalida dei clienti e magari un finanziamento di portata limitata (20.000-40.000 euro).
In questa fase i programmi di accelerazione possono essere molto utili per le startup poiché consentiranno loro di accelerare questo processo di "trial and error", grazie al contatto con professionisti con una vasta esperienza che offrono le loro competenze in aree chiavi per l'evoluzione di una startup, come il marketing, vendite, faccende legali e altro ancora.

1.3.3 Early Stage

L'obiettivo principale di questa fase del ciclo di vita di una startup è ottenere feedback dal mercato stesso e individuare il giusto product/market fit, che permetta di ottenere i primi ricavi. Quando la startup ha già il suo MVP sul mercato e ha i primi clienti/utenti è arrivato il momento di migliorare questo prodotto innovativo attraverso un processo iterativo, in cui vengono raccolti i feedback degli utenti e vengono migliorati i difetti o bug.
In questa fase, oltre alle tipologie di finanziamento viste nella fase precedente, acquisiscono più valore Venture Capital e acceleratori, che sono entità che oltre a supportare le idee emergenti, aiutano anche i nuovi modelli di business a testare le loro soluzioni e ad accedere ai clienti. Un'altra dinamica sotto i riflettori è il crowdfunding. È una fase

molto delicata in cui la capacità di attrarre investimenti fa la differenza tra la vita e la morte. Individuare quali sono le caratteristiche o le funzionalità più importanti della startup è uno dei compiti essenziali di questo ciclo, così come stabilire le prime relazioni o accordi commerciali per il futuro.

1.3.4 Early Growth

In questa fase la startup ha sviluppato un buon MVP, i clienti iniziano ad arrivare e le persone pagano per il prodotto o servizio. Ora è il momento di puntare in alto e lavorare sul business model per trovare la combinazione vincente che permetta di scalare e crescere. Il piano di marketing e la strategia commerciale sono cruciali per acquisire rapidamente clienti, espandersi in maniera capillare nel paese d'origine e avviare l'internazionalizzazione. Mentre nelle altre fasi è associato il primo round di finanziamento ovvero quello di seed, in questa fase solitamente si entra nel round di serie a, il primo vero round di finanziamenti. L'obiettivo principale è migliorare il business, consolidare il prodotto e aumentare la cerchia dei clienti

1.3.5 Growth

L'ultima fase del ciclo di vita di una startup è definita Growth o Sustained Growth, e rappresenta la fase di crescita sostenuta. Nella fase di Growth la crescita di utenti e clienti diventa esponenziale e il fatturato aumenta rapidamente. Solo poche startup raggiungono questa fase, quelle che ci riescono hanno già un prodotto con il suo mercato (product/market fit), clienti stabili e numeri positivi, nonché una strategia di crescita definita e un modo per ottenere maggiori clienti. Questo percorso è la fase in cui la startup dovrebbe

concentrarsi sulla sua crescita e aumentare sia i vantaggi che il numero di clienti. In questa fase sono protagonisti i Venture Capital e i Corporate Venture Capital; mentre il Venture Capital è una forma di finanziamento fornita da investitori istituzionali, il corporate venture capital è una forma di venture capital fornita da grandi aziende o corporazioni alla ricerca di sinergie per il loro business.
Associato a questa fase ci sono due round di finanziamento quello di serie a e il successivo quello di serie b. La fase di serie b è utilizzata per espandere ulteriormente il business, scalando le operazioni e ampliando la presenza sul mercato

1.3.6 Expansion

Una volta che il prodotto è consolidato si cerca di distribuire il prodotto su più mercati, in modo tale da portare avanti la continuità aziendale. L'azienda, quindi, apre le porte a mercati internazionali o alla stessa area geografica ma in segmenti diversi.
Questa è una parte critica, poiché ci sono molti rischi e il futuro della startup può dipendere dalla scelta del posto o del settore in cui espandersi.
Raggiungere accordi con grandi aziende già stabilite in diversi paesi o nei diversi settori che si intende raggiungere può essere un modo più semplice per eseguire questo processo.
A questa fase sono solitamente associati finanziamenti di serie b e serie c. Questi finanziamenti sono usati per espandere ulteriormente le operazioni, entrare in nuovi mercati, e migliorare l'efficienza operativa. L'obiettivo è consolidare la posizione sul mercato e preparare l'azienda per un'eventuale uscita.

1.3.7 Exit

L’exit determina il passaggio dallo stato di startup ad azienda vera e propria e rappresenta il momento di uscita degli investitori dalla proprietà della startup.
Le principali opzioni per la exit sono:

• Tramite l’Offerta Pubblica Iniziale o IPO (Public Sale Offer “OPV”) la startup mette a disposizione del pubblico le proprie azioni.
L’imprenditore quota in borsa la propria attività per accedere, rapidamente, a finanziamenti necessari per lo sviluppo.
• Acquisizione della startup da parte di un’altra azienda.
• Completamento dell’espansione e consolidamento della propria posizione sul mercato.
• Fallimento e conseguente chiusura dell’impresa
Per quanto riguarda questa fase, chiaramente, se si dovesse optare per la prima opzione ci troveremmo nel round di finanziamento Ipo.
Un'IPO rappresenta la transizione verso un'azienda pubblica, permettendo di raccogliere capitali su larga scala e offrendo liquidità agli investitori iniziali. Di contro, se ci trovassimo nell’ultima opzione l’azienda sarebbe catalogata come morta. Nelle altre opzioni saremmo nei round serie d o successivi, in cui l’azienda si espande su larga scala, attraverso acquisizioni strategiche o l’espansione internazionale.

1.4 I motivi del fallimento di una startup

La rilevanza, quando si parla di startup, viene spesso data ai grandi successi, ovvero alle startup che vengono definite unicorno, cioè le startup che hanno una valutazione maggiore di un miliardo di euro.
Spesso ciò che distingue una startup di successo da una che non lo ha è la fase di avvio. CB Insights ha scoperto che il 70% delle nuove aziende

tecnologiche fallisce. E le startup di hardware di consumo falliscono ancora più frequentemente, con il 97% che alla fine muore o diventa "zombi". Gli analisti di CB Insight volevano capire perché così tante startup falliscono, quindi hanno scavato nei "post-mortem" scritti da fondatori, investitori e giornalisti di quasi 300 fallimenti di startup.

Vi sono molte ragioni per le quali le startup falliscono e solo in pochi casi sono riconducibili ad una singola causa. Nella maggior parte dei casi il fallimento è determinato da una combinazione di fattori. CB Insights, che tramite un proprio software raccoglie dati ed elabora statistiche relativamente al mondo delle startup, ha recentemente presentato la sua ricerca sulle cause del fallimento. Tra le principali motivazioni che determinano il fallimento di una Startup lo studio annovera:

• Mancanza di liquidità: da ricondurre anche ad una cattiva gestione e ad errori nella allocazione delle risorse, è la causa di fallimento di una startup nel 38% dei casi. Tale mancanza si ricollega anche ad altre cause quali l'incapacità di soddisfare un bisogno del mercato o la mancanza di leadership.

• Mancato soddisfacimento di un bisogno del mercato: nel 35% dei casi analizzati la causa del fallimento di una startup è riconducibile nell'offerta di prodotti che non soddisfano un preciso bisogno del mercato.

• Non supera la concorrenza: Nonostante i luoghi comuni secondo cui le startup non dovrebbero prestare attenzione alla concorrenza, la realtà è che una volta che un'idea diventa interessante o ottiene la

convalida del mercato, potrebbero esserci molti concorrenti in giro. E mentre l'ossessione per la concorrenza non è salutare, ignorarli è stata la causa del fallimento nel 20% dei casi.

•modello di business debole: la mancanza di un modello di business scalabile che preveda delle attività collaterali su cui direzionarsi in caso di fallimento delle attività core, non solo non è profittevole per l'imprenditore ma genera anche dei dubbi negli investitori con conseguente difficoltà per la startup di reperire finanziamenti. Questo porta inevitabilmente al fallimento nel 17% dei casi.

• Sfide legali: A volte una startup può evolvere da una semplice idea a un mondo di complessità legali che può rivelarsi nel 18% dei casi una delle cause della chiusura di una startup.

• Pricing/costi: stabilire il giusto prezzo per un prodotto o servizio non è mai semplice e ciò vale a maggior ragione per una startup. Non dare al proprio prodotto il giusto valore tenendo conto dei costi e del burn rate iniziale è causa di fallimento nel 15% dei casi.

• Team: all'interno della startup devono essere presenti risorse umane con le competenze necessarie a realizzare la "visione" dell'imprenditore. Il team deve inoltre condividere questa visione, dare un contributo anche di idee significativo, essere coeso ed avere una grande adattabilità, questo non succedeva nel 14% dei casi. Necessaria è anche la presenza di un fondatore con una forte passione per il proprio business.

• Rilasciare il prodotto al momento sbagliato: Se rilasci il tuo prodotto troppo presto, gli utenti potrebbero considerarlo non abbastanza

buono e recuperarli potrebbe essere difficile se la loro prima impressione su di te è stata negativa. Se invece rilasci il tuo prodotto troppo tardi, potresti aver perso la tua finestra di opportunità sul mercato. Il timing è fondamentale, e causa fallimenti nel 10% dei casi.

