Politecnico di Torino

Corso di Laurea Magistrale in Ingegneria Gestionale (LM-31)

a.a. 2020/2021

Sessione di Laurea dicembre 2021

Uso dell'Intelligenza Artificiale per la predizione della business interruption

Relatore:
Prof. Guido Perboli
Correlatore:
Prof.ssa Mariangela Rosano

Candidato:
Andrea Del Pero
Matricola: 276189

Sommario

Elenco delle figure

Elenco delle tabelle

Introduzione

La stima del rischio di fallimento e la predizione della crisi d'impresa rappresentano argomenti molto significativi nell'ambito economico e della finanza. Lo stato di salute di un'azienda è di grande importanza per i suoi creditori e investitori in generale, ma non solo, infatti entra in gioco anche una componente sociale se teniamo in considerazione tutti gli altri stakeholders, tra i quali l'esempio più immediato è rappresentato dai dipendenti con la relativa importanza di mantenere un posto di lavoro, ma anche fornitori, clienti ecc.; i costi relativi a situazioni di difficoltà finanziarie e fallimenti sono altissimi e impattanti sia in ambito locale, se pensiamo alle ricadute economiche sul territorio, che in un ambito più generale, se pensiamo che possono innescare crisi a livello globale andando a compromettere la salute di interi settori.

Fino a qualche decennio fa gli unici strumenti a supporto della predizione del fallimento erano metodi basati su modelli di statistica tradizionale, negli ultimi anni, invece, grazie all'evolvere della tecnologia e allo sviluppo di nuovi paradigmi, sono entrati in gioco modelli di apprendimento automatico, più complessi e più affidabili, aprendo così l'era dell'Artificial Intelligence e del Machine Learning anche in questo ambito. Una caratteristica che ha spinto la diffusione di questi modelli è la capacità di sfruttare e processare le grandi quantità di dati che sempre più sono disponibili per le aziende e che, se ben sfruttati, possono trasformarsi in valore aggiunto. Questi strumenti, se inseriti in un contesto strutturato e in sinergia con processi di Risk Management, riescono ad esprimere tutto il loro potenziale e raggiungere un ruolo di supporto molto importante per il decision maker aziendale e non solo.

In particolare, lo scopo della predizione del fallimento è intercettare lo stato di crisi che affligge un'azienda negli anni precedenti all'evento fallimento; in altre parole, l'obiettivo è valutare lo stato di salute finanziaria col fine di prevedere le prospettive future dell'azienda. Tipicamente questo problema è affrontato attraverso modelli di classificazione in due classi, come nel caso di questo lavoro di tesi, che permettono di stabilire se un'azienda fallirà o no attraverso processi data-driven.

L'elaborato è strutturato in sette capitoli principali, descritti brevemente di seguito.

Capitolo 1 **I Rischi**: in questo primo capitolo è trattato l'argomento generale di rischio, analizzandone le definizioni e presentando i principali metodi di classificazione dei rischi.

Capitolo 2 **Gestione del rischio: Risk Management**: il capitolo presenta l'argomento della gestione del rischio illustrando i principi e il framework alla base della disciplina. Un intero paragrafo è dedicato all'analisi del processo di Risk Management attraverso la descrizione nel dettaglio di ogni sua fase.

Capitolo 3 **Artificial Intelligence e crisi d'impresa**: il capitolo in questione tratta inizialmente il concetto di crisi d'impresa e quello di fallimento per definire l'elemento principale del lavoro svolto. Il passo successivo è la presentazione del valore aggiunto dell'Artificial Intelligence e delle motivazioni che hanno portato alla sua diffusione capillare. Nell'ultimo paragrafo è redatta una rassegna dei principali metodi di predizione del fallimento presenti in letteratura attraverso una raccolta di studi.

Capitolo 4 **Data Mining**: in questo quarto capitolo è presentata una introduzione al Data Mining, processo tramite il quale si estrae conoscenza utilizzabile da grandi quantità di dati. In particolare, è descritto il processo e le varie fasi del KDD, Knowledge Discovery in Databases.

Capitolo 5 **Il Modello**: in questo capitolo sono presentate tutte le informazioni relative al modello sperimentale di predizione del fallimento proposto in questo lavoro di tesi. Nel primo paragrafo è definito l'obiettivo della ricerca, per poi passare al secondo paragrafo, che risulta essere il più importante e corposo, nel quale sono descritte e analizzate tutte le tecniche e metodologie sulle quali si basa il modello di classificazione; in particolare, si approfondiscono la raccolta e analisi dei dati, la preparazione e trasformazione dei dati, gli algoritmi della fase di data mining e le metriche utilizzate per valutare la bontà del modello. L'ultimo paragrafo è dedicato alla presentazione dell'ambiente di sviluppo del lavoro e sono illustrati i processi RapidMiner utilizzati nella realizzazione del modello.

Capitolo 6 **Risultati ottenuti**: il sesto capitolo di questo lavoro di tesi è dedicato alla presentazione dei risultati sperimentali ottenuti attraverso le diverse tecniche adottate. In particolare, è effettuata una comparazione tra vari tipi di algoritmi di apprendimento automatico e, inoltre, sono messe a confronto diverse metodologie per gestire lo sbilanciamento dei dati, per il trattamento dei missing value presenti nel dataset iniziale e presentati i risultati di due processi di feature selection.

Capitolo 7 **Conclusioni e sviluppi futuri**: l'elaborato si conclude con un riepilogo del lavoro svolto e dei relativi risultati ottenuti. Inoltre, sono proposti alcuni possibili sviluppi futuri che potrebbero essere intrapresi per cercare di migliorare il modello e facilitare l'applicazione dello stesso nel mondo reale (progettazione di un DSS – Decision Support System).

1 I rischi

In questo capitolo è presentata un'introduzione generale sul concetto di rischio, evidenziando le definizioni più comuni di rischio secondo diversi punti di vista. Successivamente, sono descritti alcuni metodi di classificazione dei rischi, ognuno dei quali prende in considerazione un criterio di divisione differente.

1.1 La definizione di rischio

Il concetto di rischio è universalmente conosciuto, ma quando cerchiamo di fornire una definizione unica e generale ci accorgiamo che è difficile. La definizione di rischio non è univoca nella letteratura, bensì in base all'approccio con cui la si affronta e al contesto di riferimento assume conformazioni differenti. Un primo filone, il più diffuso nell'accezione comune, identifica il rischio come "*Eventualità di subire un danno connessa a circostanze più o meno prevedibili*" (Dizionario di lingua italiana Treccani). La visione del rischio che emerge in questo caso è totalmente negativa e collegata a eventi avversi che recano un danno.

Un altro punto di vista deriva dalla Norma UNI ISO 31000:2018, riferimento internazionale riguardo gli standard aziendali, che definisce il rischio come "*Effetto dell'incertezza in relazione agli obiettivi*". Un elemento fondamentale emerge dalle note comprese nella norma; infatti, la Nota 1 recita che "*Un effetto riguarda ciò che potrebbe essere diverso da quanto atteso. Può essere positivo, negativo o di entrambi i segni e può affrontare, creare o avere come risultato in cascata successive nuove opportunità e minacce*". Questa seconda definizione non affronta il rischio solo con un approccio negativo, ma allarga il concetto a quello di opportunità. Questo concetto è ripreso e riproposto da numerose istituzioni e standard, ad esempio il Project Management Institute (PMI) descrive il rischio come "*An uncertain event or condition, that if it occurs, has a positive or negative effect on a project's objective*".

Un elemento aggiuntivo è introdotto dall'Institute of Internal Auditors che propone la definizione: "*The possibility of an event occurring that will have an impact on the achievement of the objectives. Risk is measured in terms of impact and likelihood*". Nel concetto di rischio è insito, oltre alla probabilità di accadimento dell'evento, anche l'impatto potenziale sugli obiettivi che quell'evento avrebbe se si concretizzasse.

Se restringiamo il campo al contesto aziendale, possiamo definire i rischi aziendali come "*l'insieme dei possibili effetti positivi (opportunità o upside risk) e negativi (minacce o*

downside risk) di un evento inaspettato sulla situazione economica, finanziaria, patrimoniale e sull'immagine dell'impresa".

Secondo Hillson (2004, 2009, 2012), esperto riconosciuto a livello internazionale in materia di rischio, è necessario considerare nell'analisi la relazione tra rischio e incertezza, tenendo in considerazione che tutti i rischi sono incerti, ma non tutte le incertezze sono rischi. In particolare, scrive: "*risk in uncertainty that matters*". In accordo con Knight (1921), sostiene che il rischio sia una incertezza misurabile, ovvero che il rischio sia relativo ad un evento di cui si conoscono le probabilità di accadimento, calcolabili e stimabili, e gli effetti che può causare qualora si concretizzasse. L'incertezza risiede nell'incognita di quale effetto si realizza nella situazione specifica. Un'altra relazione che emerge è quella tra rischio e obiettivo; infatti, non tutte le incertezze sono rischi, solo le incertezze che interessano gli obiettivi, ovvero "*that matters*", possono essere definiti rischi. Questo è rappresentato da un ulteriore definizione proposta da Hillson che esprime meglio il concetto di rischio: "*risk is uncertainty that, if it occurs, will affect achievement of objectives*".

Nell'analizzare le differenti definizioni si può trovare un punto comune a tutte: il rischio è un aspetto intrinseco con il futuro e la sua indeterminatezza. Infatti, i rischi dipendono dall'interazione tra obiettivi, ovvero cosa deve accadere, e le incertezze, cosa può accadere. Questo è declinato in due modi differenti in base all'approccio assunto: in gran parte delle definizioni è presente l'approccio negativo al rischio che si traduce in un evento che se accade reca un danno e, soprattutto, pregiudica gli obiettivi prefissati dall'organizzazione ostacolandone il raggiungimento; solo alcune definizioni, derivanti da approcci al risk management più moderni e aziendalistici, concepiscono anche un approccio opportunistico al rischio.

Come già anticipato, esistono innumerevoli definizioni di rischio, da quelle più generali a quelle definite rispetto a un determinato settore industriale o contesto; il rischio assume diversi significati se si sta parlando di medicina, piuttosto che di ingegneria o di economia. Risulta di fondamentale importanza, per ogni organizzazione qualsiasi essa sia, la scelta della definizione che meglio si adatta agli scopi della stessa e alle condizioni di contorno che la influenzano.

1.2 Classificazione dei rischi

Dopo aver definito cosa sia un rischio e facendo riferimento ad un contesto operativo di tipo aziendale, è utile specificare alcuni criteri di classificazione del rischio. Anche in questo caso esistono differenti approcci a seconda delle variabili considerate. I principali sono elencati di seguito:

1. Il primo metodo per discriminare i rischi è basato sulla valutazione della natura interna o esterna all'azienda dell'evento che li genera.

   I *rischi interni* possono essere controllati dall'organizzazione aziendale attraverso il management poiché derivano da eventi generati all'interno dell'azienda stessa. Esempi rappresentativi di questa tipologia sono rischi che riguardano aspetti commerciali e di mercato, come la relazione con i clienti, la politica sui prezzi; aspetti tecnici/operativi, intesi come tecnologie, processi; aspetti umani, legati al mondo delle risorse umane in generale.

   I *rischi esterni* sono di origine esogena e non sono controllabili da parte dell'azienda ma sono comunque contrastabili con opportune tecniche di gestione del rischio. Fanno parte di questa tipologia i rischi naturali, pensiamo ad esempio ad alluvioni, terremoti, uragani o, come è accaduto nell'ultimo anno, una pandemia; i rischi economici/finanziari, come i tassi di cambio valuta o i tassi dei prestiti bancari; e ancora i rischi politici, come l'approvazione di una specifica legge che interessa il business della azienda di riferimento.

2. Un altro criterio di classificazione si basa sul possibile segno (positivo e/o negativo) che il rischio può assumere. In questo caso individuiamo i *rischi puri* (o unilaterali) e i *rischi speculativi* (o imprenditoriali o bilaterali).

   I rischi puri derivano da eventi, che se si avverano, hanno come unico esito un danno, una perdita. Alcuni esempi sono l'insorgere di un incendio, il furto di merce o il sabotaggio industriale. Anche il mancato rispetto di leggi e normative rientra in questa tipologia.

   I *rischi speculativi*, invece, possono assumere entrambi i segni e quindi far conseguire sia utili che perdite. Entrando nell'ambito finanziario un rischio speculativo è spesso

associato ad un investimento e alla composizione di un portafoglio. Altri esempi sono il mutamento dei gusti dei consumatori o la fluttuazione dei costi delle materie prime. Tradizionalmente a loro volta, i rischi puri e speculativi possono essere ancora divisi in ulteriori sottocategorie come individuati nella Figura 1.

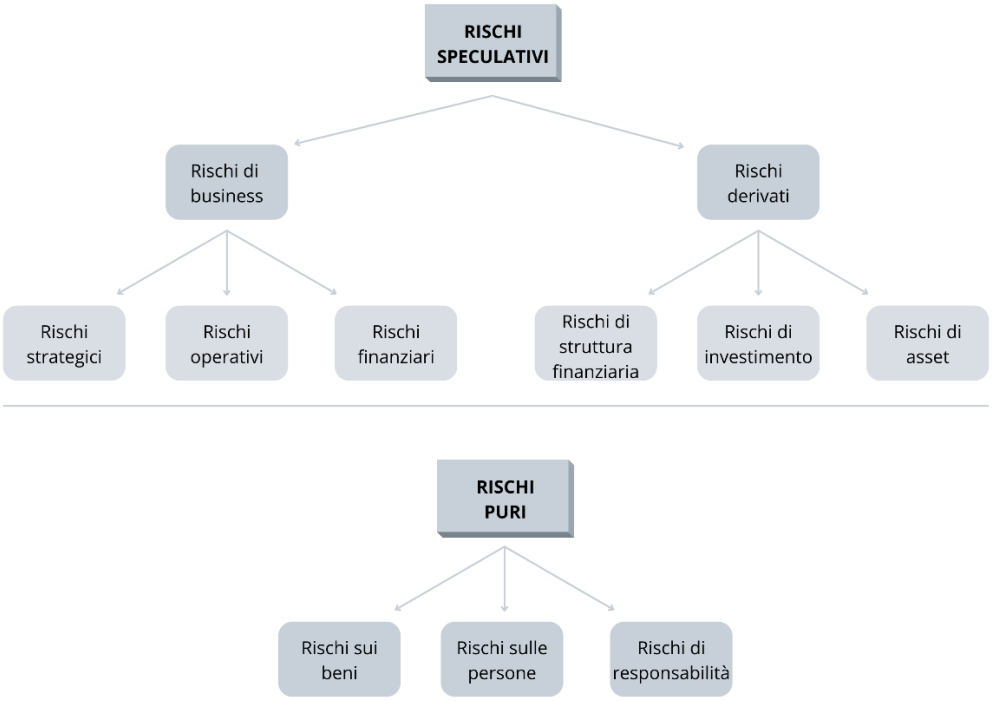

Figura 1 - Classificazione rischi speculativi e puri (Fonte: Floreani A., 2004)

3. Si possono, poi, distinguere *rischi sistematici* e *rischi specifici* in base al criterio della diversificazione.

   Un *rischio sistematico* è un rischio che non può essere eliminato applicando una diversificazione; per questo motivo questa classificazione è molto utilizzata in ambito finanziario e in particolare nella teoria del Capital Asset Pricing Model[1] (CAPM). L'impossibilità di diversificazione è generata dal fatto che questi eventi rischiosi sono correlati e dipendenti da variabili macroeconomiche, ovvero variabili economiche a livello aggregato, che quindi interessano tutte le aziende di uno stesso sistema economico senza distinzioni. Questo, però, non implica che tutte le aziende interessate

[1] Il Capital Asset Pricing Model è un modello molto conosciuto in ambito finanziario e si occupa dell'equilibrio dei mercati finanziari. La caratteristica principale del modello è la relazione lineare tra il rendimento atteso di un titolo e la sua rischiosità, ovvero la sensibilità del rendimento del titolo al rendimento di mercato.

siano toccate dall'evento con la stessa intensità. Un esempio attuale è la profonda recessione mondiale del 2020 legata alla pandemia di Covid-19; questo fenomeno ha comportato una perdita di PIL mondiale considerevole e ha interessato bene o male le aziende di tutto il mondo, prime tra tutte le aziende del settore turistico e dei trasporti (pensiamo alle compagnie aeree ad esempio) causando prevalentemente danni e perdite, ma alcuni settori come la grande distribuzione o l'industria dell'intrattenimento online hanno addirittura migliorato gli introiti e gli utili, presentando un andamento anticiclico.

Al contrario, i *rischi specifici* sono eliminabili o quantomeno controllabili tramite la diversificazione e sono peculiari di una specifica azienda. Questo tipo di rischio è gestibile combinando insieme più rischi specifici e sfruttando la proprietà di sub-additività[2] del rischio con variabili non correlate. È importante notare che, spesso, nella realtà un rischio è costituito da una componente sistematica e da una specifica (come, ad esempio, suggerito nella teoria CAPM).

4. Un'ulteriore tipologia di classificazione deriva dalla teoria dell'Enterprise Risk Management (ERM) che suddivide in quattro categorie gli obiettivi aziendali: strategici, obiettivi operativi, di reporting e di conformità. Di conseguenza si individuano quattro tipi di rischi omonimi.

   I *rischi strategici* impattano sugli obiettivi strategici di un'azienda, che sono di natura generale e definiti dal vertice della struttura organizzativa e spesso relativi al modello di business e al mercato di riferimento. Questi rischi si riferiscono alla possibilità che l'azienda non sia in grado di generare i flussi di cassa necessari alla continuità aziendale, compromettendone il valore generato. Alcuni esempi sono: i rischi legati allo sviluppo di un nuovo prodotto innovativo, i rischi legati alla scelta dei mercati su cui operare e i rischi legati alla scelta di delocalizzazione degli impianti produttivi.

   I *rischi operativi* si riferiscono a obiettivi di un livello inferiore che spesso riguardano la gestione ottimale delle risorse aziendali, siano esse procedure, risorse umane o

[2] La proprietà di sub-additività è caratteristica della funzione f tale che $f(a+b) \leq f(a) + f(b)$ per ogni valore di a e b interno al dominio della funzione.

macchinari. Un rischio di questo tipo si traduce nel peggioramento dell'efficienza, dell'efficacia e dell'economicità dei processi.

I *rischi di reporting* riguardano gli obiettivi sull'affidabilità e correttezza delle informazioni fornite dal cosiddetto reporting e comprendono la possibilità che le informazioni internamente o esternamente all'azienda non siano rilevanti, di qualità o tempestive.

I *rischi di conformità* o *compliance* tengono conto della non osservanza di leggi e regolamenti in vigore con la conseguenza di incorrere in condanne, in sanzioni e, in generale, in un danno economico e di immagine per l'azienda.

5. Nella pratica, inoltre, si utilizza anche una classificazione non formalizzata che tiene conto del tempo che intercorre tra il concretizzarsi di un evento e il momento in cui l'azienda risente dell'impatto sugli obiettivi prefissati. I rischi in questo caso si dividono in *rischi a breve*, *medio* e *lungo periodo* ai quali spesso si trova un'analogia con i rischi operativi, tattici e strategici, rispettivamente.

Il processo di classificazione dei rischi è di grande importanza per la gestione degli stessi anche in correlazione degli obiettivi aziendali interessati. Il primo passo per progettare e strutturare una risposta al rischio, argomento trattato nel prossimo capitolo, è infatti comprendere la natura e le caratteristiche del rischio e rapportarlo al contesto.

2 Gestione del rischio: Risk Management

In questo capitolo è presentata la disciplina della *gestione del rischio* o *Risk Management*. Inizialmente è fornita al lettore la definizione di Risk Management, in seguito è trattato il tema dei principi che caratterizzano il processo per la gestione del rischio all'interno di una organizzazione. Il paragrafo finale, parte principale del capitolo, illustra il processo di gestione del rischio vero e proprio, ne analizza e descrive ogni sua fase e aspetto e presenta le attività trasversali necessarie alla sua realizzazione.

2.1 La definizione di *Risk Management*

In qualsiasi organizzazione, di ogni tipologia, settore e dimensione, avviene, in maniera più o meno conscia, un processo che punta alla gestione del rischio, spesso menzionata nella letteratura col termine inglese *Risk Management*. Questo processo, che comprende svariate attività e interessa diverse risorse, ha come scopo il raggiungimento degli obiettivi prefissati. Essendo una disciplina molto diffusa è comune imbattersi nell'uso di una diversa terminologia in base al settore di business.

La Norma UNI 11230:2007, punto di riferimento nazionale per questa tematica, definisce uno standard per la gestione del rischio intesa come: "*L'insieme di attività, metodologie e risorse coordinate per guidare e tenere sotto controllo un'organizzazione con riferimento ai rischi*". Dall'analisi delle note correlate alla norma si evince che il processo sopra descritto include "*[…] la comunicazione del rischio, la contestualizzazione del rischio, la valutazione del rischio, il trattamento del rischio e il monitoraggio del rischio*" e che è finalizzato "*[…] alla prevenzione e protezione ottimizzando i benefici*".

L'*Institute of Risk Management* (IRM), invece, propone una definizione alternativa di gestione del rischio: "*Process which aims to help organizations understand, evaluate and take action on all their risks with a view to increasing the probability of success and reducing the likelihood of failure*".

2.2 I principi e la struttura di riferimento del Risk Management

La norma UNI ISO 31000:2018 fornisce una guida per l'implementazione di una gestione del rischio efficiente e proficua attraverso alcuni principi, presentati di seguito, che svolgono il ruolo di fondamenta del processo di risk management. Ricordando che l'obiettivo del risk

management è la protezione del valore aziendale, grazie ai principi che considerano contesti interni ed esterni l'organizzazione è più facilmente in grado di ridurre l'incertezza e quindi raggiungere al meglio gli obiettivi prefissati.

I principi suggeriscono che la gestione del rischio deve assumere queste caratteristiche:

- essere *Integrata*: coinvolgimento di tutte le parti interessate in tutte le attività;
- essere *Strutturata e globale*: l'approccio deve essere strutturato e globale;
- essere *Personalizzata*: appropriata al livello di rischio all'interno della organizzazione;
- essere *Inclusiva*: coinvolgimento delle parti interessate al fine di sfruttare le loro conoscenze, i loro punti di vista e le loro percezioni;
- essere *Dinamica*: è necessaria tempestività nell'intercettare gli eventi rischiosi al fine di una buona gestione del rischio;
- appoggiarsi sulle *Migliori informazioni disponibili*: il processo si basa sui dati, sia storici che attuali; è importante che la struttura disponga di dati chiari, tempestivi e affidabili;
- considerare i *Fattori umani e culturali*: è necessario tenere conto dell'influenza sulla gestione del rischio dei comportamenti, degli usi e costumi e della cultura delle parti interessate;
- operare in un'ottica di *Miglioramento continuo*: è importante apprendere dall'esperienza e migliorare il processo in modo continuativo.

Un altro elemento che emerge analizzando la norma è l'importanza della struttura di riferimento per la gestione del rischio che svolge una funzione di supporto alle attività. Riportando testualmente, "*Lo sviluppo della struttura di riferimento comprende l'integrazione, la progettazione, l'attuazione, la valutazione ed il miglioramento della gestione del rischio in tutta l'organizzazione*" (UNI ISO 31000:2018), ricordando sempre che ciò deve essere commisurato alle esigenze dell'organizzazione stessa. Come punto cardine della struttura sono individuate la leadership e l'impegno del management senza le quali è difficile implementare strategie integrate e gestire in modo allineato e adeguato il rischio.

2.3 Il processo

Il processo di risk management è composto da un sottoinsieme di altri processi che comprendono svariate attività e procedure. Anche in questo caso, diverse istituzioni presentano una propria rappresentazione del processo di gestione dei rischi; tra le varie si possono menzionare quelle elaborate dall'Institute of Risk Management (IRM), quelle proposte dagli standard BS (British Standards Institution) o quelle ideate dall'approccio COSO ERM Cube. Per mantenere coerenza con i capitoli precedenti è opportuno fornire una visione generale, applicabile a qualsiasi organizzazione e in qualsiasi ambito, e, per tanto, si è scelto di presentare e analizzare il processo presentato negli standard internazionali UNI ISO 31000 nella loro versione aggiornata del 2018 che fornisce un framework ben definito ed esaustivo.

