UNIVERSITÀ DEGLI STUDI DI BARI “ALDO MORO”

DIPARTIMENTO DI INFORMATICA

Corso di Laurea in Informatica

Tesi di laurea in

Modelli e metodi per la sicurezza delle applicazioni

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

INDICE

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 focalizzerà sulla startup, percorrendo le definizioni che sono state attribuite dai molti studiosi di questa materia, esaminando i vari e differenti tipi, facendo una panoramica su quello

che è il ciclo di vita, il ciclo di finanziamento e i motivi che portano al fallimento delle stesse.

Nel secondo 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.

Nel terzo 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.
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

1.1Definizione

Il luogo principale che viene in mente quando si parla di Startup è la California, più precisamente la Silicon Valley, riconosciuta ancora oggi per essere la terra delle opportunità. Infatti, nel 1957 nacque proprio lì la prima vera startup chiamata Fairchild dall'idea del fisico americano Shockley. Egli insieme ad altri due scienziati erano freschi vincitori del premio Nobel per uno studio sui transistor, quando a Shockley venne l'idea di lasciare tutto e fondare una propria attività proprio a Montain View in California. Nessuno dei suoi colleghi ebbe il coraggio di seguirlo e quindi fu costretto a reclutare un team di ingegneri neolaureati che ben presto segneranno un'epoca. Shockley era solito avere dei comportamenti stravaganti come il controllo delle telefonate o l'uso della macchina della verità in situazioni non troppo importanti. I dipendenti, nonostante ciò, riuscivano a sopportare il proprio capo ma quando Shockley decise di abbandonare li studi sul silicio parte del personale non si arrese. I giovani ingegneri avevano fiutato le enormi potenzialità del materiale e Montain View rappresentava terra fertile per le loro sperimentazioni perché non erano vincolati dalle restrizioni imposte dalle comunità fisiche e ingegneristiche tradizionali. Fu così che i ribelli anche chiamati "the traitorous eight" fondarono Fairchild Semiconductor e la conseguente prosperità per quella terra desolata in California che da quel giorno verrà chiamata Silicon Valley.

Il termine "startup" evoca l'immagine di un'impresa creativa avviata da giovani visionari, ma può anche evocare grandi multinazionali tecnologiche come Facebook, Google, Microsoft. In effetti ciò che accomuna tutti i paperoni come Bill Gates, Larry Page e Mark Zuckemberg è lo status sociale ottenuto dopo il successo, il punto di partenza, solitamente un garage, ma

anche la tipologia di azienda che hanno fondato. Molti esperti hanno tentato di definire teoricamente ciò che loro hanno creato praticamente ma ad oggi non c'è una definizione univoca, solo delle idee simili ma con sfumature diverse.

Secondo Eric Ries, autore di *The Lean Startup*, una startup è un'organizzazione progettata per creare nuovi prodotti o servizi in condizioni di estrema incertezza. Neil Blumenthal, cofondatore di Warby Parker, descrive una startup come un'azienda che si occupa di risolvere problemi complessi, dove la soluzione non è ovvia e il successo non è garantito. Paul Graham, CEO di Y Combinator, vede la crescita come l'elemento chiave di una startup, sostenendo che anche un'azienda con cinque anni di attività può ancora definirsi tale, mentre dieci anni potrebbero cominciare a essere eccessivi.

Adora Cheung sottolinea che una startup nasce quando le persone decidono di unirsi a un'azienda, rinunciando alla stabilità per la promessa di una crescita esponenziale futura. Dal canto suo, il dizionario Merriam-Webster definisce una startup come un'attività in fase di sviluppo o una nuova impresa commerciale. L'American Heritage Dictionary aggiunge che si tratta di un'impresa che ha appena iniziato a operare.

Paul Graham evidenzia inoltre che dopo circa tre anni di attività, molte startup smettono di essere considerate tali. Ciò può avvenire per diversi motivi: un'acquisizione, un fatturato superiore a 20 milioni di dollari, l'aumento dei dipendenti o l'uscita dei fondatori. Paradossalmente, una startup può perdere la sua etichetta di startup quando diventa redditizia.

Una criticità di questa definizione è che potrebbe includere piccole imprese familiari con generazioni di attività, o imprese nate per espandersi globalmente.

Ciò che emerge da tutte queste definizioni è che l'elemento centrale di una startup è la capacità di crescere rapidamente. Come sottolinea Graham, l'obiettivo è la crescita rapida e la capacità di scalare, che differenzia una startup da una piccola impresa tradizionale.

Alyson Shontell, caporedattrice di Business Insider US, descrive la startup come una "montagna russa emotiva", capace di portare a grandi successi o grandi fallimenti. Secondo Shontell, chi fonda una startup è spesso una persona brillante e un po' folle, pronta a sacrificare la stabilità per cambiare il mondo, anche a costo di anni di lavoro stressante.

La definizione più accettata è quella di Steve Blank, autore di *The Startup Owner's Manual*, secondo cui una startup è un'organizzazione temporanea creata per trovare un modello di business ripetibile e scalabile. Blank individua tre elementi fondamentali che caratterizzano una startup:

1. Temporaneità: Le startup sono organizzazioni temporanee che nascono con l'obiettivo di crescere rapidamente e diventare grandi imprese.
2. Ricerca del modello di business: A differenza delle imprese tradizionali, le startup non hanno un modello di business consolidato, ma lo stanno cercando e testando sul mercato. L'alto rischio deriva dal fatto che il modello imprenditoriale non è ancora definito.
3. Scalabilità e ripetibilità: Il modello di business deve essere ripetibile in diversi settori e Paesi e consentire una crescita esponenziale senza aumenti proporzionali dei costi. Un esempio di modelli scalabili possono essere gli ebook o la stampa 3D.

1.2 Tipi di Startup

Le startup presentano molte differenze tra loro. Il prodotto o servizio che vogliono realizzare determina la classificazione di quella startup in visionaria, innovativa o ordinaria.

Una startup è da considerare visionaria se rivoluziona completamente il proprio settore o crea un mercato del tutto nuovo. Una startup è innovativa se migliora un prodotto o servizio già esistente con un cambiamento rilevante. Ad esempio Airbnb ha rivoluzionato il proprio mercato introducendo la possibilità ai privati di affittare i propri appartamenti.

Ordinaria invece è una startup che non realizza grandi rivoluzioni ma si limita a migliorare o a replicare le idee esistenti. Si tratta di aziende che puntano più sull'efficienza e sulla competizione in termini di costi.
Il mercato di appartenenza è un'altra grande differenza tra le startup. Esse possono essere classificate come aziende tecnologiche ma ulteriormente divise in: aziende che operano nel settore medico(MedTech), altre che operano nel settore finanziario(FinTech) altre ancora nel settore food(FoodTech) e altre che sviluppano biotecnologie(BioTech). Il progetto analizza principalmente aziende che lavorano nel settore AITech, il mercato relativo alle intelligenze artificiali, ma, per mancanza di dati, sarà possibile trovare anche alcune aziende del settore finanziario e biotecnologico.
Un ulteriore classificazione è stata concepita da Steve Blank in un articolo risalente al 2013 nella sezione Business del sito del Wall Street Journal. Queste sono:

- Lifestyle
- Small business

- Scalabile
- Acquisibile
- Large company
- Sociale

Lifestyle Startup: Work to Live Their Passion

gli imprenditori che fondano queste aziende vivono della loro passione e amano il proprio lavoro, come i surfisti che vivono di quello dando lezioni a chi vuole imparare a stare sulla cresta dell'onda. Si può paragonare anche ai programmatori che si dedicano con piacere a progetti di sviluppo di nuovi siti web o applicazioni. In generale a tutti coloro che vanno e tornano dal proprio lavoro felici di quello che devono fare o hanno fatto.

Small business Startup: Work to Feed the Family

rappresentano la maggioranza delle startup negli Stati Uniti. Sono spesso piccole attività familiari gestite direttamente dal proprietario, il quale investe il proprio capitale o quello preso in prestito da amici e parenti.

Spesso non si tratta di attività che hanno un grande successo economico perché il loro obiettivo in realtà non è nemmeno raggiungere quel tipo di successo.

Startup scalabili: Born to Be Big

Sono aziende progettate per crescere rapidamente e diventare grandi. Aziende come Google, Uber e Facebook rientrano in questa categoria. I criteri per l'assunzione del personale sono molto stringenti e sono aziende note per avere sotto contratto i migliori ingegneri sul mercato. Il loro obiettivo è cercare trovare un modello di business ripetibile e

scalabile. Dopo averlo trovato cercano del capitale per crescere come attività.

Startup acquistabili: Acquisition Targets

nate con l'obiettivo di essere vendute a grandi aziende, spesso per cifre tra i 5 e i 50 milioni di dollari.

Large Company Startup: Innovate or Evaporate

Sono costituite dalle grandi aziende che un tempo erano startup tecnologiche che puntavano, con un prodotto rivoluzionario, a scalare. Queste sono costrette ad innovare sempre per rimanere competitive altrimenti rischiano di cadere nel baratro e di fallire.

Startup Sociali: Driven to Make a Difference

Sono startup nate per trovare soluzioni a problemi che riguardano l'intera umanità. Il loro obiettivo non è scalare e diventare grandi aziende con molti profitti, ma riuscire a sviluppare un prodotto per il bene comune che risolva un problema importante. I fondatori sono imprenditori che si distinguono dagli altri per obiettivi finanziari e motivazioni del team.

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.

1.2Ciclo di vita

Durante il suo percorso, una startup attraversa diverse fasi, ognuna delle quali è caratterizzata da sfide specifiche, esigenze differenti e obiettivi da raggiungere. Ecco un'analisi dettagliata delle principali tappe:

1.3.1 Pre-seed

Questa è la fase embrionale in cui l'idea viene creata, in cui si fa una previsione sul mercato che andrà a toccare e se è effettivamente utile per la soddisfazione di un bisogno.

Essendo una fase iniziale l'azienda non è ancora realmente partita pertanto non ha un modello di business e non ha realmente un team che sarà poi messo su dal founder e un co founder che dovranno anche occuparsi della parte legale oltre che pensare a come trasformare l'idea in qualcosa di realmente utile.

È una fase in cui si possono ottenere dei finanziamenti che sono chiamati FFF(Friends, Family and Fools) proprio ad indicare coloro che solitamente investono in questa fase ovvero amici, parenti e pazzi.

Un modo per poter sostenere questa e le altre fasi preliminari è partecipare ad un acceleratore di startup. Questo è un programma, offerto da terze parti, erogato specificatamente per finanziare le startup emergenti. Coloro che offrono questo tipo di servizio sono le grandi aziende che offrono supporto ad imprese vicine, con prodotti o servizi, a quello che loro fanno.

Altri possibili fornitori sono enti pubblici governativi o università nell'intento di rendere pubbliche le loro idee.

1.3.2 Seed

La fase di seed è considerata molto importante per il proseguio della vita di una startup. L'obiettivo di questa fase è rendere reale l'idea iniziale e

consolidare il modello di business. In questa fase si raccolgono prove, con la sperimentazione, per prendere decisioni che servono a validare l'idea. Inizialmente ci saranno delle ipotesi che, attraverso le sperimentazioni e la verifica, verranno accettate oppure rifiutate. Questo servirà a cambiare ipotesi verso un'altra sulla quale verrà eseguito nuovamente la convalida. Durante questa fase il business model viene definito insieme al business plan e si crea per la prima volta un prodotto minimo funzionante (Minimum Viable Product-MVP) che è utile per essere sottoposto al vaglio della critica del cliente. Ciò che è auspicabile è che il cliente sia talmente soddisfatto da finanziare con un importo minimo (20.000-40.000 euro) il round di finanziamento.

In questa fase sono importanti i programmi di accelerazione poiché consentono di velocizzare il processo di "trial and error". Questi programmi favoriscono il contatto con professionisti del settore che hanno un'esperienza ampia in questo settore e che mettono a disposizione le loro competenze a servizio delle startup in modo tale che possano "evolversi"

1.3.3 Early Stage

Questa fase è utile a comprendere le esigenze del mercato e individuare attraverso i feedback il miglior prodotto e il miglior mercato che servono per avere i primi ricavi. I feedback, quindi, sono parte di un processo iterativo teso a migliorare eventuali bug o difetti scoperti dagli utenti. Ovviamente in questa fase aumentano e di tanto le entità a supporto delle startup come Venture Capital e acceleratori, utili non solo per il finanziamento ma anche a testare i modelli di business. È una fase delicata in cui si segna il destino della startup per cui è molto importante attirare i clienti e di conseguenza i finanziamenti per rimanere in vita.

1.3.4 Early Growth

In questa fase il business è avviato, l'MVP è ormai stato pienamente realizzato, i clienti si sono consolidati e la gente è disposta a pagare per il prodotto o servizio. Ciò che si punta a migliorare in questo momento è il business model che deve essere migliorato affinché possa far crescere l'azienda e farla scalare. Di fondamentale importanza sono anche il piano di marketing e la strategia commerciale per incrementare esponenzialmente i clienti ed espandersi a macchia d'olio nel paese in cui si è sviluppato e magari avviare anche 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

La fase di growth è quella in cui gli utenti, i clienti e il fatturato fanno un balzo in avanti notevole e imparagonabile alle altre fasi. È una fase in cui è difficile arrivare e solo le startup che hanno un prodotto affermato sul mercato dei clienti consolidati e in termini di fatturato dei numeri positivi. In questa fase la startup dovrebbe quindi crescere di tanto e aumentare anche i clienti.

Chi risulta fondamentale in questa fase sono i Venture Capital e i Corporate venture Capital. Sono due tipi di investitori che si differenziano in quanto i Venture Capital sono istituzionali, mentre i Corporate Venture Capital sono Venture Capital di grandi aziende alla ricerca di qualcosa che possa collaborare o eventualmente supportare 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 consolidato il modello di business, il prodotto e il servizio è necessario espandersi su più mercati. In tal senso l'azienda si apre a nuove aree del mercato sia geografiche che settoriali. Anche questa fase presenta delle criticità: sbagliare il luogo o il mercato su cui catapultarsi potrebbe inficiare sui rischi a cui la startup è assoggettata. Un modo semplice per uscirne vivi da questo processo potrebbe essere certamente stringere accordi con aziende importanti presenti sul mercato sul quale si vuole tuffare.
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 è l'ultima fase che caratterizza le startup. Questa può evolversi in azienda vera diventando accessibile a tutti oppure può fallire o anche essere venduta.
Quindi ricapitolando le principali opzioni per la exit sono:

• Tramite l'Offerta Pubblica Iniziale o IPO (Public Sale Offer "OPV") le quote della startup sono disponibili al pubblico. L'imprenditore rende la propria attività acquistabile, seppur tramite azioni, in borsa e permette a chiunque di accedervi, rapidamente. Li sviluppi quindi di un'azienda in IPO sono i seguenti:
• 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 società moderna è solita a elogiare i grandi successi, nel contesto delle startup quelle di cui si parla sono solitamente le startup unicorno, che hanno una valutazione superiore al miliardo di euro. Come si è visto in precedenza ciò che ha maggiore rilevanza è la fase di avvio. Una ricerca fatta da CBInsights ha rilevato che il 70% delle nuove aziende tecnologiche fallisce nella fase di early stage. Il settore più colpito da questo fallimento prematuro è il settore che realizza hardware con il 97% che finisce per fallire. CBInsights ha realizzato uno studio nel tentativo di comprendere i motivi che provocano questo fallimento, andando a leggere gli articoli scritti da investitori e giornalisti e consultando le ultime dichiarazioni dei fondatori.

Ci sono diversi motivi che portano le startup a fallire e solo in casi esclusivi è un solo motivo che prevale. CBInsights, sito che si occupa di raccogliere ed elaborare dati per fornire statistiche utili agli investitori in startup, ha rilasciato la sua ricerca sulle cause del fallimento. Di seguito elencate le motivazioni principali che portano al fallimento secondo lo studio:

- Mancanza di liquidità: spesso associato a una gestione e distribuzione sbagliata delle risorse, è la causa di fallimento di una startup nel 38% dei casi. Oltre a ciò, i principali motivi sono da ricondurre alla mancanza di leadership e alla difficoltà

nell’accontentare le richieste del mercato.

• Mancato soddisfacimento di un bisogno del mercato: nel 35% le startup offrono prodotti o servizi che non rispondono a una domanda reale, portando al fallimento.

• Non supera la concorrenza: Come visto in precedenza, non tutte le startup sono innovative, anzi molte puntano a migliorare quello che già c’è. Quindi la startup potrebbe trovare molti concorrenti che potrebbero ostacolare il processo di crescita portando al fallimento. Certo non bisogna assolutamente concentrarsi sul confronto con le altre aziende ma ignorare completamente potrebbe portare al fallimento. Questa è stata la causa del fallimento nel 20% dei casi.

•modello di business debole: Il modello di business è estremamente importante per comprendere le attività da intraprendere in caso di fallimento del servizio principale. Questo comporta svantaggi sia per l’imprenditore che per gli investitori perché si trovano di fronte a incertezza che comporta inevitabilmente una perdita nei finanziamenti. Questo porta inevitabilmente al fallimento nel 17% dei casi.

• Sfide legali: Può capitare che la startup si trsformi da una semplice idea a una marea di complicanze legali che portano al fallimento nel 18% dei casi

• Pricing/costi: uno dei compiti più difficili è assegnare un prezzo al prodotto o servizio che si vuole vendere. Nel 15% dei casi si fallisce proprio per la mancata valorizzazione del proprio prodotto o per un ipervalutazione dello stesso che si scontra con l’offerta dei clienti, decisamente più bassa.

• Team: è necessario oltre che una buona leadership anche un discreto team al di sotto di lui che ha la formazione necessaria per soddisfare a pieno la “visione” dell’imprenditore. È necessario che il team sposi questo progetto in toto, e che si impegni affinché possa contribuire a livello significativo. Inoltre è richiesta una buona coesione e una grande flessibilità. Quando tutto questo è mancato l’azienda è fallita nel 14% dei casi.

• Rilasciare il prodotto al momento sbagliato: Cogliere l’attimo per rilasciare quello che hai sviluppato è un altro prerequisito per non fallire. Non bisogna essere precipitosi e rilasciare quando ancora il prodotto non è interamente finito perché il rischio è quello di perdere clienti e recuperarli poi dopo è davvero difficile. Se invece, al contrario si aspetta troppo ci potrebbe essere qualcun altro che porti la stessa cosa prima e potrebbe superarti. Quindi il tempismo è fondamentale ed è la causa del fallimento nel 10% dei casi.

• Scarsa qualità del prodotto Un prodotto scadente o difettoso può allontanare i clienti e rendere difficile il recupero della fiducia del mercato.. Questo può portare a feedback negativi, abbandono da parte dei clienti, e difficoltà nel ricostruire ciò che si era costruito nei mesi precedenti. La scarsa qualità è causa del fallimento nel 8% dei casi.

• Disallineamento team-investitori: Il litigio tra co-fondatori può essere fatale nel 7% dei casi. Le discrepanze non coinvolgono per forza i fondatori ma potrebbero riguardare anche gli investitori, il Board e in tal caso la situazione potrebbe diventare ingestibile. A volte capita ad esempio che un fondatore non crede nella possibile evoluzione

dell’azienda perché magari si rapporta ad un mercato estremamente di nicchia oppure ci potrebbero essere opinioni discordanti sul prezzo del prodotto. Tutto ciò potrebbe portare al fallimento se non si è in grado di ristabilire la situazione in tempo.

• Pivot andato male: Il business model come già detto è un elemento fondamentale. Una sua possibile modifica o una modifica del prodotto o servizio finale si chiama pivot. Se questo cambiamento non porta i risultati sperati il rischio del fallimento è dietro l’angolo. Infatti il fallimento per questo motivo avviene nel 6% dei casi.

• Burn out/perdita di passione: Chiaro che l’anello debole quando si parla di startup è chiaramente il fondatore. Le responsabilità sono elevatissime e i fondi almeno all’inizio molto pochi. Non sempre chi sviluppa l’idea è a conoscenza del settore su cui andranno ad operare. Talvolta capita che il lavoro che si erano immaginati si scontra con la dura realtà dei fatti che è ben diversa e potrebbe non piacere ad alcuni. Queste ragioni hanno portato al fallimento nel 5% dei casi.