• Scarsa qualità del prodotto: Il prodotto è difettoso, non affidabile o non funziona come promesso. Problemi di qualità possono portare a feedback negativi, abbandono da parte degli utenti, e difficoltà nel costruire una base clienti solida. Questo accade nell' 8% dei casi.

• Disallineamento team-investitori: La discordia con un co-fondatore è stato un problema fatale per il 7% delle startup. Questo disallineamento non è limitato al team fondatore, ma anche quando le cose vanno male con il Board o gli investitori la situazione precipita rapidamente, come nel caso di Pellion Technologies. Il suo più grande investitore, Khosla Ventures, ha perso fiducia nel fatto che Pellion potesse fare abbastanza soldi servendo un mercato di nicchia. Per questo, nel marzo 2019, Khosla ha deciso che la società sarebbe stata chiusa e ha rimosso il nome di Pellion dal suo portafoglio.

• Pivot andato male: Il Pivot, ovvero una modifica al business model o al prodotto apporta grandi cambiamenti per tutti. Pivot come Burbn su Instagram o ThePoint su Groupon possono andare straordinariamente bene. Oppure possono essere l'inizio di un percorso lungo la strada sbagliata. Questo accade nel 6% dei casi.

• Burn out/perdita di passione: Responsabilità e volume di lavoro per far crescere una startup sono un fardello pesante da portare, e spesso i fondatori lo sottovalutano. Spesso i fondatori non hanno una profonda conoscenza del settore dove operano e questo porta ad addentrarsi in

argomenti che si discostano, a volte, da ciò a cui uno è interessato dalla propria passione. Queste ragioni hanno portato al fallimento nel 5% dei casi

Figura 1**: I 12 motivi per cui le startup falliscono.* ***Fonte: *"The Top 12 reasons startup fail", CBInsights.*

1.5 I modi per finanziare una startup

I finanziamenti per startup, specie nelle fasi iniziali, sono essenziali per una startup early stage. Da vari studi emerge che più del 94% delle start up non supera il primo anno di vita, e che una delle cause è proprio la mancanza di finanziamenti. Di seguito elenco i modi usati per finanziare una startup

1.5.1 Bootstrapping

Il bootstrapping è il ricorso a capitali propri per l'avvio di una startup, in pratica è l'autofinanziamento. In fase di avvio lo startupper può avere difficoltà nel reperire finanziamenti e ricorre spesso all'utilizzo di capitali propri o capitali chiesti ad amici, parenti e a soggetti che hanno una certa disponibilità finanziaria e sono disposti a investirla in un'impresa altamente rischiosa.

1.5.2 Crowdfunding

Il crowdfunding è diventato rapidamente il modo migliore per gli imprenditori di finanziare le proprie startup. Per crowdfunding si intende la raccolta di piccole somme di denaro che vengono richieste a più persone attraverso canali diversi e con finalità diverse. Il canale di crowdfunding più diffuso è il web dove a tal fine operano le cosiddette piattaforme di crowdfunding come Kickstarter e Indiegogo. In pratica il richiedente presenta il proprio progetto alla piattaforma, la quale lo pubblica in rete. Se la campagna di crowdfunding ha successo, il

richiedente riceverà la somma richiesta; chi finanzia riceverà una ricompensa, mentre la piattaforma riceve una provvigione. Questi sostenitori non sempre avranno voce in capitolo su come viene gestita l'attività, e condividono collettivamente un rischio relativamente piccolo ciascuno, perché insieme desiderano entusiasticamente che il progetto in questione esista. Questo modello di finanziamento può essere utilizzato non solo per raccogliere fondi iniziali, ma può essere utilizzato per la successiva raccolta di fondi per prodotti e servizi futuri.

1.5.3 Business Angel

I Business angel sono persone che investono i propri soldi nella fase iniziale delle startup, in cambio di una partecipazione al capitale. Gli investimenti in startup vanno generalmente dai € 20.000 fino ad un massimo di € 100.000. Di solito svolgono anche il ruolo di mentore e offrono il loro consenso e la loro esperienza agli imprenditori. L'obiettivo dei business angel è, da un lato, ottenere un beneficio dal loro investimento, e dall'altro, aiutare gli imprenditori a portare con successo la loro idea di business sul mercato. Il ruolo di questi investitori è diventato determinante nel caso di molti progetti imprenditoriali, visto che partecipando non solo apportano denaro, ma anche esperienza, consulenza, rete di contatti e visione imprenditoriale, che possono fare la differenza tra una semplice idea e un'azienda concreta e vitale.

1.5.4 Venture capital

Il venture capital è una forma di investimento di medio-lungo termine in imprese non quotate ad alto potenziale di sviluppo e crescita (high grow companies) che si trovano nella fase di start up, effettuata prevalentemente da investitori istituzionali con l'obiettivo di ottenere un consistente guadagno

in conto capitale dalla vendita della partecipazione acquisita o dalla quotazione in borsa. Per la mancanza in fase iniziale di un prodotto e un business model definito, le startup risultano infatti troppo rischiose per ricevere finanziamenti dalle banche. L'investitore in un fondo di venture capital (chiamato Venture Capitalist) una volta scelta la startup su cui investire, entra nel capitale di rischio (quote o azioni) di quella società e vi rimane sino alla scadenza del finanziamento (3-10 anni) o in caso di vendita della società ad altra compagnia (Exit). A differenza dei Business Angel i Venture Capitalist investono generalmente molti milioni nella startup, e sebbene ci sia accettazione del rischio, sono molto selettivi su chi supportare.

1.5.5 Incubatore startup

Un incubatore di startup è un programma collaborativo progettato per aiutare le nuove startup ad avere successo. Secondo la definizione riportata su "The smart guide of innovation" promossa dalla Commissione europea, un incubatore startup "è un luogo dove gli imprenditori trovano le strutture, i servizi e le competenze necessarie ai loro bisogni ed a sviluppare le loro idee di business e trasformare queste in realtà sostenibili".

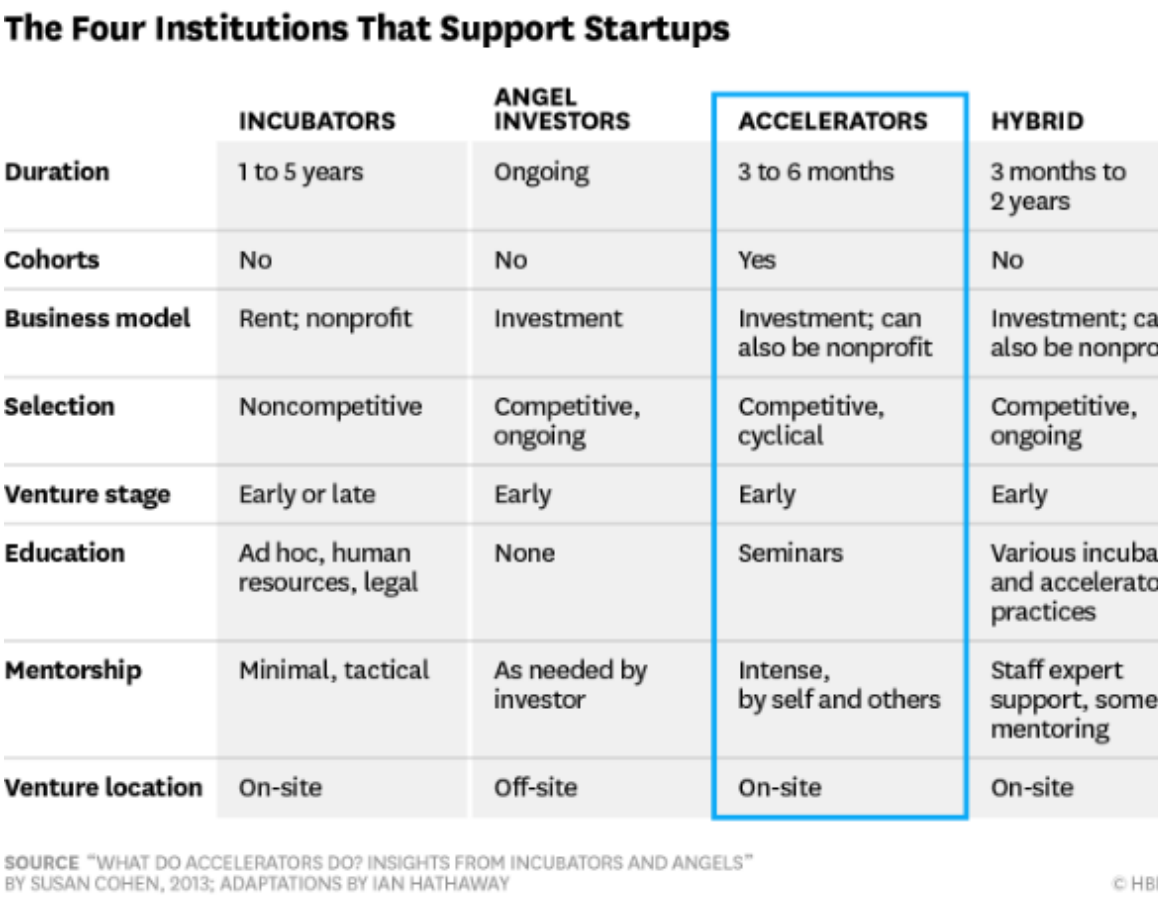
I servizi offerti dagli incubatori includono: spazi fisici e di co-working, servizi amministrativi e organizzativi, formazione, consulenza (dai cosiddetti mentor startup), accesso a finanziamenti e networking.