Il processo, presentato in Figura 2, è suddiviso in tre fasi principali che sono: 1) Campo di applicazione, contesto e criteri, 2) Valutazione del rischio (spesso conosciuta come Risk Assessment) e 3) Trattamento del rischio. Queste fasi sono accompagnate e supportate da altre attività individuate come: 4) Comunicazione e consultazione, 5) Registrazione e reporting e 6) Monitoraggio e riesame.

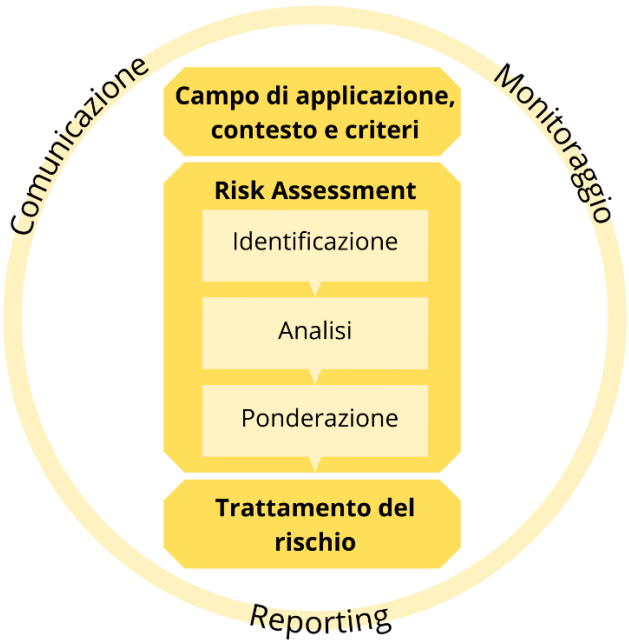

Figura 2 - Processo di Risk Management

Nei paragrafi che seguono ognuna di queste fasi verrà descritta e analizzata in modo tale da fornire al lettore una visione dettagliata del processo.

2.3.1 Campo di applicazione, contesto e criteri

La fase iniziale del processo è importante per la corretta gestione del rischio, questo perché ha il compito di stabilire il campo di applicazione, il contesto e i criteri con cui opera l'organizzazione e di conseguenza individuare la strategia che meglio si adatta agli obiettivi prestabiliti. Infatti, il tipo di organizzazione, l'ambiente che la circonda, il mercato di riferimento e tantissime altre variabili influenzano la propensione al rischio e la diversa percezione del rischio e quindi la strategia di gestione. È chiaro come in questa fase emerga l'elemento di personalizzazione e adattamento del processo che permette di caratterizzare la strategia. Pertanto, è fondamentale stabilire a quale livello dell'organizzazione si sta lavorando, l'approccio infatti è differente, ad esempio, se riguarda la gestione di rischi legati a obiettivi strategici piuttosto che a quelli operativi. L'analisi dell'ambiente avviene sia nel contesto interno all'organizzazione di riferimento sia esternamente e si riferisce allo "spazio" in cui gli obiettivi devono essere conseguiti. Stabilire i criteri di rischio significa identificare l'entità e la tipologia di rischio che l'organizzazione può o non può assumere. Nel definirli è opportuno tenere in considerazione gli obiettivi, i mezzi a disposizione e le risorse in generale e tener conto di tutti i punti di vista delle parti interessate nel processo. Sebbene questa fase risulti essere la prima del processo non è rara un'attività di riesame e revisione periodica dovuta ai cambiamenti di una delle variabili appena descritte.

2.3.2 Valutazione del rischio: risk assessment

Il risk assessment è la fase principale e più tecnica del processo di gestione dei rischi e si compone di tre attività: l'identificazione dei rischi, l'analisi dei rischi e la ponderazione del rischio.

2.3.2.1 Identificazione dei rischi

L'identificazione dei rischi ha l'obiettivo di individuare e descrivere gli eventi rischiosi e i fattori di incertezza che potrebbero verificarsi nel corso delle attività dell'organizzazione, causando un allontanamento dagli obiettivi target. La mancata o non corretta identificazione dei rischi può portare a gravi risvolti economici, finanziari e patrimoniali, andando non solo a inficiare la riuscita e la bontà di un progetto ma, in casi estremi, anche a influenzare la stabilità e la solidità dell'organizzazione stessa, compromettendone la continuità. Un corretto svolgimento di questa attività non si limita solo ad individuare le varie fonti di rischio ma punta anche

all'individuazione di quelle fonti di opportunità che ancora non sono state considerate o non sono sfruttate in tutto il loro potenziale dal management.

In questa fase è importante individuare tutti i possibili eventi, a prescindere che essi siano interni o esterni all'azienda e che siano o non siano direttamente controllabili internamente. Tra i fattori esterni più comuni si individuano quelli legati all'economia, all'ambiente, alla società e alla cultura, alla tecnologia e alla politica/normativa di riferimento. Alcuni esempi di fattori interni, invece, sono le infrastrutture, il personale, le risorse in generale e i processi.

Per svolgere correttamente questa fase è utile usare alcune tecniche di supporto che aiutino nell'identificazione delle fonti meno evidenti e comuni, tenendo conto che alcune tipologie di rischi sono più difficili da identificare rispetto ad altre, ad esempio i rischi puri rispetto a quelli speculativi. Una pratica comune è l'utilizzo di questionari e checklist ad hoc (*prompt list*) in base al settore; un altro approccio valido è l'uso di sessioni di workshop e brainstorming di gruppo nelle quali le parti interessate possono esprimere il proprio punto di visa e generare idee guidate da un moderatore, metodo che permette di far emergere rischi molto complessi e strutturati unendo il contributo di più soggetti. La fase di identificazione dei rischi, talvolta, è affidata ad aziende esterne, e, in questo caso, gli strumenti indispensabili risultano essere non solo la documentazione tecnica e contabile ma anche gli archivi storici in cui sono contenuti eventi passati relativi al processo di risk management. Questo in quanto l'analisi dell'esperienza pregressa può svolgere la funzione di punto di partenza dell'attività, tenendo comunque conto che tale approccio tende a far sovrastimare i rischi già avvenuti in passato e a sottostimare rischi che, seppur presenti, non si sono ancora verificati. Per ultimo, anche metodi visivi come flowchart o grafici dei rischi permettono di analizzare i processi e le operazioni all'interno dell'organizzazione individuandone i passaggi critici. Successivamente all'identificazione è necessario fornire una descrizione, standardizzata, delle fonti di rischio che permetta di raccogliere le caratteristiche dei rischi traducendole in informazioni per l'organizzazione.

2.3.2.2 Analisi dei rischi

La seconda fase del risk assessment è l'analisi del rischio, conosciuta anche come Risk Analysis, e ha lo scopo di stimare le probabilità e le conseguenze (l'impatto) dei rischi individuati precedentemente. Per fare questo è necessario introdurre nell'analisi incertezze, fonti di rischio, natura del rischio, fattori correlati a variabilità degli eventi e al tempo, probabilità di

accadimento degli eventi rischiosi, impatti attesi sull'organizzazione e contemplare diversi scenari, tenendo conto del sistema di controllo implementato e della relativa efficacia. Le tecniche impiegate in questa particolare fase si dividono in tecniche qualitative, quantitative e miste (semi-quantitative). L'uso di strumenti statistici e modelli matematico-statistici, grazie ai quali si estrae informazione dai dati disponibili determinando le distribuzioni di probabilità, caratterizza le tecniche quantitative, mentre nelle tecniche qualitative l'impatto e la probabilità di un rischio sono espresse tramite scale ordinali descrittive (ad esempio, la probabilità può essere alta, media e bassa mentre l'impatto: catastrofico, medio, trascurabile), che rendono il processo più soggettivo in quanto si basano solamente sull'esperienza e sulle competenze del soggetto incaricato della valutazione. Le tecniche semi-quantitative, invece, traducono i livelli descrittivi dell'approccio qualitativo in classi numeriche che però non rappresentano in senso stretto una quantificazione degli effetti economici o delle probabilità. Generalmente sono utilizzate tecniche qualitative e semi-quantitative, meno dispendiose e costose, nelle fasi iniziali dell'analisi dei rischi; mentre una volta determinate le principali tipologie di rischio si procede con un computo quantitativo più approfondito. In generale, comunque, la scelta della tecnica più opportuna si valuta attraverso un trade-off tra costi d'implementazione della tecnica e benefici in termini di qualità della conoscenza estratta.

2.3.2.3 Ponderazione del rischio

L'analisi del rischio fornisce l'input per l'ultima fase del risk assessment: la ponderazione del rischio. Questa fase ha l'obiettivo di essere di supporto nelle decisioni e implica la valutazione dei risultati dei precedenti step confrontandoli con i criteri e le soglie di rischio determinate in precedenza dall'organizzazione (vedi fase: Campo di applicazione, contesto e criteri). A seconda dell'output si possono individuare cinque diversi tipi di intervento in presenza di un rischio: la prima opzione è quella di non intervenire in alcun modo; in alternativa si può procedere al trattamento del rischio (fase che sarà approfondita nel prossimo paragrafo), una terza opzione prevede una analisi più dettagliata se quella effettuata non permette di comprendere un rischio particolarmente difficile da gestire. Si può decidere di mantenere i controlli attuali identificati nelle fasi iniziali del processo oppure, addirittura, ripensare e riformulare gli obiettivi ai vari livelli dell'organizzazione qualora si individuassero rischi impossibili da gestire.

2.3.3 Trattamento del rischio

In questa fase si stabilisce la risposta al rischio più adatta, utilizzando come input di partenza le informazioni estratte nella fase di *risk assessment.* Il criterio decisionale segue un trade off costi-benefici, ma non solo; infatti, è necessario considerare nel processo decisionale le obbligazioni e gli impegni delle parti interessate.

La norma UNI ISO 31000 presenta il trattamento del rischio come un processo iterativo che prevede una prima fase di scelta delle opzioni di trattamento (descritte in seguito), per poi passare alla pianificazione e all'attuazione del trattamento o dell'insieme di trattamenti adeguati, valutarne l'efficacia andando ad individuare il rischio residuo[3] e come ultimo step valutare se questo rischio residuo sia accettabile o necessiti di ulteriori trattamenti.

Le opzioni di intervento in letteratura sono chiamate "4Ts" e si possono riassumere in: *tolerate, treat, transfer* e *terminate*. La scelta delle azioni da intraprendere deve tener conto del livello di criticità del rischio su cui si sta lavorando e del fatto che possono essere intraprese più azioni riguardo ad uno stesso rischio. Strumenti di supporto nel processo decisionale sono le matrici di rischio che suggeriscono la risposta dominante in base alla posizione del rischio rispetto a probabilità e impatto. Come si evince dalla Figura 3:

- per rischi con basse probabilità e basso impatto la risposta principale è *accettare* il rischio (*tolerate*);
- per rischi con alte probabilità e basso impatto la risposta principale è *mitigare* il rischio (*treat*);
- per rischi con basse probabilità e alto impatto la risposta principale è *trasferire* il rischio (*transfer*);
- per rischi con alte probabilità e alto impatto la risposta principale è *evitare* il rischio (*terminate*);

[3] Il rischio residuo è il rischio che permane anche dopo il trattamento e l'applicazione del piano di prevenzione sul rischio iniziale.

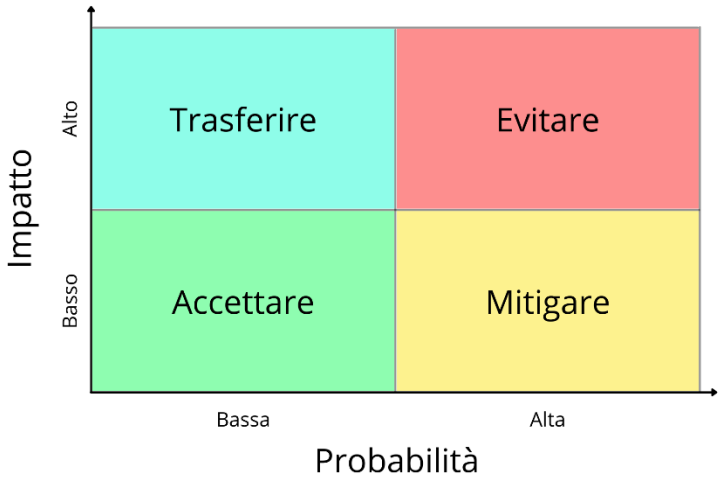

Figura 3 - Trattamento del rischio: le risposte in relazione a impatto e probabilità

Nei paragrafi successivi sono descritti i trattamenti sopra menzionati.

2.3.3.1 Accettare il rischio

L'esposizione al rischio, in questo caso, risulta essere tollerabile senza operare altre azioni. Si accettano anche quei rischi il cui trattamento comporta costi molto più alti dei benefici conseguenti all'intervento. In generale, l'adozione di questo piano d'azione implica che il rischio in esame non superi la soglia critica, decisa dall'organizzazione considerando costi-benefici e altre variabili menzionate in precedenza, e che quindi rientri nel livello di accettabilità. I rischi accettati devono comunque essere attentamente monitorati e controllati durante lo svolgimento delle attività ed è buona pratica che l'organizzazione progetti piani di recupero qualora si concretizzasse l'evento rischioso, anche attraverso l'allocazione di risorse utilizzabili per contenere gli effetti (un esempio è il cosiddetto *contingency budget* diffusa in ambito Project Management).

2.3.3.2 Mitigare il rischio

Gran parte dei rischi operativi individuati in una organizzazione è gestita con questa tipologia di trattamento. L'obiettivo di mitigare un rischio è quello di ridurre la probabilità di accadimento dell'evento rischioso e/o ridurre l'impatto qualora si concretizzasse, portandolo ad un nuovo livello accettabile attraverso azioni volte a intervenire su cause e/o effetti del rischio. Gli interventi di mitigazione sono vari e dipendono dal rischio in esame; infatti, le azioni intraprese sono strettamente legate al rischio e sono parte di un piano d'azione ad hoc. Un

esempio di mitigazione è l'aumento di controlli e di ispezioni sulle materie prime provenienti da un fornitore critico.

2.3.3.3 Trasferire il rischio

I rischi con bassa probabilità di accadimento ma con un grande impatto potenziale solitamente sono trattati con un trasferimento del rischio stesso a terze parti. Gli strumenti utilizzati sono le assicurazioni, in cui avviene un trasferimento delle conseguenze economiche in cambio di un premio assicurativo, e l'uso di contratti particolari in cui si trasferisce il rischio a committenze esterne. A differenza della mitigazione non si intraprendono azioni volte alla riduzione della probabilità del rischio e/o dell'impatto, ma ci si limita a riversare gli effetti economici su altri soggetti al di fuori dell'organizzazione. Nella realtà aziendale, e non solo, spesso accade che alcuni rischi per legge debbano essere trasferiti, un esempio è l'obbligatorietà della assicurazione contro gli infortuni dei dipendenti o l'assicurazione RCA per le autovetture. Questa tipologia di intervento è consigliata per rischi di tipo finanziario o rischi legati agli asset dell'organizzazione.

2.3.3.4 Evitare il rischio

L'ultimo piano d'azione interessa quei rischi che non possono essere mitigati o controllati a livelli accettabili. Solitamente questi rischi hanno la caratteristica di avere un'alta probabilità di accadimento e un impatto molto alto. In questi casi è appropriato eliminare l'incertezza alla base, evitando il rischio. Questa tipologia di trattamento risulta essere estrema perché in alcuni casi può addirittura portare allo stop di processi o attività, in altri casi comporta la modifica degli obiettivi dell'organizzazione. In alcuni casi, rischi con queste caratteristiche interessano attività fondamentali per lo svolgimento del business dell'organizzazione e pertanto non è possibile eliminarli interamente; risulta quindi necessario implementare misure di controllo alternative. Un esempio può essere la rinuncia all'entrata in un nuovo mercato o la cessione di una determinata business unit.

2.3.4 Comunicazione, monitoraggio e reporting

Le fasi principali del processo di risk management, descritte nei paragrafi precedenti, sono supportate e accompagnate dalle attività di comunicazione e consultazione, di monitoraggio e riesame, e di registrazione e reporting. Queste attività svolgono un ruolo fondamentale per la corretta gestione del processo e devono essere svolte durante tutte le fasi di gestione del rischio.

La comunicazione e la consultazione aiutano le parti interessate a comprendere i rischi su cui si sta lavorando e ad avvalorare le decisioni e le strategie intraprese e sono caratterizzate dallo scambio di informazioni tra le parti. La consultazione è fondamentale nel processo decisionale perché permette di attingere in maniera tempestiva a informazioni complete ed integre. Queste attività permettono, inoltre, il lavoro coordinato di diverse parti interessate con differenti competenze e garantisce criteri di rischio che rispecchino tutti i punti di vista.

Il monitoraggio continuo e il riesame hanno lo scopo di sorvegliare e misurare i parametri e i risultati della gestione del rischio e permettono miglioramenti continui nella qualità ed efficacia dei piani d'azione progettati. Sono parte delle attività che svolgono un ruolo importante nel PDCA Cycle[4], garantendo l'individuazione tempestiva di problematiche e dei relativi interventi di risposta. La fase di controllo si concretizza a diversi livelli dell'organizzazione, spaziando dai controlli dell'alto management fino a quelli più operativi. Il monitoraggio oltre a valutare efficacia ed efficienza dei piani d'azione deve tener conto anche dell'economicità degli stessi, intervenendo qualora il rapporto costi-benefici cambi durante lo svolgimento delle attività.

I risultati delle varie fasi del processo di risk management e i suoi output finali devono essere documentati con strumenti e metodologie appropriate. Le attività di registrazione e reporting hanno proprio lo scopo di immagazzinare informazioni in modo tale che siano facilmente consultabili a supporto del processo decisionale e per favorire il miglioramento del processo. Inoltre, aiutano anche l'integrazione e il dialogo tra le varie parti interessate e svolgono la funzione di supporto per gli organismi di supervisione del processo. La metodologia e i meccanismi utilizzati nel reporting devono necessariamente considerare il tipo di informazioni contenute e la loro delicatezza, tenendo conto che le informazioni immagazzinate derivano da numerose fonti diverse che possono essere interne o esterne all'organizzazione.

[4] Il PDCA Cycle, conosciuto anche come ciclo di Deming e acronimo di *Plan-Do-Check-Act,* è un metodo di gestione dei problemi che si articola in quattro step. Nato come sistema applicato al controllo qualità e alle attività manifatturiere ora è esteso a moltissimi campi, tra i quali quello di strategia aziendale e management. L'obiettivo di questa tecnica è il controllo e il miglioramento continuo spinto da una gestione iterativa e basata su azioni correttive.

3 Artificial intelligence e crisi d'impresa

Il presente lavoro di tesi è focalizzato sullo studio di modelli predittivi, basati sull'utilizzo di algoritmi di *Artificial Intelligence* (di seguito *AI*), che possano individuare e intercettare situazioni di crisi aziendale in maniera tempestiva fornendo al management uno strumento di supporto decisionale. In questo capitolo è dapprima definito il concetto di crisi aziendale, sono analizzate le disposizioni normative in materia ed il rischio di fallimento. In seguito, è analizzato il valore aggiunto fornito dalle nuove tecniche di AI, in grado di proporre strumenti sempre più validi a supporto di processi decisionali anche molto complessi. Infine, è presentata una breve review dei principali e più significativi modelli predittivi in ambito fallimentare presenti in letteratura, divisi a loro volta in metodi statistici e metodi machine learning, evidenziandone le maggiori differenze.

3.1 Crisi d'impresa e fallimento

In letteratura esistono diversi modi di intendere il fallimento di un business, modi che spesso portano a creare ambiguità rispetto alla sua definizione e ai criteri utilizzati per identificarlo. Risulta fondamentale fornire al lettore i diversi punti di vista scelti dai vari autori, definirne e analizzarne le caratteristiche ed individuare quelli maggiormente utilizzati.

Balcaen e Ooghe (2006) forniscono una prima analisi dei criteri utilizzati nella definizione di fallimento. In particolare, si evidenzia come questi criteri siano scelti in maniera arbitraria in base agli studi.

Alcuni autori prediligono l'uso della definizione di fallimento fornita dalla giurisprudenza, in inglese definita come *bankruptcy,* distinguendo le aziende fallite dalle non fallite riferendosi alla presenza di una sentenza dichiarativa di fallimento da parte delle autorità preposte o comunque in presenza di un procedimento fallimentare. Questo criterio è spesso utilizzato in quanto permette di dividere facilmente le aziende dello studio in due popolazioni, semplificando i campioni utilizzati nei modelli di predizione e inoltre consente di trattare il fallimento come un evento oggettivo con una data precisa.

Altri autori usano definizioni finanziarie, scegliendo criteri come EBIT o EBITDA negativo, utili e perdite, rapporto di copertura di interessi passivi e altri indici finanziari. (Vedi Platt e Platt, 2002; Platt e Platt, 2004). In altre occasioni invece sono utilizzati come criteri eventi correlati

al fallimento come, ad esempio, lo stato di insolvenza, la ristrutturazione dei capitali o la cessione o chiusura forzata di parti importanti dell'azienda.

Un aiuto nel definire meglio la terminologia relativa al fallimento deriva da Altman e Hotchikiss (2006) che nel loro libro "*Corporate financial distress and bankruptcy*" individuano e analizzano i quattro principali termini inglesi utilizzati in letteratura per individuare il fallimento che sono: *failure*, *insolvency*, *default* e *bankruptcy*. In particolare, con *failure* identificano il concretizzarsi di tassi di ritorno sui capitali investiti (ROI) molto al di sotto degli standard di settore e in modo continuativo se confrontati con aziende equivalenti per investimenti e rischi. L'*insolvency*, tradotta letteralmente con il termine insolvenza, rappresenta l'impossibilità dell'azienda di soddisfare gli obblighi in maniera transitoria, traducibile anche con mancanza di liquidità: avviene quando le passività superano le attività. Lo stato di *default* è uno step superiore all'insolvenza temporanea; infatti, indica l'inadempienza di obbligazioni che può comportare azioni legali da parte del creditore. Infine, è fornita una duplice definizione di *bankruptcy*: la prima fa riferimento al patrimonio netto dell'azienda mentre la seconda utilizza come criterio la dichiarazione formale di un tribunale in merito al fallimento.

Riassumendo, si possono individuare tre tipologie principali di fallimento in base al criterio scelto nei vari studi presenti in letteratura: *fallimento legale*, *fallimento tecnico* e *fallimento legato a misure di bilancio*.

Facendo riferimento, invece, alla situazione normativa italiana è utile menzionare il nuovo "Codice della crisi d'impresa e dell'insolvenza" (CCII) (decreto legislativo 12 gennaio 2019, n.14) che ha rivoluzionato le normative in materia di crisi d'impresa e procedure fallimentari. È doveroso indicare che l'entrata in vigore della quasi totalità degli articoli è stata rimandata dapprima a settembre 2021 e poi, con una nuova modifica da parte del Governo, ulteriormente posticipata a maggio 2022, fatta eccezione per gli articoli riguardanti gli strumenti di allerta che entreranno in vigore solo dal 2024. Uno dei punti cardine di questa normativa, che avvalora e giustifica lo studio proposto in questo lavoro di tesi, ha lo scopo di far emergere l'importanza dell'esistenza di procedure d'allerta e di composizione della crisi. La ratio dietro questa scelta normativa è quella di aumentare la consapevolezza riguardo l'importanza dell'intercettare in maniera tempestiva lo stato di crisi aziendale attraverso procedure, strumenti di allerta e nuovi obblighi. Risulta rilevante analizzare il concetto di crisi,

che è definito come "s*tato di difficoltà economico-finanziaria che rende probabile l'insolvenza del debitore*" e che si può manifestare come "*inadeguatezza dei flussi di cassa prospettici a far fronte regolarmente alle obbligazioni pianificate*" dall'impresa. A sua volta è importante comprendere cosa si intende per insolvenza, che è definita come "*lo stato del debitore che si manifesta con inadempimenti od altri fatti esteriori, i quali dimostrino che il debitore non è più in grado di soddisfare regolarmente le proprie obbligazioni*".

Analizzando gli articoli emerge il dovere da parte dell'imprenditore individuale di "*adottare misure idonee a rilevare tempestivamente lo stato di crisi e assumere senza indugio le iniziative necessarie a farvi fronte*" e dell'imprenditore collettivo che "*deve adottare un assetto organizzativo adeguato ai sensi dell'articolo 2086 del Codice civile, ai fini della tempestiva rilevazione dello stato di crisi e dell'assunzione di idonee iniziative.*"(Art. 3). È chiaro come queste normative spingano l'imprenditore ad un monitoraggio dell'andamento dell'azienda e delle sue performance (anche attraverso processi di risk management) al fine di individuare tempestivamente i sintomi di una crisi aziendale e la conseguente attivazione per tempo di misure volte a controllare la crisi, evitare situazioni di insolvenza di grave entità, e dunque andare verso una gestione preventiva di tali situazioni, con il fine di salvaguardare il valore dell'azienda, i posti di lavoro e evitare procedure fallimentari.

