# POLITECNICO DI TORINO

Corso di Laurea Magistrale In Ingegneria Gestionale
Percorso Finanza

Tesi di Laurea Magistrale

## Applicazione degli Algoritmi Genetici ai Modelli di Scoring

Relatore
Professoressa. Laura Rondi
Co-Relatore
Prof. Franco Varetto

Candidato
Giordano Gregnanin

Anno Accademico 2021/2022

# Indice

# 1 INTRODUZIONE

All'interno di questo elaborato si vogliono esporre i risultati ottenuti in letteratura dai modelli di scoring che utilizzano come motore decisionale gli algoritmi genetici. Gli algoritmi genetici sono un approccio ai problemi di calcolo numerico ispirati al meccanismo di sviluppo biologico e che sono tra i metodi più utilizzati dalle attuali tecnologie di machine learning, individualmente o in forma ibrida con altre metodologie.

La tesi introduce il rischio così come è stato concepito nel mondo imprenditoriale e finanziario, fornendo una tassonomia dei rischi principali e una loro definizione, con particolare focus alla valutazione del rischio di credito. Il rischio di credito è il principale rischio analizzato dagli istituti bancari e affini, la sua importanza rende la rapida valutazione della rischiosità della controparte una necessità. Sono così stati sviluppati numerosi modelli previsionali fin dai primi anni del ventesimo secolo, dai modelli statistici fino, appunto, ai più moderni modelli con metodologie di machine learning. Verrà quindi descritto come costruire un modello dall'acquisizione dei dati necessari fino alla sua messa in opera e valutazione delle performance.

Successivamente l'elaborato fornisce i concetti chiave per capire il funzionamento degli algoritmi genetici, descrivendo i principali operatori (riproduzione, incrocio e mutazione) e condividendo un esempio semplificato al fine di ottenere una maggior chiarezza espositiva. Una volta esposto il concetto di schema viene approfondita la definizione numerica dell'algoritmo in modo da poter dimostrare il Teorema Fondamentale degli AG e poterne derivare importanti considerazioni.

Infine, vengono presentati importanti risultati dei ricercatori che si sono cimentati nello sviluppo di modelli AG per il credit scoring, fornendo evidenza della loro applicabilità e delle performance ottenute. Queste saranno poi messe a confronto con i risultati ottenuti da

modelli costruiti con altri approcci al fine di dimostrare il contributo che questo algoritmo può fornire e la sua competitività.

Dal punto di vista personale la scelta del tema da trattare è stata condizionata dal quotidiano fascino scaturitomi dal processo esplorativo della naturale, ma anche artificiale, evoluzione. Questo processo cerca di essere imbrigliato all'interno di un algoritmo, accelerato ed esacerbato al fine di raggiungere uno stato ottimo di adeguatezza con l'ambiente circostante. Le capacità acquisite in materia economico-finanziaria e gli studi ingegneristici, esplorati durante l'esperienza accademica e lavorativa, mi hanno permesso di comprendere e apprezzare il complesso ramo della ricerca riguardante il machine learning. Questo ramo è attualmente in sviluppo, potenzialmente rivoluzionario e dalle applicabilità più eterogenee e trasversali.

*<<We shall see a little later that the possibility of imputing discovery to pure chance is already excluded…. On the contrary, that there is an intervention of chance but also a necessary work of unconsciousness, the latter implying and not contradicting the former…. Indeed, it is obvious that invention or discovery, be it in mathematics or anywhere else, takes place by combining ideas.>>*

Jacques Hadamard (1949)

# 2 IL RISCHIO

In questo capitolo verranno esposte le varie categorie di rischi che solitamente vengono valutate in ambito finanziario al fine di fornire un quadro complessivo, non esaustivo, ma utile all'introduzione del rischio principale di questa trattazione, il rischio di credito. Verranno esposte brevemente quelle che sono le caratteristiche della sua valutazione e si sottolineerà l'importanza di questo rischio per le istituzioni finanziarie e l'evolversi della Regulation di Basilea nel tempo.

Il termine "rischio" in italiano ha un'accezione negativa ma nella valutazione del rischio non è sempre così. Il rischio è un'aleatorietà che può portare ad eventi favorevoli così come sfavorevoli. È comunque fondamentale identificarli e capire la loro entità e il possibile impatto al fine di non essere sopraffatti dagli eventi avversi.

## 2.1 La gestione del rischio

La prima fase per una corretta gestione dei rischi parte innanzitutto dalla loro identificazione e ricordiamo che non sempre sono eventi negativi ma possono essere opportunità che bisogna essere pronti a cogliere.

Una volta identificati vanno analizzati cioè valutati nella quantità di impatto economico che possono apportare e nella probabilità di accadimento dell'evento. Tramite questi valori è possibile definire una matrice probabilità-impatto che guida nella scelta di gestione del rischio.

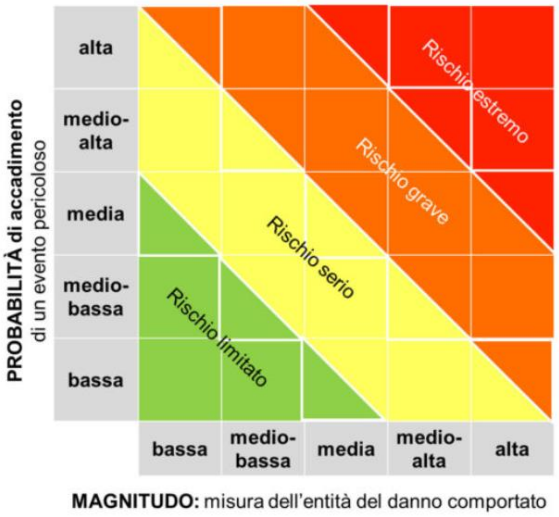


**Figura 2. 1 Matrice Probabilità-Impatto**[1]

Normalmente nella prassi aziendale prima di intraprendere un rischio si decide se è il caso di evitarlo totalmente se possibile, mitigarlo con azioni preventive, trasferirlo ad altri che magari lo gestiscono meglio o semplicemente addossarselo se di impatto e probabilità bassa.

In ambito finanziario una volta che sono stati assunti dei rischi con delle posizioni le azioni che si possono intraprendere per la loro gestione sono sempre sostanzialmente 4:

a) IGNORARE: soluzione utilizzata per i rischi ad alta frequenza e basso impatto
b) TRASFERIRE: tramite assicurazioni o prodotti derivati assimilabili come, per esempio, un CDS (Credit Default Swap)
c) COPRIRE/MITIGARE: tramite operazioni di hedging/diversificazione
d) ASSORBIRE: attraverso l'erosione dei profitti o con il *capitale*

Ed è proprio su quest'ultima risposta al rischio, ovvero la capacità di assorbire le perdite attese (e non) tramite la creazione di cuscinetti di capitale, si basano i propositi delle regulation di Basilea.

[1] Fonte: https://www.zerounoweb.it/

## 2.2 Rischio di mercato

Il rischio di mercato è sostanzialmente il rischio che il valore di qualche condizione di mercato muti diversamente da quanto atteso. Questi mutamenti sono dovuti al variare di fattori di mercato quali, per esempio, i tassi di interesse, i tassi di cambio o le loro volatilità generando così un aumento o una riduzione delle posizioni in strumenti finanziari e/o delle valute detenute.

Negli ultimi decenni il mercato finanziario è cresciuto notevolmente così come la complessità degli strumenti finanziari stessi, soprattutto i derivati, che qui vengono scambiati. I trading book delle imprese hanno assorbito questa complessità e il rischio di mercato ha assunto sempre più importanza data l'aumento di esposizione alle variazioni inattese delle condizioni di mercato.

Uno degli strumenti più utilizzati per la quantificazione di questo rischio è il *Value-at-Risk* (VaR). Il VaR è un unico valore che da solo identifica la perdita potenziale in modo che perdite più alte abbiano una probabilità di realizzarsi molto basse ed è per questo molto utile alla valutazione del capitale regolamentare.

Facendo riferimento al trading book valutato ai fini della vigilanza bancaria, il rischio di mercato comprende:

- Il rischio assunto nelle posizioni in strumenti di debito e di capitale: rischio che deriva dall'oscillazione del prezzo dei valori mobiliari dovuto alle oscillazioni del valore di mercato o alla situazione della società emittente.
- Il rischio di concentrazione: è dovuto ad una eccessiva esposizione verso un determinato strumento finanziario, un prodotto, un cliente o un insieme di questi tra loro fortemente correlati. Un portafoglio mobiliare non ben differenziato subisce maggiori fluttuazioni al variare di singoli fattori di mercato.

## 2.3 Rischio di liquidità

Il rischio di liquidità ha manifestazione evidente quando esiste uno scostamento temporale tra attività e passività a breve termine. Per coprire i debiti a breve termine con insufficienti attività mobilizzabili quindi di breve termine, la società sarà costretta ad erodere le attività di medio-lungo termine. L'uscita inattesa e affrettata di queste attività può comportare delle perdite.

Nell'identificazione del rischio di liquidità si possono evidenziare quindi due momenti e due relativi rischi che insieme concorrono a definire il rischio di liquidità:

- Il *funding risk* è la compromissione, dovuta alla cattiva gestione aziendale, della stabilità finanziaria dell'impresa che porta ad uno sbilanciamento nei confronti delle uscite finanziarie attese o una non precauzione rispetto alle possibili uscite inattese. Questo genera quindi il disallineamento tra passività di breve termine e attività di breve termine.
- Il *market liquidity risk* si presenta successivamente quando si va sul mercato cercando di liquidare qualcosa che non si aveva in programma di liquidare in quel momento (forced sale). Nella maggior parte dei casi porta ad un minore prezzo di realizzo oltre alle mancate entrate che ci si aspettava da quell'attività. L'effetto di questo secondo rischio è ovviamente dipendente dalla liquidità del mercato dell'attività in questione.

Il monitoraggio continuo della liquidità strutturale è volto a contenere la non congruità di scadenze che si genera fra attivo e passivo. La funzione di gestione della liquidità viene svolta dall'ufficio tesoreria che garantisce l'utilizzo ottimale delle risorse finanziarie, cerca di minimizzare il costo di raccolta fondi e mantiene stabile e diversificata la struttura finanziaria.

## 2.4 Rischio operativo

I rischi operativi non sono riconducibili a fattori di mercato o a eventi creditizi, bensì sono legati al fermo delle attività, al funzionamento dei sistemi informativi, alle policy interne, alle risorse umane o ad eventi esterni come per esempio furti, disastri ambientali, attacchi informatici o, come abbiamo visto recentemente, pandemie. Da qui la necessità per gli imprenditori di prevenire questi imprevisti, attraverso una valutazione dei rischi (Risk Assessment) e la pianificazione strategica in caso alcuni eventi si presentino e interrompano la normale produttività con la stesura del Business Continuity Plan e del Disaster Recovery Plan. Fin dal principio è necessario che l'impresa predisponga in anticipo come ristabilire il più in fretta possibile la normale operatività.

Il Business Continuity Plan è il documento con cui l'azienda stabilisce la strategia operativa da adottare per ripristinare la continuità aziendale in caso di eventi interruttivi non attinenti alla sicurezza IT, come un guasto macchina, un'azione legale o una momentanea poltrona libera nel management. Aiuta a gestire eventi critici in grado potenzialmente di minacciare la sopravvivenza dell'impresa e cerca di far ripristinare le attività nel più breve tempo possibile. Si tratta quindi di un piano che riassume i costi da sostenere, le attività strategiche da porre in essere e i referenti da coinvolgere per assicurare la continuità di servizi e profitti.

Il Disaster Recovery Plan fa parte del più ampio Business Continuity Plan ed è un documento specifico per affrontare gli eventi in grado di compromettere le funzionalità tecnologiche di sistemi critici e che trattano dati. È costituito dalle misure tecniche, logistiche e organizzative predisposte da un'azienda per ripristinare dati, applicazioni e sistemi informatici necessari per l'operatività del business.[2]

Il rischio operativo è un rischio che nel tempo è stato preso sempre di più in considerazione, un po' per imposizione del Comitato di Basilea che come vedremo dalla terza emanazione lo annovera tra i rischi da valutare per quantificare il requisito minimo di

[2] https://www.unocloudbackup.it/differenza-tra-disaster-recovery-plan-e-business-continuity-plan/

capitale, un po' perché ha nascosto storicamente molti cigni neri[3] e ovviamente per la ormai pervasiva presenza di sistemi informativi complessi nel settore finanziario.

Come detto il rischio operativo è proprio della realtà della singola azienda, non a fattori di mercato. Non si decide preventivamente a quali rischi operativi andare in contro, come quando per esempio si decide una posizione sui prodotti finanziari, ma sono intrinseci nell'attività stessa. Questo rischio ha inoltre la caratteristica di essere tendenzialmente negativo, ovvero di condurre principalmente a perdite.

Problema principale del rischio operativo più che la sua individuazione è la sua quantificazione che risulta complicata e perciò non accurata. Vi è inoltre pecunia di strumenti di *hedging* comunque riconducibile alla difficoltà di *pricing.* Gli strumenti principali sono prodotti assicurativi che proteggono da specifici eventi, questi però non possono essere esaustivi di ogni evento né tantomeno poter sanare gli effetti negativi non monetari degli eventi avversi come il danno reputazionale. È perciò molto importante per questo particolare rischio la prevenzione, limitare il più possibile che gli eventi si realizzino.

Il rischio operativo ha molte sfumature, è quindi possibile identificarne di comuni a tutte le imprese ma anche di specifici per settori o esclusivi di un'impresa sola. In ambito bancario si possono identificare le seguenti categorie di rischi operativi:

1) Reputazione
2) Strategico
3) Frodi interne ed esterne
4) Compliance
5) Sistemi informativi
6) Outsourcing

[3] Un cigno nero è considerato un evento del tutto inatteso e dagli impatti elevatissimi. (Talen, 2014)

### 2.4.1 Reputazione

Il rischio reputazionale ha un ruolo centrale poiché fondamentale in un mercato di scambi. Quando la storia di un'impresa riporta eventi di truffe da parte di dipendenti o dirigenti, di attacchi informatici o la partecipazione a scandali mediatici, clienti e investitori perdono fiducia almeno nel breve e medio termine con ripercussioni su tutto il business dell'impresa.

Tutti gli organi aziendali devono essere orientati al mantenimento della buona immagine e reputazione dell'azienda. Le scelte strategiche spesso sono volte ad accrescere questa caratteristica ed è importante vigilare grazie alla funzione svolta dall'audit sulla corretta operatività. Anche la scelta delle controparti viene esaminata con tali intenti poiché anche la caduta di reputazione di clienti o fornitori può inficiare ciò che è stato finora costruito.

Ex-post si cerca di creare un adeguato servizio di post-vendita in modo da soddisfare il cliente in ogni esigenza cercando di non incorrere in reclami. Questi a loro volta quando incorrono sono gestiti dall'ufficio reclami e la loro gestione deve essere celere ed efficace per mitigare il più possibile il danno

### 2.4.2 Strategico

Per quanto attiene al rischio strategico, ossia il rischio di erosione di utile/capitale derivante da scarsa o erronea reattività a variazioni del contesto competitivo, in fase di elaborazione della strategia di business, le scelte sono tendenzialmente orientate verso:

- innovazione tecnologica ed ottimizzazione operativa per incrementare i livelli di servizio
- consolidamento/crescita dimensionale al fine del raggiungimento di adeguate economie di scala

- ottimizzazione delle risorse finanziarie impiegate
- diversificazione dei prodotti collocati nonché dei canali distributivi

Fondamentale per la prevenzione di questo rischio, che è caratterizzato da alti impatti proprio perché direttamente collegato al livello di utile, è l'attività di monitoraggio che viene svolta da diverse funzioni organizzative. L'area marketing, per esempio, effettua costantemente analisi di posizionamento nel mercato rispetto competitor sul quale la strategia adottata ha posto il benchmark. Anche le altre funzioni svolgono analisi di adeguatezza alla strategia intrapresa, di rilievo in ambito bancario sono ovviamente il financial control con benchmark il business plan e il monitoraggio del rischio assunto rispetto al prestabilito Risk Appetite Framework (RAF).
Si tenga presente che tutte le scelte strategiche vengono approvate dal CdA ed è quindi rilevante il ruolo del monitoraggio in modo da poter reagire con strategie differenti in caso di necessità.

### 2.4.3 Frodi interne o esterne

Una frode può essere definita come tutto l'insieme di attività poste in essere a scopo ingannevole, direttamente o indirettamente, al fine di sottrarre valore al business e procurare un vantaggio a chi commette l'azione. È un rischio trasversale e profondamente eterogeneo, può essere compiuto internamente o esternamente all'azienda.

Le frodi interne vengono generalmente suddivise in tre macrogruppi: corruzione, appropriazione indebita e frodi di bilancio. Aspetto fondamentale è lo sviluppo di una struttura organizzativa adatta e lo svolgimento di attività volte al monitoraggio continuo.[4]

Le frodi esterne possono essere attuate da clienti o fornitori ma anche soggetti terzi come per esempio tramite attacco informatico. Per quanto riguarda il mondo bancario e

[4] https://www.filodiritto.com/le-frodi-aziendali-la-prevenzione-delle-frodi

assicurativo, per la prevenzione del rischio di frode esterna, come definito dal Decreto del Ministero dell'Economia e delle Finanze, 19 maggio 2014, è obbligatoria l'adesione al sistema SCIPAFI. Il sistema pubblico di prevenzione della frode consente il riscontro dei dati contenuti nei principali documenti d'identità e reddito con quelli registrati nelle banche dati degli enti di riferimento, quali l'Agenzia delle Entrate, il Ministero dell'Interno, il Ministero delle Infrastrutture e dei Trasporti, INPS e INAIL.[5]

### 2.4.4 Normativo e Compliance

Il rischio normativo si verifica quando si devono attuare mutamenti nell'impostazione del business al fine di adattarsi a nuove norme imposte dal legislatore. Per il rischio normativo riporto un rapido esempio dall'esperienza lavorativa personale, si tratta di una finanziaria che tra le attività di business eroga finanziamenti per prodotti di telecomunicazione. I finanziamenti erogati con contratti a 30 mesi con formule di rateizzazione atte ad avere un effetto lock-in per l'operatore partner hanno dovuto subire una modifica a seguito dell'esprimersi dell'AGCOM (Autorità per le Garanzie nelle Comunicazioni). La norma espressa ha portato ad una rimodellizzazione dei contratti ma non solo, l'adeguamento infatti deve essere attuato da tutte le direzioni di business, marketing, IT, finance.

Con rischio di compliance si intende il rischio derivante dal non adeguamento a norme aziendali e del legislatore, a regole o standard riconosciuti. Il fine è evitare di incorrere in sanzioni, perdite finanziarie o danni di reputazione. Nelle società normalmente si viene a creare a livello organizzativo una funzione di compliance che svolge attività volte a:

- Monitorare le norme che diventano applicabili al business e la misurazione del loro impatto sui processi;
- Suggerire modifiche procedurali al fine di assicurare l'adeguatezza alle norme
- Verificarne l'efficacia una volta applicate

[5] https://www.experian.it/business/identita-e-frode/prevenzione-frodi/scipafi

Vengono presidiate le norme definendo a priori quelle a maggior impatto e che quindi se non rispettate pedissequamente possono arrecare maggior danno. Per fare sempre degli esempi della realtà finanziaria e di concessione del credito è di rilievo la disciplina a tutela del consumatore, la trasparenza e i conflitti di interesse. (Banca d'Italia, 2007)[6]

### 2.4.5 Sistemi informativi

I sistemi informativi largamente utilizzati e ormai strutturali in qualsiasi attività comportano il rischio operativo dovuto al loro malfunzionamento che potrebbe comportare l'interruzione dell'attività di business. Una corretta gestione del fattore errore umano nell'utilizzo delle risorse informatiche prevede che ci sia compartimentazione e gerarchizzazione nell'utilizzo delle funzionalità degli applicativi più impattanti e che comunque l'attività sia monitorata.

Ancora più importante è la messa in sicurezza dei dati raccolti da fuoriuscite involontarie o da furti per attacchi perpetrati da terzi.

È molto frequente, soprattutto nelle società che raccolgono informazioni personali, essere presi di mira da hacker che cercano di penetrare le difese informatiche, per esempio, con le più classiche attività di phishing ai dipendenti per ottenere accesso alla rete aziendale, ma le tipologie di attacchi informatici sono varie. Spesso viene istituita nella struttura organizzativa una funzione specifica atta a mantenere la sicurezza informatica.

### 2.4.6 Outsourcing

Il rischio di Outsourcing è fondamentalmente il rischio che il fornitore non esegua correttamente quello per cui è stato ingaggiato o che il suo cattivo operato generi ripercussioni sul business dell'azienda. È importante per mitigare questo rischio

[6] Banca d'Italia (2007). Disposizioni di vigilanza in materia di conformità (compliance)

un’attenta esecuzione della fase contrattuale, soprattutto la definizione di indicatori efficaci nella valutazione del service level fornito, la previsione degli scenari rilevanti e le obbligazioni delle parti nel caso del loro avverarsi o del conseguimento degli obiettivi prefissati. Il controllo sull’operato dei fornitori tramite il rispetto del service level contrattualizzati sono di competenza degli uffici aziendali che si avvalgono del servizio all’interno della parte di processo di cui sono responsabili.

## 2.4 Rischio di credito

Principalmente il rischio di credito è dovuto ad una variazione inattesa del valore di mercato del credito che è dovuta ad una variazione del merito creditizio della controparte in debito. È il rischio tipicamente centrale per le banche e tutte quelle società che forniscono credito nelle sue diverse modalità, dal microcredito al consumo per l’acquisto di uno smartphone, ai mutui o ai più grandi finanziamenti alle imprese. Le cause della variazione del merito di credito di una controparte sono varie e di diversa natura, il concetto base è che con il peggiorare della qualità di un credito aumenta la probabilità di insolvenza o di default, ovvero che il credito non venga ripagato nella sua interezza.

<<*If you owe your bank a hundred pounds, you have a problem. But if you owe a million, it has.*>> cit. John Maynard Keynes

Prendendo ad esempio l’erogazione di un mutuo da parte di una banca, prima che venga emesso il credito, la banca effettua delle valutazioni ex-ante sul merito creditizio dell’intestatario del mutuo ed il pricing del credito viene condizionato da questa valutazione. Chi richiede il mutuo sosterrà un costo per il finanziamento ricevuto, il tasso d’interesse congruo al rischio di insolvenza nel momento in cui il mutuo è erogato. Se ci si aspetta un peggioramento della qualità creditizia, si chiederà un premio per il rischio maggiore. Il vero rischio di credito l’ente finanziatore lo corre temporalmente dopo l’erogazione del finanziamento, esso è infatti dovuto alla variazione inattesa del merito creditizio durante la

vita del mutuo che porterà ad una perdita di valore del credito in mano alla banca, perché chi sarà disposto ad acquistare il credito in questo secondo momento, valutando il rischio ora, offrirà un prezzo più basso rispetto al valore preventivato dalla banca emettitrice.

È perciò fondamentale analizzare per valutare nel modo più accurato possibile la controparte e stabilire il pricing corretto fin dal principio. È infatti la scorretta valutazione anteriore all'emissione una delle cause di una variazione inattesa a posteriori oltre le nefaste insorgenze di nuove problematiche economico-finanziarie riguardanti la controparte ma anche magari a causa dell'intero sistema economico in generale. Per effettuare la miglior stima della qualità creditizia di un soggetto è però necessario nella maggior parte dei casi raccogliere un numero esagerato di informazioni. Raccogliere informazioni ha un costo crescente con il crescere della profondità che si vuole raggiungere nell'analisi. Nasce quindi il trade-off per i costi da sostenere ex-ante per una migliore analisi e che andranno ad erodere i profitti dell'operazione in una realtà competitiva dove, comunque, chi deve farsi finanziare andrà a scegliere il finanziamento offerto alle migliori condizioni.