L'unico scopo di un incubatore di startup è aiutare gli imprenditori a far crescere il proprio business. Gli incubatori di startup sono generalmente organizzazioni senza scopo di lucro, che di solito sono gestite da enti pubblici e privati. Gli incubatori sono spesso associati alle università e ad alcune business school che consentono ai loro studenti e alunni di prendere parte a questi programmi. Ci sono molti

altri incubatori, tuttavia, formati da governi, gruppi civici, organizzazioni di startup o imprenditori di successo.

1.5.6 Acceleratore startup

Un acceleratore startup o seed accelerator è una società che supporta lo sviluppo di altre società, tipicamente startup, attraverso dei programmi di accelerazione. L'esperienza dell'acceleratore è un percorso di formazione intenso, rapido e immersivo volto ad accelerare il ciclo di vita delle giovani imprese innovative, comprimendo anni di learning by doing in pochi mesi. Susan Cohen dell'Università di Richmond e Yael Hochberg della Rice University evidenziano i quattro fattori distintivi che rendono unici gli acceleratori: sono a tempo determinato, guidati da tutoraggio e culminano in presentazioni pubbliche (public pitch event) o giornate dimostrative (demo day). Nessuna delle altre istituzioni (incubatori, Business Angels o Venture Capitalist) ha questi elementi collettivi. Gli acceleratori possono condividere con questi altri l'obiettivo di coltivare startup in fase iniziale, ma è chiaro che sono diversi, con modelli di business e strutture di incentivazione nettamente diversi.



The Four Institutions That Support Startups

	INCUBATORS	ANGEL INVESTORS	ACCELERATORS	HYBRID
Duration	1 to 5 years	Ongoing	3 to 6 months	3 months to 2 years
Cohorts	No	No	Yes	No
Business model	Rent; nonprofit	Investment	Investment; can also be nonprofit	Investment; ca also be nonpro
Selection	Noncompetitive	Competitive, ongoing	Competitive, cyclical	Competitive, ongoing
Venture stage	Early or late	Early	Early	Early
Education	Ad hoc, human resources, legal	None	Seminars	Various incuba and accelerato practices
Mentorship	Minimal, tactical	As needed by investor	Intense, by self and others	Staff expert support, some mentoring
Venture location	On-site	Off-site	On-site	On-site

SOURCE "WHAT DO ACCELERATORS DO? INSIGHTS FROM INCUBATORS AND ANGELS" BY SUSAN COHEN, 2013; ADAPTATIONS BY IAN HATHAWAY

© HB

Figura 2: Le quattro istituzioni che supportano le startup ***Fonte:*** *"What do accelerators do?" HBR.ORG*

1.5.7 Prestiti bancari

Le banche hanno di recente incominciato a supportare le startup e PMI innovative e lo hanno fatto muovendosi essenzialmente su tre binari.

• La concessione di finanziamenti tramite il fondo di garanzia per le PMI: il fondo di garanzia (istituito con la legge 662 del 1996, e dal 2013 estesa alle Startup innovative) è un fondo governativo concesso per un importo massimo di 2,5 milioni di euro e fino a un massimo dell'80% del finanziamento. Oltre alle startup innovative e agli incubatori

certificati, vi possono accedere tutte le imprese di micro, piccole e medie dimensioni, nonché i professionisti iscritti agli ordini professionali, che posseggono i requisiti richiesti.

• Tramite delle “Competition”: sono delle competizioni organizzate da Banche, fondazioni, aziende multinazionali, incubatori, fondi di investimento. Alcune banche concedono alle startup dei finanziamenti attraverso l’erogazione di mutui di importo tra i 30 ed i 250 mila euro. La Startup rimborserà il debito entro 7 anni a meno che la banca non decida di convertire in azioni il finanziamento diventando così azionista dell’impresa.

• Mutui per giovani aziende: canale questo al quale le startup difficilmente fanno ricorso.

CAPITOLO 2: CASO DI STUDIO

2.1 Il dataset

Considerando quanto detto finora, ho potuto individuare una vasta gamma di fattori in grado di determinare il successo o meno di una startup.
Inizialmente ero orientato sul creare più file csv per raccogliere informazioni multiple su alcune startup; ad esempio, avevo pensato di comprendere tutte le nuove tecnologie introdotte da un'azienda o anche pensavo di memorizzare più competitors o più acquisizioni. Dopo una breve consultazione con il professore si è pensato di optare per un file unico, eliminando di fatto le righe multiple relative ai campi nuova tecnologia, competitors e acquisizioni, e mantenendo così solo una singola quella considerata più importante.
Ho creato quindi un dataset con le seguenti feature:

- id: utile a distinguere univocamente un'azienda da un'altra. Inizialmente serviva per collegare più file csv velocemente.
- nome: il nome con cui l'azienda è conosciuta fiscalmente
- settore di mercato: il settore di mercato in cui l'azienda operava o ha operato
- descrizione: la descrizione indica di cosa si occupa l'azienda più in particolare del prodotto che vende
- finanziamenti: la somma degli importi monetari che la startup è riuscita ad accumulare durante i round di finanziamento

- numero investitori: il numero di persone fisiche che hanno contribuito ai finanziamenti ricevuti dalla startup
- valore startup: stima della valutazione monetaria dell'azienda in virtù del suo modello di business del suo fatturato e dei suoi finanziamenti
- valore mercato totale: stima del mercato in cui opera l'azienda attraverso i movimenti di denaro che coinvolgono quel mercato
- numero brevetti: numero di tecnologie introdotte per la prima volta da questa startup
- numero prodotti attivi: numero di prodotti
- stage: round a cui la startup ha ricevuto l'ultimo finanziamento
- media fatturato: media del fatturato ottenuto durante gli anni di attività
- tasso di crescita dip: crescita dei dipendenti rispetto all'anno precedente
- anni di attività: anni passati dopo la creazione dell'azienda
- anno creazione: anno in cui la startup è stata creata
- città: città in cui l'azienda ha la sede legale
- stato: stato in cui l'azienda ha sede
- continente: continente in cui l'azienda ha sede
- budget formazione: importo impiegato per formare i dipendenti a nuove tecnologie o ad un eventuale change management
- nuova tecnologia: tecnologia nuova più importante introdotta dalla startup
- tipo di miglioramento tecnologia: fattore che è stato modificato in seguito all'introduzione di una nuova tecnologia
- incremento tecnologico: miglioramento, in termini percentuali, del fattore descritto in precedenza.

- numero operatori: numero di dipendenti che svolgono l'attività di operatori
- dipendenti ingegneri: numero di dipendenti che lavorano come ingegneri
- dipendenti business: numero di dipendenti che lavorano come operatori di business
- dipendenti vendite: numero di dipendenti incaricati alle vendite del prodotto
- dipendenti design: numero di dipendenti che lavorano per il design del prodotto.
- dipendenti informatica: numero di dipendenti che lavorano come informatici
- dipendenti amministrativo: persone incaricate della parte gestionale amministrativa.
- dipendenti controllo qualità: dipendenti addetti al controllo qualità
- dipendenti ricerca: dipendenti deputati alla ricerca
- dipendenti risorse umane: dipendenti addetti all'assunzione del personale
- dipendenti assistenza: dipendenti deputati all'assistenza clienti
- concorrente: principale competitor dell'azienda che si sta esaminando
- fatturato concorrente: fatturato dell'azienda concorrente

Ho scelto di creare un dataset con 100 aziende di cui 50 fallite e 50 che sono ancora attive. Delle 50 attive la variabile stage può assumere diversi valori a seconda di qual è l'ultimo round di finanziamento ricevuto:
-seed: quando è nella fase seed ovvero sta accumulando i finanziamenti per la realizzazione e lo studio del prodotto

-serie a: il cui obiettivo principale è migliorare il business, consolidare il prodotto e aumentare la cerchia dei clienti
-serie b: l'azienda espande ulteriormente il business, scalando le operazioni e ampliando la propria presenza sul mercato
-serie c: espande ulteriormente le operazioni, entra in nuovi mercati e migliora l'efficienza operativa
-serie d: l'azienda si espande su larga scala, attraverso acquisizioni strategiche o l'espansione internazionale
-ipo: la startup mette a disposizione del pubblico le proprie azioni. L'imprenditore quota in borsa la propria attività per accedere, rapidamente, a finanziamenti necessari per lo sviluppo.
-dead: la startup si dichiara ufficialmente fallita, di conseguenza, non è più operativa

Ho potuto rilevare questo grande quantitativo di dati grazie alle risorse presenti online. L'uso di siti come cbinsight, crunchbase è stato utile per rilevare gran parte dei dati, anche se, non avendo l'account a pagamento, non ho potuto trascrivere tutti i dati che questi strumenti fornivano, proprio perché appunto alcuni contenuti erano riservati solo per coloro che avevano un account premium.