Gli articoli 12, 13, 14 e 15 aiutano a definire meglio le procedure di allerta e in particolare gli strumenti di allerta. In particolare, l'art. 13 recita: "*Costituiscono indicatori di crisi gli squilibri di carattere reddituale, patrimoniale o finanziario, rapportati alle specifiche caratteristiche dell'impresa e dell'attività imprenditoriale svolta dal debitore […] rilevabili attraverso appositi indici che diano evidenza della sostenibilità dei debiti per almeno i sei mesi successivi […] sono indici significativi quelli che misurano la sostenibilità degli oneri dell'indebitamento con i flussi di cassa che l'impresa è in grado di generare e l'adeguatezza dei mezzi propri rispetto a quelli di terzi. Costituiscono altresì indicatori di crisi ritardi nei pagamenti reiterati e significativi…*". Gli indici di cui si fa menzione sono elaborati per tipologia di attività economica e a cadenza trimestrale e fungono da indicatori di allerta della crisi. Nel caso in cui essi superino determinate soglie si concretizza la procedura di allerta che comprende diversi obblighi sia da parte dell'Agenzia delle Entrate sia dell'azienda in crisi al fine di contenerne l'impatto e risolvere la crisi.

Il “Codice della Crisi e dell’Insolvenza” non è un testo definitivo perché sarà integrato e aggiornato con le novità introdotte da nuovi interventi del legislatore e modifiche richieste dagli organi ministeriali preposti.

Risulta evidente la volontà del legislatore di evitare situazioni di crisi irreversibili utilizzando sistemi di monitoraggio tempestivi. Lo studio e la realizzazione di modelli predittivi in ambito di fallimento, sfruttando anche sistemi di algoritmi di AI, è un chiaro passo verso questa visione di gestione dei vari rischi di default, insolvency e bankruptcy, fornendo strumenti di supporto alle decisioni tempestive e correttive, proprio come richiesto dalla normativa.

Inoltre, risulta chiaro che la corretta valutazione del rischio di fallimento è di fondamentale importanza per il mantenimento degli equilibri dell’economia, permettendo agli istituti di credito e alle istituzioni di investire in aziende “sane” e con un futuro; e, come già descritto, fornire al management aziendale indicatori, che se individuati per tempo, possano aiutare ad evitare crisi aziendali e quindi andare ad aumentare il numero di quelle stesse aziende “sane”.

3.2 Valore aggiunto dell’AI: nuovi strumenti per la predizione

3.2.1 Definizione dell’AI

Per illustrare il valore aggiunto e le potenzialità dell’Artificial Intelligence (AI) è necessario capire in primis cosa si intende per AI.

L’AI, nella sua definizione più generale, è la capacità di una macchina/sistema informatico di svolgere attività e risolvere problemi attraverso la simulazione dei processi cognitivi caratteristici della mente umana tramite algoritmi logico-matematici.

Molto spesso si usano come sinonimi Intelligenza Artificiale e Machine Learning (ML) in modo errato; infatti, questi termini fanno riferimento a due approcci diversi seppur strettamente relazionati. Concretamente si può affermare che il Machine Learning, ovvero l’apprendimento automatico, sia uno dei sottoinsiemi dell’AI.

Il Machine Learning è, quindi, uno dei metodi in cui si attua un processo di intelligenza artificiale e fa riferimento a quei modelli matematici, applicati ad una macchina, in grado di ricevere dati in input e apprendere in modo autonomo da essi, progredendo di pari passo con l’esperienza. Un algoritmo che usa il Machine Learning è in grado di imparare e auto-

modificarsi, permettendo azioni per cui non è stato programmato direttamente. Attualmente esistono due principali approcci di apprendimento di algoritmi di Machine Learning:

- Machine Learning SUPERVISIONATO: in questo caso l'apprendimento avviene tramite un data scientist che interviene sui dati etichettandone almeno una parte. La macchina estrae relazioni tra input e output per poi utilizzarle per etichettare autonomamente nuovi input. Esempi di questo tipo di ML sono gli algoritmi di classificazione, le regressioni e SVM;
- Machine Learning NON SUPERVISIONATO: l'apprendimento avviene in modo indipendente dall'intervento umano attraverso dati non etichettati, la macchina è in grado di apprendere dagli errori. Esempi diffusi sono gli algoritmi di clustering e le regole di associazione.

L'AI, a differenza di cosa si possa pensare, è una disciplina che fonda le sue radici già negli anni '50 quando Alan Turing propose il suo celebre test[5] in cui si cerca di determinare se sia possibile che una macchina possa avere un comportamento intelligente, assimilabile a quello umano (Turing, 1950). Nei vari decenni successivi si concretizzano le prime teorie sulle reti neurali, sugli algoritmi di AI debole e forte, fino agli anni '80 in cui si sviluppano le prime applicazioni industriali. Ad oggi il ML e il Deep Learning[6], una sottocategoria di algoritmi ML, sono al centro dell'attenzione della comunità scientifica e soprattutto delle aziende che beneficiano di questi approcci; l'AI ha un ruolo rilevante in tutti gli ambiti della nostra vita e questo crescerà sempre di più in futuro, anche grazie alla integrazione con nuove tecnologie come l'Internet of Things[7] (IoT).

In Figura 4 è illustrata una grafica che aiuta ad individuare i differenti tipi di AI.

[5] Il test di Turnig, in breve, è superato quando una persona che manda domande attraverso un terminale a due altri terminali, uno gestito da un operatore umano e l'altro gestito da una macchina, non riesce a distinguere se la risposta di ritorno sia da parte dell'operatore umano o della macchina.
[6] Il Deep Learning, o anche apprendimento profondo, comprende una classe di algoritmi costituiti da diversi strati di reti neurali artificiali e pertanto è parte della più ampia disciplina Machine Learning.
[7] L'Internet of Things, conosciuta come IoT, è l'internet delle cose o degli oggetti, ovvero l'interconnessine di oggetti "intelligenti" attraverso la rete internet grazie alla quale ogni oggetto potenzialmente può acquisire una identità digitale e raccogliere e trasmettere dati online.

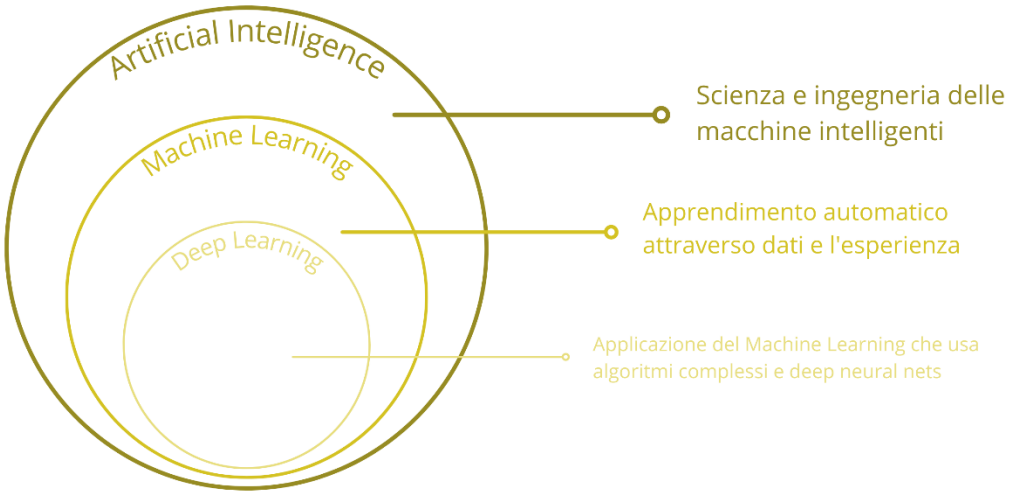

Figura 4 - Artificial Intelligence, Machine Learning e Deep Learning: le relazioni

3.2.2 Rivoluzione AI

Uno degli obiettivi dell'applicazione dell'AI in ambito aziendale, oggetto di questo lavoro di tesi, è la creazione di un DSS (Decision Support System). Un DSS è un sistema informativo aziendale usato come supporto alle decisioni. Dalla loro nascita negli anni '70 ad oggi l'approccio e le caratteristiche del concetto di DSS si è evoluto traducendosi in cambiamenti pratici dello strumento, per ultimo è stato toccato l'aspetto "intelligente" del sistema che lo riconduce in modo diretto alle tecnologie AI. Il ruolo di un DSS è analizzare e sintetizzare in informazioni grandi quantità di dati. Le informazioni così ricavate aiutano il decision-maker fornendo supporto nelle decisioni che diventano in questo modo data-driven ovvero guidate dai dati.

Questo aiuta le aziende a raggiungere anche uno degli obiettivi strategici aziendali più desiderato che è la diffusione di una cultura aziendale fondata sui dati, la cosiddetta Data-driven Culture. Una organizzazione di questo genere fonda sui dati la maggior parte delle decisioni e utilizza un approccio ai problemi guidato dai numeri e non da interpretazioni soggettive. Un'azienda con una cultura data-driven si pone come principio fondante l'estrazione di valore e di nuova conoscenza dai dati.

La correlazione tra AI e dati è evidente e merita un approfondimento. La rivoluzione AI nelle aziende ha come prerequisito la disponibilità di grandi quantità di dati per il processo di business intelligence. Non per caso, infatti, l'esplosione delle tecniche di AI si è sviluppata di

pari passo con il diffondersi dei cosiddetti Big Data e delle relative tecnologie di raccolta e gestione. Infatti, quando si parla di Big Data si fa implicitamente riferimento anche alle tecniche rivoluzionarie in ambito storage e di manipolazione dei dati, che insieme alla sempre maggiore potenza computazionale delle macchine, a nuovi approcci di sistema come il Cloud Computing e la disponibilità di software/servizi industriali che facilitano lo sviluppo e l'implementazione di modelli AI, hanno permesso un abbattimento dei costi che ha portato a una crescita importante e repentina della diffusione di questo approccio di supporto "intelligente" alle decisioni nelle aziende.

Si può affermare che le tecniche AI, e in particolare il Machine Learning, permettono ai dati di acquisire un nuovo valore che porta a nuovi modelli di business e quindi valore aggiunto per l'azienda, innescato dalla "Data Acceleration", attraverso una trasformazione dei suoi processi.

Tra gli svariati vantaggi portati dall'applicazione dell'AI processi aziendali spicca la possibilità di gestire in modo più semplice e migliore l'elaborazione di dati strutturati e, soprattutto, non strutturati (secondo KPMG circa l'80% di dati presenti in azienda è di questo ultimo tipo e, senza tecniche AI, è difficilmente sfruttato per via di costi e tempi di gestione elevati); inoltre esiste la possibilità di implementare simulazioni di vari scenari attraverso dati storici, il che rende possibile l'individuazione di criticità e falle di sistema garantendo l'intervento preventivo al fine di evitare gli scenari peggiori; a questo è legata la possibilità di svolgere analisi predittive riguardo problemi o crisi aziendali fornendo quindi uno strumenti di monitoraggio, volendo real-time, in grado di intercettare le problematiche per tempo, usando il passato per prevedere il futuro. La potenza dei sistemi AI è insita anche nella possibilità di rendere questi processi automatizzati, andando a ridurre costi, tempi e sforzo del personale, seppur ricordando che rimane di fondamentale importanza il giusto bilanciamento tra risorse umane e sistemi AI.

La facilità di gestione di questi sistemi ha portato allo sviluppo di modelli predittivi sempre più complessi, sia in riferimento agli algoritmi utilizzati sia per quanto riguarda la quantità e la varietà di dati utilizzati, e con performance migliorate rispetto ai sistemi classici.

L'AI diventa, in questo modo, un driver della strategia aziendale, andando ad impattare a livello organizzativo su tutta l'azienda e favorendo un approccio proattivo nei confronti del

business, anticipandone gli eventi attraverso strumenti di previsione. Già ad oggi l'integrazione di sistemi AI all'interno di un'organizzazione è motivo di vantaggio competitivo e, secondo ricerche di Accenture e Frontier Economics, nel 2035 potrebbe diventare il "motore della crescita", riuscendo a raddoppiare la crescita economica di aziende e paesi, grazie ad un aumento di produttività che in alcuni casi arriverebbe fino al +40%.

3.3 Rassegna dei principali metodi di predizione del fallimento

La predizione del fallimento è un argomento molto significativo nell'ambito della ricerca scientifica e nel mondo della finanza e management aziendale. Lo studio di modelli di previsione del fallimento e della crisi economica, infatti, è diffuso ormai da più di un secolo; l'esigenza di predire e discernere le attività redditizie e i business solidi da realtà non performanti e con il rischio di crisi aziendali e di default ha interessato moltissimi studiosi che in base all'avanzamento tecnologico e alle tecniche disponibili hanno sviluppato diversi modelli e teorie che impiegano diverse assunzioni e metodologie nell'approcciare il problema. Tendenzialmente questi metodi possono essere suddivisi in due grandi categorie: i metodi tradizionali, che utilizzano modelli statistici, e i metodi intelligenti, ovvero costruiti con l'utilizzo dell'AI (machine learning).

Recentemente la predizione del fallimento è ritornata in auge negli ambienti di ricerca e questo in grande parte è dovuto al tracollo finanziario del 2008 che ha portato ad una recessione globale innescata dal fallimento, il più grande fino ad oggi, di una delle banche d'affari più importanti d'America: la Lehman Brothers Holdings Inc. Questo evento, con effetto domino, ha interessato moltissime aziende che hanno dovuto gestire anni di depressione mondiale e le conseguenti situazioni di crisi aziendale. Inoltre, è stato attuato anche un irrigidimento della regolamentazione bancaria e delle normative in ambito aziendale da parte dei policy maker (l'esempio italiano è stato descritto nel paragrafo 3.1). Secondo studi ex post, tra i quali Hauser e Booth (2011), la crisi di questa grande banca poteva essere intercettata alcuni anni prima se si fossero osservati i numerosi segnali premonitori della crisi. In questo caso, l'utilizzo di modelli predittivi avrebbe potuto evitare il concretizzarsi della fase più acuta della crisi, fornendo strumenti di supporto alle decisioni basati sui dati.

Un altro evento che ha messo a dura prova le aziende di tutto il mondo e ha favorito situazioni di crisi aziendale è la pandemia Covid-19 diffusasi nei primi mesi del 2020 andando a colpire

duramente anche aziende e settori industriali che godevano di un ottimo stato di salute. In questo caso, modelli predittivi di fallimento possono aiutare il Governo nella giusta erogazione di fondi e ristori, indirizzando i trasferimenti ai settori più colpiti e meno resilienti che senza un intervento pubblico sono destinati a fallire con il conseguente problema della perdita di posti di lavoro.

Ragionando in ordine cronologico, quindi partendo dai modelli di tipo statistico, e differenziando per tipologie di modelli, è qui presentata una review dei principali studi in materia di previsione di fallimento.

I primi studi disponibili in letteratura risalgono agli inizi degli anni '30 quando Patrick (1932) presentò un'analisi multi-variabile su un cluster di 20 aziende.

L'attenzione via via crescente da parte dei ricercatori in materia ha portato ad un primo picco di studi verso la fine degli anni '60.

Tra i principali troviamo Beaver (1966) che utilizzò per primo alcuni indici finanziari per predire il fallimento attraverso un t-test[8]. In particolare, lo studio è sviluppato attraverso un modello che considera una sola variabile (modello Univariate Discriminant Analysis) molto semplice da implementare, che non richiede una particolare conoscenza della statistica, ma che si fonda su assunzioni molto forti che ne limitano l'utilizzo reale.

In contemporanea a Beaver, Tamari (1966) propose un modello di previsione basato su un indice di rischio: il Risk Index Model. Questo modello valuta la situazione finanziaria dell'azienda, attraverso vari indici finanziari che sono pesati in base all'importanza, su una scala da 0 a 100, con punteggio migliore 100. Questa tipologia di modello è stata utilizzata anche nello studio Moses e Liao (1987).

Uno dei modelli più conosciuti e citati in letteratura è stato sviluppato da Altman (1968) e fa parte di un gruppo di modelli basati sull'analisi di più variabili, la Multiple Discriminant Analysis (MDA).

Questi metodi sono caratterizzati da una funzione del tipo $Z = v_1X_1 + v_2X_2 + \cdots + v_nX_n$ dove: X_i corrispondono alle variabili indipendenti e v_i i coefficienti discriminanti, con *i* che va

[8] Il t-test, conosciuto anche come test t di Student, è un test d'ipotesi statistico che ha lo scopo di verificare se il valore medio di una distribuzione si discosta significativamente da un valore di riferimento.

da 1 a n, dove n è uguale alla numerosità delle variabili indipendenti. In particolare, Altman propose un indice conosciuto come Z-score, dipendente da cinque variabili indipendenti che secondo Altman rappresentano al meglio la situazione finanziaria dell'azienda:

X1 = Capitale circolante / Totale attivo

X2 = Utili portati a nuovo / Totale attivo

X3 = EBIT / Totale attivo

X4 = Valore di mercato Equity / Valore di bilancio del Debito

X5 = Fatturato / Totale attivo

Numerosi lavori utilizzano questo modello come punto di partenza con l'intento di migliorarlo, vedi Altman (1968), Altman et al. (1977) e Altman e Narayanan (1997), ed è spesso utilizzato come baseline nella comparazione della bontà di nuovi modelli di predizione del fallimento. Anche questa tipologia di modello si fonda su forti assunzioni, ad esempio sulla tipologia di dati utilizzati, che ne limita l'utilizzo.

Per questo motivo negli anni '80 si è sviluppata un'altra corrente di studio che utilizzava altri metodi statistici, definiti modelli di probabilità condizionata, come la regressione (logit, probit e modelli lineari di probabilità). Il modello di Ohlson (1980), O-Score, è uno dei primi che sfrutta nella previsione del fallimento aziendale la Logit Analysis, ovvero la regressione logistica, che è un modello di regressione non lineare, che si usa per definire la probabilità di una variabile dipendente di tipo binario. Le variabili indipendenti che caratterizzano il modello sono nove. Tra i tanti studi che sfruttano il modello Logit si menziona anche Martin (1977), che attraverso un modello di regressione logistica prevede il fallimento in ambito bancario.

Un altro studio significativo è pubblicato da Zmijewski (1984) che adopera la Probit Analysis, simile al modello Logit, dal quale differisce per la forma funzionale scelta nel calcolo della probabilità. Il modello di Zmijewski è anche conosciuto come X-Score e utilizza tre variabili per definire la probabilità di fallimento.

Nella letteratura il metodo Logit è molto più diffuso del gemello Probit e questo è dovuto al fatto che la complessità computazionale del secondo risulta essere più elevata (Balcaen e Ooghe, 2006). Come il modello Z-Score di Altman, anche i modelli di Ohlson e Zmmijewski sono punti di partenza per molti studi con lo scopo di migliorarne le performance (Hillegeist et al., 2004; Chen et al., 2010).

Dall'analisi della letteratura appena svolta si possono individuare quattro principali tecniche statistiche impiegate nei modelli di predizione del fallimento: *analisi monovariabile*, *Risk Index Model*, *Multiple discriminant analysis (MDA)* e i *modelli di probabilità condizionata* (come Probit e Logit). Tutti questi modelli statistici hanno sicuramente delle caratteristiche positive, ad esempio la determinazione di una probabilità di fallimento e sono in grado di valutare l'impatto di ogni variabile indipendente singolarmente (Perboli e Arabnezhard, 2020), ma sono soggetti anche a svariate problematiche e criticità. Il cosiddetto "paradigma classico" non tiene conto, per caratteristica dei modelli, di molti aspetti della previsione del fallimento ed è caratterizzato da una certa arbitrarietà del ricercatore, sia su cosa intende per fallimento (vedi paragrafo precedente) sia nella scelta dei criteri del modello. Inoltre, spesso, questa tipologia di modelli è affetta da over-modeling, ovvero l'ottimizzazione esasperata del modello rispetto ad un ristretto campione specifico e spesso non-random, cosa che rende il modello non scalabile e instabile. Un altro limite dei modelli statistici classici è l'assunzione, forte, che le relazioni tra le variabili indipendenti e quella dipendente siano stazionarie e stabili nel tempo. Anche l'orizzonte temporale di predizione in questi modelli è limitato, solitamente a 12 mesi, per mantenere output affidabili.

L'esigenza di superare tali problematiche, la crescente evoluzione nella raccolta di grandi serie di dati e l'avanzamento della tecnologia ha spinto i ricercatori ad affrontare il problema della predizione del fallimento tramite algoritmi di AI, in particolar modo si è sviluppato l'utilizzo di tecniche di machine learning, che hanno permesso un miglioramento generale delle performance dei modelli. Di seguito sono presentate le tecniche di AI più diffuse in letteratura ed i relativi studi associati.

Tra i primi algoritmi intelligenti applicati in questo campo troviamo le *reti neurali*, *Neural Networks (NN)*, che cercano di emulare il comportamento del cervello umano a cui si ispirano, imitando il modo in cui i neuroni si scambiano le informazioni. La struttura della rete è composta da vari nodi che formano strati collegati tra loro, in particolare esistono tre tipologie di strati: lo strato di input, lo strato nascosto, core della struttura, e infine, lo strato di uscita. I nodi, i neuroni, ricevono in input dei dati e delle informazioni e, dopo averle elaborate ad esempio assegnando loro un peso, le trasmettono modificate ai nodi dello strato successivo.

Gli studi che sfruttano le NN applicati alla previsione del fallimento aziendale, che hanno iniziato ad emergere a fine anni '80 e sono stati protagonisti nel decennio successivo, da subito

hanno ottenuto buoni risultati indicando buone capacità del modello in questo contesto. Il primo studio, presentato dai ricercatori Odom e Sharada (1990), fece da apripista a numerosi altri studi come: Tam e Kiang (1992), Udo (1993), Coats e Fant (1993), Altman et al. (1994), Wilson e Sharada (1994), Boritz e Kennedy (1995), Lee et al. (1996), Alam et al. (2000), Lee et al. (2005). Anche recentemente numerosi ricercatori hanno utilizzato reti neurali, ottimizzando e migliorando i modelli sviluppati nel passato, ad esempio Tsai e Wu (2008) e Zhao et al. (2015).

Altri algoritmi di AI utilizzati sono i *classificatori* come *Decision Tree*, *Naive Bayes*, *K-Nearest Neighbor (KNN)*, *Genetic algorithm* e alcuni studi utilizzano i modelli *SVM (Support Vector Machine)*, ad esempio nello studio di Shin, Lee e Kim (2005).

L'attenzione dei ricercatori, però, si è rivolta maggiormente verso i cosiddetti *modelli ensamble* (strong learner) che usano tecniche di apprendimento d'insieme (ensamble learning). I modelli ensamble sono costituiti da combinazioni di modelli base (weak learners o base models), che ne determinano il tipo, e sono conosciuti anche con il nome di multi-classifier method. La bontà di questi modelli per la predizione del fallimento aziendale è evidenziata negli studi di numerosi ricercatori, ad esempio Nanni e Lumini (2009) nel loro studio hanno ottenuto risultati e performance migliorate rispetto all'utilizzo dei classificatori stand-alone. I modelli base sono combinati principalmente per ottenere due miglioramenti: aumentare la stabilità e l'accuratezza del modello ed evitare il fenomeno di overfitting[9]. Questo perché spesso i classificatori presi singolarmente presentano delle capacità di predizione limitate per costruzione e delle caratteristiche ben definite che portano a risultati non sempre esaustivi, al contrario mettendo insieme più classificatori, dello stesso tipo o di tipi differenti, il risultato finale è costituito dalla combinazione di tanti risultati intermedi. Inoltre, benché gli strumenti di AI siano riconosciuti come *Black Box*[10], questi metodi permettono di ottenere buona accuratezza previsionale e "spiegabilità" dei risultati; in un certo senso si può affermare che il metodo ensamble permette la diversificazione della predizione attraverso più di un algoritmo. In base all'approccio usato per costruire un metodo

[9] Il problema di overfitting si presenta quando un modello è sovra-adattato ai dati di training. Questo comporta la perdita di validità generale del modello, il modello avrà performance molto alte rispetto ai dati di training ma basse per nuovi dati, rendendolo di fatto inutilizzabile.

