UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI TECNICA E GESTIONE DEI SISTEMI INDUSTRIALI

Corso di Laurea Triennale in Ingegneria Gestionale

TESI DI LAUREA

METODI STATISTICI PER IL CONTROLLO QUALITÀ

Relatore: Prof.ssa Marta Disegna

Dipartimento di Tecnica e Gestione dei sistemi industriali

Laureanda: Sara Contarin

Matricola: 2004137

ANNO ACCADEMICO 2022/2023

Abstract

Il tema del controllo della qualità è stato oggetto, nel corso degli anni, di un crescente interesse in ambito aziendale. Il cambiamento delle richieste del mercato e l'incremento delle pretese qualitative dei consumatori hanno reso necessaria l'introduzione nelle aziende di metodi statistici per il controllo della qualità. Tra questi ha riscosso maggior successo il Six Sigma, articolato nelle fasi di definizione, misurazione, analisi, miglioramento e controllo. Esso viene implementato attraverso l'utilizzo di strumenti statistici e si occupa dello studio della qualità di processo, prodotti e servizi.

Negli ultimi decenni questi metodi sono stati rivoluzionati per potersi adattare alle enormi quantità di dati da gestire. Si è iniziato a parlare di Industria 4.0, un'impostazione aziendale che prevede l'automazione dei processi e l'utilizzo delle nuove tecnologie per una continua connessione tra persone, macchine e prodotti durante la produzione. L'introduzione di pianificazione e controllo intelligente della produzione permette di aumentare la produttività degli impianti e la qualità dei prodotti tramite strumenti innovativi, quali il Machine Learning e le tecnologie Blockchain. In tale contesto, l'Internet of Things ha assunto un ruolo rilevante, tanto che si è iniziato a parlare di Industrial Internet of Things.

Il presente lavoro di tesi include una revisione della letteratura sui metodi statistici per il controllo della qualità e a seguire un confronto tra diversi casi studio nei quali sono stati applicati gli strumenti sopra citati. Lo scopo è l'analisi delle differenze tra i metodi presentati, eventuali analogie e risultati a cui conducono gli stessi. Tutti i metodi descritti ed applicati, sia quelli tradizionali che quelli di recente introduzione, sono di fondamentale supporto ai responsabili di qualità e produzione nel processo decisionale.

Indice

Introduzione

La parola “qualità” non può essere definita in modo semplice ed univoco, poiché assume diversi significati a seconda del punto di vista attraverso il quale la si considera. Può, in generale, essere intesa come l’insieme di caratteristiche che un prodotto deve possedere per essere ritenuto desiderabile dagli acquirenti e per garantire la soddisfazione delle loro richieste esplicite o implicite. Il livello di qualità, però, non si riferisce unicamente al prodotto, bensì anche al processo produttivo, il quale deve assicurare la creazione di prodotti conformi alle specifiche tecniche predefinite. Un’eventuale presenza di difetti è solitamente imputata alla variabilità del processo ed in tal caso l’azienda deve intervenire con l’obiettivo di migliorarne la capacità e ridurre eventuali inefficienze [5] [10] [31].

Nei primi anni del Novecento negli Stati Uniti si diffuse la produzione di massa, la cui logica era quella di creare grandi quantità di prodotti con varietà molto bassa. Un noto esempio è il caso dell’industria automobilistica Ford, la quale, nel 1913, iniziò a produrre in elevate quantità il modello standardizzato Ford T. Tuttavia, la produzione di massa richiedeva di accelerare il processo e quindi, per limitare le sue interruzioni, fu introdotto nel 1924 da W. Shewhart il processo di controllo statistico (SPC). Esso permetteva di disporre dell’utilizzo di grafici per monitorare la qualità in ambito industriale e di interrompere il processo produttivo soltanto in presenza di evidenti anomalie, statisticamente provate [2].

Negli anni successivi furono implementate ulteriori tecniche per il controllo e il miglioramento della qualità e venne gradualmente introdotto il controllo di qualità riferito non soltanto al prodotto, ma anche al processo e all’intero sistema, fino ad arrivare allo sviluppo del Total Quality Management (TQM), il cui principale promotore fu Edward Deming. Le nuove tecniche ideate da quest’ultimo trovarono immediata applicazione nelle aziende giapponesi, i cui prodotti, dopo la Seconda Guerra Mondiale, riscossero grande successo nel mercato proprio grazie al loro elevato livello di qualità. A seguito della competitività che avevano generato, le nuove metodologie si diffusero anche negli Stati Uniti e successivamente nel resto del mondo [2] [33].

Tra gli anni ’80 e ’90 questi metodi iniziarono ad essere visti come un ostacolo all’attività aziendale, in quanto lenti e con effetti non immediati, inoltre il loro utilizzo spesso richiedeva la presenza di esperti di statistica all’interno dell’azienda. In questo contesto trovò terreno fertile il metodo del Six Sigma, adottato per la prima volta da Motorola nel 1987 [2]. Venne stabilito un obiettivo di 3.4 DPMO (defects per million opportunities), ovvero il 99.99966% dei pezzi prodotti doveva essere privo di difetti, percentuale che corrisponde a sei volte la deviazione standard, da cui il nome “Six Sigma” [5] [12]. A metà degli anni ’90 le aziende hanno dovuto focalizzare sempre di più la loro attenzione sul mantenimento di elevati standard di qualità di prodotti e servizi, per poter incrementare la propria competitività nel mercato. Questa necessità è cresciuta a seguito dello sviluppo della globalizzazione e del conseguente aumento delle pretese qualitative

da parte dei consumatori [27]. Per questo motivo il Six Sigma fu implementato in altre aziende, tra cui AlliedSignal, General Electric, Sony e Honeywell e la sua diffusione fu crescente nel corso degli anni seguenti. Un grande vantaggio di questo metodo è che può essere applicato in tutte le aziende, senza la necessità di supervisione da parte di statistici [2] [12].

Il recente sviluppo delle nuove tecnologie ha trovato applicazione anche nel settore industriale. Nel 1970 fu sperimentato per la prima volta l'utilizzo di microprocessori per il controllo, per poi arrivare, nel 1990, alla progressiva digitalizzazione dei processi industriali [14]. Nel corso degli ultimi anni, l'ambiente manifatturiero è diventato sempre più complesso, dinamico e volatile. I volumi, la velocità e la varietà di informazioni avanzate, che prendono il nome di Big Data, hanno portato alla nascita nel 2011 in Germania dell'Industria 4.0, la quale consente di connettere in modo intelligente persone, macchinari e prodotti durante il processo produttivo. Alcuni Paesi hanno creato nuovi progetti per permettere alle imprese di stare al passo con le nuove tecnologie, ad esempio gli Stati Uniti hanno ideato la "Smart Manufacturing Leadership Coalition" e la Cina il "China Manufacturing 2025" [1] [25] [32]. Si è passati così da pianificazione e controllo della produzione tradizionale allo Smart PPC (Smart Production Planning and Control), con l'introduzione di processi produttivi automatizzati [18]. L'Industrial Internet of Things (IIoT) permette di trasmettere tutte le informazioni in modo immediato all'interno dell'azienda, attraverso l'utilizzo di Internet. Questo nuovo approccio consente di prendere decisioni più velocemente ed eseguire indagini più specifiche e approfondite. Nel campo dell'intelligenza artificiale si distinguono diversi metodi per il controllo della qualità, attualmente in continua sperimentazione ed evoluzione, tra cui il machine learning e la tecnologia blockchain [11].

Il seguente lavoro di tesi è costituito da una prima parte che spiega a livello teorico l'applicazione del metodo tradizionale del Six Sigma, un secondo capitolo che introduce l'utilizzo delle nuove tecnologie nell'ambito del controllo della qualità e una parte finale in cui viene eseguita la manutenzione predittiva di una fresatrice attraverso l'utilizzo di algoritmi di machine learning, a cui segue un confronto tra casi studio con l'applicazione di diversi algoritmi.