CBInsights raccoglie ed elabora un'enorme quantità di dati provenienti da varie fonti, tra cui comunicati stampa, articoli di notizie, registri di brevetti, dati di finanziamento, e social media.

Come dicevo in precedenza, CB Insights è una piattaforma a pagamento, con accesso a vari livelli di abbonamento. I prezzi variano in base alla profondità delle informazioni e agli strumenti di analisi a cui si desidera accedere. Generalmente, i costi sono significativi e destinati a professionisti o aziende con necessità di analisi approfondite. Quindi purtroppo non ho potuto magari reperire tutte le informazioni di cui avrei avuto bisogno. Nonostante ciò, grazie a CBInsights ho potuto rilevare il numero di investitori, i finanziamenti

ricevuti e la stima della valutazione della startup. Non sempre però tutte queste informazioni erano note, infatti per alcune di esse, le meno rilevanti, queste non erano presenti. Per queste ed altre informazioni ho dovuto attingere a Crunchbase, il portale più autorevole per reperire informazioni relative alle startup. Crunchbase è, infatti, noto per la sua vasta banca dati di informazioni su aziende di tutte le dimensioni, dagli stadi iniziali fino alle grandi corporation. Le informazioni comprendono dettagli su fondatori, team esecutivi, finanziamenti ricevuti, acquisizioni, partner e investitori. Uno dei punti di forza di Crunchbase è la sua capacità di aggregare e aggiornare continuamente i dati, attingendo a fonti pubbliche, segnalazioni dirette da parte degli utenti, e collaborazioni con aziende e istituzioni. Questo gli consente di mantenere un database molto accurato e aggiornato, che riflette le ultime tendenze e sviluppi nel panorama delle startup. La piattaforma offre diverse funzionalità che vanno oltre la semplice raccolta di dati. Ad esempio, consente agli utenti di seguire specifiche aziende, settori, o investitori, ricevendo aggiornamenti in tempo reale sui loro movimenti e attività. Questa capacità di monitoraggio è particolarmente utile per investitori, analisti, e imprenditori che devono prendere decisioni rapide basate su informazioni aggiornate.

Crunchbase è anche una risorsa preziosa per chi è alla ricerca di opportunità di lavoro nel settore delle startup o desidera stabilire connessioni con potenziali partner commerciali. La piattaforma permette di scoprire nuove aziende emergenti, esplorare i loro team e conoscere le opportunità di investimento o collaborazione. Questa funzione ha reso Crunchbase non solo uno strumento per l'analisi, ma anche un punto di incontro per la comunità delle startup.

Ho utilizzato crunchbase spesso per sopperire alle mancanze di CBInsights, ma principalmente per ottenere il numero di prodotti attivi di una startup, un

indicatore chiave che riflette diversi aspetti della salute, della strategia e della fase di sviluppo di una startup.
Quando mi capitava di rilevare differenze importanti tra i due siti verificavo l'attendibilità dei dati su un terzo sito ovvero pitchbook.com. PitchBook raccoglie e analizza un'enorme quantità di dati provenienti da una varietà di fonti, offrendo agli utenti una panoramica completa e dettagliata sui mercati privati. PitchBook fornisce informazioni dettagliate su milioni di aziende private, compresi i loro dati finanziari, round di finanziamento, valutazioni, investitori coinvolti, e dettagli sulle operazioni di M&A. Gli utenti possono esplorare il ciclo di vita delle startup, dal finanziamento iniziale fino all'exit. Anche questo sito aveva bisogno di un abbonamento per l'accesso completo ai dati. Questo sito mi è servito per confrontare per esempio anche il numero di dipendenti con quelli presenti su altri due siti utili alla creazione del dataset: LinkedIn e Growjo.
LinkedIn è una piattaforma di social networking professionale, utilizzata per connettere professionisti, cercare lavoro, reclutare talenti e condividere contenuti aziendali. Fondata nel 2002 e acquisita da Microsoft nel 2016, LinkedIn permette agli utenti di creare profili che funzionano come curriculum digitali, consentendo di mettere in evidenza esperienze lavorative, competenze e obiettivi professionali. È ampiamente utilizzata per networking, sviluppo di carriera, e marketing B2B, oltre a essere uno strumento chiave per i recruiter nella ricerca di candidati qualificati. Io ho utilizzato LinkedIn per reperire il numero dei dipendenti e per ottenere come erano divisi nelle varie mansioni all'interno dell'azienda.
Growjo è una piattaforma che identifica e classifica le aziende in rapida crescita a livello globale, concentrandosi principalmente su startup e aziende emergenti. Utilizza algoritmi che analizzano diversi indicatori di crescita, come l'aumento del personale, il finanziamento ricevuto e la crescita delle entrate,

per creare classifiche e liste delle aziende più promettenti in vari settori. Questo sito è stato utile innanzitutto per confrontare i dati ottenuti in precedenza, ma anche per ricavare l'incremento o decremento dei dipendenti nell'ultimo anno, importante per capire lo stato di salute di una startup, e il fatturato medio delle stesse durante tutto il loro ciclo di vita.

Infine ho fatto uso anche di language model come chatgpt e meta ai.

ChatGPT è un modello di intelligenza artificiale sviluppato da OpenAI, progettato per comprendere e generare testo in linguaggio naturale. Basato sull'architettura GPT (Generative Pretrained Transformer), ChatGPT è in grado di rispondere a domande, creare contenuti, assistere con traduzioni, e sostenere conversazioni su una vasta gamma di argomenti. Addestrato su enormi quantità di dati testuali, ChatGPT utilizza il deep learning per elaborare contesti complessi e produrre risposte coerenti e pertinenti. Viene utilizzato in applicazioni come assistenti virtuali, chatbot, e strumenti di scrittura automatizzata, dimostrando grande versatilità nell'interazione uomo-macchina.

Chatgpt mi è stato utile per colmare le lacune di conoscenza che avevo su argomenti finanziari oltre che uno strumento utile per riassumere in breve la descrizione in breve di un'azienda.

Meta AI è un'intelligenza artificiale molto avanzata, creata da Meta (l'azienda di Facebook, Instagram e WhatsApp), che è in grado di comprendere e generare contenuti come testi, immagini e altro ancora.

L'uso di Meta Ai non è consentito in Italia motivo per cui l'unico modo per poterlo utilizzare era attraverso l'uso di una VPN. Una VPN (Virtual Private Network) è un servizio che crea una connessione sicura e criptata tra il dispositivo dell'utente e Internet. La VPN mi è servita in questo caso per nascondere il mio indirizzo IP, collegandomi ad un server situato negli Stati Uniti, che mi ha poi permesso di accedere a Meta Ai, strumento ancora

indisponibile in Italia. Nel momento in cui io facevo accesso a Meta Ai dal browser del mio computer, attraverso la VPN accedevo ad una moltitudine di server che culminavano con un server presente in America che a sua volta accedeva alla pagina web di cui avevo bisogno. Poi successivamente la pagina veniva riportata sul mio computer attraverso una serie di pacchetti. Per meta ai l'indirizzo IP con cui stavo accedendo al suo servizio era quello del server terminale locato in America per cui l'accesso era consentito.

Proprio Meta è stato uno strumento molto utile per ricavare le informazioni che non sono riuscito a trovare online. Infatti, il language model mi ha permesso di fare una stima di dati che altrimenti non avrei potuto ricavare come il numero dei brevetti, il budget allocato per la formazione, l'impatto della nuova tecnologia introdotta, la principale concorrente e il valore totale del mercato a cui un'azienda fa riferimento.