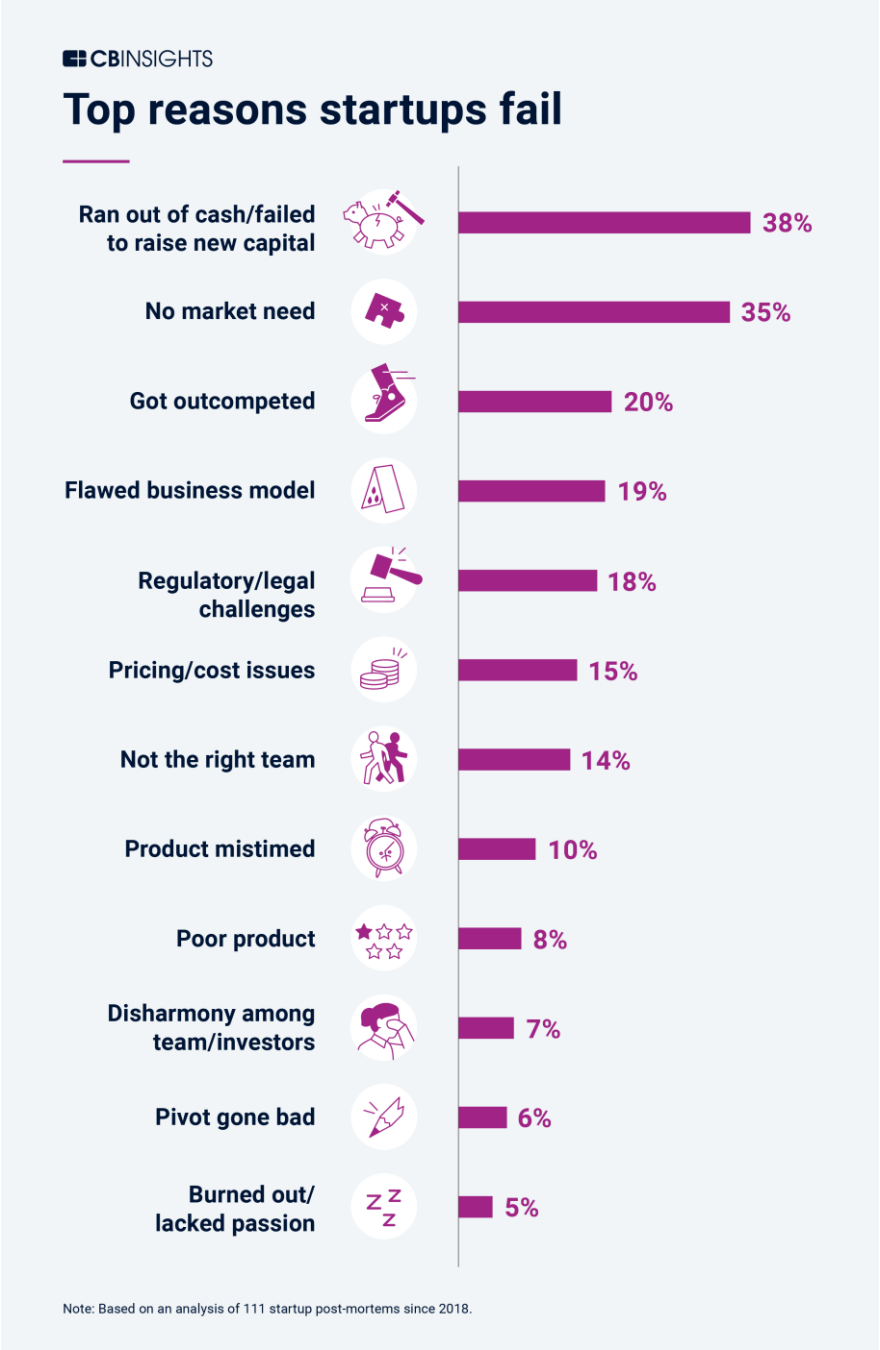


Figura 1**: I 12 motivi per cui le startup falliscono.* ***Fonte: *"The Top 12 reasons startup fail", CBInsights.*

1.5 I modi per finanziare una startup

Per una startup early stage I primi finanziamenti sono fondamentali. Secondo alcuni studi il 94% delle startup non riesce a rimanere in vita dopo il primo anno e uno dei motivi per il quale succede ciò è proprio la mancanza di finanziamenti. Perciò è importante vedere i mezzi con cui è possibile finanziare una startup che sono elencati di seguito:

1.5.1 Bootstrapping

Questi fondi servono per avviare l'impresa e sono solitamente capitali che appartengono allo stesso imprenditore.

Nella prima fase lo startupper ha difficoltà a trovare qualcuno che voglia finanziare il proprio progetto e quindi fa spesso uso dei propri risparmi o chiede aiuto a persone a lui vicine, come amici o parenti, che hanno il coraggio di investire ma soprattutto i mezzi per farlo.

1.5.2 Crowdfunding

Il crowdfunding è il modo più facile per raccogliere soldi da gente sconosciuta. Questo metodo consiste nella richiesta di una somma di denaro non troppo elevata a più persone per motivi molteplici. Il canale più usato è il web e le più famose piattaforme che operano in questo settore sono Kickstarter e Indiegogo. Colui che ha bisogno di raccogliere denaro pubblica il proprio progetto sulla piattaforma e successivamente questo diventa di dominio pubblico. La compagnia alla quale si affida riceverà ovviamente una provvigione dalla campagna creata dall'imprenditore, che potrà comunque godere della maggior parte degli incassi richiesti. Spesso questi investitori non

potranno infierire sulle scelte aziendali ma avranno solo in comune il piccolo rischio che la spesa investita non venga ripagata. Il crowdfunding è spesso usato nelle fasi iniziali, ma ciò non toglie che possa essere ripreso in fasi successive.

1.5.3 Business Angel

I Business Angel sono delle persone che offrono il proprio contributo, principalmente in termine monetario, al finanziamento di una startup nella fase iniziale. Gli investimenti si aggirano intorno ai € 20.000 fino ad un massimo di € 100.000 e comportano in una fase successiva una partecipazione al capitale.

Spesso si tratta di gente esperta, che ha esperienza, e che è pronta ad offrire il proprio contributo di imprenditore affermato a solitamente imprenditori novellini che si affacciano per la prima volta in questo campo. L'obiettivo di questi ultimi è sicuramente rientrare nell'investimento fatto ma anche aiutare i capi delle startup a portare il loro prodotto e business model sul mercato. Chiaramente il loro ruolo sta diventando sempre più importante perché offrono un bagaglio di esperienza non trascurabile ma anche conoscenze e un punto di vista diverso utile per la crescita dell'attività.

1.5.4 Venture capital

Il venture capital è un tipo di investimento che tende a guardare ad ampio raggio al futuro. Le aziende coinvolte sono in una fase di grow e godono della stima del mercato. Infatti gli investimenti venture capital sono avviati da fondi istituzionali con l'obbiettivo di guadagnare in maniera consistente dalle azioni che successivamente verranno quotate in borsa. L'insicurezza che caratterizza le startup, almeno nella fase iniziale, non permette ai fondi di investimento bancari di avvicinarsi con interesse a queste realtà. L'investitore, chiamato

Venture Capitalist, acquisisce parte delle azioni della società fino al termine del finanziamento, che può durare dai 3 ai 10 anni o nel caso in cui l'impresa venga venduta. L'investimenti che vengono fatti dai venture capital sono consistenti e riguardano generalmente diversi milioni. Di conseguenza non è affatto facile ottenere un finanziamento del genere in quanto non sempre l'azienda offre abbastanza garanzie.

1.5.5 Incubatore startup

Per aiutare le startup a raggiungere il successo solitamente le si sottopone a un programma chiamato incubatore di startup. Esso consiste come da definizione presente su "The smart guide of innovation" in "un luogo dove gli imprenditori hanno strutture, servizi, e competenze in grado di soddisfare i loro bisogni e realizzare le loro idee di business"
Ciò che un imprenditore trova quando si imbatte in un incubatore sono quindi i servizi di cui fanno parte: spazi adibiti al lavoro di squadra, servizi per la gestione e l'organizzazione, formazione, dei mentor startup ovvero delle persone che offrono consulenza, accesso ai finanziamenti e networking. Generalmente questo tipo di organizzazioni non hanno scopo di lucro e l'unico fine per cui restano in piedi è aiutare gli imprenditori a migliorare il proprio business. Questi possono essere sia pubblici che privati e sono spesso correlate a scuole private ed università che consentono agli studenti e alunni di seguire parte dei programmi. Ovviamente gli incubatori possono essere di diversa natura formati da governi, gruppi civici, insieme di startup o imprenditori di successo.

1.5.6 Acceleratore startup

Una società che fa in modo che delle startup crescano grazie a dei programmi di accelerazione è chiamata acceleratore di startup. Il percorso che queste startup svolgono punta alla rapidità, come suggerito dal nome, e serve per formare in maniera immersiva i giovani che si affacciano a questo mondo con l'obiettivo di ridurre gli errori dettati dall'inesperienza che si incontrano durante il lavoro.

> Susan Cohen dell'Università di Richmond e Yael Hochberg della Rice University hanno evidenziato i quattro punti più rilevanti per un acceleratore: è a tempo limitato, sono condotte da tutor e vengono completate con conferenze o giornate dimostrative (demo day). Questo tipo di attività è propria degli acceleratori e infatti nessun altro tipo di finanziatore (incubatori, business Angel o venture capitalist) offre queste possibilità. L'elemento comune tra tutti resta quello di far di crescere le startup ma il modo in cui lo realizzano è spesso molto diverso.

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; can also be nonprofit
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 incubator and accelerator 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

© HBR.ORG

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

1.5.7 Prestiti bancari

Le banche sono l'ultima carta che una startup si può giocare per ottenere finanziamenti. Solo di recente le banche hanno cominciato a supportare le startup e lo hanno fatto in tre diverse maniere:

• La concessione di finanziamenti tramite il fondo di garanzia per le PMI: è un fondo statale concesso alle aziende con una somma erogata di circa 2,5 milioni di euro e fino a un massimo dell'80% del finanziamento. Oltre alle startup e agli incubatori certificati, possono fare richiesta per questo fondo qualsiasi impresa di micro, piccole e medie dimensioni oltre che i

professionisti che sono regolarmente iscritti agli ordini professionali e che hanno i requisiti per poter accedervi.

• Tramite delle “Competizioni”: sono delle gare create ad hoc da banche, fondazioni, multinazionali ma anche da incubatori e fondi di investimento. Queste banche danno alle startup finanziamenti attraverso mutui che variano dai 30 ai 250 mila euro. Le startup hanno l’obbligo di rimborsare il debito entro e non oltre i 7 anni altrimenti si può optare nel trasformare in azioni il finanziamento. Così facendo la banca diventa a tutti gli effetti azionista dell’impresa.

• Mutui agevolati per imprese giovani: canale a cui difficilmente le imprese accedono.

CAPITOLO 2: L'INTELLIGENZA ARTIFICIALE

2.1 Definizione

Come per le startup, l'intelligenza artificiale ha nutrito la curiosità di molti esperti del settore che hanno anche provato a definirla secondo una moltitudine di accezioni.

Secondo un'accezione strettamente informatica, l'I.A. potrebbe essere classificata come la disciplina che racchiude le teorie e le tecniche pratiche che puntano all'implementazione di algoritmi che rendono le macchine intelligenti, in questo momento solo in particolari domini e contesti applicativi. Una seconda definizione, strettamente legata al citato 'Test di Turing', descrive l'I.A. come: «... l'impresa di costruire sistemi di simboli fisici che possono passare in maniera affidabile il Test di Turing» (M. L. Ginsberg). L'"Enciclopedia Britannica" definisce l'I.A. come segue: «Artificial Intelligence (AI) è l'abilità di un computer digitale o di un robot di eseguire un compito solitamente svolto dall'intelligenza umana. Il termine è spesso associato a sistemi di sviluppo con caratteristiche di pensiero umane, come l'abilità di ragionare, di scoprire i significati o imparare dalle esperienze passate».

Anche l'"Enciclopedia della Scienza e della Tecnica" della Treccani ha provato a dare una definizione per l'Intelligenza artificiale. Quello che recita è questo: «L'Intelligenza Artificiale studia le teorie, i metodi e le tecniche che permettono di creare sistemi hardware e software volti a creare un computer

con delle prestazioni che, a un osservatore comune, sembrerebbero essere di pertinenza esclusiva dell'intelligenza umana»

Le definizioni di I.A. sono anche differenziabili in due tipologie diverse:

- L'impostazione che è più funzionale ovvero che realizza l'intelligenza senza svilupparla a livello fisico, senza renderla simile a quella umana. Quindi si concentra ad emulare il cervello degli esseri umani e non la fisionomia.
- Viceversa, per l'impostazione 'strutturale' si copia il cervello tale e quale a come è ricostruendo a tutti gli effetti la sua struttura, la sua complessità e le sue caratteristiche.

Ulteriori distinzioni possono essere nel tipo di approccio operativo. I tipi di approcci generalmente sono due 'top down' e 'bottom up'. L'approccio top-down non considera il funzionamento del cervello umano ma cerca di codificarlo attraverso delle regole e dei simboli. Infatti li stati mentali si distinguono attraverso l'uso di simboli in un sistema sia simbolico che fisico.

L'approccio bottom-up è un approccio più reale che parte da reti di neuroni artificiali, considerati come architetture, che svolgono lo stesso ruolo di neuroni reali, per costruire strutture e modi di ragionare più complessi. Nel caso di top down è ovviamente limitato dalle regole imposte e non è capace di generalizzare pur essendo interpretabile perché basato su regole comprensibili e modificabili. Di contro il bottom-up apprende pattern complessi e si adatta a dati variabili senza una conoscenza predefinita. Ma ha bisogno di un enorme carico computazionale per funzionare oltre che di molti esempi per addestrarsi.

2.2 Intelligenza artificiale forte e debole

Considerando le funzioni di un cervello reale, un AI può compiere diverse funzioni:

- Agire come se fosse un umano non facendo risaltare la sua vera natura.
- Pensare come un umano attraverso le proprie capacità cognitive
- pensare in maniera razionale giustificando le scelte con una logica.
- Agire in maniera razionale ovvero con un procedimento che porta a cercare il miglior risultato a seconda delle informazioni che si hanno a disposizione.

Quanto detto permette di differire l'intelligenza artificiale in due diversi tipi:

Intelligenza Artificiale debole: è un sistema che permette la simulazione di parte delle funzioni cognitive dell'uomo senza però arrivare a simulare tutte le capacità che lo contraddistinguono. Ad esempio, esistono programmi che permettono la risoluzione di problemi matematici di tipo logico o permettono anche alle macchine di prendere decisioni, ma non riescono ad ampliare le proprie funzionalità su tutto ciò che contraddistingue l'uomo.

Intelligenza Artificiale forte: Nel caso di un AI forte invece si tratta di un'intelligenza che oltre a svolgere più compiti specifici è dotata anche di una comprensione della realtà che la circonda. In un certo senso è senziente comprende ciò che le sta attorno. Ad oggi questo tipo di tecnologia non è ancora stata sviluppata ed ha nella sua natura rischi etici molto importanti.

2.3 Machine learning e deep learning

Partendo dalla differenza appena illustrata tra intelligenza artificiale forte e debole è importante per introdurre la differenza che c'è tra 'Machine Learning' e 'Deep Learning', due divisioni che segnano una differenza considerevole nella disciplina dell'AI. Ciò che le distingue principalmente sono le metodologie che andranno ad adottare per realizzare il training, fondamentale per realizzare un compito o un'azione.

2.3.1 Machine learning

Il Machine Learning o apprendimento automatico è utile per realizzare una stima di una funzione complessa.

Il machine Learning utilizza un insieme di tecniche e sistemi come il riconoscimento dei pattern, la statistica computazionale, reti neurali e altre che consentono al sistema di imparare dai dati e, successivamente, prendere decisioni o predire qualche dato mancante.

Un modello di machine learning può processare un insieme di dati provenienti da diverse sorgenti che insieme formano una base di conoscenza. Questo modello può fare diversi compiti tipo realizzare una classificazione facciale riconoscendo i pattern comuni, riconoscere il parlato e gli oggetti e altri compiti legati al riconoscimento e alla classificazione come quello di prevedere il round di finanziamento di un'azienda, l'oggetto principale di questa tesi.

Un modello di machine learning ha un funzionamento completamente diverso rispetto ad un codice normale. Il codice normale viene scritto con una serie d'istruzioni che lo rendono euristico. Il machine learning invece elaborano i dati e successivamente compiono predizioni sugli stessi. Quindi gli algoritmi usati per il machine learning non sono euristici, ovvero non hanno un insieme

di istruzioni ben definite ma utilizzano un elaboratore per comprendere e apprendere dai dati in maniera solitaria e poi effettuare predizioni su di essi.

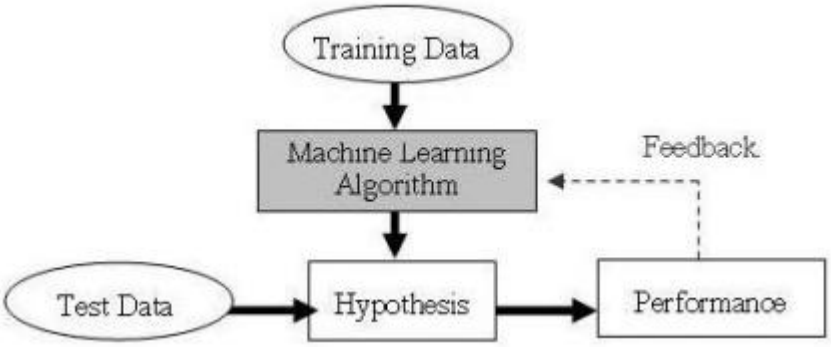


Figura 3: Schema di funzionamento del machine learning

Il Machine Learning può compiere tre diverse mansioni:

- Classificazione
- Clustering
- Predizione

Il modello di apprendimento è la parte fondamentale per avere una distinzione tra i vari tipi di machine learning. Ci sono due tipi di algoritmi che si distinguono attraverso il modello di apprendimento:

Con la supervisione didattica (Figura 4), l'apprendimento avviene mediante un insieme di istanze di input utilizzate per l'addestramento del modello. Queste istanze generano un valore predetto per la variabile target che si desidera stimare, il quale verrà confrontato con il valore reale che la variabile assume in quella specifica istanza. In seguito, gli errori di

previsione vengono utilizzati per aggiornare il modello e migliorare le sue prestazioni nel tempo

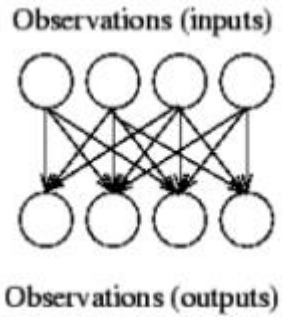


Figura 4: Machine learning con supervisione

- senza supervisione didattica (Figura 10): l'apprendimento avviene mediante l'analisi di un insieme di istanze di input senza l'uso di etichette o valori target predefiniti. L'obiettivo è quello di notare eventuali correlazioni o pattern comuni tra i dati in modo da poterli classificare. Inoltre, questo modello non può verificare le proprie previsioni ma è capace di dividere i dati in gruppi o cluster che hanno una certa somiglianza tra loro.

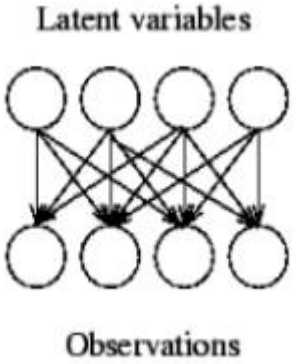


Figura 5: Machine learning senza supervisione

2.3.2 Deep learning

ll Deep Learning, o Apprendimento Profondo, si basa su modelli simili a quelli che normalmente caratterizzano la struttura del cervello biologico umano.

Mentre il Machine Learning corrisponde al metodo attraverso cui l'intelligenza artificiale viene "allenata", il Deep Learning è l'approccio che permette di simulare il comportamento della mente umana.
Il Deep Learning è una branca del Machine Learning che utilizza Reti Neurali Profonde (Deep Neural Networks), ossia delle reti che hanno diversi strati di profondità. Rispetto al machine learning impiega nuovi algoritmi per il pre-processamento dei dati e la regolarizzazione dei modelli. Queste reti traggono ispirazione dalle neuroscienze, in quanto cercano di imitare il comportamento dei neuroni umani. Ma a dispetto del cervello umano in cui i neuroni si connettono tra di loro attraverso gli assoni, i neuroni artificiali sono limitati in connessioni e le informazioni seguono un flusso predefinito.

Per poter addestrare un modello di Deep Learning c'è bisogno di un training set decisamente esteso. Pertanto, il deep learning è il modello più indicato fra tutti a soddisfare il processamento dei Big Data.