CAPITOLO 1

Six Sigma

1.1. Definizione del Six Sigma

La definizione di Six Sigma, così come quella di qualità, non è chiara e specifica. Una descrizione accurata del metodo è quella fornita da Linderman et al. (2003), secondo cui il Six Sigma è una tecnica sistematica ed organizzata per il miglioramento del processo strategico e per lo sviluppo di nuovi prodotti e servizi, che si basa su metodi scientifici e statistici per effettuare riduzioni considerevoli nei tassi di difettosità definiti dal cliente. Questo metodo può essere analizzato da due punti vista: quello commerciale, poiché permette la riduzione del tasso di difettosità e l'aumento della soddisfazione del cliente, e quello statistico, poiché riduce la variabilità del processo. Esso, infatti, permette di stabilire gli obiettivi da raggiungere basandosi sulle richieste dei clienti, anziché su considerazioni interne all'azienda, ed integra l'utilizzo di diversi strumenti statistici [17] [28].

1.2. Implementazione del metodo

Il Six Sigma si sviluppa secondo l'approccio DMAIC, ovvero si suddivide in cinque fasi, dalle cui iniziali deriva il nome:

1. Definizione (define): definisce gli obiettivi e i limiti del progetto sulla base delle richieste e delle aspettative dei clienti;
2. Misurazione (measure): misura la capacità del processo e delle tecnologie presenti nel sistema prima che vengano effettuati cambiamenti;
3. Analisi (analyze): analizza le cause dei difetti, la variabilità del processo e le fonti di variazione;
4. Miglioramento (improve): migliora il processo sulla base di informazioni ottenute ai punti precedenti, eliminando variazioni e implementando piani più avanzati;
5. Controllo (control): controlla le variazioni del processo.

Questo approccio permette di eseguire operazioni in una sequenza strutturata e sistematica, soprattutto quando il problema è di difficile individuazione. Ognuna di queste fasi può essere eseguita utilizzando diversi strumenti, alcuni dei quali sono descritti di seguito [2] [28].

1.2.1. Fase di definizione

La prima fase del Six Sigma è stata aggiunta in un secondo momento dalla società General Electric per identificare il problema a monte dell'intero processo di controllo e operare con maggior coerenza rispetto all'obiettivo da raggiungere [28]. Questa fase permette di definire il progetto per diminuire i costi derivanti dalla non qualità [12].

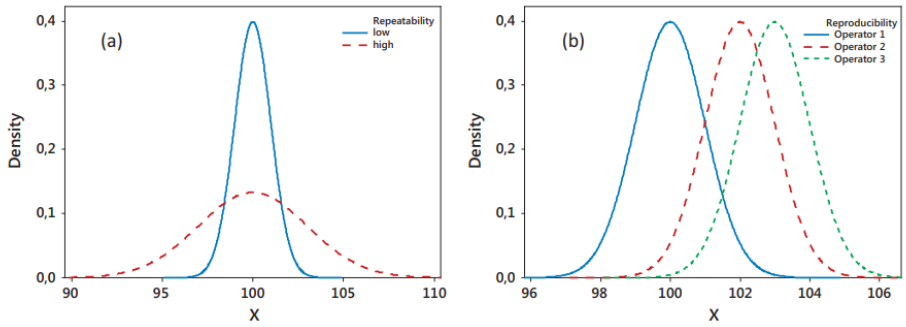
Gli strumenti più utilizzati per eseguire lo studio sono l'analisi comparativa, i diagrammi ad albero CTQ (Critical to Quality), il metodo "5 Whys", l'analisi SWOT e il diagramma di Pareto.
L'analisi comparativa, o benchmarking, consiste nel confrontare i propri prodotti con quelli della concorrenza, per avere una visione più ampia del mercato e adattarsi alle sue necessità [2].
Una CTQ, invece, è una variabile che influenza direttamente la qualità di un processo ed è ritenuta particolarmente importante per soddisfare i bisogni presenti e futuri dei clienti, oltre che per generare profitto per l'azienda e valore per gli azionisti. Il diagramma ad albero CTQ serve per capire su quali variabili è necessario intervenire per ottenere un miglioramento del processo produttivo [28] [29] [31].
Un ulteriore strumento è il "5 Whys", che permette di porsi domande sull'argomento che si sta studiando, con lo scopo di definire la provenienza delle difettosità dei processi e dei costi elevati, mentre l'analisi SWOT aiuta a capire quali sono i punti di forza, di debolezza, le opportunità e le minacce dello studio [31].
Il diagramma di Pareto permette di avere una visione completa degli attributi che possono generare difetti durante il processo produttivo, con la rispettiva frequenza di rilevazione. Questo consente di prendere decisioni più velocemente e in modo più accurato, basandosi non soltanto sulle ultime cause di non conformità, ma soprattutto su quelle più rilevanti. Secondo la "regola di Pareto", infatti, al 20% delle cause è associato più dell'80% delle non conformità [2].

1.2.2. Fase di misurazione

Spesso la non conformità di un prodotto alle specifiche tecniche predefinite viene imputata alla scarsa capacità del processo. Tuttavia, alcuni errori potrebbero essere dovuti all'inadeguatezza dei sistemi di misura, la quale può portare ad assumere decisioni errate [3] [21]. Pertanto, prima di intraprendere azioni di miglioramento è necessario controllare che il processo sia stabile e sotto controllo statistico [34]. Eventuali errori possono essere individuati e ridotti in questa fase del Six Sigma, con l'utilizzo del metodo di ripetibilità e riproducibilità di Gauge (Gauge R&R), il quale racchiude un insieme di strumenti per la valutazione dei sistemi di misura [2]. Il Gauge R&R è uno studio della variabilità delle misurazioni derivante da ripetibilità e riproducibilità. Si definisce ripetibilità la variazione osservata quando un operatore esegue una misurazione con lo stesso strumento nello stesso pezzo per più di una volta. La riproducibilità, invece, è la variazione generata quando diversi operatori eseguono una misurazione sullo stesso pezzo con lo stesso strumento [34]. La rappresentazione grafica di ripetibilità e riproducibilità è riportata in figura 1.1.
Il metodo di Gauge R&R è strutturato da una sequenza non rigida di fasi. Numerosi autori hanno individuato diversi indici che possono essere utilizzati nella procedura di analisi e valutazione del sistema di misura, oltre che della capacità del processo, a seconda delle esigenze dello studio che deve essere effettuato [24].

Innanzitutto, è necessario determinare il numero di operatori (o), il numero di parti (p) e quello delle ripetizioni (r). L'intervallo di confidenza dipende da alcuni parametri, quali la varianza degli operatori (σ_O^2), la varianza delle parti (σ_P^2), la varianza derivante dall'interazione tra operatori e parti (σ_{OP}^2) e quella delle ripetizioni (σ_R^2) [23].

Figura 1.1: (a) ripetibilità e (b) riproducibilità in un sistema di misura [24]



Per calcolare le varianze si conduce un test ANOVA a una o a due vie. Si imposta un'equazione di questo tipo:

$$y_{ijk} = x + P_i + O_j + (OP)_{ij} + \varepsilon_{ijk}$$

∀ i=1,…,p;

∀ j=1,…,o;

∀ k=1,…,r;

dove y_{ijk} rappresenta la k-esima misurazione eseguita dal j-esimo operatore sulla i-esima parte, x è il valore vero della misurazione, mentre $P_i, O_j, OP_{ij}, \varepsilon_{ijk}$ sono variabili casuali indipendenti che rappresentano rispettivamente le parti, gli operatori, l'interazione tra essi e l'errore casuale.
Il test ANOVA ha come ipotesi nulla $H_0: \sigma_{OP}^2 = 0$. Se questa ipotesi viene rifiutata, si ottengono i risultati mostrati in figura 1.2 [21] [23].