Tutte queste informazioni sono state raccolte in un file .csv, file solitamente usato per collezioni di dati numerose in forma di dataset. Generalmente è suddiviso in questa maniera:

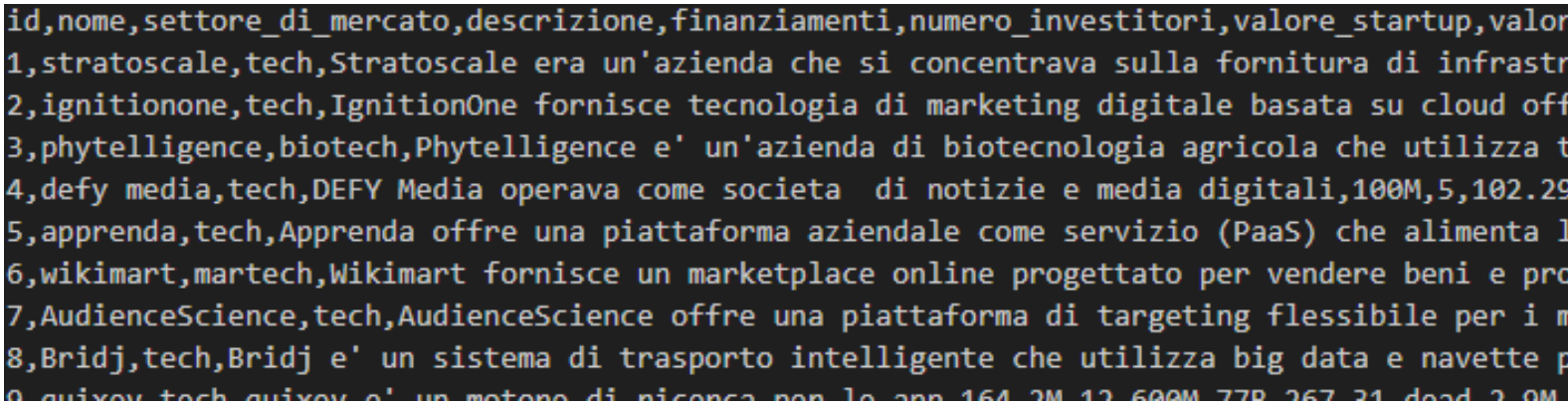
Nella prima riga è presente i nomi di tutte le caratteristiche, ovviamente comuni ad ogni tupla del dataset, separate da una virgola.

In seguito sono rappresentati i valori di ogni feature, corrispondente a quella determinata posizione, per ogni tupla.

Ogni riga avrà quindi valori diversi separati da una virgola e, una volta

terminato l’inserimento di tutti i dati relativi ad una singola istanza, a capo verrà appunto indicata la prossima.

Figura 3, un piccolo estratto del dataset oggetto di caso di studio



2.2 Il preprocessing

Dopo aver creato il dataset e averlo riempito con tutti i dati ricavati dai siti visti in precedenza, ho dovuto elaborare i dati, rimuovere eventuali ambiguità e renderlo compatibile alla lavorazione.
Ho dovuto quindi scrivere un nuovo file di python, il linguaggio su cui verrà sviluppato tutto il progetto, per preparare il dataset all'elaborazione. In prima istanza ho importato la libreria pandas, fondamentale per l'analisi e l'elaborazione dei dati. Le strutture dati usate principalmente da pandas sono due:
-series: una struttura dati simile all'array che può contenere dati di un solo tipo
-dataframe: una tabella bidimensionale con righe e colonne
Inoltre pandas supporta l'importazione e l'esportazione dei dati da e verso vari formati come .csv, json, Excel ed sql, e risulta particolarmente efficace quando si devono caricare dataset di grandi dimensioni.
Ora nel caso di studio pandas è servito particolarmente a caricare il dataset su un dataframe attraverso il comando data = pandas.read_csv('csv/main.csv').
Una volta ottenuto il dataset sotto forma di dataframe, ho dovuto fare una revisione dei dati inseriti in modo tale da poter essere elaborati successivamente. In prima battuta, considerando che tutti i campi sono di tipo object, ho formattato i dati ad un tipo di dato compatibile con il compilatore. Vale a dire che ho effettuato un cast di tipo per i campi numerici di tipo intero, un cast per i campi di tipo float e un altro per i campi di tipo category. In particolare, i campi finanziamenti, valore startup, valore mercato totale, media fatturato, tasso di crescita dipendenti, budget formazione, incremento tecnologico, fatturato concorrente, valore startup, valore mercato totale sono di tipo float. Il tipo category è usato solo per indicare il campo nuova tecnologia mentre i restanti relativi ai dipendenti e alla loro suddivisone sono

di tipo intero.
Nel dataset sono presenti delle lettere nei campi numerici tese ad indicare la grandezza del numero. Sono state usate le lettere K, M e B e in precedenza la cifra numerica per indicare rispettivamente le migliaia, i milioni e i miliardi. Quindi ho sostituito le K con tre zeri, le M con sei zeri e le B con nove zeri, nei campi in cui esse comparivano. Naturalmente questi indicatori venivano usati per indicare parametri come i finanziamenti o il valore startup e non campi con valori numerici più piccoli come il numero dei dipendenti e la loro suddivisione nei rispettivi settori.
Mentre nel dataset alla voce tasso di crescita dipendenti ponevo il simbolo percentuale(%) al termine della scrittura del numero, nel dataframe lo rimuovo sempre per la medesima esigenza ovvero rendere i dati comprensibili al compilatore.
Ho anche sostituito le stringhe del campo 'stage' con valori numerici, ovvero 'dead' con il valore 0, 'seed' con il valore 1, 'serie a' con il valore 2, 'serie b' con il valore 3, 'serie c' con il valore 4, 'serie d' con il valore 5, e 'ipo' con il valore 6.
Ho poi valutato inutile, al fine del caso di studio, alcuni campi come id, che è solo un identificativo e non incide sul successo o meno di una startup; il nome, la descrizione e infine anche il nome della concorrente. Ho deciso di rimuovere anche la città dove viene fondata perché sostanzialmente non è un parametro incisivo in quanto in città appartenenti ad uno stesso stato vigono leggi molto simili se non identiche per quanto riguarda le imprese di questo tipo.
Inizialmente avevo escluso anche continente e stato in quanto il mio dataset risulta essere non uniforme sotto questo punto di vista: le aziende americane prevalgono e sono per la maggior parte le uniche di cui è possibile trovare informazioni circa il fallimento. Ciò significa che molte aziende fallite sono

Americane proprio perché non è facile trovare informazioni riguardo startup ‘decedute’ in altri continenti.
Chiaramente questo problema avrebbe potuto causare overfitting, un fenomeno che avviene quando il sistema è troppo dipendente dai dati che vengono usati per addestrarlo.
Ho verificato che aggiungendo continente e stato le previsioni erano più accurate con un significativo aumento delle prestazioni di tutti i classificatori usati dall’esperimento.
Quindi anche se il dataset risulta ‘drogato’ di aziende americane fallite, le previsioni rimangono accurate e anzi migliorano, anche perché evidentemente nelle aziende che operano ancora, c’è una larga parte di aziende americane. Ho realizzato quindi un grafico a torta per capire meglio questa situazione ed infatti nonostante il 57% delle aziende americane sia fallita, c’è un 43% che invece è ancora attiva. L’analisi di questo grafico ha inciso quindi sulla mia scelta di considerare anche la feature del continente.

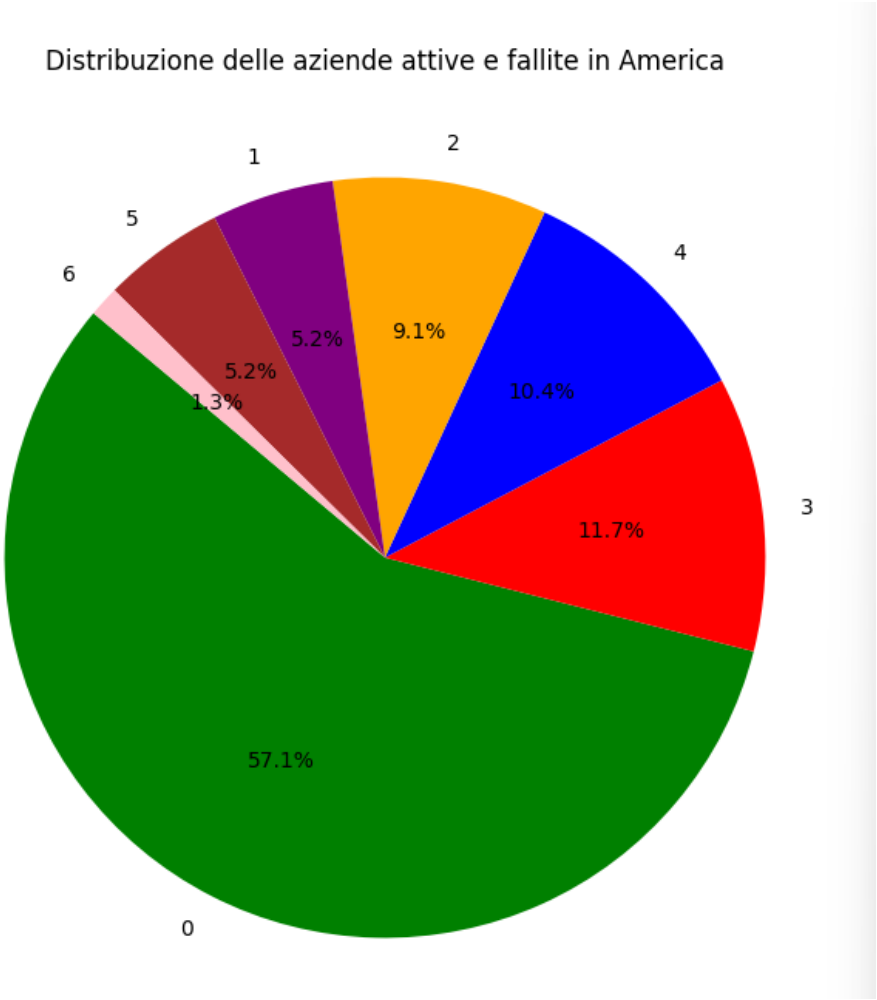


Figura 4: grafico a torta rappresentante i valori che assumono le aziende americane nel campo stage dove 0 indica l'azienda fallita e gli altri valori indicano le aziende attive

2.3 L’interfaccia grafica

Dopo aver sostanzialmente definito il dataframe su cui andrò ad operare, è la volta dell’interfaccia grafica che permetterà all’utente di inserire i dati relativi alla digital twin, un’azienda gemella di cui vogliamo simulare l’andamento. Una digital twin è una startup fittizia creata con l’intento di simulare un’azienda esistente per prevederne il suo funzionamento in un contesto reale. Lo scopo di questa simulazione è prevedere il funzionamento della startup per correggere potenziali errori che potrebbero essere fatali per la sopravvivenza dell’azienda stessa. In alternativa si potrebbe simulare l’andamento di un’azienda vera o fittizia in condizioni di change management. Le modifiche organizzative all’interno di una startup sono chiamate change management e sono il principale motivo per cui è stata realizzata la seguente tesi.