Infatti, questo nuovo metodo di machine learning deve la sua fama proprio ai Big Data, in quanto è uno dei pochi strumenti che può processarli.
Grazie alla quantità elevata di dati la rete apprende, attraverso l' addestramento, come raggiungere gli obiettivi. Un esempio di algoritmo è un modello che riesce a in primo luogo apprendere e poi successivamente riconoscere un determinato animale. Ovviamente prima di poter agire ha bisogno di un quantitativo di immagini elevate per evidenziare le differenze che ci sono tra i vari animali e riconoscere pattern che si ripetono.

Il Deep Learning può generare errori se questo è soggetto ai 'bias'. Quando le etichette sono create erroneamente, il modello apprenderà dai dati errati e di conseguenza potrebbe generare errori nelle predizioni.
Per quanto riguarda l'uso pratico molti sono gli esempi tangibili dell'uso di questa tecnologia nella nostra società: vanno dal riconoscimento testuale come l'individuazione di frodi o spam, a quello visivo come la ricerca delle immagini, ma anche quello uditivo come riconoscimento del parlato. Poi ci sono anche i sistemi di natural language process come chatgpt, i reccomandation system usati per le ricerche nei motori di ricerca come google. Ma anche la capacità di individuare pericoli nelle strade con la 'Street View Change Detection' e la capacità di tradurre in più lingue. Questi sono solo parte delle funzionalità che il Deep Learning è in grado di fare; in Google, le Reti Deep hanno già rimpiazzato decine di 'sistemi a regole' ovvero programmi con una programmazione sequenziale. Aziende come Google hanno ampiamente adottato le reti Deep per compiti che in precedenza erano affidati a programmi di tipo sequenziale.
In molte applicazioni il Deep Learning ha già ampiamente superato l'essere umano come quelle che riguardano il riconoscimento di figure comuni e l'individuazione prematura del cancro su immagini tomografiche polmonari.

2.4 funzionamento dell'intelligenza artificiale

Dal punto di vista delle abilità intellettuali, il funzionamento di un'I.A. si divide principalmente in quattro diversi livelli funzionali:

- comprensione: grazie alle capacità cognitive simulate l'AI è in grado di elaborare e riconoscere testi, tabelle, immagini e video ed estrapolarne il succo ovvero ciò che c'è di più importante.

- Ragionamento: attraverso la logica, il sistema mette insieme informazioni eterogenee raccolte. Questo è possibile grazie a precisi algoritmi e avviene in modo automatizzato
- Apprendimento: in tal caso si parla di sistemi che permettono l'analisi dei dati di input e la loro corretta restituzione in dati di output, ad esempio, i sistemi di Machine Learning che attraverso le tecniche di apprendimento automatico fanno in modo che l'AI impari e svolga varie funzioni.
- Interazione: indica la modalità di funzionamento dell'AI nell'interazione con un uomo reale.
  I pionieri di questo campo sono certamente i sistemi di Natural Language Processing che permettono all'uomo di interagire con tutta facilità e le macchine fanno con l'uomo la stessa cosa comunicando nel linguaggio naturale.

2.5 Rischi etici

Le applicazioni dell'intelligenza artificiale hanno portato sicuramente un miglioramento nell'efficienza e una riduzione di costi, che ha portato a benefici economici, oltre che aver migliorato il benessere delle persone e dei lavoratori. Per esempio, i chatbot di cui sono dotate ormai tutte le multinazionali, possono rispondere alle esigenze dei clienti in qualsiasi momento aumentando di tanto la soddisfazione del cliente verso l'azienda. Non ci sono dubbi che il rapido sviluppo di questa tecnologia e le tante applicazioni che sono realizzate grazie ad essa hanno cambiato il nostro modo di vivere, la nostra società ma ancora più in generale tutta l'umanità.
Allo stesso tempo l'AI rappresenta un possibile rischio etico per gli utenti, li sviluppatori e la società più in generale. L'AI infatti non ha sempre prodotto

ottimi risultati: ad esempio nel 2016 una tesla a guida autonoma ha ucciso un pedone dopo che l'autopilot non ha riconosciuto un camion. Il bot di Microsoft Tay.ai è stato dismesso perché era diventato razzista e sessista in meno di un giorno dopo essere stato rilasciato su twitter. Ci sono molti altri esempi che riguardano fallimenti, disparità, errori, violazione della privacy e altri problemi etici dei sistemi AI. L'AI potrebbe anche essere usata per reati o frode da criminali digitali.

Detto ciò i rischi che possono nascere con l'utilizzo dell'AI sono vari e diversi e non sono sempre noti a coloro che la utilizzano. Perciò per identificare più facilmente quelle che sono le aree di interesse che possono compromettere l'uso adeguato dell'AI si può utilizzare l'approccio sistematico di Mckinsey & Company riportato nella figura 6. Di seguito verranno analizzati i rischi uno ad uno.

Figura 6: Approccio sistematico di McKinsey e Company

2.5.1 Privacy

La privacy è il diritto di un individuo di un gruppo di individui a mantenere le proprie informazioni personali riservate. Lo sviluppo di un'AI comporta un uso massiccio di dati e spesso il consenso per poter fruire di questi ultimi non è stato concesso da coloro i cui i dati sono stati sottratti. Le aziende da qualche anno a questa parte devono fare attenzione a non diffondere i propri dati a terze parti perché questi ultimi sono ampiamente tutelati dal GDPR (General data protection regulation) che vieta l'uso indiscriminato dei dati appartenenti ai cittadini europei.

L'utilizzo improprio dei dati potrebbe portare a un abbassamento dell'affidabilità che i clienti ripongono nell'azienda con le relative sanzioni economiche e conseguenze legali, che influiscono negativamente sulla routine quotidiana del complesso aziendale.

2.5.2 Security

Il rischio legato alla sicurezza può essere visto attraverso due differenti modalità:

l'appropriazione illegittima dei dati degli utenti o la manipolazione dei dati per far sì che si inneschino meccanismi errati.

Gli attacchi che puntano all'appropriazione illecita dei dati personali sono rischi conosciuti alle aziende che hanno un'eCommerce come principale modo di rapportarsi col cliente.

Il rischio invece legato alla variazione dei dati è un problema che si è sviluppato di recente e si realizza in tre diversi modalità a seconda dei dati che vengono compromessi ovvero input, training e feedback.

- La manipolazione dei dati di input avviene modificando i dati che prendono parte all'addestramento. Un hacker può quindi fornire dei

dati estranei non coerenti con gli altri, utili a provocare un addestramento non corretto che quindi influisce inevitabilmente sulle previsioni che potranno essere sbagliate.

- Manipolazione dei dati di training: questo tipo di manipolazione avviene tramite il reverse engineering. Questo processo interroga i sistemi di AI in modo da ricevere informazioni chiave per ricreare il sistema esistente copiandolo a tutti gli effetti.
- Manipolazione dei dati di feedback: i concorrenti, se sleali, potrebbero rappresentare una minaccia alle aziende con cui si dividono la fetta di mercato. Questo tipo di azione consiste nell'alterare i dati di feedback per far sì che vengano modificate le predizioni e così facendo l'AI restituisca sempre una predizione sbagliata.

2.5.3 Fairness

Questo problema diventa evidente quando il sistema comincia a fare discriminazioni nel momento in cui deve prendere delle decisioni. Questo tipo di discriminazioni possono essere razziali, sessuali e tutto quello che riguarda la cultura woke.

I bias possono essere propri dei dataset, nel senso che già nel set di addestramento sono presenti questo tipo di discriminazioni oppure avvengono in seguito ad un modo di raccogliere i dati erroneo e in seguito l'AI viene allenata con questi deficit.

Per evitare che questo accada è necessario rendere il dataset in cui si opera completo ovvero in grado di rappresentare i gruppi minoritari in egual misura con i gruppi maggioritari presenti nella società.

2.5.4 Trasparency and explainability

Parte dei sistemi AI non riescono a fornire una spiegazione per il quale hanno ottenuto un determinato risultato. Questo tipo di sistemi che trovano delle relazioni tra i dati e forniscono previsioni abbastanza positive ma non sanno spiegare come queste vengono ricavate, vengono chiamati black box.

Ovviamente questo problema può creare più di qualche grattacapo quando si utilizzano in procedure delicate come cure mediche o quando avvengono operazioni complesse che coinvolgono grossi giri monetari. Un esempio può essere la concessione di un prestito elevato a un cliente da parte di una banca o l'autorizzazione di una transazione piuttosto anomala.

2.5.5 Safety and performance

Quando un sistema di AI entra in funzione e viene messo sul mercato necessita prima di essere rilasciato di una fase di test molto accurata perché gli errori possono inficiare negativamente sull'intera attività e causare danni maggiori rispetto ai benefici prodotti dal sistema. Questi danni potrebbero essere gravi a tal punto da, in caso di sistemi medicali, portare alla morte di colui che sta usando il servizio.

Oltre a questo la tecnologia deve costantemente aggiornarsi e allenarsi e pertanto deve essere sottoposta a un grande quantitativo di dati in modo tale che non diventi subito antiquata e difficilmente adattabile a scenari diversi, rendendo le previsioni errate e falsando il processo decisionale che si basa su di esso.

2.5.6 Third-party risks

L'implementazione di sistemi AI potrebbe richiedere l'intervento di consulenti esterni all'azienda ma esperti della materia. Il fatto di coinvolgere terze parti potrebbe non essere una soluzione sicura perché l'azienda ha di fronte diversi

rischi che potrebbero essere ridotti se il processo di realizzazione fosse svolto solo da membri interni.

2.6 Considerazioni personali

In conclusione credo che l'intelligenza artificiale sarà un invenzione che modificherà irrimediabilmente la nostra vita, segnando la fine di un epoca e realizzando una sorta di nuova rivoluzione industriale. I rischi dell'uso improprio di questa tecnologia sono elevati, così come i benefici e le semplificazioni che la nostra vita potrà avere. Sono fiducioso per l'avvenire, perché non mi spaventa il cambiamento, se questo rappresenta per me, come per tutti, un grosso miglioramento nelle nostre condizioni di vita. Nonostante ciò comprendo i timori della gente e non nego che ho più volte pensato di poter diventare in qualche modo "schiavo" di questa tecnologia e costituire il ruolo di parassita, inutile per il proseguo della società.
Onestamente credo che questa tecnologia possa aprirci un mondo e farci svolgere incarichi nuovi diversi e forse un po' più difficili. Ci aspetta una grande sfida ma come tutte le altre sfide riusciremo a superarla.

CAPITOLO 3: CASO DI STUDIO

3.1 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 più importante introdotta dalla startup
- tipo di miglioramento tecnologia: impatto che l'introduzione di nuova tecnologia ha avuto sull'azienda o sui suoi dipendenti o sui clienti.
- 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 200 aziende di cui 50 fallite e 150 che sono ancora attive. Delle 150 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

-series: che intende tutti i round di finanziamento in cui il prodotto è già stato realizzato quali:
-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.

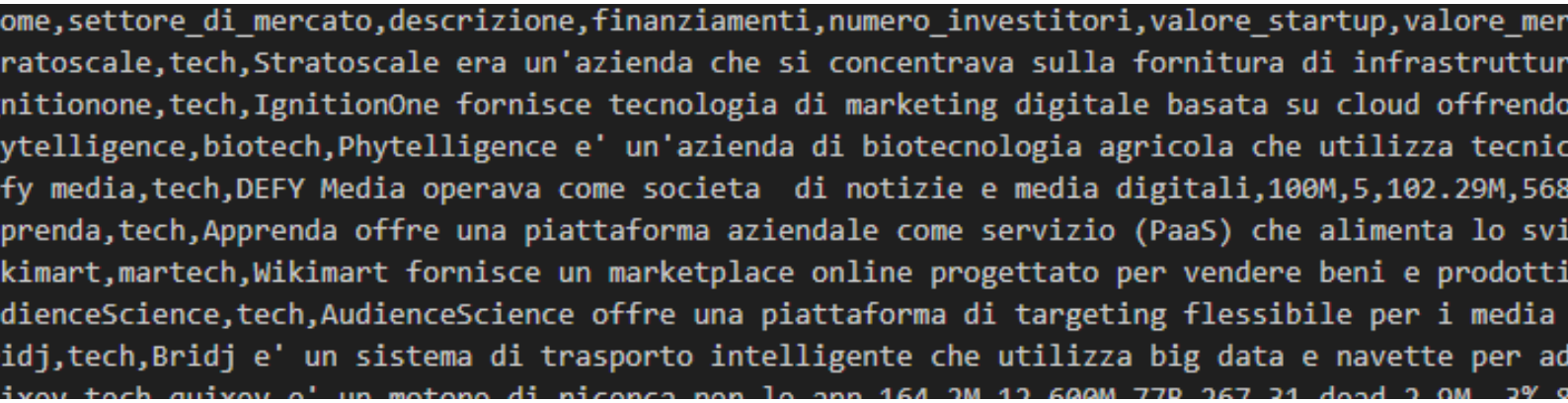


Figura 7, un piccolo estratto del dataset oggetto di caso di studio

3.2 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.
Un metodo che ho infine usato per poter ottenere una corretta scala delle variabili, ovvero per poter avere variabili con uguale influenza, è la standardizzazione. La standardizzazione permette la conversione delle feature numeriche in un intervallo che va da 0 a 1, affinché i dati abbiano una media di 0 e una deviazione standard di 1. La deviazione standard indica quanto i dati sono vicini alla media, quindi più è piccola più i dati saranno distribuiti intorno alla media. Questa si calcola con questa formula:

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2}$$

Dove n indica il numero di dati nel campione,x_i ogni singolo dato, μ la media del campione.

Mentre la formula per calcolare lo Z-score di un singolo valore x è:

$$Z = \frac{x - \mu}{\sigma}$$

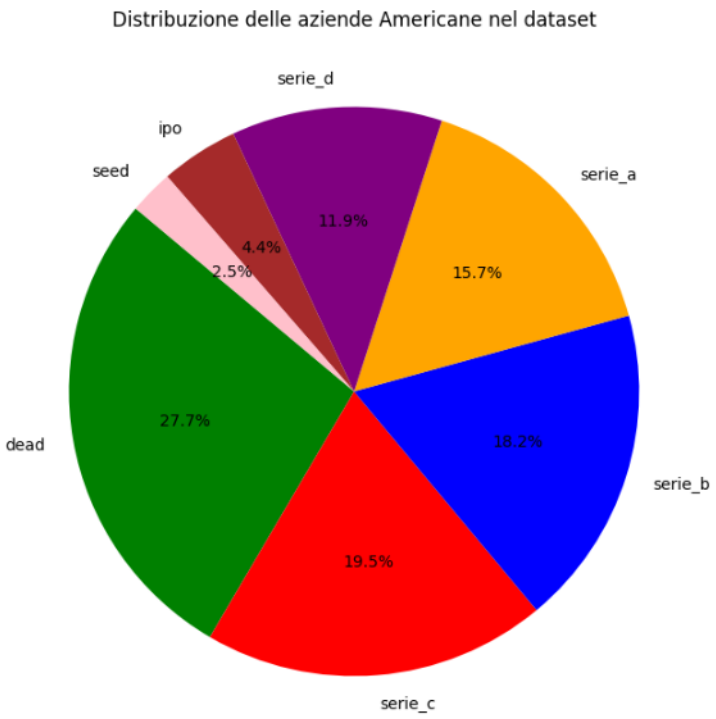


Figura 8: grafico a torta rappresentante i valori che assumono le aziende americane nel campo stage dove 0 indica l'azienda fallita e gli altri valori indicano le aziende attive

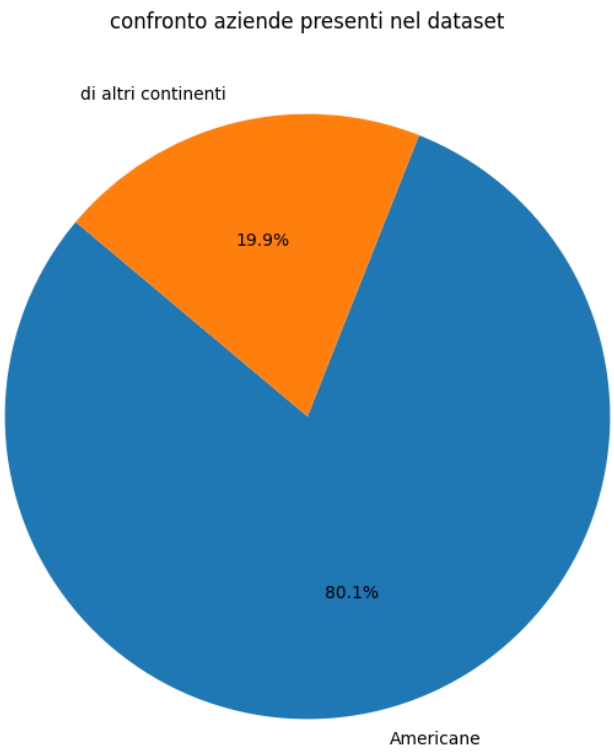
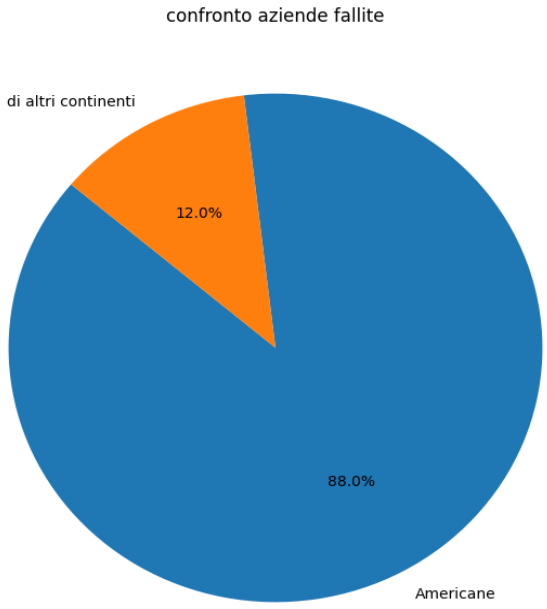


figura 9: rapporto delle aziende presenti nel dataset

figura 10: confronto tra aziende americane e estere fallite



3.3 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.