Figura 1.2: tabella per uno studio AGRR standard [3]

Source of variability	Sum of squares	Degrees of freedom	Mean squares	Expected mean squares
Part	SS_P	$d_P = p - 1$	$MS_P = \frac{SS_P}{d_P}$	$\theta_P = \sigma_R^2 + r\sigma_{OP}^2 + or\sigma_P^2$
Operator	SS_O	$d_O = o - 1$	$MS_O = \frac{SS_O}{d_O}$	$\theta_O = \sigma_R^2 + r\sigma_{OP}^2 + pr\sigma_O^2$
Operator × part	SS_{OP}	$d_{OP} = (o-1)(p-1)$	$MS_{OP} = SS_{OP}/d_{OP}$	$\theta_{OP} = \sigma_R^2 + r\sigma_{OP}^2$
Error	SS_R	$d_R = op(r-1)$	$MS_R = \frac{SS_R}{d_R}$	$\theta_R = \sigma_R^2$
Total	SS_T	$opr - 1$		

A questo punto, si possono calcolare i valori delle varianze nel seguente modo:

$$\sigma_P^2 = \frac{MS_P - MS_{OP}}{or}$$

$$\sigma_O^2 = \frac{MS_O - MS_{OP}}{pr}$$

$$\sigma_{OP}^2 = \frac{MS_{OP} - MS_R}{r}$$
$$\sigma_R^2 = MS_R$$

La varianza riferita alla ripetibilità è σ_R^2, mentre quella riferita alla riproducibilità è data dalla somma di σ_O^2 e σ_{OP}^2. Dunque, la varianza di Gauge si calcola come somma delle tre, ovvero $\sigma_R^2 + \sigma_O^2 + \sigma_{OP}^2 = \sigma_{R\&R}^2$.

Stimati i valori delle varianze, per verificare la capacità del sistema di misura possono essere utilizzate diverse misure di qualità, le più importati sono descritte di seguito.

- Precision to tolerance ratio, $PTR = \frac{6\sigma_{R\&R}}{USL-LSL}$

  dove USL e LSL sono rispettivamente i limiti superiore e inferiore della tolleranza.

  Secondo questo criterio se $PTR \leq 0{,}1$ il sistema di misura è capace, mentre se $PTR > 0{,}3$ non è capace.

- Signal-to-noise ratio, $SNR = \frac{\sqrt{2}\sigma_P}{\sigma_{R\&R}}$

  Secondo questo criterio se $SNR \geq 5$ il sistema è capace, mentre se $SNR < 2$ non è capace.

- Discrimination ratio, $DR = \sqrt{\frac{2\sigma_P^2}{\sigma_{R\&R}^2} + 1}$

  Secondo questo criterio se $DR \geq 4$ il sistema è capace, se $DR < 2$ il sistema non è capace.

In tutti e tre i casi precedenti, se il valore misurato ricade all'interno dell'intervallo compreso tra i valori limite è necessario effettuare ulteriori misurazioni. Generalmente il PTR fornisce risultati insufficienti per valutare il sistema di misura; pertanto, deve essere utilizzato in concomitanza con SNR o DR, i quali forniscono risultati simili [3] [23].

Per ottenere un risultato più preciso spesso vengono utilizzati ulteriori indici detti indici di capacità, ovvero il C_p e il C_{pk}.

Il C_p è dato dal rapporto tra l'estensione della specifica e quella del processo:

$$C_p = \frac{USL - LSL}{6\sigma_P}$$

Questo indice permette di attestare quanto bene la distribuzione del processo si adatta ai limiti di specifica del prodotto. Affinché il processo sia capace, il valore di C_p dovrebbe essere almeno di 1,67 e deve valere:

$$\sigma_P \leq \frac{USL - LSL}{6C_p}$$

Tuttavia, il C_p non considera la centratura del processo rispetto ai limiti di specifica e questo potrebbe portare alla creazione di difetti. È utile, quindi, calcolare anche un ulteriore indice, ovvero C_{pk}, che tiene conto della centratura e si calcola come valore minimo tra capacità del processo di incontrare il limite di specifica superiore (C_{pu}) e capacità del processo di incontrare il limite di specifica inferiore (C_{pl}), dove:

$$C_{pu} = \frac{USL - \bar{x}}{3\sigma_P}$$

$$C_{pl} = \frac{\bar{x} - LSL}{3\sigma_P}$$

$$C_{pk} = \min\left\{\frac{USL - \bar{x}}{3\sigma_P}, \frac{\bar{x} - LSL}{3\sigma_P}\right\}$$

con $\bar{x}$ la media del processo.

Un processo manifatturiero è ritenuto capace se C_{pk} è almeno pari a 1,33 e deve valere:

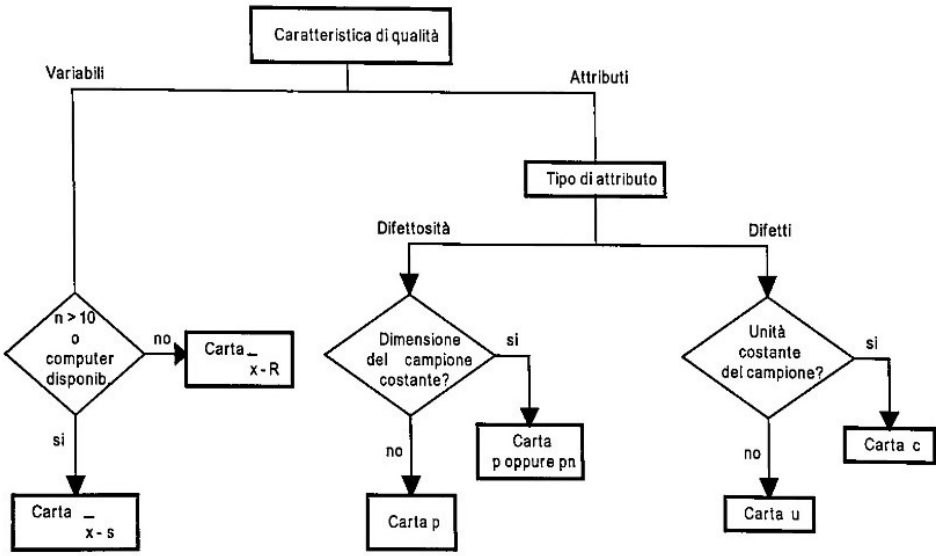
$$\sigma_P \leq \frac{\min\{USL - \bar{x}, \bar{x} - LSL\}}{3C_{pk}}$$

Per avere una visione completa della capacità del processo, questi due indici dovrebbero essere usati in concomitanza. Questo permette di intuire facilmente se il superamento dei limiti di specifica è determinato dall'aumento della dispersione o dallo spostamento verso uno dei due limiti. In particolare, se il processo è centrato, C_p e C_{pk} coincidono [3] [10].

Una volta che il processo è stato dichiarato "capace", o almeno accettabile, si possono utilizzare le misure rilevate per costruire una rappresentazione grafica delle rilevazioni, tramite l'utilizzo di carte di controllo. Queste ultime permettono di controllare il processo e segnalare eventuali anomalie attraverso il campionamento di alcuni pezzi estratti dal processo produttivo in maniera casuale [2] [10].

Le carte di controllo possono essere di diverso tipo a seconda del tipo di dati e della finalità d'uso e la scelta di applicazione di una piuttosto che dell'altra avviene secondo lo schema in figura 1.3.

Figura 1.3: Scelta carta di controllo [10]



La carta più utilizzata è quella $\bar{x} - R$, ovvero una carta di controllo per variabili, utilizzata per variabili continue. Questa carta si può creare attraverso una sequenza di passaggi semplici.