Il change management è formato da approcci e metodi utilizzati per preparare, supportare e aiutare le persone, i team e le organizzazioni a realizzare il cambiamento. Queste strategie mirano a minimizzare la resistenza al cambiamento e a garantire una transizione fluida e di successo verso nuovi processi, tecnologie, strutture organizzative o strategie aziendali. Lo scopo del change management è fornire informazioni chiare e tempestive sul cambiamento, chiarendo il motivo per il quale viene fatto e come verrà modificato il lavoro degli individui. Pertanto, è fondamentale coinvolgere fin da subito i dipendenti ascoltando le loro opinioni e preoccupazioni circa ciò che verrà modificato nel futuro prossimo, così da limare aspetti controversi prima ancora che si presentino. Insieme al cambiamento di assetto aziendale o del business model, è necessario formare il personale con corsi di formazione per fornire nuove competenze e conoscenze utili a fronteggiare le difficoltà del nuovo lavoro.

I leader devono dimostrare impegno e supporto per il cambiamento, fungendo da modello di comportamento, cercando di riconoscere ed identificare le resistenze al cambiamento e di conseguenza trovare dei modi per ridurre queste difficoltà.

Per poter inserire le caratteristiche relative alla digital twin abbiamo bisogno di un'interfaccia grafica che possa acquisire ciò che l'utente predilige.
La libreria che ho utilizzato per realizzare l'interfaccia grafica è tkiner. Tkiner è una delle librerie GUI più utilizzate per Python grazie alla sua integrazione nativa e alla facilità d'uso.
Tkinter offre una varietà di widget che possono essere usati per costruire interfacce grafiche complesse. In particolare quelle che verranno usate per questo progetto sono i bottoni, le etichette o label, le entry cioè campi per inserire il testo e i menu a tendina.
Per poter ordinare i vari widget nella pagina ho usato invece il metodo grid(), che permette di suddividere la pagina in una griglia con righe e colonne.
Nel caso di mancato inserimento o inserimento scorretto ho gestito l'eccezione attraverso un pop-up creato attraverso messagebox, un modulo interno alla libreria tkiner. Quando quindi si verificava un mancato inserimento o un inserimento scorretto veniva segnalato rispettivamente con i messaggi "Il campo non è stato inserito correttamente" e "si è verificato un errore durante l'inserimento dei dati".
Questo modulo l'ho usato anche nel caso in cui l'utente chiudesse la finestra attraverso la "X", per evitare chiusure involontarie. Il messaggio ad esso associato è un messaggio di conferma ovvero "sei sicuro di voler uscire?" .
Nell'interfaccia usata, per acquisire i dati della digital twin in condizioni di change management, sono presenti dei menu a tendina per le features categoriche dove sono elencati tutti i possibili valori presenti nel dataframe

per quella determinata feature. Non è possibile fare una previsione per feature categoriche nuove in quanto il dataframe non è abbastanza esteso e i modelli di intelligenza artificiale non sono così sofisticati da comprendere la correlazione tra l'inserimento di una nuova feature categorica e il successo o meno di un'azienda. Infatti i modelli si limitano a convertire la feature categorica in un numero e, in base a quante volte è presente, a calcolare l'incisività per la previsione della variabile target.
Per quanto riguarda le features numeriche come il budget formazione, i finanziamenti, il valore della startup, il valore mercato totale e la media fatturato è necessario inserire la cifra e in seguito anche l'ordine di grandezza. Infatti per facilitare la vita all'utente associato al campo usato per l'inserimento sono presenti dei bottoni, chiamati radio button, utili alla scelta dell'ordine di grandezza della cifra inserita. B indica i miliardi, M i milioni e k le migliaia quindi rispettivamente nove, sei e tre zeri che l'utente non avrà bisogno di inserire. Tutte le feature numeriche sono considerate float quindi è possibile inserire numeri con la virgola. Per le entry, ovvero le caselle per l'inserimento, relative agli aumenti percentuali è necessario inserire solo la parte numerica senza il simbolo percentuale.

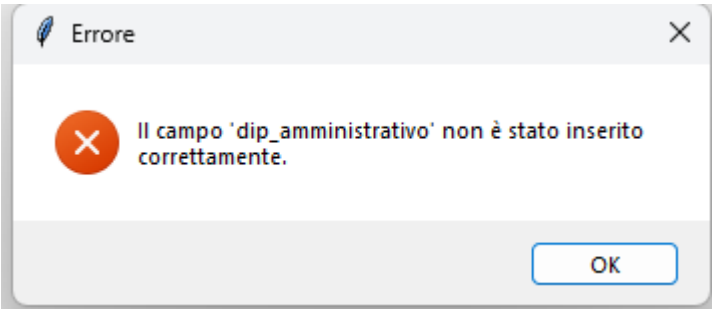


Figura 5: popup che notifica i campi non inseriti correttamente

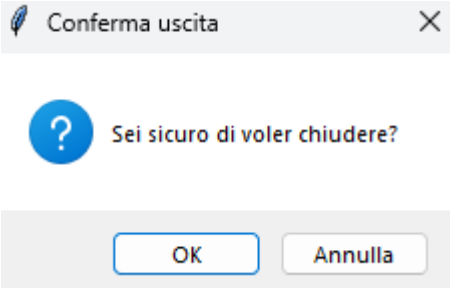


Figura 6: popup di conferma nel caso in cui si voglia chiudere la finestra


Inserimento Dati
Inserimento dati Digital Twin
inserire valore_startup
B
M
K
inserire finanziamenti
B
M
K
inserire budget_formazione
B
M
K
inserire fatturato concorrente in miliardi
altrimenti se la cifra è inferiore al miliardo
scrivere come indicato nella parentesi(0.1)
inserire media_fatturato
B
M
K
inserire settore_di_mercato
tech
inserire valore_mercato_totale
B
M
K
inserire nuova_tecnologia
virtualizzazione del server
Indicare con k le migliaia, con m i milioni e con b i miliardi
inserire tipo_di_miglioramento_tecnologia
costi
inserire incremento_tec in %
senza inserire il simbolo
inserire continente
Asia
inserire stato
Israel
inserire tasso_di_crescita_dip in %
senza inserire il simbolo
inserire numero_operatori
inserire dip_ingegneri
inserire dip_business
inserire dip_vendite
inserire dip_design
Invia
inserire dip_informatica
inserire dip_amministrativo
inserire dip_controllo_qualità
inserire dip_ricerca
inserire dip_risorseumane
inserire dip_assistenza

Figura 7: interfaccia grafica per l'inserimento della digital twin

2.4 I classificatori

Una volta realizzata l'interfaccia e dopo aver acquisito i dati, è necessario addestrare i classificatori. I classificatori usati per l'esperimento saranno l'XGBoost, il CatBoost e il Random Forest. Si è deciso di optare per i due classificatori legati al Gradient Boosting perché hanno dato risultati consistenti e rilevanti per la buona riuscita dell'esperimento. Riguardo al random forest, si è deciso di usarlo come modello vista la sua fama e data la sua versatilità su dati eterogenei.

Il Gradient Boosting è una tecnica di apprendimento supervisionato utilizzata principalmente per problemi di regressione e classificazione. Si basa sull'idea di costruire un modello predittivo forte a partire da una combinazione di modelli deboli, solitamente alberi di decisione.

Il Gradient Boosting costruisce modelli in modo sequenziale. Ogni modello cerca di correggere gli errori commessi dai modelli precedenti. La valutazione della correttezza dei modelli precedentemente generati è affidata alla loss function. Questa funzione misura quanto le predizioni del modello si discostano dai valori reali.

Generalmente, gli alberi di decisione sono utilizzati come modelli deboli perché sono alberi poco profondi, di una profondità che può variare da 3 a 5. Considerati singolarmente non sono molto performanti ma, combinati insieme, possono produrre modelli molto potenti.

Infatti, il modello finale risulta essere una somma pesata dei modelli deboli. Ogni nuovo modello è addestrato per ridurre l'errore residuo del modello

combinato precedentemente. $\hat{y}_i^{(m)}$

L'algoritmo può essere descritto in questi passaggi:

Inizializzazione: Si inizia con una predizione iniziale (ad esempio, la media dei valori di output nel caso della regressione).