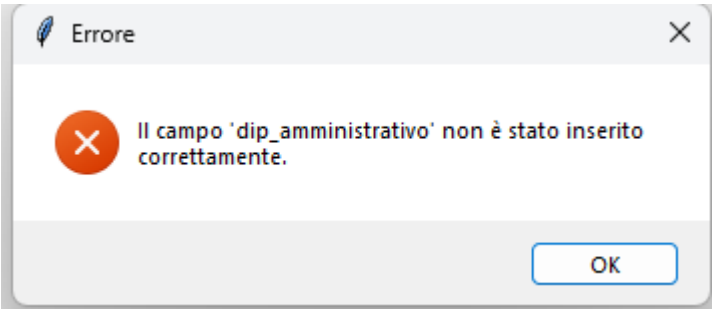


Figura 11: popup che notifica i campi non inseriti correttamente

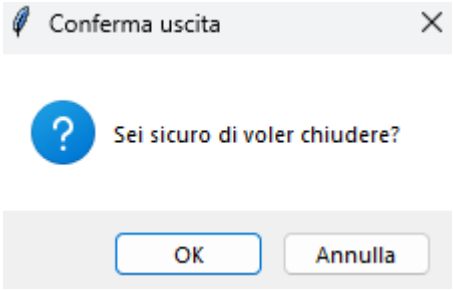


Figura 12: popup di conferma nel caso in cui si voglia chiudere la finestra

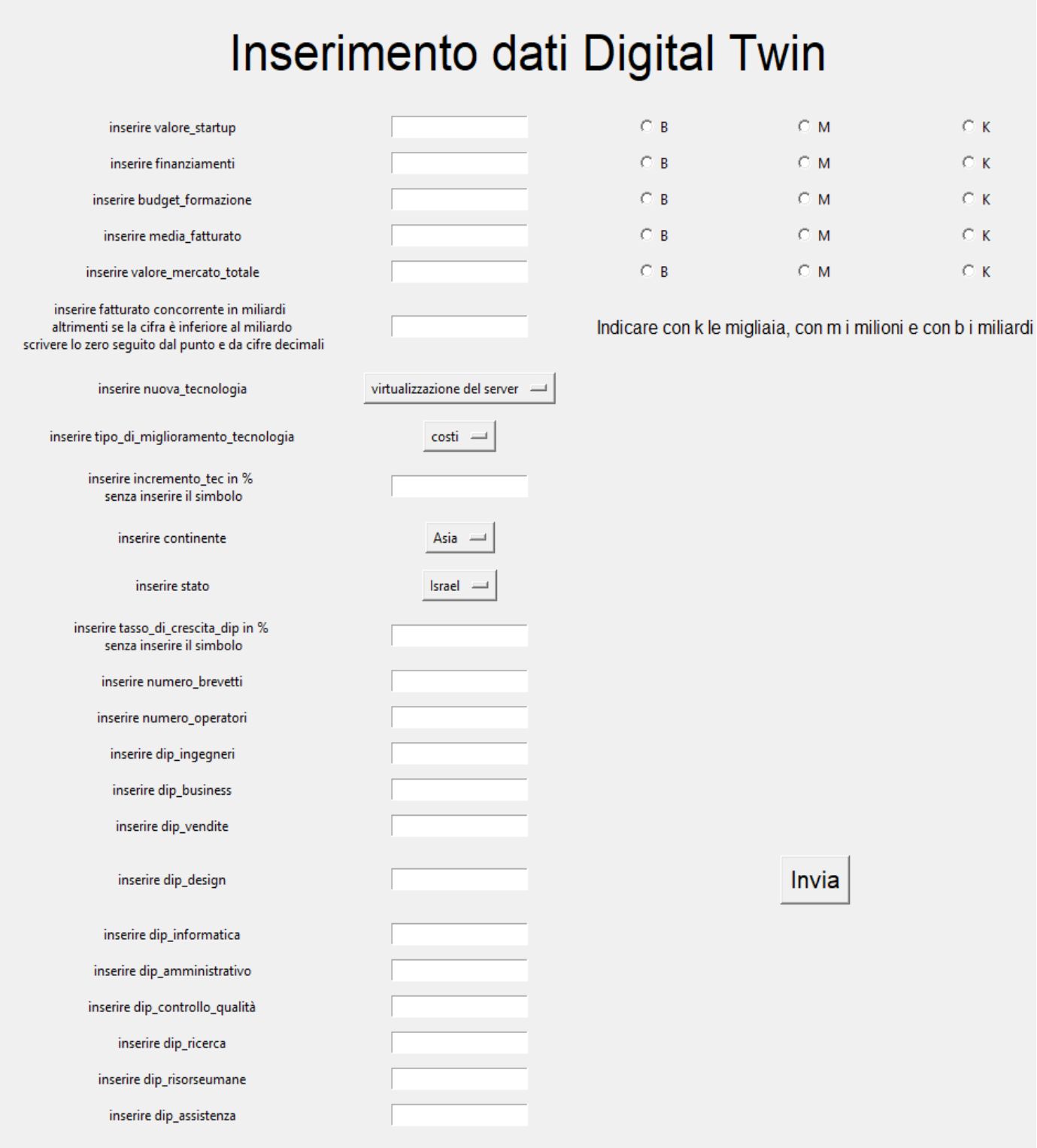


Figura 13: interfaccia grafica per l'inserimento della digital twin

3.4 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.

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 $\hat{y}_i^{(m)}$

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.

3.4.1 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 ma, per addestrare meglio entrambi i modelli, si è deciso di introdurre per entrambi la stratified k fold cross validation.
La stratified 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.

Inoltre la stratified k fold cross validation viene usata quando ci si trova di fronte un dataset, come quello in esame, sbilanciato ovvero una classe è molto più presente rispetto alle altre.
I vantaggi dell'uso di questa tecnica 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.
Poiché le classi, ovvero i valori della variabile target, sono sbilanciati, come si può notare dal grafico successivo, si è reso necessario l'adozione di un'altra tecnica ovvero l'oversampling.

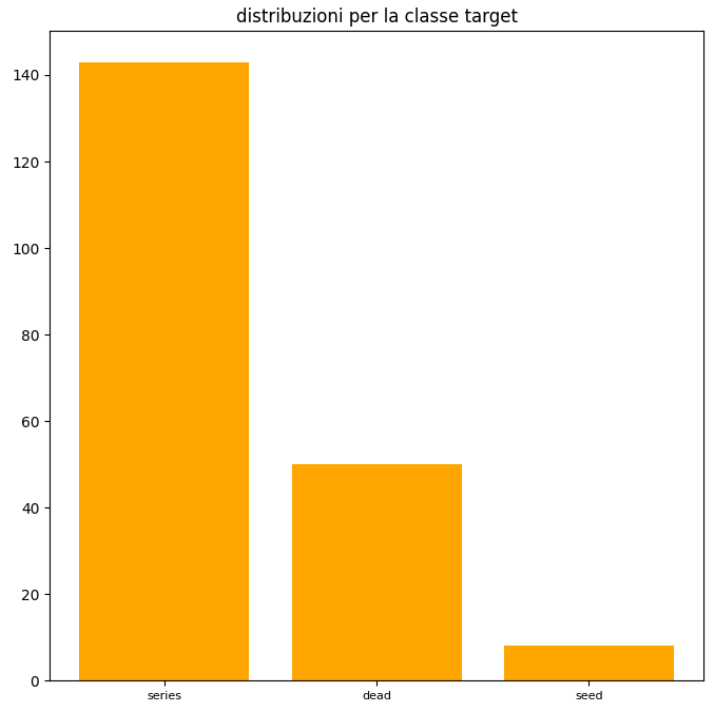


Figura che rappresenta la differenza enorme tra le classi della variabile target

L'ovesampling serve per riequilibrare le classi della variabile target aumentando le istanze delle classi minoritarie e rendendo così facendo il dataset bilanciato.

Ci sono due tipi di oversampling il random oversampling e lo smote oversampling.

Il random oversampling duplica casualmente le istanze delle classi minoritarie fino a quando non raggiungono le dimensioni della classe maggioritaria. Il vantaggio dell'uso del random oversampling sta nel fatto che è facile da implementare e aumenta la capacità del modello di generalizzare sulla classe minoritaria, lo svantaggio sta nel fatto che il modello potrebbe adattarsi troppo bene ai dati provocando un overfitting.
La tecnica SMOTE invece genera nuove istanze della classe minoritaria prendendo esempio dagli esempi già esistenti. Generando nuovi dati, il rischio di overfitting è scemato e allo stesso tempo generalizza meglio perché viene arricchito da nuovi esempi sintetici. Chiaramente il problema potrebbe nascere sui dati che genera: se i dati precedenti sono sparsi potrebbe commettere errori nella generazione e produrre dati che non riflettono la distribuzione reale.
Tornando alle differenze tra i due modelli 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.

3.4.2 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.

CAPITOLO 4: ANALISI DEI RISULTATI E CONCLUSIONI

4.1 Le metriche

Per poter valutare quale dei modelli risponde in maniera migliore, bisogna trovare dei parametri con cui valutarli. Questi parametri sono chiamati metriche di valutazione. Le metriche di valutazione adottate in questo caso di studio sono le seguenti:

4.1.1 Accuratezza

L'accuratezza è la misura delle previsioni corrette rispetto a tutte le previsioni effettuate. Si calcola con la seguente formula:

$$\text{accuratezza} = \frac{\text{veri positivi+veri negativi}}{\text{totale previsioni}}$$

dove come veri positivi si indica le previsioni corrette, veri negativi quelle sbagliate e totale previsioni è semplicemente il totale delle previsioni.

4.1.2 precisione

La precisione misura la percentuale di risultati positivi correttamente predetti rispetto a tutti quelli previsti come positivi. La formula è la seguente:

$$\text{precisione} = \frac{\text{veri positivi}}{\text{veri positivi+falsi positivi}}$$

4.1.3 richiamo

Il richiamo misura la percentuale di veri positivi individuati rispetto al totale effettivo di positivi. La formula è la seguente:

$$\text{richiamo} = \frac{\text{veri positivi}}{\text{veri positivi+veri negativi}}$$

4.1.4 f1 score

L'F1 score è la media armonica tra precisione e richiamo. La media armonica è invece una misura statistica che rappresenta il rapporto tra il numero di valori considerati (in questo caso precisione e richiamo quindi 2) e la somma tra i

reciproci dei valori numerici. Fornisce un bilancio tra queste due metriche e viene usato quando è importante trovare un equilibrio tra i falsi positivi e falsi negativi. Di seguito la formula:

$$f1=\frac{2}{\frac{1}{\text{precisione}}+\frac{1}{\text{richiamo}}}$$

4.1.5 deviazione standard

La deviazione standard è una misura statistica che descrive la dispersione o la variabilità di un insieme di dati rispetto alla loro media. Nel contesto di metriche di valutazione come precision, accuracy, recall e F1-score, la deviazione standard può essere utilizzata per valutare la stabilità e la consistenza di queste metriche. In particolare viene misurata ogni qual volta il modello viene addestrato con lo stratified k fold cross validation.

La deviazione standard si calcola con questa formula:

$$\sigma = \sqrt[2]{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

Dove:

x è ogni singolo valore del dataset

μ è la media del dataset

n è il numero di valori

4.2 I grafici

Oltre alle metriche appena elencate, si è deciso di rappresentare graficamente le predizioni dei modelli attraverso la matrice di confusione. La matrice di confusione confronta le previsioni del modello con le etichette reali.
In particolare ogni riga di questa matrice rappresenta una classe reale mentre ogni colonna rappresenta una classe predetta. Nelle celle dove i e j, rispettivamente il numero di riga e di colonna, sono uguali, sono presenti le predizioni corrette del modello. In tutte le altre celle invece sono presenti le predizioni errate, ovvero quando un determinato esempio viene predetto per una determinata classe, che non corrisponde alla classe reale.

Ho realizzato il grafico ROC per mostrare le performance dei classificatori a vari livelli di soglia. In questo caso, essendo una classificazione multiclasse, ho dovuto usare l'approccio one vs rest. Questo approccio considera la classe che si sta esaminando come positiva, mentre tutte le altre come negative, creando così una serie di problemi di classificazione binaria che corrisponde al numero di classi.

Nel grafico che sono andato a realizzare l'asse delle ascisse(x) indica il tasso di falsi positivi ovvero delle predizioni correttamente non assegnate alla classe che si sta esaminando. Mentre l'asse delle ordinate(y) indica il tasso di veri positivi ovvero i positivi correttamente classificati come positivi appartenenti alla classe che si sta analizzando. Entrambi i valori, sia il tasso di falsi positivi che il tasso di veri positivi, hanno valori che variano da 0 a 1.

L'AUC è invece l'area presente sotto la curva ROC. È un valore che indica le performance del classificatore che, anche in questo caso, va da 0 a 1. Se dovesse essere 1 il classificatore ha una perfetta separazione tra le classi, se dovesse essere 0.5 il classificatore non ha capacità discriminante per quella

classe per cui è come se fosse una classificazione casuale. Se dovesse essere minore di 0.5 il classificatore è persino peggiore di una classificazione casuale e il modello potrebbe essere invertito.

Ho costruito poi un grafico per ogni modello, che stabilisce le feature più importanti per realizzare le predizioni. Il grafico è un grafico a barre dove sull'asse delle x ci sono le features e sulle y ci sono le importanze percentuali di ogni features.

Questo grafico generale, per comprendere le feature più influenti, è stato realizzato grazie alla libreria matplotlib.

Matplotlib è una libreria di visualizzazione dei dati per Python, largamente utilizzata per creare grafici 2D in modo semplice ed efficace.
Il modulo che ho usato per creare il grafico è il modulo pyplot ed il metodo in particolare che mi ha permesso di creare il grafico a barre è il metodo bar.
I vantaggi di questa libreria sono la flessibilità, in quanto permette di creare grafici altamente personalizzabili, poi è documentata ed è possibile poter accedere al supporto online ed è anche estendibile da altre librerie che l hanno usata come base per miglioramenti stilistici sui grafici.

Gli altri grafici realizzati sono stati fatti con la libreria shap, strumento utile per spiegare le predizioni di un modello di machine learning. Shap cerca di spiegare il contributo di ogni caratteristica per le predizioni di un determinato valore del modello. Questa teoria è basata sui valori Shapley, un concetto proprio della teoria dei giochi. In pratica ogni caratteristica viene vista come un "giocatore" che contribuisce a una previsione. Ad ogni caratteristica è assegnato un valore che riflette il suo impatto sulla predizione del modello, così la somma dei valori delle caratteristiche dia la previsione del modello.

La libreria fornisce diversi tipi di grafici per visualizzare i risultati ma noi andremo a considerare solo il summary plot che mostra l'importanza globale delle caratteristiche su un intero dataset, il bar plot che mostra l'importanza media assoluta di ogni caratterisitica e infine il waterfall plot che fornisce una spiegazione dettagliata di una singola previsione del modello.

4.3 I risultati

Oltre all'analizzare le performance complessive dei classificatori, si analizzerà anche il risultato della predizione dei modelli per un caso di studio reale. Ovvero verranno introdotte tre aziende esistenti, una per ogni possibile valore della variabile target, e verrà simulato un contesto di change management, inserendo i dati presenti su internet che riguardano le aziende. Al sistema, quindi, verranno "dati in pasto" i dati relativi a queste aziende e dovrà prevedere, in base al training set con cui si è allenato, in quale round di finanziamento si trova quell'azienda. Successivamente questo valore verrà confrontato con quello effettivo e si andrà a verificare, su un contesto reale, le performance del sistema.
I risultati relativi alle metriche verranno salvati nel file output.txt, mentre quelli relativi ai grafici saranno presenti nelle relative cartelle dei classificatori presenti nella cartella image.

4.3.1 casi di studio

Ho selezionato quindi tre aziende di cui una fallita, una nella fase di seed, una nella fase series ovvero in un round di finanziamento tra il serie a e il serie d. Queste aziende sono le seguenti:

L'azienda fallita è theranos, un'azienda che aveva promesso progressi nel campo della diagnostica delle malattie. Vado ad inserire i dati ricavati dai siti elencati in precedenza nel sistema attraverso l'interfaccia grafica.

Dopo quindi aver fatto l'inserimento nella figura 13, otteniamo i riscontri delle predizioni nella figura 14. Possiamo osservare come 2 classificatori su 3 predicono correttamente.

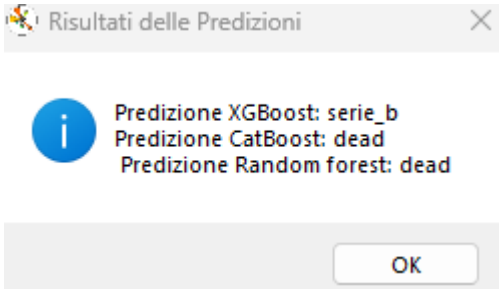


Figura 14: popup delle predizioni

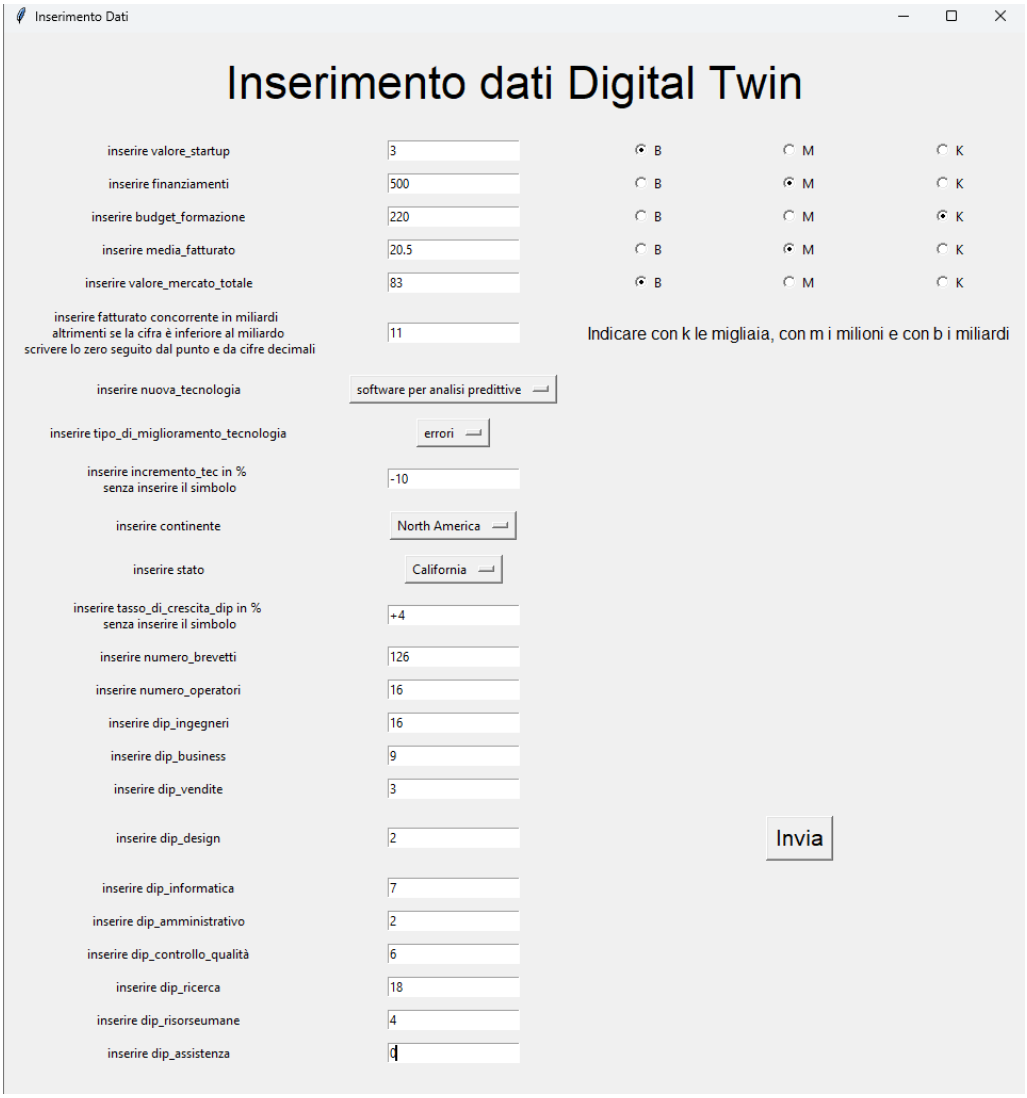


Figura 15 inserimento dati theranos

L'azienda selezionata per il feed è invece opusflow, una startup che si occupa di creare un software in grado di aiutare le aziende edili nella coordinazione dei loro progetti in maniera più efficiente. Qui invece si può notare come, a

differenza del precedente, solo un modello fa la previsione corretta e corrisponde a Xg_boost, l'unico che aveva sbagliato in precedenza.

Figura 16 inserimento opusflow

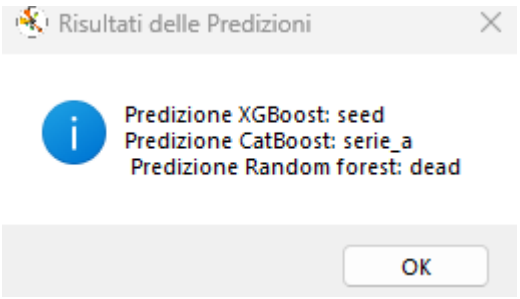


Figura 17 predizioni per opusflow

La successiva azienda rappresentativa per il series è la startup devrev. Devrev è un'azienda che ha creato una piattaforma in grado di unire le funzioni di sviluppo del prodotto con il supporto dei clienti.

In questo caso, come mostrato dalle figura 19, nessuna predizione è corretta.

Figura 18 inserimento devrev

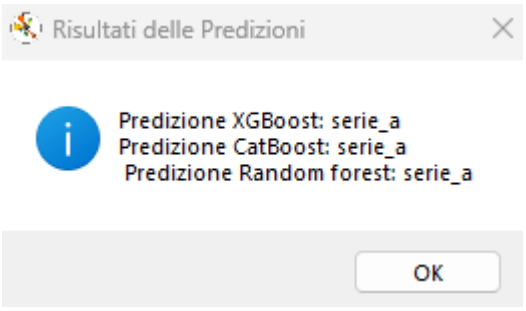


Figura 19 predizioni per devrev

Possiamo riassumere i risultati delle predizioni in questo piccolo schema, che prende ispirazione dalla matrice di confusione. Le righe rappresentano le aziende predette(per semplicità ho scritto solo le fasi senza indicare il nome delle startup), e le colonne rappresentano i classificatori. Nell'intersezione tra righe e colonne notiamo una x quando la predizione è sbagliata e una spunta quando è corretta. Di seguito il grafico:

fase	Random forest	Cat boost	Xg boost
dead	✓	✓	✓
Seed	✓	X	✓
Series	✓	✓	✓

Da questo grafico si evince che il classificatore che ha predetto le fasi in maniera più corretta è il random forest con 5 predizioni su 7 corrette, seguito dall' Xg boost con 4 predizioni su 7 corrette e il catboost con sole 2 su 7 corrette.

