POLITECNICO DI TORINO

Dipartimento di Ingegneria Gestionale e della Produzione

Corso di Laurea Magistrale
in Ingegneria Gestionale (L31)

Tesi di Laurea Magistrale

Credit Scoring mediante tecniche di machine learning

Relatori
Prof. Franco Varetto

Candidato
Emanuele Scoccia
(266589)

Dicembre 2021

Sommario

Indice delle figure

INTRODUZIONE

Questo lavoro di ricerca si occupa di applicare alcune tra le più diffuse tecniche di machine learning allo studio dei bilanci di un campione numeroso di aziende del settore del Legno-Arredo. L'obbiettivo è quello di predire efficacemente il fenomeno del fallimento o della liquidazione societaria, a partire dai principali indicatori spia dello stato di salute della società.

La scelta dell'argomento è legata al forte riscontro pratico che esso ha per gli intermediari finanziari, sempre più interessati al monitoraggio costante dello stato di salute dei crediti erogati e alla corretta scelta dei crediti da erogare. Questa necessità poggia innanzitutto sugli stringenti parametri imposti dalla regolamentazione bancaria, particolarmente severa in riferimento alla tematica del rischio di credito, e, ovviamente, sulla volontà di migliorare i risultati economici conseguiti dalla banca. È evidente che tanto più la banca sarà in grado di limitare la percentuale di crediti deteriorati, tanto più sarà una banca solida e non avrà necessita di nuovi apporti di capitale per ripianare le perdite da svalutazione crediti.

Il forte impatto che il credit scoring riveste all'interno del business bancario si accompagna all'introduzione nell'ultimo ventennio di nuovi approcci di studio sempre più sofisticati, che fanno ricorso a tecniche di machine learning e deep learning, poggiando le loro basi nell'ambito della cosiddetta "big data analytics" e dell'intelligenza artificiale, di cui ne costituiscono una delle applicazioni di punta. Queste tecniche si sono affiancate e stanno progressivamente sostituendo le tecniche tradizionali di analisi discriminante e regressione logistica e consentono di ottenere migliori risultati predittivi tanto più è grosso il campione di riferimento e il numero di caratteristiche analizzate.

Questo lavoro di tesi si concentra esclusivamente sugli algoritmi di machine learning, cercando di confrontarli tra loro, con l'obbiettivo di stabilire quali siano i più efficaci in relazione al campione preso in analisi.

L'elaborato si compone di quattro capitoli. Nel primo capitolo si espone brevemente cosa è e a cosa serve il credit scoring, utilizzato dagli intermediari finanziari per l'erogazione del credito, quando il numero di prestiti è molto elevato e si dispone di un'ampia base dati di fallimenti aziendali. Si introducono brevemente le tecniche tradizionali e quelle basate sul deep learning.

Nel secondo capitolo si fa un focus specifico sulle tecniche di machine learning, spiegandone le caratteristiche principali e l'ambito di applicazione. Queste tecniche possono essere racchiuse in due macroaree, l'apprendimento supervisionato e l'apprendimento non supervisionato, in relazione alla presenza o meno dell'etichetta di classificazione. L'etichetta consente di allenare l'algoritmo su un campione di prova, indicando quale deve essere l'output prodotto, cosicché a seguito dell'allenamento l'algoritmo sarà capace di generare l'etichetta corretta sul campione di test. Nell'apprendimento non supervisionato, essendo assente l'etichetta, si utilizzano approcci differenti, in cui l'intento è quello di creare dei cluster tra gli elementi del campione per cui sia minima la varianza intra-cluster e massima la varianza inter-cluster, di modo che l'elemento da classificare verrà inserito nel cluster con cui ha maggiore similarità. Nell'ultima sezione del capitolo, vengono introdotti brevemente gli algoritmi di classificazione delle serie temporali, in cui per ogni caratteristica si ha un vettore di valori tempo-dipendenti.

Nel terzo capitolo si conduce un'analisi sul settore di riferimento delle aziende del campione, evidenziandone l'andamento durante gli anni di pertinenza dei bilanci estratti. Si analizzano poi le caratteristiche del campione estratto (ripartizione territoriale, età media, tipologia societaria, ecc.) e si espone il pretrattamento e la trasformazione dei dati necessaria precedentemente all'implementazione degli algoritmi di machine learning.

Nell'ultimo capitolo, infine, si mostra l'esecuzione pratica dei vari algoritmi e i risultati ottenuti, confrontandoli tra loro. Particolare importanza viene data al confronto tra gli algoritmi classici e quelli basati sulle serie temporali. I primi fanno riferimento esclusivo al penultimo anno precedente al fallimento, i secondi fanno riferimento alla serie temporale composta dagli ultimi n anni precedenti al fallimento. Questo secondo approccio attualmente non è molto diffuso, quindi è interessante verificare quanto i suoi risultati possano divergere dall'approccio tradizionale.

1. CREDIT SCORING: TECNICHE IN USO

Le tecniche di credit scoring sono un insieme di tecniche adottate dalle banche e dagli intermediari finanziari per valutare le richieste di finanziamento della clientela (in genere per la concessione del credito). Esse si basano su sistemi automatizzati che prevedono l'applicazione di metodi o modelli statistici per valutare il rischio creditizio, e i cui risultati sono espressi in forma di giudizi sintetici, indicatori numerici o punteggi, diretti a fornire una rappresentazione, in termini predittivi o probabilistici, del profilo di rischio, affidabilità o puntualità nei pagamenti. Il credit scoring è una procedura che viene eseguita al momento dell'istruttoria, ovvero della presa in carico da parte di istituti bancari e società finanziarie della richiesta di finanziamento, prestito o mutuo presentata dal cliente. Il credit score equivale alla percentuale di rischio legata al finanziare un determinato soggetto. Per comprendere meglio il funzionamento del credit scoring, è necessario capire a fondo il concetto di rischio e, successivamente, di rischio di credito.

1.1. Rischio

Il concetto di rischio è alla base della teoria economica e della pratica finanziaria ed è strettamente collegato, anche se talvolta contrapposto, a quello di incertezza.

Secondo una visione prevalente, il rischio nasce quando, da situazioni incerte, derivano conseguenze negative in termini monetari o, più in generale, di utilità. In questa impostazione, il concetto di rischio e la sua misurazione debbono essere necessariamente inquadrati in uno specifico problema decisionale. In primo luogo, bisogna descrivere l'incertezza specifica del problema, che dipende dallo stato di informazione del decisore. In seguito, va tradotta l'incertezza in una misura analitica, tramite una funzione di costo (*loss function*) che associ un risultato monetario a ogni coppia di scelte del decisore e della natura.

Molto diverso, invece, è l'approccio che considera incertezza e rischio come due concetti nettamente contrapposti. Capostipite di questa impostazione fu F. Knight. Egli definì il rischio come associato a una situazione di incertezza oggettivamente probabilizzabile; ogni situazione non suscettibile di attribuzione oggettiva di probabilità sarebbe al contrario incertezza in senso proprio[1]. Nella sua opinione, le situazioni di rischio possono essere affrontate ricorrendo all'applicazione della legge dei grandi numeri, come

tipicamente viene fatto dalle compagnie assicurative. Queste ultime aggregano potenziali eventi avversi omogenei e indipendenti tra loro, trasformando un'incertezza a livello di singolo individuo in una certezza statistica a livello di collettività. In altri termini, trasformano una elevata perdita aleatoria, causata dal singolo evento avverso, in una piccola perdita certa, il premio assicurativo.

Le situazioni di incertezza, tipiche delle attività economiche, richiedono l'abilità del decisore (imprenditore). Consegue da questa impostazione che il profitto è la ricompensa per la capacità dell'imprenditore di gestire il (non assicurabile) rischio d'impresa, dove il capitale di rischio è appunto la parte delle risorse investite da un'impresa soggetta al rischio e il profitto è la parte del risultato dell'attività d'impresa che residua dopo aver compensato tutti gli altri fattori produttivi, incluso il capitale di terzi.

1.2. Rischio di credito

Fra le varie tipologie di rischio, questa tesi si concentra sul rischio di credito.

In letteratura, per rischio di credito s'intende "la possibilità che una variazione inattesa del merito creditizio di una controparte, nei confronti della quale esiste un'esposizione, generi una corrispondente variazione inattesa del valore di mercato della posizione creditoria".[2]

Esistono però diverse accezioni di rischio di credito, che distinguono l'eventualità in cui la perdita creditizia si manifesti solo in seguito all'insolvenza del debitore (default–mode paradigm), dal caso in cui la variazione del valore dell'esposizione derivi dal deterioramento del merito creditizio della controparte, trattando l'insolvenza come evento estremo (mark–to–market paradigm)

In particolare, si avranno allora:

- **Rischio di default**, ossia la possibilità che il debitore incorra in fallimento, non potendo ripagare, in tutto o in parte, il debito contratto. L'evento default è un evento dicotomico e, come tale, è modellato tramite distribuzione binomiale.
- **Rischio di migrazione**, ossia la possibilità che ci sia un deterioramento inatteso della qualità creditizia (downgrade), con conseguente aumento della probabilità di insolvenza della controparte. Tale deterioramento può trovare riscontro concreto in un declassamento del rating del debitore ad opera di un'agenzia o ad opera degli analisti fidi della banca creditrice. La conseguenza naturale è che il

valore di mercato del credito diminuirà, impattando sul prestatore nel caso in cui egli voglia cedere a terzi il credito in portafoglio (credit spread risk).

La migrazione del credito va modellata seguendo un modello multinomiale, in cui gli eventi si distribuiscono seguendo diverse probabilità che si manifesti l'evento estremo (default).

Le componenti del rischio di credito sono essenzialmente due: la perdita attesa (o Expected loss, EL) e la perdita inattesa (o Unexpected loss, UL).

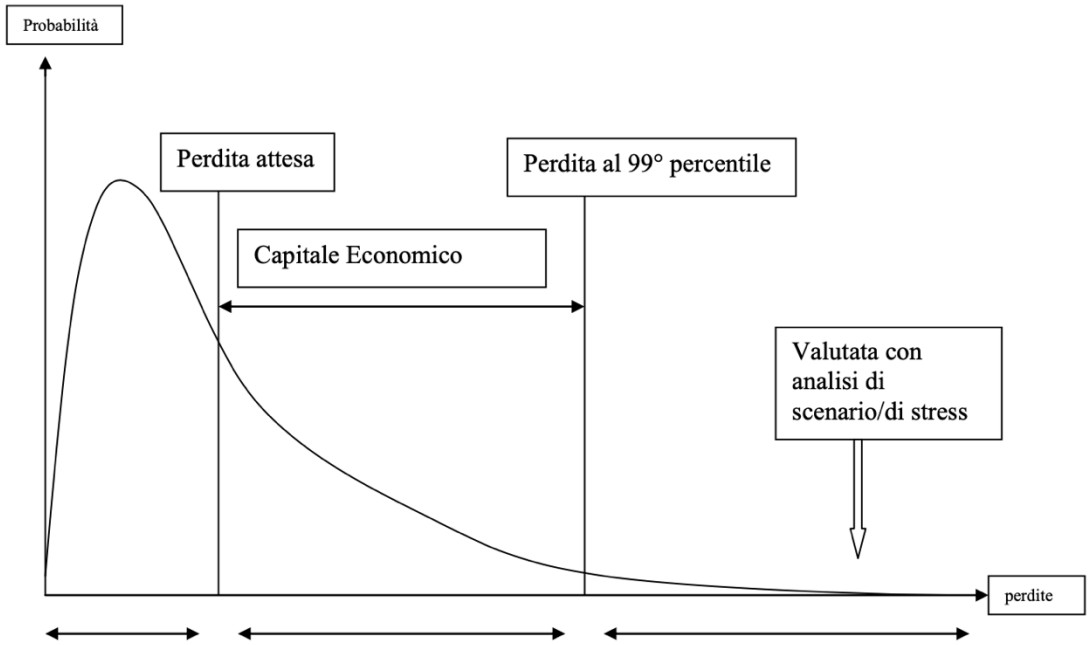

Figura 1:Distribuzione delle perdite nel rischio di credito

Per il prestatore assume rilevanza solo la componente inattesa del rischio di credito, ossia l'eventuale deterioramento non previsto della qualità del credito; questo perché le perdite attese sono già comprese negli accantonamenti prudenziali e nella determinazione del tasso d'interesse per i titoli di debito o per i prestiti, nell'ambito di quell'attività di pricing che deve riflettere in modo adeguato il profilo di rischio di un impiego. Proprio in quanto stimata a priori, la perdita attesa non costituisce il vero rischio di un'esposizione creditizia, ma si configura piuttosto come un elemento di costo per così dire "fisiologico", incorporato già nelle aspettative dell'investitore. In altri termini, essa consente di tener conto del rischio medio di insolvenza della controparte, che viene quantificato, nella determinazione del pricing, da uno spread che misura il premio rispetto ad un investimento privo di rischio.[3]

Se si considera come creditore tipicamente un'istituzione finanziaria (una banca), questa ha l'obbligo di costituire una riserva a fronte dei rischi assunti. (expected loss reserve).

Analiticamente, per perdita attesa s'intende il valor medio della perdita che una banca si attende di subire con riferimento ad un credito o portafoglio di crediti, in un certo arco temporale; mentre la perdita inattesa non è altro che il grado di volatilità del tasso di perdita intorno al proprio valore atteso.

Le due componenti del rischio di credito quindi, non solo rappresentano aspetti diversi della manifestazione delle perdite, ma hanno anche implicazioni diverse sulle politiche di bilancio della banca: nei riguardi della perdita attesa, la banca deve essere in grado di determinare il livello adeguato degli accantonamenti in conto economico; nei riguardi della perdita inattesa, la banca deve garantire un adeguato livello di patrimonializzazione dell'istituzione creditizia in grado di assorbire la perdita senza incorrere nel fallimento.

Normalmente la perdita attesa è espressa come funzione di tre elementi:

I. la probabilità di insolvenza del debitore (Probability of Default, PD)
II. la perdita in caso di insolvenza (Loss Given Default, LGD)
III. l'esposizione al momento dell'insolvenza (Exposure at Default, EAD)

$$EL = EAD * LGD * PD$$

I. EAD rappresenta il valore economico della perdita in caso di default. Se ci si riferisce ai mutui, essa si identifica con il debito residuo alla comparsa della sofferenza. Se ci si riferisce ai fidi di conto corrente, il calcolo è più complesso: si richiede di conoscere sia la quota di fido utilizzata (Drown Portion, DP), sia la quota non utilizzata (Undrown Portion, UP); quest'ultima assume importanza in quanto il debitore ha praticamente la facoltà di aumentare la sua esposizione in corrispondenza dell'insolvenza.

   Si inserisce quindi una terza variabile che prende il nome di Usage Given Default, UGD, che rappresenta la percentuale della quota inutilizzata che si ritiene venga utilizzata dal debitore in corrispondenza dell'insolvenza.

   Analiticamente avremo:

$$EAD = DP + UP * UGD$$

II. LGD rappresenta il tasso di perdita subito dai creditori in caso di insolvenza ed è complementare a RR, recovery rate o tasso di recupero.

$$RR = \frac{\sum_{t=1}^{n} \frac{ER_t - AC_t}{(1+i)^t}}{EAD}$$

Il recovery rate è il valore attuale delle somme recuperate nei vari tempi al netto dei costi amministrativi espresso in percentuale dell'EAD. I tassi di recupero dipendono da molti fattori (aleatori) e sono difficili da stimare, soprattutto per mancanza di dati storici sulle insolvenze, sui recuperi effettivi e sui tempi di recupero. Il tasso di attualizzazione usato può essere il tasso interno di trasferimento dei fondi (dalla filiale di raccolta alla filiale di impiego), il tasso originale del prestito, il tasso risk free o il tasso congruo per il rischio. Quest'ultimo sarebbe il valore più giusto da usare, ma in concreto è estremamente difficile da stimare.

L'entità di RR dipende poi anche da:

- Esistenza di garanzie (liquide o illiquide)
- Tipologia di contenzioso previsto per il recupero dei crediti insoluti
- Settore produttivo e specificità degli asset aziendali
- Paese o regione geografica e relativo sistema giudiziario
- Efficienza servizi legali interni all'azienda
- Livello dei tassi d'interesse
- Stato del ciclo economico

III. PD è la probabilità di insolvenza e discende a sua volta dal merito creditizio del debitore, cioè dalla sua capacità di reddito e quindi da fattori relativi alle condizioni economico-finanziarie, attuali e prospettiche, dell'impresa affidata e da altre variabili come la qualità del management o le prospettive di sviluppo del settore produttivo.

Il lavoro di tesi si concentra sullo sviluppo di algoritmi basati su tecniche di apprendimento supervisionato per la predizione più dettagliata possibile delle probabilità di default a partire dai bilanci delle aziende debitrici. Di seguito si espongono gli approcci più classici attualmente in uso.

1.3. Stima della probabilità di default

Gli intermediari finanziari hanno la necessità di elaborare dei modelli che consentano loro di clusterizzare i crediti in buoni e cattivi in termini di affidabilità. Questa operazione è molto complessa a causa dell'asimmetria informativa in essere tra il concedente (la banca) e il ricevente. Poiché i prestiti bancari sono molti e in maggioranza di piccola entità, per la gran parte la banca non ha abbastanza risorse per poter analizzare singolarmente ogni singolo credito erogato. Conseguentemente, i modelli di credit scoring sono nati con l'intento di trattare con un approccio statistico rigoroso ma automatizzato i crediti in aggregato. Questi modelli negli ultimi anni sono evoluti in approcci sempre più innovativi che si servono delle più moderne tecnologie di intelligenza artificiale, in particolare di machine e deep learning. Di seguito si introducono brevemente gli approcci più usati. Nel capitolo due si dedicherà poi un focus specifico sulle tecniche di machine learning classico, dando una base teorica agli approcci sperimentali sviluppati nei restanti due capitoli.

1.3.1. Modelli Univariati

L'approccio più elementare e incompleto che però si è rivelato fallace da un punto di vista teorico è l'approccio univariato. Esso esamina singolarmente i diversi indicatori (non combinandoli insieme in un'unica misura quantitativa), cercando di capire quali siano quelli maggiormente esplicativi dello stato di salute di un'impresa. L'approccio univariato fu introdotto nel 1966 da William H. Beaver nell'articolo "*Financial ratios as predictors of failure",* in cui lo studioso esaminò la capacità predittiva di alcuni indicatori, prendendo in considerazione un campione di 158 imprese (79 imprese anomale e 79 imprese sane). Beaver scelse una trentina di indicatori tra quelli più studiati in letteratura e ne analizzò la sovrapposizione delle distribuzioni calcolate separatamente sui due campioni, giungendo a determinare un punto ottimale di separazione, in grado di ridurre al minimo gli errori di attribuzione delle società ai due insiemi. Da questa ricerca, Beaver individuò nel rapporto Cash Flow/Debiti totali il miglior indicatore dello stato di salute delle imprese; esso, infatti, nell'anno precedente all'insolvenza, classificava correttamente le imprese nell'87% dei casi mentre la precisione dell'indicatore scendeva al 78% cinque anni prima del fallimento. Dallo studio emerse anche che gli indicatori con maggiore capacità predittiva erano quelli connessi alla struttura finanziaria e alla capacità di generare cassa, mentre i valori con minore potere esplicativo erano quelli legati al

circolante e alla liquidità. Il punto debole dello studio di Beaver era dato dal fatto che egli prendeva in considerazione gli indicatori singolarmente, anziché implementarli in un'unica misura in grado di riflettere la condizione globale di un'impresa[4]. Tale scopo non era facilmente raggiungibile, perché non bastava semplicemente considerare tutti gli effetti delle variabili, ma era necessario gestire i vari trade-off che si instauravano tra le varie componenti d'azienda. Ciò non sta comunque a significare che l'approccio univariato sia da considerarsi totalmente inadeguato, in quanto esso ha costituito la prima tappa per l'evoluzione in un modello più completo.

1.3.2. Analisi Discriminante

L'analisi discriminante è un tipo di analisi ad approccio multivariato introdotta per la prima volta da Fisher nel 1936. Essa ha come scopo la classificazione dei soggetti in diversi gruppi prestabiliti, utilizzando come discriminanti le loro principali caratteristiche. Pertanto, condizione indispensabile per avvalersi di questo metodo, è la disponibilità di osservazioni che siano già divise in gruppi (nel caso di analisi del merito di credito, tra unità solventi e non). [4]

Il modello si articola in una serie di passi:

I. A partire da una popolazione di n individui, dei quali si conosce l'appartenenza ad uno dei gruppi precostituiti, si selezionano, per ogni individuo, le caratteristiche ritenute responsabili dell'appartenenza ad uno degli specifici gruppi
II. Si costruisce una funzione (funzione discriminante) delle caratteristiche, con la quale si possa giungere alla determinazione del gruppo di cui l'individuo fa parte
III. Si procede alla classificazione di qualunque individuo di cui si conoscano le caratteristiche, grazie all'uso della funzione discriminante

In pratica, una volta che si hanno a disposizione tutte le caratteristiche necessarie, si procede alla configurazione di una relazione matematica capace di massimizzare le differenze tra i gruppi, minimizzando quelle infragruppo, diminuendo così la probabilità di errore. [4]

Il numero di funzioni discriminanti ottenibili varia in base al numero dei gruppi; in particolare esso è uguale a k-1, dove k è il numero dei gruppi (quindi in presenza del binomio solventi/insolventi la funzione discriminante è una). Un altro modo di leggere il

risultato del modello è quello che permette di individuare la probabilità di distanza di un soggetto rispetto alle caratteristiche medie di una data popolazione (per esempio, un nuovo soggetto può essere giudicato in base alla distanza da una popolazione di aziende insolventi, permettendo di quantificarne la qualità creditizia). [4]
Volendo classificare una serie di individui e supponendo, per semplicità, di avere due soli gruppi la relazione che si ottiene è del tipo:

$$Z = c_1 * X_1 + c_2 * X_2 + \ldots + c_n * X_n$$

Le variabili indipendenti Xi rappresentano le variabili descrittive, la variabile dipendente Z rappresenta il punteggio discriminante. I coefficienti attribuiti alle singole variabili considerate sono scelti da un algoritmo facendo in modo che i valori discriminanti ottenuti minimizzino la differenza tra le imprese dello stesso gruppo e massimizzino la differenza complessiva tra i due gruppi d'imprese. [4]

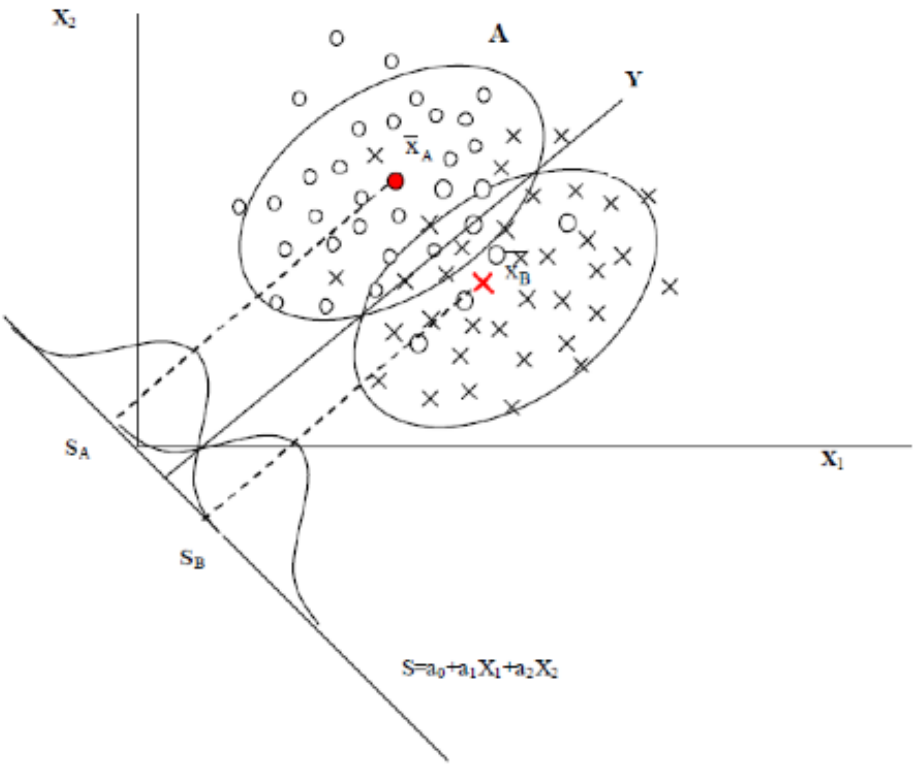

Figura 2:Individuazione dei cluster nell'Analisi Discriminante

Nel piano raffigurante i due campioni A e B, con le rispettive medie, si può notare la parziale sovrapposizione dei due insiemi; l'obiettivo dell'analisi discriminante è quello di individuare la funzione (la retta Y nel grafico) che permette di separare nel miglior modo possibile i due insiemi, commettendo cioè il minor numero di errori di attribuzione.

Un'importante proprietà di cui gode la retta Y è quella che, disegnata una retta perpendicolare ad essa, le proiezioni delle nuvole di punti su questa retta disegnano a loro volta due distribuzioni con la minor area di sovrapposizione. La retta S quindi costituisce la funzione discriminante lineare ottima. [4]

Dopo aver stabilito un punteggio soglia, sarà possibile dividere le osservazioni nei due gruppi stabiliti secondo il punteggio ottenuto da ciascuna osservazione.