1. Si raccolgono almeno 100 dati da suddividere in 20 o 25 sottogruppi;
2. si calcola $\bar{x}$ per ogni sottogruppo;
3. si calcola la media delle medie, ovvero $\bar{\bar{x}}$;
4. si calcola l'escursione R per ogni sottogruppo, ovvero la differenza tra valore massimo e valore minimo nel sottogruppo;
5. si calcola la media degli R, ovvero $\bar{R}$;
6. si calcolano i valori delle linee limite di controllo, sia per la carta $\bar{x}$:

$$LSC = \bar{\bar{x}} + A_2\bar{R}$$
$$LC = \bar{\bar{x}}$$
$$LSC = \bar{\bar{x}} - A_2\bar{R}$$

che per la carta $\bar{R}$:

$$LSC = D_4\bar{R}$$
$$LC = \bar{R}$$
$$LSC = D_3\bar{R}$$

dove A_2, D_4 e D_3 sono valori che dipendono dalla numerosità del sottogruppo e sono riportati in figura 1.4;

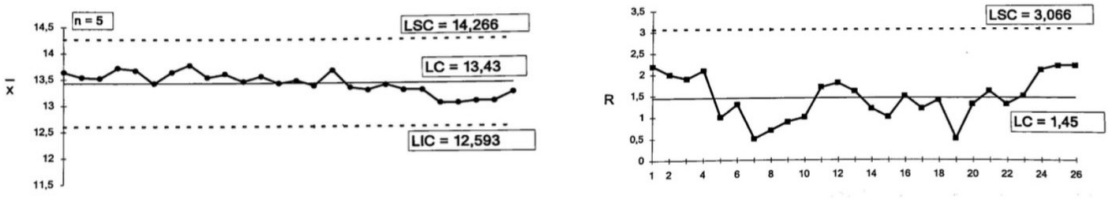
7. si disegnano i limiti di controllo trovati su un foglio di carta millimetrata;
8. si riportano sulla carta le informazioni rilevanti, come la numerosità del sottogruppo, il nome del processo, il periodo d'indagine, ecc.;
9. si riportano sulle rispettive carte i valori di $\bar{x}$ e R di ogni sottogruppo, rilevati nei punti precedenti.

Figura 1.4: tabella dei coefficienti [10]

Dimensione del campione	carta $\bar{x}$		Carta R	
n	A_1	A_2	D_3	D_4
2	3,760	1,880	-	3,267
3	2,394	1,023	-	2,575
4	1,880	0,729	-	2,282
5	1,595	0,577	-	2,115
6	1,410	0,483	-	2,004
7	1,277	0,419	0,076	1,924
8	1,175	0,373	0,136	1,864
9	1,094	0,337	0,184	1,816
10	1,028	0,308	0,223	1,777
11	0.973			
12	0,925			
13	0,884			
14	0,848			
15	0,816			
20	0,697			
25	0,619			

Rappresentati i dati, si otterrà una carta qualitativamente simile a quella in figura 1.5, che permette di verificare visivamente il loro andamento. Il processo può essere ritenuto sotto controllo se i valori rilevati non fuoriescono dai limiti e non mostrano particolari tendenze nei valori medi [2] [10].

Figura 1.5: esempio carta di controllo [10]



1.2.3. Fase di analisi

La fase di analisi si basa sul riconoscimento di una relazione causa-effetto tra gli input e gli output di un sistema. I metodi per eseguire l'analisi si basano su un'ampia gamma di dati e facilitano il processo decisionale. I più utilizzati sono il DOE (Design of Experiments) e la FMEA (Failure Mode and Effects Analysis).

Il DOE è un processo costituito da una serie di passaggi:

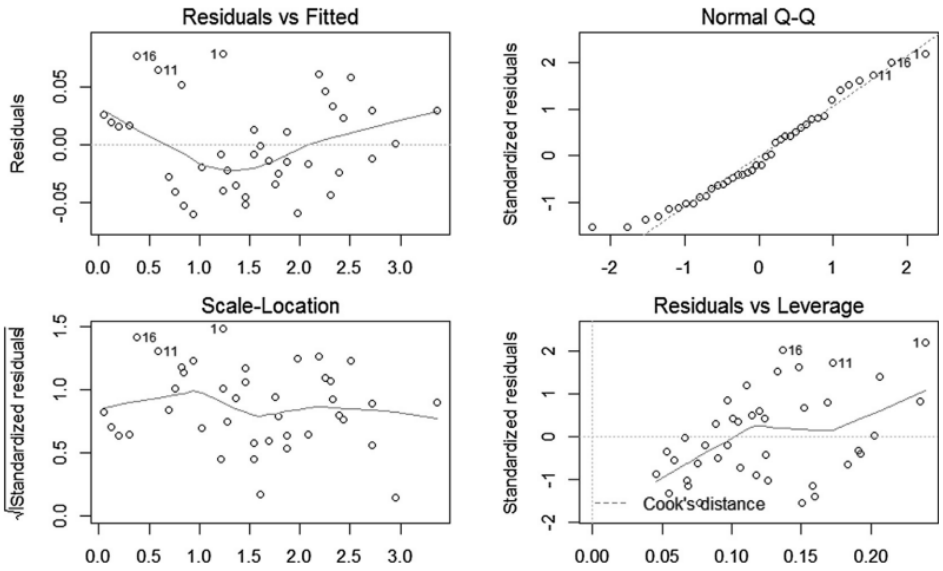
1. selezione degli input da testare;
2. esecuzione di test e registrazione degli output;
3. utilizzo di un metodo di interpolazione, come la regressione;
4. il modello viene utilizzato per prevedere nuovi output e creare nuove combinazioni di input.

I test che possono essere eseguiti sono molteplici, tra questi il t-test a due campioni e il test ANOVA. Un aspetto comune ad entrambi è l'identificazione delle variabili di input con i rispettivi range di accettabilità. Il t-test a due campioni è utile quando si vuole attestare, con un certo livello di confidenza, che una determinata alternativa è preferibile ad un'altra in termini di risposta media. Il test ANOVA, invece, permette di confrontare tra loro le medie dei dati, utilizzando le loro varianze.

All'esecuzione del test segue l'interpolazione dei dati, che può essere eseguita con un modello di regressione lineare. La regressione serve sia per individuare le cause che sono maggiormente correlate a determinati effetti, sia per effettuare previsioni future.

Prima di creare il modello è necessario verificare che i risultati che si otterranno si basino su un'analisi affidabile. Tale verifica può essere eseguita osservando i valori ottenuti dai fattori di inflazione della varianza (VIF), dal grafico dei residui e dalle statistiche riassuntive, dati rilevati dalla creazione del modello con appositi software. Il grafico dei residui, di cui un esempio in figura 1.6, permette di verificare l'eteroschedasticità, ovvero se i residui hanno o meno una deviazione standard costante [2].

Figura 1.6: Grafici dei residui [23]



Tra le statistiche riassuntive, la principale è il coefficiente R^2 aggiustato, il quale indica l'adeguatezza del modello che è stato creato. Tale coefficiente si calcola nel seguente modo: $R^2_{adj} = 1 - \frac{n-1}{n-k}\frac{SSE}{SST}$, dove SSE è la somma dei quadrati degli errori e SST è la somma dei quadrati totale.

Il metodo FMEA ha la funzione di supportare le decisioni volte a ridurre la probabilità che il processo possa fallire, individuare eventuali cause e valutare i possibili effetti dell'eventuale fallimento. La FMEA si suddivide nelle seguenti fasi:

1. si individuano le possibili cause di fallimento;
2. si ipotizzano gli effetti che ne potrebbero derivare;
3. si calcolano la severità del guasto (S), la probabilità di accadimento (P), la rilevabilità del guasto prima di conseguenze rilevanti (R);
4. si calcola il risk priority number: $RPN = S \cdot P \cdot R$;
5. si attribuisce ad ogni causa una priorità e si identificano delle azioni correttive [2].