4.3.2 I risultati dell'addestramento

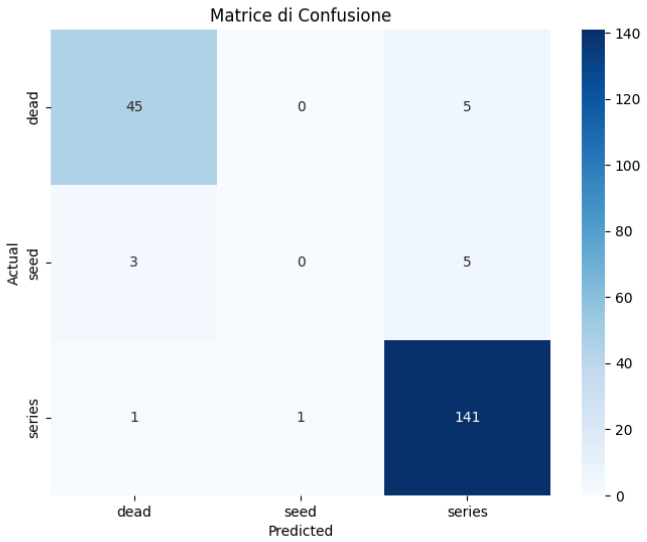
Ora verranno invece illustrati i risultati ottenuti in seguito alla divisione del dataset in training e test set.

4.3.2.1 Random Forest

Le metriche che caratterizzano il random forest sono riassunte in questo schema:

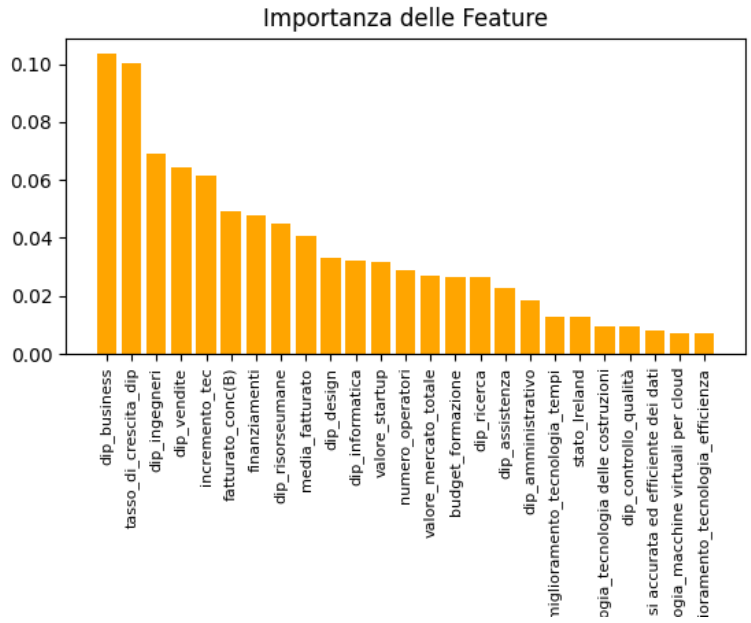
metrica	Valore
Accuratezza	0.9252
Precisione	0.6544
Richiamo	0.6589
F1	0.6527
Deviazione standard accuratezza	0.0405
Deviazione standard precisione	0.1067
Deviazione standard richiamo	0.0940
Deviazione standard f1	0.0984

La matrice di confusione, relativa all'ultima esecuzione, del random forest è la seguente:



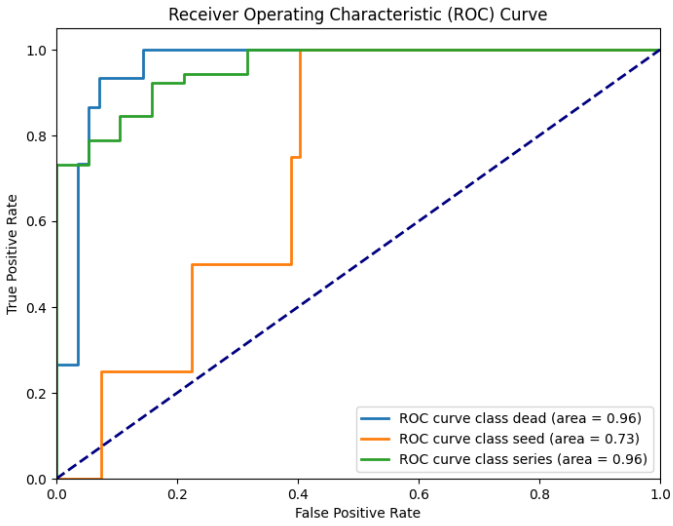
Come si può notare la parte più scura indica un maggior numero di predizioni per quella determinata classe. Per le aziende nella fase seed le predizioni sono praticamente tutte sbagliate. Infatti delle 8 presenti 5 sono state etichettate come series e 3 come seed.

Le feature che hanno influenzato maggiormente le predizioni sono le seguenti:



Si può notare come in realtà l'influenza della singola feature è davvero bassa pari al massimo al 10%.

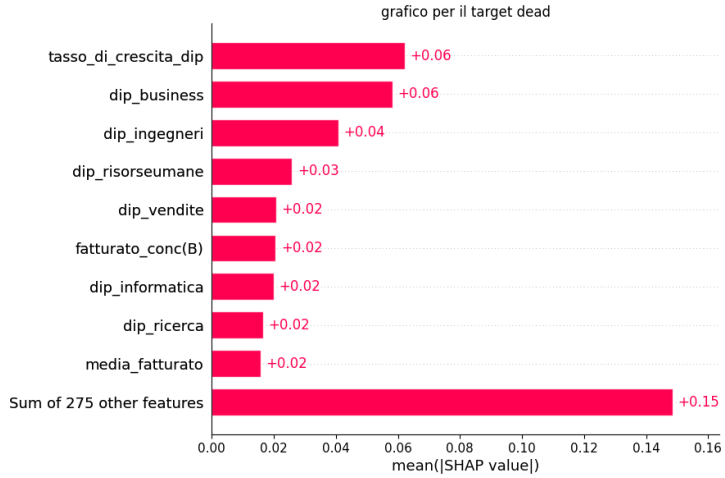
Il grafico di roc è invece il seguente:



Anche da qui si può notare come c'è un enorme discrepanza tra classi diverse: la funzione relativa al seed varia molto più velocemente rispetto alle altre perché il numero di esempi è nettamente inferiore.

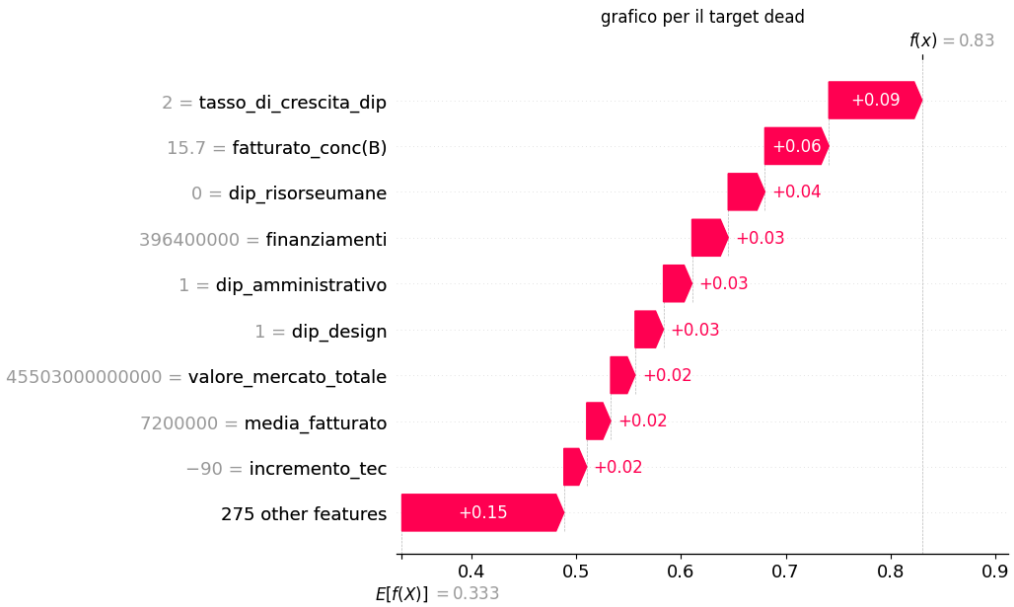
Ora verranno analizzate l'influenza delle features sulla predizione di una determinata classe. In particolare vedremo i grafici creati attraverso la libreria shap ovvero waterfall, summary e bar plot per ogni singola classe.

Per la classe dead il bar plot è il seguente:

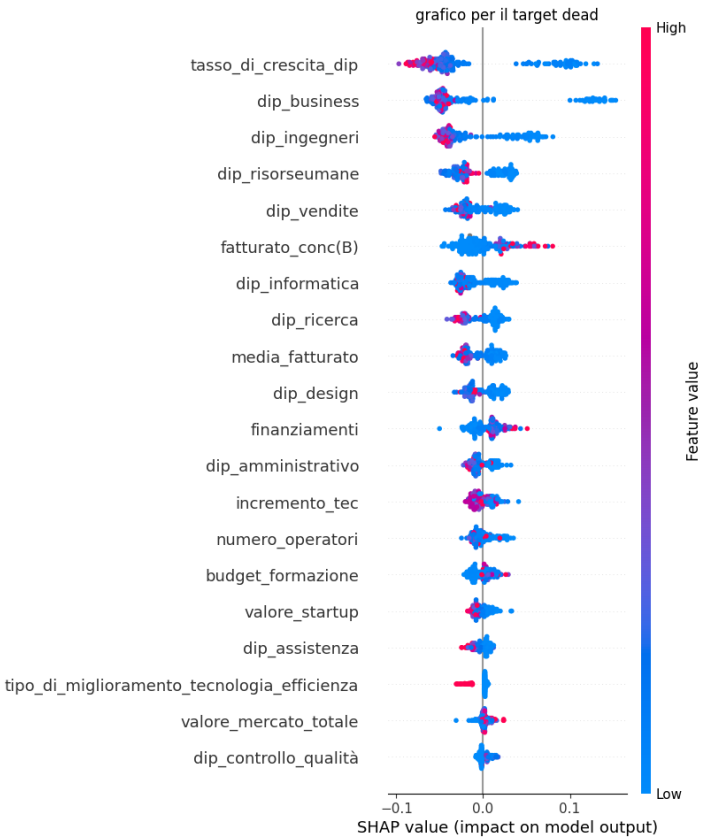


Anche qui si può notare come il tasso di crescita, i dipendenti ingegneri e business sono le feature che influenzano maggiormente.

Il waterfall plot invece si presenta così:



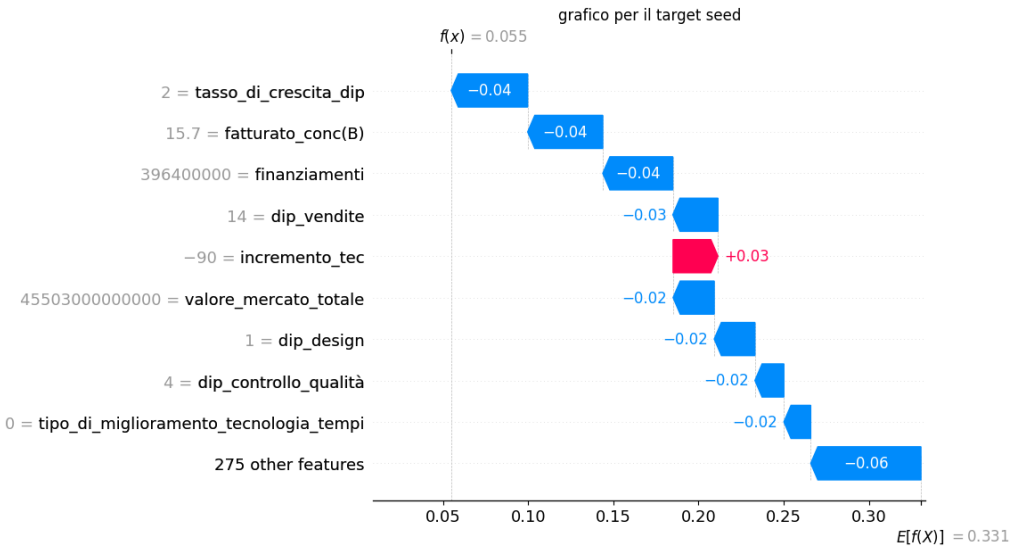
La funzione E[f(x)] rappresenta il valore che ci si aspetta come output del modello e ogni riga mostra quanto ogni features contribuisce in positivo o negativo alla predizione finale rappresentata da f(x). Anche in questo caso troviamo come feature più influenti il tasso di crescita.



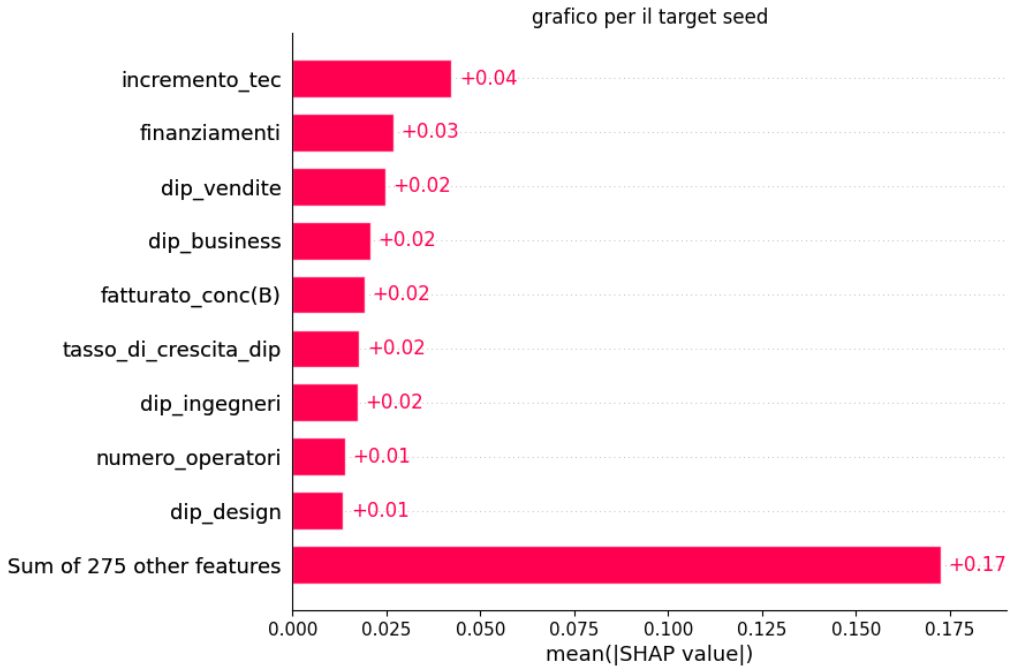
Questo di sopra è invece il grafico summary. Anche questo è ordinato in maniera decrescente dall'alto verso il basso in base al contributo delle features per la predizione. Ogni istanza è rappresentata da un pallino e la posizione è determinata dall'utilità di quella singola istanza per la predizione. Il colore invece indica il valore più è alto più è rosso, più è basso più è blu.

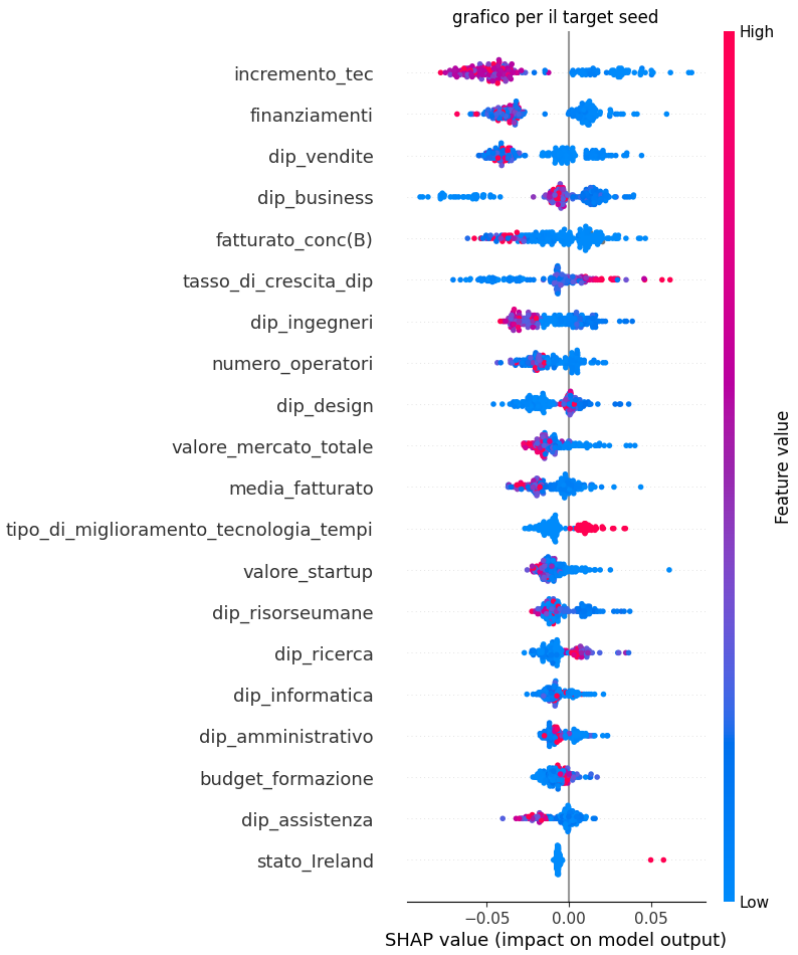
Anche in questo grafico si può notare che il tasso di crescita, gli ingegneri e i dipendenti del business sono le feature più importanti.

Ora mostrerò i grafici per la classe seed:



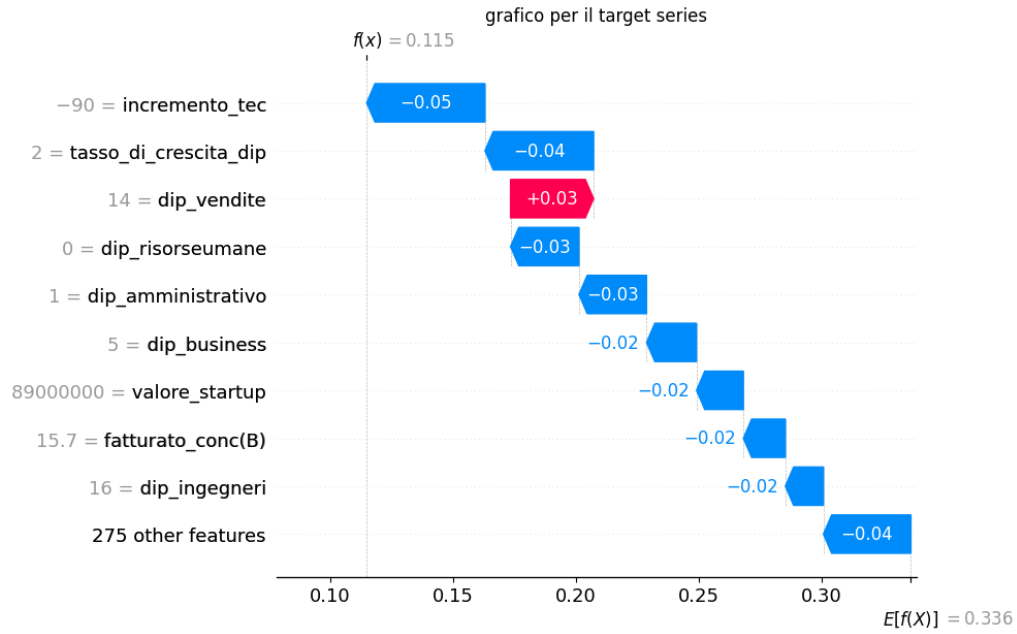
In questo caso si può notare come al diminuire del tasso di crescita, del fatturato e dei finanziamenti si determina più facilmente la predizione.