Nella costruzione del modello, è necessario fare attenzione a non incorrere in errori; in particolar modo, poiché ogni errore ha un costo, l'obiettivo è di minimizzare congiuntamente probabilità di errore e costo associato ad ogni errore. All'aumentare dell'area di sovrapposizione tra le due distribuzioni cresce l'incertezza della classificazione; ciò significa che la situazione ottimale è rappresentata da due distribuzioni totalmente separate mentre la peggiore situazione è data nel caso di perfetta sovrapposizione, dove le caratteristiche considerate sono totalmente inutili nel compito di classificazione delle imprese. [4]

Inoltre, come ricordato poc'anzi, è necessario prendere in considerazione una stima dei costi di errata classificazione. Nello specifico, è possibile commettere due tipologie di errori:

- classificare come sana un'impresa insolvente (errore del I tipo)
- classificare come insolvente un'impresa sana (errore del II tipo)

La prima tipologia d'errore è la più grave in quanto implica la concessione di un fido ad un'impresa in dissesto finanziario con la conseguente perdita di interessi e capitale, mentre l'errore del II tipo conduce ad un mancato guadagno che si sarebbe potuto ottenere finanziando un'impresa sana. Di conseguenza, il valore soglia verrà scelto dall'analista in base al costo delle due tipologie di errore. Nella maggior parte dei casi, vista la differente importanza dei due tipi di errore, il valore soglia andrebbe aumentato allo scopo di rendere meno probabile un errore del I tipo rispetto ad un errore del II tipo. In generale però, lo scegliere il valore soglia può anche essere determinato dalla volontà che il costo associato a ciascun tipo di errore si eguagli, che tradotto in formula è:

$$P_I * C_I = P_{II} * C_{II}$$

Nel caso standard in cui $C_I > C_{II}$, il cut-off point dovrà essere fissato in modo che $P_I < P_{II}$.

Analizzando più in dettaglio lo Z-score, questo modello venne formulato da Altman sotto due ipotesi riguardanti le variabili discriminanti utilizzate:

1. esse sono caratterizzate da una distribuzione normale multivariata

2. sia quelle delle imprese insolventi che quelle delle imprese sane hanno uguali matrici di varianza-covarianza.

In realtà, la distribuzione normale, essendo una distribuzione illimitata non può rappresentare variabili come molti indici economico-finanziari strutturalmente limitati tra 0 e 100, con la conseguenza che la prima ipotesi sia irrealistica. C'è da precisare, però, che le ricerche hanno mostrato come questa lacuna influisca sull'efficacia previsionale del modello e non tanto sulla capacità discriminatoria dello stesso. [4]

Il modello fu sviluppato prendendo in considerazione 66 differenti società, trascurando quelle di piccola e media dimensione, di cui 33 fallite nel periodo 1945-1965 e 33 in buona salute, estratte casualmente dagli elenchi di Moody's e altre fonti. Fu, poi, compilata una lista di ventidue indici di bilancio più significativi, raggruppati in cinque classi: liquidità, solidità patrimoniale, redditività, rotazione ed efficienza della struttura operativa. Gli indici si ridussero poi a cinque in base al potere discriminante di ognuna delle variabili indipendenti, alle correlazioni delle variabili ed al personale giudizio dell'analista. In questo processo di snellimento si è utilizzata la seguente procedura: è stato osservato il rilievo di ciascun indicatore considerando i singoli contributi che apportava alle analisi, è stata osservata l'accuratezza previsionale dei vari profili e sono state valutate le correlazioni tra le variabili rilevanti. Poiché il processo è principalmente iterativo, non esiste un nucleo di indici stabile da utilizzare, per cui è l'analista che di volta in volta cerca di combinare quelli che a suo avviso sono i più rilevanti nella situazione specifica da valutare. In generale, la selezione delle variabili può avvenire seguendo due modalità: il metodo simultaneo o diretto e il metodo stepwise. Con il metodo diretto le variabili sono selezionate sulla base di un modello teorico mentre i coefficienti vengono stimati empiricamente; il metodo stepwise, invece, prevede l'utilizzo di variabili selezionate in relazione alla propria capacità discriminante. All'interno della seconda modalità, si ravvisano tre ulteriori tecniche:

- inclusione iniziale di tutte le variabili e successiva rimozione di quelle che hanno un potere discriminante minore (backward elimination)
- inclusione iniziale di un'unica variabile seguita dall'aggiunta di quelle che dimostrano maggiore potere discriminante addizionale rispetto a quello ottenuto con le variabili già presenti (forward selection);
- uso di una procedura a fasi alterne che ingloba le caratteristiche delle tecniche precedenti (stepwise selection).

In generale, nel metodo stepwise si procede con l'inserimento di una sola variabile alla volta, la quale viene tenuta nel modello qualora disponga di un valore discriminante considerato adeguato. Inoltre, quando si aggiunge un'ulteriore variabile, è necessario osservare attentamente anche il comportamento di tutte le altre, in modo da rimuovere quelle che non sono più caratterizzate da un sufficiente potere discriminante (ciò può accadere quando due variabili spiegano lo stesso effetto economico). [4]
La prima formulazione della funzione discriminante fu:

$$Z = 1{,}2 * X_1 + 1{,}4 * X_2 + 3{,}3 * X_3 + 0{,}6 * X_4 + 1{,}0 * X_5$$

Essa prevede l'utilizzo dei seguenti cinque indicatori:

X_1= capitale circolante netto / totale attivo
X_2= utili non distribuiti / totale attivo
X_3= EBIT / totale attivo
X_4= valore di mercato del patrimonio / valore contabile debiti
X_5= fatturato / totale attivo
La prima variabile discriminante è un indice che misura la percentuale dell'attivo circolante (calcolato come differenza tra attività correnti e passività correnti) rispetto al totale dell'attivo. Questo è un indicatore della liquidità dell'impresa poiché, di norma, un'impresa che sperimenta un periodo di crisi vede ridimensionarsi il suo capitale circolante netto rispetto al totale degli attivi.
La seconda variabile discriminante descrive che percentuale di attivo è rappresentata dal reinvestimento di utili non distribuiti. Tale rapporto identifica, in sostanza, un surplus di guadagno (tolti quindi i dividendi), cioè un valore creato dall'azienda che rimane al suo interno per ulteriori investimenti. Grazie a questo valore è possibile, inoltre, supporre l'età approssimativa di un'impresa, poiché è decisamente improbabile che un'azienda appena nata sia in grado di autofinanziarsi con delle riserve di utili non distribuiti. La realtà conferma che le più suscettibili crisi di insolvenza si presentano nei primi cinque anni di vita e dunque l'indice rappresenta un'ottima indicazione di solvibilità dell'impresa. Un'ultima considerazione desumibile da questo indice è il grado di leva finanziaria dell'impresa, in quanto un maggiore autofinanziamento è indice di un minor ricorso all'indebitamento.
La terza variabile rapporta l'EBIT (Earning Before Interest and Taxes) al totale dell'attivo aziendale. Questo indicatore è particolarmente utile perché prescinde dall'analisi

cumulativa dei tassi di interesse e dalle aliquote di imposta peculiari di ogni stato, permettendo di arrivare ad un concetto univoco di profitto non inficiato da valori che possono potenzialmente distorcere il giudizio.

La quarta variabile discriminante rapporta il valore del capitale proprio (derivante dalla somma del capitale conferito e delle riserve) al valore contabile dei debiti (sia a breve che a medio/lungo termine) e costituisce un indicatore di struttura finanziaria in cui il patrimonio netto è valutato a valori di mercato; questo comporta che, perciò se il mercato azionario è capace di rispecchiare in maniera corretta le prospettive dell'impresa allora i prezzi di borsa incorporano l'aspettativa dell'insolvenza permettendo al modello di Altman di catturare implicitamente anche tale previsione Il reciproco di questo indicatore è stato successivamente utilizzato da Fisher (1959) in uno studio sui differenziali di spread nei titoli corporate. Modelli più recenti, come il KMV, utilizzano come indicatore il valore di mercato del capitale e la sua volatilità.

La quinta variabile discriminante descrive il turnover del capitale generato dall'attività di vendita ed è una misura dell'abilità manageriale dell'impresa di operare in ambienti competitivi. Sebbene questo indice sia il meno importante su base individuale (in un'analisi univariata potrebbe persino non comparire), è invece estremamente significativo per la sua relazione con le altre variabili, tanto che è in seconda posizione nella gerarchia d'importanza del contributo che fornisce alla validità del modello.

Ai fini delle capacità classificatorie, le variabili maggiormente significative sono la redditività e l'efficienza complessiva, mentre la variabile meno rilevante è la liquidità.

La capacità diagnostica del modello di Altman con riferimento all'anno immediatamente precedente all'insolvenza è molto buona; infatti, il 95% delle imprese sono classificate in maniera corretta tenendo in considerazione le due tipologie di errori che si possono commettere: l'errore di I° tipo è del 6% e l'errore di II° tipo è del 3%.

Il potere predittivo del modello però cala notevolmente se vengono utilizzati parametri degli anni precedenti all'insolvenza, sia considerando il campione d'origine sia con riferimento a campioni di controllo (la capacità di previsione scende all'82% in riferimento ai due anni precedenti all'insolvenza). [4]

La progressiva perdita di efficacia del modello è plausibile, considerato che i segnali di crisi sono ovviamente più evidenti man mano che ci si avvicina all'insolvenza. Si può affermare che le difficoltà diagnostiche del modello sono da imputare all'attenuazione delle differenze tra i due gruppi, all'allontanarsi del fenomeno insolvenza.

In generale, la probabilità di fallimento è:

- alta, se il valore dello Z-Score è minore di 1,79
- medio alta, se il valore è tra 1,8 e 2,69
- media, se il valore è tra 2,7 e 2,99
- bassa, se il valore è maggiore di 3.

Invece di identificare un unico valore soglia, è più corretto ricorrere all'individuazione di un intervallo di cut-off caratterizzato da un estremo superiore e un estremo inferiore. Un'impresa che presenta uno score discriminante inferiore all'estremo inferiore è considerata un'impresa a rischio elevato e dunque inaffidabile. Viceversa, un'impresa che presenta uno score superiore all'estremo superiore è considerata un'impresa a rischio basso e dunque affidabile. Infine, se un'impresa presenta uno score compreso fra i due estremi non si è in grado di prevedere con precisione se essa appartiene al gruppo delle imprese sane o a quello delle imprese che diverranno insolventi, quindi il modello è incapace di discriminare con estrema precisione fra le due categorie di imprese. [4]

Il modello permette di prevedere l'insolvenza delle imprese fino a due anni prima del verificarsi di questa, con un errore medio di previsione del 15%, un anno prima del fallimento, e del 17%, due anni prima.

Successivamente, nel 1977, Altman ed altri autori perfezionarono un nuovo modello detto modello Zeta, frutto delle modifiche dovute alle critiche ricevute dal precedente Z Score. Il nuovo modello era costituito da 53 società fallite e 58 società sane e presentava alcune novità rispetto alla versione precedente:

- gli indicatori usati nel modello sono stati aggiustati allo scopo di renderli maggiormente espressivi della realtà aziendale (gli aggiustamenti più importanti riguardano le riserve, la capitalizzazione dei contratti di leasing operativo e finanziario, il consolidamento delle consociate finanziarie, le attività immateriali e gli avviamenti)
- l'analisi dell'importanza dei vari indicatori è stata effettuata utilizzando sei test diversi
- sono state definite delle modalità a priori e una stima dei costi di errata classificazione

Il modello Zeta non utilizza cinque variabili come lo Z-Score, bensì sette:

- *ROA*, inteso come rapporto tra utili ante interessi e attivo totale
- *Stabilità degli utili*, calcolata attraverso lo scarto quadratico medio della stima intorno alla tendenza decennale del ROA

- *Servizio del debito*, dato dal rapporto che presenta al numeratore l'utile ante interessi e tasse e al denominatore gli oneri finanziari totali
- *Redditività cumulata*, calcolata come rapporto tra riserve di utili e attivo netto
- *Liquidità corrente*
- *Capitalizzazione*, calcolata come rapporto tra valore di mercato del patrimonio netto (media dei prezzi delle azioni degli ultimi cinque anni) e valore totale del debito
- *Dimensione*, misurata dal logaritmo dell'attivo netto

1.3.3. Modelli di regressione

In precedenza, si è visto che l'analisi discriminante lineare ha come scopo quello di individuare la combinazione lineare che consente di separare due gruppi di imprese nel modo più efficace possibile. Questa tecnica ha dei punti in comune con il modello di regressione, infatti i coefficienti dell'analisi lineare sono pari a quelli della regressione con i minimi quadrati ordinari a meno di un rapporto costante. Uno degli esempi più usati è il linear probabilistic model, dove variabili e relativi pesi vengono individuate grazie ad una regressione lineare. Lo schema seguito dal modello si può riassumere in quattro fasi:

I. *Selezione del campione*: In questa fase viene selezionato un numero sufficientemente elevato d'imprese divise in due gruppi, imprese insolventi e imprese sane. È doveroso precisare che assume una notevole importanza il numero delle imprese insolventi, che seppur inferiore a quelle sane, dovrebbe essere il più elevato possibile in maniera tale che i risultati della regressione siano statisticamente significativi; paradossalmente, dunque, una banca che voglia usare una funzione probit o logit, è avvantaggiata se in passato ha concesso un fido ad un numero elevato di imprese che si sono rivelate insolventi[5]
II. *Selezione delle variabili indipendenti*: Si calcolano le variabili casuali in grado di riflettere le informazioni quantitative rilevanti per tutte le imprese (in genere indici economico-finanziari)
III. *Stima dei coefficienti:* La variabile Y, che può assumere alternativamente valore nullo o unitario, è la variabile dipendente mentre gli indici economico-finanziari sono le variabili indipendenti. In questa fase si procede alla stima dei relativi coefficienti di ponderazione.

IV. *Stima della probabilità d'insolvenza***:** I risultati ottenuti grazie al modello vengono utilizzati per stimare la probabilità d'insolvenza di una nuova impresa.

Tale approccio, però, presenta una varianza dei residui che non è costante ma risente di un problema di eteroschedasticità (la varianza dei residui delle varie osservazioni non assume lo stesso valore). Questo problema implica l'assenza di omoschedasticità (ipotesi base della regressione lineare), portando a stime imprecise e distorte. [5]

Inoltre, un ulteriore difetto è dato dal fatto che questo approccio può generare valori non compresi tra 0 e 1, come sarebbe logico aspettarsi: valori stimati negativi o di molto superiori a uno creano errori di stima crescenti, a mano a mano che ci si allontana dall'intervallo [0;1]. [5]

Allo scopo di ottenere valori che appartengono tutti all'intervallo [0;1], è possibile utilizzare il modello logit. Tale modello non è l'unico in grado di produrre questi valori limitati, ma le sue caratteristiche lo rendono più facilmente manipolabile e quindi maggiormente usato dagli studiosi.

L'idea sottostante a questo modello è di supporre che esista una relazione tra la probabilità di un'impresa di diventare insolvente e le variabili esplicative osservabili strettamente connesse con l'evento insolvenza. Ciò significa stimare un modello usando gli indicatori di bilancio come variabili indipendenti e una variabile dicotomica Y come variabile dipendente, che assume il valore 0 se l'impresa è sana mentre assume il valore 1 se l'impresa è insolvente.

Mentre nell'analisi discriminante si ipotizza che le imprese appartengano a due universi distinti e che la rilevazione delle caratteristiche possa fornire un aiuto per determinare l'universo da cui provengono, il modello logistico (e quello probit) ipotizza che le imprese siano selezionate casualmente da un unico universo e cerca di stimare una caratteristica specifica di tali imprese (la probabilità d'insolvenza), immaginata come una variabile latente continua e della quale sono osservabili solamente le due determinazioni estreme 0 e 1. [5]

Il modello ipotizza l'esistenza di una relazione causale tra le variabili osservate e la variabile dipendente, evidenziando una relazione di causa-effetto tra i risultati economici riflessi dalle variabili indipendenti e lo stato di salute dell'impresa. In questo caso non vi è più l'intento di stimare l'appartenenza ad un gruppo, ma c'è l'intento di misurare il grado di difficoltà dell'impresa.

Un modello del tutto similare a quello logistico è il modello probit. Esso si differenzia dal modello logit poiché la forma della funzione di ripartizione cumulata della distribuzione non è la logistica, bensì la normale standardizzata. [6]
Si può effettuare un test basato sul rapporto di massima verosimiglianza, analogo al test F nella regressione lineare, sottoponendo al test l'ipotesi nulla che tutti i parametri, esclusa l'intercetta, siano nulli. I risultati di questo test, uniti al p-value, indicheranno l'efficacia del modello
Gli indicatori di bilancio generalmente utilizzati sono:

- Rapporto cash-flow su debito totale
- Rapporto reddito netto su attività totali
- Rapporto attività correnti su passività correnti
- Rapporto attività correnti su vendite nette

La probabilità di fallimento è inversamente proporzionale ai rapporti appena descritti, dunque all'aumentare dei valori che assumono questi indicatori di bilancio, la probabilità che l'impresa fallisca si riduce.
Apparentemente il modello sembra molto semplice ma ci possono essere
spiacevoli inconvenienti nel caso in cui la probabilità d'insolvenza stimata assuma valori esterni all'intevallo [0;1].

1.3.4. Reti neurali

Le reti neurali sono un insieme di tecniche innovative che usano un approccio completamente diverso da quello usato con le tecniche di analisi discriminante o regressione. Questo approccio è detto "a scatola nera", in quanto sta a significare che le variabili in input, che entrano nella scatola nera, scaturiscono nelle variabili in output, in uscita dalla scatola nera, senza che dall'esterno si possa stabilire i collegamenti, meglio ancora le relazioni, intercorse tra le variabili in input.[7] In altre parole, non è possibile fare un'analisi a posteriori in grado di comprendere perché a partire da certi input, si generino certi output.
Lo scopo delle reti neurali artificiali è di riprodurre all'interno del sistema informatico la struttura dei sistemi nervosi biologici, costituiti da una moltitudine di neuroni connessi in rete; l'idea sottostante è di usufruire del meccanismo di apprendimento che caratterizza la memoria e la conoscenza dell'uomo.

I primordi di questa disciplina si riscontrano a partire dal secondo dopoguerra, negli studi di J. Von Neumann e D.Hebb. Negli anni Sessanta vengono costruite le prime macchine in grado di presentare forme di apprendimento, come il Perceptron di Frank Rosemblatt. Alla fine degli anni Sessanta lo sviluppo delle reti neurali subisce uno stop, a causa della pubblicazione di un'analisi da parte degli studiosi del MIT Marvin Minsky e Seymour Papert, in cui venivano criticate le macchine di tipo Perceptron, in quanto dotate di gravi limitazioni nella risoluzione di alcune tipologie di problemi, per i quali, secondo gli studiosi, l'unica soluzione era rappresentata da reti neurali omniconnesse in cui ciascun neurone è connesso con tutti gli altri neuroni della rete. [8]

Il migrare degli interessi dalle reti neurali al campo dell'Intelligenza Artificiale (apparentemente più promettente) fu causato anche dal fatto che la tecnologia allora disponibile rendeva molto difficoltosa o addirittura impossibile la sperimentazione nel campo delle reti neurali e non vi erano computer abbastanza veloci in grado di simulare reti neurali complesse. Per questi motivi, in quegli anni, le applicazioni create furono scarse e le ricerche vertevano maggiormente su ambiti teorici. Solo negli anni Ottanta, grazie al progresso della tecnologia con l'affinamento delle capacità di calcolo, fu dapprima elaborata la prima rete neurale con neuroni nascosti e in seguito introdotto l'algoritmo di retropropagazione dell'errore, in base al quale i pesi delle connessioni sono sistematicamente modificati allo scopo di migliorare l'output finale della rete. Nel 2000 gli stessi Minsky e Papert rivederono le loro critiche e contribuirono ad identificare nuove direzioni di sviluppo.[8] Attualmente i settori in cui le reti neurali sono più utilizzati sono:

- La visione e il riconoscimento di forme
- La comprensione del linguaggio naturale
- Le memorie associative e la robotica
- L'inferenza e la risoluzione di particolari problemi computazionali

Tutte le reti neurali presentano alcune proprietà:

- *Parallelismo*: tenuto conto dell'indipendenza di ciascun neurone, è possibile recuperare le informazioni disponibili in parallelo. Tale proprietà è tipica del meccanismo neurale di apprendimento e, infatti, il cervello è in grado di risolvere con grande velocità dei problemi che considerano un elevato numero di dati (un esempio è dato dal riconoscimento visivo di oggetti o persone)
- *Conoscenza distribuita*: la connessione tra neuroni e tra strati diversi permette una distribuzione ampia di conoscenze

- *Tolleranza:* Gli input utilizzabili possono essere anche non perfettamente adeguati al problemi da risolvere, senza che la qualità dell'output ne risenta
- *Non linearità*: Potendo sfruttare diverse tipologie di relazioni, l'insieme delle connessioni sarà non lineare
- *Approssimazione universale*: Grazie alle proprietà precedenti, si può ricostruire una legge all'interno del dominio analizzato, anche se con margini di errore. L'analista avrà il compito di valutare la tollerabilità dell'errore in relazione al problema affrontato
- *Precisione sfumata*: Anche se apparentemente questi due termini appaiono in netta contrapposizione tra loro, secondo il principio della logica fuzzy o sfumata, il migliore risultato di un problema non deve soddisfare necessariamente tutte le proprietà attese
- *Classificabilità*: Un'altra proprietà delle reti è l'abilità nella classificazione delle osservazioni contenute nel dataset, permettendo il raggiungimento di risultati non parametrici e quindi non dipendenti da ipotesi sulla distribuzione dei dati.
- *Generalizzabilità*: Grazie a questa proprietà è possibile risolvere i problemi legati alla previsione poiché le regole individuate dalle reti valgono sia per le informazioni contenute negli input, sia per i dati non presenti nel campione

Ogni rete neurale è organizzata in vari strati:

- Strato di input: questi neuroni, il cui numero varia in base al tipo e alla quantità di informazioni in entrata, hanno compiti di trasmissione dei dati agli strati successivi
- Strati nascosti: questi strati, il cui numero dipende dalla complessità della situazione, contengono i neuroni che svolgono la funzione di elaborazione. È sufficiente una rete con un solo strato nascosto nel caso in cui si debba approssimare una funzione di tipo lineare mentre per problemi più complessi sono necessari due o tre livelli
- Strato di output: ha la funzione di restituire il valore elaborato.

La stragrande maggioranza dei lavori empirici utilizza un solo strato nascosto in quanto è sufficiente ad approssimare funzioni non lineari con elevato grado di accuratezza. Tuttavia, questo approccio richiede un elevato numero di neuroni, andando a limitare il

processo di apprendimento. Risulta, quindi, essere più efficace l'utilizzo di reti con due strati nascosti, soprattutto per previsioni su dati ad alta frequenza. Questa scelta, oltre ad essere suggerita da un'apposita teoria, è supportata dall'esperienza, la quale mostra come, d'altro canto, un numero di strati nascosti superiore a due non produce miglioramenti nei risultati ottenuti dalla rete. In merito al numero di neuroni da assegnare agli strati nascosti, è necessario minimizzare il rischio di overfitting, in base al quale un numero troppo elevato di neuroni delinea quasi ottimamente il pattern della serie storica ma, d'altro canto, riducendo il contributo degli input, non è in grado di generale una previsione affidabile; in questi casi, è come se la rete avesse "imparato a memoria" le risposte corrette, senza essere in grado di generalizzare. Al contrario, un numero troppo basso di neuroni riduce il potenziale di apprendimento della rete. Occorre trovare una soluzione di trade-off fra un numero troppo basso o troppo elevato di neuroni. Il numero degli strati nascosti viene scelto in base alla natura del problema da risolvere ed è strettamente legato al numero di neuroni presenti in ciascun strato nascosto.

La funzione lineare viene in genere utilizzata per lo strato che contiene l'output della rete neurale: la ragione è che questa funzione, pur essendo più rigida delle alternative, evita che il risultato tenda verso il minimo o il massimo. Meno efficace è, invece, l'utilizzo della funzione lineare negli strati nascosti, soprattutto se questi sono caratterizzati da un elevato numero di neuroni, che risulterebbero così connessi proprio su una base funzionale che si vuole superare con l'utilizzo della rete stessa.

La funzione logistica e quella logistica simmetrica presentano la caratteristica di variare, rispettivamente, negli intervalli (0; 1) e (−1; 1). La prima è particolarmente utile negli strati nascosti delle reti applicate alle serie storiche finanziarie. Alcuni problemi presentano caratteristiche dinamiche che sono colte in misura più precisa dalla funzione simmetrica, soprattutto nello strato di input e in quelli nascosti. La maggior parte della letteratura empirica presenta l'utilizzo di questa funzione nello strato nascosto, anche se manca una robusta motivazione teorica. Il metodo più diffuso di interconnessione dei neuroni è quello di piena connessione (fully connected neural network), dove ogni singolo neurone di un livello è connesso con tutti i neuroni del livello successivo e dove non vi sono connessioni tra neuroni appartenenti allo stesso strato.

In merito al problema della connessione tra i vari strati di neuroni, sono diffuse tre varianti:

a) connessioni standard, che prevedono connessioni dirette tra input e output passando attraverso uno o più strati nascosti

b) connessioni a salti, che implicano collegamenti anche tra neuroni presenti in strati non adiacenti con il conseguente aumento della ramificazione al crescere del numero degli strati nascosti

c) connessioni ripetute, che permettono ai neuroni degli strati nascosti di ricollegarsi ai neuroni di variabili input con processi iterativi in modo da attribuire precisamente i pesi

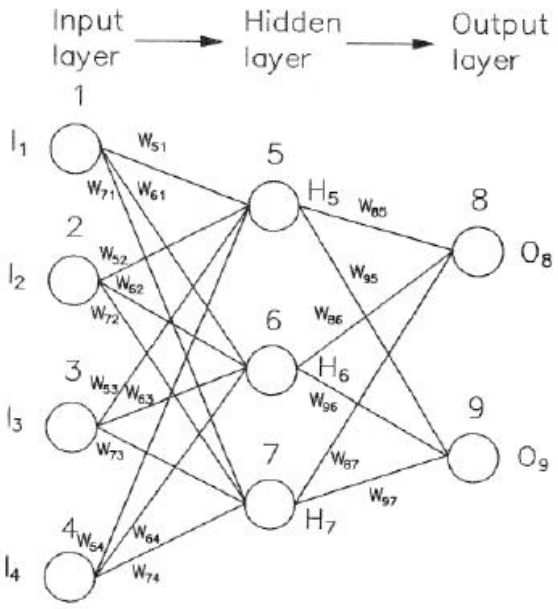

Figura 3:Rete Neurale

I pesi sinaptici sono definiti in modo che la funzione che collega gli input agli output sia la migliore possibile; tale funzione è ottenuta da un processo di apprendimento basato su dati empirici che può essere di tre tipi: supervisionato, non supervisionato o per rinforzo. Il metodo standard più usato in letteratura in riferimento agli algoritmi supervisionati è l'algoritmo di back-propagation, il quale, attraverso opportune modifiche dei pesi della rete, cerca di minimizzare, in termini assoluti, una predefinita funzione di performance.