[10] Un sistema Black Box permette di vedere solo gli input e gli output di un modello senza conoscerne i processi e le scelte interni.

ensamble distinguiamo principalmente tre tipologie di ensamble metohod: il *Bagging*, il *Boosting* e lo *Stacking*. I più diffusi nella letteratura risultano essere il Bagging e il Boosting (Qu, Quan, Lei, Shi; 2019).

Il metodo Bagging, Breiman (1996), conosciuto anche come Bootstrap aggregating, utilizza un numero arbitrario di classificatori base che forniscono ognuno in output una predizione. Ogni predizione ha la stessa importanza, quindi lo stesso peso, nella definizione della predizione finale. Questo metodo è implementato attraverso la tecnica di campionamento Bootstrap che si basa sul campionamento casuale con rimpiazzo del dataset iniziale per ottenere tanti subset, di dimensione uguale, quanti sono i classificatori base; questo significa che una stessa istanza può essere presente più di una volta nello stesso subset. Il processo di Bagging avviene in parallelo per ogni classificatore. Uno dei modelli più conosciuti e usati che sfrutta questa tecnica è l'algoritmo Random Forest che utilizza classificatori base del tipo Decision Tree che in output forniscono una predizione ciascuno, e il cui ultimo passo è l'aggregazione dei risultati attraverso il criterio majority voting.

Un esempio di questa tipologia di modello ensamble è illustrato in Figura 5.

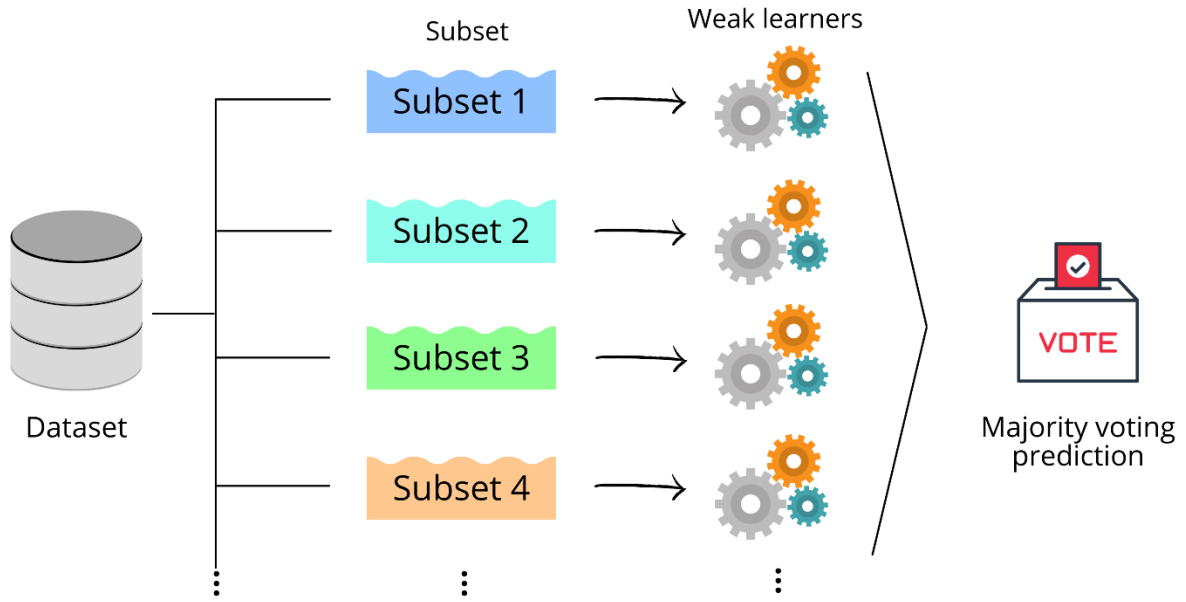

Figura 5 - Processo di Bagging

Il modello Boosting, Freund et al. (1999), è costituito da un processo sequenziale. In particolare, i classificatori sono in sequenza, a differenza del metodo Bagging, e a ogni predizione il modello individua gli errori di classificazione e cerca di correggerli nell'iterazione

successiva assegnando un peso maggiore alle istanze classificate in modo errato. Il processo termina quando è raggiunto il numero massimo di iterazioni scelte oppure quando l'error rate scende al di sotto di una soglia limite. L'idea dietro questo modello è che una sequenza di classificatori deboli che imparano dagli errori dei classificatori precedenti costruiscono un classificatore forte. Questa tecnica è utilizzata nell'algoritmo AdaBoost (Adaptive Boosting) e negli algoritmi Gradient Boosting, come Xgboost.

Nelle Figure 6 e 7 è illustrato il processo iterativo dell'algoritmo AdaBoost e il processo generale che caratterizza i metodi Boosting.

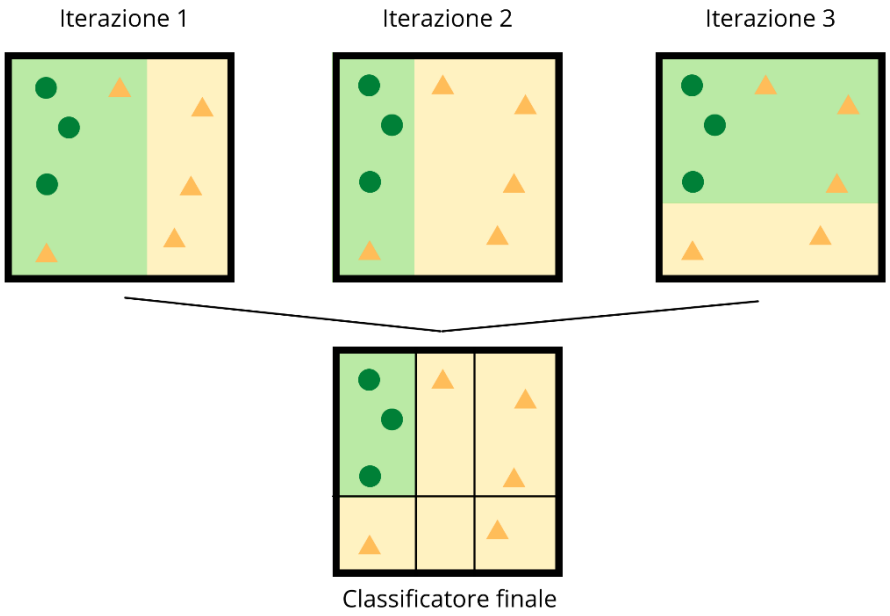

Figura 6 - Visualizzazione algoritmo Adaboost

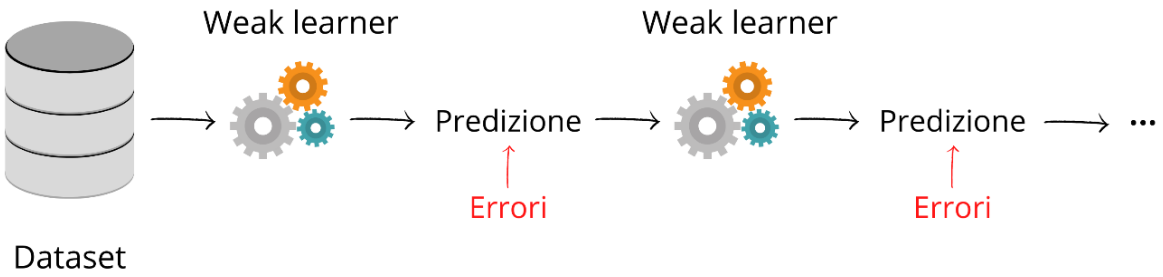

Figura 7 - Processo sequenziale Boosting

L'ultimo tipo di modello ensamble è lo Stacking. Questa tipologia si compone di un insieme di classificatori base che operano in modo parallelo e forniscono in output l'input per un ulteriore classificatore, chiamato meta-model, che svolge la previsione finale basandosi sui risultati dei weak learners. Una caratteristica di questo modello è la possibilità di usare classificatori di tipo diverso tra loro, a differenza dei modelli precedentemente illustrati che utilizzano classificatori omogenei. Un modello Stacking di esempio potrebbe essere formato da 3 weak learners, come KNN, Random Forest e SVM, e utilizzare come Stacking Model Learner una rete neurale. È possibile costruire Multi-levels Stacking aggiungendo alla struttura standard uno o più livelli di meta-models intermedi che precedono il final meta-model, in questo caso il modello può diventare data expensive o time expensive.

In Figura 8 è presentata una grafica che illustra il metodo Stacking:

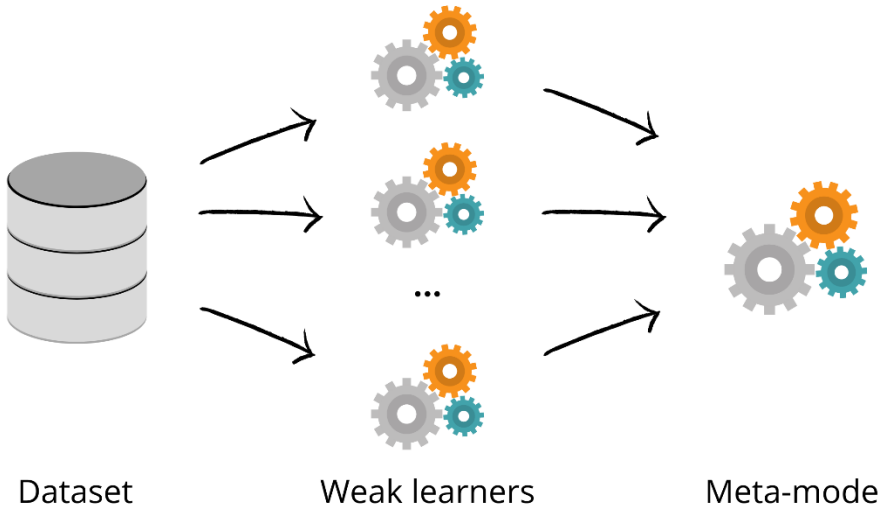

Figura 8 - Processo di Stacking

Tra i numerosi studi che utilizzano tecniche ensamble si menzionano: Friedman (2001) incentrato sull'utilizzo dell'algoritmo Gradient Boosting; Kruppa et al. (2013) validano l'algoritmo Random Forest (Bagging) nella valutazione del rischio di credito; Kim e Upneja (2015) utilizzano l'algoritmo Adaboost per la predizione di crisi finanziarie; Zieba, Tomczak e Tomczak (2016) comparano vari metodi e algoritmi, il migliore risulterà essere l'Extreme Gradient Booststing (XGBoost); Barboza, Kimura e Altman (2017) confermano le migliori performance dei metodi AI rispetto a quelli classici; Carmona, Climent e Momparler (2018) sviluppano un modello XGBoost applicato al settore bancario.

Per approfondire la storia e l'evoluzione dei metodi e degli studi riguardo la predizione del fallimento e delle crisi aziendali è possibile fare riferimento alle seguenti review presenti in letteratura: Bellovary, Giacomino e Akers (2007); Balcen e Ooghe (2006); Shi e Li (2019); Kumar e Ravi (2007); Clement (2020); Qu, Quan, Lei e Shi (2019); Hauser e Booth (2011) e Perboli e Arabnezhad (2020). Il contributo complessivo permette di coprire un periodo temporale che inizia nel 1930 circa e termina nel 2020.

4 Data Mining

In questo capitolo è presentata una breve introduzione al Data Mining e alle fasi del KDD (Knowledge Discovery in Databases), utile alla comprensione della struttura del modello predittivo, oggetto del lavoro di tesi, presentato nel prossimo capitolo.

4.1 Introduzione al Data Mining

Il necessario sfruttamento delle grandi quantità di dati disponibili presso le aziende al fine di avviare processi data-driven è reso possibile anche grazie al Data Mining. Il termine Data Mining si utilizza per indicare specificamente il processo, reso automatico tramite opportuni algoritmi, per mezzo del quale si estraggono dai grandi archivi di dati informazioni implicite, non note a priori e utili.

Il processo di Data Mining è una delle fasi che caratterizza il processo generale di ricerca di nuova conoscenza a partire dai dati comunemente conosciuto come Knowledge Discovery in Databases (KDD). In particolare, con il termine KDD si fa riferimento all'intero processo che dai dati grezzi arriva all'interpretazione dei pattern estratti che permette di sviluppare conoscenza. Tipicamente il KDD è un processo iterativo che usa il metodo "Trial and Error[11]" per fornire una soluzione.

[11] Il metodo euristico "Trial and Error", o anche "prova e sbaglia", punta a trovare una soluzione ad un problema effettuando tentativi diversi fino al raggiungimento della soluzione desiderata.

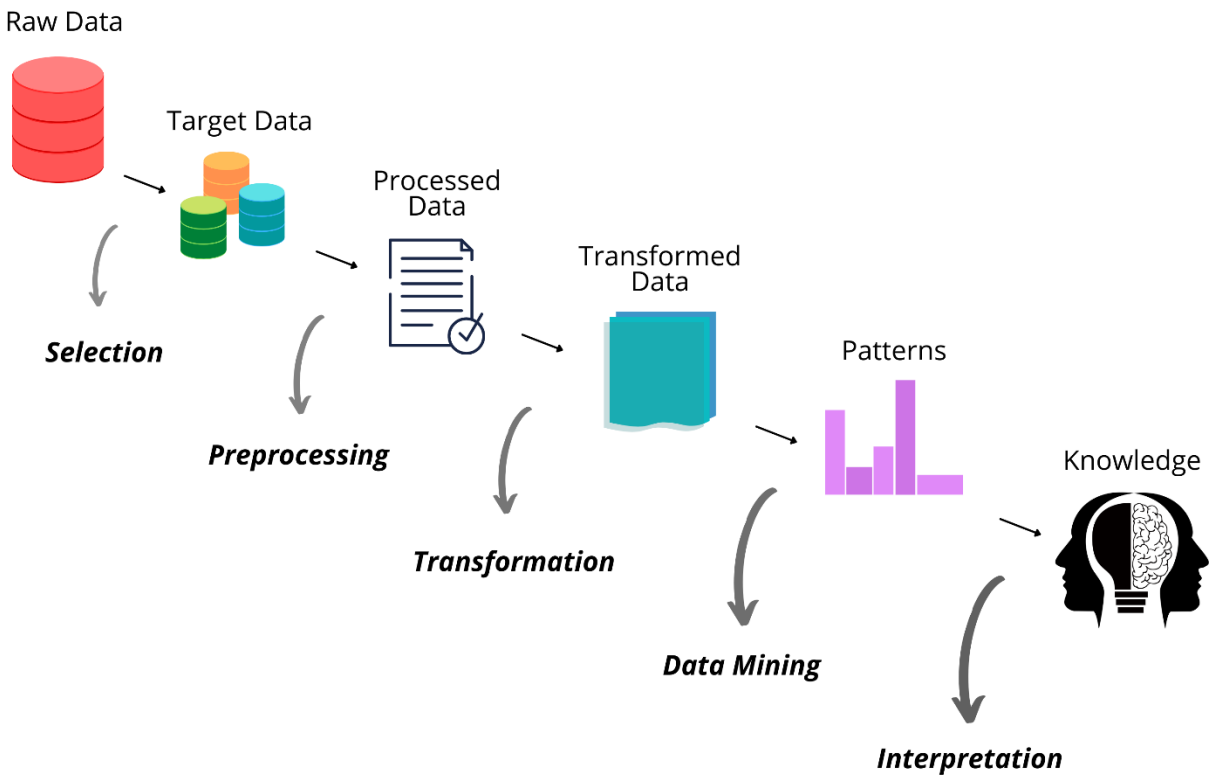

Figura 9 - Knowledge Discovery in Databases: processo e attività del KDD

Come si evince dalla Figura 9, l'input del processo KDD è costituito dai dati grezzi di partenza (Raw Data) che attraverso una selezione di dati e variabili, che deve tener conto del dominio di applicazione a cui si fa riferimento e degli obiettivi prefissati, produce una serie di dataset che saranno oggetto dell'elaborazione. Il passo successivo prende il nome di preprocessing e consiste nel "pulire" i dati. Infatti, nel mondo reale i dati disponibili risultano essere "sporchi", affetti da rumore, inaccurati, talvolta mancanti e di bassa qualità; attraverso il Data Cleaning si effettua una prima manipolazione dei dati volta a ridurre ed eliminare dove possibile queste problematiche (riduzione del rumore, individuazione di outlier[12], intercettazione dei dati inconsistenti). Un ruolo importante in questa fase è giocato anche dalla Data Integration che punta all'integrazione dei dati di partenza con quelli provenienti da altre risorse per gestire il problema dei dati mancanti (missing values), utilizzando metadati per arricchire i dati disponibili e andando anche ad eliminare possibili problematiche di conflitto tra i dati e di ridondanza. Ottenuti i dati preprocessati avviene il processo di trasformazione che ha lo scopo di modellare e adattare i dati per gli studi successivi attraverso vari metodi quali la

[12] Un outlier è un valore che in un insieme di osservazioni risulta essere anomalo e distante dalle altre osservazioni

normalizzazione, la binarizzazione o la categorizzazione. Le ultime due fasi descritte sono di fondamentale importanza per la qualità dei risultati ottenuti in output del KDD e risultano onerose in termini computazionali e anche a livello di tempo sono tra le più impegnative. In base all'obiettivo prefissato (problema di categorizzazione, problema di clustering, problema di regressione ecc.) si scelgono gli algoritmi e i metodi per estrarre informazioni nel processo di Data Mining e i relativi parametri. Gli algoritmi impiegati solitamente fanno riferimento al mondo AI e a tecniche di Machine Learning, ma possono talvolta anche derivare da tecniche classiche statistiche. In questa fase si ricavano i pattern significativi che sono in seguito interpretati e validati, anche con l'uso di tecniche di Data Visualization, permettendo l'estrazione di conoscenza e quindi di valore per l'azienda.

5 Il Modello

Questo capitolo raccoglie tutte le informazioni e le descrizioni dei processi relativi al modello sperimentale sviluppato in questo lavoro di tesi e si occupa dei seguenti argomenti: obiettivo della ricerca (5.1), metodologia usata (5.2) e ambiente di sviluppo del modello (5.3). Il primo paragrafo è fondamentale per definire lo scopo del lavoro svolto e la sua struttura; nel paragrafo 5.2, che risulta essere il più corposo, sono presentate le varie tecniche e metodologie usate nella realizzazione del modello che segue un approccio "*trial and error*", mentre nell'ultimo paragrafo è presentato l'ambiente di sviluppo della ricerca e nel dettaglio il processo *RapidMiner* con gli operatori relativi alle varie fasi.

5.1 L'obiettivo della ricerca

Obiettivo di questa tesi è proporre un modello predittivo del fallimento aziendale a 36 mesi attraverso variabili di tipo finanziario, sfruttando le potenzialità dell'AI. Il modello utilizza algoritmi per la classificazione di tipo binario (azienda fallita e azienda ancora attiva) ed è incentrato sull'ambito nazionale italiano in quanto tratta dati provenienti dalla banca dati AIDA[13]. L'idea alla base di questo lavoro è quella di progettare uno strumento capace di individuare con un anticipo di almeno tre anni il rischio di fallimento attraverso un processo di apprendimento automatico e *data-driven*, fornendo al management aziendale sufficiente tempo per ideare e implementare azioni correttive in modo tale da evitare il fallimento aziendale.

Nella ricerca sperimentale questo si traduce nella predizione della variabile di interesse "Fallita" attraverso un processo di classificazione binaria. Tale processo consiste nell'assegnare un'etichetta a insiemi di dati, in origine non etichettati, attraverso l'uso di un classificatore[14], costruendo, dapprima, in modo induttivo un modello basato su insiemi di dati etichettati a priori (fase di apprendimento), per poi applicare, in una fase successiva, tale modello a insiemi di dati non etichettati ottenendo, secondo un processo di tipo deduttivo, la

[13] AIDA (Analisi Informatizzata delle Aziende Italiane) è un database realizzato da Bureau van Dijk S.p.A. ed è il punto di riferimento per la ricerca strutturata su dati aziendali nel contesto italiano. Contiene i dati anagrafici, i bilanci, gli indici economico-finanziari e numerose altre informazioni utili riferite a società di capitale italiane attive e fallite, con serie storiche fino a 10 anni. Il programma di ricerca, consultazione ed esportazione delle informazioni del database permette di sfruttare la risorsa in modo semplice e personalizzato.

[14] In questo caso il termine "classificatore" è usato in senso generale e si riferisce all'insieme delle tecniche di classificazione.

predizione della classe di appartenenza associata a ciascun insieme. In Figura 10 è illustrato il processo di classificazione.

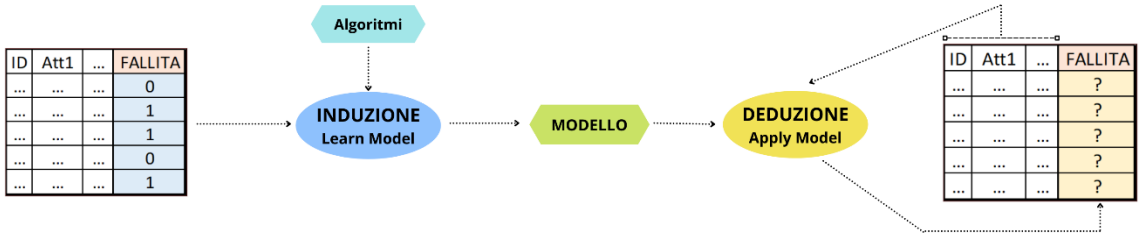

Figura 10 - Classificazione

La Figura 11, invece, illustra il framework del modello sperimentale sviluppato.

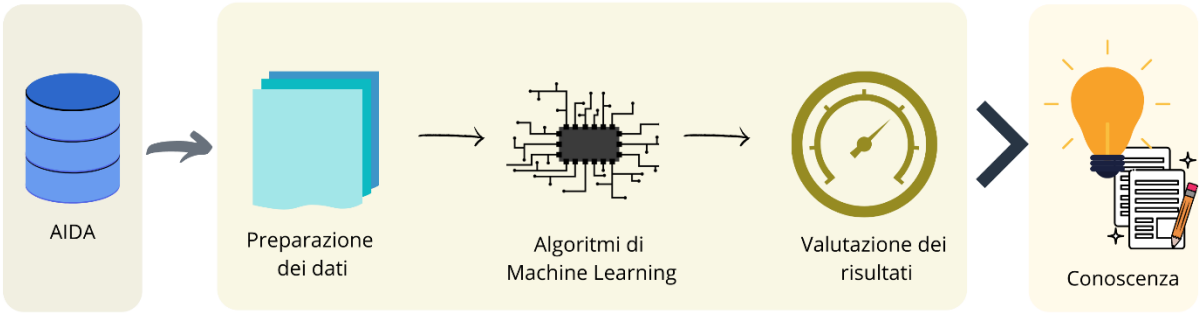

Figura 11 - Framework del modello

5.2 Metodologia

5.2.1 Selezione e analisi dei dati iniziali

Il dataset di partenza alla base di questo studio si riferisce ad aziende italiane ed è stato estratto dalla banca dati AIDA. I dati in oggetto rappresentano circa 137 mila[15] aziende e per ognuna di esse sono presenti 55 attributi caratterizzanti. Le aziende considerate possono essere distinte in due classi principali: aziende ancora attive e aziende fallite; per semplicità ci riferiamo alla Classe 0 per le aziende attive e alla Classe 1 per le aziende fallite. Questa informazione è contenuta nell'attributo binario "Fallita" che risulta quindi essere la variabile d'interesse del modello.

Gli altri attributi possono essere distinti per tipologia di informazione e sono principalmente dati anagrafici e indici riguardanti l'azienda. In particolare, il dataset è composto, oltre che

[15] Aziende totali: 136.781

dalla variabile di interesse "Fallita", da 17 attributi che contengono informazioni principalmente anagrafiche e di governance (come la ragione sociale, i vari identificativi univoci, la classificazione ATECO[16] e il tipo di bilancio presentato). Gli altri 37 attributi sono degli indici di varia tipologia: indici finanziari, indici della gestione corrente, indici di redditività e altri dati significativi. Nella Tabella 1 è presentata la lista completa di tutti gli attributi presenti nel dataset iniziale e la relativa tipologia di informazione.