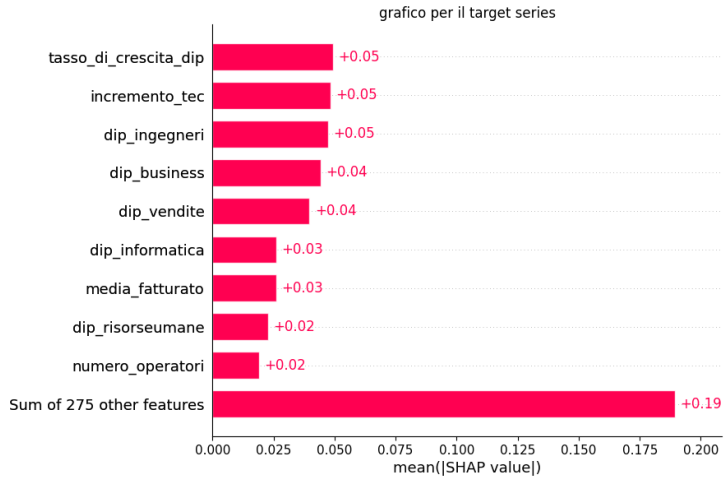


Il motivo per cui questi grafici si presentano in questo modo sta nel fatto il dataset non ha troppi esempi di aziende nella fase seed, pertanto non sono accurati e affidabili quelli realizzati, considerando che il numero di istanze usati per il test della classe seed sono sole 8.

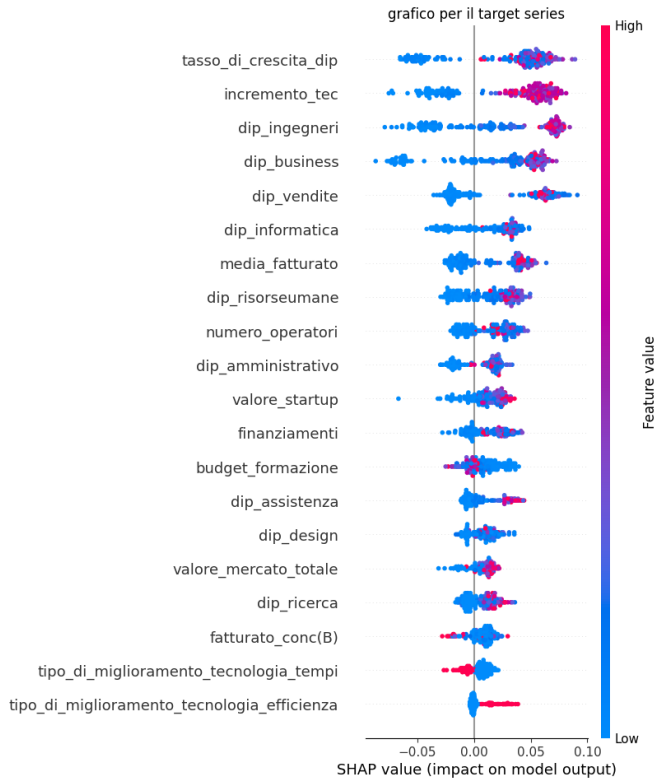
Ora verranno mostrati i grafici che riguardano il valore series per la variabile target:



L'incremento tecnologico, che è la features che indica l'apporto percentuale sull'impatto che una nuova tecnologia ha in termini di benefici, è molto importante. Qui al diminuire di quest'ultimo e del tasso di crescita è possibile prevedere i valori del target series.



Anche il bar plot evidenzia come feature più importanti il tasso di crescita e l'incremento tecnologico.



Il summary plot conferma quanto detto in precedenza rilevando come le feature più importanti per questa classe assumono spesso valori elevati (la maggior parte delle istanze è colorata di rosso).

valori di stage	Feature più importanti
Dead	Tasso di crescita dipendenti, dipendenti business e dipendenti ingegneri
Seed	Tasso di crescita dipendenti, incremento tecnologico
series	Tasso di crescita dipendenti, incremento tecnologico, dipendenti ingegneri

In generale possiamo concludere che per il random forest una feature importantissima per la previsione è il tasso di crescita dei dipendenti. Il fattore dipendenti è presente anche con i dipendenti ingegneri e business utili a prevedere la classe dead e series. Un fattore chiave quindi per la previsione in questo caso risulta essere proprio la parte legata ai dipendenti dell'azienda.

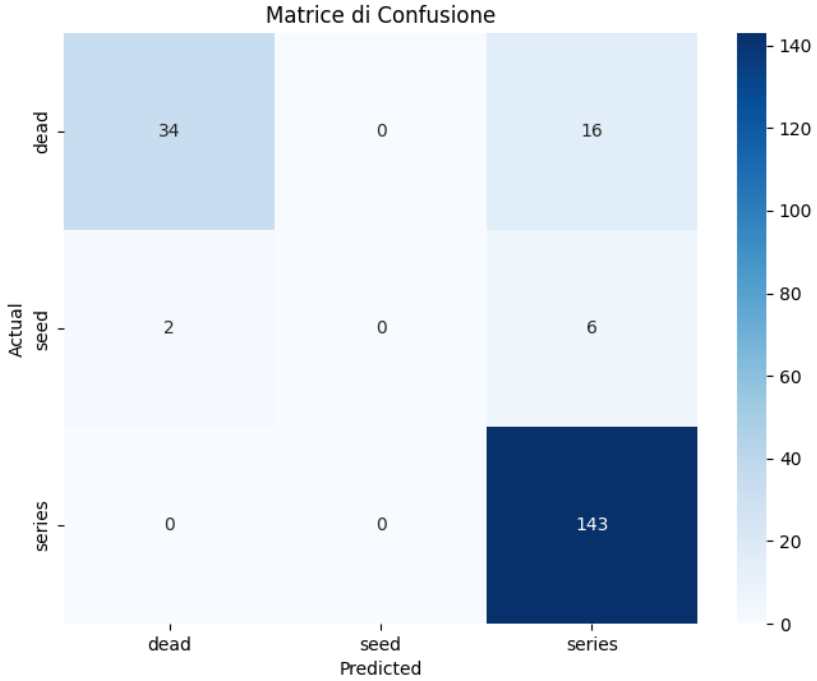
4.3.2.2 Cat boost

Il cat boost si presenta come il modello più debole sulla carta, viste le scarse prestazioni ottenute in fase di test sui casi di studio. Di seguito i valori delle metriche riassunti nella tabella:

metrica	Valore
Accuratezza	0.8802
Precisione	0.6240
Richiamo	0.6067
F1	0.6015
Deviazione standard accuratezza	0.0717
Deviazione standard precisione	0.1365
Deviazione standard richiamo	0.1236
Deviazione standard f1	0.1323

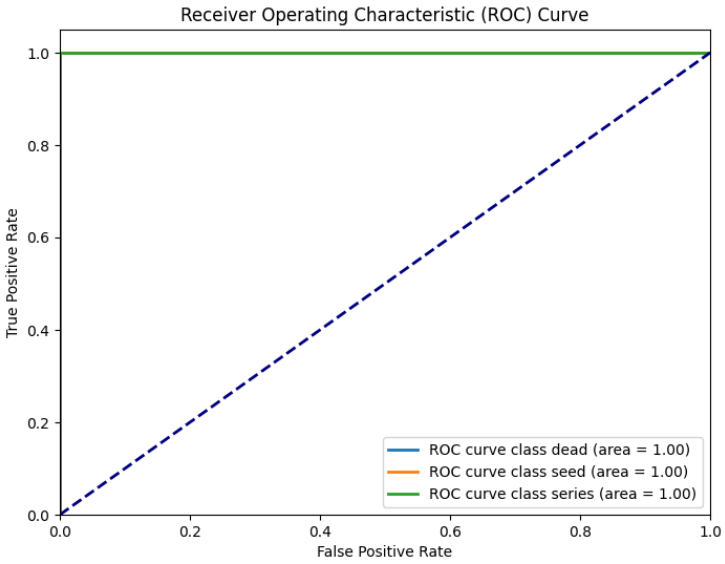
Come si può notare rispetto al random forest le prestazioni anche in termini di metriche sono decisamente inferiori.

La matrice di confusione è la seguente:

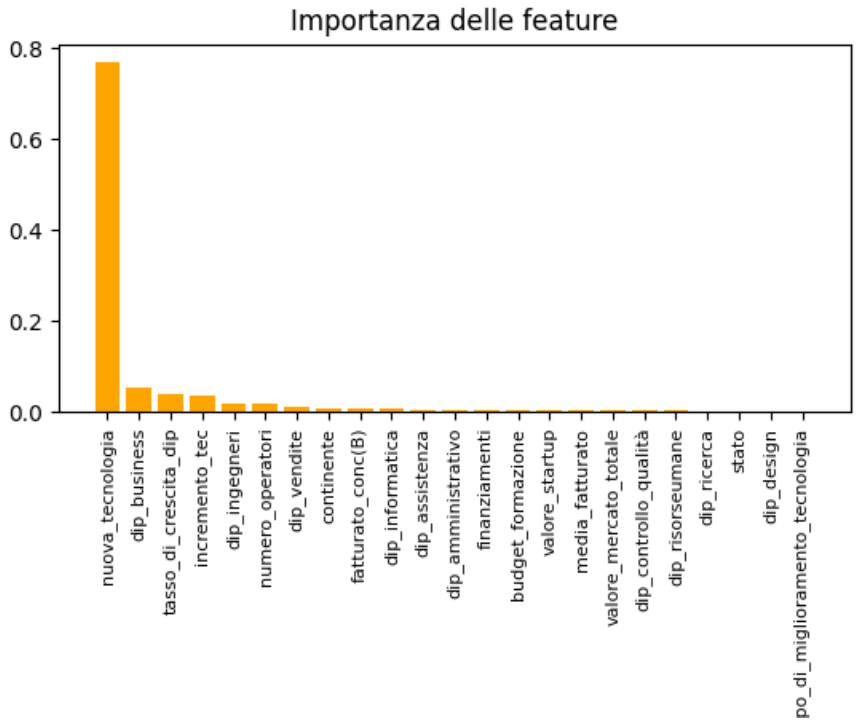


Anche qui si può notare una netta differenza sulle predizioni delle classi, la classe dead ha sole 34 previsioni azzeccate a dispetto delle 50 complessive, mentre la classe seed ne ha 0. La classe series invece è stata predetta al 100% correttamente.

Il grafico della curva di roc è il seguente:

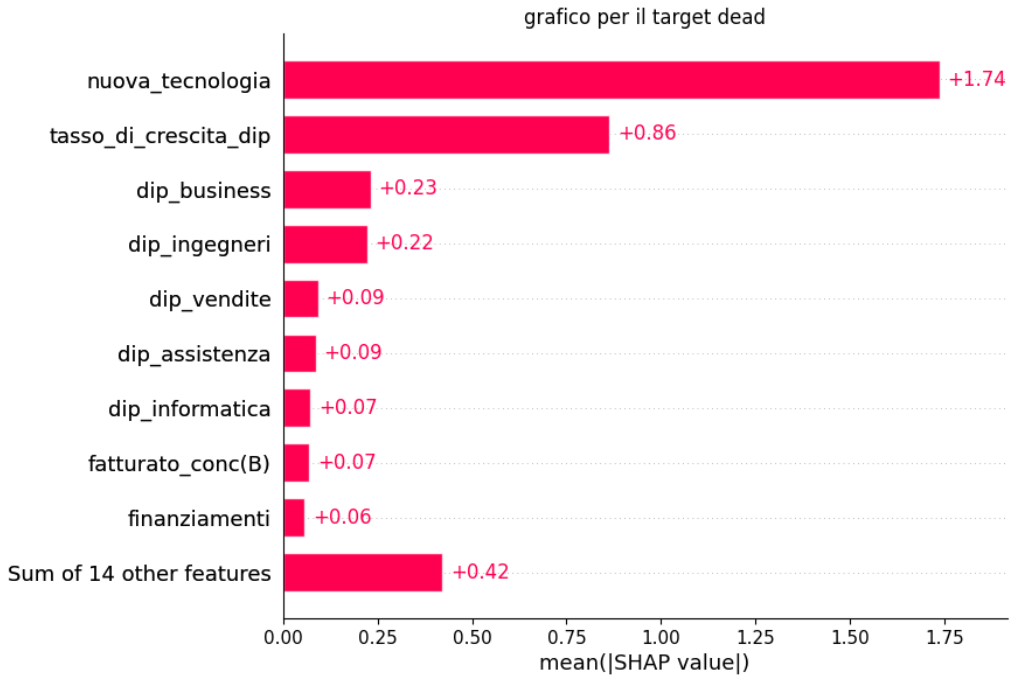


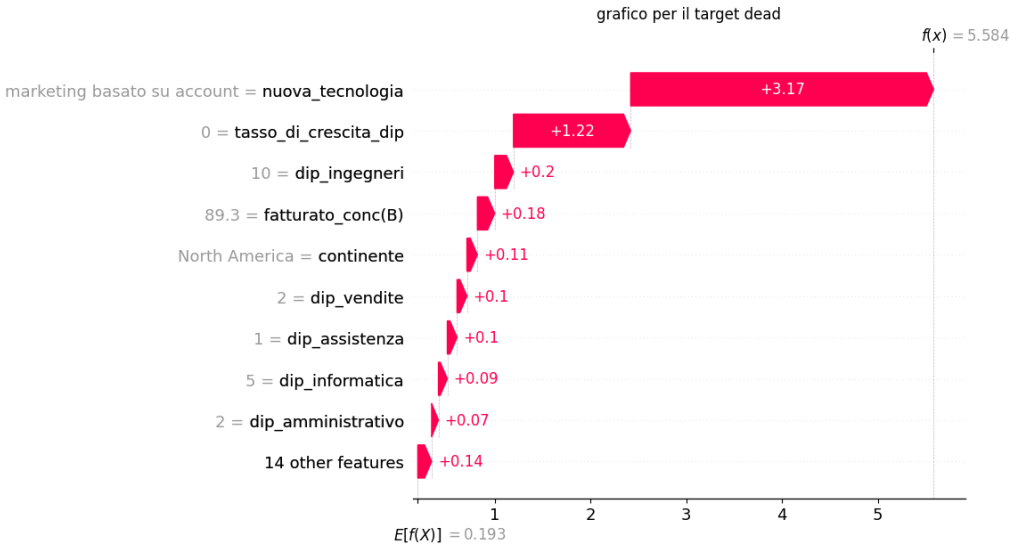
In questo grafico si nota più di tutte una cosa: tutti i possibili valori della variabile target coprono tutta l'area del grafico, ciò significa che tutte le predizioni sono corrette.



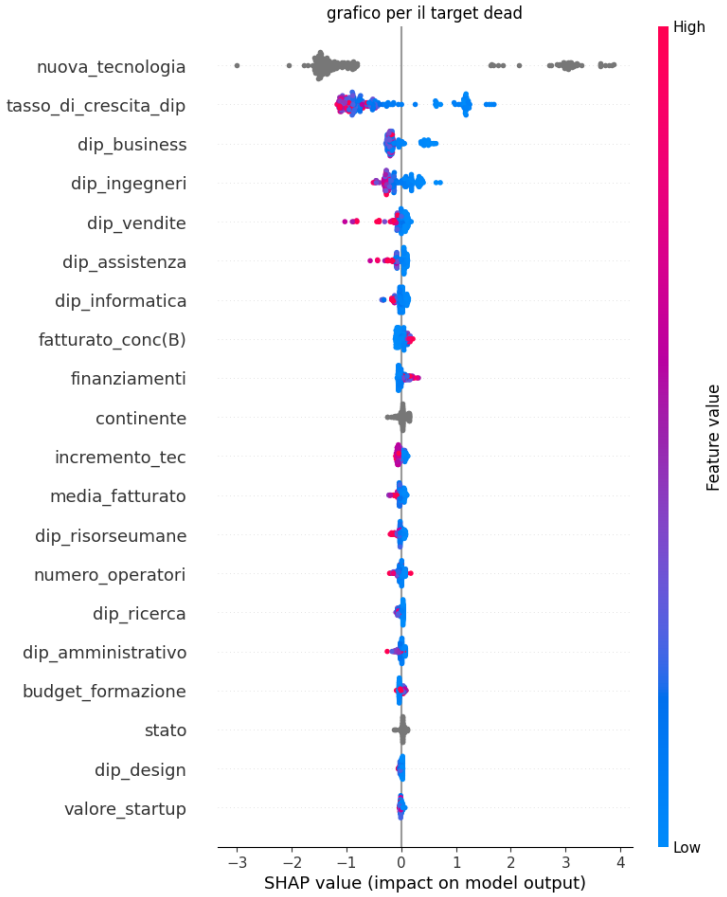
Qui invece si nota come il modello sia in realtà molto legato ai dati di addestramento: la feature prevalente è nuova tecnologia con un'influenza pari quasi a 0.8. Le altre sono davvero poco rilevanti.

Di seguito mostro i grafici bar plot, waterfall e summary per l'explanability della classe dead per il modello catboost.



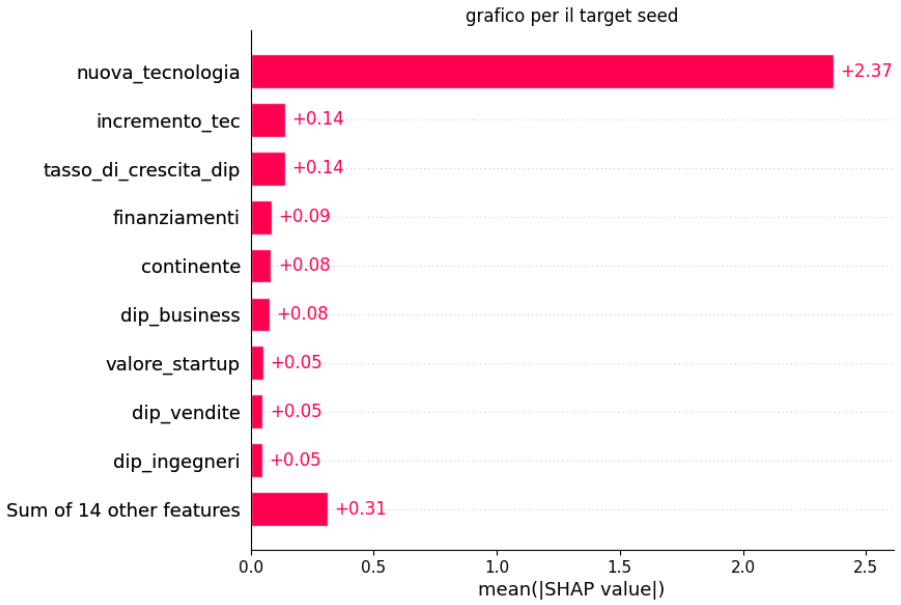


Come si può notare la feature categorica nuova tecnologia va ad incidere positivamente sulla previsione. Anche il tasso di crescita dipendenti è importante mentre le altre feature contano ben poco.