La funzione più usata è la regola del gradiente, dove l'errore è dato dall'errore quadratico medio registrato ogni volta che l'output parziale della rete non coincide con quello atteso. È presente, inoltre, un'ulteriore possibilità di correzione dell'errore, chiamata momentum, attraverso la quale si decide la proporzione con cui si aggiunge la variazione dell'ultimo peso raggiunto al nuovo peso.

La frequenza di cambiamento con la quale la rete deve modificare i pesi dei neuroni rispetto alla significatività dell'errore commesso è un parametro fondamentale che determina il successo o meno del modello. Adottando una frequenza di apprendimento

troppo elevata, c'è il rischio che la rete oscilli in modo troppo ampio, condizionante l'evoluzione corretta della serie storica. Un modo alternativo di procedere è quello di assegnare un valore corretto al momentum, permettendo alla rete un apprendimento a tasso elevato ma limitando l'oscillazione grazie al recupero di un'elevata quota dell'ultimo peso raggiunto. L'analista dovrà poi decidere il livello iniziale del peso da dare alla connessione fra neuroni e valutare se le osservazioni sono caratterizzate da un elevato tasso di rumore. Il processo di apprendimento della rete passa naturalmente per l'individuazione progressiva del valore più adatto di questi parametri.

2. TECNICHE DI MACHINE LEARNING

Con l'avvento delle tecnologie digitali e in special modo del web, la mole di dati generata e resa disponibile è esplosa, così da far definire la società contemporanea come "la società dei dati"[9]. Il possesso e il controllo dei dati è diventato fattore abilitante di vantaggio competitivo per le più grandi società tecnologiche mondiali, in special modo statunitensi e cinesi, le cosiddette "big tech", come Apple, Google, Facebook, Amazon, Tencent, Alibaba, etc. Da questa esigenza è nata la moderna disciplina nota come data science, che negli ultimi anni è rapidamente cresciuta passando dall'essere una piccola branca dell'information technology, ad una scienza a sé in cui si intrecciano conoscenze di molteplici discipline diverse (matematica, computer science, statistica, machine learning, ricerca operativa…) necessarie per interpretare e dare valore ai dati.

L'estrazione delle informazioni a partire dai dati si realizza tramite il processo di data mining, in cui si applicano una serie di algoritmi in grado di estrarre pattern specifici, sequenze di informazioni che permettono di massimizzare la funzione obbiettivo (es. minimo costo, massima utilità).

Il data mining a sua volta va inquadrato in un processo più ampio di estrazione di conoscenza a partire da una base di dati grezza, noto come KDD (knowledge discovery from data)[10].

Il KDD si articola in cinque step:

- SELEZIONE: Si scelgono quali dati e quali variabili sono significativi all'interno del database
- PRE-TRATTAMENTO: Viene operata una pulizia dei dati al fine di eliminare o correggere dati verosimilmente sbagliati (outliers), di gestire dati mancanti o di correggere errate attribuzioni
- TRASFORMAZIONE: I dati pretrattati vengono trasformati nella maniera più funzionale alla futura estrazione dei pattern. In altre parole, la rappresentazione dei dati deve essere resa la più opportuna in relazione agli obbiettivi della ricerca. Generalmente si può ricorrere a tecniche di normalizzazione, discretizzazione, aggregazione, standardizzazione e gerarchizzazione dei dati
- DATA MINING: Si estraggono i pattern di interesse in una particolare forma di rappresentazione o su un set di rappresentazioni diverse (regole di classificazione,

alberi decisionali, regressione, clustering...). Il risultato del processo di data mining è considerevolmente influenzato dalla correttezza delle fasi precedenti

- INTERPRETAZIONE-VALUTAZIONE: Si analizzano i pattern trovati e, se non si è soddisfatti, si itera nuovamente tutto il processo. Se invece i risultati sono soddisfacenti, si generano report e documenti utili alle parti interessate. Inoltre, quando il lavoro è di una certa rilevanza, si deve assistere ad un processo di consolidamento della conoscenza estratta al fine di incorporare tale conoscenza e renderla disponibile all'esterno. Questo lavoro include anche il controllo per la risoluzione di contraddizioni con la conoscenza precedentemente disponibile

Il KDD quindi, tramite il data mining, consente di estrarre informazione e conoscenza per informare le decisioni di business, consentendo un'ottimizzazione delle stesse e aumentando l'efficienza dei processi aziendali che se ne servono.

Il data mining può usare molteplici algoritmi di natura diversa, tra questi vi sono molti algoritmi di machine learning. Il machine learning è l'insieme delle tecniche che consentono di istruire delle macchine all'apprendimento di determinate conoscenze che poi utilizzeranno per lo svolgimento di specifiche attività, senza essere esplicitamente programmate per farlo. L'applicazione degli algoritmi di machine learning è preceduta dall'estrazione delle feature, caratteristiche descrittive dei dati significative nel particolare contesto.

Il machine learning si divide in tre branche: apprendimento supervisionato, non supervisionato e semi supervisionato.

2.1. Apprendimento supervisionato

Un algoritmo è supervisionato quando viene utilizzato un dataset che contiene dati annotati tramite un'etichetta[11]. In base al dataset, e tenendo come "ancora" l'etichetta, l'algoritmo apprenderà (in vari modi) come classificare un dataset completamente nuovo, ma che non contiene l'etichetta, in base alle informazioni del primo dataset già classificato. La variabile etichetta indica per ogni istanza del dataset, che essa è da intendersi in un dato modo (spam/non spam, vince/perde), mentre tutte le altre variabili sono utilizzate per costruire una logica di classificazione. Questa logica di classificazione verrà quindi utilizzata per classificare nuovi dataset non etichettati. A seconda del tipo di output, la classificazione sarà binaria, se include solo due classi (es. spam/non spam), oppure multiclasse se è possibile avere più di due output (es.

bambino/ragazzo/anziano…). Parallelamente si possono avere più variabili etichetta, fra loro indipendenti, e in tal caso la classificazione è detta multilabel.
Le tecniche di machine learning in cui si usano algoritmi di classificazione hanno la variabile etichetta che assume valori discreti. Gli algoritmi in cui l'etichetta assume valori numerici nel continuo sono invece algoritmi di regressione. L'output della regressione è sempre di tipo quantitativo e permette di individuare una funzione nel continuo.
La variabile etichetta rappresenta il valore da predire, funzione delle altre variabili (feature). Tanto più il modello sarà addestrato correttamente, tanto più il valore predetto da tale modello si avvicinerà al valore della variabile etichetta. Facendo un focus sugli algoritmi di classificazione binari, la variabile etichetta potrà assumere il valore VERO/FALSO (o 0/1) e nella previsione del modello si potranno generare due diverse tipologie di errore, i falsi positivi e i falsi negativi. I falsi positivi si hanno quando il valore predetto è vero, ma il valore reale è falso, i falsi negativi viceversa.
Nella valutazione delle prestazioni di un algoritmo di classificazione, si possono usare metriche diverse, a seconda del focus di interesse dell'applicazione concreta. L'accuratezza è il numero di elementi classificati correttamente dal modello in rapporto agli elementi totali. La recall è il numero di veri positivi in rapporto al numero di veri positivi sommato al numero di falsi negativi. La precisione è il numero di veri positivi in rapporto ai veri positivi più i falsi positivi. La recall può essere vista come una misura di copertura, indicando la quota di veri positivi indicata dal modello sui positivi effettivi. Essa è quindi una misura utile quando ci si vuole concentrare sulla capacità del modello di evitare falsi negativi, ad esempio nelle diagnosi di un cancro, in cui la mancata diagnosi è molto più dannosa di una diagnosi positiva che poi si rivela errata. La precisione, al contrario, si utilizza quando si vogliono evitare falsi positivi, ma è accettabile che il modello sbagli individuando falsi negativi.
Infine, c'è una quarta metrica, la F-misura (o F1) che tiene conto contemporaneamente di precisione e recall secondo la formula:

$$F_1 = \frac{2 * P * R}{P + R}$$

Si passa ora ad analizzare in dettaglio gli algoritmi di apprendimento supervisionato più usati.

2.1.1. Alberi decisionali

Un albero di decisione è un sistema con n variabili in input e m variabili in output. Le variabili in input (attributi) sono derivate dall'osservazione dell'ambiente. Le variabili in output, invece, identificano la decisione / azione da intraprendere. Negli alberi decisionali profondi le variabili in output intermedie, in uscita dai nodi genitori, coincidono con le variabili in input dei nodi figli e condizionano il percorso verso la decisione finale.
Il processo decisionale è rappresentato con un albero logico rovesciato dove ogni nodo è una funzione condizionale. Ogni nodo verifica una condizione (test) su una particolare proprietà dell'ambiente (variabile) e ha due o più diramazioni verso il basso. Il processo consiste in una sequenza di test. Comincia sempre dal nodo radice, il nodo genitore situato più in alto nella struttura, e procede verso il basso. Ogni nodo è un punto di scelta, rappresentato dai valori che può assumere la variabile. A seconda dei valori rilevati in ciascun nodo, il flusso prende una direzione oppure un'altra e procede progressivamente verso il basso. Man mano che il processo di selezione prosegue verso il basso, lo spazio delle ipotesi si riduce perché gran parte dei rami decisionali dell'albero sono eliminati. La decisione finale si trova nei nodi foglia terminali, quelli più in basso dove, dopo aver analizzato le varie condizioni, l'agente giunge alla decisione finale. Per scegliere in modo efficiente l'ordine delle variabili da usare nell'albero decisionale, si devono usare inizialmente le variabili più discriminanti, cioè quelle in grado di minimizzare l'entropia massimizzando il guadagno informativo, e poi via via tutte le altre fino alla meno discriminante.
In un albero decisionale le variabili possono essere discrete o continue e daranno origine rispettivamente ad una classificazione o ad una regressione. La rappresentazione delle variabili continue è più complessa, ma si adatta meglio alla logica sfumata (fuzzy logic) e alla logica in condizioni di incertezza, quando non esiste una distinzione netta tra i valori che può assumere la variabile etichetta[12]. Gli alberi logici hanno l'indiscusso vantaggio della semplicità. Sono facili da capire e da eseguire. Rispetto alle reti neurali l'albero decisionale è facilmente comprensibile dagli esseri umani. Pertanto, l'uomo può verificare come la macchina giunge alla decisione. Eventualmente dissentire. Ad esempio, se un albero decisionale applicato alla medicina fornisce delle diagnosi, essendo una decisione importante per il paziente, è sempre opportuno che un medico verifichi il processo di classificazione che ha portato la macchina a prendere quella decisione. Potrebbero anche esistere criteri decisionali più efficienti, più adatti alla logica delle macchine ma meno

comprensibili dall'uomo e non adatti a specifici contesti. Inoltre, gli alberi decisionali booleani sono facilmente sviluppabili sotto forma di codice di programmazione, perché possono essere rappresentati con qualsiasi linguaggio proposizionale, e tutte le funzioni booleane possono essere rappresentate come albero decisionale. Quest'ultima caratteristica è detta espressività degli alberi decisionali[12].

Gli alberi decisionali hanno però degli svantaggi che ne limitano il campo di applicazione. La rappresentazione ad albero decisionale è poco adatta per i problemi complessi, perché lo spazio delle ipotesi diventa troppo grande. La complessità spaziale dell'algoritmo potrebbe diventare proibitiva. Nel caso più semplice di un albero booleano, per n attributi in input (es. le variabili in input A, B, C) occorrono 2n combinazioni (cammini nelle ramificazioni) ossia 2^n righe in una tavola di verità per determinare la decisione finale (SI/NO). Per questo, se un nodo avesse più attributi (numero classi = numero rami in uscita) o fosse una variabile non booleana (es. temperatura), le combinazioni sarebbero decisamente molte di più.

Inoltre, un albero decisionale può descrivere soltanto una relazione tra una funzione di verità e le combinazioni logiche di attributi. Non riesce a rappresentare più funzioni. Per farlo occorre creare un altro albero decisionale e associarlo al precedente. Ogni funzione di verità è un albero decisionale a sé stante. Infine, l'albero decisionale non riesce a rappresentare tutti i tipi di funzioni come, ad esempio, una funzione che restituisce 1 quando tutti gli attributi sono veri (funzione di parità) o quando la maggioranza degli attributi sono veri (funzioni di maggioranza)[12].

2.1.2. Naive Bayes

Il Naive Bayes è un algoritmo di classificazione che si fonda esclusivamente sui teoremi del calcolo probabilistico. Esso calcola la probabilità di ogni etichetta per un determinato oggetto osservando le sue caratteristiche. Poi, sceglie l'etichetta con la probabilità maggiore. E' detto "naive", ossia ingenuo, perché le ipotesi di partenza sono molto semplificate.

Dati due eventi A e B, dove A è la classe e B sono le variabili attributo, si definisce probabilità condizionata P(A|B) la probabilità del verificarsi dell'evento A, condizionata al fatto che sia verificato l'evento B. Questa probabilità è pari alla probabilità che i due eventi siano entrambi verificati rapportata alla probabilità che sia verificato l'evento B. Ovvero:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Da questa formula possiamo ricavare il teorema di Bayes, mettendo in relazione le due probabilità condizionate $P(A|B)$ e $P(B|A)$ e ottenendo:

$$P(A|B) = P(B|A) * \frac{P(A)}{P(B)}$$

Questo risultato è molto importante, in quanto permettere di esprimere la probabilità a posteriori P(A|B) che una determinata ipotesi A (l'appartenenza ad una determinata classe) sia verificata in presenza del verificarsi di una evidenza B (ad esempio l'occorrenza di una determinata frase da classificare).

Una assunzione "naive" ma assai utile per molte applicazioni è quella di considerare le caratteristiche dell'evidenza B stocasticamente indipendenti tra loro. In altre parole, si assume che non ci sia correlazione tra le caratteristiche di B.[13]

Per esempio, si consideri un frutto che può essere classificato nella classe "mele" se è di colore rosso, ha un diametro di circa 10 cm e ha una forma rotonda. Sono tre caratteristiche differenti. Un algoritmo di classificazione Naive Bayes considera ognuna di queste caratteristiche come un contributo indipendente alla probabilità complessiva che il frutto sia una mela. Non considera le possibili correlazioni tra le caratteristiche (colore, diametro, forma).

Questa ipotesi è chiaramente una forzatura, ma semplifica notevolmente il modello e risulta accettabile per molte applicazioni concrete di algoritmi di classificazione.

La P(B) è la somma delle singole probabilità delle caratteristiche ed è detta evidenza. Essa può essere ignorata perché non influisce attivamente sul processo di classificazione.

La P(A) è nota come probabilità a priori, cioè la probabilità sull'intera popolazione che il classificatore dia un certo output.

La P(B/A) è nota come verosimiglianza e rappresenta la probabilità che si verifichi l'evidenza B data la classe A. La verosimiglianza può essere stimata ricorrendo a diversi tipi di distribuzioni. Generalmente si ricorre alla distribuzione Gaussiana e, in tal caso, il classificatore è ribattezzato Gaussian Naive Bayes. La Gaussiana usata in Bayes è però non standard, quindi vanno stimate media e varianza per poterla rappresentare adeguatamente. Media e varianza sono stimate tramite un processo numerico, noto come "Stima della massima verosimiglianza". [13]

La massimizzazione della verosimiglianza consente contemporaneamente di massimizzare la probabilità a posteriori P(A|B), data la natura fissa della P(A). Questa sarà l'etichetta individuata dal classificatore come la più probabile.

2.1.3. K-Nearest Neighbors

L'algoritmo K-NN (K-Nearest Neighbors) è un algoritmo di riconoscimento dei pattern basato sulla vicinanza dei dati. L'algoritmo è molto semplice e la fase di addestramento non richiede troppe risorse. Il KNN è uno strumento non parametrico, ossia non fa alcuna ipotesi sulla distribuzione dei dati che analizza. Di conseguenza, KNN potrebbe essere una delle prime scelte per uno studio di classificazione quando c'è poca o nessuna conoscenza precedente sulla distribuzione dei dati, ovviamente presupponendo che la rappresentazione stessa dei dati consenta di confrontarli tramite una qualche misura di similarità o di distanza all'interno di uno spazio normato.

Il principio di base che esso utilizza è che, se un oggetto A ha caratteristiche simili a un oggetto B, probabilmente A appartiene alla stessa classe (categoria) di B.

Si parte da un dataset contenente istanze annotate. Per prima cosa si rendono le unità di misura delle variabili tra loro comparabili, per fare in modo che l'influenza sull'algoritmo sia correttamente bilanciata. Si può ricorrere a varie tecniche, come ad esempio la normalizzazione.

A questo punto, si prende l'istanza da classificare e si calcola la distanza dalle istanze già classificate. Si scelgono solo le k istanze più vicine e la classe più frequente sarà quella associata all'istanza (da ciò, nasce l'esigenza di scegliere un K dispari, per evitare situazioni di parità). Nella regressione, invece, il KNN sceglie il valore medio dei valori dei vicini più prossimi come valore previsto. Fondamentale per le prestazioni del modello è la scelta del parametro k. In linea teorica, esso può oscillare tra uno e il numero di istanze. Se K è troppo piccolo, l'esempio è assegnato in base ai pochi esempi più vicini e la classificazione è influenzata dal rumore nei dati: stiamo limitando la regione di una determinata previsione e costringendo il nostro classificatore ad essere "più cieco" rispetto alla distribuzione generale.[14]

Se K è troppo grande, l'esempio viene classificato nella categoria più numerosa nella popolazione. Inoltre, la complessità computazionale dell'algoritmo aumenta perché il calcolo della distanza è un'operazione molto onerosa in termini di risorse e tempo. In generale quindi, si sceglie come k un numero dispari non molto elevato. Alcuni autori

suggeriscono di impostare k uguale alla radice quadrata del numero di osservazioni nel set di dati di addestramento.

L'idea di distanza o vicinanza può perdere significatività in spazi altamente dimensionali (molte variabili di input) che possono influire negativamente sulle prestazioni dell'algoritmo sul problema di interesse. Questo fenomeno viene chiamato la maledizione della dimensionalità (curse of dimensionality).[14]

2.1.4. Support-vector machines

L'SVM è un algoritmo basato sull'idea di trovare un iperpiano di separazione che divida al meglio un set di dati in due classi. Si definisce iperpiano il luogo geometrico dei punti di dimensione n-1 che separa lo spazio di dimensione n in cui giace in due semispazi. Esempio, un iperpiano in uno spazio tridimensionale è un piano, un iperpiano in uno spazio bidimensionale è una retta. Si definiscono vettori di supporto i punti del dataset più vicini all'iperpiano scelto dall'algoritmo. Tali punti dipendono dal set di dati che si sta analizzando e se vengono rimossi o modificati alterano la posizione dell'iperpiano divisorio. Per questo motivo, possono essere considerati gli elementi critici di un set di dati. Si definisce margine la distanza tra i vettori di supporto di due classi differenti più vicini all'iperpiano. Alla metà di questa distanza viene tracciato l'iperpiano, o retta nel caso si stia lavorando a due dimensioni.[15]

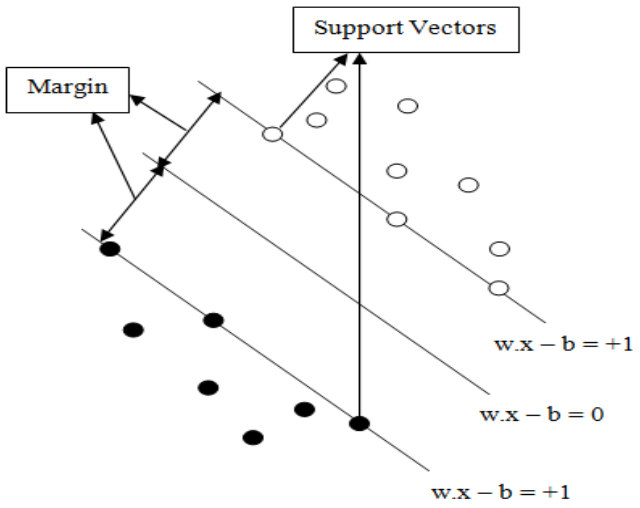

Figura 4: Support Vector Machine

Il Support Vector Machine ha l'obiettivo di identificare l'iperpiano che meglio divide i vettori di supporto in classi. Per farlo esegue i seguenti step:

1. Cerca un iperpiano linearmente separabile o un limite di decisione che separa i valori di una classe dall'altra. Se ne esiste più di uno, cerca quello che ha margine più alto con i vettori di supporto, per migliorare l'accuratezza del modello.
2. Se tale iperpiano non esiste, SVM utilizza una mappatura non lineare per rappresentare i dati di allenamento in uno spazio a dimensionalità superiore (es. da uno spazio a due dimensioni ad uno spazio a tre dimensioni). In tal modo, si può individuare un iperpiano in questo nuovo spazio che riesca a separare i dati, i quali, alla fine del processo, verranno nuovamente proiettati nello spazio originale. Questa tecnica è nota come trucco del Kernel. La scelta della funzione del kernel (radiale, polinomiale, ecc…) tra gli altri fattori potrebbe influire notevolmente sulle prestazioni di un modello SVM. Così come gli altri iperparametri, come ad esempio quello che regola il trade-off tra accuratezza e margine, anche la funzione kernel può essere scelta ricorrendo a tecniche di tuning tra cui le più usate sono la grid-search, randomized-search e gli algoritmi genetici.

L'algoritmo SVM è molto efficace in dimensioni spaziali elevate ed è molto efficiente dal punto di vista computazionale, grazie al fatto che solo un sottoinsieme dei punti di allenamento viene utilizzato nel processo di assegnazione dei nuovi membri. Infine, è anche molto versatile, poiché la capacità di applicare nuovi kernel consente una sostanziale flessibilità per i limiti decisionali, portando a una maggiore performance di classificazione.[15]

D'altro canto, la sua interpretazione può risultare molto complessa con dei risultati poco trasparenti e difficili da visualizzare, per cui si ricorre a tecniche di visualizzazione grafica. In aggiunta, il metodo SVM non è un metodo probabilistico, quindi non si può fornire una stima diretta della probabilità dell'oggetto di appartenere al gruppo, ma si può usare la distanza dal confine per avere un'idea qualitativa. Se si vuole ottenere la probabilità associata alla classificazione, infine, si può ricorrere ad una tecnica definita calibrazione dell'SVM.

2.1.5. Modelli di regressione

Il modello più semplice è quello di regressione lineare, usato quando si vuole prevedere un valore continuo. Se la variabile di input è solo una allora la regressione lineare si dice semplice, altrimenti in caso contrario la regressione lineare si dice multipla. Si ha una regressione logistica quando si vuole prevedere a quale classe appartiene l'osservazione che si sta analizzando. In effetti tale tecnica, strettamente connessa alla funzione sigmoidea, si colloca (nonostante l'appellativo "regressione") tra le tecniche di classificazione.

La regressione più utilizzata è la regressione dei minimi quadrati (OLS, Ordinary Least Square). È un metodo statistico di analisi che stima la relazione tra una o più variabili indipendenti e una variabile dipendente. Il metodo stima la relazione attraverso degli stimatori (definiti stimatori dei minimi quadrati), che minimizzano la somma dei quadrati della differenza tra i valori osservati e previsti della variabile dipendente configurata come linea retta. Gli stimatori dei minimi quadrati determinano i valori minimi del modello di regressione lineare risolvendo l'equazione:

$$\beta = arg\ min_b \Sigma_{i=1}^{n} (y_i - x_i * b)^2$$

Con arg min b, si intende l'argomento del minimo di b, cioè l'insieme dei punti per i quali una data funzione raggiunge il suo minimo. Gli stimatori dei minimi quadrati hanno bias nullo, ma possono avere varianza molto alta e quindi essere poco accurati. Esistono delle efficaci tecniche di regolarizzazione che consentono di ridurre significativamente la varianza del modello predittivo, a patto di aumentare leggermente il bias. Tra le tecniche più usate si annoverano Ridge, Lasso ed Elastic Net.

La regressione Ridge aggiunge alla normale regressione dei minimi quadrati una quantità quadratica che funge da perturbazione. La logica sottostante è che se il modello impara a minimizzare la funzione di costo in presenza di una perturbazione, a maggior ragione sarà capace di farlo in sua assenza e quindi potrà essere considerato più robusto. Dal punto di vista pratico, inoltre, l'introduzione di una perturbazione (o penalità) aumenta il bias e di conseguenza presenta un effetto regolarizzante.