Calcolo del residuo: Per ogni iterazione, si calcola il residuo $r_i^{(m)}$ come la differenza tra il valore osservato y_i e la predizione attuale del modello

Addestramento del modello debole: Si addestra un modello debole utilizzando i residui come target.

Aggiornamento del modello: Il modello finale è aggiornato aggiungendo il nuovo modello debole, moltiplicato per un fattore di apprendimento v(learning rate) che riduce l'impatto di ogni singolo modello debole.

Ripetizione: Si ripetono i passaggi dal 2 al 4 per un numero prefissato di iterazioni.

Per la buona riuscita dell'algoritmo è importante impostare i parametri per l'algoritmo descritto in precedenza nella maniera corretta:

Il numero di alberi, ovvero il numero di modelli deboli da addestrare, non deve essere troppo elevato perché in tal caso porterebbe ad un problema di overfitting, mentre se fosse troppo basso potrebbe creare un modello sottostimato.

Il learning rate o tasso di apprendimento stabilisce quanto rapidamente il modello apprende aggiungendo nuovi alberi. Se il learning rate è basso il nuovo albero generato si discosterà poco da quello generato in precedenza, viceversa, con un learning rate alto ci sarà una differenza tangibile tra i due alberi, permettendo al modello di adattarsi più rapidamente, ma aumentando anche il rischio di overfitting.

L'ultimo parametro è il minimum sample split che definisce il numero di campioni richiesti per suddividere un nodo dell'albero. Se un nodo ha un

numero di campioni inferiore a quello definito da questo valore diventa foglia. Viceversa, se ha un numero di campioni pari o superiore a questo valore è considerato un nodo normale. Ovviamente più sarà alto il valore, meno definito sarà l'albero creato e viceversa.
Solitamente il gradient boosting viene usato in contesti dove è necessario produrre modelli molto accurati, in casi in cui c'è bisogno di gestire dei dati mancanti. Inoltre, è utile sia in problemi di regressione che di classificazione, ma non è adatto per dataset troppo grandi perché richiede un'elevata potenza di calcolo, e può facilmente innescare l'overfitting.

2.4.1 I classificatori catBoost e XGBoost

Ora focalizziamoci sui classificatori usati per le predizioni: CatBoost e XGBoost sono due popolari librerie di gradient boosting utilizzate per problemi di classificazione e regressione. Entrambe offrono ottimizzazioni avanzate rispetto all'implementazione standard del gradient boosting e sono note per la loro alta efficienza e performance.
CatBoost (Categorical Boosting) è una libreria sviluppata da Yandex che si distingue per la sua capacità di gestire in modo efficiente le caratteristiche categoriche senza bisogno di pre-elaborazione. È progettata per essere facile da usare, veloce e robusta contro l'overfitting. CatBoost permette di lavorare direttamente con i dati categorici in quanto ha integrato uno strumento di conversione chiamato target statistic. Questa tecnica avanzata, usata nel preprocessing, implica la sostituzione di variabili categoriche con valori numerici ricavati da un'altra caratteristica associata alla caratteristica precedente.
Ad esempio, prendiamo due features del nostro dataset nuova_tecnologia e incremento_tec sono due feature dipendenti tra loro in quanto l'incremento varia a seconda di quale nuova tecnologia viene introdotta. Si sceglie un target

con cui procedere tra media, mediana o tendenza e lo si calcola attraverso la feature numerica ovvero incremento tecnologico. Successivamente si sostituisce la feature categorica con la somma tra il valore associato della feature numerica con il target scelto.
Questo metodo comporta diversi vantaggi:
-Le target statistics riducono l'alta cardinalità delle caratteristiche categoriche, trasformandole in valori numerici più gestibili.
-La sostituzione delle variabili categorie con quelle statistiche può aiutare i modelli a captare meglio le relazioni tra le caratteristiche e il target.
-Utilizzando statistiche derivate dai dati, anziché valori arbitrari o indici, si riduce il rischio di overfitting.
Per prevenire l'overfitting il catboost ha integrato anche la k-fold cross validation.
La K-Fold Cross Validation è una tecnica di validazione incrociata utilizzata per valutare la capacità predittiva di un modello di machine learning e garantire che esso generalizzi bene su tutti i dati e non solo con una parte di essi.
Questa tecnica è particolarmente utile per evitare l'overfitting e per fornire una stima più accurata delle performance del modello.
inizialmente si divide il dataset in k parti approssimativamente della stessa dimensione, chiamati fold.
Il modello viene addestrato k volte e, in ogni iterazione, viene sostituito il fold utilizzato per il set di test e i restanti k-1 fold vengono usati come set di addestramento.
I vantaggi sono molteplici:

- il modello risulta essere robusto e ben addestrato in quanto ogni osservazione viene usata sia per l'addestramento che per il test

- viene ridotto il bias poiché ogni dato viene usato come test almeno una volta e anche perché vengono effettuate diverse iterazioni. Di conseguenza viene ridotta anche la varianza

Ovviamente ci sono anche degli svantaggi legati al costo computazionale, perché richiede k iterazioni che possono essere computazionalmente costosi specialmente per modelli costosi o dataset grandi. L'altro svantaggio è legato al valore di k; maggiore sarà il valore di k maggiori saranno le iterazioni e di conseguenza maggiore sarà il costo computazionale.
Il catboost è ottimizzato per essere veloce e scalabile e quindi adatto a dataset grandi, oltre che è possibile trasferire la mole di calcolo sulla gpu per l'addestramento.
A differenza del catboost l' XGBoost (Extreme Gradient Boosting) è una libreria sviluppata da Tianqi Chen. Questo modello deve la sua popolarità alle sue eccellenti performance in competizioni di machine learning e la sua versatilità. XGBoost è anche noto per essere estremamente efficiente, flessibile e utilizzabile in vari contesti.
L'XGBoost utilizza il parallelismo a livello di thread per essere più veloce ed efficiente nel processo di addestramento. Il parallelismo a livello di thread è una tecnica di programmazione che consente di eseguire più thread simultaneamente all'interno di un singolo processo per migliorare l'efficienza e le prestazioni delle applicazioni. I thread sono piccole unità di esecuzione che condividono lo stesso spazio di indirizzamento del processo principale, il che permette loro di comunicare e condividere risorse più facilmente rispetto a processi separati. Questa tecnica comporta vantaggi computazionali, in quanto i core della CPU vengono sfruttati al meglio attraverso l'esecuzione in parallelo, di conseguenza ne risente anche il tempo di esecuzione, nettamente inferiore rispetto a un'esecuzione seriale. Di contro i thread vengono realizzati

attraverso una programmazione complessa che potrebbe causare problemi di sincronizzazione. Questi problemi avvengono quando più thread accedono alle stesse risorse condivise simultaneamente, come strutture dati o lo stesso processore, senza un adeguato coordinamento. Senza una corretta sincronizzazione, l'accesso concorrente può portare a comportamenti indesiderati, risultati errati e bug difficili da rilevare e risolvere. I problemi più comuni che si verificano sono i seguenti:

- Race condition: quando due o più thread accedono ad una risorsa condivisa contemporaneamente e il risultato finale dipende dall'ordine in cui i thread accedono a quella risorsa. Questo può portare a comportamenti imprevedibili e risultati non corretti.
- Il deadlock si verifica quando due o più thread rimangono bloccati in attesa di risorse che sono tenute l'uno dall'altro. Nessuno dei thread può procedere, portando l'applicazione a uno stato di stallo.
- Il livelock si verifica quando due o più thread continuano a cambiare stato in risposta agli stati degli altri senza mai progredire. Sebbene i thread non siano bloccati, non riescono comunque a portare a termine il loro lavoro.
- La starvation si verifica quando un thread non riesce ad accedere alle risorse necessarie per proseguire la sua esecuzione perché altre risorse sono continuamente date a thread prioritari.