È da notare quanto sia più evidente questo problema della raccolta delle informazioni per quelle società dove i rapporti fra banche e imprese sono riassumibili dal termine *Transaction Banking,* dove cioè la relazione tra banca e impresa è improntata alla convenienza ad effettuare una specifica operazione. In questo contesto la banca non ha interesse ad approfondire la conoscenza dell'impresa, affrontando i costi necessari a farlo, ma preferisce tutelarsi con garanzie, diversificando i portafogli o cedendo il rischio scorporandolo dal credito stesso. Inoltre, nei momenti di difficoltà di una controparte, la banca cercherà di svincolarsi il più velocemente possibile al fine di evitare l'aggravarsi delle perdite.

Di contro, nelle società in cui il rapporto tra la banca e l'impresa è caratterizzato da una relazione di lungo periodo si parla di *Relationship Banking.* Come si può intuire una banca ha maggior interesse nel raccogliere informazioni e mantenerle aggiornate nel tempo sicura del fatto che potrà ammortizzarli grazie alle numerose operazioni che porterà a termine nel lungo termine e l'impresa potrà beneficiare dell'appoggio di una banca che non ha nessun interesse a lasciarla in brutte acque perché perderebbe l'investimento fatto fino a quel punto.

Ne deriva, pertanto, che:

- rischio di credito non significa solo possibilità di insolvenza della controparte, in quanto anche il semplice deterioramento del merito creditizio deve considerarsi una manifestazione del rischio;
- un'attenta valutazione delle controparti è attività imprescindibile per la corretta allocazione degli impieghi.

Il processo di emissione del credito, al fine di mantenere la stabilità degli intermediari finanziari, è normalmente caratterizzato dalle seguenti fasi:

1. Istruttoria, l'acquisizione delle informazioni e della documentazione del cliente;
2. Delibera a seguito della valutazione del merito creditizio anche tramite la consultazione dei SIC;
3. Liquidazione della quota che si è deciso di finanziare
4. Monitoraggio dell'andamento del credito, delle esposizioni deteriorate e la loro gestione

La fase di istruttoria è l'acquisizione delle informazioni necessarie ad effettuare una corretta valutazione del merito creditizio ex-ante del cliente e la sua corretta esecuzione è la base per un data entry efficiente, utilizzabile nel calcolo automatico del punteggio di scoring per ogni singola pratica inserita ma anche per poter interrogare correttamente i SIC (Sistemi di Informazioni Creditizie) e con i dati andamentali relativi a possibili pratiche già in essere con lo stesso cliente.

Lo scoring attribuito ad un cliente o una società è un indicatore elaborato da un algoritmo che utilizza una combinazione di informazioni disponibili al momento della richiesta di finanziamento. Il suo calcolo ha come componente principale la definizione della "Probability of default" (PD), ossia la probabilità che il richiedente finanziamento diventi insolvente. Definito il rating della controparte e delineate delle fasce di accettazione sui rating raccolti è possibile definire se accettare o meno di emettere il finanziamento. È sempre possibile rettificare l'esito dello score attribuito tramite operazioni di modifica successive, definite azioni di override.

Il monitoraggio del rischio di credito dopo l'accettazione avviene attraverso una reportistica di dettaglio in grado di garantire l'intercettazione di eventuali andamenti anomali. I principali Key Risk Indicator vengono raccolti con periodicità definite; alcuni esempi sono:

- Delinquency, percentuale di finanziamenti con almeno due pagamenti insoluti contenuti nel portafoglio;
- Recovery Rate, valore stimato di recupero del credito in caso di default;
- Bad Rate, i definiti "cattivi pagatori" diviso i "buoni pagatori" presenti in portafoglio;
- Reject Rate, tasso di rifiuto dei prestiti richiesti, spesso anche valutato per canale di acquisizione, zone o metodo di pagamento utilizzato;
- No-start Rate, percentuale di pratiche liquidate che sono insolute fin dalla prima rata.[7]

Nel caso gli indicatori segnalino un andamento della posizione di rischio peggiorata vengono intraprese azioni mirate alla mitigazione del rischio assunto tramite la rielaborazione dei parametri di accettazione o direttamente l'innalzamento del cut-off di accettazione. Inoltre, dal R.A.F. sono definite le soglie di Risk Appetite, Risk Capacity e Risk Tolerance per tutti i rischi e definiscono i benchmark da tenere in considerazione.

### 2.4.1 Expected Loss

L'Expect loss è il valore atteso della distribuzione delle perdite. Poiché la EL rappresenta qualcosa di prevedibile è inclusa in fase di pricing del credito, seguendo la logica già descritta che rischio più alto definisce perdita attesa maggiore, quindi costo maggiore per potersi finanziare ovvero tassi più alti.

Per il calcolo dell'EL si segue la formula:

$$EL = EAD * LGD * PD \tag{2.1}$$

[7] https://www.investopedia.com/

Dove EAD (Exposure At Default) è l'ammontare dell'esposizione, LGD (Loss Given Default) è la percentuale dell'esposizione che non è possibile recuperare in alcun modo e che quindi definisce la vera perdita in caso di default, il tutto ancora moltiplicato per la probabilità che il default si verifichi (PD).

La LGD è anche definibile come il complemento a uno del Recovery Rate, che quindi è la percentuale di exposure che si crede di recuperare anche in caso di default:

$$LGD = 1 - RR \tag{2.2}$$

L'EL è quindi direttamente proporzionale all'EAD e alla PD ma se nel caso limite il Recovery Rate fosse pari a 1, il che vuol dire che anche in caso di default il credito verrebbe interamente recuperato, allora ovviamente l'EL sarebbe nulla. Garanzie, ipoteche, fideiussori hanno proprio l'obiettivo di mantenere il più alto possibile il Recovery Rate.

### 2.4.1 Exposure-At-Default

È l'entità del finanziamento che tuttavia non è sempre facile da individuare. Lo è per esempio per un mutuo o un prestito, che non dipendono nel tempo da scelte del debitore dove quindi l'EAD corrisponde all'ammontare finanziato e durante il tempo corrisponde al capitale a scadere non ancora restituito. Diventa più complesso in caso, per esempio, di un fido bancario dove una determinata somma è messa a disposizione del cliente ed è lui a decidere quando e quanto utilizzarne. Questa aleatorietà sull'ammontare prestato rende più difficile identificare a priori, ovvero prima della concessione del fido, l'EAD per il calcolo dell'EL e quindi per la definizione del costo da richiedere per mettere a disposizione la somma. L'esposizione effettiva sarà nota al creditore solo nel momento del default ed è anche da tenere in considerazione che in prossimità del default il debitore probabilmente consumerà quanto più riesce del fido concessogli. Un metodo per il calcolo dell'exposure di un fido è:

$$EAD = DP + UP * UGD \tag{2.3}$$

Dove Drawn Portion (DP) è la quota di fido attualmente utilizzata a cui si somma la Undrawn Portion (UP), quota non utilizzata, moltiplicata per la percentuale che ci si aspetta venga utilizzata precedentemente al default, l'Usage Given Default (USD).

### 2.4.2 Loss Given Default

Per poter capire come definire la Loss Given Default, bisogna passare dal Recovery Rate e analizzare i fattori che ne caratterizzano un aumento o una diminuzione. Fattori principali sono:

- Il tipo di credito, per esempio se o meno il finanziamento è coperto da garanzie o da assicurazioni terze. Quando il credito è rivolto a imprese il livello di seniority del credito, definisce le priorità di rimborso in caso di default dell'impresa debitrice,
- Le caratteristiche della banca, come la propensione al rischio oppure la capacità della Collection Business Unit che si occupa del recupero crediti;
- Le caratteristiche dell'impresa finanziata. Un'impresa con all'attivo molti asset di valore anche nel mercato secondario fornisce più garanzia di recupero maggiore rispetto ad una società con un attivo immateriale e dubbio valore in caso di default;
- I fattori esterni, come per esempio degli aspetti normativi eccezionali come visto di recente in occasione dei prestiti garantiti dallo stato a seguito del lockdown ma anche lo stato del ciclo economico in corso.

Per il calcolo del RR, come visto ci sono molti elementi difficilmente quantificabili, ma una volta arrivati ad una stima dell'Expected Recovery (ER) è possibile utilizzare la seguente formula:

$$RR = \frac{\sum_{t=1}^{n} \frac{ER_t - AC_t}{(1+i)^t}}{EAD} \tag{2.4}$$

Dove AC sono i costi amministrativi del periodo, in genere per effettuare il recupero, *i* è il tasso di attualizzazione; il tutto calcolato al momento del default e quindi *n* è il tempo

stimato per completare il recupero. Generalmente il tasso di attualizzazione utilizzato può essere:

- Il tasso di Interesse risk-free
- Il tasso del finanziamento andato in default
- Il tasso pagato dalla banca per finanziarsi (tasso base)
- Un tasso congruo per il rischio calcolato tenendo conto dei rischi sull'Expected Recovery

Data la difficoltà di calcolo c'è il rischio che l'effettivo recupero differisca da quello stimato con questo metodo analitico.

Utilizzando un approccio statistico su dati raccolti in precedenza è corretto utilizzare una distribuzione BETA per il calcolo del Recovery Rate, utilizzando così media e volatilità dei RR osservati e fare una stima più robusta dei futuri.

### 2.4.2 Default Probability

Rappresenta la probabilità che la controparte debitrice vada incontro ad un evento di default che comporta l'impossibilità di ripagare interamente o parzialmente il debito contratto. Bisogna però definire cosa si intende per evento default, in che momento un credito può essere considerato del tutto insolvente.

Dal 1° gennaio 2021 è entrata in vigore la nuova definizione di default prevista dal Regolamento europeo relativo ai requisiti prudenziali per gli enti creditizi e le imprese di investimento (articolo 178 del Reg. UE n.575/2013). Ai fini del calcolo dei requisiti minimi obbligatori, i debitori sono classificati come deteriorati al ricorrere di almeno una delle seguenti condizioni:

a) Il debitore è insolvente da almeno 90 giorni (180 per le amministrazioni pubbliche) se l'obbligazione è "rilevante" (a seguire si delinea cosa si intende per rilevante);
b) La banca giudica improbabile che, senza il ricorso all'escussione delle garanzie se presenti, il debitore adempia per intero.

Un debito scaduto va considerato *rilevante* quando l'ammontare dell'arretrato supera entrambe le seguenti soglie:

1. *Soglia assoluta:* 100 euro per le esposizioni al dettaglio, altrimenti 500 euro;
2. *Soglia relativa:* l'1% dell'esposizione complessiva.

Superate entrambe le soglie inizia il conteggio per i 90, o 180, giorni consecutivi di scaduto che portano al default del credito.[8]

Per poter definire la probabilità che un default si verifichi vengono utilizzati modelli statistici ma anche algoritmi più complessi che sfruttano le maggiori capacità di calcolo e le potenzialità del machine learning. Il risultato è comunque una probabilità compresa tra 0% e 100%, intervallo che per normativa viene poi suddiviso in classi omogenee di rischio, e quindi con una determinata PD, atte al collocamento delle posizioni assunte.

Una banca o un ente finanziatore potrebbe non disporre di queste capacità internamente, quindi la normativa prevede differenti tecniche che si possono utilizzare per il calcolo della PD (articolo 180 del Reg. UE n.575/2013) utilizzabili anche in forma ibrida:

- È possibile stimare la PD tramite i dati storici dei default che si sono verificati nel passato. Se i dati non sono sufficienti si deve applicare un margine di cautela;
- È possibile utilizzare scale di rating importate dagli ECAI (External Credit Assessment Institution);
- Utilizzando modelli statistici che rispettino determinati criteri

### 2.4.2 Unexpected Loss

Abbiamo definito la perdita come una variabile casuale e l'EL come il valor medio della distribuzione delle perdite; la perdita inattesa è la volatilità della perdita attorno al valor medio.

L'UL rappresenta il vero fattore di rischio di credito proprio perché aleatorio.

[8] https://www.bancaditalia.it/media/fact/2020/definizione-default/index.html

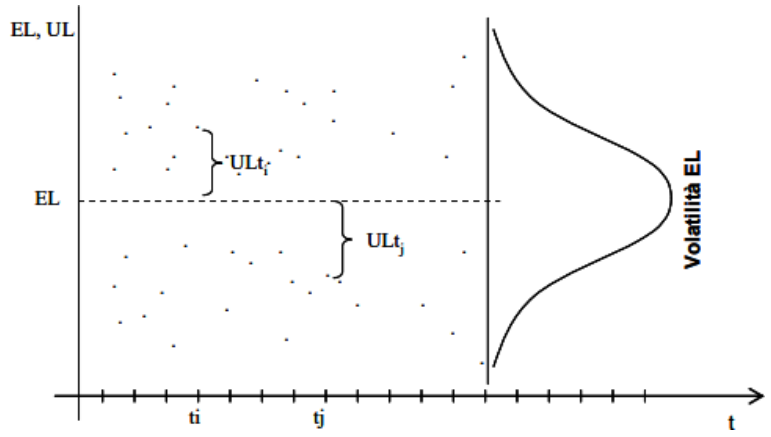


**Figura 2. 2 Evidenza dei possibili valori dell'UL**[9]

Ogni singolo credito nel portafoglio concorre in piccola parte alla volatilità delle perdite, così prende un ruolo centrale alla riduzione delle perdite inattese la diversificazione. La diversificazione si attua detenendo in portafoglio strumenti poco correlati tra loro in modo che se un particolare titolo, o settore per esempio, avesse un calo inaspettato, questo non si ripercuota su tutto il portafoglio. Minore è la correlazione tra le perdite inattese delle varie posizioni minore sarà la perdita inattesa del portafoglio globale.

In figura è possibile osservare la distribuzione di probabilità delle perdite, la curva non è simmetrica, si verificano frequentemente perdite minori e con probabilità più basse le perdite di gran lunga più impattanti. L'area A rappresenta la perdita attesa che viene coperta mediante il corretto pricing del titolo e ad opportuni accantonamenti, l'area B fino alla perdita inattesa deve essere coperta da capitale. Il limite superiore dell'intervallo di fiducia normalmente esclude le perdite catastrofiche che si verifica solo con l'1% di probabilità.

[9] Fonte: Varetto, corso di "Mercati Rischi e Strumenti Finanziari", A.A. 2020/2021, Politecnico di Torino.

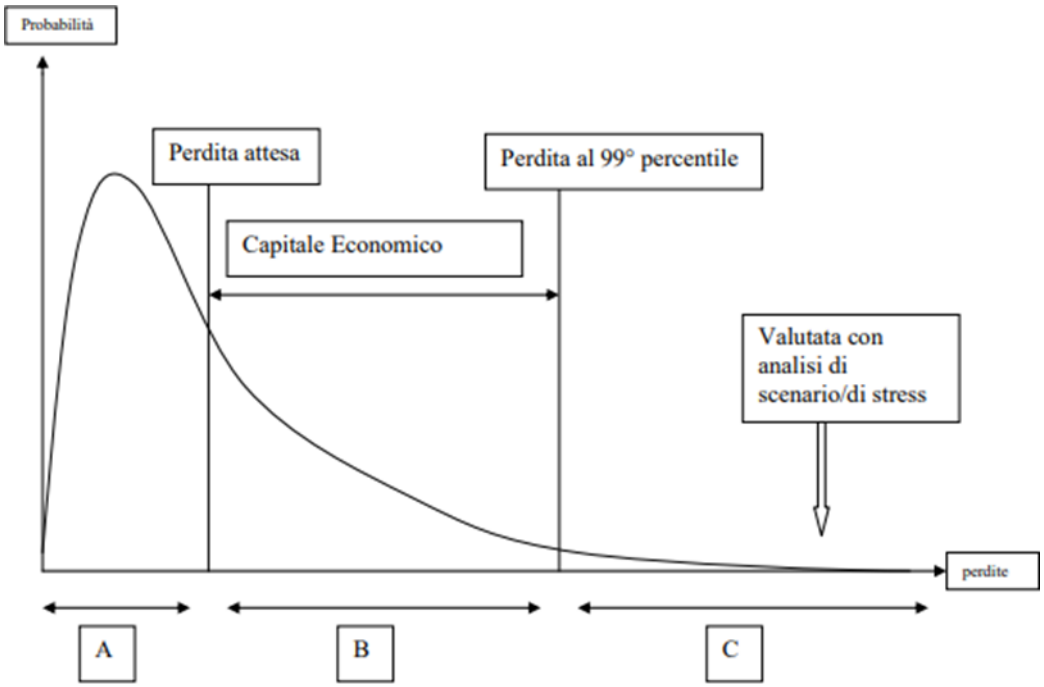


**Figura 2. 3 Copertura delle perdite con capitale di rischio**[10]

Per l'effettivo calcolo dell'Unexpected Loss, il modello di riferimento è quello binomiale che prevede due possibili eventi, l'evento default che si verifica con probabilità PD e l'evento non default con probabilità 1-PD. Di seguito la matrice con i payoff, si tenga presente che l'orizzonte temporale preso in considerazione su cui sono valutate le variabili è di un anno.

| | EVENTI | |
|---|---|---|
| | Default | Non default |
| Probabilità | PD | 1-PD |
| Perdita | LGD | 0 |

**Figura 2. 4 Matrice eventi e probabilità di accadimento**[11]

Supponendo deterministica la LGD e vincolando l'EAD a 1 per semplicità, bisogna calcolare valore atteso e scarto quadratico medio della distribuzione binomiale:

$$E[X] = EL = PD * LGD + (1 - PD) * 0 = PD * LGD \qquad (2.5)$$

[10] Fonte: Varetto, corso di "Mercati Rischi e Strumenti Finanziari", A.A. 2020/2021, Politecnico di Torino.
[11] Fonte: Varetto, corso di "Mercati Rischi e Strumenti Finanziari", A.A. 2020/2021, Politecnico di Torino.

$$\sigma = UL = LGD * \sqrt{PD * (1 - PD)} \tag{2.6}$$

Mentre, considerando la LGD, come variabile casuale lo scarto quadratico medio si ottiene applicando:

$$UL = \sigma = \sqrt{PD * (1 - PD) * LGD^2 + PD * \sigma_{LGD}^2} \tag{2.7}$$

Dove $\sigma_{LGD}$è lo scarto quadratico medio della distribuzione della Loss Given Default. Ricordiamo che l'exposure è supposta pari a 1.

Questa era la valutazione per un singolo credito, per poter espandere la valutazione ad un portafoglio crediti è possibile utilizzare le seguenti formule:

$$EL_p = \sum_i a_i * EL_i \tag{2.8}$$

$$UL_p = \sqrt{\sum_i \sum_j a_i a_j UL_i UL_j \rho_{ij}} \tag{2.9}$$

Dove $a_i$ e $a_j$ sono le quote di portafoglio del credito i-esimo e j-esimo del portafoglio preso in esame e $\rho_{ij}$ la correlazione tra le rispettive perdite inattese.

### 2.4.3 Tipologie di Rischio di Credito

Il rischio di credito, data l'importanza per gli istituti e per l'economia, viene analizzato in tutte le sue sfaccettature. Viene quindi suddiviso in numerose sottocategorie:

- Rischio di insolvenza, rischio che la controparte non restituisca l'importo del credito emesso;
- Rischio di migrazione o downgrading, con il deteriorarsi del merito creditizio della controparte diminuisce il valore di mercato del credito stesso;

- Rischio di esposizione, che al momento del default l'esposizione sia alta;
- Rischio di concentrazione, dovuto alla bassa diversificazione di portafoglio;
- Rischio paese, esposizioni verso controparti con sede in paesi instabili:
- Rischio di sostituzione, sostituzione dei derivati con sottostanti i crediti nel caso durante la vita dello stesso derivato si debba andare sul mercato a cercarne un altro ma in una posizione più svantaggiosa rispetto la precedente. Un esempio può chiarire questo punto: ho un credito sul quale firmo un CDS con ente terzo alla stipula, l'ente terzo va in default e io dovrò rinegoziare un secondo CDS in un momento in cui il credito magari è deteriorato; troverò condizioni svantaggiose e quindi pagherò un premio più oneroso;
- Rischio di spread, nel caso cresca l'avversione al rischio del mercato i tassi aumentano provocando una diminuzione del valore dei crediti in essere.

## 2.5 Il rischio sistemico

Inserisco qui una definizione del rischio sistemico fondamentale per comprendere la ratio dietro la regolamentazione che affronteremo successivamente. Il rischio sistemico non è un rischio finanziario come il rischio di credito, il rischio di mercato o di liquidità e non è neanche un rischio operativo. Il perimetro di osservazione non è quello interno alla società e dei fattori che hanno effetto su di essa ma al contrario ingloba tutti i perimetri dei componenti di tutto il tessuto economico-finanziario. Lo scopo della valutazione del rischio sistemico è quello di mantenere la stabilità e continuità del sistema economico globale. Viene infatti definito come il rischio di un'ampia rottura dell'equilibrio del sistema finanziario, spesso sotto forma di una serie di insolvenze correlate che si verificano in un breve periodo di tempo, con la possibilità che le difficoltà di un singolo operatore siano trasmesse ad altri, da un segmento di mercato ad un altro.

Una crisi sistemica può per esempio essere innescata da una crisi di liquidità; una società non riesce ad onorare i propri debiti e cerca di liquidare le proprie attività il più in fretta

possibile, causando un crollo dei prezzi con relative perdite per altri operatori e per chiunque abbia garanzie o derivati collegati. Società che potrebbero, a loro volta, diventare a rischio di insolvenza.

La crisi finanziaria ha messo in evidenza come questo rischio sia ancora più pericoloso qualora vi siano delle interconnessioni finanziarie dirette come nel caso delle banche che scambiano denaro fra di loro quotidianamente.

# 3 MODELLI DI SCORING

Emettere credito è uno dei core business delle banche di tutto il mondo. La prima fondamentale decisione da prendere è se concedere un prestito ad un potenziale cliente. È quindi primaria la necessità di poter differenziare in modo accurato i buoni pagatori dai cattivi pagatori. Questa capacità è limitata dalle informazioni disponibili alla banca nel momento in cui il finanziamento viene richiesto e dalla necessità di determinare score per una grande platea di potenziali clienti. Sono stati sviluppati molti modelli di scoring per assistere le banche a svolgere questa attività in modo automatico e i più comuni sono basati sulla regressione logistica o l'analisi discriminante lineare.

I modelli di scoring sono utilizzati per prevedere l'insolvenza di una possibile controparte utilizzando metodologie di carattere statistico-quantitativo. Lo score associato alla clientela vuole rappresentarne l'affidabilità creditizia e al contempo posizionarla in una classe di rating. Il rischio di default di ogni singola controparte è anche importante in ottica regolamentare, ogni finanziamento incide infatti sul rischio dell'intero portafoglio detenuto, portafoglio monitorato dagli organi di vigilanza.

Nel corso del capitolo verranno esposti le principali metodologie e il processo con il quale sviluppare e testare un modello.