1.2.4. Fase di miglioramento

La quarta fase del processo DMAIC consiste nel miglioramento del sistema già esistente. Gli obiettivi aziendali possono essere impostati come un problema di ricerca operativa, da risolvere ottimizzando un modello di programmazione lineare. Ad esempio, l'obiettivo potrebbe essere la massimizzazione del profitto, soggetta a vincoli di costo. La relazione tra le variabili può essere quella ottenuta tramite il modello di regressione lineare sviluppato nel DOE.

Generalmente, a questo punto del processo il numero di dati è sostanzioso, così come l'ammontare di opzioni; pertanto, è preferibile utilizzare strumenti tecnologici a supporto del processo

decisionale. È possibile impostare il problema in un foglio di lavoro Excel ed ottenere immediatamente la soluzione ottimale attraverso l'utilizzo del risolutore [2]. Un esempio di impostazione e risoluzione di un problema con l'utilizzo del risolutore è riportato in figura 1.7 e la sua rappresentazione grafica in figura 1.8 [20].

Può accadere che la soluzione ottimale sia data da una combinazione di fattori che dipendono da diverse aree aziendali e che, di conseguenza, sia necessario garantire un elevato grado di confronto e comunicazione tra esse. Inoltre, talvolta esistono vincoli più "informali", cioè difficilmente traducibili in formule matematiche, che potrebbero condurre ad una soluzione ottimale non attuabile. Pertanto, nonostante il supporto tecnologico sia fondamentale, è importante l'intervento umano per tradurre il risultato ottenuto nella realtà aziendale. L'utilizzo del risolutore Excel permette anche di verificare quanto è sensibile il risultato ottimale ad eventuali variazioni dei dati. Quindi, la soluzione può essere adeguata alle esigenze più informali, pur rimanendo il più vicino possibile a quella ottimale [2].

Figura 1.7: risoluzione di un problema su Excel

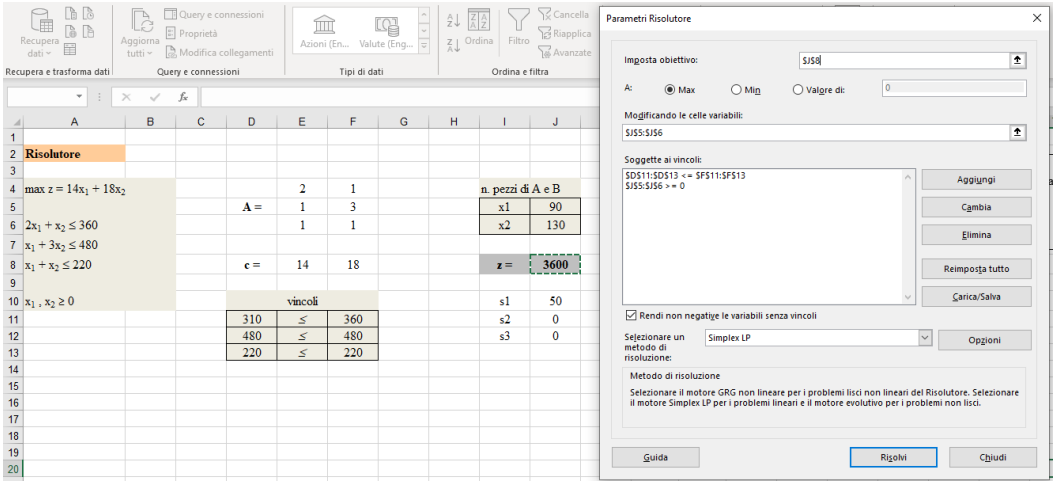
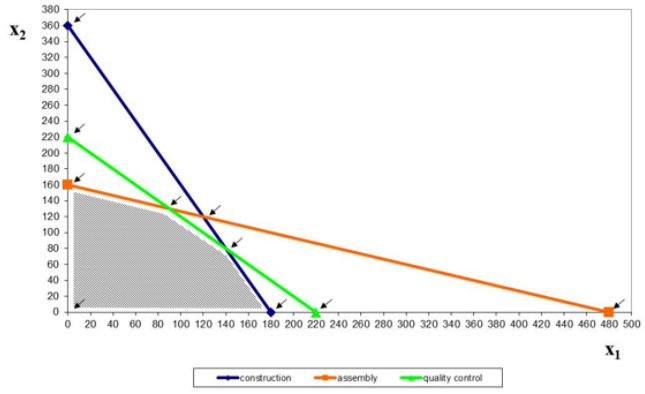


Figura 1.8: rappresentazione grafica del problema [20]



1.2.5. Fase di controllo

La fase finale del Six Sigma è quella del controllo ed ha lo scopo di assicurare che gli obiettivi raggiunti con l'applicazione del metodo siano mantenuti anche nel lungo termine [6]. I due metodi principali applicabili in questa fase sono la pianificazione del controllo e l'accettazione per campionamento.

La pianificazione del controllo consiste nel continuo monitoraggio dei dati attraverso l'utilizzo di grafici. Questo processo può includere molti degli strumenti utilizzati nelle fasi precedenti ed è implementato secondo il seguente algoritmo:

1. si individuano gli output più critici e li si ispeziona con l'utilizzo delle carte di controllo;
2. il metodo di Gauge R&R viene iterato fino a quando le caratteristiche critiche sono considerate accettabili;
3. per ogni caratteristica critica viene assegnato ad una o più persone un "piano di reazione" da attuare nel caso in cui gli output siano fuori controllo;
4. vengono definiti la dimensione del campione e il periodo di ispezione;
5. per ogni caratteristica critica viene impostato un grafico specifico;
6. se il grafico è ritenuto inaccettabile vengono aggiustati la dimensione del campione e il periodo di ispezione;
7. si valutano e si registrano i valori di C_{pk} e della varianza del processo.

Si consideri, infine, che non sempre è conveniente l'applicazione di questo metodo. Infatti, può accadere che una caratteristica abbia un valore di C_{pk} molto elevato e, di conseguenza, la non conformità sia talmente rara che il suo monitoraggio rischierebbe di comportare uno spreco di denaro.

L'accettazione per campionamento è un metodo utile per controllare anche le variabili che altrimenti non verrebbero ispezionate. Esso consiste nello studio di un piccolo campione rappresentativo dell'intera popolazione, che permetta di prendere decisioni su quest'ultima. Indubbiamente, essendo lo studio eseguito su un numero limitato di dati, prevede l'accettazione di una probabilità di rischio non nulla [2].

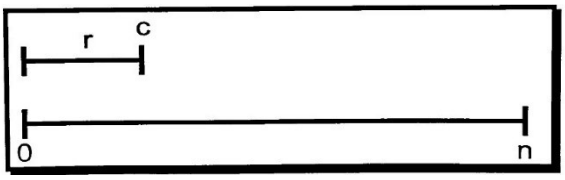
Il campionamento può essere di diverso tipo, i due principali sono quello semplice e quello doppio.

Il piano di campionamento semplice (figura 1.9) è definito da:

- n: dimensione del campione;
- c: numero di accettazione, ovvero di unità difettose ammesse nel campione;
- r: numero di unità difettose rilevate.

Se $r \leq c$, il lotto è da accettare, altrimenti è da rifiutare.

Figura 1.9: schema di funzionamento piano di campionamento semplice [10]



La dimensione del lotto N è definita da alcune regole, ovvero:

- se il processo produttivo è sotto controllo conviene avere lotti grandi (N grande) per diminuire, a parità di n, il numero di elementi controllati;
- al contrario, se il processo non è sotto controllo conviene formare lotti più piccoli (N piccolo) per avere un controllo più pronto;
- se non si possiedono sufficienti informazioni, è preferibile iniziare con lotti piccoli e poi eventualmente aumentare la dimensione del lotto [10].