L'equazione diventa:

$$\beta = arg\ min_b \Sigma_{i=1}^{n} (y_i - x_i * b)^2 + \lambda * \Sigma_{k=1}^{K} b_k^2$$

L'equazione è composta dalla somma dei residui quadrati più una penalità, definita dalla somma dei coefficienti quadrati b e riscalata di un fattore Lambda. Per uno stimatore ridge, la selezione di un buon valore per λ è fondamentale. Quando λ è nullo, il termine di penalità non ha alcun effetto e la regressione ridge produrrà i coefficienti minimi quadrati classici. Tuttavia, quando λ aumenta all'infinito, l'impatto della penalità aumenta e i coefficienti di regressione si avvicinano allo zero. Un importante vantaggio della regressione ridge è che si comporta ancora bene, rispetto al normale metodo dei minimi quadrati, in una situazione in cui si hanno molti dati multivariati con il numero di predittori maggiore del numero di osservazioni. Uno svantaggio della regressione ridge, invece, è che includerà tutti i predittori nel modello finale senza fare una selezione.
L'equazione di una regressione Lasso è leggermente diversa:

$$\beta_\lambda = arg\ min_b \Sigma_{i=1}^{n}(y_i - x_i * b)^2 + \lambda * \Sigma_{k=1}^{K}|b_k|$$

In questo caso la perturbazione introdotta è di tipo lineare. Al crescere di Lambda, i coefficienti di regressione vengono ridotti con l'effetto di forzare alcune delle stime dei coefficienti, con minore contributo al modello, a essere nulle. In altre parole, con il modello Lasso si realizza una selezione dei predittori con riduzione della complessità del modello. Quando Lambda è piccolo, il risultato è molto vicino alla stima dei minimi quadrati. All'aumentare di lambda, si verifica una contrazione in modo da poter eliminare le variabili che sono a zero. Un ovvio vantaggio della regressione lasso rispetto alla regressione ridge è che produce modelli più semplici e più interpretabili che incorporano solo un insieme ridotto di predittori. In generale, il lasso potrebbe funzionare meglio in una situazione in cui alcuni predittori hanno coefficienti elevati e i restanti predittori hanno coefficienti molto piccoli. La regressione ridge funzionerà meglio quando il risultato è una funzione di molti predittori, tutti con coefficienti di dimensioni approssimativamente uguali[16]. C'è infine un modello intermedio noto come Elastic Net, con la seguente funzione di costo:

$$\beta_{enet} = \frac{arg\ min_b \Sigma_{i=1}^{n}}{2n} + \lambda(\frac{1-\alpha}{2}\Sigma_{k=1}^{K} b_k^2 + \alpha\Sigma_{k=1}^{K}|b_k|)$$

Oltre a impostare e scegliere un valore Lambda, l'elastic net ci consente anche di ottimizzare il parametro alfa dove α = 0 corrisponde alla regressione ridge e α = 1 alla regressione lasso.[17]In altre parole la penalità introdotta è riconducibile, con buona

approssimazione, ad una combinazione convessa delle penalità descritte nei due precedenti approcci.

2.1.6. Modelli di ensemble learning

Il concetto di apprendimento Ensemble richiama l'utilizzo di diversi modelli differenti, uniti in un certo modo per riuscire a massimizzarne le prestazioni usando i punti di forza di ognuno e limitando le debolezze dei singoli. Alla base del concetto di Ensemble Learning ci sono i singoli classificatori, detti classificatori deboli. Sommati in un certo modo tra di loro, i classificatori deboli permettono di costruire un classificatore forte. I vantaggi principali del classificatore forte creato sono la maggiore capacità di generalizzazione e la maggiore robustezza rispetto agli outlier. Intuitivamente, pescando da più modelli, se anche uno di essi dovesse essere più sensibile agli outlier, gli altri permetterebbero di smorzare quest'effetto. Infatti, tutti i modelli hanno una certa quantità di errore, ma gli errori di un modello saranno diversi dagli errori prodotti da un altro modello. Quando tutti gli errori vengono esaminati, non saranno raggruppati attorno a una risposta o all'altra, ma saranno sparsi. Le ipotesi errate sono essenzialmente distribuite su tutte le possibili risposte sbagliate, e tendono in buona misura ad annullarsi a vicenda. Nel frattempo, le ipotesi corrette dei diversi modelli saranno raggruppate attorno alla risposta vera e corretta. Ne deriva che, quando vengono utilizzati metodi di ensemble, la risposta corretta può essere trovata con maggiore affidabilità.
Ci sono tre diversi tipi di ensemble:

1. ***Bagging***

   Nel bagging più modelli dello stesso tipo (alberi decisionali nel caso della tecnica Random Forest) vengono addestrati su dataset diversi, ciascuno ottenuto dal dataset iniziale tramite campionamento casuale con rimpiazzo. Di conseguenza, ogni singolo classificatore si addestra su una porzione casuale di caratteristiche e quindi alcune di esse possono comparire contemporaneamente in più modelli mentre altre potrebbero non comparire mai.[18] L'allenamento su una parte delle caratteristiche consente ad ogni modello di limitare l'overfitting e migliorare le capacità predittive. Il bagging può essere usato per casi di classificazione, regressione e clustering. Nella classificazione, il criterio usato dal classificatore forte è di scegliere la classe predetta a maggioranza dai singoli classificatori deboli. Nella regressione si prendono invece i valori in uscita da ogni modello

debole (media, moda, varianza, mediana…) e si aggregano ricavandone la distribuzione del modello forte.

2. ***Gradient Boosting***

Nel gradient Boosting si utilizza un sistema di raffinazioni successive in cui si inizia a partire da un primo modello analizzandone le prestazioni. In seguito, viene creato un secondo modello in sequenza, che si concentra sulla previsione accurata dei casi in cui il primo modello ha prestazioni scarse. La combinazione di questi due modelli dovrebbe essere migliore di entrambi i modelli presi singolarmente. Il processo viene ripetuto molte volte. Ogni modello successivo tenta di correggere le carenze dell'insieme boosting combinato di tutti i modelli precedenti.[19]

In linea generare, l'algoritmo Gradient Boosting ragiona al contrario della maggior parte degli algoritmi esaminati fin'ora. Anziché puntare a prevedere il risultato della variabile target, si prefigge di prevedere gli errori del modello. Questi errori vengono stimati dai residui, ossia dalla differenza tra il valore realmente osservato e il valore previsto. I residui vengono calcolati da alberi di regressione a più di un livello di profondità. Una volta calcolati i residui, il Gradient Boosting si avvale di un altro modello (solitamente sempre un albero di regressione) per prevedere i nuovi residui. E così via ad ogni iterazione: il modello punta a stimare i residui che vengono utilizzati per compensare gli errori commessi. Ciò causa la progressiva diminuzione dei residui stessi e l'aumento delle performance del modello.

Alla conclusione dell'algoritmo, che avviene dopo un certo numero di iterazioni, l'errore complessivo del modello è ridotto rispetto l'inizio. Quando si implementa un algoritmo Gradient Boosting si devono tenere a mente due parametri da cui dipenderà l'accuratezza del modello: la scelta del numero di alberi da utilizzare e il tasso di apprendimento[19]. Il primo corrisponde semplicemente al numero di alberi che saranno adattati in serie per correggere gli errori di predizione. Il tasso di apprendimento, detto anche Learning Rate (LR), corrisponde alla velocità con cui l'errore viene corretto da ciascun albero al successivo ed è un semplice moltiplicatore che ricade nell'intervallo]0,1] (0 escluso e 1 incluso). Ad esempio, se la previsione corrente per un campione particolare è 0,2 e l'albero successivo prevede che dovrebbe effettivamente essere 0,8, la correzione sarebbe + 0,6. A un tasso di apprendimento di 1, la previsione aggiornata sarebbe esattamente la 0,2 +

1 * (0,6) = 0,8, mentre una velocità di apprendimento di 0,1 aggiornerebbe la previsione a 0,2 + 0,1*(0,6) = 0,26. È consigliabile impostare un tasso di apprendimento basso piuttosto che alto, in modo da ridurre la varianza complessiva del modello finale.

3. ***Stacking***

Nello stacking si lavora su più livelli diversi. Al primo livello, si addestrano in parallelo più classificatori deboli spesso non omogenei tra loro sul dataset di allenamento. Al secondo livello, il modello prende come input i valori in output del primo livello e apprende come meglio combinarli tra loro per ottimizzare la predizione. [20]

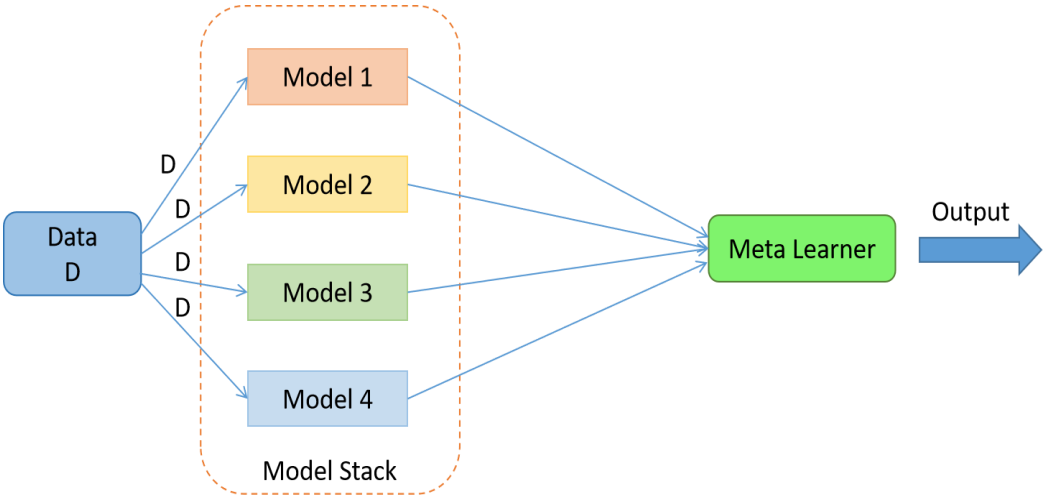

Figura 5: Stacking

Gli output generati dai modelli base possono essere numeri reali nel caso siano modelli di regressione oppure etichette nel caso siano classificazioni.

A differenza del bagging, tipicamente nello stacking si lavora con modelli diversi tra loro allenati sullo stesso dataset. A differenza del boosting, si utilizza un unico modello, noto come meta-modello, per imparare come meglio combinare i risultati provenienti dai modelli di primo livello (modelli base).

Lo stacking è molto indicato quando molti modelli base diversi sono adatti in diverso modo a trattare il dataset. In altre parole, gli errori commessi nella predizione dai vari modelli base sono tra loro scarsamente o per nulla correlati. Spesso, quindi, al primo livello si usano modelli complessi e diversi tra loro. Al contrario, il meta-modello tendenzialmente può essere un modello semplice e

nella maggioranza dei casi si va ad utilizzare la regressione lineare per algoritmi di regressione e la regressione logistica per algoritmi di classificazione.

Lo stacking è usato per migliorare le performance di predizione, ma non è detto che ci riesca[20]. Se il set di dati di allenamento non è abbastanza complesso da conferire un vantaggio pratico all'apprendimento e i modelli base non sono sufficientemente scorrelati tra loro, la capacità predittiva dello stacking potrebbe essere simile o inferiore rispetto ai modelli base. Data la complessità e l'onerosità a livello computazionale, in questi casi non vale la pena ricorrere allo stacking.

2.2. Apprendimento non supervisionato

Un algoritmo è non supervisionato quando viene utilizzato un dataset in cui non è presente un'etichetta. L'assenza dell'etichetta, su cui allenare l'algoritmo a riconoscere i cluster, determina approcci molto diversi da quelli usati nell'apprendimento non supervisionato

2.2.1. Clustering

L'analisi dei cluster è un insieme di tecniche di analisi multivariata dei dati volte alla selezione e raggruppamento di elementi omogenei in un insieme di dati (cluster). L'obbiettivo è avere dei dati che abbiano alta similarità interna al cluster e bassa similarità esterna al cluster. In molti approcci questa similarità, o meglio, dissimilarità, è concepita in termini di distanza in uno spazio multidimensionale. La bontà delle analisi ottenute dagli algoritmi di clustering dipende molto dalla scelta della metrica, e quindi da come è calcolata la distanza[21]. Gli algoritmi di clustering raggruppano gli elementi sulla base della loro distanza reciproca, e quindi l'appartenenza o meno a un insieme dipende da quanto l'elemento preso in esame è distante dall'insieme stesso.

Le tecniche di clustering si possono basare principalmente su due "filosofie":

- Dal basso verso l'alto (metodi aggregativi o bottom-up):

  Questa filosofia prevede che inizialmente tutti gli elementi siano considerati cluster a sé, e poi l'algoritmo provvede ad unire i cluster più vicini. L'algoritmo continua ad unire elementi al cluster fino ad ottenere un numero prefissato di cluster, oppure fino a che la distanza minima tra i cluster non supera un certo valore, o ancora in relazione ad un determinato criterio statistico prefissato.

- Dall'alto verso il basso (metodi divisivi o top-down):
  All'inizio tutti gli elementi sono un unico cluster, e poi l'algoritmo inizia a dividere il cluster in tanti cluster di dimensioni inferiori. Il criterio che guida la divisione è naturalmente quello di ottenere gruppi sempre più omogenei. L'algoritmo procede fino a che non viene soddisfatta una regola di arresto generalmente legata al raggiungimento di un numero prefissato di cluster.

Esistono varie classificazioni delle tecniche di clustering comunemente utilizzate. Una prima categorizzazione dipende dalla possibilità che un elemento possa o meno essere assegnato a più cluster[21]:

- clustering esclusivo: ogni elemento può essere assegnato ad uno e ad un solo gruppo. Quindi i cluster risultanti non possono avere elementi in comune. Questo approccio è detto anche hard clustering.
- clustering non-esclusivo, in cui un elemento può appartenere a più cluster con gradi di appartenenza diversi. Questo approccio è noto anche con il nome di soft clustering o fuzzy clustering, dal termine usato per indicare la logica fuzzy.

2.2.2. K-means

Il K-means è uno degli algoritmi di clustering più diffuso e più performante. Esso si basa sui cosiddetti centroidi. Il centroide è un punto appartenente allo spazio delle feature che media le distanze tra tutti i dati appartenenti al cluster ad esso associato. Rappresenta quindi una sorta di baricentro del cluster ed in generale, proprio per le sue caratteristiche, non è uno dei punti del dataset.

L'obiettivo che l'algoritmo si prepone è di minimizzare la distanza intra-cluster e di massimizzare la distanza inter-cluster[22]. Per farlo, si segue una procedura iterativa con una serie di passi ripetuti.

1. Si decide il numero K di cluster in cui si vuole dividere il dataset. Si individueranno in modo casuale K centroidi con la condizione che non coincidano e possibilmente siano abbastanza distanziati tra loro, in modo che l'algoritmo converga più rapidamente.
2. Si calcola la distanza di ogni punto del dataset rispetto ad ogni centroide e il punto viene assegnato al cluster collegato al centroide più vicino. Si crea così una partizione dello spazio in cui ogni punto può appartenere ad uno ed un solo cluster in quella che è detta tassellatura di Voronoi

3. Si ricalcola la posizione di ogni centroide facendo la media delle posizioni di tutti i punti del cluster associato
4. Si itera dal punto due finché non ci sarà alcun punto che cambia di cluster o quanto meno il numero di dati che si muove è inferiore ad una certa soglia percentuale. Tale condizione assicura una convergenza efficiente anche in casi di oscillazioni in fase di terminazione.

L'algoritmo K-means presenta due importanti limiti. Esso dà luogo a cluster di forma sempre poliedrica, in ossequio alla tassellatura di Voronoi. Non è quindi in grado di predire adeguatamente i cluster, quando questi abbiano distribuzioni non convesse nello spazio. Esso inoltre vuole in ingresso il numero k di cluster da utilizzare, che non è noto a priori e nella gran parte dei casi non è così immediato da individuare. Si utilizza generalmente un metodo maggiormente oggettivo per decidere questo numero, il cosiddetto metodo del gomito. [22]

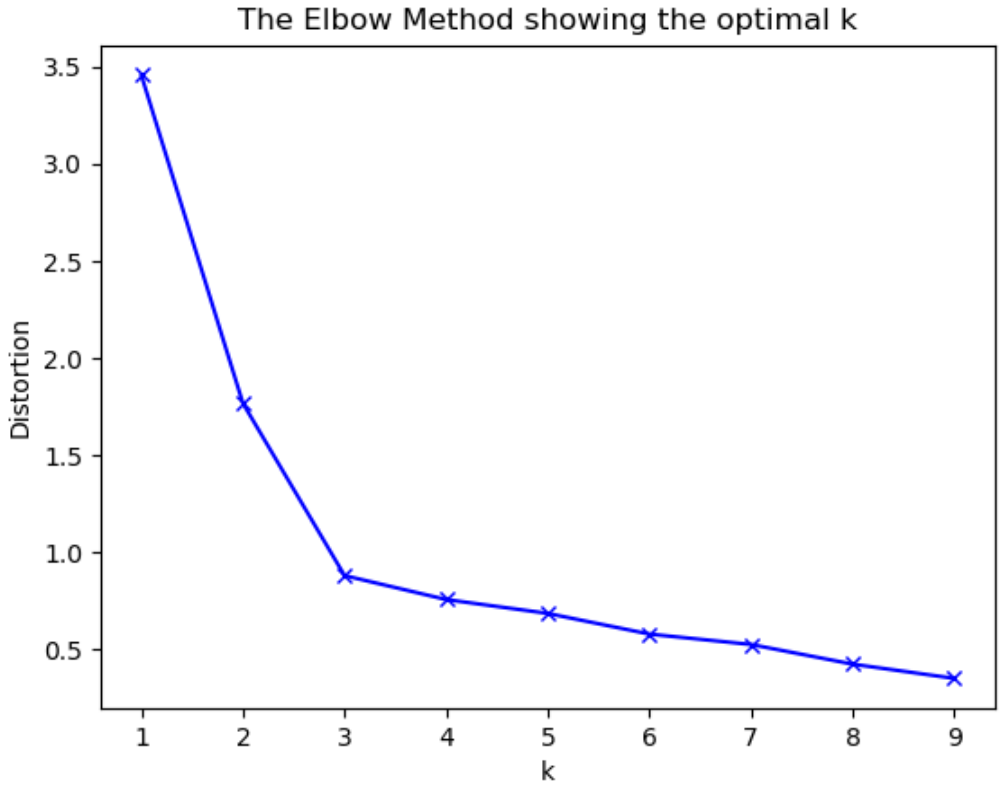

Figura 6: Ricerca del numero di cluster tramite il metodo del gomito

In pratica si itera il K-means per diversi valori di K ed ogni volta si calcola la somma delle distanze al quadrato tra ogni centroide ed i punti del proprio cluster.

Graficando i valori di K (asse orizzontale) e i valori della somma delle distanze al quadrato (asse verticale), si ottiene un grafico simile a quello in figura. Il numero ottimale di cluster è quello in cui è posizionato il gomito.

2.2.3. DBSCAN

Il DBSCAN è un algoritmo di machine learning basato sul concetto di densità particolarmente utile per individuare cluster dalle forme arbitrarie senza rimanere influenzato dal rumore di fondo.

Nel caso del DBSCAN non è necessario fornire il numero di cluster in quanto sarà l'algoritmo stesso a scoprire quanti cluster esistono in base a due iperparametri:

- epsilon o eps (ε): dato un punto *p*, identifica il raggio dell'intorno sferico in cui si verifica il numero di punti vicini a *p*
- min_pts: numero minimo di punti vicini a *p* all'interno del raggio *eps* affinché esso venga considerato un core point. In generale a questo iperparametro va dato un valore maggiore o uguale al numero dimensioni dataset più uno. Ad esempio, se il dataset è costituito da valori in due dimensioni (punti su un piano), allora min_pts sarà maggiore o uguale a tre. Valori maggiori sono usualmente migliori per data set con rumore.

Fondamentale sarà la scelta dei due iperparametri. Il risultato della clusterizzazione dipenderà dai parametri utilizzati per istanziare la classe DBSCAN. In base a come cambiano questi valori, si otterranno cluster diversi.[23]

I punti del dataset possono appartenere a tre categorie. Un core point è, come descritto precedentemente, un punto che ha attorno a lui (entro la distanza eps) un numero di altri punti almeno pari a min_pts. Un core point definisce un cluster.

Un border point è un punto attorno al quale si trovano un numero di punti minore di min_pts, ma uno di questi è un core point. Per questo motivo questo border point viene assegnato al cluster identificato dal core point a lui vicino.

Un noise point è un punto attorno al quale si trovano un numero di punti minore di min_pts e nessuno di questi è un core point. Questo significa che il noise point è a distanza maggiore di eps da qualunque core point. Per questo motivo non viene associato a nessun cluster e resta per conto suo, come outlier, cioè come valore anomalo.

Due punti si dicono connessi in base alla densità se esiste un cammino tra essi identificato da una catena di core-points. Sfruttando tale nozione di connessione, si individuano i vari

cluster raggruppando i core points connessi fra loro, con i relativi punti di bordo che descriveranno i confini dei loro cluster di appartenenza.
Al contrario degli algoritmi tradizionali, privi della nozione di outliers, il DBSCAN è in grado di riconoscerli evitando che questi finiscano assegnati ad un cluster pur non appartenendo a nessuno di essi. Spesso viene creato un cluster fittizio, denominato cluster spazzatura, ottenuto mettendo insieme tutti i punti anomali.
L'altro grosso vantaggio è che il DBSCAN riesce a gestire cluster non sferici, anche se, come tutti gli altri algoritmi di clustering, sfrutta una distanza (che quindi è calcolata su uno spazio circolare, sferico o in generale iper-sferico)[23]. Questo perché un core point può essere vicino ad un altro core point identificato in precedenza, il che porta alla fusione di più cluster in uno solo.
Lo svantaggio principale è invece DBSCAN non è in grado di classificare insiemi di dati con grandi differenze nelle densità, dato che la combinazione minPts-epsilon non può poi essere scelta in modo appropriato per tutti i cluster. Se si aumenta eps per adattarsi ai cluster a densità più basse, come conseguenza potrebbero essere accorpati cluster a densità più alta. Quando le differenze di densità sono molto elevate, questo algoritmo è chiaramente sconsigliato e si preferiscono altri approcci resolution-based o gerarchici.

2.2.4. E-M

L'algoritmo E-M (dall'inglese Expectation-Maximization) è un tipo di clustering non esclusivo. In esso si ipotizza che i pattern siano stati generati da una mistura di distribuzioni: ogni classe ha generato dati in accordo con una specifica distribuzione, ma al termine della generazione i pattern appaiono come prodotti da un'unica distribuzione multi-modale.
Obiettivo del clustering con E-M è risalire, a partire dai pattern del training set, ai parametri delle singole distribuzioni che li hanno generati. A tal fine si ipotizza nota la forma delle distribuzioni e si assume, per semplicità, che esse siano tutte dello stesso tipo. Il caso più frequente è quello di misture di distribuzioni multinormali (gaussiane), di cui si vogliono stimare i parametri di definizione.[24]
La stima dei parametri avviene secondo il criterio della stima della massima verosimiglianza (MLE). In generale la verosimiglianza corrisponde alla probabilità che i dati(osservazioni) siano stati generati da una certa distribuzione data. Per ragioni di stabilità numerica, al posto della verosimiglianza, si massimizza il suo logaritmo.

L'algoritmo è iterativo e si ripetono due fasi:

- Expectation: Si calcola la verosimiglianza per ogni punto del training set. In altre parole, si trova quale è la probabilità con cui i vari punti appartengono a ciascuna distribuzione (cluster)
- Maximization: Si massimizza la verosimiglianza trovando lo stimatore della massima verosimiglianza per medie e varianze di ciascuna distribuzione. Infine, si calcolano i parametri delle nuove distribuzioni

Questi due passi vengono eseguiti iterativamente (fino a convergenza), e può essere assimilato ad una versione probabilistica del K-Means.

2.2.5. Regole di associazione

Le regole di associazione sono un insieme di tecniche che permettono di scoprire le relazioni esistenti all'interno di un set di dati. Queste regole sono molto utilizzate in svariati ambiti:

- nel settore della grande distribuzione organizzata, o nel piccolo negozio al dettaglio, dove si devono scegliere come posizionare i prodotti tra i vari scaffali
- nel campo dell'istruzione, per scoprire regole e pattern nascosti nell'apprendimento degli studenti
- nel campo medico, ad esempio per eseguire analisi specifiche del paziente, in modo da identificare la combinazione delle caratteristiche del paziente e dei farmaci che portano a effetti collaterali negativi
- nelle aziende tecnologiche, in cui l'algoritmo Apriori è utilizzato da Amazon come sistema di raccomandazione di prodotti visualizzati e comprati da clienti simili e da Google per la funzionalità di completamento automatico in modo da capire l'associazione tra le parole cercate nel più famoso motore di ricerca

L'obiettivo di determinare una regola di associazione non è quello di estrarre le preferenze di un individuo, ma piuttosto quello di trovare relazioni tra un insieme di elementi di ogni transazione distinta[25]. Ad esempio, nel campo della grande distribuzione organizzata, se chi compra spesso il prodotto A, compra spesso anche il prodotto B (e viceversa), allora conviene posizionare vicini tra loro questi prodotti nelle scaffalature. Quindi, si potrebbe valutare di:

- Mettere i due item insieme in modo tale che quando un cliente acquista uno dei prodotti non deve andare lontano per acquistare l'altro prodotto
- Focalizzarsi sulle persone che acquistano uno dei prodotti in modo da avvalersi di campagne pubblicitarie per invogliarli ad acquistare l'altro
- Offrire sconti collettivi su questi prodotti se il cliente li acquista entrambi
- Impacchettare insieme sia A che B (ove possibile)

Esistono tre indicatori che possono essere utilizzati congiuntamente per misurare la forza dell'associazione ed estrarre le relative regole. Si consideri un paniere composto da X prodotti e da un numero totale di rilevazioni pari a N. Si ipotizzi di avere N=4 transazioni con X=2 prodotti acquistati: pane e burro.

Transazione	Prodotti acquistati
1	Pane
2	Pane, Burro
3	Burro
4	Pane

1. Il supporto è un'indicazione della frequenza con cui gli articoli compaiono nei dati. Matematicamente, il supporto di un prodotto A è la frazione rispetto al numero totale di transazioni in cui si trova il prodotto:

$$Supporto = \frac{Freq(A)}{N}$$

$$Supporto(Pane) = \frac{3}{4}$$

2. La confidenza è invece una misura che spiega come l'item B è comprato quando è comprato l'item A. Essa è utile a scremare le transazioni, aiutando a rispondere alla domanda: di tutte le transazioni che contengono A, quante contengono anche B? La formula di calcolo è la seguente:

$$Conf(A \rightarrow B) = \frac{Supporto(A \cup B)}{Supporto(A)}$$

$$Conf(Pane \rightarrow Burro) = \frac{1}{3}$$

Può dare alcuni spunti importanti, ma ha anche un grosso svantaggio. Tiene conto solo della popolarità del set di elementi A e non della popolarità di B. Se B è popolare come A, allora ci sarà una maggiore probabilità che una transazione contenente A conterrà anche B, aumentando così la confidenza. Difatti questo valore cambia se si inverte la relazione. La confidenza di B -> A, risulta differente rispetto a prima (poiché cambia la frequenza dell'item B)

$$Conf(Burro \rightarrow Pane) = \frac{1}{2}$$

Per ovviare a questo inconveniente esiste un'altra misura chiamata lift.