## 3.1 Requisiti di un modello

Per sviluppare un modello è necessario procurarsi innanzitutto un campione di soggetti su cui lavorare. Il campione deve essere rappresentativo della popolazione e perciò includere sia soggetti "buoni" che soggetti "cattivi", di questi inoltre si devono possedere i dati storici. Dai dati in possesso si identificano le variabili quantitative, nei

modelli più complessi vengono utilizzate anche variabili qualitative, che meglio discriminano le controparti insolventi dalle solventi. Nei modelli più semplici lo score per ogni controparte è il risultato di una funzione con variabili quelle selezionate precedentemente, lo score è un numero per definizione ordinale ed è quindi possibile fare una classifica delle parti dalle meno rischiose alle più rischiose. La scala di punteggio così ottenuta può essere suddivisa in fasce per poter, in future applicazioni, poter associare direttamente una parte alla classe e prendere una decisione. Il valore soglia, cut-off, è la linea di demarcazione, oltrepassata scatta il rifiuto della richiesta di finanziamento. La scelta del valore soglia dipende dalla propensione al rischio decisa a livello strategico per il sostenimento del business sempre rispettando i limiti imposti dal regolatore.

Obiettivo del modello è principalmente la stabilità delle performance previsionali nel tempo e nelle diverse fasi del ciclo economico, si rende quindi necessaria una verifica periodica dell'affidabilità del modello.

## 3.2 Approcci metodologici

Negli ultimi decenni molta letteratura è stata sviluppata attorno alla creazione di modelli previsionali della stabilità finanziaria di una controparte. Questa grande varietà rende anche complicato classificarli in macrocategorie. Considero adeguata la classificazione di alto livello che effettua una suddivisione in modelli statistici, modelli intelligenti e ibridi.

I principali modelli statistici sono l'analisi univariata, l'analisi discriminante lineare o il modello logit. Essi sono basati su teorie statistiche e sono caratterizzati dalla facilità di utilizzo e dal basso dispendio di tempo e risorse. Beaver propose in prima battuta i modelli univariati che analizzano l'impresa affidandogli uno score sulla base di un indicatore preso in esame. Egli riuscì a dimostrare la correlazione tra valori di alcuni

indici e la probabilità di default e che più era vicino l'evento maggiore era il gap fra indicatori per le imprese sane e quelli per le imprese anomale. Altman per primo utilizzò la 'multivariate discriminant analysis' (MDA) che, come suggerisce il nome, tiene in considerazione contemporaneamente diverse variabili. Utilizzando questo modello Altman costruì il famoso modello Z-score, una funzione lineare composta da 5 indicatori finanziari che fornisce buoni risultati nella previsione di default ad un anno dall'evento. I modelli MDA sono modelli di regressione lineare che assumono indipendenza tra le variabili e matrici di covarianza tra popolazione sana e popolazione anomala uguali fra di loro. Se i dati del campione non rispecchiano queste due assunzioni il modello non è applicabile anche se nella pratica si ottengono risultati attendibili anche in caso non siano rispettate. Ulteriore limite è dovuto alla definizione della variabile dipendente della regressione che è continua e non significativa per valutare la probabilità di default che essendo una probabilità deve cadere nell'intervallo [0,1]. Risolvono in gran parte questi problemi i modelli basati sulla regressione logistica che definiscono la variabile dipendente all'interno dell'intervallo consono ad una probabilità. Inoltre, non sono necessarie le assunzioni che caratterizzano i modelli MDA, unica assunzione al modello è la non multicollinearità delle variabili, che significa la non presenza di una forte relazione lineare tra le variabili dipendenti.

I modelli intelligenti sono, invece, tutti quei modelli che utilizzano algoritmi volti a ricercare la soluzione ottima al problema senza informazioni aggiuntive ad eccezione dei dati stessi che sono forniti al problema. Stilandone un elenco non esaustivo si annoverano le metodologie basate su neural network, case-based reasoning, support vector machines, decision tree, rough set theory e evolutionary algorithm (EA) tra cui gli algoritmi genetici. Vantaggio principale è dovuto all'assenza di stringenti assunzioni sulla distribuzione dei valori delle variabili richieste dai modelli statistici.

La maggior parte degli argomenti qui citati esulano dallo scopo di questa trattazione, eccezion fatta per gli algoritmi genetici.

## 3.3 Fasi del processo di stima di un modello di scoring

Il processo di stima è diviso nelle seguenti fasi:

1. Selezione del campione;
2. Selezione delle variabili di analisi;
3. Pre-processing dei dati;
4. Stima del modello;
5. Calibrazione e Master Scale;
6. Backtesting.

### 3.3.1 Selezione del campione

La prima fase prevede la selezione di un numero sufficiente di soggetti (supponiamo imprese), che vengono suddivise in gruppi (sane o anomale), sulla base di una predeterminata definizione per la discriminazione dei gruppi. Questi sono poi identificati da una variabile binaria. È necessario avere a disposizione i dati di un elevato numero di imprese insolventi, al fine di garantire risultati significativi. Paradossalmente una banca che ha concesso più crediti a clienti insolventi sarà avvantaggiata rispetto a una che ne ha concessi di meno, in quanto potrà ottenere un modello previsionale più efficace.

Una volta che è stato selezionato il campione molto generico è necessario ricavarne un portafoglio necessario a stimare un modello coerente. Per farlo si seguono dei drivers di segmentazione come possono essere i codici sintetici SAE (Settore di Attività Economica) e RAE (Ramo Attività Economica) forniti da Banca d'Italia, il codice ATECO (ATività ECOnomica) forniti dall'ISTAT, il fatturato, la rischiosità, l'area geografica.

Deve inoltre essere scelta una distanza temporale dal default in cui analizzare le variabili che verranno scelte. Come detto precedentemente, parte del campione sarà composto da imprese anomale e parte da imprese sane, un modello predittivo deve però prendere i valori delle variabili antecedentemente al default. Due approcci diversi sono:

- l’utilizzo di una finestra temporale fissa, utilizzando le informazioni per esempio ad un anno dal default dell’impresa;
- Utilizzare una finestra temporale variabile, selezionando una data di riferimento fissa per tutta la popolazione e utilizzare i dati del campione a quella data indipendentemente dalla data di default. Nel campione vi saranno imprese andate in default dopo 1 mese o magari dopo 2 anni.

### 3.3.2 Selezione delle variabili di analisi

La fase di identificazione del set di indicatori da usare nel modello deve essere in grado di valutare quelle più rilevanti. I dati utilizzabili possono essere:

- Dati anagrafici ovvero informazioni personali;
- Sociodemografici, riguardanti l’aspetto ambientale in cui è inserito il soggetto;
- Qualitativi, come la struttura organizzativa o la tipologia di business e il settore in cui opera, ultimamente viene utilizzato l’indice ESG (Environmental, social and governance) che cerca di racchiudere in un univo valore molti aspetti considerati qualitativi;
- Dati di bilancio, che forniscono gli indicatori classici utilizzati nell’analisi fondamentale di un’impresa
- Andamentali interni, raccolti nel tempo dal rapporto banca-impresa intercorso
- Andamentali esterne, informazioni sul rapporto tra l’impresa e il sistema bancario in generale e sono forniti dalle banche dati autorizzate come la Centrale dei Rischi appartenente alla Banca d’Italia o società come CRIF o Experian.

Ovviamente il numero di variabili potenzialmente utilizzabili è altissimo, ma in un modello che ha bisogno di stabilità nel tempo e la necessità di essere generalizzabile a

diversi soggetti con diversi parametri non può sfruttarli tutti, ne risulterebbe infatti ridotta l'efficacia incorrendo nel rischio di produrre un modello che sia overfitting ovvero troppo calibrato sul campione utilizzato e non efficiente su altri soggetti. Questo però esula dallo scopo che infatti è creare un modello dal campione al fine di utilizzarlo con informazioni che ancora non abbiamo ed estenderlo a tutta la platea di possibili clienti. Conviene, quindi, eliminare dall'analisi le variabili che presentano una correlazione molto intensa tra loro. Un metodo per farlo consiste prima di tutto nel calcolare i valori di media e varianza delle variabili prese in esame assumendo una distribuzione normale, dopodichè si procede al calcolo dell'indice di correlazione di Pearson:

$$\rho = \frac{cov_{xy}}{\sqrt{var_x \cdot var_y}} \tag{3.1}$$

Dove:

- $cov_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ è la covarianza tra le due variabili *x* e *y*
- $var_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$ è la varianza di una variabile
- $n$ è il numero di elementi nel campione

Calcolando i valori $\rho$ per ogni coppia di variabili è possibile costruire una matrice triangolare in cui sono riportati tutti gli indici di correlazione. Essendo la correlazione di una variabile con sé stessa pari a 1 i valori in diagonale saranno tutti 1 mentre per i restanti incroci l'indice di Pearson assume valori compresi tra -1 e 1. Valori prossimi a 1 e a -1 indicano una forte correlazione tra le variabili in esame rispettivamente positiva o negativa ed in questo caso è preferibile eliminare una delle variabili in considerazione, mentre invece valori prossimi a 0 indicano una scarsa o assente correlazione.

È ora necessario quindi attribuire alle variabili un peso al fine di ottenere una funzione che porti ad uno score. Per far questo si procede alla stima del modello anticipato però da una fase di pre-processing dei dati.

### 3.3.3 Pre-processing dei dati

In qualsiasi lavoro avente ad oggetto operazioni su un insieme di dati è opportuno fare delle analisi sugli stessi prima di iniziare la fase operativa vera e propria. Vi possono essere dati errati o mancanti per diverse ragioni e per questo è opportuno molta attenzione alla natura dei dati e provvedere ad effettuare delle azioni correttive per "pulire" i dati con cui si lavorerà.

#### 3.3.3.1 Analisi dei missing

Prima della stima del modello è importante effettuare dei controlli sul campione da utilizzare, tra questi vi è l'analisi dei missing. I missing sono dei valori di variabili mancanti per alcuni soggetti, possono essere dovuti all'assenza vera e propria dell'informazione ma anche a mancanze in fase di data entry o errori di archiviazione. Un elevato numero di missing pregiudica la corretta stima del parametro assegnato alla variabile in questione durante la stima del modello, portando anche alla successiva non significatività della variabile stessa. Nella scelta delle variabili è quindi importante valutare anche la capacità di recepire l'informazione per assicurare al modello il mantenimento nel tempo. Per gestire i missing sono possibili diversi approcci:

- Eliminazione del record e quindi esclusione totale del soggetto dalla popolazione;
- Forzatura del valore missing, per esempio, al valor medio riscontrato dal resto della popolazione;
- Se il numero di missing per una variabile è superiore ad una certa sogli è preferibile non utilizzare la variabile in questione.

#### 3.3.3.2 Analisi degli outliers

Un altro controllo porta all'analisi degli outliers. Gli outlier sono valori estremi, che si scostano molto da quelli normalmente osservati per la variabile in esame e per questo

vanno trattati al fine di non produrre una distorsione dell'analisi. È necessario svolgere un'analisi approfondita sulla causa della presenza del dato anomalo per evitare che sia dovuto ad errori evitabili o una non corretta selezione del campione. Dopodichè vengono applicare le metodologie seguenti per evitare la distorsione nella stima del modello:

- Troncatura delle code in modo simmetrico della distribuzione dei valori della variabile oltre una predeterminata soglia e quindi eliminazione delle informazioni;
- Definizione di una funzione di ponderazione dei dati appartenenti alla distribuzione, in modo da attribuire un peso irrilevante agli outlier e vanificarne quindi l'effetto distorsivo pur mantenendo l'informazione.

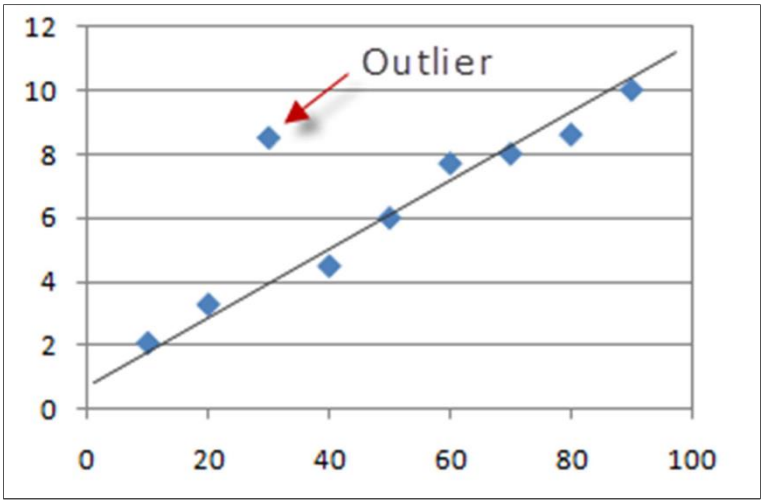


**Figura 3.1 Outlier**[12]

## 3.3.4 Stima del modello

Un modello viene prima stimato e poi testato. Per farlo è necessario quindi dividere il campione in due, un campione di stima tramite il quale verranno calcolati i parametri del modello stesso ed un campione di validazione per la verifica dell'efficacia dei parametri

[12] Fonte: https://www.mathopenref.com/outlier.html

scelti. In due campioni devono mantenere una distribuzione dei drivers di segmentazione simile a quella di partenza.

La best practice per effettuare questa suddivisione viene chiamata in campo statistico *cross-validation*, si suddivide il campione osservato in gruppi di egual numerosità e si esclude iterativamente un gruppo alla volta per l'addestramento del modello e utilizzando il gruppo scartato per la validazione. É possibile in questo modo evitare problemi di overfitting ma anche di campionamento asimmetrico che conducono a distorsione[13]. L'overfitting è un problema cruciale per i modelli con algoritmi genetici, occorre quando i parametri si adeguano troppo al campione di addestramento e ne memorizzano talmente bene il rumore che però sarà diverso da quello del campione di validazione sul quale le performance del subiranno un drastico calo. Utilizzando un grande set di dati, generalmente, è meno probabile cadere nell'overfitting ma non è sempre così, soprattutto se la complessità del problema da ottimizzare aumenta(Dos Santos et al., 2009)[14]

In caso la numerosità del campione sia limitata non è possibile utilizzare la cross-validation, è però possibile utilizzare le seguenti logiche per selezionare il campione di validazione:

- *Out of time:* il campione di validazione è sulle stesse imprese ma osservate in un periodo di tempo diverso rispetto a quello utilizzato per la definizione del campione di stima;
- *Out of sample:* il campione di validazione è composto da imprese non appartenenti al campione di partenza;
- *Out of universe:* il campione di validazione è composto da imprese che non sono quelle del campione di stima e l'orizzonte temporale è diverso.

La stima del modello conduce alla definizione dei parametri delle variabili scelte in modo

[13] https://en.wikipedia.org/wiki/Cross-validation_(statistics)
[14] Dos Santos, E. M., Sabourin, R., & Maupin, P. (2009). Overfitting cautious selection of classifier ensembles with genetic algorithms. Information Fusion, 10, 150–162

che sarà possibile, successivamente, applicate la funzione ai valori di un altro soggetto per definirne lo score. Come visto precedentemente le metodologie applicabili sono numerose, statistiche o con algoritmi più complessi.

Ottenuta la stima del modello si eseguono alcune azioni di controllo:

- È necessario verificare la coerenza dei segni delle variabili in relazione allo score. In presenza di segno positivo, all'aumentare della variabile indipendente aumenta la variabile dipendente, il significato economico di tale relazione deve essere coerente
- Al fine di verificare la significatività di un indicatore viene effettuato il test statistico t. L'ipotesi nulla è che la variabile non sia statisticamente significativa e che quindi il parametro sia uguale a zero. Affinché si possa rigettare l'ipotesi nulla è necessario che il P-value sia inferiore al livello di significatività $\alpha$ scelto.
- Si effettua il calcolo dell'$R^2$. L'$R^2$ valuta quanta della varianza della variabile dipendente è spiegata dal modello stimato

È possibile a questo punto definire il valore soglia, il cut-off, che scerne gli score ottenuti dal modello nelle due categorie di imprese, quelle considerate sane e quelle che invece sono anomale. Per scegliere il valore di cut-off vi sono diverse metodologie che si differenziano anche sulla base del modello utilizzato. Per fare qualche esempio, Altman propone per il suo modello il calcolo del valore soglia tramite la seguente relazione:

$$Cut - off = \ln \frac{q_a C_1}{q_s C_2} \qquad (3.2)$$

Dove:

- $q_a \; e \; q_s$ sono le probabilità a priori che l'impresa sia rispettivamente anomala o sana, ovvero la probabilità che l'impresa appartenga ad una delle categorie prima ancora che si osservino i valori delle variabili;
- $C_1 \; e \; C_2$ sono i costi degli errori di prima e seconda specie, verranno introdotti

nel prossimo paragrafo.

È inoltre possibile selezionare il valore soglia in modo da massimizzare gli indici di performance del modello, come l'accuracy ratio o l'indicatore derivante dalla curva ROC.

### 3.3.5 Test del modello

Una volta sviluppato il modello sul campione di stima, viene testato sul campione di validazione, al fine di valutare la bontà dello score assegnato e l'assenza di overfitting, ovvero che i risultati non siano troppo adattati al campione su cui è stato effettuato lo sviluppo e non sia invece abile per campioni diversi. Seguendo la logica della cross-validation, in realtà verranno eseguiti diverse stime e diversi test su campioni differenti. Il fine è valutare l'efficacia su un campione che non è quello con cui è stato costruito il modello e per farlo vengono calcolati gli indicatori di performance adeguati al tipo di modello utilizzato.

Preliminare a qualsiasi altra considerazione è il calcolo della matrice di confusione, detta anche tabella di errata classificazione, in cui vengono rappresentate come sono state classificate le imprese con la previsione del modello rispetto a come si sono comportate nella realtà. Per farlo si utilizza appunto una matrice come quella sotto riportata:

| | | **Classificazione ottenuta** | | **Totale** |
|---|---|---|---|---|
| | | Anomala | Sana | |
| **Situazione reale** | Anomala | Corretta classificazione Anomale<br>Vero Positivo (VP) | Errore di prima specie<br>Falso Negativo (FN) | Imprese realmente insolventi |
| | Sana | Errore di seconda specie<br>Falso Positivo (FP) | Corretta classificazione Sane<br>Vero Negativo (VN) | Imprese realmente sane |
| **Totale** | | Imprese classificate insolventi | Imprese classificate sane | Totale imprese considerate |

Tabella 3.1 Confusion Matrix

L'errore di prima specie si palesa quando viene giudicata dal modello un'impresa sana quando questa si è rivelata anomala nella realtà mentre l'errore di seconda specie è la classificazione da parte del modello di un'impresa tra le anomale mentre questa in realtà si è dimostrata sana. La gravità dei due errori è evidentemente diversa perché se può non essere economicamente debilitante il mancato guadagno dovuto alla non erogazione di un finanziamento ad un'impresa sana classificata come anomala, lo è invece la perdita dell'investimento in caso si decida di finanziare un'impresa che si rivela in un secondo momento anomala. Dalla matrice di confusione è possibile ricavare degli indicatori di performance. Consideriamo l'impresa che si dimostra anomala come un caso 'positivo' per avere delle nomenclature maneggevoli nelle formule e attribuiamo dei valori arbitrari perché siano più chiari i concetti:

| | |
|---|---|
| 10<br>Vero Positivo (VP) | 2<br>Falso Negativo (FN) |
| 3<br>Falso Positivo (FP) | 20<br>Vero Negativo (VN) |

**Tabella 3.2 Confusion Matrix. Valori di esempio**

- Tasso di errore (Error Rate), misura il totale degli errori rispetto al totale delle previsioni fatte

$$ERR = \frac{FP + FN}{VP + VN + FP + FN} = \frac{5}{35} = 14{,}3\% \qquad (3.3)$$

  Il modello ha attribuito a categorie sbagliate il 14,3% del campione;

- Accuratezza (Accuracy), percentuale di misure esatte sul totale. Corrisponde al complemento a 1 del tasso di errore

$$ACC = \frac{VP + VN}{VP + VN + FP + FN} = \frac{30}{35} = 85{,}7\% \quad (3.4)$$

Il modello ha attribuito l'85,7% delle imprese nel campione alla categoria corretta

- Precisione (Precision), misura la percentuale di corretta previsione di imprese anomale sul totale delle previsioni di imprese anomale

$$PR = \frac{VP}{VP + FP} = \frac{10}{13} = 76{,}9\% \quad (3.5)$$

Il 76,9% delle imprese definite anomale dal modello è veramente anomala;

- Sensibilità (Sensitivity), percentuale delle imprese previste positive correttamente sul totale delle reali positive

$$Sensitivity = \frac{VP}{VP + FN} = \frac{10}{12} = 83{,}3\% \quad (3.6)$$

L'83,3% delle imprese realmente anomale sono state rilevate dal modello;

- Specificità (Specificity), previsioni corrette di sane sul totale delle imprese sane nel campione

$$SP = \frac{VN}{VN + FP} = \frac{20}{23} = 86{,}96\% \quad (3.7)$$

L'86,96% delle imprese realmente sane sono state identificate come tali dal modello.

- Tasso dei Falsi Positivi (False Positive Rate), percentuale di imprese sane che sono state erroneamente categorizzate come anomale

$$FPR = \frac{FP}{VN + FP} = \frac{3}{23} = 13{,}04\% \tag{3.8}$$

Il 13% delle imprese sane è stato attribuito alla categoria delle anomale

Oltre a questi indicatori vi sono numerosi metodi per valutare la performance di un modello ma i più frequentemente utilizzati sono l'Accuracy Ratio (o indice di Gini) e la curva ROC.

L'Accuracy Ratio è una misura per descrivere la capacità discriminante di un modello, quanto il modello ottenuto si discosta da un modello che assegna a tutte le imprese lo stesso score. Un modello che per qualsiasi mix di variabili indipendenti assegna la stessa variabile dipendente è un modello che non ha nessun valore predittivo. Il coefficiente è calcolato a partire dalla definizione della curva CAP (Cumulative Accuracy Profile) nota anche come curva di Gini, curva di potenza o curva di Lorenz. In ascissa è riportata la percentuale di popolazione in esame ordinata per score crescenti; quindi, raggiunto il 10% vuol dire che si è preso in esame il 10 % della popolazione e questa è caratterizzata dai peggiori score; sulle ordinate la cumulata delle imprese considerate anomale, se la curva cresce rapidamente significa che il modello identifica velocemente le imprese anomale perché queste sono le prime prese in considerazione dato l'ordinamento delle ascisse.

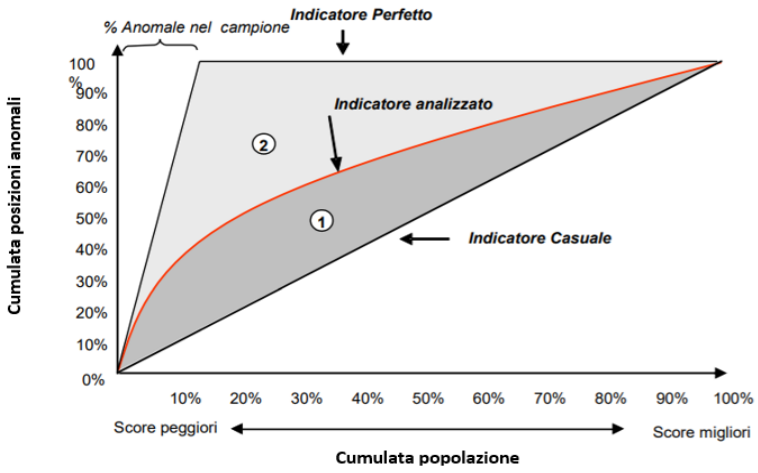


**Figura 3.2 Rappresentazione grafica delle curve CAP**[15]

[15] Fonte: https://www.mathisintheair.org/wp/2021/08/il-coefficiente-di-gini-e-suo-cugino-laccuracy-ratio/

Come mostrato in figura 3.2, supponendo che nella popolazione vi sia il 10% delle imprese anomale, vengono identificate 3 curve per il calcolo dell'indice:

- La curva del modello ideale il quale riesce a discriminare il 100% delle anomale con il primo 10% degli score (i peggiori) attribuiti;
- La curva rappresentante il modello casuale che non è in grado di discriminare in quanto, per esempio, con il 50 % degli score peggiori individua solo il 50% delle anomale, difficile fare peggio quando un'impresa può essere sana o anomala,
- La curva del modello valutato che si colloca tra i due casi estremi, più questa curva è ripida vicino all'origine maggiore è la capacità predittiva del modello.