Tabella 1 - Elenco degli attributi del dataset iniziale

ID	Attributo (• tipo)	ID	Attributo (• tipo)
X1	• Fallita	X29	• Grado di copertura degli interessi passivi
X2	• Ragione sociale	X30	• Oneri finanz. su fatt. (%)
X3	• Numero CCIAA	X31	• Indice di indip. finanz. (%)
X4	• Partita IVA	X32	• Grado di indip. da terzi
X5	• Ragione sociale	X33	• Posizione finanziaria netta
X6	• Forma giuridica	X34	• Debt/Equity ratio
X7	• Anno di costituzione	X35	• Debt/EBITDA ratio
X8	• ATECO 2007 codice	X36	• Rotaz. cap. investito (volte)
X9	• ATECO 2007 descrizione	X37	• Rotaz. cap. cir. Lordo (volte)
X10	• Ultimo modello di contabilità - Bilancio	X38	• Incidenza circolante operativo (%)
X11	• Indicatori d'indipendenza BvD	X39	• Giac. media delle scorte (gg)
X12	• N° di azionisti presenti	X40	• Giorni copertura scorte (gg)
X13	• Azionisti Ricavi delle vendite (Fatt.)(k€)	X41	• Durata media dei crediti al lordo IVA (gg)
X14	• CSH - First name	X42	• Durata media dei debiti al lordo IVA (gg)
X15	• CSH - Last name	X43	• Durata Ciclo Commerciale (gg)
X16	• CSH - Also a manager	X44	• EBITDA (k€)
X17	• Data di chiusura della procedura	X45	• EBITDA/Vendite (%)
X18	• Procedura/Cessazione	X46	• Redditività del totale attivo (ROA) (%)
X19	• Indice di liquidità	X47	• Redditività del capitale investito (ROI) (%)
X20	• Indice corrente	X48	• Redditività delle vendite (ROS) (%)
X21	• Indice di indebitam. a breve	X49	• Redditività del capitale proprio (ROE) (%)
X22	• Indice di indebitam. a lungo	X50	• Incid. oneri/Proventi extrag. (%)
X23	• Indice di copertura delle immob. (patrim.)	X51	• Capitale circolante netto (k€)
X24	• Grado di ammortamento	X52	• Margine sui consumi (k€)
X25	• Rapporto di indebitamento	X53	• Margine di tesoreria (k€)
X26	• Indice di copertura delle immob. (fin.)	X54	• Margine di struttura (k€)
X27	• Debiti v/banche su fatt. (%)	X55	• Flusso di cassa di gestione (k€)
X28	• Costo denaro a prestito (%)		

• Var. interesse •Anagrafica e governance •Indici fin. •Indici gestione corrente •Indici redditività •Altri dati significativi

[16] La classificazione ATECO (*ATtività ECOnomiche*) è una tipologia di classificazione delle attività economiche adottata dall'ISTAT (Istituto nazionale di statistica [italiano]) che si compone di un codice alfanumerico che indica, con sei livelli di dettaglio, i settori economici.

Gli attributi da X19 a X55 si riferiscono all'*ultimo anno disponibile-3* fornito da AIDA: per le aziende attive si tratta dei dati relativi al terzo anno antecedente all'ultimo anno disponibile, corrispondente all'anno dell'ultimo bilancio depositato, mentre per le aziende fallite ci si riferisce ai dati relativi al terzo anno antecedente la dichiarazione di fallimento. Questa scelta deriva dal fatto che il fallimento di un'azienda non è un evento improvviso ma solitamente è preceduto da segnali di crisi negli anni precedenti. I dati degli anni immediatamente prossimi al fallimento risulterebbero, inoltre, essere "drogati" da dinamiche prefallimentari e di redazione dei bilanci e solitamente rappresentano in modo evidente l'imminente e inevitabile fallimento. In questo lavoro di tesi, invece, l'obiettivo è quello di fornire uno strumento di monitoraggio e controllo della crisi aziendale a medio termine, fornendo così l'opportunità di intercettare i segnali per tempo in modo tale da poter intervenire con azioni correttive al fine di evitare il fallimento.

Da una prima analisi delle caratteristiche dei dati risultano evidenti alcune criticità. In particolare, il dataset risulta essere sbilanciato se si prende in considerazione la variabile d'interesse "Fallita", il numero di aziende che risultano essere attive (Classe 0) è circa il doppio di quelle fallite (Classe 1), come si evince in Figura 12.

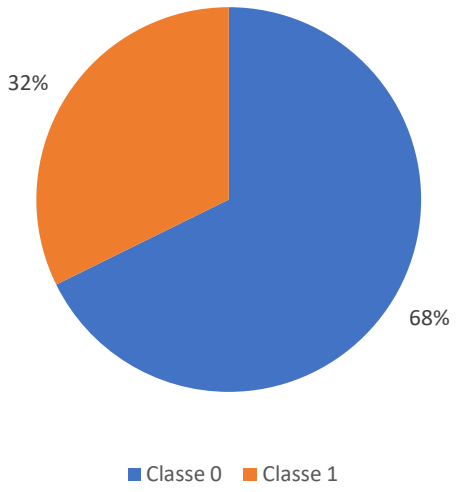

Figura 12 - Sbilanciamento Dataset

Un altro limite dei dati in esame è l'elevato numero di valori mancanti, o *missing value*, che caratterizzano il dataset. Questo problema, in generale, è molto diffuso per dati provenienti dal mondo reale. In particolare, si evidenzia una distribuzione dei missing value non omogenea all'interno delle varie feature. Gli attributi con il numero di dati mancanti più elevato, con una

percentuale *missing value*/dati totali maggiore del 15%, sono in ordine: X43, X40, X28, X24, X39, X47, X29 e X50 (in Tabella 2 sono presentate le rispettive percentuali). Gli altri attributi presentano un numero di *missing value* più contenuto (dall'8% allo 0%) con media percentuale complessiva di circa 2.5% e pertanto risultano più facilmente gestibili con tecniche di stima dei valori mancanti e con tecniche statistiche.

Tabella 2 - Attributi con missing value > 15%

ID	***Attributo***	***n° missing***	***% missing***
X43	Durata Ciclo Commerciale (gg)	67010	48.99%
X40	Giorni copertura scorte (gg)	61607	45.04%
X28	Costo denaro a prestito (%)	56750	41.49%
X24	Grado di ammortamento	55883	40.86%
X39	Giac. media delle scorte	50070	36.61%
X47	Redditività del capitale investito (ROI) (%)	39530	28.90%
X29	Grado di copertura degli interessi passivi	32884	24.04%
X50	Incid. oneri/Proventi extrag. (%)	23782	17.39%

Risulta interessante approfondire l'analisi dei valori mancanti separatamente per le aziende fallite (Classe 1) e attive (Classe 0) e le feature per determinare se lo stato dell'azienda influisca su queste variabili. Dallo studio emerge una tendenza generale e diffusa che evidenzia, a parità di feature analizzata, un numero più elevato di *missing value* per le aziende di Classe 1 rispetto a quelle di Classe 0 e questo andamento è più rimarcato per alcuni attributi presentati nella Tabella 3. In generale la percentuale di *missing value* per le aziende di Classe 1 è infatti pari al 12.24% mentre quella delle aziende di classe opposta è ridotta all'8.32%. Questo aspetto è da tenere in considerazione nelle fasi successive del modello e può suggerire che le aziende che risultano essere più "sane" forniscono dati mediamente più completi a differenza delle aziende in cattive condizioni.

Tabella 3 - Missing value per feature distinti per classe

ID	Attributo	Missing value (Classe 0)	Missing value (Classe 1)	Delta (Classe1 - Classe0)
X47	Redditività del capitale investito (ROI) (%)	23.5%	40.2%	16.7%
X50	Incid. oneri/Proventi extrag. (%)	12.4%	27.8%	15.4%
X28	Costo denaro a prestito (%)	37.0%	50.8%	13.8%
X49	Redditività del capitale proprio (ROE) (%)	4.1%	16.8%	12.7%
X29	Grado di copertura degli interessi passivi	20.1%	32.4%	12.3%
X43	Durata Ciclo Commerciale (gg)	45.6%	56.0%	10.4%
X54	Margine di struttura (k€)	2.6%	12.4%	9.8%
X32	Grado di indip. da terzi	4.1%	13.5%	9.5%

Anche l'analisi delle singole aziende e del rispettivo numero di valori mancanti evidenzia un fenomeno simile. La distribuzione delle aziende in base al numero dei missing value, illustrata nella Figura 13, evidenzia una concentrazione differente per i due tipi di classi.

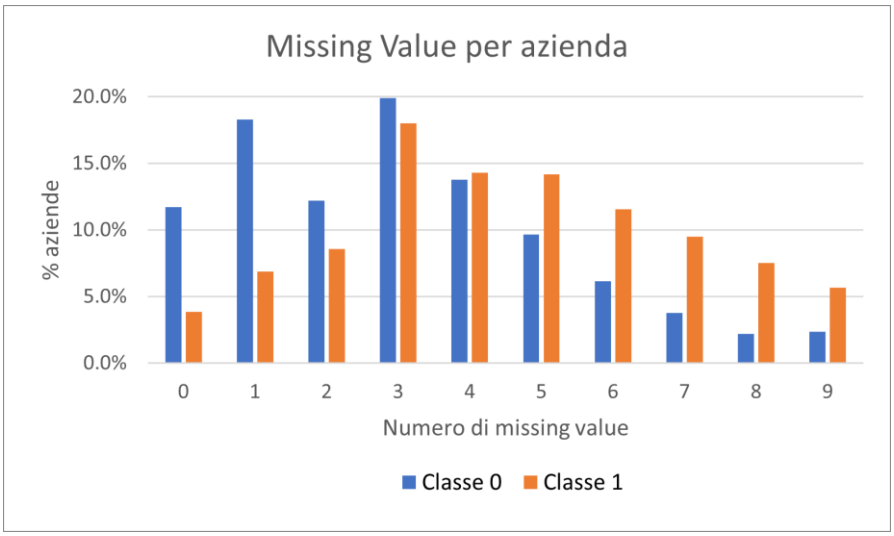

Figura 13 - Distribuzione aziende per numero di missing value

La percentuale di aziende di Classe 0 che ha *al più 3 attributi mancanti* è pari a circa il 62.1%, mentre per quanto riguarda la Classe 1 è solamente il 37.3%. Quanto emerge dalla figura indica che in media i dati delle aziende attive sono più completi e dettagliati in relazione ai dati provenienti da aziende destinate al fallimento.

5.2.2 Preprocessing e Transformation

Una delle fasi più importanti nella costruzione di modelli di Machine Learning è quella del preprocessamento e trasformazione dei dati. Questo passaggio permette la preparazione dei dati per i passi successivi di apprendimento e permette di ottimizzare i risultati ottenuti.

5.2.2.1 Feature extraction

Un primo intervento di *Feature Transformation* sui dati ha permesso di sfruttare l'attributo X8, "*ATECO 2007 codice"*, attraverso una aggregazione e categorizzazione. Il codice ATECO, che permette la classificazione delle attività economiche, riportato da AIDA è composto da sei cifre con un conseguente livello di dettaglio troppo elevato. Al fine dell'analisi si è provveduto a creare una suddivisione per settore di attività attraverso una categorizzazione delle prime due cifre del codice. In particolare, si è creato un nuovo attributo X56 "*SettoreAz"* che rappresenta il settore in cui opera l'azienda e che può assumere quattro valori diversi: *Industria, Pubblico, Commercio e Servizi*. La categorizzazione effettuata è la seguente:

- Codici ATECO da 1 a 33, 41, 42 e 43 corrispondono al settore *Industria;*
- Codici ATECO da 35 a 39, 84, 85 e 99 corrispondono al settore *Pubblico;*
- Codici ATECO da 44 a 47 corrispondono al settore *Commercio*;
- Codici ATECO da 49 a 82 e da 86 a 97 corrispondono al settore *Servizi*.

Dall'analisi di questo nuovo attributo emerge che il settore *Industria* è il più rappresentativo del dataset, con circa 55 mila aziende, e pertanto, anche in ottica di riduzione della mole di dati e dello sforzo computazionale, si è deciso di analizzare separatamente le aziende di tale settore (dopo una prima analisi estesa a tutti i settori).

5.2.2.2 Feature scaling

I dati di tipo numerico sono stati trattati attraverso un processo di normalizzazione che ha permesso l'avvicinamento delle distribuzioni degli attributi alla distribuzione Gaussiana che facilita la comparazione tra i diversi dati. In particolare, è stata usata la tecnica *Z-Transformation* che segue la trasformazione:

$$Z = \frac{x - \mu_x}{\mu_x}$$

dove x è la variabile da trasformare, μ_x è la media dei dati e μ_x è la loro varianza.

5.2.2.3 Trattamento di dataset sbilanciati

Lo sbilanciamento del dataset, presentato nel paragrafo 5.2.1, ha reso necessario un intervento sui dati fondamentale al fine di migliorare le performance del modello. Questa esigenza deriva dal fatto che, spesso, un classificatore tende a sovrastimare la classe

maggioritaria quando è addestrato con un dataset sbilanciato. L'obiettivo di questo trattamento dei dati punta a garantire che il training del modello sia effettuato su dataset con una adeguata rappresentazione di entrambe le classi. Per affrontare questo problema sono state individuate due metodologie: l'*Undersampling* e il *Cost-Sensitive*.

L'*Undersampling* è una tecnica di campionamento ed è tra le più usate per affrontare problemi di classificazione con classi sbilanciate grazie alla sua facilità di implementazione unita alla sua efficacia. Ricordando che nel nostro caso siamo di fronte a un problema di classificazione di tipo binario, la tecnica opera sul training set con lo scopo di ribilanciare le due classi andando ad effettuare un campionamento della classe maggioritaria (Classe 0) e permette di sfruttare tutte le istanze della classe minoritaria (Classe 1) nel processo di addestramento. Alla fine del processo il training set è composto per il 50% da dati della classe maggioritaria e per il restante 50% da dati della classe minoritaria. In Figura 14 è illustrata la tecnica di Undersampling applicata alla classe maggioritaria (Classe 0).

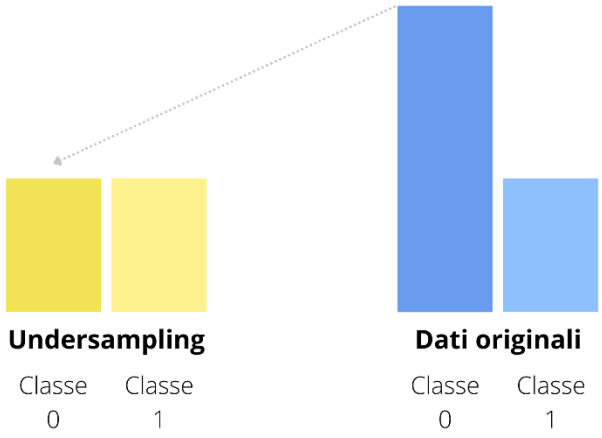

Figura 14 – Undersampling

Un'altra tecnica usata è basata sull'approccio *Cost Sensitive.* La maggior parte degli algoritmi di apprendimento valuta tutti gli errori di classificazione nello stesso modo, attribuendo lo stesso peso; questo non si adatta a molti problemi reali in cui sbagliare predizione di un caso positivo o minoritario è più grave che classificare in modo errato un'istanza della classe negativa o maggioritaria. Nel caso del fallimento aziendale, in un modello di previsione della crisi aziendale come strumento di allerta, ad esempio, un errore nel predire un'azienda che

fallirà come sana comporta rischi maggiori, e quindi costi maggiori, rispetto ad un errore dovuto alla predizione di un'azienda sana come in crisi.

Il *Cost Sensitive Learning* prende in considerazione diversi tipi di pesi in base al tipo di errore attribuendo un costo specifico alla predizione errata durante la fase di apprendimento del modello rendendo l'uso di questa tecnica molto comune e affine nella risoluzione di problemi con dataset sbilanciati. I costi, o pesi, relativi ai diversi tipi di errore del classificatore, o anche delle classificazioni corrette, sono contenuti nella Matrice dei Costi, strutturata partendo dalla *Confusion Matrix* descritta nel dettaglio al paragrafo 5.2.4. In Tabella 4 è presentata la Matrice dei Costi.

Tabella 4 - Matrice dei Costi

		CLASSE PREDETTA	
		Classe=1 Positiva	Classe=0 Negativa
CLASSE REALE	Classe=1 Positiva	Ricavo TP	Costo FN
	Classe=0 Negativa	Costo FP	Ricavo TN

I quattro valori contenuti nella matrice di confusione rappresentano i seguenti costi/ricavi:

- *Ricavo TP* (*True Positive*): ricavo associato alla corretta classificazione di un'istanza Positiva;
- *Costo FP* (*False Positive*): costo associato alla classificazione errata di un'istanza Negativa nella realtà classificata come Positiva;
- *Ricavo TN* (*True Positive*): ricavo associato alla corretta classificazione di un'istanza Negativa;
- *Costo FN* (*False Negative*): costo associato alla classificazione errata di un'istanza Positiva nella realtà classificata come Negativa.

In particolare, nel problema affrontato in questo studio è stato attribuito un costo maggiore ai falsi positivi (FP), per trattare il problema dello sbilanciamento delle classi e si è lavorato su un ricavo maggiore per la corretta previsione della classe positiva (Classe 1) (TP) e sul costo

dei falsi negativi (FN) per cercare di ottenere un equilibrio tra Recall e Precision della classe positiva (vedi paragrafo 5.2.4.1).

5.2.2.4 Trattamento Missing Value

I dati mancanti all'interno del dataset sono numerosi, come indicato nel paragrafo 5.2.1, in quanto molti indici e variabili non sono disponibili per una buona parte delle aziende. Ricordando che le prestazioni di modelli di machine learning sono strettamente legati alla percentuale di dati mancanti si è deciso di intervenire attraverso tre diverse strategie di trattamento dei *missing value*: l'eliminazione di attributi con dati mancanti maggiori del 15%, la sostituzione dei *missing value* con la media dei dati disponibili della feature e, infine, la strategia di non intervenire affatto, ignorando i *missing value* nell'analisi.

La prima strategia esclude gli attributi con una percentuale di dati mancanti troppo elevata (soglia definita >15%), che risultano essere otto: X43, X40, X28, X24, X39, X47, X29 e X50. Questo permette di ottenere un set di dati più completo e compatto.

Una strategia molto utilizzata, derivante dalla statistica, consiste nel sostituire i dati mancanti con una costante specifica per ogni attributo: nello svolgimento di questo lavoro di tesi come costante sostitutiva si usa la media dei dati disponibili (*mean imputation*).

Nel corso della creazione del modello queste due strategie sono state combinate insieme al fine di creare un dataset completo, senza valori mancanti, ma allo stesso tempo senza snaturare le informazioni degli attributi evitando di stimare i *missing value* degli attributi più incompleti.

Una terza via percorsa è stata quella di ignorare i *missing value* e svolgere l'addestramento del modello con i dati così come forniti basandosi sulla capacità di funzionamento di alcuni algoritmi non sensibili ai *missing values*.

Un quarto potenziale approccio consiste nella stima dei dati mancanti attraverso modelli d'apprendimento che utilizzano un *learner* per la stima dei valori di ogni attributo (ad esempio, KNN imputation "k-nearest neighbor"). Queste tecniche risultano, però, molto pesanti in termini computazionali e temporali e pertanto non sono state utilizzate.

Un'ultima strategia inizialmente applicata prevedeva l'eliminazione tout-court delle istanze con *missing value* (o che superavano un certo numero stabilito di dati mancanti), ma questo

approccio comportava una riduzione molto significativa del dataset con perdita di molte informazioni, portando a performance non accettabili e, pertanto, è stato abbandonato.

5.2.2.5 Feature selection

Il processo di *feature selection* è uno dei passi più importanti nel processo di preparazione dei dati e ha l'obiettivo di selezionare gli attributi più significativi per la classificazione e nello stesso tempo di permettere una riduzione della dimensionalità dei dati rendendo più semplice e interpretabile il modello, riducendone la varianza e limitando il fenomeno di *overfitting*.

In prima battuta sono stati individuati gli attributi irrilevanti, ovvero feature che non contengono informazioni utili nel processo di estrazione della conoscenza. La selezione degli attributi significativi è stata svolta escludendo dai dati in input del modello i dati anagrafici e di governance; un'eccezione è stata fatta per l'attributo X8, che, dopo la rielaborazione descritta nel paragrafo 5.2.2.1, è stato viceversa sfruttato nella selezione delle aziende utilizzate nel modello. Il dataset utilizzato per la costruzione del modello è quindi composto dalla variabile di riferimento X1 "Fallita" e 37 attributi che comprendono gli indici finanziari, di gestione corrente, di redditività e i dati della categoria "Altri dati significativi". Le analisi svolte in seguito prendono in considerazione gli attributi: X1 e da X19 a X55.

Un secondo processo di feature selection è stato intrapreso per individuare gli attributi maggiormente importanti tra quelli rimasti ed è stato effettuato utilizzando due tecniche di tipo wrapper.

Il primo approccio usa la *backward elimination*: partendo dal set di attributi completo si esclude dall'analisi un attributo per volta, si valuta la perdita di performance relativa (nel nostro caso attraverso un algoritmo Random Forest valutato tramite Cross Validation, argomenti approfonditi nel paragrafo 5.2.3.1 e 5.2.4.2) e dopo aver processato tutti gli attributi si individua quello la cui eliminazione comporta la diminuzione di performance minore che, a questo punto, viene escluso dalla selezione ottimale. Il processo si ripete iterativamente lavorando sulla nuova selezione ottenuta di volta in volta e termina al raggiungimento di una soglia prestabilita di iterazioni o di massima perdita di accuratezza. Questo tipo di algoritmo rientra nella famiglia dei cosiddetti algoritmi *greedy* che si caratterizzano per la ricerca di una soluzione globale attraverso l'individuazione di soluzioni ottime locali ad ogni passo.

Il secondo approccio utilizza una tecnica di tipo evolutivo sfruttando algoritmi genetici (GA), ovvero euristiche che, partendo da una soluzione, la fanno evolvere attraverso modifiche casuali, usando tecniche di ricombinazione, mutazione, selezione e crossover, convergendo ad una soluzione migliore. In particolare, le tecniche più usate dai GA nel problema di feature selection sono quella di mutazione che prevede l'inclusione o meno di un attributo nella selezione testata (attributo X on-off) e quella di crossover che si basa sullo scambio degli attributi usati nella selezione. La selezione invece è la fase che permette l'individuazione degli attributi per la fase successiva di crossover e nel nostro caso segue il metodo di *tournament selection*. Anche in questo caso l'algoritmo utilizzato nel processo di valutazione è il Random Forest tramite Cross Validation. Questo metodo richiede uno sforzo computazionale considerevole e pertanto è stato applicato su un campione stratificato ristretto di circa ottomila istanze.

Inoltre, attraverso l'uso di un classificatore di tipo Random Forest è possibile ottenere un peso relativo per ogni attributo che corrisponde alla sua importanza all'interno del modello (*Feature Importance*). Questo permette di interpretare il modello e di individuare le variabili che ne determinano i risultati.

5.2.2.6 Sampling

L'utilizzo di un algoritmo di *sampling* è un approccio comunemente usato per la selezione di un subset di dati da analizzare, nota come *Data Reduction*, ed è impiegato per abbattere i costi computazionali e in termini di tempo che avrebbe l'analisi se svolta sul dataset iniziale e completo. Un buon campionamento, per essere significativo, deve essere rappresentativo del dataset da cui sono estratti i dati e può essere ottenuto attraverso diverse tecniche di campionamento probabilistiche, tra le quali, le più conosciute sono: il campionamento semplice con rimpiazzo, il campionamento semplice senza rimpiazzo e il campionamento stratificato.

Nel lavoro di tesi è utilizzato il campionamento stratificato che prevede la divisione della popolazione in sottopopolazioni, o strati, in modo tale da garantire omogeneità all'interno dello stesso strato; successivamente si estrae un campione da ogni strato e si compone il campione stratificato che in questo modo mantiene la significatività e la distribuzione delle classi della popolazione iniziale.

5.2.3 Data Mining: Algoritmi

La fase di *Data Mining* prevede la scelta degli algoritmi di classificazione, che hanno un ruolo cardine nella costruzione del modello, e il loro utilizzo per estrarre pattern e conoscenza dai dati. Inizialmente è stata adottata una strategia che prevede l'utilizzo di un solo classificatore, di tipo *Random Forest,* che è stato comparato con algoritmi come *Gradient Boosted Trees* e *Logistic Regression,* mentre, successivamente si è passati all'utilizzo di modelli ensamble di tipo stacking. L'ottimizzazione dei classificatori e dei relativi parametri (*Hyperparameter*[17]) è avvenuta tramite l'uso della tecnica di *Grid-Search Optimization* che prevede una ricerca esaustiva dei parametri migliori in un determinato sottoinsieme dello spazio dei possibili parametri.