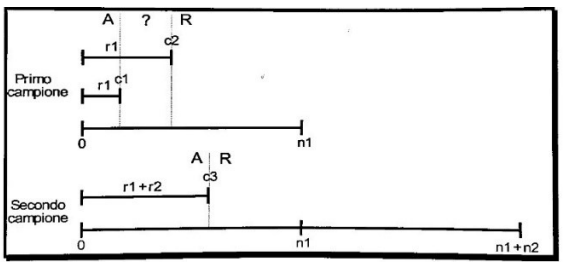
Il piano di campionamento doppio (figura 1.10) è più complesso ma garantisce un miglior compromesso tra costi e rischi derivanti dall'ispezione [2]. Esso è definito da:

- n_1, n_2: dimensioni dei due campioni;
- c_1, c_2, c_3: numeri di accettazione, con $c_3 \geq c_2$ e $c_2 > c_1$;
- r_1, r_2: numero di unità difettose rilevate nei due campioni.

L'algoritmo in questo caso è:

1. prelevamento di n_1 pezzi dal primo campione
2. si valuta:
   a. se $r_1 \leq c_1$ il lotto è da accettare, il processo termina
   b. se $r_1 > c_2$ il lotto è da rifiutare, il processo termina
   c. se $c_1 < r_1 < c_2$ si deve effettuare il secondo campionamento di dimensione n_2
3. si effettua un secondo campionamento di dimensione n_2 da cui si estraggono r_2 unità difettose
4. si valuta:
   a. se $r_1 + r_2 \leq c_3$ il lotto è da accettare
   b. se $r_1 + r_2 > c_3$ il lotto è da rifiutare
5. in entrambi i casi il processo termina [10].

Figura 1.10: schema di funzionamento del piano di campionamento doppio [10]



CAPITOLO 2
Intelligenza Artificiale

2.1. Le nuove tecnologie

L'ultimo decennio ha visto uno sviluppo eccezionale delle nuove tecnologie che hanno rivoluzionato la vita privata dei consumatori. Così come nella sfera privata, questo enorme cambiamento si è ripercosso anche nell'intero processo industriale.

Alcune delle tecnologie che hanno contribuito allo sviluppo di questa nuova era industriale sono:

- l'IoT, il quale permette una continua connessione tra le funzioni aziendali, facilitando la trasmissione orizzontale delle informazioni e la decentralizzazione del processo decisionale;
- il Cloud, che attraverso il collegamento ad Internet permette a più persone l'accesso agli stessi dati, nonostante queste non si trovino fisicamente nello stesso luogo;
- i Big Data, i quali permettono di disporre di grandi quantità di informazioni per identificare in anticipo i problemi o creare modelli più affidabili;
- i siti non presidiati e le operazioni da remoto, che rendono più autonomi il processo operativo e le operazioni di controllo.

Con lo sviluppo dell'IoT è possibile inserire le informazioni in un Cloud e renderle accessibili a chiunque in qualsiasi momento. La creazione di una piattaforma aziendale permette a tutti i membri delle diverse aree di accedere agli stessi dati, rendendo così più ampia la visione degli obiettivi e facilitando il processo decisionale. Inoltre, i dati possono essere elaborati e si possono ottenere soluzioni ai problemi più complessi in tempi rapidi.

I nuovi sistemi di automazione devono garantire affidabilità. Spesso la modellizzazione e l'ottimizzazione dei processi non sono immediate, poiché alcune informazioni potrebbero essere soggette a segreto aziendale e quindi nascoste ai progettisti, oppure alcuni modelli potrebbero essere troppo complessi.

Infine, in uno scenario sempre più automatizzato è importante considerare il ruolo dell'operatore. Se quest'ultimo dovesse essere indispensabile nel processo produttivo, si dovrebbe studiare la sua interazione con i macchinari, con l'obiettivo di non svalutare le competenze e conoscenze umane [8].

2.2. Industria 4.0 e smart PPC

Il termine "Industria 4.0" è nato nel 2011, in Germania, durante la quarta rivoluzione industriale. Si riferisce ad un sistema che mira ad una produzione intelligente, connessa e decentralizzata. L'aspetto caratterizzante questa rivoluzione è la comunicazione tra uomini, macchine e prodotti, garantita dai sistemi cyber-fisici (CPS).

La condizione principale per garantire la qualità di un prodotto è il costante monitoraggio del processo produttivo, spesso impossibile da eseguire in modo diretto su tutti i prodotti. Per questo motivo, l'utilizzo di sensori che monitorano e rielaborano i dati sta diventando sempre più frequente e permette lo sviluppo di servizi innovativi, come la manutenzione predittiva [1]. Con l'avvento dell'Industria 4.0, la pianificazione e il controllo della produzione (PPC) hanno subito un grande riadattamento ai nuovi sistemi digitali, tanto che si è iniziato a parlare di "smart PPC". Il PPC include le attività di pianificazione, controllo, monitoraggio, organizzazione e riprogrammazione della produzione, eseguite attraverso la redazione di piani di previsione della domanda, tra cui MPS (Master Production Scheduling) e MRP (Material Requirements Planning). Gli attuali cambiamenti repentini del mercato e le enormi quantità di dati da gestire impongono la necessità di adottare strumenti all'avanguardia per la gestione delle operazioni aziendali. L'implementazione dell'Industria 4.0 necessita di diversi elementi, quali i CPS, l'IoT, analisi dei Big Data e intelligenza artificiale (BDA/AI), produzione in cloud (CMg) e manifattura additiva (AM). L'integrazione tra PPC e Industria 4.0, con l'introduzione di nuovi metodi per la previsione della domanda, per la pianificazione e il controllo della capacità produttiva e del magazzino, ha permesso un aumento della flessibilità e della qualità e una riduzione di costi e dei tempi di consegna. I benefici principali di questa rivoluzione sono stati l'incremento di automazione, tracciabilità, ottimizzazione, sincronizzazione dei processi, la semplificazione delle attività e dei processi decisionali [6].

Le quattro caratteristiche principali dello smart PPC sono:

- gestione dei dati in tempo reale;
- pianificazione e riprogrammazione dinamica della produzione;
- controllo autonomo della produzione, che prevede il coordinamento e la condivisione di informazioni tra le diverse parti del sistema;
- apprendimento continuo, che porta ad un costante incremento delle conoscenze [18].

2.2.1. Le fasi per l'adozione dello smart PPC

I passi per adottare una soluzione di smart PPC sono i seguenti:

1. Studio preliminare: determinare gli obiettivi e le priorità.
   Lo studio generalmente inizia nel momento in cui si identifica un problema aziendale o la mancata opportunità di mercato. Solitamente l'obiettivo che si stabilisce è influenzato dall'ambiente interno ed esterno in cui avviene l'attività produttiva e spesso si devono raggiungere compromessi tra le necessità delle diverse aree aziendali.
2. Definizione dei requisiti del sistema e degli indicatori di performance.
   In questa seconda fase si stabiliscono i dettagli dell'obiettivo definito al punto precedente. Spesso le priorità sono stabilite dal team di gestione dell'azienda. Tuttavia, può essere utile confrontarsi con i lavoratori che interagiscono direttamente con il sistema produttivo per

prendere decisioni più adeguate. Inoltre, in questa fase vengono stabiliti degli indicatori di performance che denotano la qualità dell'analisi e l'affidabilità del sistema.