Il valore dell'indice viene ottenuto graficamente come rapporto tra l'area sottesa alla curva del modello in valutazione e l'area identificata dal modello ideale non considerando in entrambi i casi l'area del modello casuale. L'indicatore assume valori compresi tra 0 e 1 e generalmente sono considerati buoni modelli quelli che ottengono un'accuracy ratio superiore al 50%.

Per quanto riguarda la curva ROC, invece, viene posto sull'asse delle ordinate la sensibilità come definita precedentemente, la percentuale delle imprese che sono realmente anomale e che il modello ha identificato come tali, gli allarmi veri; sulle ascisse l'FPR, complemento a 1 della specificità, ovvero i falsi allarmi. Nel caso dei modelli applicati al rischio di credito è preferibile avere un modello il più sensibile possibile, come già affermato il costo dovuto alla non identificazione di un'impresa anomala è maggiore. La costruzione della curva ROC si ottiene modificando la soglia di cut-off con la quale si discrimina un'impresa sana da una anomala, ad ogni variazione si calcola la matrice di confusione e i due indicatori, la sensitivity e il false positive rate, identificando così i punti nel piano rappresentato in figura:

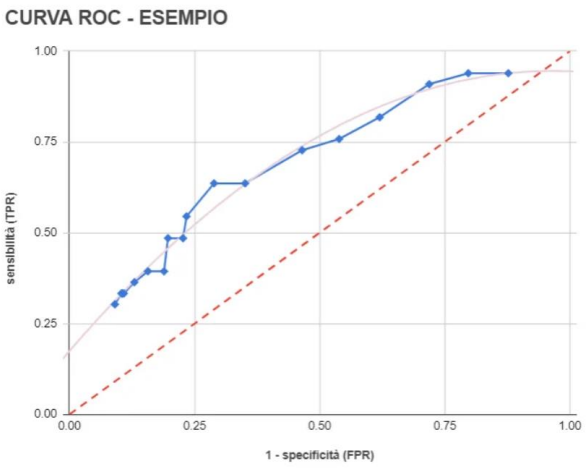


**Figura 3.3 Esempio curva ROC[16]**

I valori in diagonale rappresentano il modello casuale che non è in grado di discriminare, mente i valori del semi-contorno superiore il modello ideale. Il valore ottimo di cut-off sul modello è quello che genera valori di falsi positivi e di sensibilità più prossimi all'angolo superiore sinistro perché massimizza contemporaneamente sensibilità e specificità del modello. Per confrontare e valutare modelli diversi, dopo aver costruito la curva, si calcola l'indice AUC (Area Under the Curve). Il modello non informativo ha AUC = 0,5. Per valori crescenti il modello risulta sempre più performante e in genere sono considerati modelli attendibili quelli che registrano valori maggiori del 75%. È possibile anche che due modelli abbiano lo stesso valore AUC ma che la curva ROC sia differente (Figura 3.4 in basso). In questi casi è possibile valutare che tipo di test si vuole più sensibile o più specifico, per i modelli di scoring è sicuramente preferibile un modello maggiormente sensibile e quindi il modello migliore risulterebbe quello con curva ROC identificata in figura dalla lettera A.

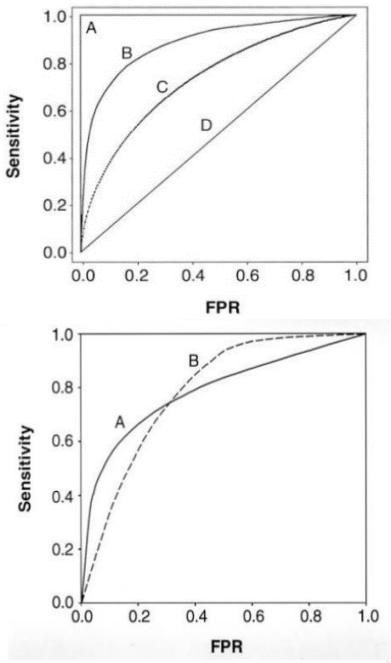


**Figura 3. 4 Differenti tipi di curve[16]**

[16] Fonte: https://www.med4.care/curva-roc-receiver-operating-characteristic-introduzione-e-applicazione-ai-test-diagnostici/

### 3.3.6 Backtesting del modello

Il backtesting di un modello viene effettuato alla fine di valutare ex-post la bontà del modello utilizzato e viene fatto mediante l'analisi dei dati passati. Vengono analizzate 3 caratteristiche del modello:

1. Il potere discriminante, attraverso la valutazione della capacità del modello di attribuire score alle controparti sulla loro effettiva qualità creditizia che ora è notta. Vengono utilizzati i test di performance già utilizzati in fase di test, l'accuracy ratio, la curva ROC e gli errori di prima e seconda specie. In fase di backtesting viene anche utilizzato un adattamento del test di Kolmogorov-Smirnov, questo test misura la massima deviazione verticale esistente tra la distribuzione cumulata di frequenza delle controparti in bonis rispetto alla rispettiva dei clienti in default.

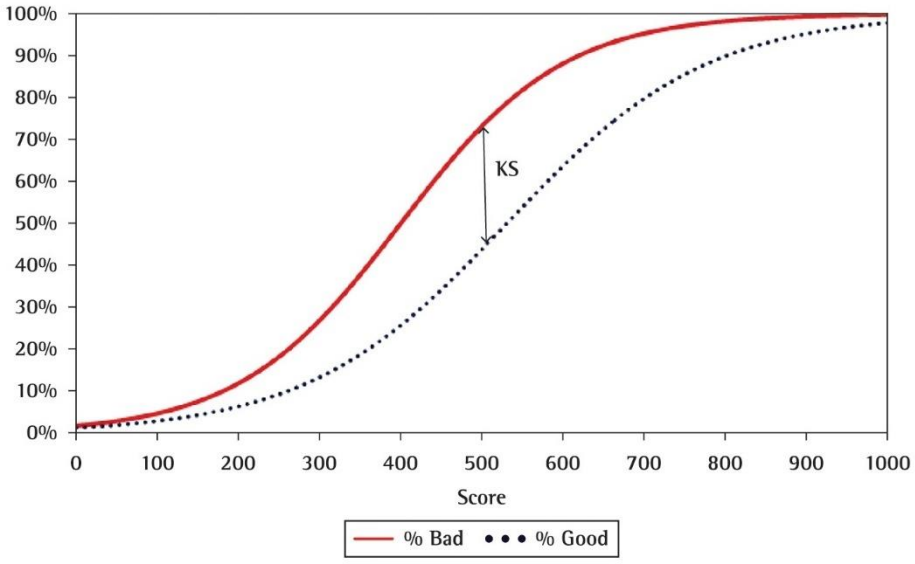


**Figura 3.3 Kolmogorov-Smirnov test**[17]

2. La calibrazione, tramite il calcolo della differenza tra PD stimata con il modello e i tassi di default osservati ex-post. Si vuole verificare che questa differenza sia riconducibile alla casualità o ad un atteggiamento sistemico. La corretta calibrazione può essere valutata utilizzando alcuni test statistici come:

[17]Fonte: https://corninaber.nl/2017/01/06/credit-risk/

- Il test $\chi^2$ che determina se c'è una relazione tra distribuzione osservata realmente e quella stimata;
- Il test binomiale verifica l'ipotesi nulla che la probabilità di default attribuita ad una classe di rating corrisponda al tasso di default osservato alla realtà contro l'ipotesi alternativa che il modello sia sotto/sovrastimato (se si sta eseguendo un test bilaterale)
- Il Reliability diagram, uno strumento grafico che permette di relazionare i tassi di default osservati con quelli previsti per ogni classe di rating in modo da rappresentare sul grafico una spezzata. Più la spezzata ottenuta si approssima alla bisettrice del quadrante, migliore la calibrazione del modello.

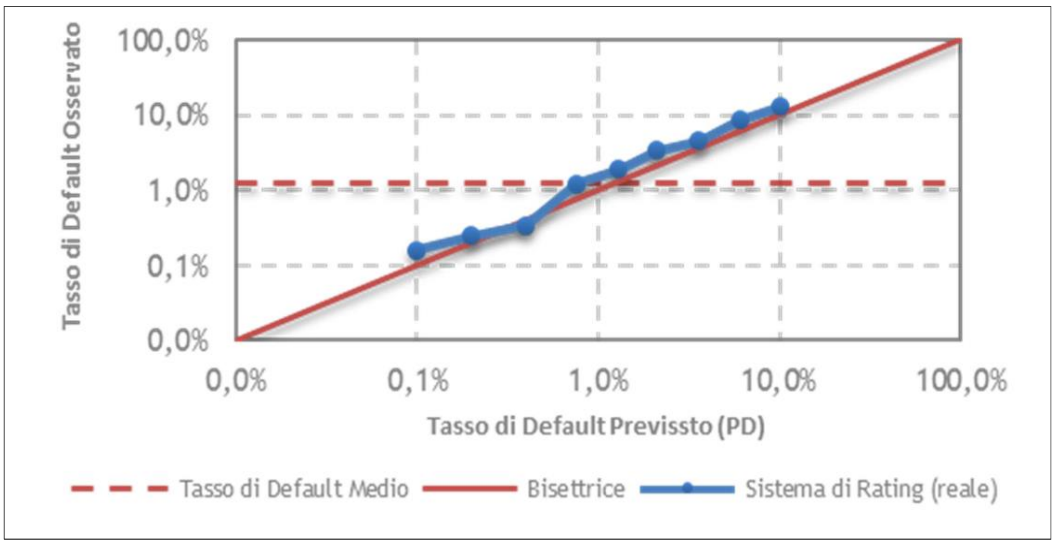


**Figura 3.4 Reliabillity diagram**

- La stabilità delle classi di rating nel tempo. Per valutare la stabilità è possibile utilizzare le matrici di transizione. Una transition matrix è utile ad indicare la probabilità che una controparte in una certa classe di rating resti stabile nella sua classe o migri in un'altra classe. In figura 3.5 la matrice di transizione per principali aree geografiche pubblicata nel 2021 da Standard & Poor's

| 2020 One-Year Corporate Transition Rates By Region (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| From/to | AAA | AA | A | BBB | BB | B | CCC/C | D | NR |
| Global | | | | | | | | | |
| AAA | 100.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| AA | 0.00 | 87.27 | 9.63 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 3.11 |
| A | 0.00 | 0.00 | 92.88 | 4.05 | 0.00 | 0.07 | 0.00 | 0.00 | 3.00 |
| BBB | 0.00 | 0.05 | 0.59 | 90.13 | 4.47 | 0.22 | 0.00 | 0.00 | 4.53 |
| BB | 0.00 | 0.00 | 0.00 | 0.78 | 78.20 | 11.40 | 0.85 | 0.93 | 7.84 |
| B | 0.00 | 0.00 | 0.00 | 0.05 | 0.96 | 71.99 | 12.56 | 3.51 | 10.92 |
| CCC/C | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 5.46 | 34.45 | 47.48 | 12.61 |
| U.S. | | | | | | | | | |
| AAA | 100.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| AA | 0.00 | 90.30 | 8.96 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.75 |
| A | 0.00 | 0.00 | 93.86 | 3.79 | 0.00 | 0.00 | 0.00 | 0.00 | 2.35 |
| BBB | 0.00 | 0.14 | 0.98 | 90.17 | 4.49 | 0.56 | 0.00 | 0.00 | 3.65 |
| BB | 0.00 | 0.00 | 0.00 | 0.90 | 76.13 | 15.55 | 0.72 | 1.27 | 5.42 |
| B | 0.00 | 0.00 | 0.00 | 0.09 | 1.20 | 73.05 | 12.36 | 3.61 | 9.70 |
| CCC/C | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 4.58 | 35.95 | 49.02 | 10.46 |
| Europe | | | | | | | | | |
| AAA | 100.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| AA | 0.00 | 85.71 | 10.20 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 4.08 |
| A | 0.00 | 0.00 | 93.13 | 3.56 | 0.00 | 0.00 | 0.00 | 0.00 | 3.31 |
| BBB | 0.00 | 0.00 | 0.46 | 89.47 | 5.49 | 0.00 | 0.00 | 0.00 | 4.58 |
| BB | 0.00 | 0.00 | 0.00 | 0.96 | 75.60 | 9.09 | 1.44 | 0.96 | 11.96 |
| B | 0.00 | 0.00 | 0.00 | 0.00 | 0.87 | 74.78 | 12.17 | 3.04 | 9.13 |
| CCC/C | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 8.16 | 32.65 | 44.90 | 14.29 |
| Emerging markets | | | | | | | | | |
| AAA | 100.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| AA | 0.00 | 93.18 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 6.82 |
| A | 0.00 | 0.00 | 91.94 | 5.49 | 0.00 | 0.37 | 0.00 | 0.00 | 2.20 |
| BBB | 0.00 | 0.00 | 0.19 | 89.60 | 4.91 | 0.00 | 0.00 | 0.00 | 5.29 |
| BB | 0.00 | 0.00 | 0.00 | 0.42 | 81.10 | 7.64 | 0.64 | 0.64 | 9.55 |
| B | 0.00 | 0.00 | 0.00 | 0.00 | 0.54 | 68.12 | 11.99 | 3.54 | 15.80 |
| CCC/C | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 4.00 | 28.00 | 44.00 | 24.00 |

**Figura 3.4 Transition Matrix**[18]

[18] Fonte: https://www.spglobal.com/ratings/en/research/articles/210407-default-transition-and-recovery-2020-annual-global-corporate-default-and-rating-transition-study-11900573

# 4 GLI ALGORITMI GENETICI

Gli algoritmi genetici sono stati sviluppati da John Holland, i suoi colleghi e i suoi studenti presso l'università del Michigan. L'obiettivo della loro ricerca era duplice: astrarre e spiegare rigorosamente il processo adattivo del sistema naturale, progettare un software che replicasse l'importante meccanismo naturale. Questo lavoro ha condotto ad importanti scoperte in entrambe le scienze, quella che si occupa della natura e quella dei sistemi artificiali.

Gli algoritmi genetici sono basati sul meccanismo di selezione naturale e genetica. Essi combinano la sopravvivenza della più compatibile tra le strutture di stringhe con una procedura per la selezione dall'informazione randomica al fine di generare un algoritmo di ricerca. In ogni generazione, un nuovo set di strutture artificiali (stringhe) è creato usando un pezzo delle migliori della generazione precedente e occasionalmente una nuova parte è inserita casualmente per migliorarne le capacità. Questo algoritmo sfrutta efficientemente l'informazione storica per speculare sulle caratteristiche con performance migliori.

Tema centrale è la robustezza dei risultati, la loro applicabilità a differenti ambienti, la loro efficacia ed efficienza. Se si prendono in considerazione ambienti in cui performance robuste sono le desiderate, allora bisogna tenere in conto che la natura lo fa meglio.

Per la stesura del capitolo mi sono avvalso delle informazioni apprese dal testo scritto da Goldberg (Goldberg, 1989)

## 4.1 Efficienza

Non è solo eleganza, la robustezza dei sistemi di algoritmi genetici è provata sia a livello teorico che empirico.

Gli algoritmi genetici hanno infatti trovato numerosi campi di applicazione, sono computazionalmente semplici, potenti nella loro ricerca di miglioramento e non sono limitati da assunzioni che restringono il loro spazio di azione. Investigheremo meglio le

ragioni di queste qualità. Ma prima bisogna esplorare e confrontare quelli che sono gli altri metodi di ricerca per avere un quadro generale.

Ci sono principalmente tre metodi di ricerca della soluzione ottima: quelli basati sul calcolo numerico, quelli enumerativi e i modelli randomici.
I modelli di calcolo numerico sono stati studiati approfonditamente e si suddividono in due classi: diretti ed indiretti. I metodi indiretti ricercano massimi e minimi locali tramite la risoluzione di sistemi di equazioni non lineari il cui risultato è ottenibile impostando il gradiente della funzione obiettivo pari a zero. Questa è la generalizzazione multidimensionale dell'elementare calcolo di massimi e minimi per una funzione alla ricerca di un possibile picco.

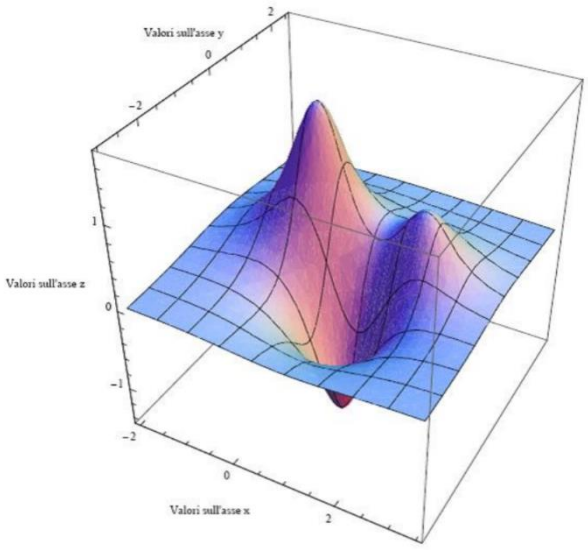


**Figura 4. 1 Funzione multimodale**[19]

I metodi diretti ricercano l'ottimo scorrendo sulla funzione nella direzione suggerita dal gradiente nel punto da cui si parte. Qualche semplice esempio mostra la mancanza di robustezza di questi metodi.
Primo, entrambi i metodi ritrovano l'ottimo in una zona ristretta in un intorno del punto di partenza. Secondo, i metodi basati sul calcolo dipendono interamente dall'esistenza delle derivate delle funzioni che quindi devono avere una ben definita forma e scorrevolezza.

[19] Fonte: https://www.unirc.it/documentazione/materiale_didattico/1465_2013_353_18089.pdf

I matematici del 18esimo e 19esimo secolo dipingono un mondo fatto esclusivamente di funzioni obiettivo quadratiche e derivabili in ogni punto, purtroppo il mondo reale della ricerca è pieno di discontinue e multimodali funzioni che sono descritte da equazioni molto meno amichevoli e trattabili.

I metodi enumerativi sono molto sbrigativi, lavorano con uno spazio finito o uno spazio infinito ma discreto e l'algoritmo di definizione dell'ottimo consiste semplicemente nell'osservare tutti i punti della funzione obiettivo, uno alla volta. Nonostante la semplicità e completezza del metodo, ovviamente la robustezza cade nella carenza di efficienza. La maggior parte degli spazi di ricerca sono semplicemente troppo grandi per essere analizzati interamente.

Gli algoritmi di ricerca randomici hanno incrementato la loro popolarità fra i ricercatori una volta che sono stati percepiti i limiti dei precedenti metodi. Ma anche loro nel lungo termine risultano inefficienti poiché la continua ricerca di valori migliori random, ripetendo l'algoritmo più e più volte, tende a rendere questi metodi approssimabili a quelli enumerativi.
Gli AG (algoritmi genetici) sono un esempio di procedura di ricerca che usa la scelta random come strumento di guida per l'esplorazione dello spazio di dominio, definendo in questo modo i parametri del codice. Bisogna infatti dire che la ricerca randomizzata non implica necessariamente una ricerca senza direzione definita.
L'aver definito i metodi precedenti come non robusti, non implica che siano totalmente inutili. Anzi per alcune precise applicazioni possono avere delle notevoli performance.

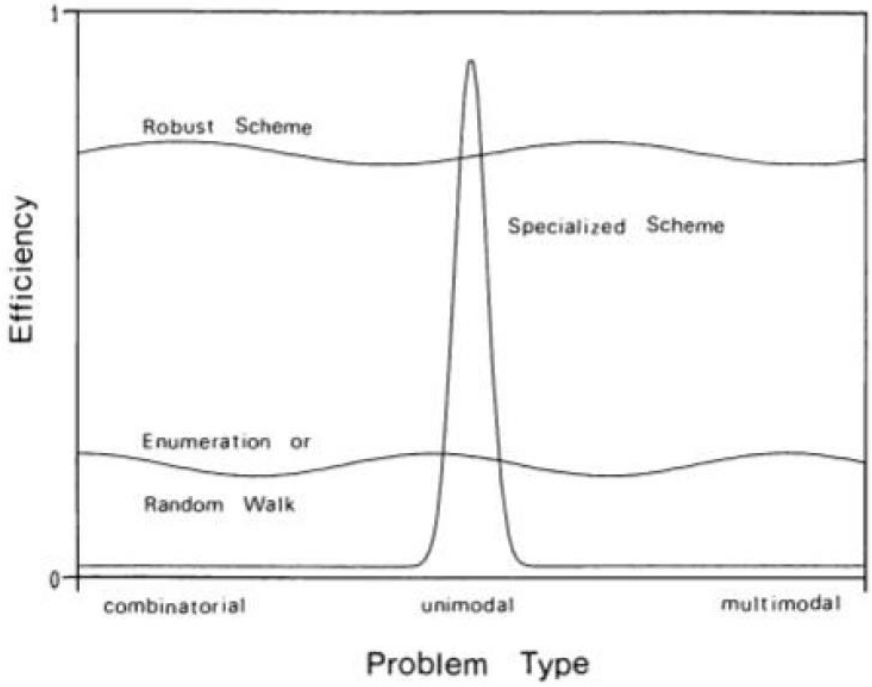


**Figura 4. 2 Efficienza dei metodi di ricerca per problemi affrontati[20]**

La figura aiuta qualitativamente a capire qual è l'efficienza dei metodi in relazione al tipo di problema. Si può notare come su una funzione obiettivo che nella realtà è unimodale, il metodo basato sul calcolo raggiunge un'efficienza molto elevata che però è praticamente nulla in altri casi. Metodi enumerativi e random performano con stoica uguaglianza di performance nei vari casi.

Un ideale schema robusto invece riporta alta efficienza in tutti i campi seppur non raggiunga il massimo di efficienza in nessuno di essi.

La teoria dell'ottimizzazione comprende lo studio quantitativo dei punti di ottimo e i metodi per trovarli. Per giudicare un metodo di ottimizzazione di solito ci concentriamo esclusivamente sul risultato ovvero il raggiungimento del vero ottimo dimenticandoci della performance per raggiungerlo.

Quali criteri possiamo utilizzare per definire se abbiamo fatto un buon lavoro? Di solito possiamo dirlo quando abbiamo fatto un'adeguata selezione nei tempi e con le risorse che avevamo a disposizione o che ci eravamo prefissati di utilizzare. In molte applicazioni quello

[20] Goldberg, D., Genetic Algorithms in Search, Optimization, and Machine Learning

che conta è raggiungere un risultato in modo migliore relativamente agli altri. Il più importante obiettivo dell'ottimizzazione è il miglioramento continuo.