3. Il lift ci mostra come l'item B è comprato quando è comprato l'item A, mentre viene controllato quanto è popolare l'item B. In pratica è simile alla misura di confidenza, solo che a denominatore viene aggiunto anche il supporto del prodotto B. La formula è così rappresentata:

$$Lift = \frac{Supporto(A \cup B)}{Supporto(A) * Supprto(B)}$$

Il valore del lift per entrambi i prodotti sarà allora:

$$Lift(Pane, Burro) = \frac{1}{6}$$

Resta da chiarire come vengono calcolate le regole di associazione usando questi tre indicatori. L'algoritmo più usato è l'algoritmo Apriori. Per capirlo meglio, è utile ricorrere ad un esempio concreto, con sei transazioni e cinque prodotti.

Transazione	Lista di prodotti
1	Mela, Pera, Latte
2	Pera, Latte, Burro
3	Burro, Pane
4	Mela, Pera, Burro
5	Mela, Pera, Latte, Pane
6	Mela, Pera, Latte, Burro

L'algoritmo utilizza un approccio iterativo articolato in quattro step[25].

1. Inizialmente si sceglie una soglia di supporto minima. La soglia è moltiplicata per il numero di transazioni per valutare il valore del supporto minimo. Al di sotto di tale valore, gli item saranno considerati non frequenti. In questo caso, fissiamo al 50% la soglia, per cui il valore minimo sarà 3.
2. Vengono conteggiati gli item singolarmente in tutte le transazioni, ottenendo la seguente tabella

Prodotti	Conteggio
Mela	4
Pera	5
Latte	4
Burro	4
Pane	2

   Si verificano gli articoli che non superano il valore minimo fissato dal supporto e si effettua la cosiddetta "potatura"
3. Ogni prodotto viene associato agli altri presenti nella tabella precedente al netto della potatura, in modo da poter valutare le relazioni tra essi. Si ottengono i seguenti risultati

Prodotti	Conteggio
Mela, Pera	4
Mela, Latte	3
Mela, Burro	2
Pera, Latte	4
Pera, Burro	3
Latte, Burro	2

4. Anche in questo caso, gli item {Mela, Burro} e {Latte, Burro} non raggiungono il valore minimo e vengono esclusi. I restanti vengono ordinati in maniera decrescente

Prodotti	Conteggio
Mela, Pera	4
Pera, Latte	4
Mela, Latte	3
Pera, Burro	3

La procedura può poi essere ripetuta verificando le relazioni esistenti tra tre articoli, quattro articoli ecc. Vale però la regola per cui tutti i sottoinsiemi di un set di articoli frequente devono anche essere frequenti.

Prodotti	Conteggio
Mela, Pera, Latte	3

L'unica associazione frequente è {Mela, Pera, Latte} perché {Mela, Pera}, {Pera Latte} e {Mela Latte} sono a loro volta frequenti.

Uno dei principali fattori che limita l'utilizzo di quest'algoritmo è che risulta computazionalmente costoso[25]. Anche se l'algoritmo Apriori riduce il numero di itemset candidati da considerare, questo numero potrebbe comunque essere enorme quando gli inventari dei negozi sono grandi o quando la soglia di supporto è bassa.
Una seconda limitazione è dovuta alla creazione di associazioni spurie, ossia associazioni che appaiono collegate casualmente ma che in realtà non lo sono[25].
L'analisi di grandi inventari comporterebbe più configurazioni di set di elementi e potrebbe essere necessario abbassare la soglia di supporto per rilevare determinate associazioni.
Tuttavia, abbassando la soglia di supporto potrebbe anche aumentare il numero di associazioni spurie rilevate. Per garantire che le associazioni identificate siano generalizzabili, potrebbero essere prima distillate da un set di dati di formazione, e sucessivamente valutate in un set di dati di test separato.

2.3. Classificazione di serie temporali

Le serie temporali (o serie storiche) necessitano di un approccio ad hoc per poter essere affrontate mediante algoritmi di machine learning. Al loro interno, infatti, i dati sono ordinati rispetto alla variabile tempo, cioè esprimono la dinamica di un certo fenomeno nel tempo. Le serie storiche permettono così di cogliere l'andamento futuro del fenomeno sotto osservazione, scomponendolo in una serie di componenti:

- Componente tendenziale (o trend): mostra l'andamento di lungo periodo del fenomeno. I metodi più utilizzati per determinare il trend sono il metodo dei minimi quadrati e il metodo delle medie mobili
- Componente ciclica: si manifesta con fluttuazioni periodiche o non periodiche attorno alla curva di trend
- Componente stagionale: determina variazioni che avvengono negli stessi mesi di anni successivi
- Componente casuale: è l'insieme delle piccole oscillazioni dovute ad eventi casuali
- Componente occasionale: consiste in fenomeni rari, ma ad alto impatto, capaci di interrompere o addirittura invertire dei trend. Si pensi a pandemie, guerre, etc.

Le serie storiche possono essere di due tipi:

I. Deterministico, se i valori della variabile possono essere esattamente determinati sulla base dei valori precedenti
II. Stocastico, se i valori delle variabili possono essere determinati sulla base dei valori precedenti solo in misura parziale. La maggioranza delle serie storiche è di tipo stocastico e non è possibile elaborare previsioni prive di errore

Una soluzione comune, ma problematica, alla classificazione delle serie temporali consiste nel trattare ogni punto temporale come una caratteristica separata e applicare direttamente un algoritmo di apprendimento standard. Questo approccio ha un limite piuttosto evidente: l'algoritmo ignora le informazioni contenute nell'ordine temporale dei dati. [26]

Si può più correttamente ricorrere all'uso delle reti neurali, in particolare le reti neurali ricorrenti, in grado di estrarre le caratteristiche dinamiche delle serie temporali, quelle cioè varianti rispetto al tempo. Tuttavia, le reti neurali, accanto ad una spiccata efficacia data da un'alta capacità di generalizzazione, presentano alcune sfide che ne rendono a volte complicata l'applicazione, come la necessità di lavorare su una mole di dati

massiccia per massimizzare i risultati di predizione, le ingenti risorse di calcolo necessarie, la selezione di un'architettura efficiente compresa nel tuning degli iperparametri.[26]
Ci sono poi alcuni algoritmi di machine learning che si dedicano alla classificazione delle serie temporali, riadattando algoritmi classici.

2.3.1. Classificatori distance-based

Usano come metrica la distanza per determinare l'appartenenza di un elemento ad una classe. Sono un adattamento del KNN per le serie temporali, in cui la metrica di distanza Euclidea viene sostituita con la metrica di deformazione dinamica del tempo (in inglese dynamic time warping o DTW). Quest'ultima consente di misurare la somiglianza tra due o più sequenze temporali non esattamente allineate in termini di tempo, velocità o lunghezza.[27] Ad esempio, se due serie temporali considerate sono altamente correlate, ma sono di diversa lunghezza e sfasate temporalmente tra loro, la distanza euclidea non riuscirebbe a cogliere la correlazione, al contrario della metrica DTW.
Questo è possibile poiché mentre la distanza euclidea è calcolata come radice quadrata della somma delle distanze al quadrato tra ogni elemento della serie temporale ed il suo corrispettivo delle altre serie temporali, la distanza DTW è calcolata come radice quadrata della somma delle distanze al quadrato tra ogni elemento di una serie temporale e ogni elemento delle altre serie temporali.

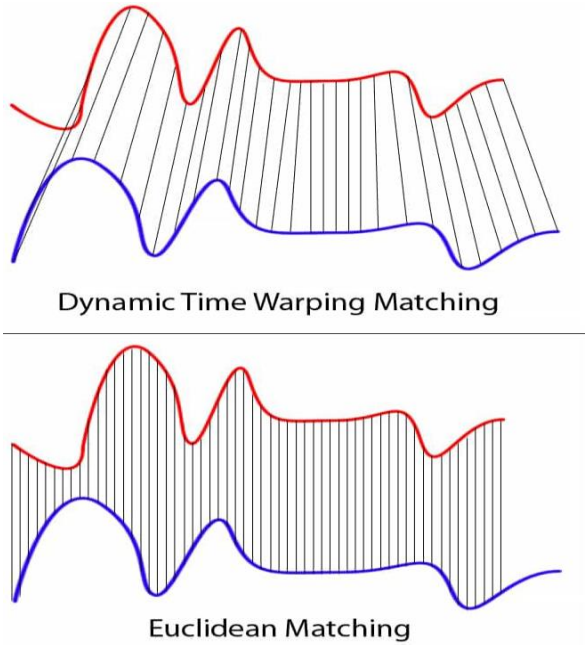

Figura 7: Dynamic Time warping

Questo approccio è spesso usato come benchmark per valutare tutti gli altri algoritmi di classificazione delle serie temporali perché è semplice e robusto, ma al tempo stesso richiede un elevato sforzo di calcolo, dovendo confrontare ogni elemento con tutti gli altri elementi del set di allenamento, e può performare male quando il rumore nei dati è elevato.[27]

2.3.2. Classificatori interval-based

Basano il processo di classificazione sulle informazioni contenute dalle serie temporali adottando un approccio rivisitato degli alberi decisionali. L'algoritmo si articola su quattro passi:

I. Si dividono le serie temporali in intervalli casuali, con posizioni di inizio e lunghezze casuali
II. Per ogni intervallo si estraggono le feature di riepilogo (media, deviazione standard, ecc) creando un vettore di feature
III. Si addestra l'albero decisionale sulle feature estratte
IV. Si ripetono i primi tre passi finché il numero fissato di alberi è stato creato

Studi sperimentali hanno dimostrato che i classificatori basati su intervalli temporali possono performare meglio di altri approcci come il DTW e sono anche molto efficienti dal punto di vista computazionale. [28]

2.3.3. Classificatori shapelet-based

Basano il processo di classificazione sull'individuazione di sotto-sequenze di serie temporali che sono rappresentative delle classi. Consentono quindi di riconoscere le somiglianze localizzate tra serie della stessa classe. I classificatori che si basano sull'individuazione di sotto-sequenze sono alla ricerca di sotto-sequenze con alta capacità discriminante. La presenza di una sotto-sequenza più discriminante di un'altra rende più probabile l'appartenenza alla rispettiva classe[29].

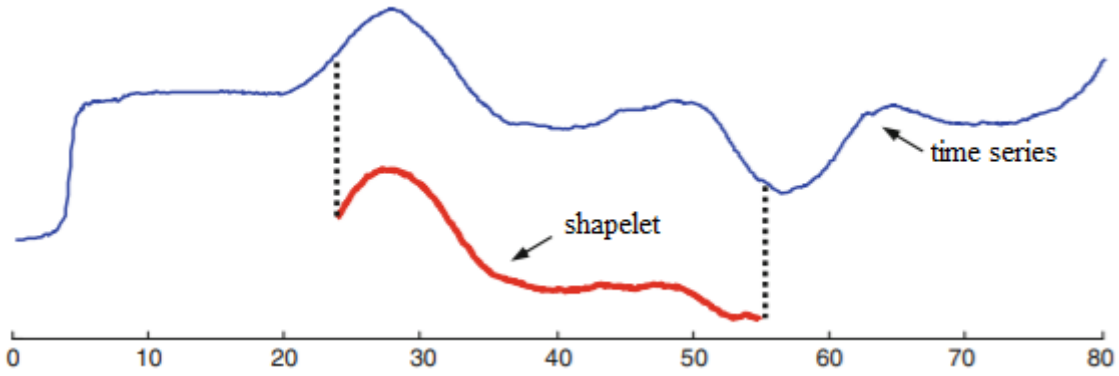

Figura 8:Esempio di sottosequenza in una serie temporale

L'algoritmo prima di tutto identifica k sotto-sequenze nel dataset. Successivamente, calcola k feature per il nuovo dataset come la distanza della serie temporale da tutte le sotto-sequenze con una colonna di valori per ogni sottosequenza. Il nuovo dataset può essere trattato infine tramite classici algoritmi di classificazione.

3. RACCOLTA E ANALISI DEL CAMPIONE

Questo capitolo è dedicato all'analisi delle prime tre fasi del KDD. Come si è già accennato nel capitolo 2, il KDD (knolwedge discovery from data) è quel processo che si articola in cinque fasi e consente di estrarre conoscenza a partire da una base dati grezza. Le prime tre fasi del KDD (selezione, pretrattamento e trasformazione dei dati) scaturiscono nella creazione della base dati pulita, pronta per essere data in pasto agli algoritmi di machine learning nel processo di data mining, che sarà oggetto del capitolo 4.

Prima di questa analisi, si dedicherà un breve paragrafo all'analisi del settore del legno/mobile-arredamento, che è il settore da cui sono stati estratti i bilanci oggetto di analisi. L'intento è quello di capire l'importanza relativa che questo settore occupa nell'economia italiana e di coglierne le maggiori peculiarità.

3.1. Analisi del settore del legno/mobile-arredamento

Il settore del Legno-Arredo è uno dei settori più storici e trainanti del made in Italy, rientrando a pieno titolo all'interno di quelle eccellenze del Paese molto apprezzate all'estero per via dell'elevata qualità delle maestranze e l'alta dose di artigianalità. Questo settore fa parte delle cosiddette quattro A, cioè i quattro comparti di punta dell'industria manifatturiera, che rendono l'Italia famosa e competitiva nel mondo (abbigliamento, arredamento, alimentare, automazione).

Massima espressione del made in Italy, grazie ad elementi quali la qualità dei materiali, l'accuratezza delle lavorazioni e il design unico, il comparto del legno-arredo contribuisce alle eccellenze dei prodotti finali di consumo esportati dal nostro Paese, riconducibili sotto il concetto di "bello e ben fatto". Negli ultimi anni il comparto legno-arredo è stato protagonista di una fusione sempre più integrata con un'altra eccellenza del Paese, la moda. Giganti di spicco come Armani, Fendi e Gucci, tra gli altri, hanno lanciato la propria linea d'arredo per un prodotto sempre più esclusivo e di alta gamma.

La filiera del legno-arredo è ampia e articolata, comprendendo tutte le attività che permettono il passaggio dalla materia prima, il legno appunto, al prodotto finito, nelle sue diverse forme, da quelle più semplici fino ai prodotti di design.

La catena di produzione si articola in due macroaree, a monte il sistema del legno, a valle il sistema dell'arredamento:

- il settore del legno comprende il commercio dello stesso e le lavorazioni intermedie, ma anche i prodotti e le finiture per l'edilizia, così come tutti gli imballaggi lignei, il comparto del sughero e la produzione di tende e pannelli.
- il settore dell'arredamento racchiude tutte le tipologie di mobili, dagli arredi per la casa e il bagno, agli elementi per gli uffici, le collettività e gli spazi commerciali, fino all'outdoor e gli elementi di illuminazione.

A queste due macroaree si aggiunge tutto il comparto della distribuzione, sia nella forma tradizionale, sia attraverso la Grande Distribuzione Organizzata (GDO) l'e-Commerce.

Figura 9:Catena produttiva dei settori del Legno e dell'Arredo

Il mercato globale dell'arredamento ha raggiunto un valore di oltre 730 miliardi di dollari nel 2019 ed è stato protagonista negli anni Dieci di una dinamica fortemente positiva.[30] L'industria del mobile europea rappresenta oltre un quarto della produzione globale ed è leader mondiale per il segmento di fascia alta. Quasi due prodotti di arredamento d'alta gamma su tre venduti nel mondo sono fabbricati in Europa. L'Italia è esponente di punta del mercato comunitario, primo paese con un valore della produzione superiore ai 23 miliardi di euro per il solo comparto dell'arredamento. [30]

La filiera italiana del legno e arredo si contraddistingue per la vocazione all'export, soprattutto extra UE (46%) e la sostenibilità. La produzione di mobili italiana è prima in Europa per economia del riciclo e ultima per emissioni climalteranti.

La filiera del Legno-Arredo italiana ha una profonda tradizione nella manifattura italiana ed è molto radicata sul territorio. Essa si basa su importanti distretti industriali con fortissima prevalenza di piccole e medie imprese con alta vocazione all'export, che esprimono forte vitalità in termini di occupazione e fatturato sul pil delle regioni che li ospitano. Tuttavia, la struttura produttiva non è omogenea all'interno del Paese, nel nord della penisola si concentrano, infatti, oltre i tre quarti delle aziende dell'intero macrosettore Le due regioni più importanti per questa filiera sono il Veneto e la Lombardia, che rappresentano la quota maggiore della produzione italiana del Legno-Arredo ed esportano da sole quasi la metà dei mobili destinati alla Germania, alla Francia e agli USA, i tre principali mercati di riferimento. [31]

In Lombardia i distretti più importanti sono quello del mobile della Brianza e quello del Pannello del Mantovano. Nella regione il 10% delle imprese manifatturiere appartiene alla filiera e genera un saldo commerciale molto elevato di 2,1 Miliardi. In Veneto, invece, è ubicato il distretto di Treviso, il più grande in Italia, con una quota elevatissima del 56% del totale della produzione nazionale nel segmento dell'arredo. Al terzo posto per fatturato si colloca il Friuli-Venezia Giulia con una quota di 3,5 Miliardi di euro, che si distingue per la presenza di aziende di dimensioni più grandi per numero di addetti e fatturato generato. Nella regione ben il 64% delle imprese censite appartiene al settore del Legno-Arredo con un contributo sul PIL regionale del 15%. Le altre tre regioni che si distinguono per la specializzazione sul Legno-Arredo, pur raggiungendo volumi produttivi molto inferiori, sono Emilia-Romagna, Marche e Puglia. La Puglia è la regione che più si distingue fra quelle del Meridione, con un fatturato di 1,3 miliardi, concentrato per circa l'80% del totale nel distretto degli imbottiti di Bari. Le Marche realizzano un fatturato ancora più alto di 2,5 Miliardi (circa il 10% del totale manifatturiero), e hanno sperimentato negli ultimi anni un forte incremento dell'export, in aumento del 65% dal 2009. Più del 60**%** del fatturato prodotto nel settore arredo viene dalla provincia di Pesaro Urbino. [31]

3.2. Selezione del campione

Il campione delle serie storiche dei bilanci delle aziende del settore del Legno-Arredo che sono state utilizzate in questa tesi è stato estratto dal Database AIDA, a cui è garantito l'accesso gratuito per gli studenti del Politecnico di Torino.

AIDA è la banca dati, realizzata e distribuita da Bureau van Dijk S.p.A., azienda di proprietà di Moody's, contenente i bilanci, i dati anagrafici e merceologici di tutte le società di capitale italiane attive e fallite (ad esclusione di Banche, Assicurazioni ed Enti pubblici). La banca dati include il programma di ricerca, consultazione ed esportazione dei dati, con cui si può accedere a:

- Informazioni anagrafiche e finanziarie dettagliate su circa 980.000 imprese aggiornate all'ultimo anno disponibile;
- Serie Storica di bilanci contenuti fino a 10 anni;
- Dati su Azionariato e Partecipazioni delle società fino al 10° livello
- Esponenti
- Oltre 400 chiavi di ricerca a disposizione
- Possibilità di effettuare ricerche attraverso classificazione per codici attività nazionali (ATECO) ed Internazionali (NACE, NAICS, US SIC, UK SIC).

Per ogni società sono estratte le seguenti informazioni (oltre ai dati di bilancio):

- ragione sociale;
- partitiva IVA;
- regione;
- codice ATECO e descrizione dell'attività svolta;
- stato giuridico ("ditta attiva", "ditta in liquidazione", "ditta in fallimento", "ditta sospesa", "ditta inattiva", "ditta cessata", "ditta cessata per trasferimento")
- l'eventuale procedura

Su questa base è stato calcolato un flag 0/1, in cui lo zero individua le aziende sane e l'uno le aziende incorse in fallimento o in liquidazione. Le procedure che determinano il fallimento sono:

- Concordato preventivo
- Fallimento
- Amministrazione giudiziaria
- Accordo di ristrutturazione dei debiti
- Chiusura del fallimento
- Liquidazione giudiziaria
- Stato di insolvenza
- Sequestro giudiziario
- Concordato fallimentare
- Amministrazione controllata

- Cancellazione per comunicazione piano di riparto
- Amministrazione straordinaria
- Chiusura per fallimento o liquidazione
- Decreto cancellazione tribunale
- Liquidazione coatta amministrativa
- Scioglimento per atto dell'autorità
- Sequestro conservativo di quote
- Bancarotta

Le procedure che determinano la liquidazione sono:
- Liquidazione volontaria
- Scioglimento e liquidazione
- Scioglimento
- Chiusura della liquidazione
- Chiusura dell'unità locale
- Cessazione di ogni attività
- Cancellata d'ufficio ai sensi art. 2490 c.c. (bilancio di liquidazione)
- Liquidazione
- Scioglimento e messa in liquidazione
- Chiusura per liquidazione
- Scioglimento senza messa in liquidazione
- Cessazione delle attività nella provincia
- Cessazione d'ufficio

Siccome il campione estratto non è bilanciato, data la prevalenza del numero di imprese sane sul numero di imprese fallite o in liquidazione, e, in particolare, il numero di imprese fallite rappresenta solo il 5% del campione estratto, a fronte di un 33% di imprese in liquidazione, si è deciso di accorpare in un'unica classe le imprese fallite e quelle in liquidazione

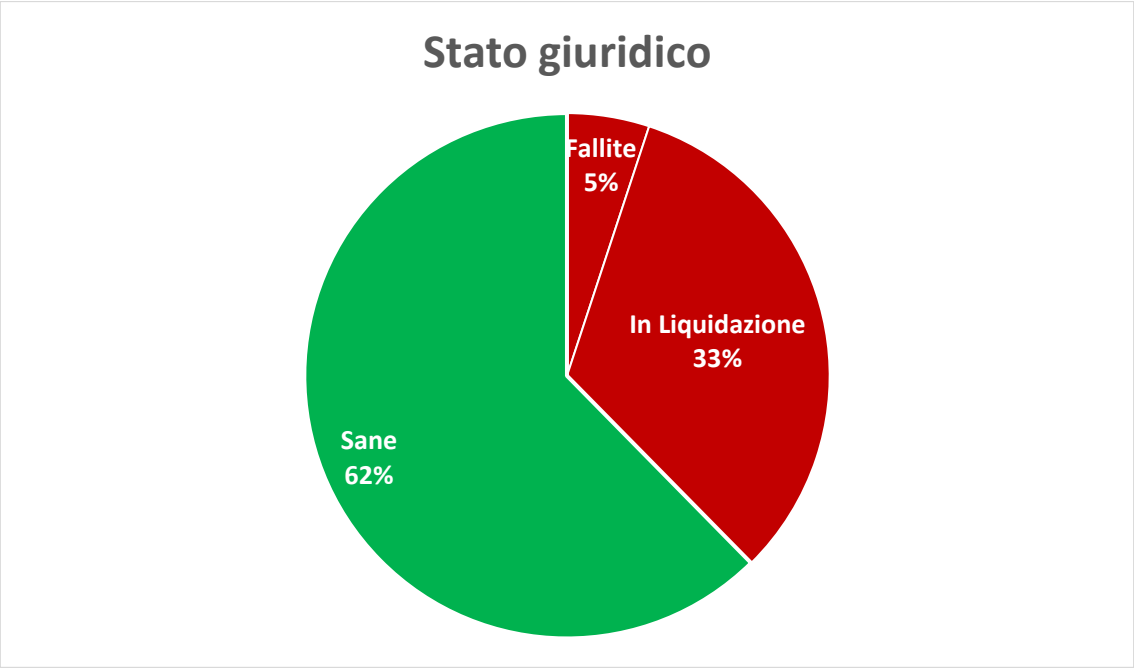

Figura 10: Distribuzione percenutale di aziende sane e anomale nel campione

Sommate insieme, le imprese fallite o in liquidazione rappresentano il 38% del campione, a fronte del 62% di imprese sane. In termini assoluti, il campione è formato da 121 aziende fallite, 778 aziende in liquidazione e 1490 aziende sane, per un totale di 2389 aziende e 20980 osservazioni disponibili.
Ogni azienda ha un numero variabile di bilanci disponibili, secondo i numeri visibili in tabella.

Num.bilanci disponibili	**Numero aziende**
3	13
4	157
5	145
6	135
7	100
8	100
9	111
10	1628

I bilanci estratti fanno riferimento agli anni compresi tra il 2000 e il 2019, con una netta prevalenza degli anni successivi al 2000, come si può notare nel grafico.

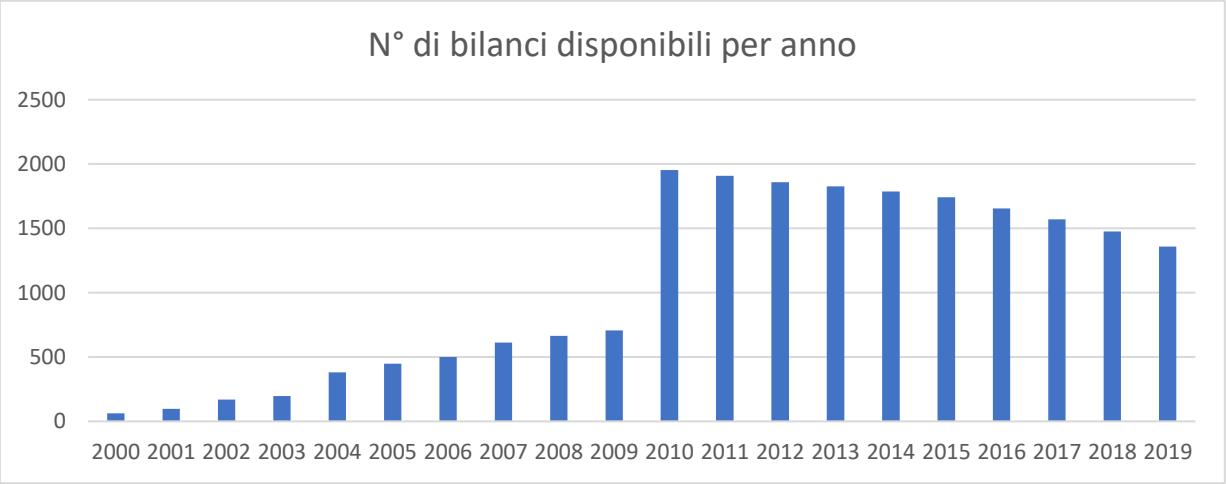

Figura 11:Numero di bilanci disponibili per anno delle aziende del campione

Questo si giustifica con il fatto che, per le società ancora attive, il Database Aida fornisce al massimo gli ultimi dieci bilanci disponibili. Si comprende, quindi, come mostrato nel grafico seguente, che la prevalenza di bilanci successivi al 2009 sia dovuta solo alle società sane (in verde), quindi ancora attive, piuttosto che alle società fallite o in liquidazione (in rosso), che hanno cessato l'attività.