3. Identificazione delle fonti dei dati e degli algoritmi di apprendimento che risolvono il sistema. Questa fase può essere seguita da un team tecnico che include un ingegnere esperto di ML ed ha l'obiettivo di determinare una visione completa sia del problema aziendale, che di quello tecnico. Di solito, inizialmente il problema viene affrontato con algoritmi di ML basilari, mentre successivamente si adottano modelli ibridi che combinano le diverse soluzioni di base.
4. Progettazione del sistema tenendo conto degli strumenti disponibili.
   L'integrazione di diversi sistemi di smart PPC consente di ottenere una soluzione più efficace rispetto all'adozione di un sistema monolitico. I processi di elaborazione dei dati, sviluppo dei modelli e previsione possono essere eseguiti senza l'utilizzo della manodopera, automatizzando i processi tramite strumenti come il ML.
5. Sviluppo di metodologie che considerino l'innovazione continua e l'adattabilità a scenari futuri.
   In questa fase si scelgono i software da adottare, si decide come gestire i servizi cloud e le tecnologie utilizzate e come sviluppare il sistema in modo che supporti l'innovazione continua. Questo significa che l'infrastruttura IT consolidata elimina i processi manuali e apporta miglioramenti al sistema di lavoro rendendolo più agile e diminuendo i tempi di fermo impianto. Per questo motivo vengono introdotte sempre di più, in azienda, figure come ingegneri di ML e sviluppatori di software che consentono il perfezionamento dei modelli [22].

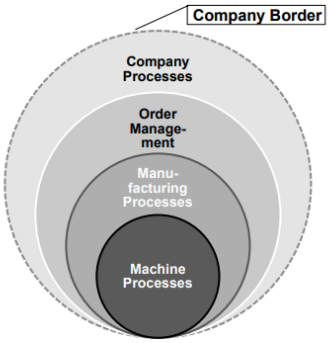
2.3. Processo per il controllo della qualità

Il processo di definizione dei requisiti prevede lo sviluppo di un modello generico per avere una visione completa dello stato attuale e della domanda. A seconda della domanda si scelgono i fornitori e i macchinari da utilizzare. In seguito, tutte le informazioni rilevanti sono condivise con gli stakeholder, le loro richieste vengono combinate in modo tale da indentificare univocamente gli obiettivi. Queste ultime fasi sono rilevanti per definire i requisiti che il prodotto deve avere, anche in funzione del successivo controllo di qualità.

Il controllo della qualità viene eseguito su quattro livelli (figura 2.1):

- i processi interni della macchina, che influenzano parti sostanziali della qualità del prodotto;
- il processo manifatturiero, che amplia la visuale a tutti i processi da monte a valle;
- la gestione degli ordini, che si concentra sulle richieste dei clienti e su come queste influenzano le decisioni;
- tutti i processi dell'azienda che entrano in contatto con quello manifatturiero.

Figura 2.1: i quattro livelli di dettaglio per il controllo della qualità [1]



La definizione del sistema di obiettivi si suddivide in tre fasi:

1. fase di inizializzazione
2. fase di analisi dello stato attuale e del controllo qualità
3. fase di sintesi del sistema di obiettivi

La prima fase si concentra sui processi che avvengono all'interno dei confini aziendali e riconosce le caratteristiche del prodotto rilevanti per il sistema di controllo qualità intelligente.

L'analisi dello stato attuale ha lo scopo di definire i sistemi di controllo qualità già presenti all'interno dell'azienda. In questa fase è consigliato compilare questionari strutturati che rilevano i problemi di qualità che si presentano, le modalità di risoluzione, i valori degli indicatori rilevati nello stato attuale e quelli da raggiungere. I dati raccolti vengono poi inseriti in un documento descrittivo, per facilitarne la comprensione.

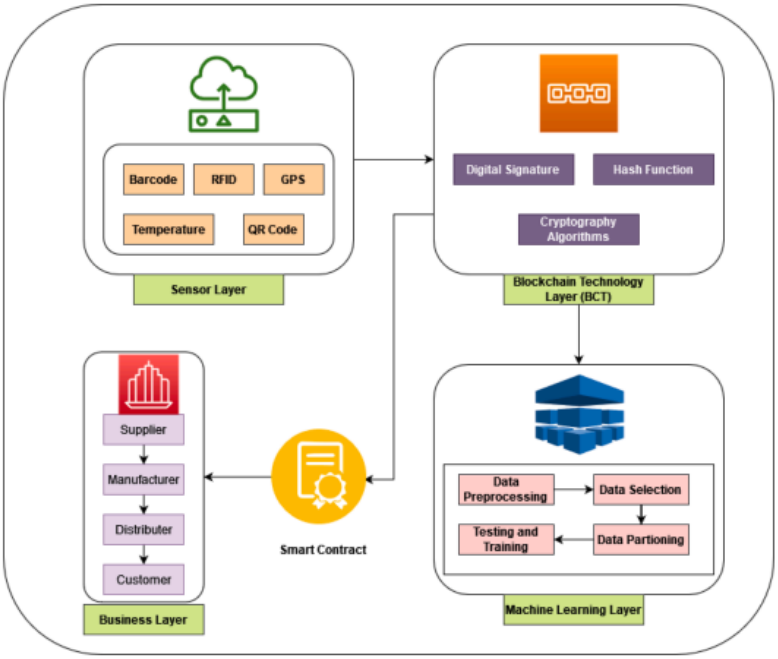
La fase di sintesi del sistema di obiettivi si suddivide in due parti. Nella prima vengono identificati gli obiettivi da raggiungere e i vincoli da rispettare, i quali potrebbero comportare delle modifiche dei requisiti tecnici del sistema di controllo qualità. In una seconda parte, invece, viene redatto il documento dei requisiti [1].

2.4. Tecnologie Blockchain

Il sistema di controllo qualità basato sull'utilizzo di tecnologie blockchain è costituito da quattro livelli, come mostra la figura 2.2:

- al primo livello ci sono i sensori, che permettono di conoscere la posizione degli oggetti. Oltre a questi possono essere utilizzati anche sensori termici o di umidità;
- al secondo livello le tecnologie blockchain, che permettono di valutare la qualità dei dati;
- il terzo livello, detto anche "contratto intelligente", è dove le informazioni sono registrate e condivise per rendere più efficiente la supply chain. Spesso queste ultime devono essere ristrette ad un gruppo limitato di persone per motivi di privacy;
- l'ultimo è il livello di business, che include tutti i collegamenti tra fornitori, produttori, distributori e consumatori, controllati tramite blockchain [11].

Figura 2.2: architettura del sistema proposto [11]



2.4.1. Controllo qualità in tempo reale

Lo sviluppo delle tecnologie blockchain aumenta con l'incremento di imprese che scambiano i loro dataset all'interno e all'esterno dell'impianto. Il funzionamento di questo sistema si basa sulla capacità del produttore di definire i propri obiettivi e coinvolge tutta la supply chain, dal fornitore, al produttore, al distributore, offrendo al fornitore contratti intelligenti sulla base delle informazioni possedute dalla blockchain. Tali informazioni non sono visibili a tutti, in particolare non sono condivise ai fornitori rivali, per questo assicurano un elevato livello di sicurezza, garantito da firma digitale e tecniche di crittografia. Il messaggio possiede caratteristiche di integrità, validità e non ripudio e può essere aperto soltanto attraverso un processo di autenticazione riservato alle persone autorizzate. I contratti intelligenti permettono di riconoscere qualsiasi uso fraudolento del database. Per questo motivo, l'amministrazione deve registrare gli utenti che hanno la facoltà di accedere ai dati ed essi possono, quindi, eseguire transazioni. Le richieste di transazione possono essere eseguite, convalidate o respinte. Se sono autorizzate, vengono registrate sui blocchi [11].

2.4.2. L'utilizzo delle blockchain

L'uso delle blockchain è cresciuto negli ultimi anni in diversi settori:

- Gestione della catena di approvvigionamento: serve per creare una registrazione trasparente della catena di approvvigionamento, dalle materie prime a produzione, trasporto e distribuzione. Questo porta a processi più rapidi ed efficienti.