## *4.2 L'algoritmo in breve*

Gli AG differiscono dagli altri metodi di ricerca operativa per quattro motivi:

- Lavorano con una *codifica* di set di parametri, non coi parametri stessi
- Ricercano a partire da una *popolazione* di punti, non da un singolo punto
- Usano come informazione i *payoff* della funzione obiettivo stessa e non il risultato di derivate o altre conoscenze ausiliarie
- Usano regole di transizione *probabilistiche* e non regole deterministiche

Gli AG richiedono di codificare i parametri come una stringa finita
Seguiamo un esempio per fissare meglio il concetto. Si vuole massimizzare la funzione f(x) = x2 nel l'intervallo [0 , 31]. Con i metodi tradizionali dovremmo metterci a lavorare con il parametro x, come se volessimo sistemare le antenne di una televisione fino al raggiungimento di una buona visione del programma. Con gli AG il primo passo consisterebbe nel codificare il parametro x come una stringa di lunghezza finita.
Consideriamo una scatola nera che presenta sulla superficie 5 interruttori. Per ogni settaggio degli interruttori c'è un determinato output *f*, matematicamente *f* = f(s) dove s e il vettore che definisce il settaggio degli interruttori. L'obiettivo è trovare il settaggio corretto al fine di ottenere il massimo valore di f possibile. Un semplice metodo per codificare il settaggio degli interruttori è quello di formare un vettore composto di 1 e di 0 dove ogni numero rappresenta lo stato di uno degli interruttori.

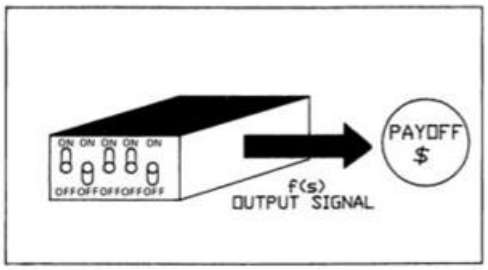


**Figura 4. 3 Black box**[21]

Con molti metodi di ottimizzazione, ci muoviamo a partire da un singolo punto verso un altro usando una determinata regola di transizione per l'identificazione del prossimo punto (per esempio la derivata). Questo metodo punto a punto è pericoloso perché è la ricetta perfetta per incagliarsi in un massimo o minimo locale di uno spazio multimodale.

Al contrario, gli algoritmi genetici lavorano a partire da un ricco database di punti (una popolazione di vettori o stringhe), simultaneamente, scalando diversi picchi in parallelo, così la possibilità di fermarsi in un estremo locale sono ridotte rispetto ai metodi prima descritti.

Tornando al problema della scatola, gli AG iniziano da una popolazione di stringhe di settaggio degli interruttori e li utilizza per generare una successiva popolazione di stringhe.

Per esempio, una popolazione iniziale creata casualmente, tramite il lancio di una moneta (che quindi segue una distribuzione binomiale), può essere utilizzato per la creazione della prima popolazione per esempio formata da 4 stringhe (vettori) (4 vettori di partenza sono pochi per gli standard ma è un esempio):

00011
11010
11100
01010

Dopo questo inizio a partire da diversi punti dello spazio, le popolazioni successive composte da vettori che rappresentano sempre il settaggio degli interruttori, saranno generate dall'algoritmo genetico.

Molti metodi di ricerca alternativi necessitano di molte altre informazioni ausiliari per funzionare correttamente. Gli algoritmi genetici sono ciechi. Non guardano dentro la

[21] Goldberg, D., Genetic Algorithms in Search, Optimization, and Machine Learning

scatola. Loro necessitano esclusivamente del payoff generato associato al vettore di parametri in ingresso.
Gli algoritmi genetici usano delle regole di transizione probabilistiche ma ciò non vuol dire che siano assimilabili a dei processi randomici di ricerca, poiché usano esclusivamente la casualità come uno strumento per guidare la ricerca attraverso le regioni dello spazio sconosciuto nel qual potrebbe celarsi un miglioramento.
Insieme queste quattro differenze rispetto ai metodi numerici ed enumerativi - diretto uso di stringhe, ricerca a partire da una popolazione, non utilizzo di informazioni ausiliarie e operatori randomici - contribuiscono alla robustezza di un AG.

### 4.2.1 Gli Operatori

Il meccanismo è sorprendentemente semplice e questa semplicità legata alla potenza dei suoi effetti rendono il metodo molto attraente.
Ricordando l'esempio della black box la popolazione delle 4 stringhe iniziali erano state scelte in modo casuale da 20 lanci di una moneta equilibrata consecutivi. Dobbiamo ora definire una serie di operazioni che da questa popolazione iniziale ne generino di successive che si spera siano migliori e che possano ulteriormente migliorare con l'applicazione reiterata delle operazioni stesse.
Un semplice AG che porta a buoni risultati in molti problemi pratici è composto da tre semplici operatori:

1. Riproduzione
2. Incrocio (Crossover)
3. Mutazione

La riproduzione è un processo nel quale una singola stringa è copiata in linea con la sua funzione obiettivo chiamata dai biologi la 'fitness function'. Definiamo *f* la funzione di utilità o profitto che vogliamo massimizzare. Con l'operatore riproduzione le stringhe che generano un alto valore *f* hanno più probabilità di contribuire per uno o più discendenti della

futura generazione. Questo operatore, ovviamente, è una versione artificiale della selezione naturale, una darwiniana sopravvivenza del più adatto, del fittest value.

Verrà qui utilizzata la codifica binaria per esprimere dei valori di input dell'algoritmo e l'esempio considererà come funzione obiettivo la semplice $f(stringa) = x^2$ con x∈ [0 , 31]. I valori di partenza scelti casualmente risultano essere rispettivamente 13, 24, 8 e 19.

| n | Stringa | Fitness | % del totale |
|---|---|---|---|
| 1 | 01101 | 169 | 14,4% |
| 2 | 11000 | 576 | 49,2% |
| 3 | 01000 | 64 | 5,5% |
| 4 | 10011 | 361 | 30,9% |
| Totale | | 1170 | 100,0% |

**Tabella 4. 1 Fitness della popolazione**

In natura sono le abilità o le caratteristiche a definire chi sopravviverà nel lungo termine. Nel nostro mondo artificiale è il risultato della funzione obiettivo a stabilire chi vive e chi muore. Implementare in un algoritmo l'operatore di riproduzione è semplice. Si pensi ad una roulette le cui parti in cui è suddivisa la ruota sono di grandezza proporzionale al peso del valore della fitness function. Per esempio, il primo valore estratto 13 utilizzato nella fitness function darà come payoff 169. Ripetendo per tutta la popolazione di partenza e calcolando i rispettivi pesi relativi, il valore ottenuto dalla prima stringa coprirà il 14.4% = 169/1170 della roulette.

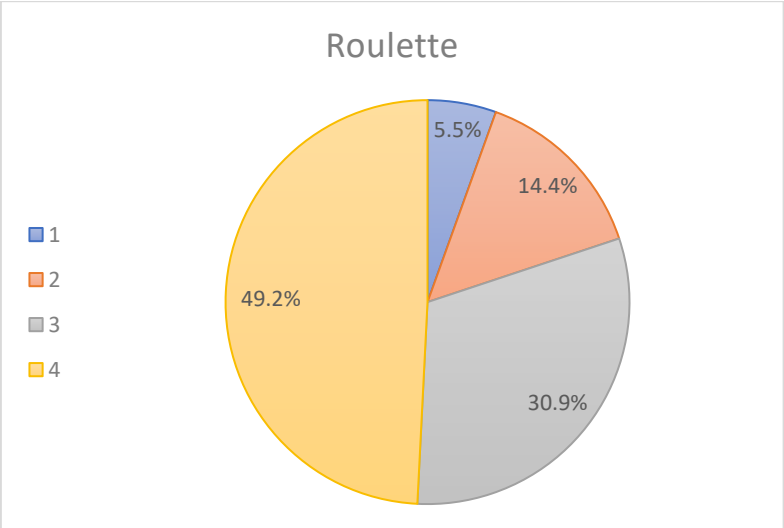


**Figura 4. 4 Rappresentazione della roulette ponderata**

A questo punto entra in gioco un ulteriore elemento casuale che è rappresentabile dal far girare 4 volte la ruota al fine di selezionare le stringhe che potranno riprodursi. L'esito potrebbe essere vario, per esempio potrebbero accedere alla riproduzione tutte e 4 le stringhe o addirittura solo una. Questi eventi estremi sarebbero molto poco probabili nel caso avessimo utilizzato una popolazione più numerosa in partenza come in realtà faremmo nella pratica.

Una volta che una stringa è stata selezionata per la riproduzione una sua esatta copia viene creata e aggiunta all'insieme di stringhe che potranno accoppiarsi. Qui entra in gioco l'operatore di incrocio (crossover) che può essere considerato composto da due step:

- il primo è la scelta casuale di quelli che saranno gli accoppiamenti tra le stringhe
- il secondo step prevede la scelta casuale di una posizione k all'interno dei vettori stringa.

Le nuove stringhe verranno create a partire dalle stringhe madri tagliate fra la posizione *k* e la k+1, i pezzi delle stringhe madri tagliate vengono incrociate dando vita (nella maggior parte dei casi) a vettori figli diversi da quelli di partenza.

Nell'esempio seguente sono state casualmente scelte per il crossover la prima e la seconda stringa, mentre k è uguale a 4. Le stringhe risultanti saranno quindi rispettivamente uguali a quelle di partenza ad eccezione dell'ultimo bit che è stato appunto incrociato.

$$A_1 = 0\ 1\ 1\ 0 \mid 1$$
$$A_2 = 1\ 1\ 0\ 0 \mid 0$$

$$A'_1 = 0\ 1\ 1\ 0\ 0$$
$$A'_2 = 1\ 1\ 0\ 0\ 1$$

L'informazione scambiata nell'incrocio dà agli AG molto del loro potere anche se può sembrare strano.

Gli algoritmi genetici sfruttano la ricchezza delle informazioni della generazione precedente riproducendo quelle di alta qualità secondo le loro performance e combinandole con altre informazioni di alta qualità da altre stringhe.

Si scambiano nozioni al fine di generare nuove idee. Cos'è un'idea innovativa? Molto spesso è una giustapposizione di cose che hanno funzionato bene in passato. Allo stesso modo gli operatori riproduzione e incrocio combinano informazioni al fine di ricercare idee innovative.

Se la riproduzione dei più adatti, combinati con l'incrocio, dà agli AG la massa principale della loro potenza, qual è lo scopo dell'operatore mutazione? La mutazione è necessaria perché, gli operatori riproduzione e incrocio occasionalmente possono diventare troppo zelanti e perdere dell'informazione utile nei passaggi di generazione. L'operatore mutazione protegge dalle perdite irreversibili di informazione. La mutazione è una rara alterazione di un valore di una stringa.
Per essere tale e per dare un'idea generica, al fine di ottenere buoni risultati negli studi empirici di AG la frequenza è dell'ordine di una mutazione ogni migliaia di bit trasferiti.

### 4.2.2 Applicazione

Continuiamo a considerare il semplice problema della massimizzazione della funzione $f(x) = x^2$ con x ∈ [0 , 31] e applichiamo gli operatori visti. Per questo problema abbiamo codificato la variabile x con una codifica binaria senza segno a formare stringhe di lunghezza 5 bit. Con questa codifica possiamo ottenere i numeri dallo 0 (00000) al 31 (11111).
Per iniziare abbiamo selezionato una popolazione iniziale formata da 4 individui, lo facciamo scegliendo i bit grazie a 20 lanci di moneta e supponiamo di ottenere le 4 stringhe precedentemente viste.

| n | Popolazione iniziale | $x_i$ | $f_i(x)$ $x_i^2$ | Probabilità di selezione |
|---|---|---|---|---|
| 1 | 01101 | 13 | 169 | 14,4% |
| 2 | 11000 | 24 | 576 | 49,2% |
| 3 | 01000 | 8 | 64 | 5,5% |
| 4 | 10011 | 19 | 361 | 30,9% |
| **Somma** | | | 1170 | 100% |
| **Media** | | | 293 | 25% |
| **Max** | | | 576 | 49% |

**Tabella 4. 2 Probabilità di selezione**

Una generazione successiva viene definita applicando l'operatore riproduzione. È quindi necessario girare la roulette le cui aree sono date dai pesi dei payoff ottenuti con la funzione obiettivo al fine di selezionare quali stringhe della precedente generazione saranno scelte per l'accoppiamento (verranno cioè inserite nel 'mating pool'). La simulazione di questo processo può portare per esempio all'estrazione delle stringhe 1 e 4 per due volte, ad una estrazione della stringa 2 e a zero volte la stringa 3. Risultato simile a quello che ci aspettiamo se arrotondiamo il risultato atteso che per la prima stringa è 0.58 = 169/293. Tendenzialmente i migliori acquisiscono più copie, quelli vicino alla media restano costanti, i peggiori vengono esclusi dal mating pool.

| n | Popolazione iniziale | X | f(x) $x^2$ | Probabilità di selezione | Conteggio atteso | Estrazione dalla roulette |
|---|---|---|---|---|---|---|
| 1 | 01101 | 13 | 169 | 14,4% | 0,58 | 1 |
| 2 | 11000 | 24 | 576 | 49,2% | 1,97 | 2 |
| 3 | 01000 | 8 | 64 | 5,5% | 0,22 | 0 |
| 4 | 10011 | 19 | 361 | 30,9% | 1,23 | 1 |
| **Somma** | | | 1170 | 100% | 4 | |
| **Media** | | | 293 | 25% | 1 | |
| **Max** | | | 576 | 49% | 1,97 | |

**Tabella 4. 3 Estrazione delle stringhe per la riproduzione**

Il crossover più semplice, come detto avviene in due step: le stringhe scelte per la riproduzione vengono accoppiate in modo casuale e in modo casuale viene scelto il punto della stringa in cui si effettuerà il taglio. Supponiamo che vengano scelte per l'accoppiamento le prime due stringhe con punto per il crossover dopo il quarto bit e che le rimanenti due vengano accoppiate con punto di crossover dopo il secondo bit.

| Mating Pool | stringa crossover | crossover site | Nuova Popolazione | x | f(x) $x^2$ |
|---|---|---|---|---|---|
| 0110\|1 | 2 | 4 | 01100 | 12 | 144 |
| 1100\|0 | 1 | 4 | 11001 | 25 | 625 |
| 11\|000 | 4 | 2 | 11011 | 27 | 729 |
| 10\|011 | 3 | 2 | 10000 | 16 | 256 |
| **Somma** | | | | | 1754 |
| **Media** | | | | | 439 |
| **Max** | | | | | 729 |

**Tabella 4. 4 Crossover**

L'ultimo operatore, la mutazione, ovvero il cambio di uno 0 in 1 o viceversa, viene effettuata a livello di bit sulla totalità dei bit trasferiti alla generazione successiva. Vuol dire che se assumiamo che la probabilità di mutazione, sia per esempio 0.001, con il trasferimento di 20 bit alla generazione successiva ci aspettiamo 20*0.001 = 0.02 bit che subiscono la mutazione. In pratica in questo esempio nessun bit subisce la mutazione.

A seguito dei tre operatori possiamo testare la nuova popolazione, per farlo decodifichiamo semplicemente i valori delle stringhe create dall'algoritmo e calcoliamo la fitness function. Nonostante trarre delle conclusioni da una singola iterazione di un processo stocastico sia un azzardo, è possibile iniziare a osservare come gli algoritmi genetici combinino le informazioni sulla performance precedente per ottenere ulteriori migliori performance. Confrontando le due tabelle sia il valore medio che il valore massimo della nuova popolazione sono migliorati rispetto a quello della popolazione di partenza.

La migliore stringa della prima generazione 11000 è stata presente nella mating pool 2 volte per la sua performance sopra la media. Durante l'accoppiamento casuale con le altre stringhe in un caso grazie all'incrocio al secondo bit (scelto di nuovo casualmente) ha dato origine alla stringa 11011. In questo caso l'ottimo buon risultato può sembrare ancora come qualcosa di puramente euristico ma a breve vedremo come gli algoritmi genetici si possono dimostrare un metodo di ricerca robusto. Per capirlo dobbiamo esaminare i dati disponibili e capire come è possibile siamo riusciti a sfruttare le similitudini della codifica degli stessi dati per rendere più efficiente la ricerca. Questo ci conduce a sviluppare l'importante nozione di schema (o similarity template) che ci condurrà alla building block hypothesis.

### 4.2.2 Schemata

In un processo di ricerca come quello visto, basato soltanto sul payoff ottenuto da una generazione (fitness values), quali informazioni sono ottenibili dalla popolazione di stringhe di partenza e dalla funzione obiettivo al fine di indirizzare e migliorare la ricerca?

| Popolazione iniziale | x | f(x) $x^2$ |
|---|---|---|
| 01101 | 13 | 169 |
| 11000 | 24 | 576 |
| 01000 | 8 | 64 |
| 10011 | 19 | 361 |

**Tabella 4. 5 Popolazione prima generazione**

Certe stringhe hanno buoni risultati e viene spontaneo pensare di incrociare queste. Possiamo notare alcune similarità tra le stringhe, per esempio che alte performance sono associate a quelle stringhe che iniziano con 1. Questa informazione è utile e lo sappiamo per certo data la banalità dell'esempio. Quello che intuitivamente è stato fatto dal nostro processo cognitivo è composto da due step. Per prima cosa abbiamo cercato somiglianze tra le stringhe nella popolazione, dopodichè abbiamo guardato se c'era una relazione tra questi gruppi e gli alti payoff. Nel farlo abbiamo quindi tratto delle utili informazioni per guidare la ricerca.

Gli schemata sono tutti gli schemi, che a loro volta rappresentano un insieme di stringhe caratterizzate da certe somiglianze o similarità poiché hanno in comune certi bit della stringa. Per poterli rappresentare aggiungiamo un terzo simbolo al nostro alfabeto binario che ora diventa {0 , 1 , *}, dove * è un simbolo che ci dice che in quella posizione della stringa può esserci qualsiasi valore tra 0 e 1 indifferentemente. Per esempio, lo schema 0*1** rappresenta tutte le 8 stringhe di lunghezza 5 che iniziano con 0 e hanno 1 in terza posizione:

{ 00100 , 01100 , 00110 , 00101 , 01110 , 01101 , 00111 , 01111 }

Il simbolo utilizzato è un metasimbolo e non verrà processato dall'algoritmo, è semplicemente una notazione.

È possibile ovviamente enumerare i possibili schemi che possono esserci in una stringa di lunghezza 5 con 3 valori possibili (1, 0 e *), essi sono $3^5$ = 243 diversi possibili schemi di questo tipo. Generalizzando su un alfabeto di k elementi non metasimboli a comporre stringhe di lunghezza l è possibile individuare $(k+1)^l$ schemi.

Ad un primo sguardo sembra che gli schemi abbiano reso computazionalmente più complessa la ricerca in quanto lo spazio di ricerca è più grande. Si è passati da $2^5$ a $3^5$ punti nello spazio.

Dobbiamo però considerare l'informazione che riusciamo a trarre da una popolazione di partenza e per farlo dobbiamo prima contare quanti sono gli schemi che questa può generare. Una determinata stringa, per esempio 11111, può essere rappresentante di $2^5$ = 32 schemi. Ogni sua posizione può essere un 1 o una * e i 32 schemi così creati devono poter dare la stringa di partenza. In generale una stringa definita può fornire $2^l$ schemi e una popolazione di n stringhe ne contiene massimo $n * 2^l$. Massimo perché in una popolazione ci possono essere anche due stringhe uguali che quindi forniscono gli stessi schemi, dipende quindi dalla diversità all'interno della popolazione. Anche in una popolazione piccola sono contenute delle discrete quantità di informazioni dovute alle somiglianze e vedremo come gli algoritmi genetici sfruttano efficientemente queste informazioni.

Ma quanti degli schemi tra il valore minimo $2^l$ , nel caso di popolazione di stringhe uguali, e $n * 2^l$ sono efficientemente sfruttati dagli algoritmi genetici? Per rispondere dobbiamo considerare cosa succede grazie agli operatori riproduzione, crossover e mutazione. Grazie alla riproduzione gli schemi più performanti hanno più probabilità di finire nella mating pool e quindi in media nelle generazioni successive osserveremo sempre di più i migliori. L'operatore crossover può lasciare illeso uno schema oppure distruggerlo. Lo schema 1***0 è molto probabile che venga distrutto al contrario dello schema *11**. Si può quindi dedurre che gli schemi con alte performance con i bit definiti ravvicinati si propagheranno molto più facilmente attraverso le generazioni. Gli schemi con queste caratteristiche durevoli agli operatori verranno chiamati "Building Blocks".

Scopriremo che il numero di schemi efficacemente processati dagli AG e dell'ordine di $n^5$ dove n è la numerosità della popolazione. Il concetto di schemi efficacemente processati prende il nome di "*parallelismo implicito".*

## *4.3 Definizione numerica degli AG*

Quantitativamente abbiamo trovato che ci sono un gran numero di somiglianze da sfruttare in una popolazione di stringhe. Abbiamo visto come gli algoritmi genetici sfruttano in parallelo le molte similarità contenute nei building blocks ovvero nei brevi schemata che conducono ad alte performance. In questo capitolo vedremo questi risultati in un modo più rigoroso.

Per prima cosa conteremo gli schemi rappresentati all'interno di una popolazione di stringhe e considereremo quali aumenteranno e quali no ad ogni generazione. Per farlo consideriamo gli effetti degli operatori di riproduzione, crossover e mutazione su un particolare schema. Quest'analisi condurrà al fondamentale teorema degli algoritmi genetici che quantifica gli aumenti e le diminuzioni degli schemi più precisamente di quanto visto fin'ora. La forma di questa spiegazione si connette ad un importante e classico problema della teoria decisionale, il "two-armed bandit problem" e la sua estensione "k-armed bandit problem". Contando il numero di schemi processati utilmente rivela l'importante efficacia nel processare i building blocks.

Infine, ci porremo un'importante domanda: come sappiamo che la combinazione dei building block conduce ad una migliore performance nei problemi decisionali?

### 4.3.1 Il Teorema Fondamentale degli Algoritmi Genetici

Si inizia con la scelta casuale di una popolazione di n stringhe, che vengono copiate con una certa inclinazione verso le stringhe migliori, accoppiate e incrociate in una loro parte, mutando qualche bit occasionale per ottenere delle migliori misure. Precedentemente abbiamo iniziato a riconoscere che l'esplicito processare delle stringhe causa un implicito processare di molti schemi contenuti nella popolazione.

Consideriamo le stringhe come un vettore V = {0 , 1}. Ci riferiremo all'intera stringa con la lettera maiuscola mentre i singoli valori al suo interno verranno riferiti con la lettera minuscola con pedice la posizione degli stessi all'interno del vettore. Per esempio, la stringa di 5 bit A = 01100 viene rappresentata simbolicamente come segue:

$$V = a_1 a_2 a_3 a_4 a_5$$

Consideriamo poi una popolazione di j stringhe Aj di una determinata generazione al tempo t, A(t). Ci serve inoltre definire con una notazione gli schemi contenuti nelle stringhe, quindi consideriamo lo schema H come facente parte di un vettore coi caratteri V+ = {0 , 1 , *} dove l'asterisco indica il metasimbolo con valore 0 o 1 indifferentemente. Per esempio, considerando lo schema H = *11* 0** sappiamo che la stringa V = 1111001 è una manifestazione dello schema H.

Ricordando che per stringhe di lunghezza l ci sono $3^l$ schemi, in generale per stringhe di alfabeti composti da codifica a k elementi ci sono $(k+1)^l$ schemi. In una popolazione con *n* stringhe già definite ci sono al massimo $n * 2^l$ schemi come visto precedentemente.

Alcuni schemi sono più specifici di altri. Per esempio, lo schema 011*1** è più definito rispetto a 0******. Alcuni schemi, invece, coprono lunghezze maggiori di altri come per esempio 1****1* rispetto 1*1****. Per quantificare questa idea introduciamo due proprietà degli schemi: l'ordine e la lunghezza definita.

L'ordine di uno schema H, definito o(H) è semplicemente il numero di posizioni fissate presenti nel template dello schema. In 011*1** è o(H) = 4.