Per evitare questi problemi, vengono utilizzati diversi meccanismi di sincronizzazione che assicurano che i thread accedano alle risorse condivise in modo controllato e coordinato.
Una di queste è la mutua esclusione che permette a un solo thread alla volta di accedere a una risorsa critica. Quando un thread acquisisce un lock, quindi

una risorsa critica condivisa, altri thread devono attendere finché il lock non viene rilasciato.
Un altro meccanismo per gestire la sincronizzazione è il semaforo. A differenza della mutua esclusione, i semafori permettono a un numero limitato di thread di accedere contemporaneamente alla risorsa. Questo numero è definito dal contatore del semaforo. Quando un thread acquisisce il semaforo, il contatore viene decrementato. Quando un thread rilascia il semaforo, il contatore viene incrementato. Un thread che desidera accedere alla risorsa chiama il metodo `acquire()` del semaforo. Se il contatore è maggiore di zero, il thread decrementa il contatore e procede. Se il contatore è zero, il thread viene bloccato finché il contatore non diventa positivo. Quando un thread ha terminato di utilizzare la risorsa, chiama il metodo release del semaforo, incrementando il contatore e permettendo ad altri thread bloccati di procedere. Esiste anche una variante del semaforo che è il semaforo binario in cui solo un thread alla volta può accedere ad una risorsa.
Una barriera,invece, è un meccanismo di sincronizzazione utilizzato per coordinare l'esecuzione di un insieme di thread. Serve a far sì che un gruppo di thread si blocchi in un punto determinato del programma finché tutti i thread del gruppo non abbiano raggiunto quel punto. Solo quando tutti i thread sono arrivati alla barriera, essi possono proseguire la loro esecuzione. Questo è utile per garantire che tutti i thread abbiano completato una fase di lavoro prima di procedere alla fase successiva.
Le condition variables (variabili di condizione) sono meccanismi di sincronizzazione utilizzati per consentire ai thread di attendere che determinate condizioni vengano soddisfatte. Sono utilizzate in combinazione con un mutex per coordinare l'esecuzione dei thread in base a condizioni specifiche. Il mutex protegge l'accesso alla variabile di condizione e ai dati condivisi associati. Un thread può bloccare una condition variable invocando

la funzione `wait()`. Durante l'attesa, il mutex viene rilasciato, permettendo ad altri thread di acquisire il mutex e modificare la condizione. Quando la condizione è soddisfatta e la variabile di condizione viene notificata, il thread viene svegliato e il mutex viene riacquisito. Un thread che modifica la condizione attende la notifica ad uno o più thread in attesa sulla variabile di condizione. Questo può essere fatto usando le funzioni notify_one() o notify_all().

Tornando all'Xg boost, esso utilizza un'altra tecnica per prevenire l'overfitting che si chiama regolarizzazione.

XGBoost utilizza diverse forme di regolarizzazione per controllare la complessità del modello e prevenire l'overfitting:

La regolarizzazione L1 (Lasso) penalizza la somma dei valori assoluti dei coefficienti delle caratteristiche (features).

La regolarizzazione L2(Ridge) Penalizza la somma dei quadrati dei coefficienti delle caratteristiche.

Il termine di regolarizzazione sulla complessità dell'albero in XGBoost è un meccanismo che penalizza la complessità dei singoli alberi. Questo termine agisce come un deterrente per la crescita incontrollata degli alberi, assicurando che il modello rimanga semplice e generalizzabile. Gli alberi con molte foglie vengono penalizzati in quanto possono catturare rumore nei dati di addestramento. Anche gli alberi aventi foglie con pesi elevati vengono penalizzati.

La formula della regolarizzazione è la seguente:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

dove T è il numero di foglie dell'albero, wj sono i pesi delle foglie e γ e λ sono i parametri di regolarizzazione.

Di seguito i parametri spiegati:

omega: Controlla la regolarizzazione L1 (Lasso). Valori più grandi causano una maggiore varietà nei pesi.
lambda: Controlla la regolarizzazione L2 (Ridge). Valori più grandi causano una riduzione dei pesi, ma li mantengono non zero.
gamma: Controlla la regolarizzazione sulla complessità degli alberi. Valori più grandi rendono il modello più conservativo riducendo la crescita degli alberi.
Un altro strumento di regolarizzazione per quanto riguarda l'Xgboost è l early stopping.
L'early stopping monitora le prestazioni del modello su un set di validazione (validation set) durante l'addestramento e interrompe l'addestramento quando le prestazioni non migliorano più o iniziano a peggiorare. Questo punto è considerato il punto ottimale per fermare l'addestramento, in quanto oltre questo punto il modello tende ad adattarsi troppo ai dati di addestramento, portando a un overfitting.
Il modello viene addestrato iterativamente per un numero predefinito di boosting round e dopo ogni iterazione le prestazioni del modello vengono valutate secondo una metrica di valutazione.
Le metriche in grado di valutare un modello potrebbero essere per esempio, l'errore quadratico medio, l'accuratezza o l'AUC. Se le prestazioni sul validation set non migliorano per un certo numero di iterazioni consecutive (patience), l'addestramento viene interrotto. Chiaramente il vantaggio più grande apportato da questa tecnica è la prevenzione dell'overfitting in quanto si interrompe l'addestramento prima che il modello inizi ad adattarsi troppo ai dati con cui viene addestrato e riduce anche il tempo di addestramento fermando il processo non appena il miglioramento si stabilizza, evitando iterazioni inutili.
Mentre il catboost gestisce le caratteristiche in modo nativo ed automatico, l' XG Boost ha bisogno di una pre-elaborazione delle feature categoriche. Nel

caso di studio si è usato come convertitore l'One Hot encoding. L'One-Hot Encoding è una tecnica di preelaborazione dei dati utilizzata per convertire le variabili categoriali in un formato numerico che può essere utilizzato dagli algoritmi di machine learning. Questa tecnica rappresenta ogni valore di una variabile categorica come un vettore binario univoco.
Gli algoritmi di machine learning generalmente lavorano meglio con dati numerici. Le variabili categoriche hanno invece valori discreti come "tech", "biotech"," fintech" e "martech", e devono essere convertite in una forma che gli algoritmi possano utilizzare. L'One-Hot Encoding è una delle tecniche più comuni per fare questa conversione. Il funzionamento dell'One hot encoding è il seguente:
consideriamo per semplicità la feature categorica con il minor numero di valori possibili, in questo caso settore di mercato, che ha 4 valori diversi ovvero tech, biotech, martech e fintech. L'One-Hot Encoding trasforma questa variabile in quattro variabili binarie (dummy variables), una per ciascuna categoria.
Esempio:

Settore di mercato	Tech	Martech	Fintech	biotech
Tech	1	0	0	0
Martech	0	1	0	0
Fintech	0	0	1	0
Biotech	0	0	0	1

Tornando alle differenze tra cat boost e xg boost entrambi sono veloci e offrono la possibilità di usare la GPU per alleggerire i processi sul processore. Il catboost può usare la gpu per l'addestramento mentre l xg boost la può usare per realizzare il parallelismo in fase di esecuzione.

Per quanto riguarda la prevenzione dell'overfitting il catboost include tecniche come il cross validation e il gradient boost ordinato per ridurlo, mentre l'xg boost utilizza la regolarizzazione.
Il catboost è certamente più indicato quando ci sono tante variabili categoriche, dato che ha un sistema integrato per la conversione, mentre l'XG Boost richiede un po' più di lavoro di preelaborazione ma offre grande flessibilità e controllo sui parametri.

2.4.2 Il classificatore random forest

Il Random Forest è un algoritmo di machine learning utilizzato per problemi di classificazione e regressione. È basato su una collezione di alberi decisionali, da cui il nome "foresta". L'idea centrale è quella di creare un modello più robusto e accurato combinando le previsioni di molti alberi decisionali, ciascuno dei quali viene costruito in modo leggermente diverso.
Gli alberi decisionali sono modelli che suddividono ripetutamente i dati in sottoinsiemi basati su caratteristiche specifiche, creando una struttura ramificata dove ogni nodo rappresenta una decisione basata su un attributo dei dati. Un singolo albero decisionale, sebbene potente, tende ad essere molto soggetto all'overfitting, soprattutto quando è profondo e complesso. La Random Forest affronta questo problema combinando i risultati di molti alberi decisionali, ognuno dei quali viene addestrato su un campione casuale del set di dati originale. Questa tecnica è nota come bagging (Bootstrap Aggregating). Ogni albero viene costruito utilizzando un campione casuale, i cui dati potrebbero essere stati usati in parte già per addestrare altri alberi.
Per ogni nodo, logicamente, viene considerato solo un sottoinsieme casuale di caratteristiche, anziché tutte le caratteristiche disponibili.

Una volta generati tutti gli alberi si effettua la previsione considerando tutti gli alberi. Per i problemi di classificazione, come quello del caso di studio, la previsione avviene attraverso la votazione a maggioranza: ogni albero predice una classe, poi la classe più “gettonata” diviene la previsione finale. Per i problemi di regressione la previsione finale è la media delle previsioni di tutti gli alberi.

I vantaggi della random forest sono i seguenti:

- Robustezza contro l'overfitting: Combinando molti alberi, la Random Forest tende a generalizzare meglio rispetto a un singolo albero decisionale.
- Buona accuratezza: È spesso uno degli algoritmi più accurati per una vasta gamma di problemi di classificazione e regressione.
- Gestione delle caratteristiche: Può gestire dati con molte caratteristiche e non richiede molta preelaborazione (ad esempio, scaling delle variabili).
- Stima dell'importanza delle caratteristiche: La Random Forest può valutare l'importanza delle diverse caratteristiche nei dati, fornendo informazioni utili su quali variabili influenzano maggiormente il risultato.

I limiti sono invece i seguenti:

- Maggiore complessità e tempi di calcolo: A differenza di un singolo albero decisionale, la Random Forest richiede più risorse computazionali, sia in termini di tempo che di memoria.
- Interpretabilità ridotta: Anche se ogni albero decisionale è interpretabile, combinare centinaia o migliaia di alberi rende il modello finale meno trasparente.