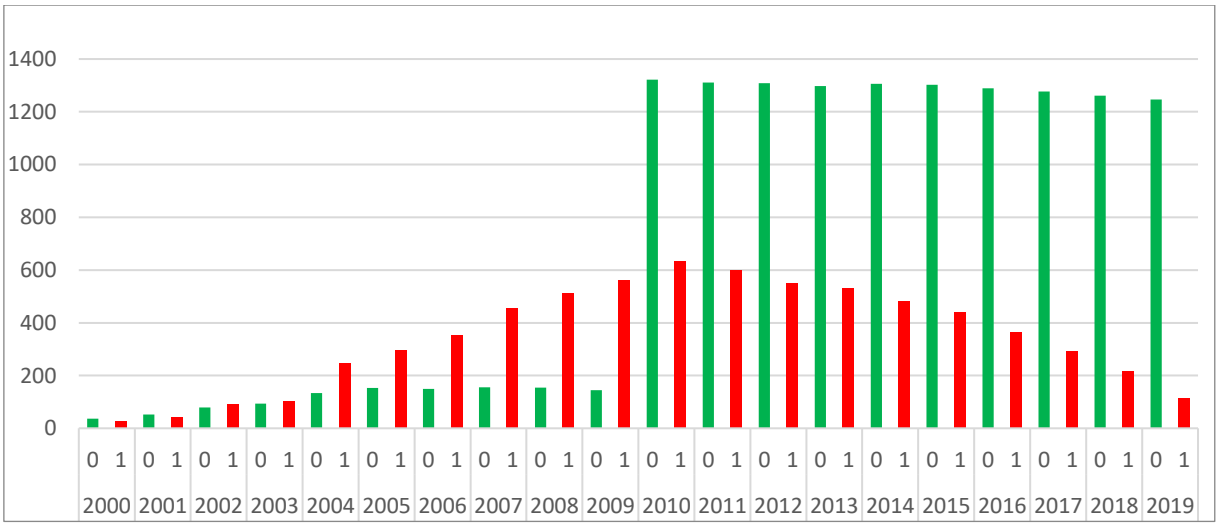

Figura 12: Ripartizione dei bilanci disponibili tra aziende sane e anomale

Il campione estratto rispetta i numeri citati dello studio di Federlegno Arredo, da cui si evince una nettissima prevalenza delle imprese con sede legale nel Nord del Paese.

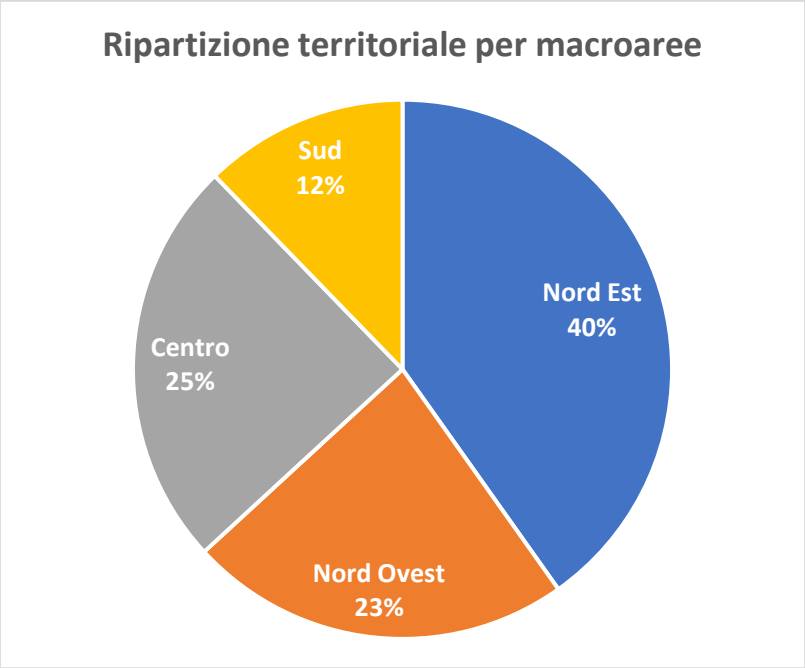

Figura 13:Ripartizione territoriale per macroaree

La predominanza delle aziende del Nord nel campione è ancor più evidente guardando alla ripartizione regionale, dove quattro delle prime cinque regioni sono nel Nord.

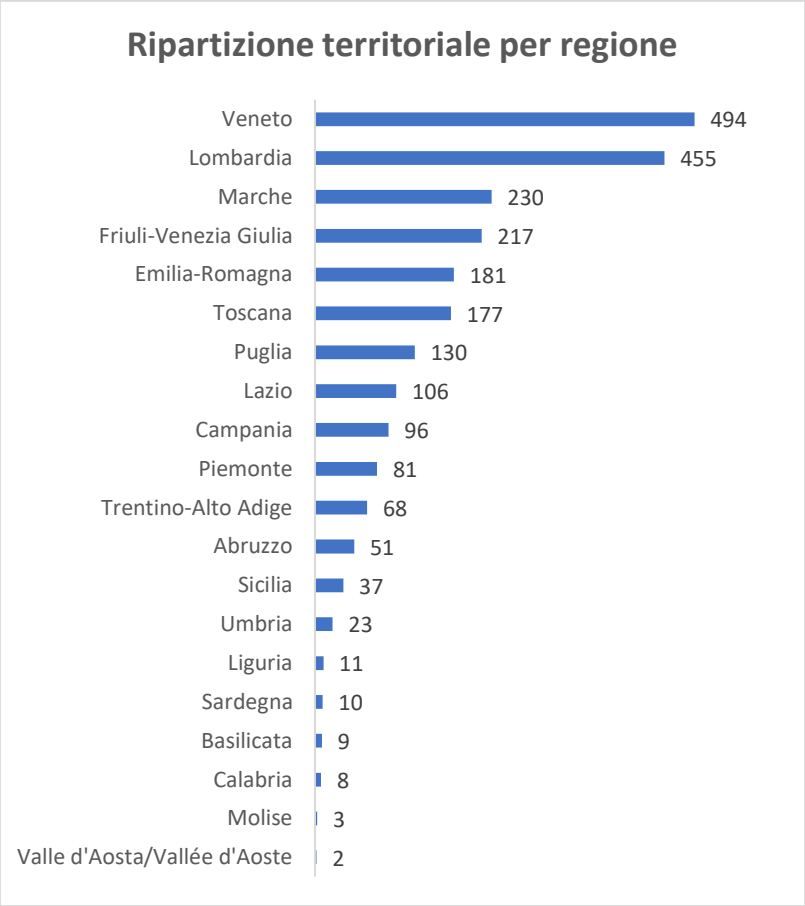

Figura 14:Ripartizione territoriale per regione

Questo dato è la conferma di come questo settore sia allocato in Italia in maniera altamente asimmetrica, concentrandosi in un certo numero di distretti, quasi tutti ubicati nelle regioni del Nord Est e in Lombardia.

Le aziende del campione che hanno avviato una procedura di fallimento o liquidazione, sono aziende generalmente mature, attive sul mercato da più di cinque anni, come visibile nel grafico seguente:

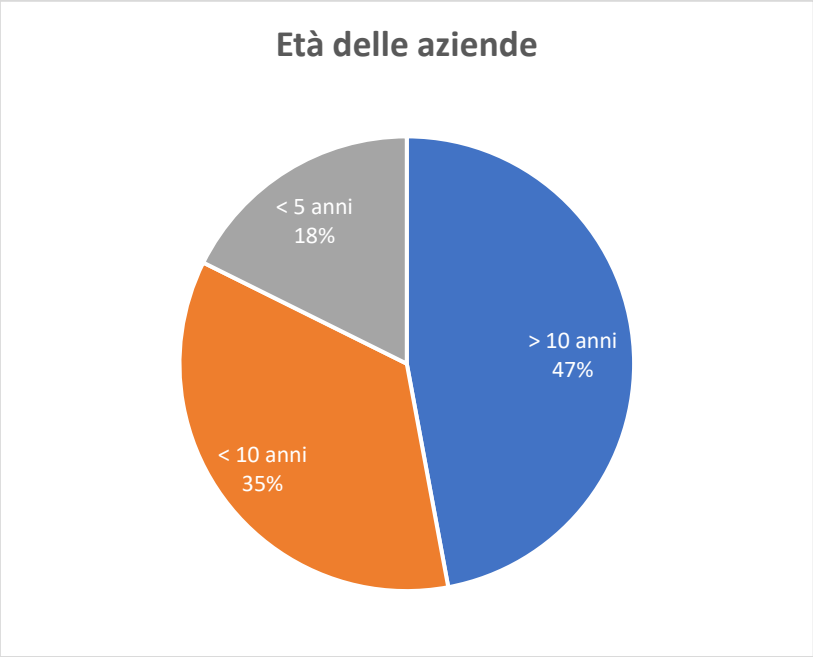

Figura 15:Distribuzione per età delle aziende del campione

Questa scelta è dovuta al fatto che in questa tesi si è voluto attuare un approccio di classificazione basato sulle serie temporali, che necessita di un numero di bilanci minimo, da cui ne deriva che sono state escluse le aziende incorse in fallimento entro i tre anni dalla fondazione. Ciò va contro l'evidenza empirica per cui il fenomeno del fallimento è più probabile nei primissimi anni dalla fondazione della stessa.

La forma giuridica prevalente nel campione è la S.R.L., come si può vedere nel grafico seguente.

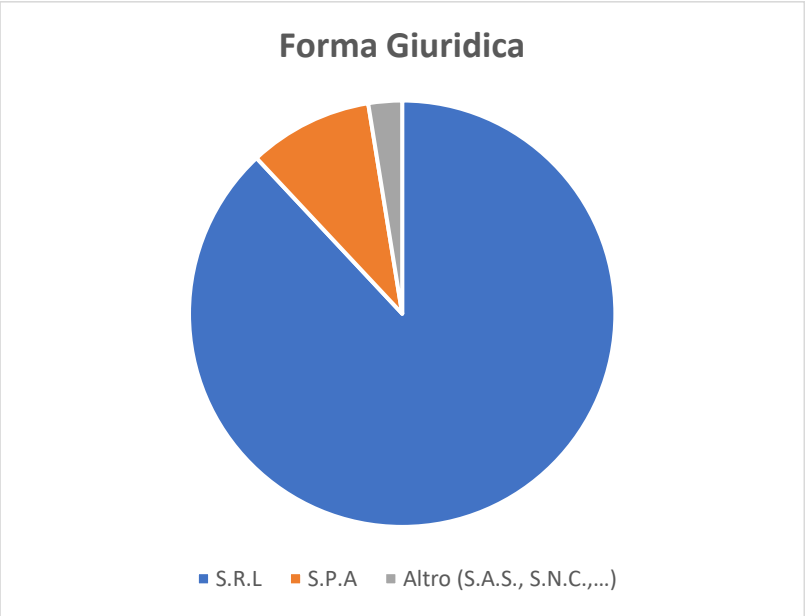

Figura 16: Distribuzione per forma giuridica delle aziende del campione

3.3. Pretrattamento e trasformazione

I bilanci scaricati da AIDA presentavano alcuni errori, per cui si è proceduto ad una pulizia, sfruttando una serie di test di controllo. Innanzitutto, si sono eliminati tutti i valori n.d., ossia non disponibili, rimpiazzandoli con uno zero e si è verificato che l'attivo di bilancio fosse strettamente maggiore di zero. In molti casi però, restavano alcuni bilanci in cui il valore dell'attivo non coincideva con il valore del passivo o in cui la somma dei valori delle singole voci di bilancio non coincideva con il valore delle classi o sottoclassi di appartenenza. Molto spesso, cioè, chi aveva redatto il bilancio non aveva aggregato correttamente le singole voci del bilancio nelle classi. Dato che gli indici di bilancio attingono dalle classi, l'eventuale scorrettezza dei valori delle classi avrebbe inficiato i risultati dell'analisi.

L'analisi di bilancio è utile per determinare lo "stato di salute" dell'impresa e il suo posizionamento rispetto a tre equilibri:

1. economico: la capacità dell'impresa di produrre reddito, per un tempo sufficientemente ampio, in grado di remunerare tutti i fattori della produzione
2. patrimoniale: l'equilibrio fra attività e passività

3. finanziario: la capacità di un'impresa di rispondere in modo tempestivo agli impegni assunti

A questo punto, si è proceduto al calcolo degli indici da dare in input agli algoritmi di machine learning. Purtroppo, all'interno del campione, il numero di bilanci dettagliati (4613) è molto inferiore rispetto al numero dei bilanci abbreviati (16367). Per questo motivo, alcuni indicatori che richiedevano la compilazione di determinate voci di bilancio, non presenti nei bilanci abbreviati, non sono stati calcolati.
Gli indici utilizzati sono suddivisi in una serie di macrocategorie:

- Indici di sviluppo (5): pongono in risalto la variazione annua delle grandezze fondamentali che qualificano la crescita dimensionale di un'azienda. I valori espressi da tali indici consentono di evidenziare condizioni che caratterizzano lo sviluppo dell'attività ed il rafforzamento dell'iniziativa sociale.
- Indici di redditività (9): misurano l'attitudine di un'impresa a produrre un reddito sufficiente a coprire i costi e a generare profitti, in misura tale da mantenere un equilibrio che giustifichi gli investimenti effettuati
- Indici di produttività e struttura operativa: esprimono il grado di efficienza dei fattori produttivi
- Liquidità e struttura finanziaria: servono a verificare in che misura la combinazione impieghi-fonti è in grado di produrre nel breve periodo flussi monetari equilibrati, ossia tali da consentire di far fronte in ogni momento agli impegni di uscita che la gestione richiede
- Indici di allerta per le crisi d'impresa: sono gli indicatori spia di un futuro possibile dissesto aziendale

Gli indici così suddivisi hanno subito un pretrattamento in cui le osservazioni con un valore minore del quinto percentile o maggiori del novantacinquesimo percentile sono state forzate ad assumere i valori di quei percentili. Questo processo è servito ad evitare che la presenza di outliers andasse ad influenzare il risultato di analisi.
In seguito, per ogni indicatore, i valori delle singole osservazioni sono stati normalizzati assumendo un valore compreso tra 0 e 1.
Dopo queste operazioni preparatorie, la base dati è pronta per poter essere analizzata tramite le tecniche di machine learning.

4. DATA MINING E INTERPRETAZIONE DEI RISULTATI

Il database estratto e pretrattato presenta una elevata numerosità di dati (2389 serie storiche) e si caratterizza per la presenza dell'etichetta per ogni serie, che per ogni azienda comunica se nel periodo di riferimento è fallita o è stata liquidata. Queste sono le condizioni ideali per portare a termine un'analisi predittiva tramite algoritmi di machine learning supervisionati. Al fine di rendere il lavoro di tesi completo, si è deciso di utilizzare anche un algoritmo non supervisionato, in modo da poter evidenziare similitudini e differenze in termini di processo di data mining e di risultati di classificazione. Entrambi questi approcci sono stati portati avanti considerando per ogni azienda sana l'ultimo anno a disposizione e per ogni azienda anomala l'anno precedente all'evento default/liquidazione.

Come ultima opzione, si utilizza poi un approccio più innovativo, in cui si sfruttano algoritmi di machine learning capaci di analizzare per ogni azienda l'intera serie storica ricavandone l'evoluzione temporale.

Questi tre differenti approcci danno origine a risultati molto diversi tra loro e vengono confrontati nell'ultima parte del capitolo, allo scopo di evidenziare quale meglio si adatta al database in analisi in termini di precisione e robustezza dei risultati.

Nell'applicazione di tutte e tre le categorie di algoritmi adottati, per prima cosa si è proceduto a bilanciare il numero di esempi per classe. Come già detto nel capitolo 3, il numero di bilanci di aziende sane del campione è maggiore della somma del numero dei bilanci di aziende fallite o in liquidazione. In dettaglio, il campione è composto da 121 aziende fallite e 778 aziende in liquidazione, per un totale di 899 aziende anomale opposte a 1490 aziende sane. Per questa ragione si è proceduto ad una esclusione dei bilanci delle aziende sane eccedenti il numero di aziende anomale, nel processo noto come Undersampling. Questo processo consente il ribilanciamento delle classi, che di conseguenza concorreranno nella stessa percentuale (50%) all'addestramento dell'algoritmo. Si è scelto di adottare un undersampling randomico, in cui cioè la scelta dei bilanci delle aziende sane da dare in pasto all'algoritmo sia del tutto casuale.

4.1. Algoritmi supervisionati

Nell'applicazione degli algoritmi supervisionati, dopo aver importato il file completo, si è proceduto alla selezione dell'ultimo anno per le aziende sane e del penultimo anno per le aziende anomale.

Dopodiché il campione è stato diviso in due gruppi. Il primo gruppo è quello di train, quello cioè in cui si è allenato l'algoritmo fornendo delle feature in input e delle etichette in output da predire. Il secondo gruppo è quello di test, in cui l'algoritmo ha ricevuto in input le feature, restando però all'oscuro della classe di appartenenza dell'elemento del campione. Nella fase di test è stato quindi possibile calcolare la reale precisione dell'algoritmo verificando quanti bilanci sono stati correttamente classificati sul totale testato. Analogamente si è potuto ricavare altre misure significative per la valutazione delle performance, come la recall o la F1-score. Nei casi elencati si tiene conto della quantità di falsi positivi e di falsi negativi registrati e tramite la F1 si premiano gli algoritmi tanto più il numero di falsi positivi e il numero di falsi negativi è basso e bilanciato.

Oltre alle fasi di train e test c'è una terza fase, nota come fase di Validation. In questa fase, l'obbiettivo è quello di assicurarsi che l'allenamento abbia funzionato come voluto. In questo caso, dalla letteratura è emerso che il modo migliore per rendere le previsioni dell'algoritmo generalizzabili fosse quello di ripetere più volte la divisione in gruppi di train e test, variando ogni volta i parametri di inizializzazione. I risultati ottenuti sono quindi una media dei valori ottenuti da ogni singolo train-test split e quindi non soffrono della variabilità che potrebbe essere causata da un'inizializzazione particolarmente favorevole o sfavorevole.

Nella scelta delle dimensioni, si è scelto di dedicare il 75% del campione al gruppo di training e il restante 25% del campione al gruppo di test.

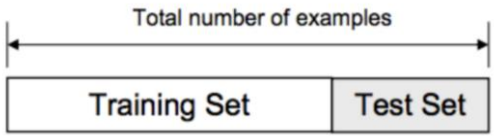

Figura 17:Ripartizione del campione tra gruppo di Allenamento e gruppo di Test

La definizione di queste percentuali si basa su prassi consolidate nella letteratura, utilizzate come strumento per prevenire o, perlomeno, ridurre al minimo l'overfitting.

Overfitting significa che il modello si è addestrato "troppo bene" ai dati di training e ora è troppo vicino al set di dati di addestramento. Questo di solito accade quando il modello è troppo complesso (cioè troppe caratteristiche / variabili rispetto al numero di osservazioni). Questo modello sarà appiattito sui dati di addestramento, ma probabilmente perderà di capacità di generalizzazione su dati nuovi o non addestrati; quindi, non consentirà di fare efficace inferenza su nuovi dati. Fondamentalmente, quando si ha overfitting, il modello descrive il "rumore" nei dati di addestramento invece delle relazioni effettive tra le variabili nei dati. Questo rumore, ovviamente, non fa parte di nessun nuovo set di dati e non può essere quindi generalizzato sul gruppo di test o sul resto della popolazione.

4.1.1. Scelta delle k-feature

Al fine di ridurre l'overfitting è utile anche mantenere il modello il più possibile semplice, selezionando accuratamente il numero di feature da utilizzare. Per tale motivo si sono analizzate le 50 feature disponibili alla ricerca delle più importanti. L'obbiettivo è stato quello di comprendere quante e quali feature fossero in grado di interpretare maggiormente la varianza dei dati. In altre parole, si è voluto capire quante e quali feature fossero necessarie per avere buoni risultati in termini di precisione, recall, F1-misura.
Per la conduzione di questa analisi, l'idea di partenza è stata quella di condurre un'analisi statistica in grado di evidenziare l'indice di correlazione tra le variabili oggetto di studio. A tal fine, si sono confrontate tra loro a due a due tutte le 50 variabili usando l'indice di correlazione di Pearson.
Date due variabili statistiche *X* e *Y*, l'indice di correlazione di Pearson è definito come la loro covarianza divisa per il prodotto delle deviazioni standard delle due variabili:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x * \sigma_y}$$

Questo coefficiente può avere un valore compreso tra 1 (perfetta correlazione positiva) e -1 (perfetta correlazione negativa).
Nella pratica si distinguono vari "tipi di correlazione.
Se $\rho_{xy} > 0$; le variabili *X* e *Y* si dicono direttamente correlate;
Se $\rho_{xy} = 0$; le variabili *X* e *Y* si dicono incorrelate;
Se $\rho_{xy} < 0$; le variabili *X* e *Y* si dicono inversamente correlate.

Inoltre, per la correlazione diretta (e analogamente per quella inversa) si distingue:

- Se $0 < |\rho_{xy}| < 0{,}3$ si ha correlazione debole;
- Se $0{,}3 < |\rho_{xy}| < 0{,}7$ si ha correlazione moderata;
- $|\rho_{xy}| > 0{,}7$ si ha correlazione forte.

Per facilitare la comprensione dei risultati di analisi, si è scelto di realizzare una mappa di calore, che mostra il grado di correlazione in una scala che va dal bianco (correlazione diretta massima) al nero (correlazione inversa massima).

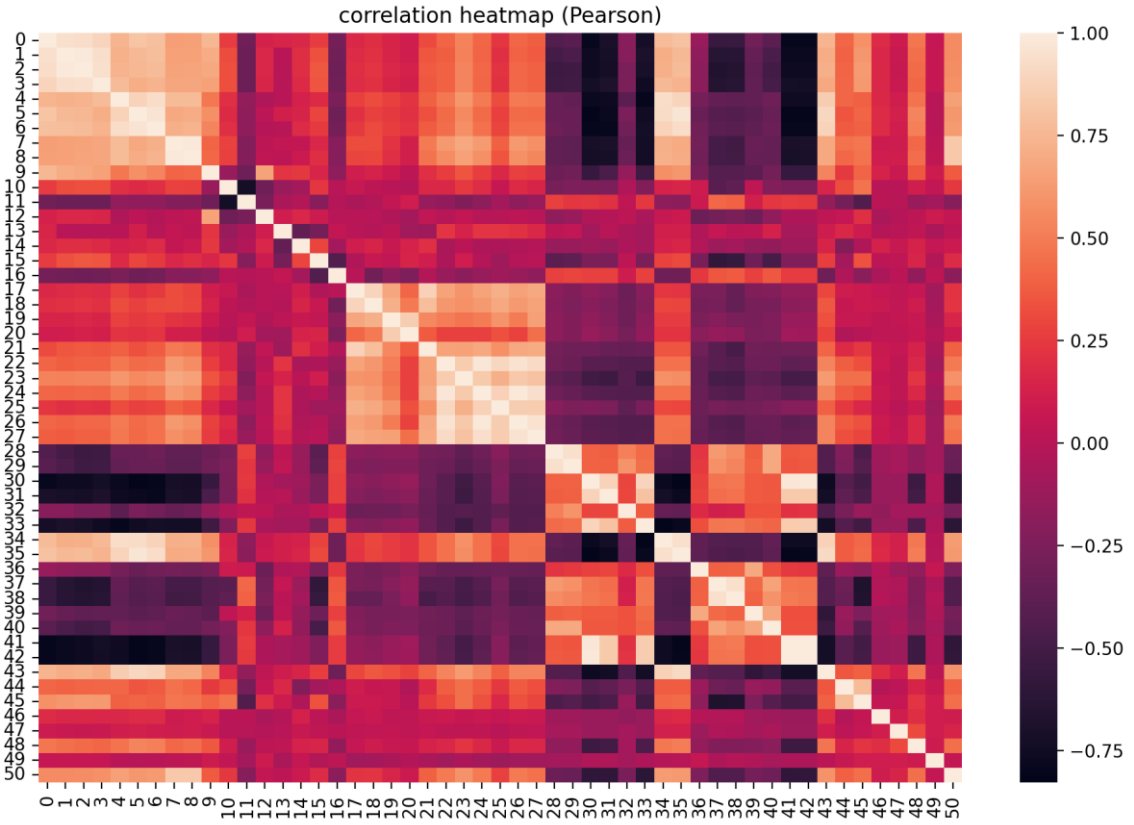

Figura 18:Mappa di correlazione lineare tra variabili (di Pearson)

La mappa di calore agevola la comprensione delle relazioni tra variabili e mostra in maniera piuttosto eloquente la presenza di cluster di variabili che si comportano in maniera molto simile tra loro. La logica conseguenza è quella di utilizzare per ogni cluster un'unica variabile, eliminando le altre che sono ridondanti. Non va però dimenticato che l'indice di correlazione di Pearson è in grado di catturare solo le correlazioni lineari tra le variabili, quindi può aiutare a comprendere i legami, senza tuttavia essere lo strumento adatto a scegliere le variabili da utilizzare.