5.2.3.1 Random Forest

L'algoritmo *Random Forest* è un metodo ensamble di tipo supervisionato derivante dalla teoria dei metodi Bagging ed è utilizzato in problemi di regressione e di classificazione. Il funzionamento alla base di questo algoritmo è l'utilizzo di una moltitudine di *weak learners*, nello specifico *Decision Tree,* che producono diversi alberi di decisione non correlati l'uno con l'altro, sviluppati in parallelo e senza alcuna dipendenza, utilizzando la tecnica del campionamento *Bootstrap* (approfondito nel paragrafo 3.3). L'algoritmo Random Forest è in grado di superare vari limiti imposti dal semplice uso di alberi delle decisioni da soli, risolvendo il problema molto comune di overfitting e ottenendo risultati di accuratezza maggiore. Una particolarità molto importante e caratterizzante del Random Forest è l'utilizzo di una ristretta selezione random di feature nel processo di apprendimento ad ogni possibile split, il che fornisce maggior casualità al modello in aggiunta a quella relativa al tipo di subset su cui è addestrato ogni albero.

L'algoritmo Random Forest si sviluppa nella sua fase di costruzione dei weak learners con i seguenti passi:

sia D la numerosità del dataset originale;
definito S il numero arbitrario di alberi decisionali utilizzati nel processo.
Per s=1 fino a s=S:

[17] Gli *Hyperparameter* sono quei parametri dei classificatori che caratterizzano il processo di apprendimento nel machine learning

1) *Estrazione di un sottoinsieme di dati D_s di numerosità N attraverso la tecnica Bootstrap dal dataset di partenza D;*
2) *Costruzione di un albero decisionale addestrato sul sottoinsieme di dati D_s attraverso i seguenti passaggi:*
    a) *Inizializzazione attraverso tutte le osservazioni disponibili in un singolo nodo;*
    b) *Per ogni nodo ripetizione ricorsiva dei seguenti passaggi fino al raggiungimento del criterio di stop:*
        i. *Selezione di k feature tra le p disponibili (con k<p);*
        ii. *Selezione della migliore feature per lo split del nodo;*
        iii. *Divisione del nodo in due nodi figli usando lo "split-point" selezionato nel punto ii.*

Alla fine del processo il modello combina attraverso un sistema di voti (nel caso della classificazione) i vari risultati ottenuti dai diversi Decision Tree producendo un singolo output.

5.2.3.2 Gradient Boosted Trees

L'algoritmo *Gradient Boosted Trees* è un metodo ensamble di tipo Boosting (vedi paragrafo 3.3) che utilizza come weak learner alberi decisionali, cosa che lo accomuna all'algoritmo Random Forest, ma a differenza di questo utilizza un approccio di tipo sequenziale. Partendo da un set di dati è costruito un albero decisionale di cui si calcolano gli errori di predizione, questo permette la definizione di un nuovo set di dati aggiustato tramite pesi da utilizzare nella costruzione del successivo albero decisionale; quindi si ripete il processo utilizzando volta per volta il dataset aggiustato derivante dal weak learner precedente, con l'obiettivo di correggere di volta in volta la predizione. Alla fine del processo si ottiene uno strong learner che apprende dai propri errori e gradualmente migliora le predizioni.

5.2.3.3 Logistic Regression

I modelli di regressione logistica, conosciuti anche come modelli *Logit,* fanno parte dei metodi statistici, come già illustrato nel paragrafo 3.3. La regressione logistica utilizza una funzione di tipo logistico, che assume valori compresi tra 0 e 1, per calcolare la probabilità di un evento dicotomico (ovvero una variabile nominale che può assumere due modalità) attraverso l'analisi delle relazioni tra le variabili indipendenti, che possono essere di tipo numerico o nominali, e la variabile di interesse dipendente. Questo classificatore è stato inserito nel lavoro di tesi per avere un metro di confronto tra algoritmi tradizionali e intelligenti.

5.2.3.4 Ensamble model: Stacking

Un modello ensamble di tipo stacking permette l'utilizzo di weak learners eterogenei tra loro nel processo di apprendimento. Il funzionamento nello specifico è descritto precedentemente nel paragrafo 3.3. I *base learner* utilizzati sono gli algoritmi Random Forest, Deep Learning e Gradient Boosted Trees, mentre il meta-model (*stacking model learner)* è un altro Random Forest; la struttura del modello è illustrata in Figura 15.

Gli algoritmi Random Forest e Gradient Boosted Trees sono stati presentati nei paragrafi precedenti, rispettivamente 5.2.3.1 e 5.2.3.2.

Gli algoritmi di tipo *Deep Learning*, tradotto letteralmente con apprendimento profondo, fanno parte di una sottocategoria del Machine Learning e basano il loro processo su più livelli di reti neurali artificiali (presentate nel paragrafo 3.3), nel nostro caso di tipo *feedforward*. Nel campo del Deep Learning si definisce feedforward una rete neurale artificiale caratterizzata da connessioni tra i nodi di tipo lineare; le informazioni si propagano dallo strato di input a quello di output in una sola direzione attraverso lo strato nascosto senza la creazione di loop. Il framework caratteristico del Deep Learning si compone di uno strato di input dei dati, una serie di strati nascosti, chiamati *Hidden Layer*, e infine uno strato di output. La particolarità di questi modelli è che ogni nodo elabora le informazioni che provengono dai nodi dello strato precedente e fornisce l'input per i nodi dello strato successivo.

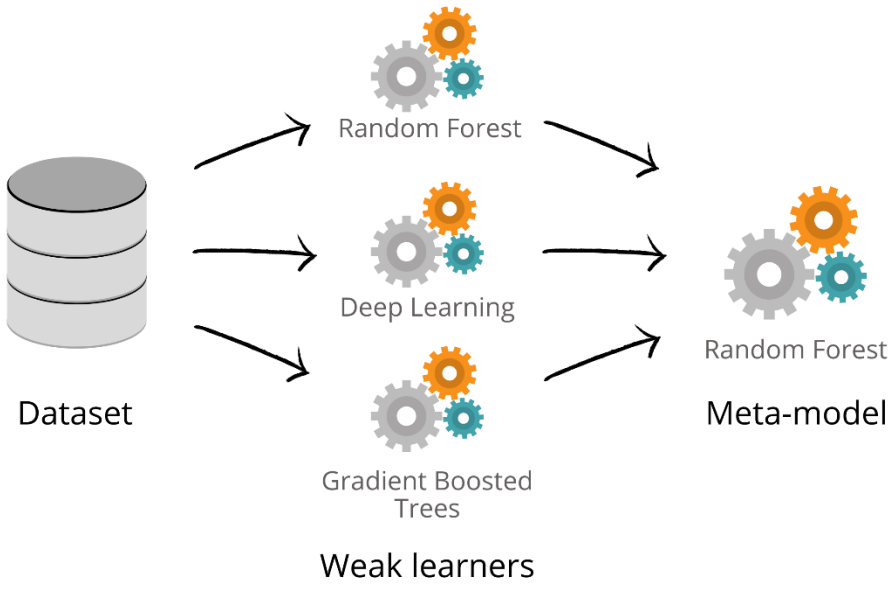

Figura 15 - Modello Stacking

5.2.4 Validazione del modello

La fase di validazione di un modello di classificazione è fondamentale per capire quanto bene il processo di predizione funzioni. Per fare questo, solitamente, si divide il dataset etichettato in due subset: il set di *training,* utilizzato nella fase di addestramento del modello e il set di *testing,* utilizzato nella fase di validazione. L'utilizzo di questo ultimo set di dati, composto da istanze mai utilizzate nella costruzione del modello in fase di data mining, è reso necessario per evitare fenomeni di *bias*, perché la valutazione del modello attraverso i dati di training perderebbe di significatività essendo il modello costruito proprio su quegli stessi dati. Quindi si procede attraverso la predizione della classe dei dati di testing per poi compararla con l'etichetta nota a priori; successivamente, attraverso il calcolo di diverse metriche, presentate nel prossimo paragrafo, si procede con una valutazione quantitativa della bontà della classificazione. La tecnica impiegata nella partizione del dataset in set di testing e di training è illustrata nell'ultimo paragrafo.

5.2.4.1 Metriche di performance

Le metriche di prestazione presenti nella teoria sono varie e ognuna di esse ha caratteristiche che la rendono più o meno adatta nella valutazione di un certo tipo di problema. Di seguito sono presentate le metriche utilizzate in questo lavoro di tesi.

La valutazione di un classificatore attraverso la misura dell'accuratezza (*Accuracy*) è molto utilizzata e si calcola attraverso la formula:

$$Accuracy = \frac{Numero\ di\ predizioni\ corrette}{Numero\ di\ predizioni\ totale}$$

Questa metrica ha il vantaggio di valutare globalmente con un'unica misura le performance di un modello e spesso è utilizzata con la metrica duale *l'Error Rate*:

$$Error\ Rate = 1 - \text{Accuracy} = \frac{Numero\ di\ predizioni\ errate}{Numero\ di\ predizioni\ totali}$$

Nei problemi di classificazione affetti da sbilanciamento dei dati, come nel nostro caso, queste metriche non performano bene e non risultano essere del tutto significative perché tendono a risultare ottime anche quando il classificatore attribuisce un peso maggiore alla classe maggioritaria. Ad esempio, se si considera un problema binario in cui la cardinalità della Classe A è pari a 990 e la cardinalità della Classe B è pari a 10, e se per assurdo si ipotizza che il

modello classifichi tutte le istanze con la Classe 0 otteniamo: $Accuracy = 990/(990 + 10) = 99\%$. In questo caso l'*Accuracy* risulta essere fuorviante perché il modello non rileva nemmeno un'istanza della Classe B.

Per ovviare a questo problema si utilizza la *Confusion Matrix,* una tabella che riporta la relazione tra la Classe Reale (*Actual Class*) delle istanze e la Classe Predetta (*Predicted Class*) dal modello. Da tale matrice si possono estrarre innumerevoli informazioni. La Tabella 5 riporta la matrice di confusione di un problema di tipo binario.

Tabella 5 - Confusion Matrix

		CLASSE REALE	
		Classe=1 Positiva	Classe=0 Negativa
CLASSE PREDETTA	Classe=1 Positiva	TP	FP
	Classe=0 Negativa	FN	TN

I quattro valori contenuti nella matrice di confusione rappresentano i seguenti record:

- *TP* (*True Positive*): numero di istanze positive (Classe=1) classificate correttamente come positive (Classe=1);
- *FP* (*False Positive*): numero di istanze negative (Classe=0) classificate erroneamente come positive (Classe=1) , chiamato anche *Errore di Tipo 1*;
- *TN* (*True Negative*): numero di istanze negative (Classe=0) classificate correttamente come negative (Classe=0);
- *FN* (*False Negative*): numero di istanze positive (Classe=1) classificate erroneamente come negative (Classe=0), chiamato anche *Errore di Tipo 2*.

L'accuratezza può essere calcolata anche attraverso i valori riportati nella confusion matrix:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Le metriche più indicate nella valutazione di un dataset sbilanciato sono misure specifiche di ogni classe, tra cui individuiamo: la *Recall*, la *Precision* e la *F-measure*.

La *Recall* è una metrica che indica la frazione di istanze correttamente assegnate alla classe generica C rispetto al numero di istanze totali appartenenti a quella stessa classe e si calcola come:

$$Recall = \frac{Numero\ di\ istanze\ correttamnete\ assegnate\ a\ C}{Numero\ di\ istanze\ appartenenti\ a\ C}$$

In breve, la *Recall* risponde alla domanda: "Quale è la percentuale di istanze della classe reale C che sono state effettivamente classificate correttamente?".

Nel caso binario distinguiamo tra TPR (*True Positive Rate*) e TNR (*True Negative Rate*) a seconda della classe su cui si opera; se la classe di interesse è quella positiva (Classe=1) si ha:

$$TPR = \frac{TP}{TP + FN} = \frac{TP}{Totale\ classe\ reale\ positiva}$$

Nel caso invece di classe negativa (Classe=0) si ha:

$$TNR = \frac{TN}{TN + FP} = \frac{TN}{Totale\ classe\ reale\ negativa}$$

Queste due misure indicano rispettivamente la frazione di istanze positive classificate correttamente come positive (TPR) e la frazione di istanze negative classificate correttamente come negative (TNR).

La Precision, invece, risponde alla domanda: "Quale è la percentuale di istanze classificate come classe C che effettivamente appartengono alla classe C?".

$$Precision = \frac{Numero\ di\ istanze\ correttamente\ assegnate\ a\ C}{Numero\ di\ istanze\ totali\ assegnate\ a\ C}$$

Nel caso binario e riferendoci alla classe positiva si ha:

$$Precision = \frac{TP}{TP + FP}$$

Per fare un esempio relativo alla classificazione del fallimento aziendale, possiamo identificare la *Recall* come la percentuale di aziende classificate come *Fallite* dal modello tra tutte quelle che lo sono nel dataset, mentre la Precision indica la percentuale di aziende effettivamente fallite nel dataset tra tutte le aziende classificate *Fallite* dal modello.

Un'altra metrica utile è la *F-measure*, nota anche come *F1 score*, che tiene conto del contributo delle due metriche viste precedentemente, la Recall e la Precision, facendone una media armonica. La formula è la seguente:

$$F - measure = 2 * \frac{Recall * Precision}{Recall + Precision}$$

Nel caso della classe positiva risulta essere:

$$F - measure = \frac{2 * TP}{2 * TP + FN + FP}$$

Questa misura è utilizzata quando è necessario avere un bilanciamento tra la Recall e la Precision.

L'ultima tecnica utilizzata in questo lavoro di tesi è la curva ROC, *Receiver Operating Characteristic Curve*, e la relativa AUC, *Area Under the ROC Curve*. La ROC è un grafico bidimensionale che rappresenta la capacità di predizione di un modello di classificazione a varie soglie (*thresholds*) di classificazione. Il grafico è caratterizzato sull'asse delle ordinate dal TPR True Positive Rate, illustrato in precedenza, e sull'asse delle ascisse dal FPR, ovvero il *False Positive Rate*, che è definito come $FPR = \frac{FP}{FP+TN}$ e indica quanti falsi positivi si verificano tra tutte le istanze di classe reale negativa. La curva ROC è dunque basata sulla sensibilità (TPR) e su *1 - specificità (FPR)* del classificatore.

Un esempio di curva ROC è illustrato in Figura 16.

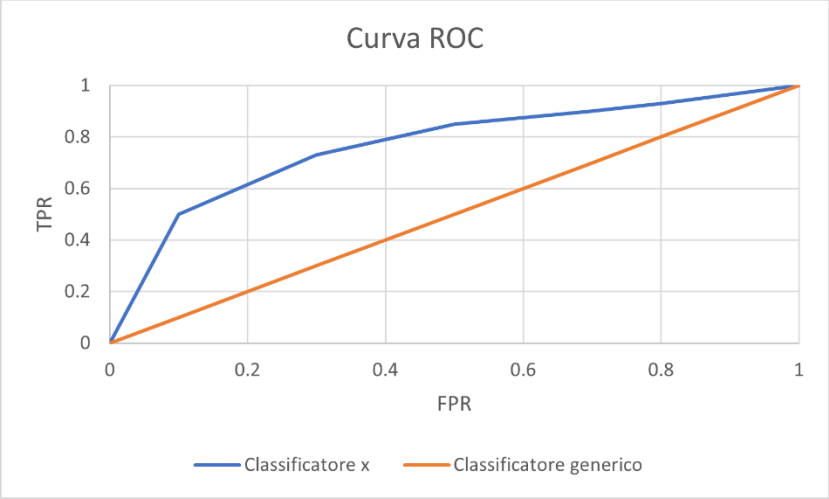

Figura 16 – Curva ROC

Un classificatore ideale ha TPR=1 e FPR=0 ovvero si posiziona nell'angolo in alto a sinistra del grafico. Altri punti rappresentativi del grafico sono: TPR=0 e FPR=0, che rappresenta un classificatore che non effettua mai una previsione di classe positiva, e TPR=1 e FPR=1, che rappresenta un classificatore che, al contrario, effettua sempre una predizione positiva. La diagonale TPR=FPR indica un classificatore che utilizza una strategia random (ovvero che etichetta come positiva metà delle istanze e negativa l'altra metà nel caso di problema binario). La curva ROC si costruisce valutando incrementalmente diverse soglie di probabilità. Una pratica comune è condensare le informazioni della curva ROC nell'AUC, che si calcola come l'area al di sotto della curva ROC. L'obiettivo è la massimizzazione di questa area, tenendo conto che un classificatore che segue una logica random in un problema binario ha l'AUC pari a 0.5 su un massimo ideale di 1.

5.2.4.2 Cross Validation

Il partizionamento del dataset disponibile in un training set, utilizzato nella fase di apprendimento, ed in un test set, da utilizzare in fase di valutazione, può essere fatto utilizzando diverse tecniche: le due prese in considerazione nel lavoro di tesi sono state l'*Holdout* e la *Cross Validation*.

La tecnica Holdout prevede la divisione del dataset in partizioni fisse, tipicamente 2/3 per il set di apprendimento e 1/3 per quello di testing, ed è appropriata per dataset con grandi quantità di istanze. Nella fase di partizionamento è importante mantenere i dati dei due set il più rappresentativi possibile ed è possibile farlo attraverso tecniche di campionamento stratificato.

La tecnica Cross Validation, che risulta essere la più usata, utilizza tutte le istanze sia come training set sia come test set. In particolare, la *k-fold Cross Validation* consiste nella creazione di un numero *k* di partizioni, o *fold*, della stessa grandezza e con *k* scelto arbitrariamente (solitamente *k*=10), da utilizzare in un processo di *k* iterazioni. In dettaglio, in ognuna delle iterazioni *k-1* partizioni sono utilizzate per addestrare il modello e la rimanente per testarlo; questo è ripetuto per tutte le fold. Da questo consegue che ogni partizione è usata *k-1* volte come training set e una volta come test set. Anche in questo caso, le partizioni sono create attraverso un campionamento di tipo stratificato per mantenere la significatività del dataset di partenza. Un esempio di *k-fold Cross Validation* con *k*=5 è illustrato nella Figura 17.

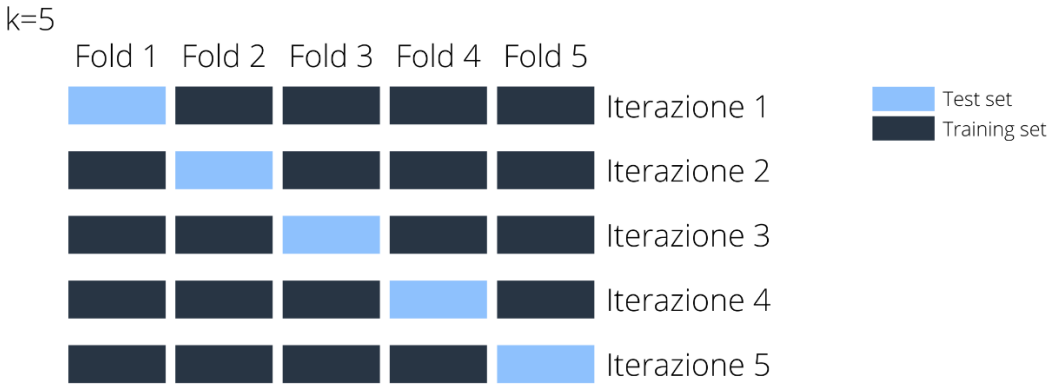

Figura 17 – Esempio di 5-fold Cross Validation

Questa tecnica è molto diffusa perché permette di minimizzare i rischi di overfitting evitando problemi di sovra-adattamento del modello in quanto il training set varia ad ogni iterazione.

5.3 Ambiente di sviluppo

L'ambiente di sviluppo utilizzato per lo svolgimento di questo lavoro di tesi è composto principalmente dal software *RapidMiner* con il supporto del pacchetto *Office* di *Microsoft*.

5.3.1 RapidMiner

RapidMiner (logo in Figura 18) è una piattaforma di *Data Science* nata nel 2006 che permette l'analisi di grandi quantità di dati attraverso strumenti che facilitano l'esplorazione e la preparazione dei dati, il data mining, analisi predittive con l'uso di tecniche machine learning e deep learning e la visualizzazione dei pattern. Il software permette dunque di svolgere tutte le attività tipiche del KDD.

Figura 18 - Logo RapidMiner (Fonte: https://rapidminer.com/)

RapidMiner fornisce un'interfaccia grafica (GUI, *Graphical User Interface*) attraverso la quale progettare e validare processi costruiti con l'utilizzo di operatori. Gli operatori hanno ognuno una funzione prestabilita e svolgono una determinata attività, dispongono di parametri da ottimizzare in base alle esigenze e possono essere collegati tra loro; in questo modo è possibile creare un flusso di lavoro che utilizza come input di un operatore l'output di quello

precedente. Una serie di operatori è direttamente disponibile in RapidMiner ed è possibile scaricare estensioni e operatori aggiuntivi di terze parti dal marketplace presente in rete o progettare e sviluppare operatori in autonomia attraverso la realizzazione di script. L'uso degli operatori nella costruzione dei processi permette di ridurre notevolmente i tempi di sviluppo riducendo anche la probabilità di errore dovuta alla quasi assente necessità di scrivere codice.

5.3.2 Processo RapidMiner

In questo paragrafo sono presentati i processi di preparazione dei dati e di addestramento e valutazione del modello. I processi sono illustrati presentando gli operatori RapidMiner utilizzati nelle varie pipeline. Gli operatori in grassetto, **nome operatore**, sono utilizzati in tutti i modelli, mentre quelli tra parentesi quadre, [nome operatore], sono i passaggi effettuati per le varianti del modello.

Inoltre, sono presentati gli operatori relativi agli algoritmi di classificazione usati nei vari modelli e quelli usati nella costruzione dei due processi di feature selection.

5.3.2.1 Preparazione dati

Il processo di preparazione dei dati è stato il primo passo nello svolgimento della ricerca sperimentale e ha permesso di ottenere il dataset definitivo usato nell'addestramento dei modelli nei vari test effettuati. In particolare, dopo aver importato i dati in RapidMiner sono stati selezionati gli attributi rilevanti e le feature di tipo numerico sono state normalizzate. Poi è stata settata come etichetta da predire (label) l'attributo di interesse "Fallita" e, infine, è stato effettuato un campionamento in ottica di riduzione della numerosità dei dati.

Di seguito è presentato nel dettaglio il processo RapidMiner e i relativi operatori:

Read Excel: questo operatore permette di leggere i dati da uno specifico file Excel, permettendo anche la definizione del tipo di attributi che vengono importati (integer, nominal, ecc.) attraverso una finestra d'importazione dedicata.

Select Attributes: questo operatore ha permesso la selezione degli attributi di interesse per l'analisi, presentati nel paragrafo 5.2.2.5, e che corrispondono alle feature X1 e da X19 a X55 (e X56 dove previsto).

[Filter Examples]: operatore impiegato per filtrare i dati rispetto all'attributo X56 "SezioneAz"; come illustrato nel paragrafo 5.2.2.1, si è scelto di effettuare un'analisi separatamente per il settore più significativo "Industria".

[Select Attributes (2)]: questo operatore è impiegato per l'eliminazione dell'attributo X56 in seguito al filtraggio delle istanze.

Normalize: l'operatore in questione permette la normalizzazione Z-Trasformation come definita nel paragrafo 5.2.2.2.

Set Role: questo operatore permette la definizione nel ruolo di *Label* dell'attributo X1 "Fallita". La Label indica al classificatore la variabile da predire, passo indispensabile per la fase di apprendimento.

Sample (Stratified): l'operatore permette di effettuare un campionamento stratificato in ottica di Data Reduction (vedi paragrafo 5.2.2.6). In particolare, si è deciso di proporre un campionamento di tipo relativo: 30% del dataset.

Store: questo operatore è impiegato nel processo di salvataggio del dataset derivante dai passaggi appena descritti, in un archivio dati (Data repository). Questo permette l'utilizzo dello stesso dataset nella costruzione dei vari modelli e ne facilità la comparabilità dei risultati.

In Figura 19 è presente il processo RapidMiner realizzato per la preparazione dei dati.