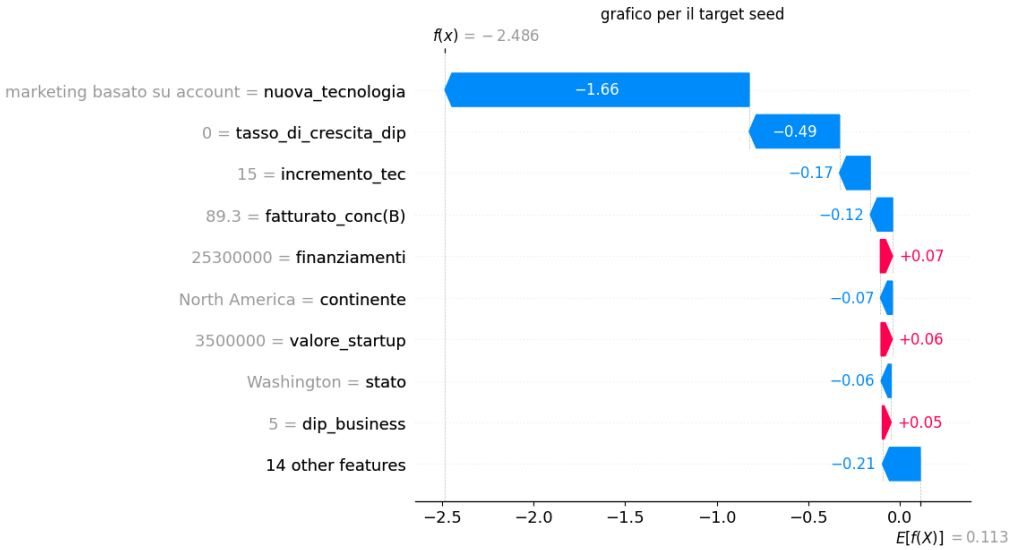


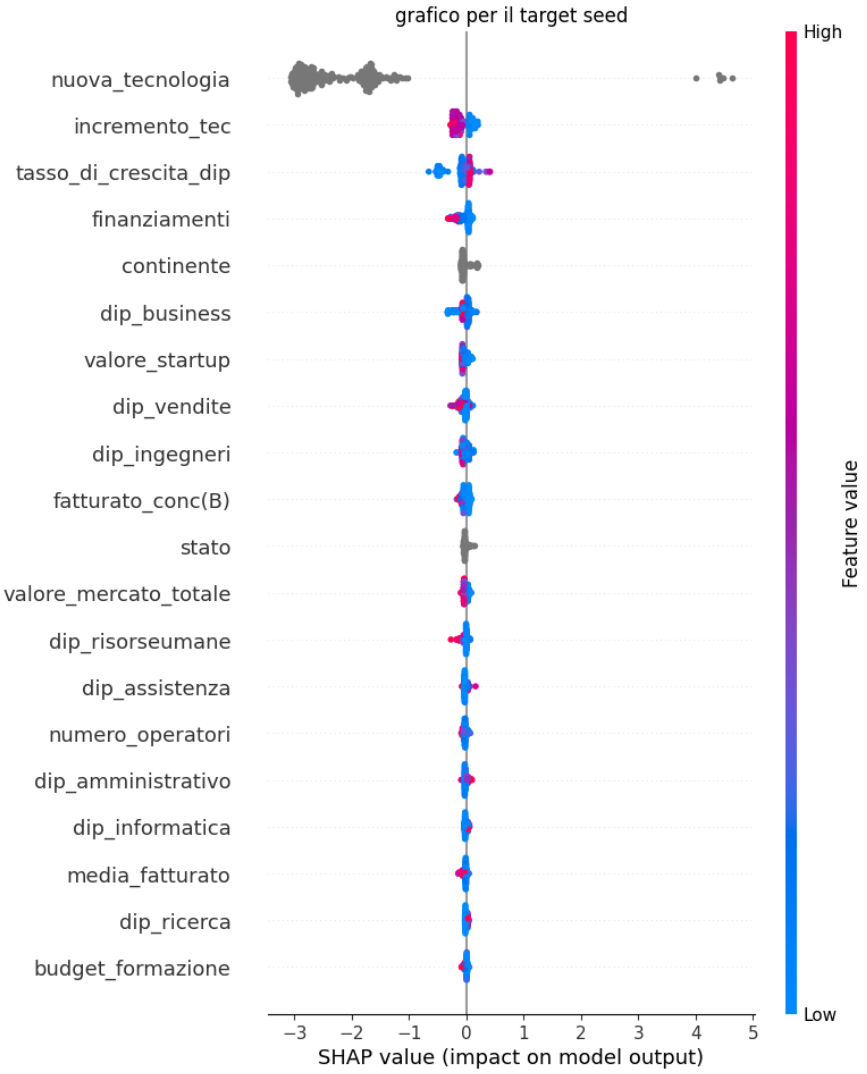
Anche il summary plot conferma ciò che si è affermato per gli altri 2 grafici: la nuova tecnologia e il tasso di crescita dipendenti sono decisamente le features più influenti.

Ora vediamo per la classe seed:



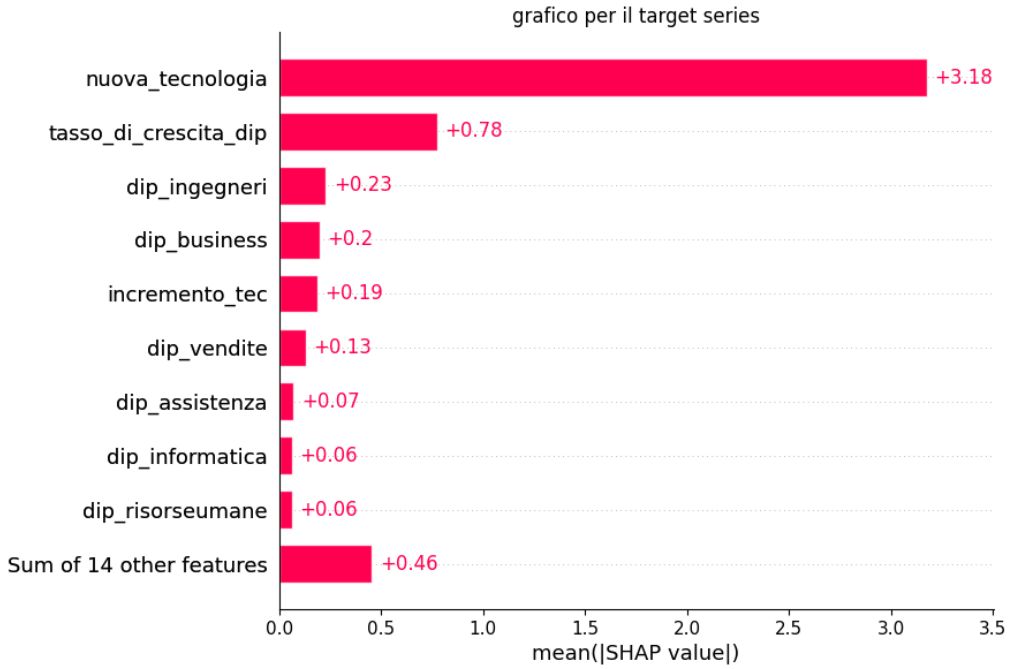
Ancora una volta la nuova tecnologia molto influente come si può notare da entrambi i grafici.

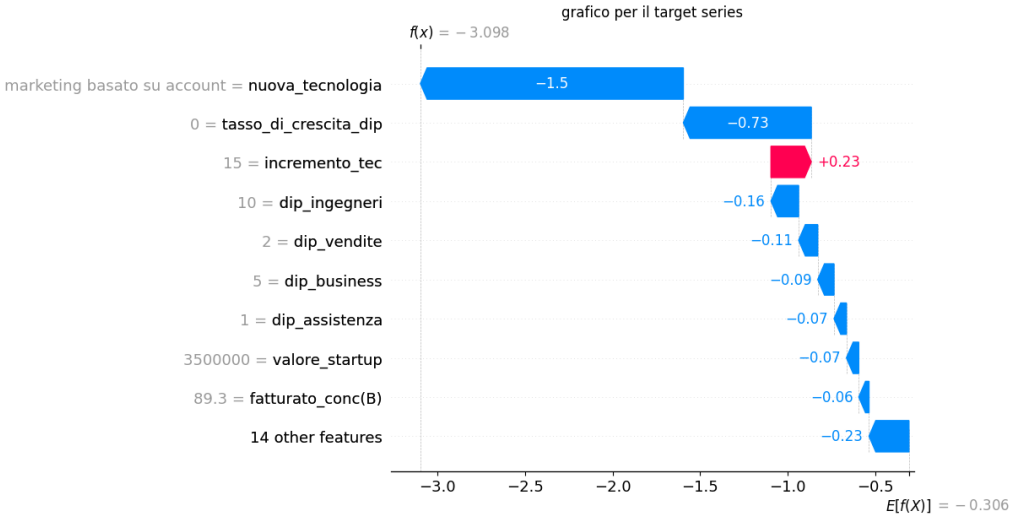




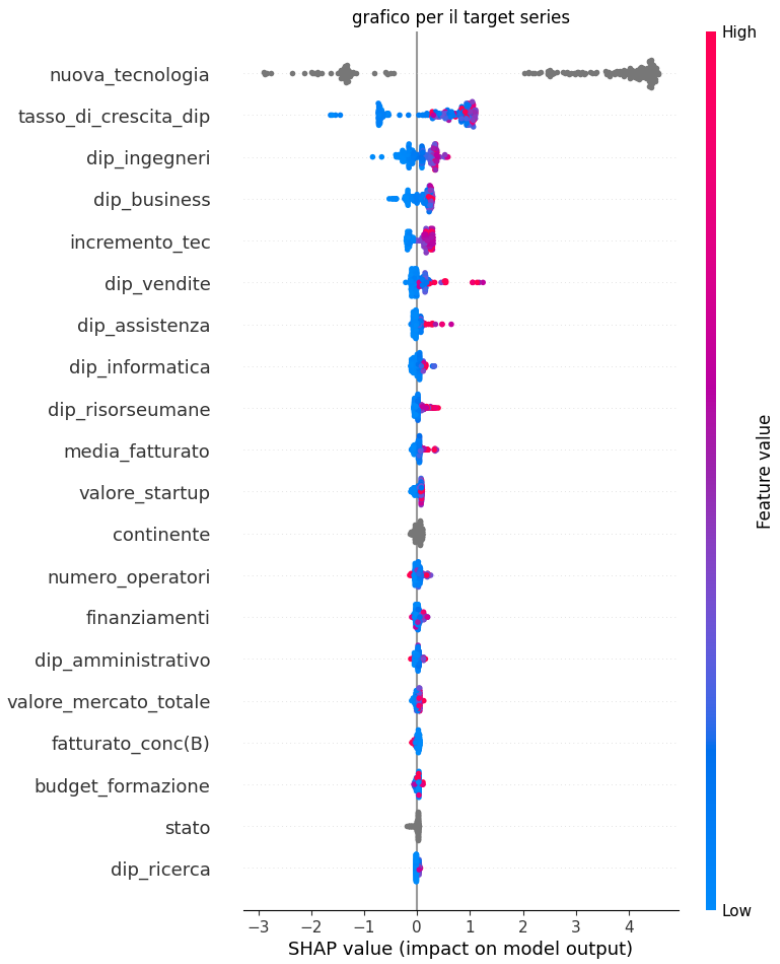
Anche per questo modello vale lo stesso discorso fatto in precedenza: gli esempi per la classe seed sono troppo pochi per poter determinare con esattezza le feature più influenti.

Si procede con l'analisi dei risultati dei grafici della classe series:





Ancora una volta la nuova tecnologia è molto importante per la predizione, solo che questa volta influenza negativamente la determinazione del risultato.



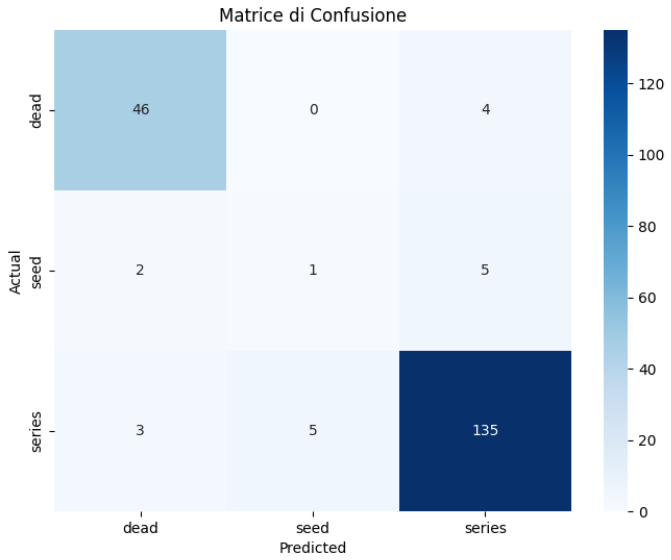
Senza far uso della tabella riassuntiva è possibile subito concludere come la nuova tecnologia sia la feature più importante seguita a ruota dal tasso di crescita dipendenti.

4.3.2.3 Xg boost

L’xg boost invece è un modello che ha avuto dei risultati intermedi sia nella simulazione con i casi di studio sia nel confronto con il test set.

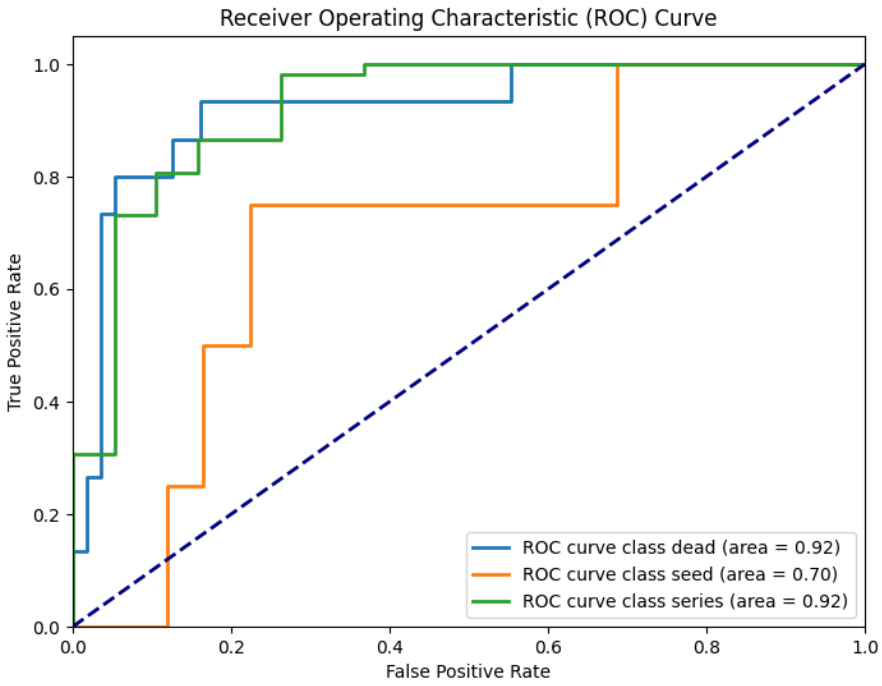
metrica	Valore
Accuratezza	0.9055
Precisione	0.6441
Richiamo	0.6496
F1	0.6453
Deviazione standard accuratezza	0.0199
Deviazione standard precisione	0.0559
Deviazione standard richiamo	0.0466
Deviazione standard f1	0.0506

La matrice di confusione è la seguente:



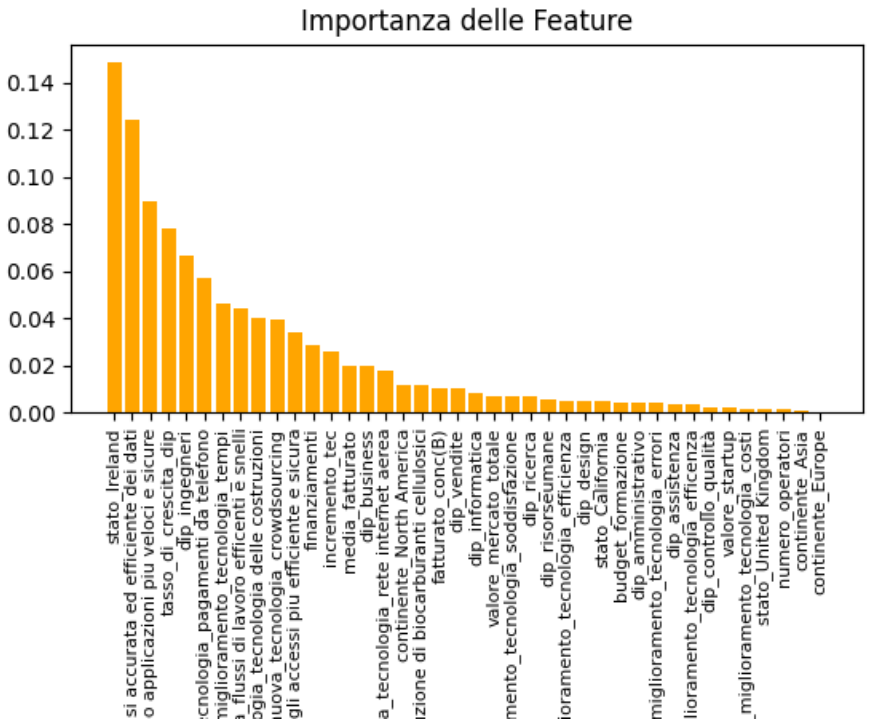
Da questo schema si può rilevare come il modello sia finalmente riuscito a predire un'istanza della classe seed. Buone le prestazioni anche per quello che riguarda la classe dead e la classe series.

Di seguito il grafico riguardante la curva roc:



Qui si può notare, oltre alle considerazioni legate al dataset affrontate nel random forest, come nessuna classe è stata predetta con precisione massima.

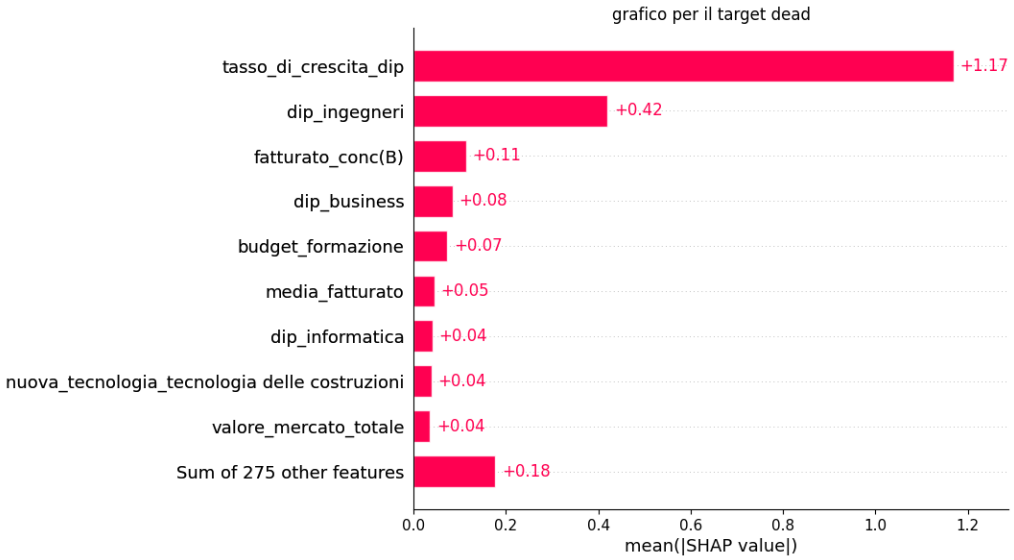
Di seguito il grafico riassuntivo delle features:

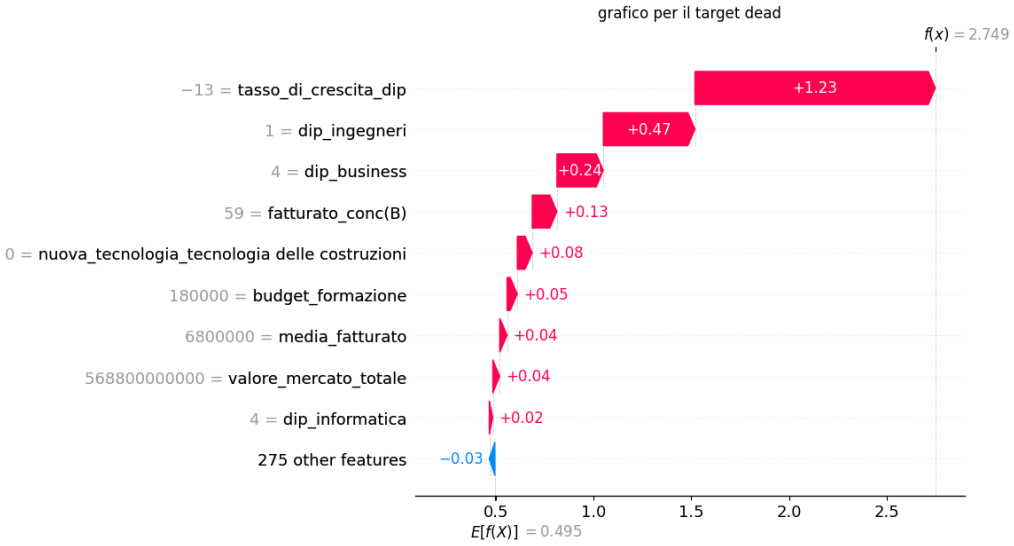


Nel grafico delle feature realizzato con matplotlib si può notare la presenza dello stato Ireland e di alcuni benefici all'azienda introdotti dalle nuove tecnologie, riferiti quindi alla feature tipo_miglioramento_tecnologia, oltre che il tasso di crescita dei dipendenti.