L'altra evidente osservazione che si può fare è che all'interno della mappa di calore ci sono ampie zone con colori molto scuri o molto chiari. Queste zone, come detto, rappresentano correlazioni lineari forti in senso positivo o negativo. Analizzando la singola variabile, se questa è fortemente correlata alla gran parte delle altre variabili, aggiungerà poca informazione a quella già disponibile. Ciò significa che questa variabile potrebbe essere omessa, senza intaccare più di tanto le prestazioni del modello.
Infine, si può notare la presenza di alcuni cluster di variabili che hanno gli stessi colori, cioè sono correlate in maniera molto simile alle altre variabili e quindi si comportano in maniera quasi uguale tra loro. Questi cluster di variabili ci indicano chiaramente che alcune variabili sono ridondanti e potrebbero essere omesse.
Dopo aver fatto questa analisi preliminare, in concreto, si è scelto di utilizzare l'Analisi Fattoriale per effettuare la scelta delle k-feature. Alla base del concetto di Analisi Fattoriale, c'è l'intuizione di Spearman secondo cui le correlazioni tra le variabili osservate possano essere spiegate da un numero molto minore di variabili latenti, in grado di "sintetizzare" l'informazione contenuta nelle variabili osservate.
Rispetto alla tecnica dell'estrazione delle componenti principali, le variabili latenti non sono tra loro ortogonali, quindi non sono tra loro completamente scorrelate. In altri termini, è come se le componenti principali venissero ruotate in modo da farle corrispondere ai cluster individuati dalle variabili osservate. Ogni variabile latente andrà a disporsi in modo tale da "interpretare" una serie di variabili osservate tra loro molto correlate.
Graficamente, nell'immagine seguente si vede come le componenti principali ruotino nello spazio al fine di disporsi in modo da approssimare nel migliore dei modi i due cluster.

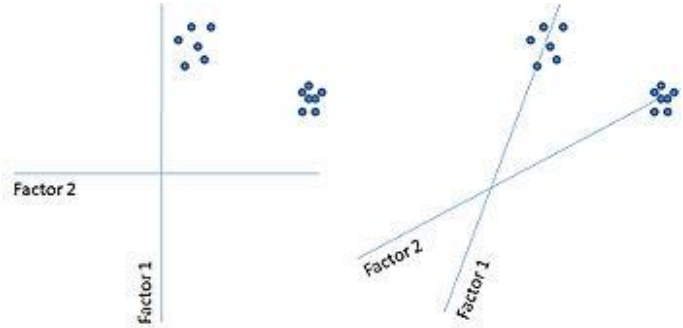

Figura 19: Analisi Fattoriale

Dai test condotti usando k=4 feature, si sono estratti gli indicatori più correlati a ciascuna variabile latente, in modo che però essi appartenessero a categorie diverse. Come detto, si sono usate cinque categorie di variabili per un totale di 51 indicatori:

1. sviluppo (6)
2. redditività (9)
3. produttività e struttura operativa (8)
4. liquidità e struttura finanziaria (18)
5. allerta per le crisi di impresa (10)

Dall'analisi condotta emerge chiaramente come gli indicatori di gran lunga migliori sono quelli relativi alle categorie 4 e 5. In particolare, i due indicatori più utili a predire il fallimento/liquidazione aziendale sono l'indicatore ln (AN) e l'indicatore OFN/Autofinanziamento Lordo. Il primo indicatore mostra che in caso di fallimento/liquidazione, c'è un evidente deterioramento delle condizioni aziendali che provoca una diminuzione delle attività, in special modo delle attività correnti, come evidente anche in letteratura.

Va però ipotizzata la presenza di un bias nel campione estratto, legato al fatto che le aziende del campione attive sono state estratte in base al fatturato decrescente, quelle fallite/liquidate sono state estratte in relazione alla presenza di un numero di bilanci superiori a 4. La conseguenza è che le aziende anomale del campione hanno dimensioni molto minori rispetto alle aziende sane, con un Attivo Netto aziendale già minore anche in periodi precedenti allo stress finanziario. Per questo si è deciso di andare a verificare la dimensione media dell'Attivo Netto quattro anni prima dell'evento default/liquidazione. I risultati sono esposti nella tabella che segue:

	sane	anomale
attivo netto medio	7572506	1376799

L'evidenza che emerge è che l'Attivo Netto medio delle aziende sane supera i 7 milioni, contro un Attivo Netto medio delle aziende anomale che non arriva ad 1,5 milioni. Questa differenza potrebbe essere molto maggiore di quella rinvenibile usando un campione di aziende sane non ordinato per fatturato.

Per questo motivo, si è deciso di escludere questo indicatore da tutti i calcoli che verranno esposti di seguito.

Analogamente, il secondo indicatore più correlato è ln (Ric) ed anch'esso è affetto dallo stesso bias.

	sane	anomale
Ricavi (media)	6376148	1259927

Come emerge dalla tabella, i Ricavi delle aziende sane superano mediamente i 6 milioni, contro quelli delle aziende anomale che non arrivano ad 1,5 milioni.

Anche questo indicatore si è deciso di escluderlo da tutta l'analisi condotta in seguito.

Alla fine, per quanto riguarda l'indicatore di Allerta per la Crisi d'Impresa, è stato selezionato l'indicatore Debiti Totali/Ebitda, che non è afflitto da questo tipo di problematiche. Questo è un indicatore di solvibilità ed è molto importante nella sua semplicità, poiché permette di capire quante volte deve essere generato il valore del Margine Operativo Lordo per poter ripagare i Debiti Totali contratti dalla società. L'indicatore può essere declinato in anni o in mesi e tanto più è grande, tanto più l'azienda è rischiosa.

Il secondo indicatore, Oneri Finanziari Netti/Autofinanziamento Lordo, è un indicatore interessante. Da una parte gli Oneri Finanziari Netti ci indicano quanto spende l'impresa per finanziare il debito. Evidentemente al crescere degli oneri finanziari pagati, la capacità dell'impresa di ripagare il debito tende a diminuire. Dall'altro lato l'autofinanziamento Lordo indica la capacità dell'impresa di finanziare il suo business senza ricorrere a fonti esterne (debito verso banche, obbligazioni, capitale apportato dagli azionisti…). In genere l'autofinanziamento è una fonte poco usata dalle aziende nata da poco, le quali però non sono comprese all'interno del database data la volontà di studiare anche le serie temporali. Più questo rapporto è alto, più l'azienda fatica a restare in piedi con le proprie gambe e rischia di entrare in crisi.

Gli altri due indicatori più significativi sono per la categoria redditività il ROE, per la categoria sviluppo la Variazione percentuale di Patrimonio Netto. Il ROE indica il ritorno degli azionisti sul Capitale investito nella società. Tanto maggiore è il ROE, tanto più l'investimento effettuato dall'azionista genera un profitto e conseguentemente l'azienda è in buone condizioni di salute. Ciò significa che, se anche un'azienda ha una leva finanziaria alta, se essa riesce a conseguire un ROE alto è chiaramente molto difficile che incorra in fallimento nell'anno seguente. Questa deduzione è valida solamente in un arco temporale estremamente limitato, poiché al deteriorarsi della redditività negli anni seguenti, l'alta leva finanziaria potrebbe comportare l'entrata in crisi della società.

La variazione percentuale del Patrimonio Netto è un indicatore anch'esso molto significativo, poiché se un'azienda tende ad aumentare il Patrimonio Netto, in genere

tramite il conseguimento di un utile non distribuito, conseguentemente la maggior patrimonializzazione funge da tutela verso un eventuale fallimento ed è sinonimo di solidità aziendale. In casi particolari, es. aumento di capitale, la variazione può essere dovuta non ai positivi risultati aziendali, ma all'apporto di capitale da parte dei soci. Anche in questo caso, l'operazione consente di mettere l'azienda fuori dal pericolo default nell'anno successivo.

Infine, gli indicatori di produttività e struttura operativa hanno dimostrato un impatto quasi nullo sull'efficacia della predizione, quindi non sono poi stati utilizzati.

Ricapitolando, di seguito una tabella con i quattro indicatori scelti ai fini dell'impiego delle tecniche di machine learning supervisionato. L'ordine con cui sono esposte indica l'ordine di priorità quando il numero di feature fosse minore di quattro.

	Indicatore	**Categoria**
1	OFN/Autof lordo	liquidità e struttura finanziaria
2	Debiti Totali/Ebitda	allerta per le crisi di impresa
3	ROE	redditività
4	Var % patrimonio Netto	sviluppo

4.1.2. Risultati

La prima tecnica implementata è stata la regressione logistica. Questa tecnica ha dimostrato altissima capacità predittiva con uno score che già con una sola feature raggiunge livelli molto elevati di accuratezza.

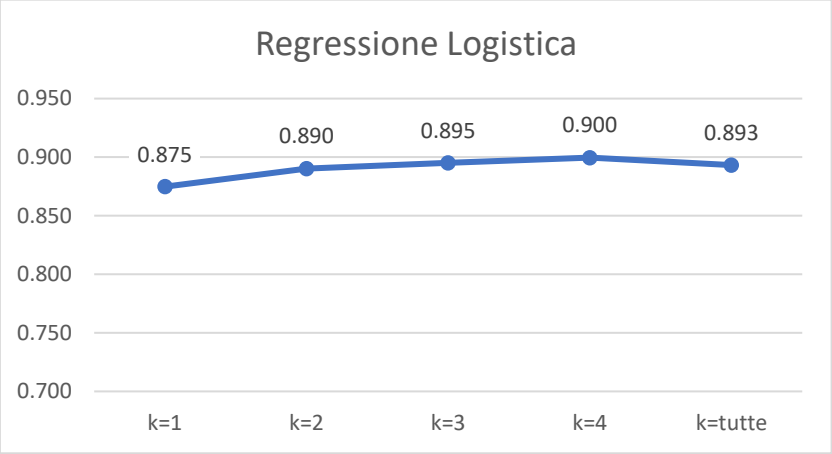

Figura 20:Regressione Logistica

L'evidenza chiara che si ha è che l'utilizzo di più di due feature è assolutamente non necessario.
La seconda tecnica implementata è stata il Gradient Boosting. Questa tecnica ha mostrato risultati in termini di accuratezza ancora migliori, ad eccezione del caso con una sola variabile. Poiché si raggiunge un livello di accuratezza del 90% con sole due feature, è consigliabile usare k=2 per avere il miglior trade-off tra accuratezza elevata e semplicità di calcolo in termini di effort computazionale.

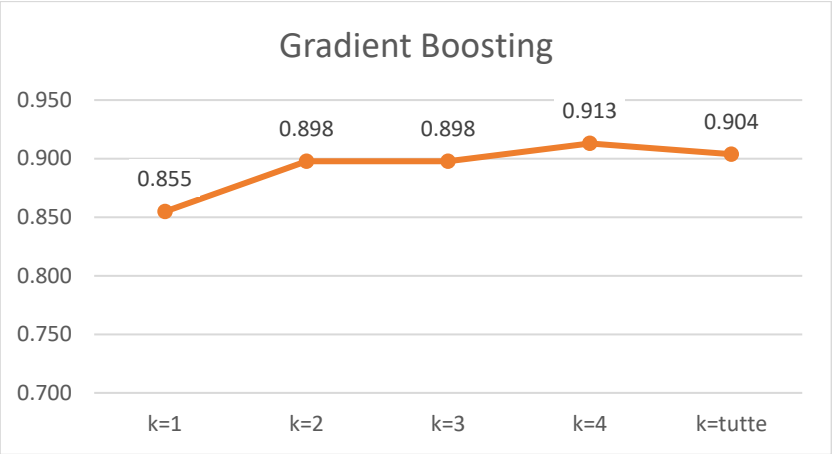

Figura 21:Gradient Boosting

La terza tecnica implementata è la cosiddetta K-Nearest Neighbors. Questa tecnica sembra funzionare sorprendentemente bene anche con un solo cluster, raggiungendo un'accuratezza superiore all'85%, ma mostra i risultati migliori con un numero di feature compreso tra 2 e 4. In generale, quindi, il miglior trade-off si raggiunge usando due feature.

A margine, è molto interessante notare come i risultati in termini di accuratezza degradano notevolmente quando si usano tutte le features. Questo è un esempio concreto di overfitting. All'aumentare delle informazioni fornite e quindi della complessità della base dati, l'algoritmo K-Nearest Neighbors perde parte della sua capacità di predizione. In effetti è confermato in letteratura che un numero troppo elevato di variabili usate conduce all'overfitting. Questo effetto è stato comunque in parte mitigato dall'utilizzo in fase di Validazione dei valori mediati di una serie di train-test split con diversa inizializzazione.

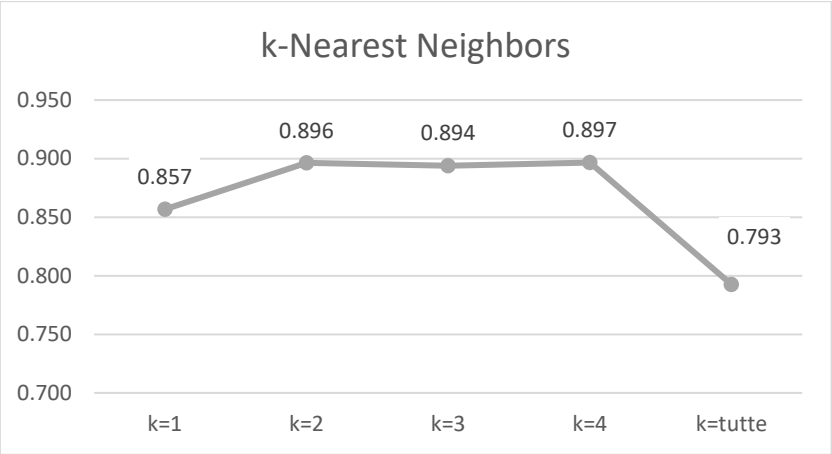

Figura 22: k – Nearest - Neighbors

L'algoritmo Gaussian Naive-Bayes conduce a risultati abbastanza particolari. Esso raggiunge dei risultati altissimi già con una sola feature, che però tendono a decrescere all'aumentare del numero di variabili. Nel caso in cui si utilizzano tutte le variabili, i risultati perdono quasi il 15% della qualità in termini di accuratezza. Questo fenomeno non dovrebbe verificarsi, poiché l'algoritmo Naive Bayes è in teoria uno dei più robusti e meno soggetti al fenomeno dell'overfitting. La considerazione che può essere fatta è che il processo di feature selection è riuscito talmente bene, che una sola feature è più che sufficiente per ottenere risultati estremamente positivi.

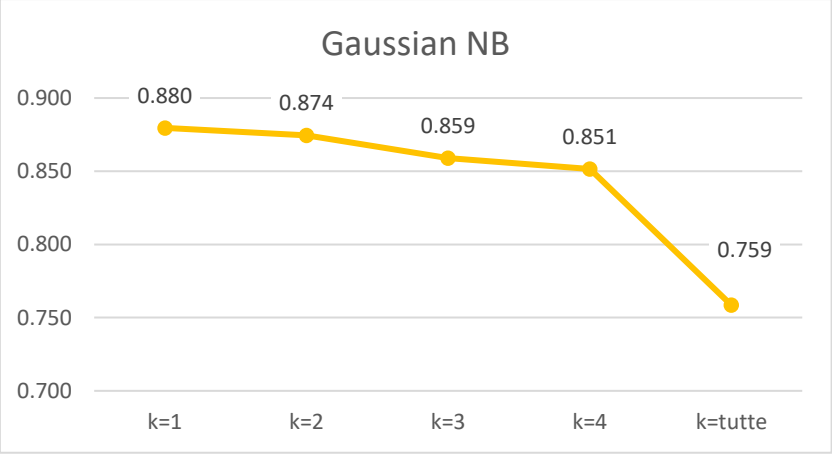

Figura 23:Gaussian Naive Bayes

L’algoritmo Decision Tree mostra dei risultati un po' meno elevati, ma comunque più che discreti. Si nota un alto incremento delle sue prestazione se il numero di feature usate passa da una a due. All’aumentare ulteriore del numero di variabili, le prestazioni dell’algoritmo crescono solo in maniera marginale, quindi k=2 rappresenta il migliore trade-off possibile.

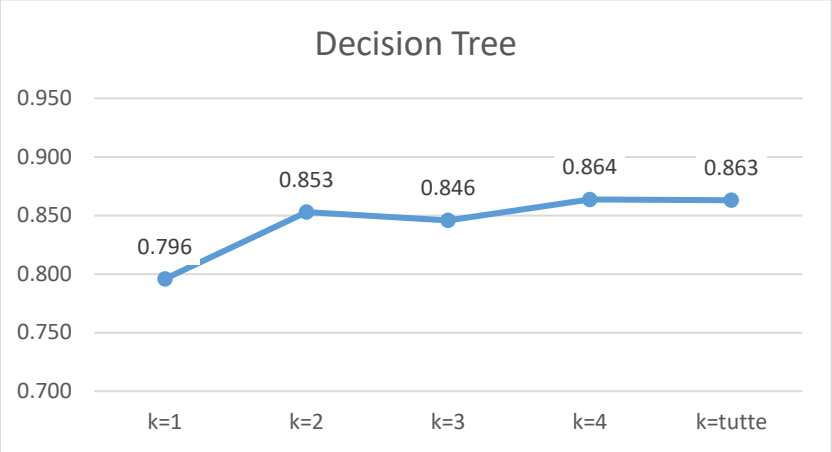

Figura 24: Decision Tree

L’algoritmo Random Forest si dimostra essere l’unico in cui c’è un incremento, seppur non significativo, delle prestazioni quando si usano tutte le variabili, raggiungendo in assoluto i valori più alti. Questo comportamento è spiegato dal fatto che il Random Forest è una tecnica di ensemble learning in cui si utilizzano più alberi decisionali addestrati su porzioni del dataset tramite campionamento casuale con rimpiazzo. Di conseguenza, ogni singolo classificatore si addestra su una porzione casuale di caratteristiche e quindi alcune di esse possono comparire contemporaneamente in più modelli mentre altre potrebbero non comparire mai. L’allenamento su una parte delle caratteristiche consente ad ogni modello di limitare l’overfitting e migliorare le capacità predittive.
Per come è costruito il Random Forest, esso consente di limitare l’overfitting e quindi ottiene i risultati migliori quando il numero di feature utilizzate è elevato.

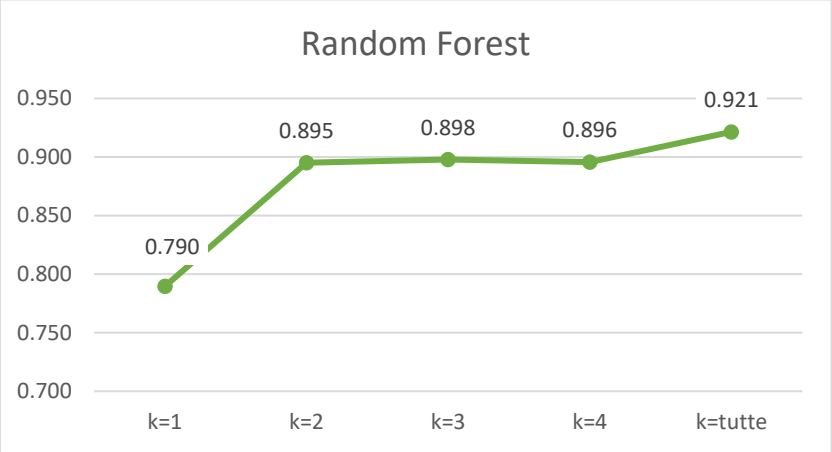

Figura 25: Random Forest

Infine, l'algoritmo SVM (o Support Vector Machines) mostra dei risultati molto buoni e stabili. Una sola feature è già più che sufficiente ad ottenere risultati estremamente buoni, senza la necessità di usarne di più.

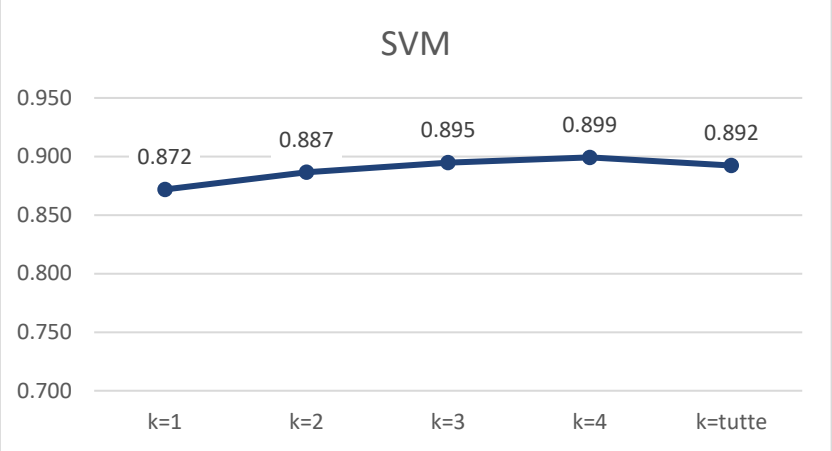

Figura 26: Support Vector Machines

Riassumendo, si vede chiaramente come tutti questi algoritmi riescono a beneficiare del processo di feature selection per cui, una volta trovate le feature che meglio interpretano i dati, si può scegliere di usare solo quelle, evitando di aggiungere "rumore" con le altre feature disponibili. Queste ultime in alcuni casi sono inutili, in altri casi addirittura confondono l'algoritmo inficiando la qualità del risultato in termini di accuratezza.
Il processo di feature selection, che nel deep learning non viene effettuato poiché le reti neurali sono esse stesse in grado di estrarre queste informazioni, è quindi molto utile quando si applicano algoritmi di machine learning, come mostrato nel capitolo.

Di seguito una vista di sintesi, da cui emerge che i risultati migliori sono stati ottenuti con l'algoritmo Random Forest con tutte le feature (accuratezza del 92,1%), ma che in generale il compromesso migliore tra complessità del modello e accuratezza del risultati si ottiene con k=2.

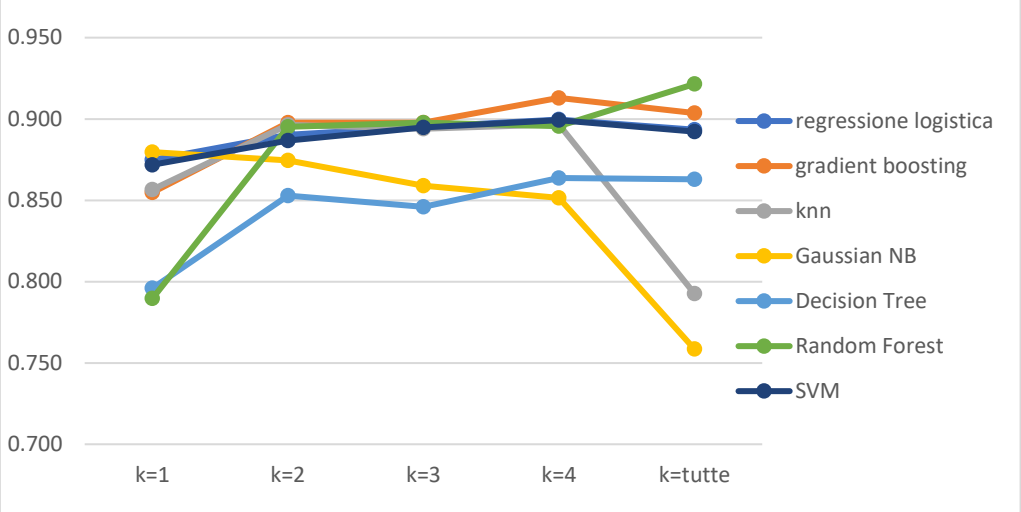

Figura 27: Accuratezza degli algoritmi supervisionati al variare del numero di feature

Infine, ribaltando la prospettiva e proiettando su un grafico i risultati in termini di accuratezza raggiunti per ogni algoritmo al variare di k, si può notare innanzitutto che quando k è compreso tra 2 e 4 i risultati sono molto solidi, non variando in maniera molto significativa al variare dell'algoritmo usato. Al contrario, quando k è uguale a 1 vanno evitati gli algoritmi Random Forest e Decision Tree; quando si usano tutte le k feature vanno evitati K-Nearest Neighbors e Gaussian Naive Bayes.

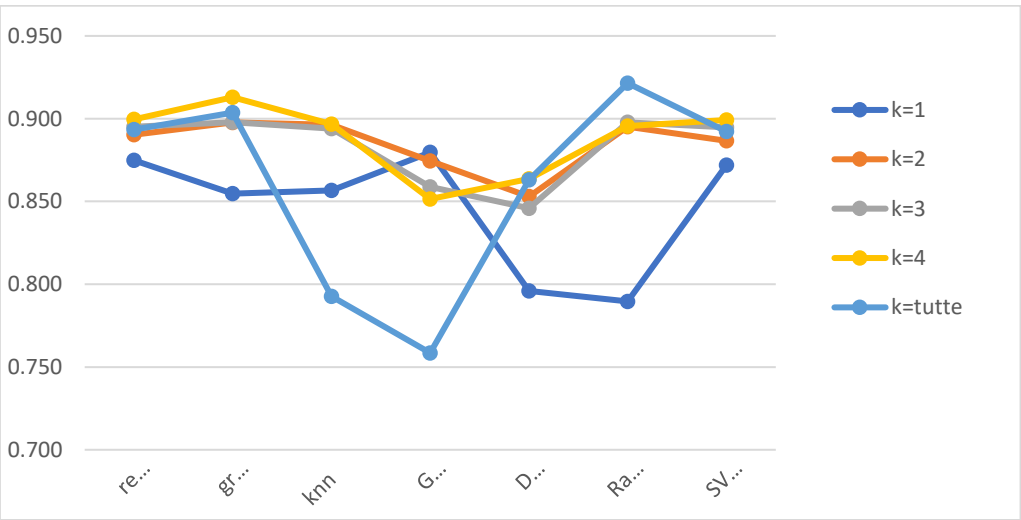

Figura 28: Accuratezza degli algoritmi supervisionati

4.2. Algoritmi non supervisionati

In questa sezione si utilizza un approccio non supervisionato, in cui cioè non si tiene conto dell'etichetta durante la fase di addestramento.