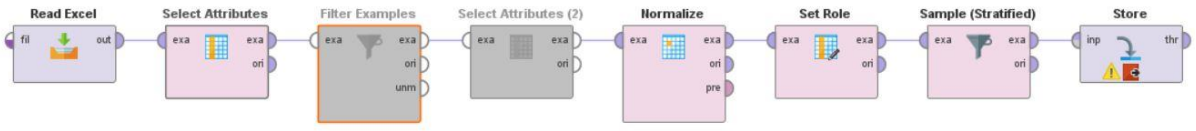

Figura 19 - Processo di preparazione dei dati

5.3.2.2 Processo di addestramento e valutazione modello

Il processo di addestramento del modello e la relativa valutazione risultano essere la parte centrale della ricerca sperimentale. In questa fase sono state testate le diverse configurazioni di preprocessing (trattamento dei missing value e applicazione della feature selection) ed è stata effettuata la comparazione dei diversi algoritmi di addestramento del modello al fine di individuare il classificatore migliore per il problema trattato. La valutazione dei modelli è stata

effettuata tramite il metodo Cross Validation e il calcolo delle metriche presentate nel paragrafo 5.2.4.1.

Il processo RapidMiner predisposto per realizzare questa fase è descritto di seguito:

Retrieve: l'operatore retrieve gestisce l'accesso al data repository permettendo il caricamento e l'utilizzazione dei dati processati come visto nel precedente paragrafo 5.3.2.3.

[Select Attributes]: questo operatore permette di intervenire sugli attributi andando a eliminare gli attributi con missing value maggiori del 15% (vedi paragrafo 5.2.2.4) e/o eliminare dall'analisi gli attributi con bassa importanza definiti dai processi di feature selection (vedi paragrafo 5.2.2.5)

[Replace Missing Values]: l'operatore in questione permette la gestione dei valori mancanti sostituendoli con la media dei dati, come descritto nel paragrafo 5.2.2.4. Questo è possibile settando il relativo parametro di default su: *Average.*

Cross Validation: si tratta di un operatore nidificato che si divide in due sottoprocessi: *fase di Training*, in cui è inserito il classificatore e costruito il modello, e *fase di Testing*, in cui il modello è applicato al testing set e si valutano le performance. Questo operatore permette di utilizzare la tecnica *k-fold Cross Validation* presentata nel paragrafo 5.2.4.2. Nel nostro caso i parametri sono settati su *k=10* (numero di fold) e su un campionamento di tipo *stratified_sampling* (che garantisce un fold stratificate e significative).

- *Fase di Training*

    - **Sample**: questo operatore permette di applicare la tecnica di Undersampling per la gestione dei dataset sbilanciati descritta nel paragrafo 5.2.2.3. Il settaggio dei parametri prevede:

      (a) sample = "relative"*;*

      (b) balance data = on*;*

      (c) sample ratio per class:

        - Classe 1 (Fallita) = *1*
        - Classe 0 (Attiva) = $\frac{(n° \; aziende \; della \; Classe \; 1)}{(n° \; aziende \; della \; Classe \; 0)}$

- [MetaCost]: questo operatore utilizza la tecnica cost sensitive per il bilanciamento dei dataset sbilanciati descritta nel paragrafo 5.2.2.3. I parametri principali che lo caratterizzano sono relativi alla costruzione della Matrice dei Costi.

    - **Operatori degli algoritmi di classificazione** (vedi prossimo paragrafo dedicato 5.3.2.3)

- *Fase di Testing*

    - **Apply Model:** questo operatore permette di applicare ai dati di testing il modello di apprendimento elaborato nella fase di training. L'obiettivo è la predizione della classe "Fallita" delle istanze del test set.

    - **Performance (Binominal Classification):** questo operatore permette di valutare con metriche statistiche la classificazione binaria dopo l'applicazione del modello sul test set etichettato. In particolare, l'output dell'operatore consiste nelle seguenti metriche: accuracy, classification error, AUC, F-measure e la Confusion Matrix comprendente anche recall e precision relativamente ad ognuna delle due classi.

Nelle figure seguenti è presentato il processo di addestramento e validazione del modello.

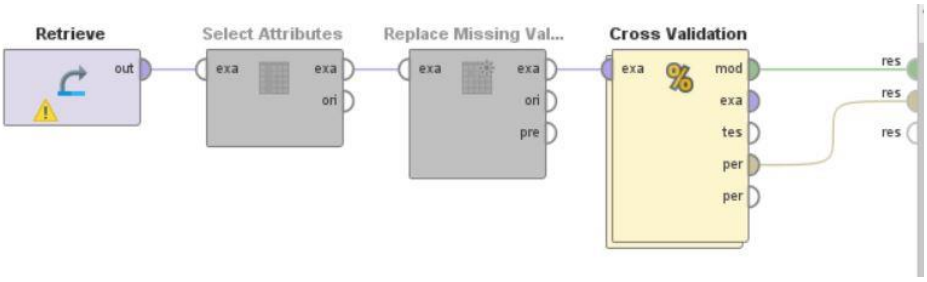

Figura 20 - Processo di addestramento e validazione modello (1/2)

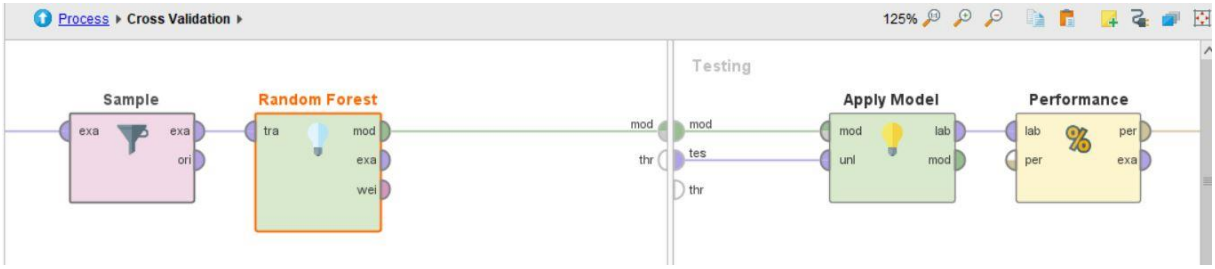

Figura 21 - Processo di addestramento e validazione modello (2/2)

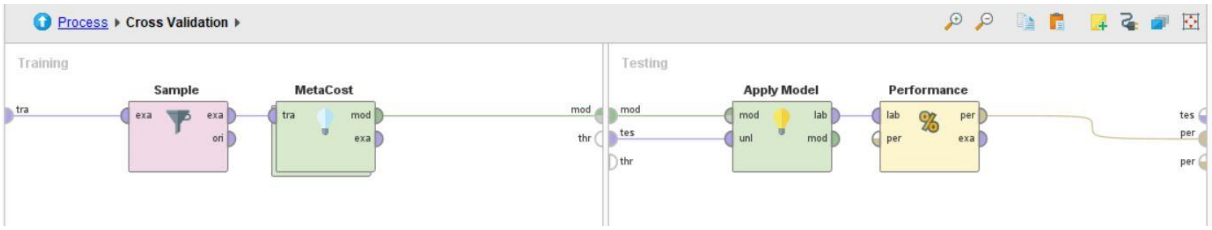

Figura 22 - Processo di addestramento e validazione modello (Caso tecnica Cost Sensitive)

5.3.2.3 Algoritmi di classificazione

A seconda del tipo di classificatore impiegato come learner nella fase di apprendimento automatico è stato utilizzato un differente operatore, ognuno caratterizzato dai seguenti parametri e settaggi.

Random Forest: questo operatore permette la costruzione di un modello Random Forest (vedi paragrafo 5.2.3.1) caratterizzato dagli *Hyperparameters*: numero di alberi, definito col parametro *number_of_trees* che nel nostro caso è pari a 200, criterio con il quale gli attributi sono selezionati ad ogni split, definito come *criterion* = information_gain, la profondità degli alberi che nel nostro caso è stato settato attraverso *maximal_depth* = 15 e la strategia di predizione dell'output finale attraverso voting_strategy = majority_vote.

Gradient Boosted Trees: questo operatore permette la costruzione di un modello di classificazione basato sugli alberi di decisione con le caratteristiche illustrate nel paragrafo 5.2.3.2. Tra i parametri più importanti per l'ottimizzazione troviamo: *number_of_trees* = 300, *maximal_depth* = 15 e *learning_rate* = 0.01 (ovvero la velocità con cui il classificatore svolge l'apprendimento; questo parametro può assumere valori da 0 a 1)

Logistic Regression: questo operatore implementa un modello di regressione logistica come descritto nel paragrafo 5.2.3.3.

Per la creazione del modello stacking, presentato nel paragrafo 5.2.3.4, è utilizzato l'operatore nidificato **Stacking** che è composto da due sottoprocessi. Nel primo operano in parallelo i base

learner (di tipo eterogeneo), nel secondo è utilizzato l'output del primo processo per addestrare un ulteriore classificatore chiamato *stacking model learner.* Per quanto riguarda il primo sottoprocesso sono utilizzati gli operatori **Random Forest** (*number_of_trees = 200, criterion* = information_gain, *maximal_depth* = 15, voting_strategy = majority_vote)e **Gradient Boosted Trees** (*number_of_trees* = 100, *maximal_depth* = 10 e *learning_rate* = 0.01), appena presentati, ed inoltre l'operatore **Deep Learning** caratterizzato dai seguenti parametri: *activation* = rectifier (scelta della funzione di attivazione) e *hidden layer size* = 60 (il numero e la grandezza di ogni strato di nodi).

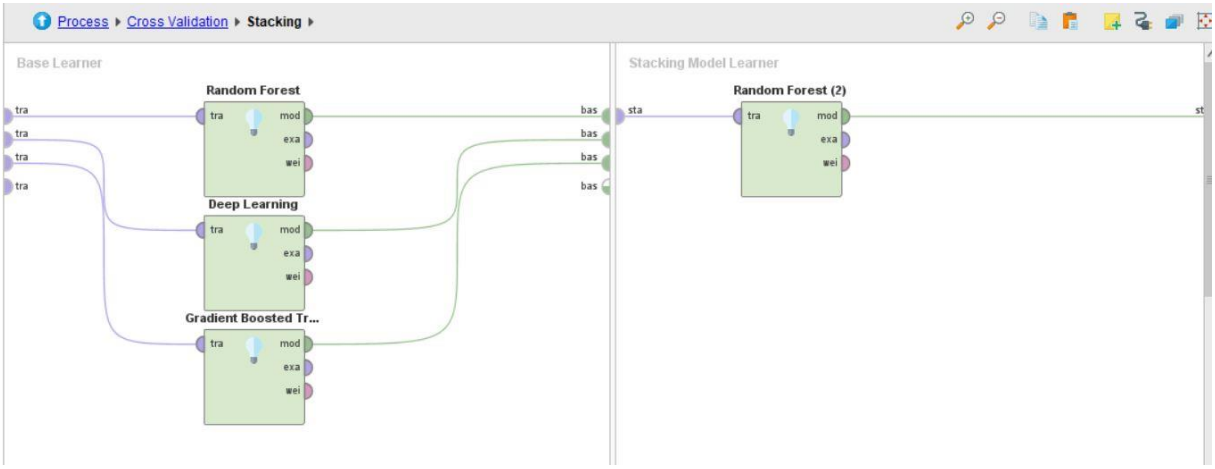

Figura 23 - Costruzione del modello stacking

5.3.2.4 Processo di Feature Selection

La Feature Selection è stata effettuata tramite due processi, presentati nel paragrafo 5.2.2.5.

Il primo utilizza l'operatore nidificato **Optimize Selection** che permette di individuare gli attributi rilevanti attraverso la tecnica di backward elimination con l'uso dell'operatore Cross Validation, illustrato precedentemente, che utilizza l'algoritmo Random Forest.

Il secondo metodo utilizza l'operatore nidificato **Optimize Selection (Evolutionary)** che applica algoritmi di tipo genetico ed evolutivi nella individuazione del subset di attributi più rappresentativi. Anche in questo caso è utilizzato l'operatore dell'operatore Cross Validation che utilizza l'algoritmo Random Forest.

Un esempio di processo di Feature Selection del tipo evolutivo è presentato nella Figura 24.

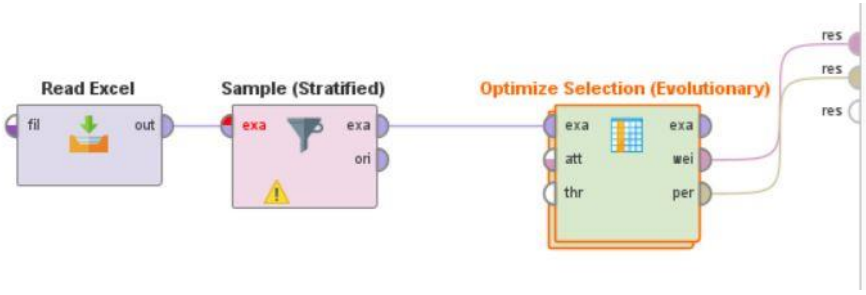

Figura 24 - Processo Feature Selection (algoritmi genetici)

6 *Risultati ottenuti*

In questo capitolo sono presentati i risultati sperimentali ottenuti utilizzando i modelli implementati secondo le tecniche e le metodologie descritte nel capitolo 5. Nel corso della ricerca sono stati sviluppati, tra modelli principali e loro varianti, più di cento processi RapidMiner, ma, per semplicità di lettura, nei seguenti paragrafi verranno illustrati solamente i risultati più rilevanti e significativi rispetto all'obiettivo del lavoro. I modelli presentati sono stati costruiti facendo riferimento alle aziende del settore Industria (come giustificato nel paragrafo 5.2.2.1), ma il comportamento dei modelli basati su tutti i settori non evidenzia differenze sostanziali.

6.1 Confronto tra algoritmi

Una prima analisi svolta riguarda la comparazione del risultato ottenuto applicando diversi tipi di algoritmi nella fase di data mining. Come descritto nel paragrafo 5.2.3, sono stati implementati quattro modelli che utilizzano quattro diversi algoritmi nella fase di apprendimento e classificazione: Random Forest, Logistic Regression, Gradient Boosted Trees e modello Ensamble Stacking. Le metriche impiegate nella valutazione dei modelli sono l'Accuracy, l'F-measure della classe positiva, l'AUC (derivate dalle curve ROC illustrate in Figura 25, 26, 27 e 28) in aggiunta alle informazioni contenute nella Confusion Matrix (TP, FP, TN, FN) che, per facilità di lettura, comprende anche i valori di Recall e di Precision per ogni classe; una descrizione completa delle metriche è presente nel paragrafo 5.2.4.1.

La Tabella 6 permette di comparare i vari modelli sviluppati nel lavoro di tesi attraverso le metriche Accuracy, F-measure e AUC.

Tabella 6 - Confronto tra algoritmi predittivi

Algoritmo	Accuracy	F-measure	AUC
Random Forest	74.40%	65.17%	0.835
Logistic Regression	68.13%	57.65%	0.753
Gradient Boosted Trees	73.36%	64.29%	0.827
Stacking Ensamble	75.66%	65.00%	0.809

Risulta evidente che gli algoritmi di Machine Learning (Random Forest, Gradient Boosted Trees e il modello Stacking Ensamble) forniscono migliori performance rispetto al modello statistico tradizionale (Logistic Regression). Infatti, tutte e tre le metriche risultano decisamente

peggiori nel caso della regressione logistica. Questo risultato conferma quanto previsto nel paragrafo 3.3 ed è avvalorato da numerosi studi che sostengono che i modelli di tipo AI sfruttano metodi di apprendimento dai dati migliori, avendo al contempo vincoli e prerequisiti molto meno marcati e stringenti, riuscendo così a performare in modo migliore.

Per il confronto tra gli altri algoritmi è utile osservare anche le Confusion Matrix, presentate nelle seguenti tabelle di numerazione da 7 a 10, che risultano più adatte nella valutazione di problemi derivanti da sbilanciamento tra le classi (per completezza si riporta anche la Confusion matrix relativa all'algoritmo Logistic Regression).

Tabella 7 - Confusion Matrix Random Forest

Random Forest		CLASSE REALE		
		Classe=1 Positiva	Classe=0 Negativa	CLASS PRECISION
CLASSE PREDETTA	Classe=1 Positiva	3926	3085	56.00%
	Classe=0 Negativa	1111	8266	88.15%
	CLASS RECALL	77.94%	72.82%	

Tabella 8 - Confusion Matrix Logistic Regression

Logistic Regression		CLASSE REALE		
		Classe=1 Positiva	Classe=0 Negativa	CLASS PRECISION
CLASSE PREDETTA	Classe=1 Positiva	3556	3742	48.73%
	Classe=0 Negativa	1481	7609	83.71%
	CLASS RECALL	70.60%	67.03%	

Tabella 9 - Confusion Matrix Gradient Boosted Trees

Gradient Boosted Trees		CLASSE REALE		
		Classe=1 Positiva	Classe=0 Negativa	CLASS PRECISION
CLASSE PREDETTA	Classe=1 Positiva	3930	3259	54.67%
	Classe=0 Negativa	1107	8092	87.97%
	CLASS RECALL	78.02%	71.29%	

Tabella 10 - Confusion Matrix Stacking Ensamble

Stacking Ensamble		CLASSE REALE		
		Classe=1 Positiva	Classe=0 Negativa	CLASS PRECISION
CLASSE PREDETTA	Classe=1 Positiva	3704	2656	58.24%
	Classe=0 Negativa	1333	8695	86.71%
	CLASS RECALL	73.54%	76.60%	

Ricordando che la Classe 1 corrisponde alle aziende fallite, mentre, la Classe 0 a quelle attive, è possibile ipotizzare che, dovendo affrontare un problema di identificazione di aziende in difficoltà, e che dunque siano destinate al fallimento entro 3 anni, sia corretto attribuire un'importanza prioritaria alla metrica Recall della Classe 1 rispetto alle altre. Questo perché questa metrica raccoglie l'informazione "Quale percentuale di aziende realmente fallite sono state classificate come fallite dal modello?", ovvero ci dice quante aziende in crisi sono state intercettate dal modello, informazione che risulta essere di primaria importanza per raggiungere l'obiettivo del lavoro, vincendo nel trade-off con la Precision. La Precision è una metrica comunque rilevante perché indica la percentuale di aziende classificate come fallite che effettivamente nella realtà lo sono; in ottica di uso al supporto decisionale aziendale un valore troppo basso di questa metrica potrebbe far perdere la fiducia del manager nell'uso del modello.

Dall'analisi incrociata delle tabelle sopra presentate emerge che l'algoritmo Random Forest e quello Gradient Boosted Trees assumono comportamenti simili per quanto riguarda Recall, Precision e AUC, mentre il Random Forest presenta prestazioni di poco migliori se si comparano l'Accuracy e la F-measure. Il modello Stacking Ensamble, contrariamente a quanto aspettato, non ottiene risultati significativamente migliori rispetto agli altri modelli, infatti si comporta meglio solo se teniamo conto dell'Accuracy migliorata dell'1% e della Precision.

L'algoritmo Random Forest risulta essere il migliore dei quattro modelli per quanto riguarda la F-measure e l'AUC.

Nelle successive analisi sono stati utilizzati come base di rifermenti il modello Random Forest e Ensamble Stacking.

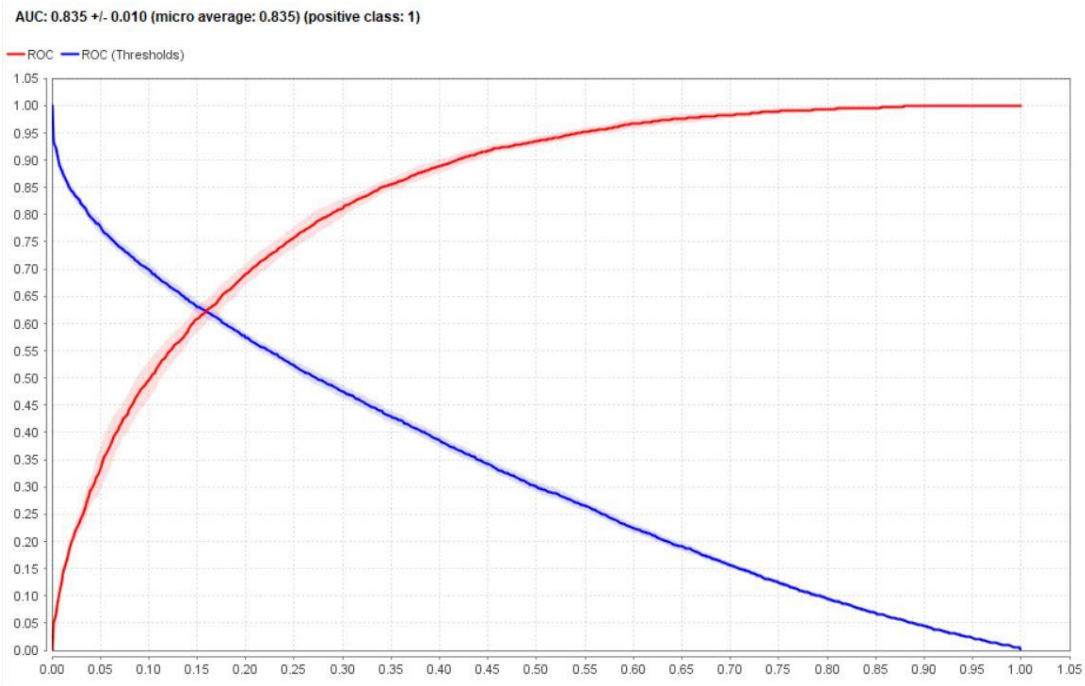

Figura 25 - ROC Random Forest

Figura 26 - ROC Logistic Regression

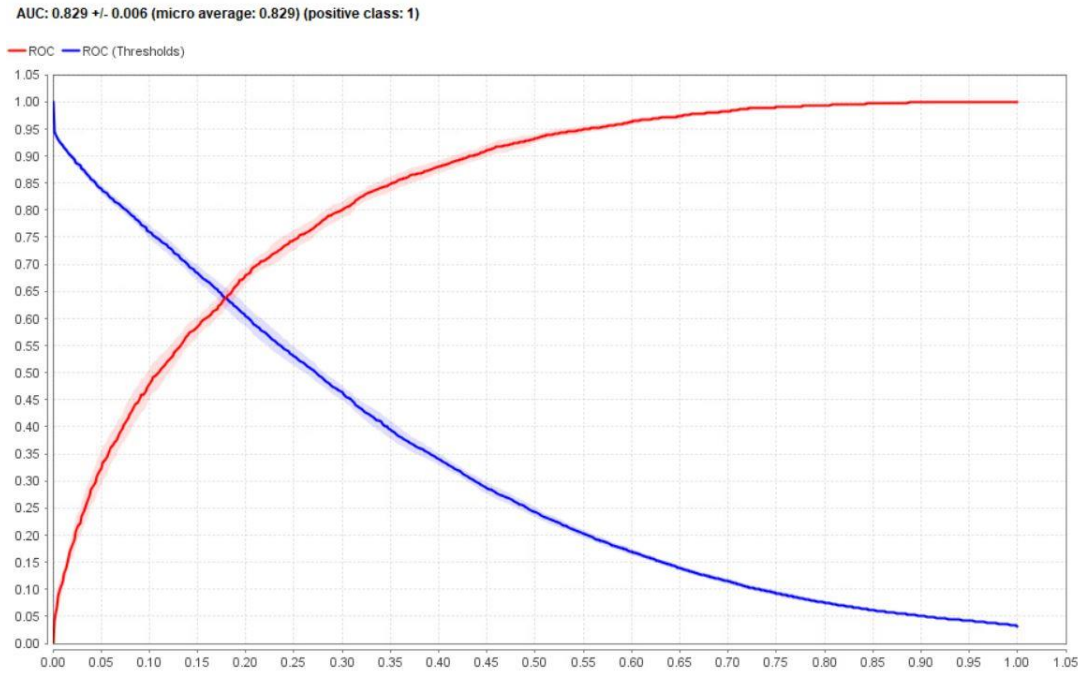

Figura 27 - ROC Gradient Boosted Trees

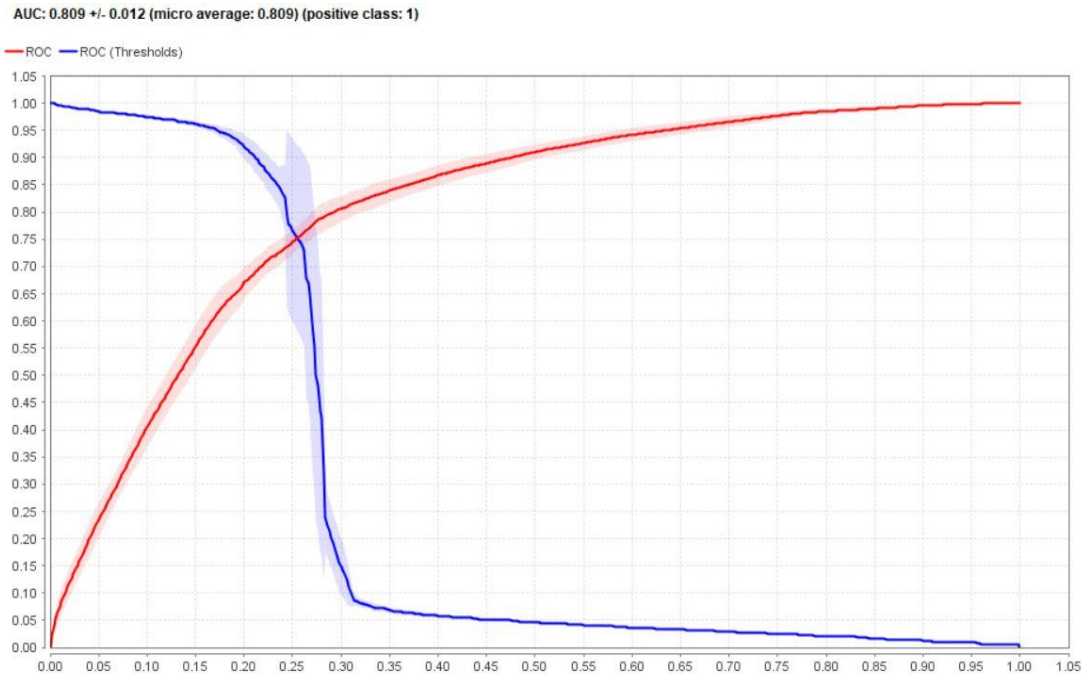

Figura 28 - ROC Stacking Ensamble Model

In ottica di *Explainable AI*, ovvero riguardo all'importanza di sviluppare modelli il più "spiegabili" possibile rendendo i processi di AI e Machine Learning meno assimilabili a meri processi "black box", risulta interessante capire quali siano gli attributi che forniscono un contributo più determinante nel processo di classificazione.