La lunghezza definita dello schema H, definito δ(H), è la distanza tra la prima e l'ultima posizione specificata. In 011*1** è 4, nello schema 0****** invece la lunghezza definita è δ(H) = 0.

L'effetto dell'operatore di riproduzione sul numero atteso di schemi nella popolazione è piuttosto semplice da calcolare. Supponiamo che ad un determinato step t ci siano m esemplari di uno schema H contenuti nella popolazione A(t), m = m (H, t). Durante la

riproduzione, una stringa viene copiata in base alla sua fitness, è selezionata con probabilità $p_i = f_i / \sum f_j$. Dopodichè ci aspettiamo di avere $m(H, t+1)$ esemplari dello schema H nella popolazione al tempo t+1.

$$m(H, t+1) = m(H, t) \cdot n \cdot \frac{f(H)}{\sum f_j} \tag{4.1}$$

Dove $f(H)$ è la fitness media di tutte le stringhe che sono rappresentative dello schema H nella popolazione al tempo t. Considerando che la fitness media dell'intera popolazione è calcolata come $\bar{f} = {\sum f}/{n}$ possiamo riscrivere la precedente come:

$$m(H, t+1) = m(H, t) \cdot \frac{f(H)}{\bar{f}} \tag{4.2}$$

Quindi un determinato schema al tempo $t$ cresce in una generazione di un fattore moltiplicativo dato dal rapporto tra la fitness media dello schema e la fitness media dell'intera popolazione. In altre parole, gli schemi con una fitness media superiore alla fitness media della popolazione riceveranno un numero crescente di rappresentanti nella popolazione successiva, invece diminuiranno gli schemi per cui questo non è vero. È particolarmente interessante osservare che questo particolare fenomeno si manifesta per tutti gli schemi che sono rappresentati dalla popolazione di stringhe, in parallelo, simultaneamente, solo grazie all'operatore di riproduzione.

Supponiamo ora che la fitness media di uno schema sia superiore alla fitness media della popolazione di un ammontare $c\bar{f}$ con c una costante. Allora:

$$m(H, t+1) = m(H, t) \cdot \frac{\left(\bar{f} + c\bar{f}\right)}{\bar{f}} = (1 + c) \cdot m(H, t) \tag{4.3}$$

Cominciando con un t = 0 e ipotizzando che la costante c sia stazionaria nel tempo

$$m(H, t) = m(H, 0)(1 + c)^t \tag{4.4}$$

L’effetto della riproduzione è ora quantitativamente chiaro.

Ma la riproduzione da sola non promuove l’esplorazione di nuove regioni dello spazio di ricerca. Copiare solamente vecchie strutture senza nessun cambiamento non rende possibile il testare nuove combinazioni. È qui che entra in gioco l’operatore crossover, creando nuove stringhe a partire dalle selezionate ma effettuando una piccola erosione del lavoro di scelta effettuato dall’operatore riproduzione.

Per vedere quali schemi subiscono maggiormente l’effetto del crossover e quali meno consideriamo una particolare stringa con $l = 7$ e due schemi diversi che lo possono rappresentare:

$$A = 0\ 1\ 1\ 1\ 0\ 0\ 0$$

$$H_1 = *\ 1\ *\ *\ *\ *\ 0$$

$$H_2 = *\ *\ *\ 1\ 0\ *\ *$$

Il crossover procede dunque con la selezione casuale degli accoppiamenti e del punto in cui si effettua il taglio. Per questo esempio supponiamo che per il sito di crossover esca il numero 3 e quindi il taglio avverrà tra la posizione 3 e la 4. Vediamo quali sono gli effetti sui due schemi:

$$A = 0\ 1\ 1\ |\ 1\ 0\ 0\ 0$$

$$H_1 = *\ 1\ *\ |\ *\ *\ *\ 0$$

$$H_2 = *\ *\ *\ |\ 1\ 0\ *\ *$$

Lo schema $H_1$ viene distrutto perché i suoi elementi definiti andranno in due discendenti diversi, mentre lo schema $H_2$ sopravvive nelle generazioni successive. È chiaro che il primo schema ha molte meno probabilità di sopravvivenza data la sua struttura, il sito di crossover ha molti punti in cui può capitare per causarne la distruzione. Mentre invece il secondo schema viene distrutto solo se per il sito di crossover esce 4. Per quantificare osserviamo che il primo schema ha una lunghezza definita pari a 5 (posizione 7 meno posizione 2), i

possibili siti per il crossover sono $l-1$ ovvero 6. Lo schema $H_1$ viene quindi distrutto con probabilità pari a $p_a = \delta(H_1)/(l-1) = 5/6$ . In modo analogo lo schema $H_2$ che ha lunghezza definita $\delta(H_2) = 1$ ha probabilità di essere distrutto pari a $1/6$. Se poi l'operatore crossover non viene applicato a tutte le stringhe ma supponendo venga applicato con probabilità $p_c$ allora la probabilità che lo schema venga distrutto è:

$$p_a = p_c \cdot \delta(H_1)/(l-1) \tag{4.5}$$

Si noti che se $p_c = 1$ vuol dire che l'operatore crossover è applicato per qualsiasi accoppiamento.

Considerando il complemento a 1 si ottiene la probabilità di sopravvivenza di un generico schema:

$$p_s = 1 - p_c \cdot \frac{\delta(H)}{(l-1)} \tag{4.6}$$

È possibile combinare gli effetti di riproduzione e crossover. Assumendo indipendenza fra i due operatori si ottiene dalla semplice moltiplicazione dei due risultati ottenuti:

$$m(H, t+1) = m(H, t) \cdot \frac{f(H)}{\bar{f}} \left[1 - p_c \cdot \frac{\delta(H)}{(l-1)}\right] \tag{4.7}$$

Lo schema H aumenta (o diminuisce) nella generazione successiva di un fattore moltiplicativo a seconda che la sua fitness media sia superiore alla fitness media della popolazione e se ha una lunghezza definita bassa.

L'ultimo operatore, mutazione, è un'alterazione di una singola posizione con probabilità $p_m$. Affinchè lo schema H sopravviva, le posizioni definite $o(H)$ dello schema devono mantenersi e ognuna sopravvive con probabilità $(1-p_m)$. Considerando sempre l'indipendenza fra le probabilità di mutazione all'interno di uno stesso schema, la probabilità di sopravvivenza di uno schema all'operatore mutazione è:

$$p_s = (1-p_m)^{o(H)} \tag{4.8}$$

Per valori molto piccoli di $p_m$, come avviene nella pratica per definizione stessa dell'operatore, l'equazione è approssimabile a:

$$p_s = 1 - o(H) \cdot p_m \tag{4.9}$$

Mettendo tutto assieme, sempre assumendo indipendenza tra gli operatori, si ottiene il valore atteso del numero di schema H dopo una generazione:

$$m(H, t+1) = m(H, t) \cdot \frac{f(H)}{\bar{f}} \left[1 - p_c \cdot \frac{\delta(H)}{(l-1)}\right] (1 - o(H) \cdot p_m) \tag{4.10}$$

Ma ignorando il contributo, tendente a 0, del prodotto:

$$p_c \cdot \frac{\delta(H)}{(l-1)} * \ o(H) \cdot p_m \to 0 \tag{4.11}$$

La formula può essere approssimata a:

$$m(H, t+1) = m(H, t) \cdot \frac{f(H)}{\bar{f}} \left[1 - p_c \cdot \frac{\delta(H)}{(l-1)} - o(H) \cdot p_m\right] \tag{4.12}$$

In conclusione, gli schemi con lunghezza definita breve, di ordine basso (o(H)), con media dei valori payoff delle stringhe rappresentanti lo schema in popolazione superiore alla media della popolazione, ricevono un incremento notevole nelle generazioni successive. Questa conclusione è così importante che gli viene attribuito un nome speciale: ***'Schema theorem'*** o Teorema Fondamentale degli Algoritmi Genetici.

### 4.3.2 Schema Processing

Osserviamo ora come vengono processati gli schemi dall'algoritmo genetico con un esempio sulla base del precedente. Teniamo presente solo 3 schemi:

| | |
|---|---|
| $H_1$ | 1 * * * * |
| $H_2$ | * 1 0 * * |
| $H_3$ | 1 * * * 0 |

**Sting Processing**

| n | Popolazione iniziale | x | f(x) $x^2$ | Probabilità di selezione | Conteggio atteso | Estrazione dalla roulette | Mating Pool | stringa crossover | crossover site | Nuova Popolazione | x | f(x) $x^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 01101 | 13 | 169 | 14.4% | 0.58 | 1 | 0110\|1 | 2 | 4 | 01100 | 12 | 144 |
| 2 | 11000 | 24 | 576 | 49.2% | 1.97 | 2 | 1100\|0 | 1 | 4 | 11001 | 25 | 625 |
| 3 | 01000 | 8 | 64 | 5.5% | 0.22 | 0 | 11\|000 | 4 | 2 | 11011 | 27 | 729 |
| 4 | 10011 | 19 | 361 | 30.9% | 1.23 | 1 | 10\|011 | 3 | 2 | 10000 | 16 | 256 |
| **Somma** | | | 1170 | 100% | 4 | | **Somma** | | | | | 1754 |
| **Media** | | | $\bar{f}$ **293** | 25% | 1 | | **Media** | | | | | 439 |
| **Max** | | | 576 | 49% | 1.97 | | **Max** | | | | | 729 |

**Schema Processing**

| | | | reproduction | | crossover | |
|---|---|---|---|---|---|---|
| | | Schema Avarage Fitness | Conteggio Atteso | Conteggio da estrazione | Conteggio atteso | Conteggio attuale |
| $H_1$ | 1 * * * * | **469** | 3.2 | 3 | 3.2 | 3 |
| $H_2$ | * 1 0 * * | **320** | 2.18 | 2 | 1.64 | 2 |
| $H_3$ | 1 * * * 0 | **576** | 1.97 | 2 | 0 | 1 |

Tabella 4. 6 Schema Processing

Durante la fase di riproduzione, le stringhe vengono copiate in accordo con la probabilità dovuta al loro fitness values. Lo schema $H_1$ che nella popolazione iniziale è rappresentato da due stringhe (la stringa numero 2 e la numero 3) vedrà in tutto 3 rappresentazioni nella mating pool. Può il teorema fondamentale dare lo stesso risultato? Ci aspettiamo:

$$m(H_1, t+1) = m(H_1, t) \cdot \frac{f(H)}{\bar{f}} = 2 \cdot \frac{\frac{576+361}{2}}{293} = 2 \cdot \frac{468.5}{293} = 3.2 \; schemi$$

| | | | reproduction | |
|---|---|---|---|---|
| | | **Schema Avarage Fitness** | **Conteggio Atteso** | **Conteggio da estrazione** |
| $H_1$ | 1 * * * * | **469** | 3.2 | 3 |
| $H_2$ | * 1 0 * * | **320** | 2.18 | 2 |
| $H_3$ | 1 * * * 0 | **576** | 1.97 | 2 |

**Tabella 4. 7 Operatore riproduzione applicato agli schemi**

L'operatore crossover non può avere nessun effetto sullo schema $H_1$ dato che ha lunghezza definita pari a zero. Per quanto riguarda gli altri due schemi invece le cose sono differenti. Prendiamo lo schema $H_3$ che ha lunghezza definita maggiore e pari a 4, l'operatore crossover distrugge questo schema con probabilità:

$$p_c = \frac{\delta(H_3)}{(l-1)} = 1$$

$$m(H_3, t+1) = m(H_3, t) \cdot \frac{f(H)}{\bar{f}} \left[1 - \frac{\delta(H_3)}{(l-1)}\right] = 1 \cdot \frac{576}{293}[1-1] = 0 \; schemi$$

| | | | reproduction | | crossover | |
|---|---|---|---|---|---|---|
| | | Schema Average Fitness | Conteggio Atteso | Conteggio da estrazione | Conteggio atteso | Conteggio attuale |
| $H_1$ | 1 * * * * | **469** | 3,2 | 3 | 3,2 | 3 |
| $H_2$ | * 1 0 * * | **320** | 2,18 | 2 | 1,64 | 2 |
| $H_3$ | 1 * * * 0 | **576** | 1,97 | 2 | 0 | 1 |

**Tabella 4. 8 Operatore crossover applicato agli schemi**

Questi semplici e banali calcoli sembrano confermare il teorema visto precedentemente, gli schemi con basso ordine, di lunghezza definita breve e con media maggiore a quella della popolazione hanno maggiore probabilità di sopravvivere. Ma perché questa strategia di allocazione è la giusta strada da seguire? Perché la crescita esponenziale degli individui più performanti è efficiente al fine di avvicinarsi all'ottimo?

### 4.3.2 K-Armed Bandit Problem

Il problema della slot-machine a due leve (estendibile a k leve) è un importante problema della teoria decisionale. La soluzione ottima a questo problema è molto simile alla soluzione utilizzata dall'algoritmo genetico.

Supponiamo di avere inizialmente una slot a due leve, utilizzando la leva di sinistra otterremo un payoff pari a $\mu_1$ e varianza $\sigma_1^2$ , differenti valori utilizzando la leva di destra ma con $\mu_1 \geq \mu_2$ .

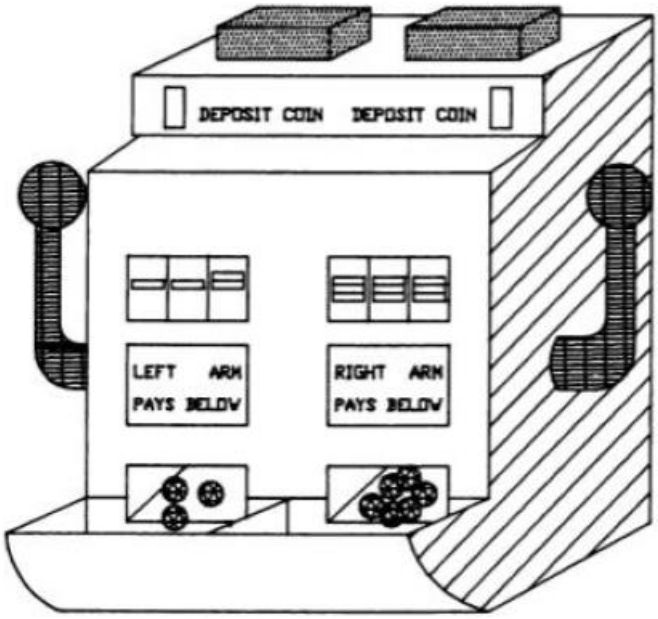


**Figura 4. 5 2-Armed Bandit**[22]

A primo avviso verrebbe da utilizzare la leva 1 anche se non si è a conoscenza della relazione tra le varianze. Abbiamo a disposizione alcune opportunità e non dobbiamo solo prendere la decisione di quale leva tirare ma a ogni opportunità dobbiamo anche raccogliere informazioni utili per i futuri tentativi. Questo compromesso tra la raccolta e lo sfruttamento dell'informazione è un tema ricorrente e fondamentale nella teoria dei sistemi adattivi.

Supponiamo di avere un totale di *N* tentativi da allocare sulle due leve. Possiamo utilizzare in prima battuta *n* tentativi per ogni braccio con $2n < N$ durante questa prima fase di esplorazione. Alla fine di questa raccolta di informazioni prenderemo una decisione e allocheremo i restanti tentativi ($N - 2n$) al braccio valutato migliore. In questo modello semplificato, la decisione viene presa in un momento solo alla fine della sperimentazione dopodichè non può essere modificata.

Possiamo calcolare l'expected loss come:

[22] Fonte: Goldberg, D., Genetic Algorithms in Search, Optimization, and Machine Learning

$$L(N,n) = \mid \mu_1 \; - \; \mu_2 \mid * [(N-n)q(n) + n\big(1 - q(n)\big)] \qquad (4.13)$$

Dove *q(n)* è la probabilità che la leva peggiore sia quella che abbiamo valutato come migliore dopo la fase esplorativa.

Possiamo vedere dall'equazione che vi sono due fonti di perdita associate alla procedura che scegliamo. La prima perdita è dovuta all'aver usato *n* tentativi nella fase di sperimentazione sulla leva sbagliata. La seconda è la conseguenza in caso si scelga, per concludere i tentativi restanti, la leva a cui corrisponde il minor payoff nonostante la fase di sperimentazione. Non è infatti certo che alla fine della fase di sperimentazione opteremo per la scelta migliore.

Al fine di ottimizzare la scelta della numerosità di tentativi da utilizzare per l'esplorazione è necessario derivare l'equazione precedente e porla uguale a 0. Avendo supposto la distribuzione delle scelte sulla leva corretta o sbagliata dopo la fase di sperimentazione approssimabile alla distribuzione normale, si ottiene (Holland 1975)[23]:

$$n^* \cong \; b^2 \ln \left[\frac{n^2}{8\pi b^4 \ln N^2}\right] \quad \text{dove} \quad b = \; \sigma_1/(\mu_1 - \; \mu_2) \qquad (4.13)$$

Questo valore definisce quanto dovrebbe essere il numero di tentativi da allocare per braccio nella fase di sperimentazione al fine di minimizzare la perdita attesa. L'approccio seguito mostra come esponenzialmente meno prove vengono attribuite alla leva sbagliata all'aumentare dei tentativi utilizzati in sperimentazione. Un metodo che si avvicina all'allocazione ideale dei tentativi di prova è proprio dato dall'utilizzo degli algoritmi genetici a 3 operatori che grazie alla teoria degli schemi garantisce un maggior numero di tentativi ai building block più performanti, rivelandosi un ottimo metodo per la ricerca tra soluzioni alternative.

[23] Holland, J. H. (1975).Adaptation in natural and artificial system. Ann Arbor:University of Michigan Press

Tramite gli algoritmi genetici non ci limitiamo a risolvere il problema della slot a 2 leve ma simultaneamente risolviamo il problema di diverse slot multi-leva (k-armed bandit). Per esempio, consideriamo il seguente insieme di schemi di grado 3 sulle posizioni 2,3 e 5:

* 0 0 * 0 * *

* 0 0 * 1 * *

* 0 1 * 0 * *

* 1 0 * 0 * *

* 0 1 * 1 * *

* 1 1 * 0 * *

* 1 0 * 1 * *

* 1 1 * 1 * *

Ogni leva della slot rappresenta uno degli schemi. Al fine di ottimizzare la funzione obiettivo abbiamo bisogno di allocare sempre più scelte per la riproduzione sullo schema maggiormente performante. L'applicazione dell'algoritmo genetico, oltretutto, non permette solo di risolvere questo singolo problema per questo tipo di schema, vengono infatti processati in parallelo diversi k-armed bandit problem. Per esempio, con 3 posizioni fissate su stringhe di lunghezza 7 ci sono $\binom{7}{3} = 35$ schemi diversi con 8 possibili configurazioni ciascuno. Parliamo di 35 slot a 8 leve ciascuna.

### 4.3.3 Schemi Processati Efficientemente

Precedentemente è stato detto che data una popolazione di *n* stringhe di lunghezza $l$ si possono identificare da un minimo di $2^l$ ad un massimo di $n2^l$ schemi. Ma come sappiamo è molto probabile che molti schemi non siano processati per via della loro stessa struttura, l'operatore crossover tende a distruggere quegli schemi che hanno lunghezza definita alta.

Holland definì una relazione molto semplice per quantificare gli schemi processati efficientemente dagli AG. Attraverso l'elaborazione di *n* stringhe a ogni generazione, l'algoritmo genetico processa una quantità di schemi dell'ordine di $n^3$. Questa relazione tra popolazione processata e schemi processati in parallelo dagli AG prende il nome di *parallelismo implicito*. Si verifica una sorta di leva computazionale, senza la necessità di inserire altri dati al di fuori della popolazione di stringhe già fornita.

Nonostante la probabile distruzione degli schemi con ordine e lunghezza definita elevata da parte dell'operatore crossover, gli algoritmi genetici sono in grado di processarne una grande quantità a partire da una relativa piccola quantità di stringhe di partenza e di ricavarne utili informazioni per la generazione successiva.

### 4.3.4 La Building Block Hypothesis

Con building blocks si intendono tutti quegli schemi di lunghezza definita breve, di ordine basso e con una fitness maggiore della fitness media della popolazione. Lavorando con questi particolari schemi si può ridurre la complessità del problema; invece di costruire stringhe di alte prestazioni provando ogni tipo di combinazione, si costruiscono sempre migliori stringhe a partire dalle soluzioni parziali migliori delle generazioni precedenti (i building blocks).

Ci sono molte evidenze empiriche a favore dell'efficacia computazionale di utilizzare i building blocks. A partire da semplici problemi unimodali arrivando fino a più ostici problemi multimodali soggetti a rumore, si sono raggiunti ottimi risultati usando un semplice algoritmo genetico con i tre semplici operatori di reproduction, crossover e mutation.

Bethke e Holland hanno contribuito alla definizione teorica analitica di questa ipotesi e i concetti dietro queste scoperte saranno esposti in maniera grafica al fine di capire la regolarità implicita nel processare i building blocks. Per farlo continueremo ad usare l'esempio di stringhe a 5 bit utilizzato al fine di massimizzare la funzione $f(x) = x^2$.

Consideriamo il semplice schema $H_1 = 1 * * * *$ , come mostrato in figura X dall'area grigia, i possibili valori assunti da questo schema rappresentano la metà superiore del dominio. Di controparte lo schema $H_2 = 0 * * * *$ comprende i valori del dominio identificati dall'area bianca.

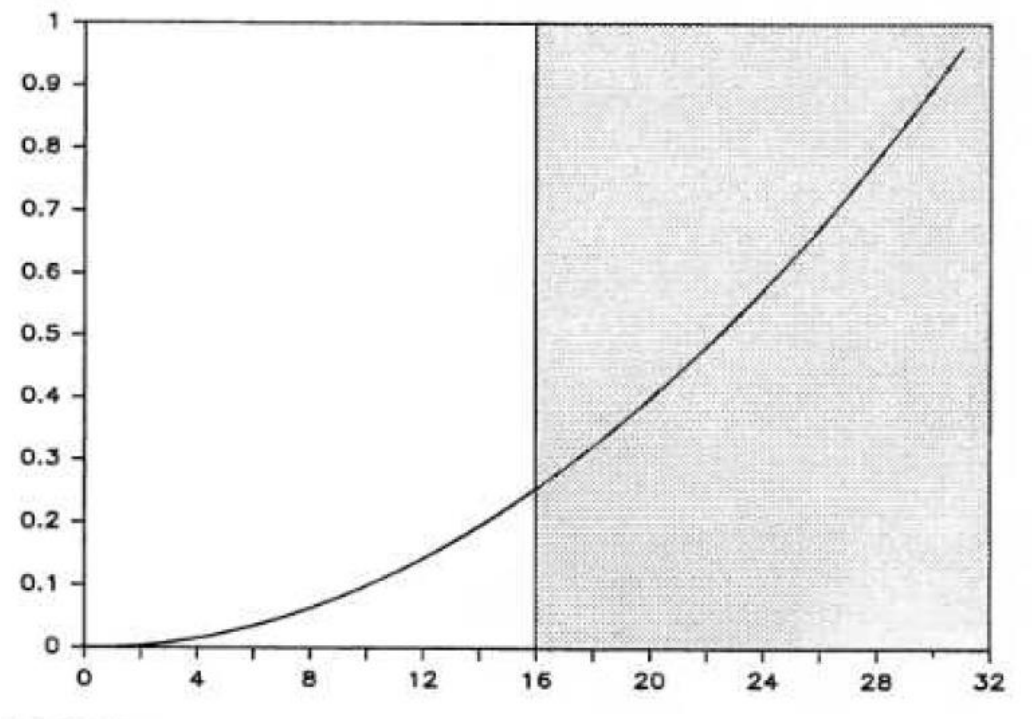


**Figura 4. 6 Dominio dello Schema di ordine 1** [24]

Un altro schema ad un bit definito come $H_3 = * * * * 1$, come mostrato in figura 4.7, rappresenta sempre metà del dominio e più precisamente tutti i numeri dispari (00001=1, 00011=3, 00101=5, etc.). In bianco lo schema $H_4 = * * * * 0$.