Nella sezione precedente si sono analizzati gli algoritmi supervisionati. In essi si crea un gruppo di addestramento in cui l'algoritmo riceve in input le variabili e la classe di appartenenza (0/1). Grazie a queste informazioni, l'algoritmo impara a riconoscere le relazioni fra le variabili che determinano l'appartenenza alla specifica classe. Applicando queste relazioni al gruppo di test, l'algoritmo riesce a predire la classe di appartenenza con una precisione più o meno marcata.

Il funzionamento degli algoritmi non supervisionati è completamente diverso. Essi non ricevono in input l'informazione sulla classe di appartenenza delle aziende del campione; quindi, il loro scopo è di trovare delle logiche secondo cui aggregare le aziende del campione, secondo quell'operazione denominata Clustering.

In questo lavoro di tesi, avendo a disposizione l'informazione riguardo la classe degli elementi del campione, ha poco senso ricorrere ad algoritmi non supervisionati.

Si offre quindi solo un esempio di algoritmo non supervisionato, per avere uno strumento di comparazione rispetto agli algoritmi supervisionati usati in precedenza.

L'algoritmo non supervisionato utilizzato è l'algoritmo E-M (expectation-maximization). In esso si ipotizza che i pattern siano stati generati da una mistura di distribuzioni: ogni classe ha generato dati in accordo con una specifica distribuzione, ma al termine della generazione i pattern appaiono come prodotti da un'unica distribuzione multi-modale.

Obiettivo del clustering con E-M è risalire, a partire dai pattern del training set, ai parametri delle singole distribuzioni che li hanno generati.

La stima dei parametri avviene secondo il criterio della stima della massima verosimiglianza (MLE). In generale la verosimiglianza corrisponde alla probabilità che i dati(osservazioni) siano stati generati da una certa distribuzione data. Per ragioni di stabilità numerica, al posto della verosimiglianza, si massimizza il suo logaritmo.

L'algoritmo è iterativo e si ripetono due fasi:

- Expectation: Si calcola la verosimiglianza per ogni punto del training set. In altre parole, si trova quale è la probabilità con cui i vari punti appartengono a ciascuna distribuzione (cluster)

- Maximization: Si massimizza la verosimiglianza trovando lo stimatore della massima verosimiglianza per medie e varianze di ciascuna distribuzione. Infine, si calcolano i parametri delle nuove distribuzioni

Questi due passi vengono eseguiti iterativamente (fino a convergenza). L'algoritmo può essere assimilato ad una versione probabilistica del K-Means.
I risultati ottenuti tramite l'uso dell'algoritmo sono comunque abbastanza soddisfacenti, come si può vedere nella tabella seguente, con valori di recall e precisione superiori al 70%.

F1-misura	Recall	Precisione
0,735	0,7191	0,7517

Questi risultati sono stati ottenuti usando un software specifico per il data mining, ELKI, che ha consentito anche la creazione di un grafico 3D in cui sono mostrati i due cluster delle aziende sane e fallite in relazione alle due variabili più correlate Debiti Totali/EBITDA e Oneri finanziari netti/Autofinanziamento Lordo. L'idea di fondo è che usando solo due variabili si può rappresentare graficamente su un piano la distribuzione spaziale dei punti appartenenti ai due cluster

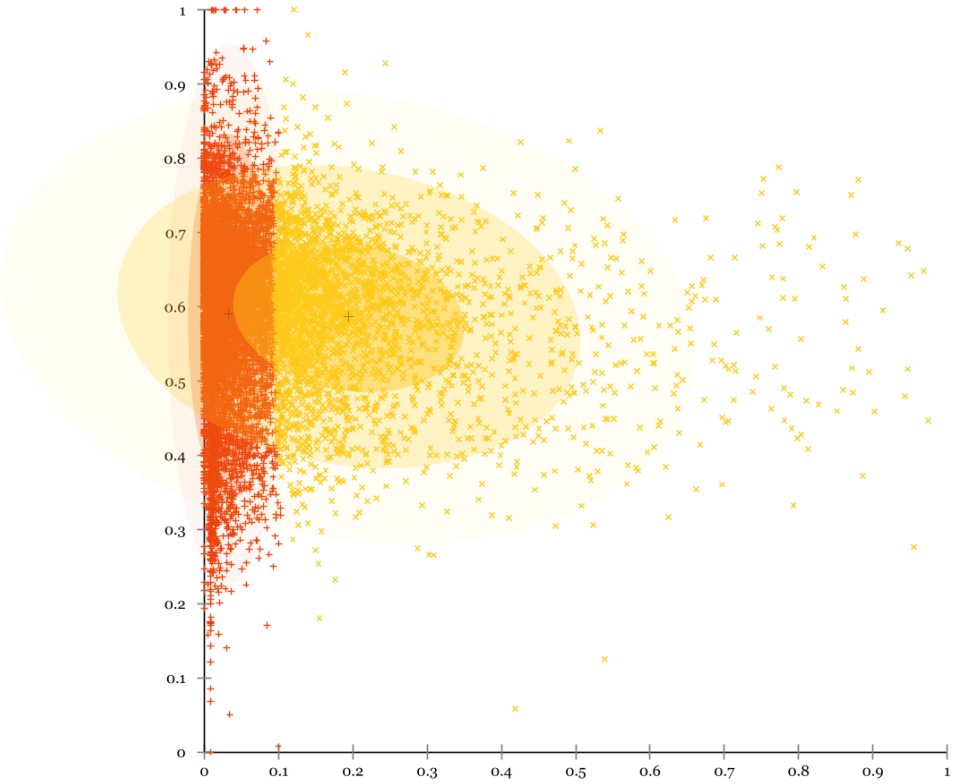

Figura 29: Creazione dei cluster su un piano nell'algoritmo E-M

Dal grafico si vede come la variabile sulle ascisse, ossia Oneri finanziari netti/Autofinanziamento Lordo contribuisca fortemente alla ripartizione delle osservazioni all'interno dei cluster. Le aziende sane si dispongono tutte all'interno dei valori compresi tra 0 e 0,1, le aziende anomale si distribuiscono invece più uniformemente.

4.3. Algoritmi basati sulle serie temporali

L'ultima opzione esplorata è stata quella di considerare per ogni azienda tutta o parte della serie storica dei bilanci disponibili. L'osservazione alla base che ha motivato questo tentativo è stata quella di capire se l'algoritmo di machine learning fosse in grado di imparare dall'evoluzione temporale avuta dalle variabili osservate dell'azienda, riuscendo a cogliere il degradamento nel tempo dei parametri messi sotto osservazione. L'obbiettivo è stato quindi quello di fornire serie temporali con un numero di anni sempre maggiore e vedere, al crescere del numero di anni messi a disposizione dell'algoritmo, quale fosse l'evoluzione in termini di accuratezza della predizione.

Il primo algoritmo usato è il l'algoritmo Distance Based. Questo algoritmo mostra scarsa capacità di apprendimento. Infatti, pur se i risultati si mantengono su alti valori assoluti in termini di accuratezza, con valori che oscillano tra 0,772 e 0,835, si nota come l'estensione temporale della serie storica non favorisca l'incremento dell'accuratezza dell'algoritmo. Ora, considerando che usando serie storiche di tre anni, si raggiunge un risultato in termini di accuratezza di 0,82, l'osservazione che si può fare è che questo algoritmo non è in grado di apprendere dall'evoluzione temporale dei dati e quindi non è adatto a predire quali aziende siano sane o anomale in questo specifico database. Mancando la principale caratteristica che deve avere un algoritmo basato sulle serie storiche, tanto vale usare l'approccio standard basato sull'analisi esclusiva dell'anno precedente al fallimento/liquidazione.

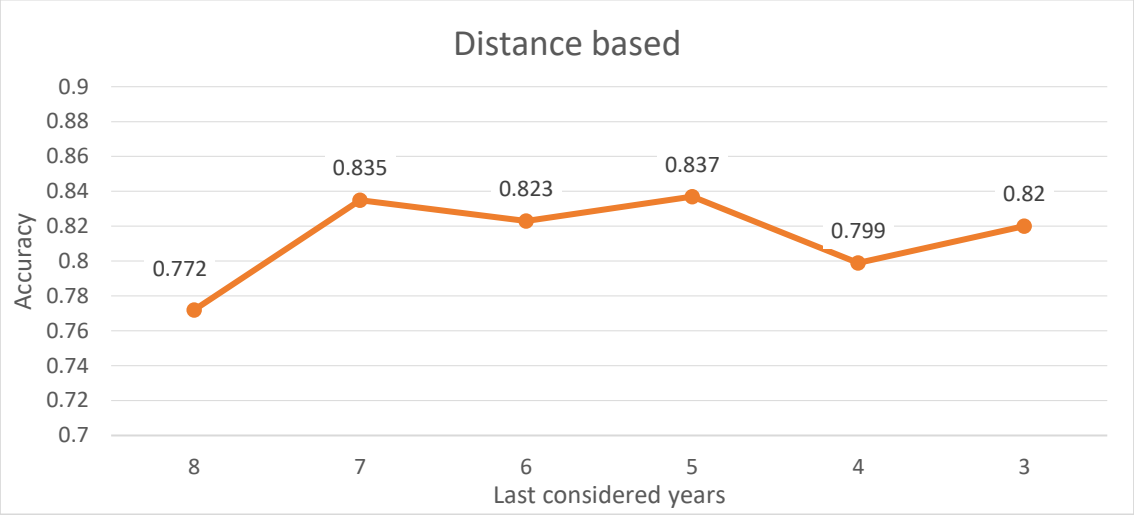

Figura 30: Algoritmo Distance Based

Al contrario, l'algoritmo Shapelet Based mostra un'evoluzione molto più interessante. Al crescere del numero di anni considerati nella serie storica, si nota un significativo incremento della capacità di apprendimento dell'algoritmo, con un valore di accuratezza che passa da 0,795 con serie storiche di 3 anni a 0,878 con serie storiche di 8 anni. Questo fattore lascia presupporre che l'algoritmo sia in grado di apprendere in maniera non trascurabile l'evoluzione temporale delle variabili analizzate facendo inferenza e arrivando a livelli di accuratezza elevati.

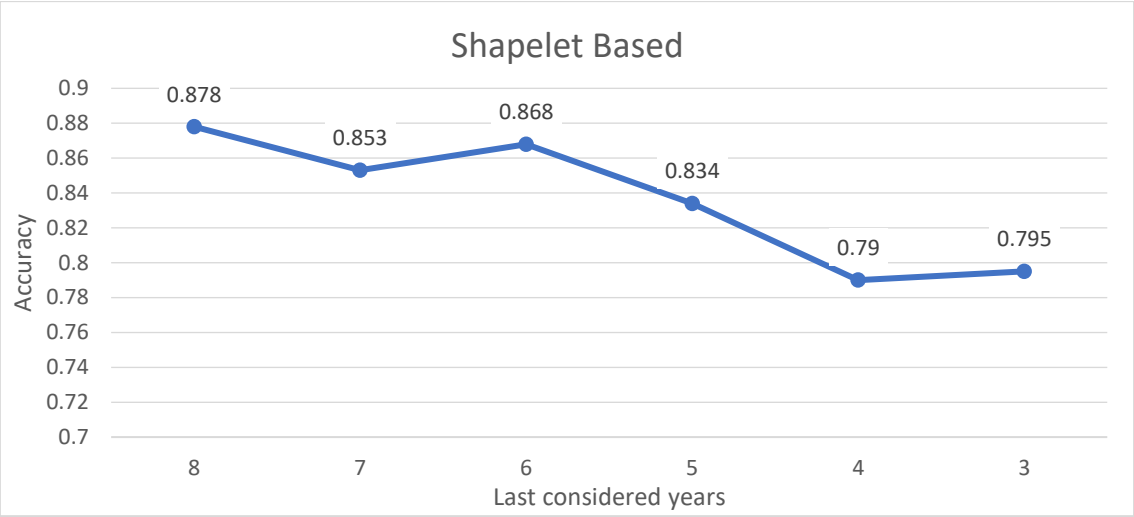

Figura 31:Algoritmo Shapelet Based

Questo tipo di algoritmo raggiunge dei risultati tali da lasciar presupporre che con serie temporali sufficientemente lunghe i risultati raggiungibili possano essere migliori di quelli ottenibili considerando solamente l'anno precedente al fallimento. Esso, infatti, riesce a sfruttare nel modo più corretto una mole di dati riuscendo a cogliere l'evoluzione nel tempo dei bilanci aziendali, che mediamente per un'azienda anomala provocano un decadimento pronunciato e prolungato negli anni delle variabili analizzate.

Durante l'analisi uno dei problemi che si sarebbe potuto verificare era che il numero di bilanci di aziende fallite si riducesse in maniera significativa andando a variare i supporti nel gruppo di test all'interno del campione. Questo avrebbe potuto determinare che l'aumento dell'accuratezza potesse essere causato dall'aumento percentuale delle aziende sane effettivamente classificate come sane, senza tuttavia effettuare una altrettanto corretta classificazione delle aziende anomale. Di seguito, viene mostrata una tabella in cui è evidente come, anche se la dimensione del gruppo di test del campione si riduce, la percentuale fra aziende sane e aziende anomale resta più o meno bilanciata. Analogamente, oltre all'accuratezza sono state analizzate anche la recall e la precisione. La recall calcola il numero di aziende fallite classificate correttamente sul totale di aziende fallite. La precisione calcola il numero di aziende effettivamente fallite sul totale delle aziende che l'algoritmo ha classificato come fallite. Ne consegue che se i valori di recall o di precisione fossero molto bassi, l'algoritmo potrebbe far fatica a riconoscere i falsi positivi o i falsi negativi. Nel caso specifico, se il numero di aziende fallite fosse basso, ma il numero di falsi positivi fosse elevato, l'accuratezza resterebbe alta, mentre

la precisione dovrebbe ridursi di molto. Come si vede nella tabella che segue, ciò non succede, a testimonianza della bontà dell'algoritmo.

N. anni serie storica	Stato	supporto	precisione	recall
8 anni	sane	129	0,87	0,9
8 anni	fallite	125	0,89	0,86
7 anni	sane	140	0,84	0,85
7 anni	fallite	152	0,86	0,86
6 anni	sane	169	0,88	0,85
6 anni	fallite	164	0,85	0,88
5 anni	sane	196	0,84	0,84
5 anni	fallite	189	0,84	0,84
4 anni	sane	234	0,81	0,79
4 anni	fallite	213	0,77	0,79
3 anni	sane	230	0,8	0,8
3 anni	fallite	220	0,79	0,8

Riepilogando, fra i due approcci sperimentati solamente l'approccio Shapelet Based ha mostrato di essere valido e può costituire una valida alternativa agli approcci standard supervisionati quando le serie storiche sono lunghe a sufficienza. Nel campione sembra esserci un incremento in termini di accuratezza quando le serie storiche sono maggiori o uguali a cinque anni.

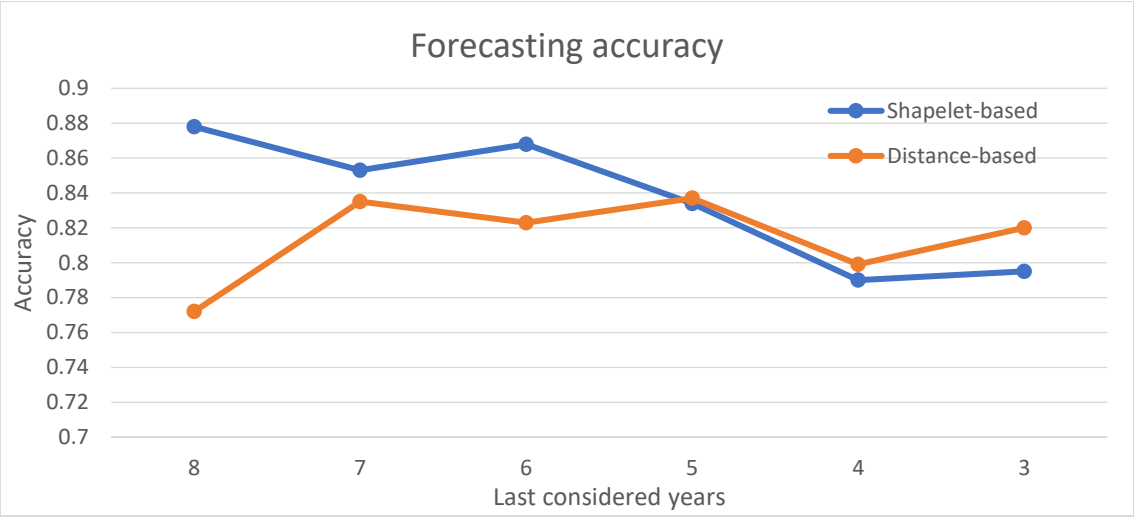

Figura 32: Accuratezza del modelli basati su serie temporali

CONCLUSIONE

In questa tesi sono stati applicati vari approcci di machine learning per predire lo stato di fallimento o liquidazione di aziende del settore del Legno-Arredo usando una serie di indicatori calcolati a partire dai bilanci di esercizio.

La scelta dell'argomento è stata dettata dal forte interesse che il mondo bancario riversa sempre più sulla analisi dei crediti deteriorati e sulle analisi da condurre prima di erogare credito. Questa necessità poggia innanzitutto sugli stringenti parametri imposti dalla regolamentazione bancaria, particolarmente severa in riferimento alla tematica del rischio di credito, e sulla volontà di migliorare i risultati economici delle banche commerciali.

L'analisi dei bilanci aziendali è ormai negli ultimi vent'anni sempre più permeata dalle nuove tecniche di data science, in particolare machine learning e deep learning. In questo lavoro ci si è concentrati solo sulle tecniche di machine learning, confrontandole tra loro al fine di stabilire quali fossero le migliori per questo specifico campione.

Ciò che è emerso è che le tecniche di machine learning hanno ottenuto dei risultati soddisfacenti.

In particolare, i sette algoritmi supervisionati utilizzati, hanno raggiunto livelli di accuratezza anche superiori al 90% senza mai scendere sotto il 75%. Nella quasi totalità dei casi, inoltre, non era necessario sottoporre all'algoritmo tutti gli indicatori, ma era sufficiente considerarne solo alcuni tra i più significativi in un range tra uno e quattro. Tutto ciò potrebbe semplificare di molto il lavoro di preparazione della base dati da elaborare. L'evidenza è che questi algoritmi, a differenza di quello che potrebbe succedere nel deep learning, non riescono ad incrementare i livelli di accuratezza al crescere del numero di variabili fornite in input.

È poi stato usato un algoritmo non supervisionato, noto come algoritmo E-M (Expectation–Maximization). Questa tipologia di algoritmo non era la più adatta. Gli algoritmi non supervisionati sono infatti pensati per essere applicati su basi dati sprovviste di etichetta, nelle quali cioè non è disponibile l'informazione su come sono classificati effettivamente gli elementi del campione. È naturale quindi che i risultati in termini di accuratezza siano stati più scadenti, pur arrivando a superare il 70%.

Infine, è stato tentato un approccio più innovativo, utilizzando tecniche di machine learning dedicate specificatamente all'analisi delle serie storiche. Negli approcci precedenti, infatti, si erano utilizzate solo le informazioni relative all'anno precedente al

default, ma il database creato aveva per ogni azienda serie storiche con un numero di anni compreso tra 4 e 10. Una delle due tecniche sperimentate ha dimostrato di saper sfruttare l'informazione contenuta anche negli altri anni disponibili. L'evidenza è stata infatti che l'algoritmo fosse in grado di aumentare la sua precisione all'aumentare del numero di anni di osservazione. Questo approccio è stato tuttavia sperimentato considerando tutte le variabili; quindi, risulta più complesso e necessita di un base dati molto articolata, con una serie storica di almeno 5 anni per ogni azienda e un numero di variabili elevato. Più si ha la disponibilità di serie storiche molto lunghe, più questo approccio risulta una valida alternativa agli approcci standard.

In conclusione, la panoramica andrebbe completata applicando le reti neurali al database ivi raccolto, per verificare se e di quanto la qualità dei risultati può essere ulteriormente incrementata. In ogni caso, i risultati raggiunti con gli algoritmi sono soddisfacenti e vista la maggiore semplicità di implementazione e lettura dei risultati, questo lavoro suggerisce che l'approccio standard per alcune applicazioni potrebbe essere sufficiente.

BIBLIOGRAFIA

[1] R. N. Langlois and M. M. Cosgel, “FRANK KNIGHT ON RISK, UNCERTAINTY, AND THE FIRM: A NEW INTERPRETATION.”

[2] J. A. Lopez and M. R. Saidenberg, “Evaluating credit risk models,” 2000. [Online]. Available: www.elsevier.com/locate/econbase

[3] F. Andersson, H. Mausser, D. Rosen, and S. Uryasev, “Digital Object Identifier (DOI) 10.1007/s101070000201,” *Math. Program., Ser. B*, vol. 89, pp. 273–291, 2001, doi: 10.1007/s101070000201

[4] F. ’Varetto and G. ’Szego, *“Il rischio creditizio: misure e controllo.”* UTET Università, 1999

[5] S. Y. Sohn, D. H. Kim, and J. H. Yoon, “Technology credit scoring model with fuzzy logistic regression,” *Applied Soft Computing Journal*, vol. 43, pp. 150–158, Jun. 2016, doi: 10.1016/j.asoc.2016.02.025

.

[6] J. A. Dubin and D. Rivers, “PASADENA, CALIFORNIA 91125 SELECTION BIAS IN LINEAR REGRESSION, LOGIT AND PROBIT MODELS SELECTION BIAS IN LINEAR REGRESSION, LOGIT AND PROBIT MODELS LOGIT AND PROBIT MODELS,” 1989.

[7] J. Schmidhuber, “Deep Learning in neural networks: An overview,” *Neural Networks*, vol. 61. Elsevier Ltd, pp. 85–117, Jan. 01, 2015. doi: 10.1016/j.neunet.2014.09.003.

[8] G. Strawn and C. Strawn, “Masterminds of Artificial Intelligence: Marvin Minsky and Seymour Papert,” *IT Professional*, vol. 18, no. 6, pp. 62–64, Nov. 2016, doi: 10.1109/MITP.2016.116.

[9] I. el Naqa and M. J. Murphy, “What Is Machine Learning?,” in *Machine Learning in Radiation Oncology*, Springer International Publishing, 2015, pp. 3–11. doi: 10.1007/978-3-319-18305-3_1.

[10] U. Fayyad, D. Haussler, and P. Stolorz, “KDD for Science Data Analysis: Issues and Examples,” 1996. [Online]. Available: www.aaai.org

[11] IEEE Electron Devices Society, Institute of Electrical and Electronics Engineers, and Vaigai College of Engineering, *Proceeding of the 2018 International Conference on Intelligent Computing and Control Systems (ICICCS) : June 14-15, 2018*.

[12] M. Brijain, R. Patel, M. Kushik, and K. Rana, “A Survey on Decision Tree Algorithm For Classification,” 2014. [Online]. Available: www.ijedr.org

[13] H. Zhang, “The Optimality of Naive Bayes.” [Online]. Available: www.aaai.org

[14] L. Jiang, Z. Cai, D. Wang, and S. Jiang, “Survey of Improving K-Nearest-Neighbor for Classification.”

[15] W. : Www and A. Pradhan, “International Journal of Emerging Technology and Advanced Engineering SUPPORT VECTOR MACHINE-A Survey,” 2012. [Online]. Available: www.ijetae.com

[16] L. E. Melkumova and S. Y. Shatskikh, “Comparing Ridge and LASSO estimators for data analysis,” in *Procedia Engineering*, 2017, vol. 201, pp. 746–755. doi: 10.1016/j.proeng.2017.09.615.

[17] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” 2005.

[18] L. Bbeiman, “Bagging Predictors,” 1996.

[19] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, “A comparative analysis of gradient boosting algorithms,” *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/s10462-020-09896-5.

[20] S. D. ˇ Zeroski, “Is Combining Classifiers with Stacking Better than Selecting the Best One?,” 2004.

[21] T. Soni Madhulatha, “AN OVERVIEW ON CLUSTERING METHODS,” vol. 2, no. 4, pp. 719–725, 2012, [Online]. Available: www.iosrjen.org

[22] X. Yang, Y. Lv, H. Sun, L. Fang, M. Wang, and X. Ma, “An improved K-means algorithm in topic detection,” in *6th International Conference on Soft Computing and Intelligent Systems, and 13th International Symposium on Advanced Intelligence Systems, SCIS/ISIS 2012*, 2012, vol. 2012-January, pp. 2366–2369. doi: 10.1109/ICCSN.2011.6014384.

[23] B. K. Sriperumbudur and I. Steinwart, “Consistency and Rates for Clustering with DBSCAN,” 2012.

[24] O. Cappé, “Online Expectation-Maximisation,” Nov. 2010, [Online]. Available: http://arxiv.org/abs/1011.1745

[25] Q. Zhao, “Association Rule Mining: A Survey.”

[26] N. K. Ahmed, A. F. Atiya, N. el Gayar, and H. El-Shishiny, “An Empirical Comparison of Machine Learning Models for Time Series Forecasting.”

[27] A. Abanda, U. Mori, and J. A. Lozano, “A review on distance based time series classification,” *Data Mining and Knowledge Discovery*, vol. 33, no. 2, pp. 378–412, Mar. 2019, doi: 10.1007/s10618-018-0596-4.

[28] Y. Zhou, H. Ren, Z. Li, and W. Pedrycz, “An anomaly detection framework for time series data: An interval-based approach,” *Knowledge-Based Systems*, vol. 228, Sep. 2021, doi: 10.1016/j.knosys.2021.107153.

[29] J. Hills, J. Lines, E. Baranauskas, J. Mapp, and A. Bagnall, “Classification of time series by shapelet transformation,” *Data Mining and Knowledge Discovery*, vol. 28, no. 4, pp. 851–881, 2014, doi: 10.1007/s10618-013-0322-1.

[30] “Lo stato di salute della filiera del legnoarredo 2016: si consolida la ripresa.”