Attraverso l'uso dell'algoritmo Random Forest è stato possibile attribuire ad ogni attributo un peso corrispondente all'importanza nella fase di apprendimento del modello (peso complessivo pari a 1). In particolare, nella Figura 29 è presentato un grafico riepilogativo.

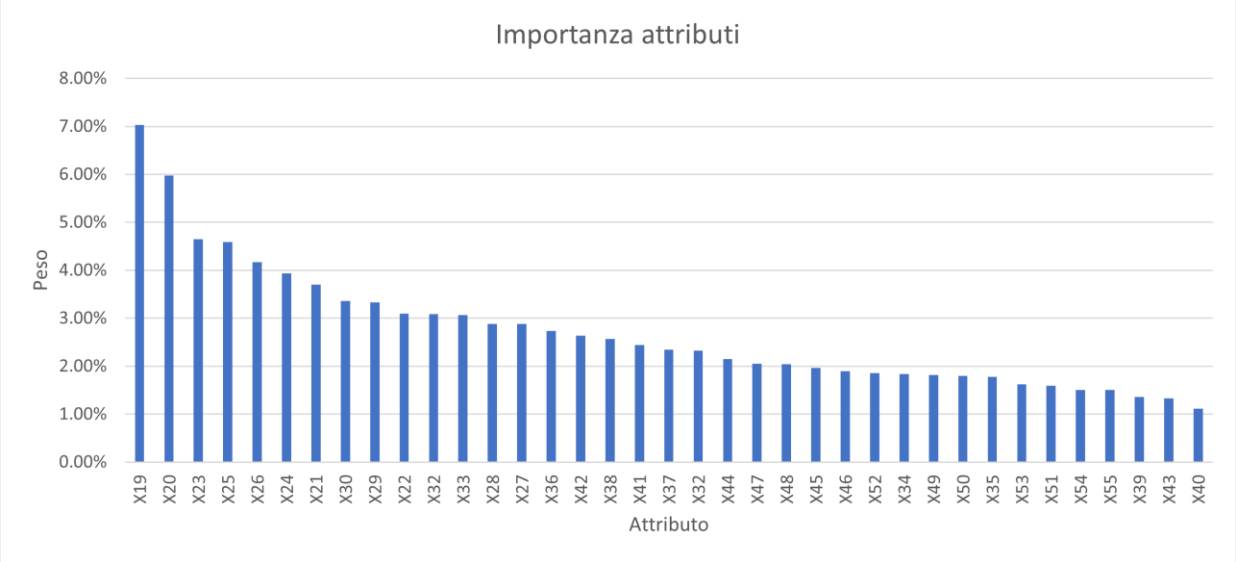

Figura 29 - Importanza degli attributi (Random Forest weight)

Come evidenziato dal grafico gli attributi più influenti sono l'X19 "Indice di liquidità" e l'X20 "Indice corrente" che fotografano la situazione di liquidità e le risorse monetarie in cui versa l'azienda. I meno importanti risultano essere invece indici della gestione corrente relativi a misure del magazzino come X39 "Giacenza media delle scorte", X43 "Durata ciclo commerciale" e X40 "Giorni di copertura scorte"; tra l'altro, questi ultimi tre attributi rientrano nella lista delle feature con una percentuale di missing value maggiore del 15%.

6.2 Trattamento dello sbilanciamento del dataset

Come già approfondito nel paragrafo 5.2.2.3, il ruolo del trattamento dello sbilanciamento del dataset è risultato fondamentale nella costruzione del modello.

Nella Tabella 11 è illustrata la comparazione delle metriche nel caso di applicazione della tecnica Undersampling, risultata la più efficace ed efficiente, nella fase di preprocessing e nel caso di non applicazione, entrambe implementate sia sul modello Random Forest che sul

modello Stacking Ensamble. Le metriche considerate nella comparazione sono la F-measure e la Recall relativa alle due classi; l'Accuracy non è presentata perché non è adatta alla valutazione di problemi sbilanciati (vedi paragrafo 5.2.4.1) mentre l'AUC risulta insensibile al trattamento dello sbilanciamento e pertanto rimane invariata.

Tabella 11 - Effetti Undersampling

Algoritmo	Undersampling	F-measure	Recall Classe 1	Recall Classe 0
Random Forest	Si	65.17%	77.94%	72.82%
Random Forest	No	58.98%	51.68%	89.53%
Stacking Ensamble	Si	65.00%	73.54%	76.60%
Stacking Ensamble	No	59.57%	52.15%	89.82%

Entrambi i modelli costruiti con i due algoritmi diversi risultano sensibili al problema del dataset sbilanciato nelle classi; in particolare, a risentirne di più è la Recall della Classe 1 che senza Undersampling diminuisce drasticamente (più di 26 punti percentuali in meno nel caso del modello Random Forest, passando da 77.94% a 51.68%). Il classificatore nel caso di sbilanciamento attribuisce un "peso" maggiore alla classe maggioritaria (Classe 0) nella fase di apprendimento, ciò si traduce in una tendenza all'aumento della predizione dell'etichetta Classe 0 come si evince dalla metrica Recall Classe 0 che cresce significativamente in entrambi i modelli.

Una seconda tecnica di bilanciamento dati è stata implementata attraverso un approccio Cost Sensitive che prevede la massimizzazione di una funzione di profitto attribuendo un ricavo alle istanze correttamente classificate e un costo alle predizioni errate attraverso la matrice dei costi. Il lavoro di tesi ha evidenziato come, nella ricerca sperimentale, questa tecnica conduca a risultati molto simili alla più semplice tecnica di Undersampling ma attraverso un uso di risorse computazionali significativamente più elevato e pertanto si è deciso di utilizzare l'Undersampling come tecnica di riferimento.

6.3 Importanza dei missing value

La problematica dei valori mancanti all'interno del dataset è stata presentata con un approfondimento nel paragrafo 5.2.1.

Nella fase di preprocessing si è optato per il trattamento dei missing value attraverso l'uso di tre strategie differenti, descritte più approfonditamente nel paragrafo 5.2.2.4, che diventano quattro, se si considera la loro applicazione in contemporanea.

Le evidenze sperimentali di questo lavoro di tesi, testate sul modello che utilizza l'algoritmo Random Forest, ci dicono che l'imputazione dei dati mancanti attraverso l'uso della media porta a risultati simili ai modelli in cui i missing value non sono trattati e l'analisi è svolta sui dati così come disponibili. La tecnica che prevede l'eliminazione degli attributi con una percentuale di valori mancanti superiore al 15%, invece, non risulta essere buona portando a una diminuzione sia di Accuracy del modello sia della F-measure della classe positiva dovuta ad un abbassamento della Recall della Classe 1 di circa 3 punti percentuali. Infine, la tecnica che utilizza contemporaneamente l'imputazione con la media e l'eliminazione degli attributi critici presenta l'Accuracy più bassa del confronto e le stesse problematiche riscontrate con la tecnica dell'eliminazione degli attributi.

Quanto appena descritto è chiaro osservando la Tabella 12.

Tabella 12 - Confronto tecniche di trattamento dei missing value

Tecnica gestione missing value	Accuracy	F-measure	AUC
1) Stima valori con media	74.46%	65.07%	0.830
2) Eliminaz. Attrib. >15% missing	74.01%	63.89%	0.823
1) + 2)	71.23%	64.43%	0.823
3) Nessun trattamento	74.40%	65.05%	0.836

Ai fini della ricerca sperimentale si è deciso di procedere con il trattamento dei missing value attraverso la sostituzione con lo stimatore media, evitando l'eliminazione degli otto attributi con missing value >15%, questo perché alcuni algoritmi, come il Deep Learning, usano nativamente come opzione di default la mean imputation per una miglior performance.

6.4 Feature selection

Le tecniche di feature selection utilizzate nella ricerca sperimentale sono state presentate nel paragrafo 5.2.2.5 e comprendono un processo di backward elimination e l'uso di algoritmi genetici.

L'applicazione di queste due tecniche ha portato a subset di attributi più importanti molto diversi sia per numerosità che per tipologia di attributo.

La tecnica di backward elimination è stata attuata attraverso l'algoritmo Random Forest e ha portato a scartare solamente tre attributi: X32, X39 e X46. In totale, oltre la variabile d'interesse X1, il modello lavora con 34 attributi rispetto ai 37 iniziali.

La tecnica che utilizza algoritmi evolutivi, sempre tramite modello Random Forest, ha invece eliminato 19 attributi, che equivalgono a più della metà degli attributi di partenza. Nello specifico il subset di feature selezionate è presentato in Tabella 13:

Tabella 13 - Attributi selezionati con la tecnica di Feature Selection evolutiva

ID	*Attributo (• tipo)*	*ID*	*Attributo (• tipo)*
X21 •	Indice di indebitam. a breve	X42 •	Durata media dei debiti al lordo IVA (gg)
X24 •	Grado di ammortamento	X44 •	EBITDA (k€)
X27 •	Debiti v/banche su fatt. (%)	X46 •	Redditività del totale attivo (ROA) (%)
X30 •	Oneri finanz. su fatt. (%)	X47 •	Redditività del capitale investito (ROI) (%)
X32 •	Grado di indip. da terzi	X48 •	Redditività delle vendite (ROS) (%)
X36 •	Rotaz. cap. investito (volte)	X50 •	Incid. oneri/Proventi extrag. (%)
X39 •	Giac. media delle scorte (gg)	X51 •	Capitale circolante netto (k€)
X40 •	Giorni copertura scorte (gg)	X52 •	Margine sui consumi (k€)
X41 •	Durata media dei crediti al lordo IVA (gg)	X53 •	Margine di tesoreria (k€)

•Indici fin. •Indici gestione corrente •Indici redditività •Altri dati significativi

Le performance ottenute applicando queste tecniche sono entrambe buone ma la più apprezzabile è la ultima feature selection presentata. Questo perché la riduzione del numero di feature risulta essere importante e permette di effettuare indirettamente anche *data reduction*, semplificando il modello e aumentandone la spiegabilità. Le performance delle tecniche a confronto sono disponibili nella Tabella 14.

Tabella 14 - Confronto Feature Selection

Feature Selection	**Accuracy**	**F-measure**	**AUC**
Tutti gli attributi	74.40%	65.17%	0.835
Back elimination	74.49%	65.27%	0.834
Algoritmi Genetici (GA)	74.60%	65.13%	0.832

Come si evince le performance dei modelli sono praticamente identiche e anche le metriche Recall Classe 1 e Precision Classe 1 differiscono al massimo dell'1% e comunque rimangono allineate al livello ottenuto nel modello base, che utilizza tutti gli attributi, come indicato dalla F-measure.

La scelta sperimentale ricade quindi sulla tecnica evolutiva che permette di snellire il modello utilizzando solo gli attributi con il maggior contributo informativo.

7 Conclusioni e Sviluppi futuri

In questo lavoro di tesi si è affrontato il tema del fallimento aziendale e della possibilità di fornire strumenti atti all'individuazione della crisi aziendale e alla gestione del rischio di fallimento. Il percorso seguito nello studio comprende tre filoni principali.

Il primo consiste nello studio della teoria relativa ai rischi aziendali e ai processi di risk management, andandoli a definire nell'ambito aziendale. Questo ha permesso di individuare e comprendere le dinamiche e le metodologie di gestione dei rischi che risultano utili ad affrontare il problema presentato nella tesi.

Il filone centrale ha toccato nello specifico il rischio di fallimento e di crisi aziendale oltre che presentare un approccio intelligente, attraverso tecniche di Artificial Intelligence e Machine Learning, nella costruzione di strumenti di allerta e predizione del fallimento. In particolare, è stato svolto un lavoro di ricerca relativamente ai principali studi presenti in letteratura riguardanti modelli di predizione del fallimento, su un intervallo di tempo dal 1930 ad oggi, lavoro che ha permesso di individuare e presentare le principali tecniche e modelli finora utilizzati.

Il terzo ed ultimo filone verte su una ricerca sperimentale in cui si è cercato di costruire un modello AI di classificazione per la predizione del fallimento a 36 mesi. Il modello è stato costruito attraverso dati di aziende italiane estratti da AIDA, dati che sono stati preprocessati e preparati alla fase di data mining con l'obiettivo di ottenere conoscenza utile nella individuazione di crisi aziendali premonitrici del fallimento. In questa fase sono stati utilizzati differenti algoritmi e tecniche di preprocessing che sono state testate e comparate tramite metriche di performance. Il modello che ha ottenuto risultati migliori e ulteriormente migliorabili è quello implementato con l'algoritmo Random Forest.

Alcuni possibili sviluppi futuri di questo lavoro di tesi sono presentati di seguito:

- come detto nel paragrafo 5.2.1, si è in presenza di molti dati mancanti all'interno del dataset utilizzato per cui sarebbe interessante attingere dati anche da altre fonti in modo tale da ottenere set di dati più completi. Inoltre, andando ad allargare la ricerca anche ad altri tipi di dati, si potrebbe implementare una imputazione dei dati mancanti

attraverso algoritmi più completi (ad esempio stime con l'utilizzo dell'algoritmo KNN) che, però, richiedono uno sforzo computazionale notevolmente maggiore;

- è possibile estendere i modelli addestrandoli su serie temporali piuttosto che, come nel caso della tesi, limitatamente ad uno specifico anno. Questo potrebbe permettere di intercettare trend ed allungare il periodo di predizione, fornendo uno strumento più utile ancora al management aziendale e alle istituzioni;
- in questo lavoro di tesi ci si è limitati a costruire un modello predittivo del fallimento, il passo successivo per rendere facilmente utilizzabile la conoscenza estratta è l'implementazione di un vero e proprio sistema DSS a supporto del decision-maker attraverso la generazione di report e l'uso di applicativi dedicati (software, app, ecc.).

BIBLIOGRAFIA

Alam, P., Booth, D., Lee, K., & Thordarson, T. (2000). The use of fuzzy clustering algorithm and self-organizing neural networks for identifying potentially failing banks: an experimental study. Expert Systems with Applications, 18(3), 185-199.

Altman E. I., & Hotchkiss E. (2006). Corporate financial distress and bankruptcy. Wiley.

Altman, E. I. (1968). FINANCIAL RATIOS, DISCRIMINANT ANALYSIS AND THE PREDICTION OF CORPORATE BANKRUPTCY. The Journal of Finance (New York), 23(4), 589-609.

Altman, E. I., & Narayanan, P. (1997). An International Survey of Business Failure Classification Models. Financial Markets, Institutions & Instruments, 6(2), 1-57.

Altman, E. I., Haldeman, R. G., & Narayanan, P. (1977). ZETATM analysis A new model to identify bankruptcy risk of corporations. Journal of banking & finance, 1(1), 29-54.

Altman, E. I., Marco, G., & Varetto, F. (1994). Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). Journal of banking & finance, 18(3), 505-529.

Balcaen, S., & Ooghe, H. (2006). 35 years of studies on business failure: An overview of the classic statistical methodologies and their related problems. The British Accounting Review, 38(1), 63-93.

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. Expert Systems with Applications, 83, 405-417.

Beaver, W. H. (1966). Financial ratios as predictors of failure. Journal of accounting research, 71-111.

Bellovary, J. L., Giacomino, D. E., & Akers, M. D. (2007). A review of bankruptcy prediction studies: 1930 to present. Journal of Financial education, 1-42.

Boden, M. A. (2018). Artificial Intelligence: A Very Short Introduction. Oxford University Press.

Boritz, J. E., & Kennedy, D. B. (1995). Effectiveness of neural network types for prediction of business failure. Expert Systems with Applications, 9(4), 503-512.

Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.

British Standard Institutions (BSI), 2021 – raccolta delle norme EN, ISO, BSI. **URL** https://www.bsigroup.com/en-ID/Standards/ (Ultima consultazione 20/09/2021)

Bureau van Dijk, 2021. AIDA – Base dati aziende italiane. **URL** http://aida.bdvinfo.com/ (Ultima consultazione 17/10/2021)

Carmona, P., Climent, F., & Momparler, A. (2019). Predicting failure in the US banking sector: An extreme gradient boosting approach. International Review of Economics & Finance, 61, 304-323.

Chen, J., Chollete, L., & Ray, R. (2010). Financial distress and idiosyncratic volatility: An empirical investigation. Journal of Financial Markets, 13(2), 249-267.

Clement, C. (2020). Machine Learning in Bankruptcy Prediction–a Review. Journal of Public Administration, Finance and Law, (17), 178-196.

Coats, P. K., & Fant, L. F. (1993). Recognizing financial distress patterns using a neural network tool. Financial management, 142-155.

Crovini, C. (2019). Risk management in small and medium enterprises. Routledge.

Floreani, A. (2004). Enterprise Risk Management: I rischi aziendali e il processo di risk management. EDUCatt-Ente per il diritto allo studio universitario dell'Università Cattolica.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence, 14(771-780), 1612.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Hauser, R. P., & Booth, D. (2011). Predicting bankruptcy with robust logistic regression. Journal of Data Science, 9(4), 565-584.

Hillegeist, S. A., Keating, E. K., Cram, D. P., & Lundstedt, K. G. (2004). Assessing the probability of bankruptcy. Review of accounting studies, 9(1), 5-34.

Hillson, D. (2009). How groups make risky decisions. Project Management Institute.

Hillson, D., & Murray-Webster, R. (2004, November). Understanding and managing risk attitude. In Proceedings of 7th Annual Risk Conference, held in London, UK (Vol. 26).

Hillson, D., Murray-Webster R. (2012). *Managing group risk attitude.* Gower Publishing, Ltd.

Hopkin, P. (2018). Fundamentals of risk management: understanding, evaluating and implementing effective risk management. Kogan Page Publishers.

Kantardzic, M. (2011). Data mining: concepts, models, methods, and algorithms. John Wiley & Sons.

Kim, S. Y., & Upneja, A. (2014). Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models. Economic Modelling, 36, 354-362.

Knight, F. H. (1921). *Risk, Uncertainty and Profit*. Houghton Mifflin Co.

Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. Expert Systems with Applications, 40(13), 5125-5131.

Kumar, P. R., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques–A review. European journal of operational research, 180(1), 1-28.

Lee, K. C., Han, I., & Kwon, Y. (1996). Hybrid neural network models for bankruptcy predictions. Decision Support Systems, 18(1), 63-72.

Lee, K., Booth, D., & Alam, P. (2005). A comparison of supervised and unsupervised neural networks in predicting bankruptcy of Korean firms. Expert Systems with Applications, 29(1), 1-16.

Martin, D. (1977). Early warning of bank failure: A logit regression approach. Journal of banking & finance, 1(3), 249-276.

Medium, 2021 - Towards data science **URL** https://towardsdatascience.com/ (Ultima consultazione 26/10/2021)

Moses, D., Liao S. (1987). *On developing models for failure prediction*. Journal of commercial bank Lending 69.7.

Mueller, J. P., & Massaron, L. (2021). Artificial intelligence for dummies. John Wiley & Sons.

Nanni, L., & Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. Expert systems with applications, 36(2), 3028-3033.

Neapolitan, R. E., & Jiang, X. (2018). Artificial intelligence: With an introduction to machine learning. CRC Press.

Odom, M. D., & Sharda, R. (1990, June). A neural network model for bankruptcy prediction. In 1990 IJCNN International Joint Conference on neural networks (pp. 163-168). IEEE.

Ohlson, J. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. Journal of Accounting Research, 18(1), 109-131.

Patrik P. (1932). *A comparison of ratios of successful industrial enterprises with those of failed firms*. Certified Public Accountant 2.

Perboli, G., & Arabnezhad, E. (2021). A Machine Learning-based DSS for mid and long-term company crisis prediction. Expert Systems with Applications, 174, 114758.

Platt H. D., Platt M. B. (2004). *Industry-relative ratios revisited: the case of financial distress*. Paper presented at the FMA 2004 Meeting, New Orleans (USA).

Platt, H. D., & Platt, M. B. (2002). Predicting corporate financial distress: reflections on choice-based sample bias. Journal of economics and finance, 26(2), 184-199.

Qu, Y., Quan, P., Lei, M., & Shi, Y. (2019). Review of bankruptcy prediction using machine learning and deep learning techniques. Procedia Computer Science, 162, 895-899.

Qu, Y., Quan, P., Lei, M., & Shi, Y. (2019). Review of bankruptcy prediction using machine learning and deep learning techniques. Procedia Computer Science, 162, 895-899.

RapidMiner, 2021 **URL** https://rapidminer.com/ (Ultima consultazione 2/11/2021)

Shi, Y., & Li, X. (2019). An overview of bankruptcy prediction models for corporate firms: A systematic literature review. Intangible Capital, 15(2), 114-127.

Shin, K. S., Lee, T. S., & Kim, H. J. (2005). An application of support vector machines in bankruptcy prediction model. Expert systems with applications, 28(1), 127-135.

Tam, K. Y., & Kiang, M. Y. (1992). Managerial applications of neural networks: the case of bank failure predictions. Management science, 38(7), 926-947.

Tamari, M. (1966). Financial ratios as a means of forecasting bankruptcy. Management International Review, 15-21.

Tan P., Steinbach M., Karpatne A., & Kumar V. (2019). *Introduction to Data Mining.* Pearson Education Limited*,* Seconda edizione.

Tsai, C. F., & Wu, J. W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. Expert systems with applications, 34(4), 2639-2649.

Turing A. M. (1950). *Computing machinery and intelligence. Mind*, Vol. LIX, Issue 236.

Udo, G. (1993). Neural network performance on the bankruptcy classification problem. Computers & industrial engineering, 25(1-4), 377-380.

UNI – Ente Italiano di Normazione, 2021 - raccolta completa delle norme UNI. **URL** https://www.uni.com/ (Ultima consultazione 24/09/2021)

Wilson, R. L., & Sharda, R. (1994). Bankruptcy prediction using neural networks. Decision support systems, 11(5), 545-557.

Zhao, Z., Xu, S., Kang, B. H., Kabir, M. M. J., Liu, Y., & Wasinger, R. (2015). Investigation and improvement of multi-layer perceptron neural networks for credit scoring. Expert Systems with Applications, 42(7), 3508-3516.

Zięba, M., Tomczak, S. K., & Tomczak, J. M. (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. Expert systems with applications, 58, 93-101.

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. Journal of Accounting research, 59-82.