[24] Fonte: Goldberg, D., Genetic Algorithms in Search, Optimization, and Machine Learning

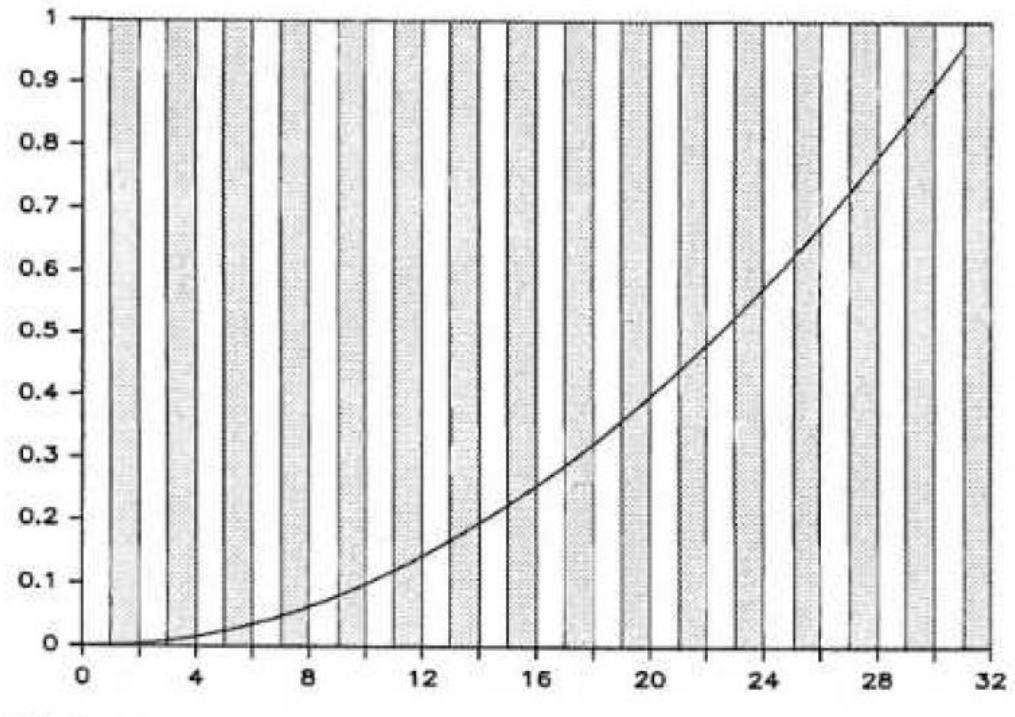


**Figura 4. 7 Dominio dello Schema di ordine 1 (2)**[25]

In figura X invece, lo schema $H_5 = * * * 0 *$, anch'esso copre metà dominio ma tramite fasce che coprono più valori (00000=0, 00001=1,00100=4, 00101=5, 8 e 9 ,12 e 13, etc.).

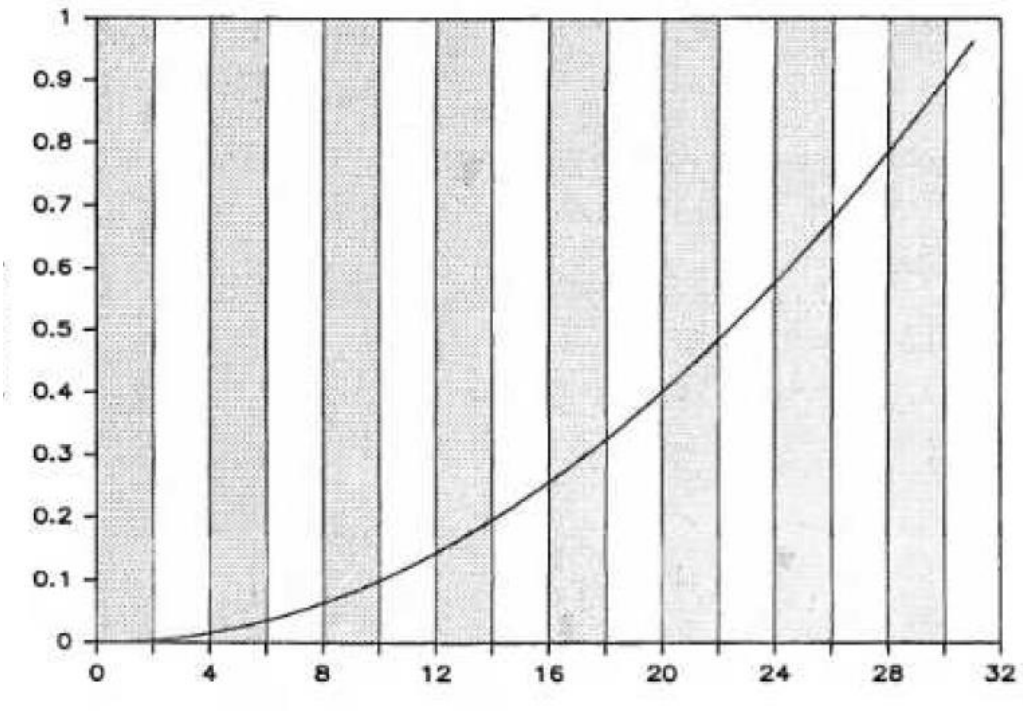


**Figura 4. 8 Dominio dello Schema di ordine 1 (3)**[26]

[25] Fonte: Goldberg, D., Genetic Algorithms in Search, Optimization, and Machine Learning
[26] Fonte: Goldberg, D., Genetic Algorithms in Search, Optimization, and Machine Learning

Sembra che uno schema ad un bit copra sempre mezzo dominio ma la frequenza delle oscillazioni dipenda dalla posizione del bit definito.

Consideriamo invece lo schema a due bit $H_6 = 1\ 0\ * * *$ mostrato in figura 4.9. Esso copre solo un quarto del dominio, più precisamente il terzo quarto. Il quarto quarto è coperto dallo schema che inizia con 1 1, il primo e il secondo rispettivamente dagli schemi che iniziano per 0 0 e 0 1.

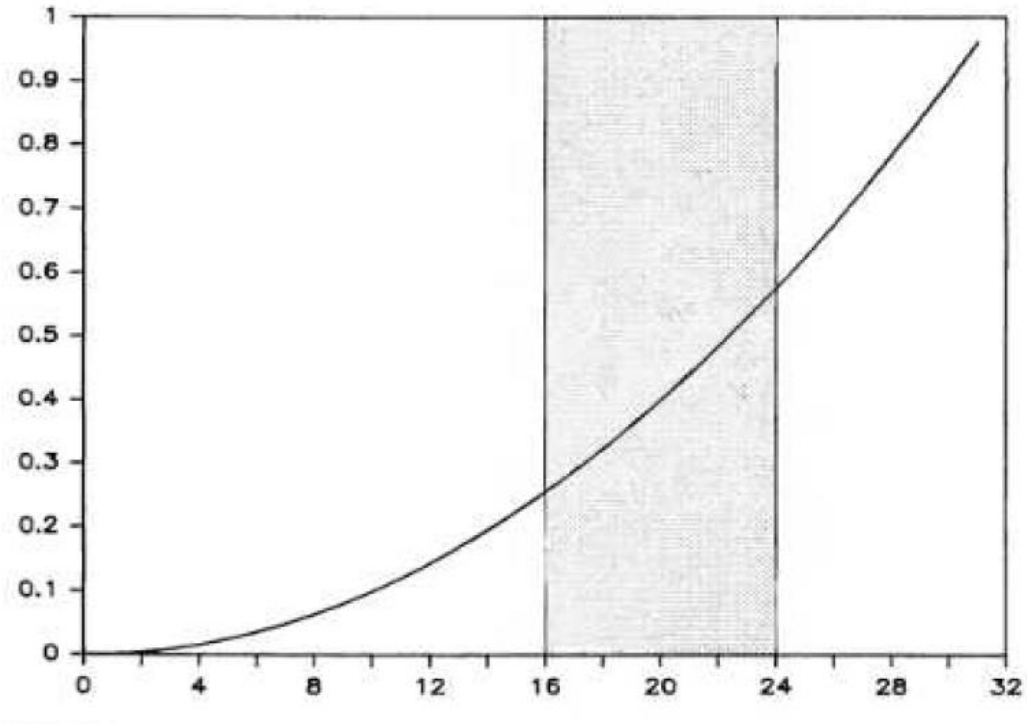


**Figura 4. 9 Dominio dello Schema di ordine 2**[27]

Un ultimo esempio più particolare può essere lo schema $H_6 = * * 1\ * 0$ che mostra una maggior frequenza e copre sempre un quarto del dominio ma in maniera discontinua.

[27] Fonte: Goldberg, D., Genetic Algorithms in Search, Optimization, and Machine Learning

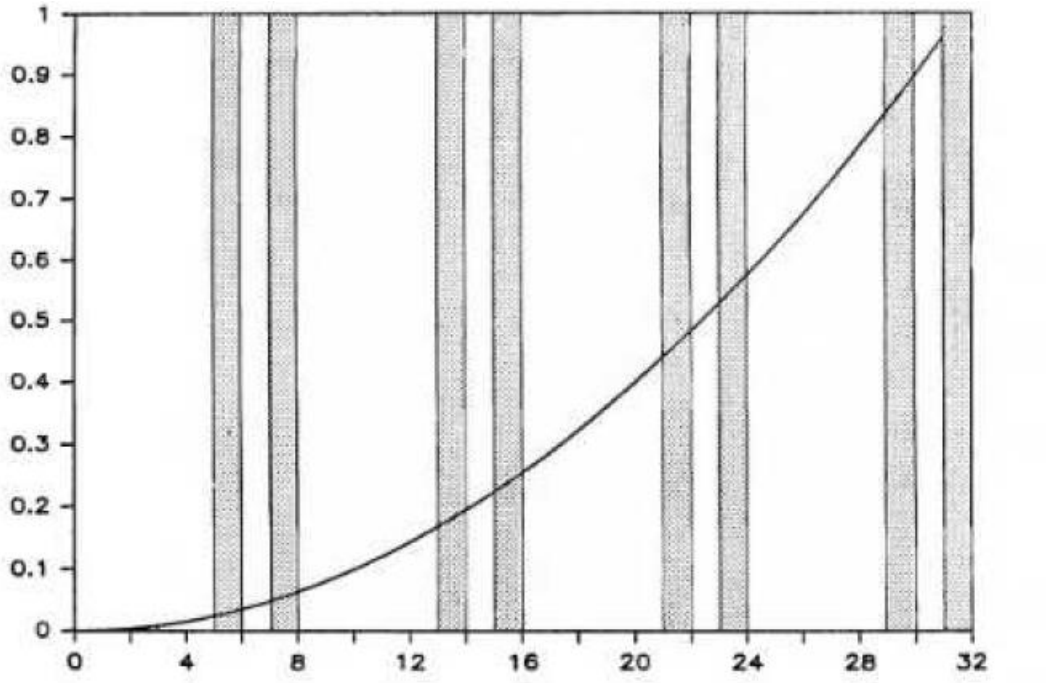


**Figura 4. 10 Dominio dello Schema di ordine 2 (2)[28]**

Betchke ha studiato la periodicità degli schemi generando potenti strumenti matematici applicabili però in specifici casi che non approfondiamo. Purtroppo, la generalizzazione a un varietà più ampia di casi è difficile. Betchke ha però definito alcuni casi molto particolari che sono ingannevoli per gli algoritmi genetici a tre operatori. Si tratta principalmente di quei casi in cui l'ottimo è isolato, ovvero circondato da soli valori pessimi. Ma non ci dobbiamo preoccupare eccessivamente perché la maggior parte delle funzioni rappresentanti i problemi del mondo reale non hanno questa particolarità, anzi sono proprio le regolarità che caratterizzano le funzioni e che vengono sfruttate in maniera efficiente dagli AG per raggiungere l'ottimo.

[28] Fonte: Goldberg, D., Genetic Algorithms in Search, Optimization, and Machine Learning

# 5 APPLICAZIONI

Data l'importanza del credit scoring e il suo potenziale impatto nel business di una banca, non è sorprendente che i tradizionali metodi utilizzati per giudicare la rischiosità di una controparte siano costantemente aggiornati. Numerosi studi comparano le performance dei metodi tradizionali, basati sulla regressione logistica o sull'analisi discriminante lineare, con i moderni approcci che sono e saranno sviluppati.

Le tecniche ispirate ai processi biologici come le reti neurali e gli algoritmi genetici sono sempre più popolari. Il loro potere predittivo viene messo in relazione con alcuni dei modelli più tradizionali e alcuni studi mostrano che possono produrre previsioni più accurate mentre altri suggeriscono il contrario.

Gli algoritmi genetici, come precedentemente detto, furono introdotti da Holland come un'astrazione dell'evoluzione biologica (Holland, 1975)[29]. L'algoritmo permette di far evolvere una popolazione di partenza in una successiva tramite gli operatori, ognuna è costituita da un insieme di codici genetici che definiscono le particolari soluzioni di ogni individuo al problema in esame. Ad ogni individuo è quindi attribuibile un fitness score che influenza la sua capacità di riprodursi e trasmettere le sue informazioni alla prossima generazione.

Le performance degli algoritmi genetici dipendono in larga parte dai parametri sotto il controllo del ricercatore che imposta il modello. Questi parametri e anche la fitness function richiedono di essere attentamente selezionati al fine di ottenere le alte performance attese da un modello di scoring.

Al momento gli algoritmi genetici applicati ai modelli di scoring vengono utilizzati in due modi differenti. La prima area di applicazione ne prevede l'utilizzo con un approccio ibrido

[29] Holland, J. H. (1975). Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence. Ann Arbor: University of Michigan Press

in cui gli AG sono usati in concerto con altri metodi, principalmente le reti neurali. Per esempio, sono spesso usati per selezionare le variabili poi utilizzate da reti neurali o dalla regressione logistica (Šušteršič et al., 2009)[30](Oreski et al., 2012)[31].

La seconda applicazione vede l'utilizzo degli AG standalone per la formulazione del modello. Alcuni di questi, tratti dalla letteratura verranno di seguito esposti per poter apprezzare le diversità e risultati ottenuti al fine di trarre successivamente le dovute conclusioni.

## *5.1 Perché gli Algoritmi Genetici?*

Successivamente agli studi di Altman nel 1968, numerosi metodi sono stati utilizzati per prevedere l'insolvenza di un'impresa. Inizialmente l'analisi discriminante multivariata (MDA) fu la tecnica dominante nonostante i vincoli di cui soffre ovvero la linearità, la necessità della distribuzione normale dei dati e l'indipendenza tra le variabili. Considerando che la violazione di questi limiti capita spesso quando si tratta i diversi dati finanziari di un'impresa, i quali tendenzialmente non godono di linearità, normalità e soprattutto non sono indipendenti gli uni dagli altri, l'attenzione si è spostata sulla regressione logistica (LR). I modelli LR richiedono infatti meno vincoli, viene assunto solo che le variabili seguano una distribuzione logistica, nonostante ciò, offrono un miglior potere discriminante. Negli anni '80 le tecniche ad intelligenza artificiale, in particolar modo Artificial Neural Network (ANNs) e Support Vectore Machine (SVM) sono state efficacemente utilizzate per predire l'insolvenza grazie alle loro capacità di identificare e rappresentare le relazioni non lineari all'interno dei dati analizzati. Da numerosi studi sembrerebbe che l'indice di accuratezza dimostrato dai modelli ANNs sia migliore rispetto a quello raggiunto dagli altri metodi ma vengono rilevati alcuni difetti come la dipendenza dalle capacità del ricercatore per

[30] Sustersic, M., Mramor, D., & Zupan, J. (2009). Consumer credit scoring models withlimited data. Expert Systems with Applications, 36(3), 4736–4744.http://dx.doi.org/10.1016/j.eswa.2008.06.01

[31] Oreski, S., Oreski, D., & Oreski, G. (2012). Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. Expert Systems with Applications, 39(16), 12605–12617.

selezionare i parametri, la difficoltà nell'evitare l'overfitting e il fatto che risultano incomprensibili agli utenti le regole stabilite del modello (Gordini, 2014)[32].

Gli algoritmi genetici mimano i principi Darwiniani per risolvere problemi di ottimizzazione non lineari e non convessi. Offrono numerosi vantaggi rispetto le altre tecniche:

1. sono algoritmi adattivi, riuscendo a ottimizzare la soluzione anche al mutamento delle condizioni esterne;
2. utilizzano solo informazioni che determinano il fitness value dell'individuo, le variabili di input. Non richiedendo informazioni aggiuntive, per esempio, per la definizione del gradiente utilizzato come direzione verso la soluzione;
3. collegato al punto due è la capacità degli algoritmi di poter raggiungere l'ottimo globale anche in presenza di numerosi ottimi locali. Mentre le tecniche tradizionali esplorano lo spazio delle soluzioni da un singolo punto rischiando di convergere in un massimo locale, gli AG conducono la ricerca da più punti dello spazio contemporaneamente minimizzando questo rischio;
4. permettono di estrarre regole per la definizione dello score facilmente comprensibili agli utenti che sono gli analisti che hanno il compito di usare la loro esperienza per definire se accettare o rifiutare un credito.

Utilizzati e confrontati con le altre tecniche, utilizzando gli indici di performance analizzati, l'utilizzo degli AG ai modelli di scoring si è rivelata una tecnica promettente.

### 5.1.1 Impostazione generale di un algoritmo

Il credit scoring può essere descritto come un problema di classificazione. Tradizionalmente i clienti vengono classificati in due gruppi, buoni e cattivi.

Gli algoritmi genetici differiscono dalle altre tecniche di ottimizzazione non lineare in quanto durante la ricerca della soluzione, invece di continuare a modificare una sola soluzione, ne

[32] Gordini, N. (2014). A genetic algorithm approach for SMEs bankruptcy prediction: Empirical evidence from Italy. Expert Systems with Applications, 41(14),6433–6445.

mantengono una popolazione dalle quali viene creata una popolazione di soluzioni ancora più performanti e così via fino alla convergenza del modello. La popolazione è costituita da stringhe (cromosomi) ognuna delle quali rappresenta una possibile soluzione al problema. L'evoluzione prende il via dalla fase definita 'Initialization stage' nella quale generalmente viene scelta una popolazione di soluzioni in maniera totalmente casuale.

Come visto nel capitolo 3 vengono selezionate delle variabili, in genere KPI finanziari derivanti dai dati di bilancio ma anche informazioni da banche dati e di carattere sociodemografico. Le variabili vengono selezionate in base alla loro effettiva capacità di fornire informazioni utili al processo di previsione, ma è anche importante che non siano ridondanti fra di loro perché lo scopo di aggiungere una variabile è quello, appunto, di aggiungere informazione.

Selezionate le variabili e raccolti i dati è necessario "pulirli", sono infatti molte le inesattezze che si possono riscontrare nei dati e che si rivelano esclusivamente nocive alla corretta stima del modello, rischiando di vanificare lo sforzo fatto e sfociando in pessime performance.

Ogni cromosoma viene valutato da una fitness function definita dal ricercatore, è il modo in cui vengono misurate le performance individuali della popolazione di soluzioni in modo che se ne possa stilare una classifica. La definizione della fitness function è uno dei momenti critici dello sviluppo dell'algoritmo, essa infatti può influenzare notevolmente l'accuratezza dei risultati ottenuti, come dimostrato da Kozeny (Kozeny, 2015)[33] che utilizza tre diversi tipi di fitness function a parità di dati e parametrizzazione del modello, ottenendo tre valori di accuracy statisticamente differenti. Se per esempio la regressione logistica fornisce come risultato solo una funzione, con gli AG si ottengono diverse funzioni progettate, inoltre, in modo da identificare i migliori o, più precisamente, i più adatti.

Impostata la funzione che permette di scegliere chi potrà contribuire nella generazione successiva, vengono definiti i parametri degli operatori propri degli algoritmi genetici. È necessario quindi stabilire le regole della riproduzione. I cromosomi che hanno occupato le

[33] Kozeny, V. (2015). Genetic algorithms for credit scoring: Alternative fitness function performance comparison. Expert Systems with Applications, 42(6), 2998–3004.

primissime posizioni potrebbero essere copiati tali e quali nella generazione successiva, oppure potrebbero seguire tutti l'algoritmo e quindi la selezione, per esempio tramite roulette wheel. Quando vi è la riproduzione, il crossover può essere applicato con una predeterminata probabilità, a singolo punto di "taglio" o multiplo. Viene inoltre decisa la probabilità di mutazione. Tutte queste scelte inficiano sulle performance dell'algoritmo e sono scelte dal ricercatore sulla base dell'esperienza ma soprattutto su riflessioni. Negli esempi di applicazione successivi sarà possibile notare anche quali sono le motivazioni dietro la scelta degli operatori.

## *5.2 Esempi di applicazione*

### 5.2.1 Applicazione con fitness function polinomiale

Vengono qui esposti i risultati ottenuti utilizzando una fitness function polinomiale da Kozeny (Kozeny, 2015).

Kozeny imposta il modello come un problema di classificazione a due vie, ovvero con due possibili risultati, cliente buono o cliente cattivo. Per farlo ogni cliente *x* è caratterizzato dalla valorizzazione di *D* variabili $x = (x_1, x_2, \dots\ , x_D)$, tramite una funzione *f* ogni cromosoma di un individuo viene attribuito ad una delle due classi $f(\overrightarrow{x_i}) \rightarrow \{good, bad\}$ effettuando una previsione per ogni individuo della popolazione. Ogni elemento del campione avrà il suo stato reale, per ogni individuo siamo infatti a conoscenza se si sia poi rivelato un cliente buono o uno cattivo, definiamo con $y_i\ \epsilon\{good, bad\}$ le osservazioni reali. La fitness function deve mettere in relazione il risultato previsto con l'esito reale dell'individuo e in questa applicazione utilizzata la forma polinomiale:

$$f(x) = \ a_1 x_1^{b1} + \ a_2 x_2^{b2} + \cdots + a_D x_D^{bD} + c \qquad (5.1)$$

Il cromosoma dell'individuo è quindi definito dai parametri codificati di questa funzione: $a_1, \dots, a_D, b_1, \dots, b_D, c$. Come vedremo, verranno utilizzate 11 variabili indipendenti e quindi il cromosoma di un individuo è composto da 23 *"geni"* in totale. Per la fase di inizializzazione questi valori sono attribuiti ad ogni individuo della popolazione in modo casuale fissando solo il range di possibili valori che è per i parametri *a* e *b* tra -10 e +10 mentre per il parametro *c tra -100 e +100*. Per ogni individuo viene così calcolato il fitness value e se *f(x) > 0*, si considera appartenente alla classe dei *good,* se *f(x) ≤ 0* appartiene alla classe *bad.* Una volta attribuita la classe ad ogni individuo, per poter valutare l'attuale parametrizzazione dell'algoritmo viene valutato il fitness score $\Phi$, l'accuratezza definita come il numero di previsioni corrette, in cui $f(\overrightarrow{x_i}) = y_i$, diviso il totale degli individui nella popolazione *N:*

$$\Phi = \frac{\#(i|f(\overrightarrow{x_i}) = y_i)}{N} \tag{5.2}$$

Per la costruzione del modello sono stati utilizzati come campione, i dati dei clienti di una banca del Venezuela che posseggono una carta di credito. Una riflessione però pone l'attenzione sul fatto che utilizzare i dati di clienti di cui si conosce la storia creditizia richiede che questi siano stati precedentemente accettati dall'istituto di credito e che quindi abbiano superato incolumi i precedenti modelli di scoring, questo potrebbe interferire sull'analisi con un bias che però non viene qui tenuto in considerazione.