Adesso si va più in profondità con i grafici XAI:

per quanto riguarda la classe dead i risultati sono i seguenti:

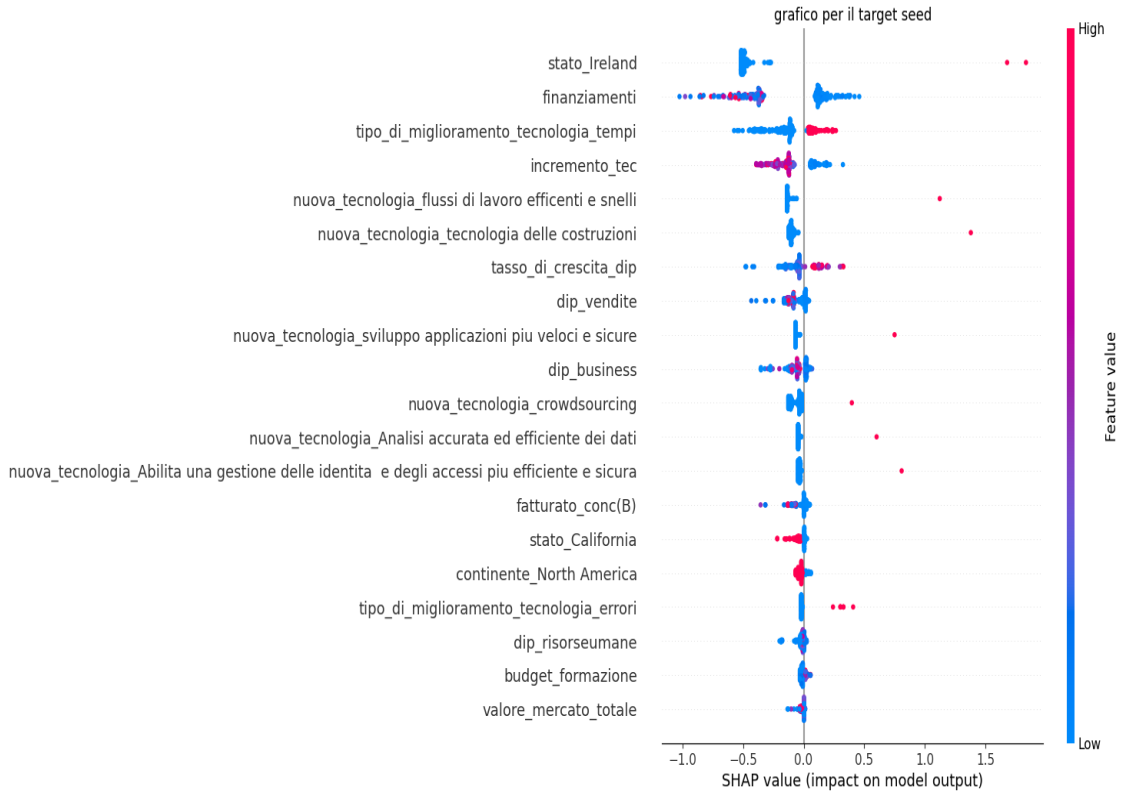




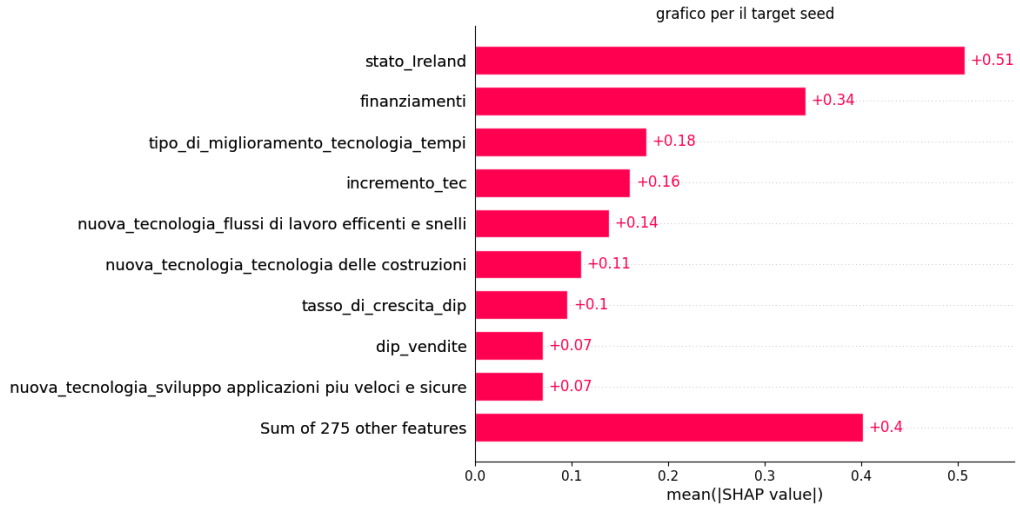
Qui il tasso di crescita e i dipendenti ingegneri ne fanno da padrone.

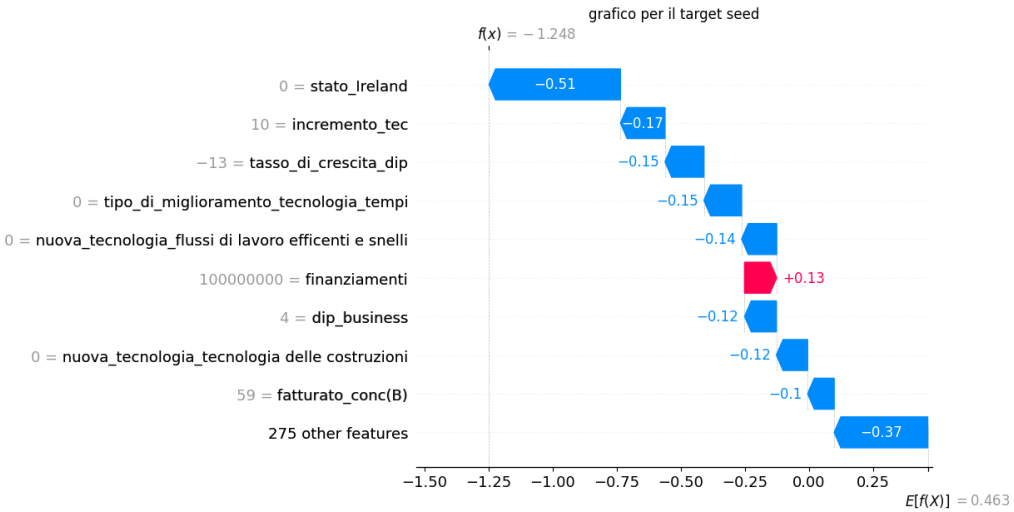
Anche nel summary si può notare come il tasso di crescita dipendenti e i dipendenti ingegneri siano le features più rilevanti.

Ora vediamo per quanto riguarda il seed:

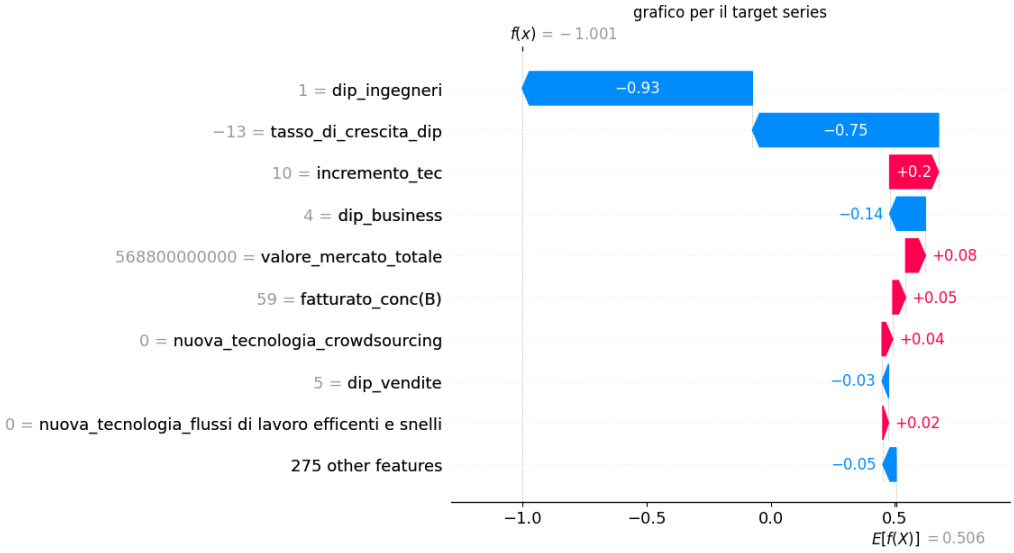


I valori sono molto lontani tra loro e a conferma di ciò c'è il fatto che sono davvero poche le istanze presenti.



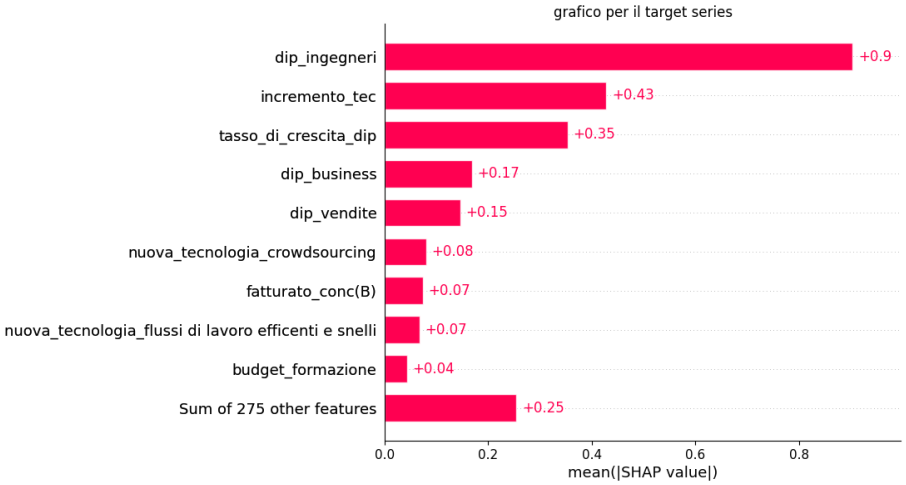


Infine i grafici per la classe series:



I dipendenti ingegneri e il tasso di crescita sono nettamente i più influenti qui

Qui si può notare come la feature che riguarda i dipendenti ingegneri presenta diversi esempi non distanti tra loro, diversamente da quanto mostrato in precedenza per il summary plot del seed



Ancora una volta i dipendenti ingegneri sono la feature più importante.

valori di stage	Feature più importanti
Dead	Tasso di crescita dipendenti, dipendenti ingegneri
Seed	Stato ireland, finanziamenti
Series	Dipendenti ingegneri, incremento tecnologico

Da questo schema possiamo dedurre che una parte importante per comprendere il round di finanziamento di una startup sono i dipendenti ingegneri. Nonostante ciò non si trova una feature troppo più importante di altre, in grado di determinare maggiormente da sola lo stage in cui l'azienda si trova.

Conclusioni

L'obiettivo di questa tesi era quello di analizzare le startup in un contesto di change management e prevedere il loro andamento in un contesto reale.

Si è partiti analizzando in primis cosa è una startup e cosa è l'intelligenza artificiale approfondendo gli ambiti delicati che essa tocca. Si è poi visto dal punto di vista prettamente pratico come è stato realizzato il dataset e quali sono i classificatori utilizzati.

Chiaramente il numero di istanze realizzate per il dataset è irrisorio, 250 istanze sono davvero poche per ottenere risultati confortanti. Nonostante ciò, il random forest ha risposto piuttosto bene ai test a cui è stato sottoposto. Ovviamente per poter rendere il progetto applicabile ad un caso reale c'è bisogno di ampliare il dataset di almeno altre 1000 istanze ed è anche necessario renderlo uniforme ovvero avendo quasi lo stesso numero di aziende per ogni classe della variabile target. Nel mio caso la ricerca su internet non mi ha fornito dati omogenei per tutte le classi della variabile target per cui c'è un evidente disparità numerica di istanze di classe che la variabile target può assumere.
Inoltre il sistema non fa una previsione sulla possibile evoluzione dell'azienda ma realizza una previsione su una "fotografia" scattata in quel momento. Non prevede quindi il successo o meno di un'azienda ma stabilisce soltanto in quale stage la startup si trova attraverso le proprie caratteristiche. Di conseguenza non è realmente in grado di prevedere l'impatto del change management sui ricavi, o sui finanziamenti, ma potrebbe dire, con una nuova configurazione aziendale, se l'azienda andasse avanti, ottenendo nuovi finanziamenti con un nuovo round, o tornasse indietro.

Oltre a ciò, il progetto è molto settorializzato per il settore tech, di conseguenza, questo progetto non è possibile applicarlo a tutte le startup ma

solo a quelle che appartengono al settore tecnologico. Di conseguenza le features che influenzano le predizioni potrebbero certamente variare se lo studio prendesse in esame startup appartenenti ad un altro settore.

I risultati ottenuti non permettono di emanare una sentenza sulle feature che maggiormente influiscono sulla scelta delle classi.
Il modello xg boost evidenzia infatti che un fattore scatenante per la comprensione del round di finanziamento sono i dipendenti che essi siano business, ingegneri o addetti alle vendite così come il tasso di crescita di questi ultimi.
Nel caso del cat boost la feature più influente è senza dubbio la nuova tecnologia seguita dal tasso di crescita dipendenti, come evidenziato dallo schema dei grafici shap ma anche dalla rilevanza delle features realizzata con matplotlib.

I grafici shap realizzati per il random forest non evidenziano una feature prevalente, denotando in realtà quanto l'insieme dei parametri è importante per la predizione. Al contempo il grafico realizzato con l'uso della libreria matplotlib per le features evidenzia come le features più rilevanti sono i dipendenti business, ingegneri e il tasso di crescita dei dipendenti.
In conclusione si può affermare che questo studio evidenzia come i dipendenti siano la parte fondamentale dell'azienda e un numero elevato degli stessi determina un maggiore rilievo per l'azienda nel mercato in cui lavora. Infatti in tutti i modelli oggetto di studio le features più rilevanti riguardavano sempre i dipendenti: per l'xg boost i dipendenti business, per il cat boost il tasso di crescita dei dipendenti mentre per il random forest i dipendenti business, gli ingegneri e il tasso di crescita dei dipendenti.
Per il resto i risultati si sono presentati piuttosto ambigui, pertanto, non si può

determinare una feature che influenza maggiormente ma si può concludere che le feature che riguardano i dipendenti sono le più importanti.

Infine, sebbene il progetto presenti diversi limiti legati principalmente al reperimento dei dati, questo progetto potrebbe essere usato come spunto per sviluppi futuri da imprenditori o analisti per simulare con buona efficacia l'andamento di una startup in un contesto di change management.

Ringraziamenti

Ringrazio mamma e papà e papà e mamma per avermi messo al mondo, per essere stati sempre presenti nella mia vita, per l'educazione che ho ricevuto, per tutte le volte che mi hanno aiutato e sostenuto economicamente ed emotivamente.

Ringrazio i nonni, presenza costante nella mia vita come poche altre, per tutte le curiosità sul passato che hanno provato a tramandarmi, per gli aneddoti e i tantissimi racconti e per avermi viziato molto bene con il cibo che in questi anni mi hanno preparato. Nonna Teresa oggettivamente cucina molto meglio di nonna Maria, di contro nonna Maria mi ha forzato a mangiare tutto e ad esplorare nuovi sapori che se non per costrizione, quello è il piatto quello devi mangiare, non avrei mai assaggiato.

Ringrazio zii e cugini per avermi cresciuto, per avermi aperto a mondi da me ancora inesplorati e per avermi stimolato con la loro attitudine al lavoro, la loro brillantezza e la loro conoscenza. Anche se non lo dico vi stimo tanto.

Ringrazio mio fratello per avermi fatto compagnia per gran parte della mia vita, per avermi donato opportunità che altrimenti non avrei mai pensato di avere e per avermi stimolato ogni giorno a migliorare me stesso con confronti e dibattiti il più delle volte costruttivi. Grazie per aver creduto in me fin da subito e per confidare ancora molta fiducia nonostante a volte io non sia stato proprio il massimo dell'affidabilità.

Un grazie particolare a tutti i miei amici chi dal giorno 0, chi dal giorno 1, chi dal giorno 100 non avete mai fatto mancare la vostra allegria e la vostra vicinanza. Grazie Stino, compagno di viaggi, avventure, risate,sciocchezze,ca**ate e stati mentali alterati. Da quando ti ho conosciuto la mia vita ha avuto una piega decisamente positiva, mi auguro valga lo stesso anche per te.

Grazie Michele per aver costituito l'esempio da seguire per tutta la mia vita e per avermi sempre coinvolto anche quando non c'era il bisogno. Grazie Mitolo per avermi cambiato o provato a cambiare, per avermi persuaso a scoprire l'Internet e per incoraggiarmi a farlo ancora oggi. Sei stato una spalla su cui piangere, tra i pochi in grado di capirmi e tra i pochissimi con cui non ho peli sulla lingua. Grazie Vincenzo per avermi insegnato a prendere in giro la gente senza che loro ne risentano (la maggior parte delle volte) e per avermi involontariamente indicato la strada da seguire. Grazie Robles per essere diventato tanto in poco tempo, per aver stimolato di molto il mio coraggio che senza dubbio è più consistente a causa tua.
Grazie Tiziano, Marco, Fabio e Antonello per aver rappresentato un supporto affidabile per il mio percorso universitario, oltre che amici con cui condividere momenti anche al di fuori delle quattro mura dell'università.
Grazie Lenny per essere stato un secondo padre, un secondo fratello, un amico su cui contare. Per avermi aperto ad una nuova cultura, avermi presentato dei pareri divergenti sulla nazione in cui vivo e per esserti aperto con me a 360 gradi come nessuno lo aveva fatto prima.
Grazie Michele Palmiotto perché, nonostante tutto, mi hai fatto crescere come persona e perché con la tua fiducia hai migliorato la mia autostima.
Grazie Diego per aver tradotto a parole ed in musica ciò che si era sempre manifestato nel mio cervello ma non si era mai tradotto in qualcosa. Mi hai dato la forza per sopportare le critiche ai miei vestiti, al mio taglio di capelli e alle mie scelte. Grazie per avermi fatto sfogare e per avermi regalato sprazzi di felicità ogni giorno. Grazie Epico podcast per avermi intrattenuto nell'ultimo anno facendomi scoprire tante nuove cose interessanti e tanti modi di pensare leggermente differenti.

BIBLIOGRAFIA

- Baldridge R., Curry B. «What Is A Startup?». Forbes Advisor, 2021. URL: https://www.forbes.com/advisor/investing/what-is-a-startup/
- Barrette A. «Business Startup Funding: A beginner's guide». Foundr, 2020. URL: https://foundr.com/articles/building-a-business/finance/funding-a-startup
- Blank S. «Steve Blank: The 6 Types of Startup». The Wall Street Journal, 2013. URL: https://www.wsj.com/articles/BL-232B-1094
- CB Insights, «The Top 12 Reasons Startup Fail». URL: https://www.cbinsights.com/research/startup-failure-reasons-top/.
- Lee Yohn D. «Why Startup Fail». Forbes, 2019. URL: https://www.forbes.com/sites/deniselyohn/2019/05/01/why-startupsfail/?sh=7cc794da28a5.
- Ian Hathaway What startups accelerators really do. URL: https://hbr.org/2016/03/what-startup-accelerators-really-do
- Landabaso M. The smart guide to Innovation-Based Incubators URL: https://op.europa.eu/it/publication-detail/-/publication/aaf1dc0e-41f8-4889-94ab-6eeac68d8b51
- Korunka, C., Frank, H., Lueger, M., & Mugler, J. (2003). The Entrepreneurial Personality in the Context of Resources, Environment, and the Startup Process—A Configurational Approach. Entrepreneurship Theory and Practice, 28(1), 23-42. https://doi.org/10.1111/1540-8520.00030

- Zoltan J. Acs,José Ernesto Amorós, Introduction: The Startup Process URL https://repositorio.uchile.cl/bitstream/handle/2250/127715/Zoltan_J.Acs.pdf?sequence=1
- Bbva, «What is a business angel?». Bbva, 2021. URL: https://www.bbva.com/en/what-is-business-angel/
- Borsa Italiana. « Glossario finanziario - venture capital». URL: https://www.borsaitaliana.it/borsa/glossario/venture-capital.html.
- Kevin Buehler, Rachel Dooley, Liz Grennan, and Alex Singla Getting to know—and manage—your biggest AI risks URL https://www.mckinsey.com/capabilities/quantumblack/our-insights/getting-to-know-and-manage-your-biggest-ai-risks
- C. Huang, Z. Zhang, B. Mao and X. Yao, "An Overview of Artificial Intelligence Ethics," in IEEE Transactions on Artificial Intelligence, vol. 4, no. 4, pp. 799-819, Aug. 2023, doi: 10.1109/TAI.2022.3194503.keywords: {Artificial intelligence;Ethics;Guidelines;Privacy;Government;Systematics;Security;Artificial intelligence (AI);AI ethics;ethical issue;ethical theory;ethical principle}
- Copeland, B.J.. "artificial intelligence". Encyclopedia Britannica, 19 Sep. 2024, https://www.britannica.com/technology/artificial-intelligence. Accessed 20 September 2024.