Il campione è composto da 344 clienti rivelatisi buoni e 145 che invece si sono dimostrati cattivi pagatori, per un totale di 489 individui. Importante la soglia di definizione di una controparte *bad, u*n cliente è definito *bad* in questo caso se è risultato inadempiente per almeno 90 giorni.

Le variabili prese in considerazione, trattandosi di dati per l'emissione della carta di credito, non possono essere totalmente finanziarie perché il soggetto non è un'impresa ma un individuo che non dispone di un bilancio. Sono state selezionate per il modello le seguenti:

1. Età
2. Stato civile
3. Nazionalità
4. Livello di educazione
5. Livello lavorativo
6. Durata in anni dell'attuale impiego
7. Occupazione del marito/moglie
8. Salario
9. Riferimenti da un'altra carta di credito
10. Stato residenziale
11. Possesso di autoveicolo

I dati così raccolti vengono preparati al fine di utilizzare la tecnica della cross validation. Il campione viene diviso nel campione per l'addestramento dell'algoritmo e in quello per il test, quest'ultimo contiene 35 esemplari *good* e 15 *bad.* I restanti 439 individui saranno parte del campione di addestramento.

La suddivisione viene ripetuta 10 volte in modo differente, l'algoritmo verrà quindi addestrato e testato in 10 diversi modi a partire dallo stesso campione di partenza. I risultati complessivi saranno valutati come la media delle osservazioni fatte.

Il modello è inizializzato con la creazione casuale di 200 diversi cromosomi composti ciascuno dai 23 geni corrispondenti ai parametri della fitness function. Dopo l'inizializzazione si susseguono gli step dell'algoritmo:

1. Per ogni cromosoma viene valutata la fitness function su ogni elemento del campione di addestramento utilizzando la fitness function polinomiale (5.2.1) e successivamente viene calcolato il fitness score come dalla (5.2.2) per la soluzione in esame;
2. Ogni soluzione viene posizionata in classifica e il primo 5% viene scelto e copiato senza modifiche nella prossima generazione. Ogni 20 generazioni viene invece

copiata nella generazione successiva il 20% dei migliori cromosomi. Questa tecnica di copia diretta dei migliori nella generazione successiva viene chiamata elitismo;

3. Viene applicato l'operatore riproduzione al resto della popolazione da cui ne viene selezionata una parte attraverso un metodo simile a quello della roulette esposto nel capitolo 3. La roulette è suddivisa in *m* spicchi, dove *m* è il numero di cromosomi su cui opera l'operatore riproduzione, l'ampiezza degli spicchi è proporzionale al fitness valute della soluzione. In questo modo le soluzioni più performanti hanno maggior probabilità di essere selezionate:

$$p_i = {\Phi(f_i)}\Big/{\sum_{i=1}^{m} \Phi(f_i)} \qquad (5.3)$$

   Nel caso, però, vi sia una soluzione ampiamente pi performante delle altre si può notare come il modello tenda a convergere rapidamente ad una soluzione e, poiché in questo caso è stato già scelto di utilizzare anche l'elitismo, per favorire una migliore diversità nelle generazioni successive viene utilizzato un approccio simile alla roulette wheel ma con una lieve differenza. Invece di lanciare la pallina della roulette *n* volte una per ogni selezione di cromosoma per la riproduzione, vengono lanciate *n* palline una volta sola con il vincolo che restino equidistanti fra di loro

4. Vengono applicati gli operatori crossover e mutation. Il primo applicato con probabilità $p_c = 80\%$ e singolo punto di incrocio selezionato casualmente. Sempre per compensare la scarsa diversità procurata dall'elitismo che potrebbe condurre il modello ad incastrarsi un punto di ottimo locale, viene impostato una probabilità di mutazione piuttosto alta, $p_m = 15\%$.

Il modello è vincolato a 110 generazioni oppure termina dopo 50 volte consecutive in cui non muta la soluzione ottima trovata.

Come metodo di valutazione della prestazione del modello sono stati utilizzati gli indicatori di accuratezza, sensibilità e specificità a partire dalla confusion matrix così come sono state

esposte nel capitolo 3.3.5. I valori sono stati raccolti per ogni osservazione effettuata con i campioni di addestramento.

Vengono riportati in tabella 5.1 i valori di accuracy rilevati sul campione di test rispettivamente ai parametri ottenuti con il rispettivo campione di addestramento:

| Osservazione | Accuracy % |
|---|---|
| 1 | 74.18 |
| 2 | 74.42 |
| 3 | 73.92 |
| 4 | 73.96 |
| 5 | 73.53 |
| 6 | 73.92 |
| 7 | 74.14 |
| 8 | 74.38 |
| 9 | 73.98 |
| 10 | 74.6 |
| Media | 74.1 |

**Tabella 5.1 Accuracy dei campioni di addestramento**

Allo stesso modo vengono calcolati i valori di specificità e sensibilità. Data la differenza di costo tra l'errore di prima specie e quello di seconda specie, la sensibilità nei modelli di scoring ha maggior rilevanza poiché valuta la quantità di imprese *bad* correttamente predette sul totale delle imprese *bad* presenti nel campione.

| Accuracy | Specificity | Sensitivity |
|---|---|---|
| 74.10% | 87.05% | 43.89% |

**Tabella 5.2 Performance medie**

### 5.2.2 Confronto con modelli LR e SVM

I valori appena riportati possono non avere molto significato ad un occhio non conscio dei normali valori risultanti da altri modelli di scoring. Per rendere l'idea della possibile utilità nell'applicazione dei modelli generati con algoritmi genetici e dunque necessario un confronto. La letteratura dimostra che non sempre è possibile definire quale modello sia migliore perché le conclusioni variano a seconda del problema analizzato, del tipo di dati presi in considerazione, quindi dalla tipologia di soggetti nella popolazione che può essere molto varia ma anche molto specifica e soprattutto dalla capacità nell'impostare il modello di chi lo costruisce.

Le metodologie con il quale poter costruire un modello sono numerose. Verranno esposti i risultati ottenuti facendo il confronto tra tecnica con algoritmi genetici, regressione logistica (LR) e macchine a supporto vettoriale (SVM) applicati ad una popolazione composta da SME (Small Medium Enterprise) italiane (Gordini, 2014)[34].

Il data set utilizzato è, infatti, composto da una selezione casuale di 3584 piccole e medie imprese italiani operanti nel settore manifatturiero. Questi dati provengono dalla banca dati CERVED che contiene e archivia le informazioni delle società iscritte alla Camera di Commercio italiana. I dati raccolti riguardano un set di indicatori finanziari e, per mantenere l'analisi più stabile e accurata, una volta ordinati, vengono troncati l'1% superiore e inferiore riducendo così il campione a 3100 imprese, 1500 delle quali sono risultate inadempienti. I financial ratio sono raccolti uno, due e tre anni prima l'evento default definito in questo caso come l'inizio dei procedimenti legali causate da inadempienze.

Al fine di testare il potere predittivo dei modelli, il set di dati è suddiviso in due in modo casuale: il campione di addestramento è composto da 2170 imprese, di cui 1050 default, mentre il campione di test è composto da 930 imprese, di cui 450 default.

[34] Gordini, N. (2014). A genetic algorithm approach for SMEs bankruptcy prediction: Empirical evidence from Italy. Expert Systems with Applications, 41(14),6433–6445.

In questo modello viene utilizzata una fitness function con variabile dipendente definita come una dummy, una variabile binaria che acquisisce valore 1 quando l'impresa viene valutata come anomala e 0 quando si predice sana. Le variabili indipendenti sono state selezionate a partire da un set iniziale di 28 indicatori, poi ridotto effettuando delle analisi su quelle che si rivelano più predittive e al contempo non affette da multicollinearità utilizzando il metodo VIF (Variance Inflaction Factor) (Montgomery & Peck, 1992)[35] e successivamente la stepwise analysis per selezionare le più significative. Le variabili che saranno quindi utilizzate sono:

1. Return on Equity (ROE)
2. Return on Investment (ROI)
3. Ebitda/Turnover
4. Interessi passivi/Ebitda
5. Cash flow/Total debt
6. Financial debts /Equity
7. Debiti totali/Ebitda
8. Current ratio

L'obiettivo del modello è quello di definire una regola che conduce ad un esito binario. A partire dalle variabili indipendenti attraverso dei valori di cut-off e i segni di relazione (maggiore o minore uguale), si definisce un'impresa insolvente se vengono rispettate le condizioni poste(Varetto, 1998)[36]. La regola seguirà la seguente struttura logica:

IF [R1 is GREATER THAN OR EQUAL TO (LESS THEN) C1,

AND R2 is GREATER THAN OR EQUAL TO (LESS THEN) C2,

AND ….. ,

AND RN is GREATER THAN OR EQUAL TO (LESS THEN) CN]

[35] Montgomery, D. C., & Peck, E. A. (1992). Introduction to linear regression analysis. New York, NY: John Wiley and Sons

[36] Varetto, F. (1998). Genetic algorithms applications in the analysis of insolvency risk. Journal of Banking and Finance, 22, 1421–1439

THEN Prediction = “DEFAULT”

ELSE Prediction = “GOOD”.

Dove $R_i$ sono i valori delle variabili indipendenti mentre $C_i$ i valori cut-off che vengono ricercati dagli algoritmi genetici per ogni financial ratio. L’operatore logico AND garantisce che solo se tutte le condizioni sono contemporaneamente verificate allora l’impresa verrà classificata come inadempiente.

La popolazione iniziale sarà quindi composta dai valori codificati di segni e cut-off selezionati casualmente. La fitness function è il modo in cui si cerca di misurare la performance di ogni individuo della popolazione, potendoli poi classificare al fine di poter selezionare i risultati migliori. È necessario che quindi si trasformino le regole che vengono prodotte dall’algoritmo genetico in un valore numerico basato sulla performance della regola stessa. Dato che l’obiettivo del modello è trovare una regola che, nel momento in cui venga applicate a tutte le imprese del campione, fornisca il maggior rapporto di successo della previsione con la manifestazione reale, viene definita la fitness function proprio come il rapporto di successo che deve essere massimizzato. Più semplicemente, ogni stringa dell’algoritmo rappresenta una regola, le regole uscenti da una generazione vengono applicate a tutto il campione di addestramento dell’algoritmo contenente i financial ratio delle imprese. Per ogni regola viene così stabilito il rapporto di successo della previsione e stilata la classifica in modo da poter proseguire con l’algoritmo.

A questo punto sarà possibile applicare gli operatori reproduction, crossover e mutation. I parametri degli operatori vengono posti come range il cui valore applicato a ogni generazione è scelto casualmente. Il range per la probabilità di effettuare il crossover è compreso tra 0.5 e 0.7 mentre la probabilità di mutazione tra 0.06 e 0.12. Per decidere quando interrompere l’algoritmo ci sono tre metodi principalmente utilizzati:

- Il primo prevede di imporre un limite fisso massimo al numero di generazioni calcolate dall’algoritmo. Per utilizzare questo metodo bisogna tenere in considerazione che più è grande la popolazione iniziale e complesso il problema,

maggiore sarà il numero di generazioni necessarie a far convergere la soluzione. Questo metodo viene utilizzato in questa applicazione fissando un limite a 3000 generazioni;

- Il secondo criterio consiste nel misurare la convergenza e stoppare il modello quando effettivamente smette di migliorare. È possibile osservando per quante generazioni consecutive la soluzione non muta o muta infinitesimamente;
- Infine, il metodo ibrido così come visto nell'applicazione precedente, dove viene fissato un limite massimo di generazioni ma al contempo l'algoritmo viene interrotto precedentemente se la misurazione della convergenza suggerisce che questa sia già avvenuta.

Lo scopo dell'ottimizzazione è trovare una soluzione che sia la migliore, l'esito dell'algoritmo genetico è una popolazione di regole dal quale tramite la fitness function si ricava quella più performante. In accordo con (Shin & Lee, 2002)[37], nei problemi che si incontrano realmente una sola soluzione non è spesso sufficiente. Si presta bene in questo caso il modello con algoritmo genetico in quanto permette di poter selezionare diverse regole tra le più performanti dalla popolazione di convergenza e per questa applicazione ne sono state selezionate 6. Le condizioni che seguono lo schema logico precedentemente esposto sono:

| Ratio | ROI | ROE | Ebitda/Turnover | Interest charges/Ebitda | Cash flow/tot debts | Finanzial debt/equity | Current ratio | Total debts/Ebitda |
|---|---|---|---|---|---|---|---|---|
| sign | < | < | < | >= | < | >= | < | >= |
| Rule 1 | 0.432 | - | 0.751 | 0.534 | 0.155 | 0.595 | 0.565 | - |
| Rule 2 | - | 0.56 | 0.69 | - | 0.15 | 0.697 | 0.52 | 0.575 |
| Rule 3 | 0.432 | - | - | 0.697 | 0.185 | 0.56 | 0.515 | 0.59 |
| Rule 4 | 0.525 | 0.535 | - | - | 0.182 | 0.577 | 0.52 | 0.595 |
| Rule 5 | 0.432 | - | 0.73 | 0.585 | 0.125 | 0.59 | - | 0.59 |
| Rule 6 | 0.69 | 0.615 | - | 0.58 | 0.13 | - | 0.503 | 0.62 |

**Tabella 5.3 Regole più performanti**

[37] Shin, K. S., & Lee, Y. J. (2002). Genetic algorithm application in bankruptcy prediction modeling. Expert System with Applications, 23, 321–328

La tabella 5.4 mostra i risultati ottenuti dall'applicazione del modello con i dati delle imprese, contenute nel campione di test, 3 anni prima l'evento default. È rappresentata una matrice di confusione con i valori percentuali e l'accuracy rate per tutti i modelli messi a confronto.

| | | | Confusion Matrix % | | |
|---|---|---|---|---|---|
| Modello | Realtà | | Predizione | | Accuracy |
| | | | 1 | 0 | |
| AG | Default | 1 | 78.8 | 21.2 | 71.5 |
| | Sana | 0 | 35.8 | 64.2 | |
| SVM | Default | 1 | 77.1 | 22.9 | 69.5 |
| | Sana | 0 | 38.1 | 64.2 | |
| LR | Default | 1 | 76.7 | 23.3 | 66.8 |
| | Sana | 0 | 43.1 | 56.9 | |

**Tabella 5.4 Confusion Matrix e Accuracy Rate a 3 anni dal default**

Evidenziato in blu vi è la percentuale di imprese anomale correttamente predetto dal modello sul totale delle imprese anomali presenti nel campione, è quindi la sensibilità del modello che si attesta al 78.8%. Mentre evidenziato in verde è la specificità ovvero le imprese sane nel campione correttamente classificate dal modello.

Il complemento a 1 di questi valori sono le classificazioni errate. Sapendo che il costo per l'errore di prima specie è maggiore, assume notevole importanza il valore evidenziato in rosso che rappresenta la percentuale di imprese classificate sane che in realtà si sono dimostrate anomale. Nell'ultima colonna è invece rappresentato il valore di Accuracy dei modelli, dal confronto di questi valori si può desumere che il modello con algoritmi genetici ha sovraperformato le altre metodologie ottenendo addirittura un distacco di due punti percentuali dal modello SVM. Si dimostra vincente anche per quanto riguarda una maggiore sensibilità e soprattutto una minor valore di errore di prima specie.

La stessa tabella e riportata di seguito i cui valori derivano dall'utilizzo dei dati a due (tabella 5.5) ed a un anno (tabella 5.6) di distanza dal default.

| | | | Confusion Matrix % | | |
|---|---|---|---|---|---|
| Modello | Realtà | | Predizione | | Accuracy |
| | | | 1 | 0 | |
| AG | Default | 1 | 79.2 | 20.8 | 73.9 |
| | Sana | 0 | 31.4 | 68.6 | |
| SVM | Default | 1 | 78.3 | 21.7 | 72.1 |
| | Sana | 0 | 34.1 | 65.9 | |
| LR | Default | 1 | 77.6 | 22.4 | 67.9 |
| | Sana | 0 | 41.8 | 58.2 | |

**Tabella 5.5 Confusion Matrix e Accuracy Rate a 2 anni dal default**

| | | | Confusion Matrix % | | |
|---|---|---|---|---|---|
| Modello | Realtà | | Predizione | | Accuracy |
| | | | 1 | 0 | |
| AG | Default | 1 | 79.6 | 20.4 | 74.5 |
| | Sana | 0 | 30.5 | 69.5 | |
| SVM | Default | 1 | 78.7 | 21.3 | 73.3 |
| | Sana | 0 | 32.1 | 67.9 | |
| LR | Default | 1 | 78.3 | 21.7 | 69.1 |
| | Sana | 0 | 40.2 | 59.8 | |

**Tabella 5.6 Confusion Matrix e Accuracy Rate a 1 anni dal default**

I risultati mostrano come l'accuratezza di tutti i modelli migliori progressivamente con il ridursi della finestra temporale utilizzata prima dell'evento default e che il modello con algoritmi genetici continua ad avere una maggiore efficacia rispetto alle altre due metodologie in tutte le misure di performance.

Infine, è necessario testare se le differenze tra i modelli sono statisticamente significative. Per farlo si possono utilizzare diverse tecniche statistiche come, per esempio, il t-Test confronta i modelli presi a coppie, l'ipotesi nulla è che non c'è differenza tra i modelli e può essere svolta con i soliti livelli di confidenza 1%, 5% e 10%. Essendo che questo test assume che i dati analizzati seguano una distribuzione normale è buona norma effettuare precedentemente un test sulla effettiva normalità dei dati, utilizzando per esempio il Shapiro-Wilk normality test.

In questo caso è stato applicato il McNemar test che testa l'ipotesi nulla che due variabili dicotomiche messe a confronto abbiano la stessa media. Nella tabella seguente sono rappresentati i valori del p-value del test per il confronto a coppie dei modelli. Si riscontra che per i risultati ripostati il modello AG ha performance migliori di quello SVM utilizzando un livello di significatività del 10%, mentre sembra sovraperformare il modello LR con un livello di significatività dell'1%. Per quanto riguarda il modello SVM confrontato con il modello LR, le performance del primo non superato il test di significatività e quindi non è possibile affermare che sia più efficacie nella classificazione.

| | SVM | AG |
|---|---|---|
| LR | 0.116 | 0.004** |
| SVM | | 0.085* |

* Livello di significatività al l10%

** Livello di significatività al 5%

**Tabella 5.5 McNemar test**

# *CONCLUSIONI*

Recenti ricerche hanno approfondito tutte le tecniche di machine learning applicabili ai modelli di credit scoring. I metodi attualmente più accreditati sono quelli ibridi in cui spesso gli algoritmi genetici assolvono ad un ruolo di supporto per altre tecniche, in particolar modo per le reti neurali. Questo trend è dovuto alle migliori performance ottenute dai ricercatori utilizzando tecniche ANNs rispetto ad altri metodi. Tuttavia, come affermato nell'elaborato, è da tener presente che le performance di un modello sono soggette alla capacità di impostazione del problema da parte del ricercatore e alcuni metodi non riescono a fornire regole comprensibili a coloro che li utilizzano nel quotidiano. Una corretta scelta della fitness function, dei parametri degli operatori e dei metodi di codifica dei dati da elaborare, possono incrementare parecchio i risultati di un modello AG in cui la tecnica è usata standalone.

Principale limite della ricerca nell'evolvere questi modelli è la scarsità di dati disponibili sulle controparti e la capacità computazionale necessaria all'analisi con algoritmi dalla complessità elevata. La complessità computazionale è, inoltre, notevolmente amplificata quando i dati non sono numerici ma magari informazioni qualitative non classificabili in macroclassi o non facilmente codificabili, come per esempio un indirizzo o il luogo di nascita.

Raggiungere sempre performance migliori per questi modelli è importante, l'errore di prima specie è deleterio per l'impresa creditrice o, nel momento in cui questa può essere considerata sistemica, per l'intera rete finanziaria. L'errore di seconda specie, l'errore di classificare una controparte come anomala quando questa invece è sana, seppur normalmente considerato meno rilevante, lo diventa per quanto riguarda le controparti che si vedono affidato un giudizio negativo nelle banche dati. Questo giudizio negativo oltre a creare un danno reputazionale, rende difficile per il soggetto richiedere finanziamenti al resto del sistema e macroeconomicamente conduce all'effetto definito 'Credit crunch'. È

quindi necessario continuare a svolgere ricerca in questo ambito per perfezionare le tecniche o aggiungerne di altre.

# BIBLIOGRAFIA


Dos Santos, E. M., Sabourin, R., & Maupin, P. (2009). Overfitting cautious selection of classifier ensembles with genetic algorithms. *Information Fusion*, *10*(2), 150–162. https://doi.org/10.1016/J.INFFUS.2008.11.003

Goldberg, D. E. (1989). Genetic Algorithms in Search, Optimization, and Machine Learning. *Addison-Wesley Publishing Company, Inc*.

Gordini, N. (n.d.). *A genetic algorithm approach for SMEs bankruptcy prediction: Empirical evidence from Italy*. https://doi.org/10.1016/j.eswa.2014.04.026

Kozeny, V. (2015). Genetic algorithms for credit scoring: Alternative fitness function performance comparison. *Expert Systems with Applications*, *42*(6), 2998–3004. https://doi.org/10.1016/J.ESWA.2014.11.028

Oreski, S., Oreski, D., & Oreski, G. (2012). Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. *Expert Systems with Applications*, *39*(16), 12605–12617. https://doi.org/10.1016/J.ESWA.2012.05.023

Shin, K. S., & Lee, Y. J. (2002). A genetic algorithm application in bankruptcy prediction modeling. *Expert Systems with Applications*, *23*(3), 321–328. https://doi.org/10.1016/S0957-4174(02)00051-9

Šušteršič, M., Mramor, D., & Zupan, J. (n.d.). Consumer credit scoring models with limited data. *Expert Systems With Applications*, *36*, 4736–4744. https://doi.org/10.1016/j.eswa.2008.06.016

Varetto, F. (1998). Genetic algorithms applications in the analysis of insolvency risk. *Journal of Banking & Finance*, *22*, 1421–1439.

Banca d'Italia. (2007). Disposizioni di vigilanza in materia di conformità.

Holland, J. (1975). Adapration in natural and artificial system. *University of Michigan Press*.

Montgomery, D., & Peck, E. (1992). *Introduction to linear regression analysis.* New York, NY: John Wiley and Sons.

Talen, N. (2014). *Il cigno nero. Come l'improbabile governa la nostra vita.* Il saggiatore.

Varetto, corso di "Mercati Rischi e Strumenti Finanziari", A.A. 2020/2021, Politecnico di Torino.

# *SITOGRAFIA*


https://www.unocloudbackup.it/differenza-tra-disaster-recovery-plan-e-business-continuity-plan/

https://www.filodiritto.com/le-frodi-aziendali-la-prevenzione-delle-frodi

https://www.experian.it/business/identita-e-frode/prevenzione-frodi/scipafi

https://www.investopedia.com/

https://www.bancaditalia.it/media/fact/2020/definizione-default/index.html

https://en.wikipedia.org/wiki/Cross-validation_(statistics)

https://www.med4.care/curva-roc-receiver-operating-characteristic-introduzione-e-applicazione-ai-test-diagnostici/