- Manutenzione predittiva: consente alle aziende di monitorare la salute e le prestazioni delle apparecchiature in tempo reale.
- Gestione dell'energia: gestisce il flusso di energia tra produttori e consumatori in modo decentralizzato.
- Contratti intelligenti: la tecnologia Blockchain può automatizzare l'esecuzione di contratti intelligenti tra dispositivi IoT. In questo modo, gli accordi tra dispositivi vengono eseguiti in modo trasparente e a prova di manomissione [11].

2.5. Machine Learning

Il machine learning (ML) è un sottoinsieme dell'intelligenza artificiale che si occupa di trovare un sistema che permetta ai computer di imparare dall'esperienza, ovvero reagire a eventi prevedibili e imprevedibili a partire da dati storici o ottenuti in tempo reale, contenuti in un training set, in italiano "insieme di addestramento". L'apprendimento automatico è stato introdotto nei processi industriali negli ultimi decenni per prevedere azioni future e facilitare il processo decisionale. L'utilizzo di algoritmi di ML permette di dare un punto di vista diverso al settore manifatturiero, soprattutto per quanto riguarda il rilevamento dei difetti [11].
Per eseguire l'analisi a partire da un dataset, si deve scegliere l'algoritmo di ML più adatto. Il tipo di apprendimento può essere suddiviso in due categorie: supervisionato e non supervisionato. Fanno parte dell'apprendimento supervisionato gli algoritmi di regressione e classificazione, mentre di quello non supervisionato l'algoritmo di raggruppamento (o clustering).

2.5.1. Apprendimento supervisionato

L'apprendimento supervisionato viene implementato attraverso un algoritmo in cui le macchine hanno lo scopo di prevedere gli output sulla base di dati di input, inseriti nel training set. Il processo, quindi, prevede una fase iniziale di addestramento del modello, una fase di valutazione e infine la previsione dell'output.
I modelli di apprendimento supervisionato possono essere di regressione o di classificazione. I primi sono applicati quando i dati sono variabili continue, mentre i secondi quando i dati sono variabili categoriali e si possono suddividere in due o più classi [7] [13] [22].
Il modello di regressione viene solitamente implementato attraverso la regressione lineare. Questa permette di individuare la relazione tra variabile dipendente (y) e variabile indipendente (x). Considerando l'immagine in figura 2.3, la retta è data dall'equazione $y = a_0 + a_1 x + \varepsilon$, dove y è la variabile dipendente (variabile target), x è la variabile indipendente (variabile predittiva), a_0 è l'intercetta, a_1 è il coefficiente di regressione lineare ed ε è l'errore [43].

Nell'algoritmo di classificazione il programma, invece, deve suddividere le osservazioni in un certo numero di classi o gruppi. Ad esempio, se si avessero due classi, una classe A e una classe B la situazione sarebbe quella rappresentata in figura 2.4 [36].

Figura 2.3: regressione lineare [43]

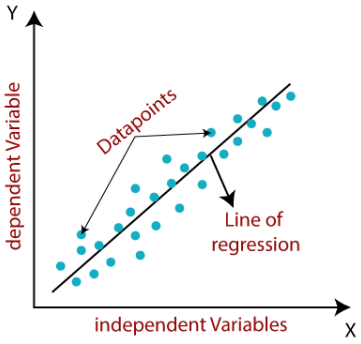
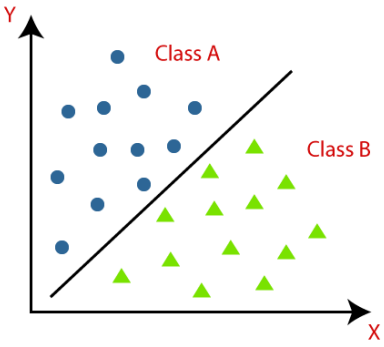


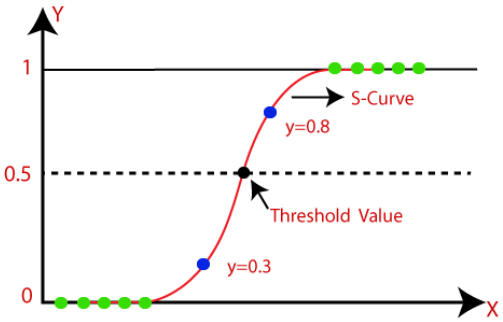
Figura 2.4: algoritmo di classificazione [36]



Gli algoritmi di classificazione possono essere di diverso tipo, i principali sono i seguenti:

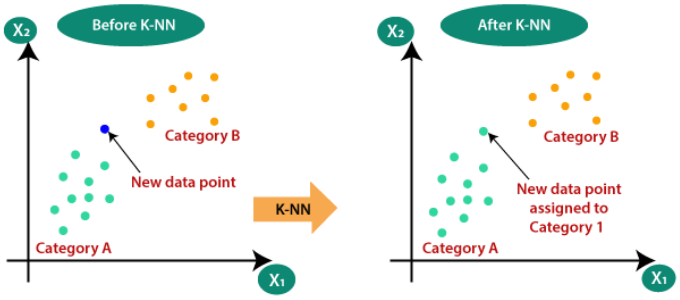
- Regressione logistica: sulla base dei dati contenuti nel training set, genera valori probabilistici che si collocano tra 0 e 1 formando una curva ad "s" come quella in figura 2.5 [44].

Figura 2.5: regressione logistica [44]



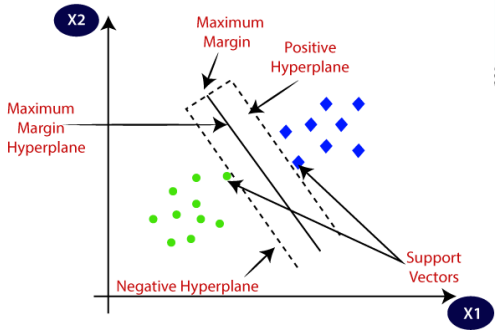
- K valori più vicini (KNN): si basa sulla classificazione dei nuovi dati sulla base delle similitudini con i K valori più vicini nel dataset. L'algoritmo calcola la probabilità che tali valori siano vicini ai nuovi dati. L'algoritmo è rappresentato graficamente in figura 2.6 [42].

Figura 2.6: KNN [42]



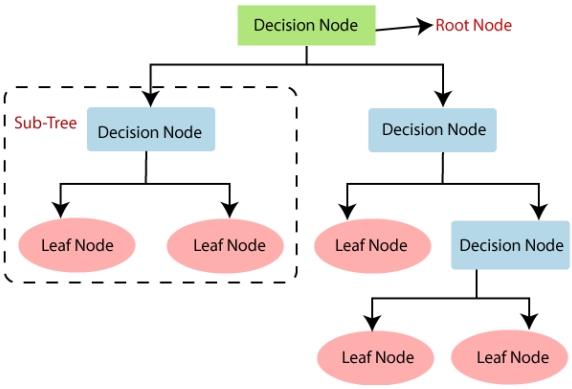
- Macchina vettoriale di supporto (SVM): mira alla ricerca di un confine di classificazione in grado di separare le due classi. La macchina sceglie i punti più estremi, detti "vettori di supporto", che permettono di individuare l'iperpiano come in figura 2.7 [47].

Figura 2.7: algoritmo SVM [47]



- Albero decisionale: è costituito da nodi e rami; ogni nodo conta un ramo entrante e due (o talvolta più) rami uscenti, eccetto per il nodo iniziale e i nodi test (cioè i nodi senza rami uscenti). Più alberi decisionali vengono collegati tra loro per migliorare l'apprendimento e vengono utilizzati per avere un quadro definito delle possibili soluzioni ad un problema. Un esempio è rappresentato in figura 2.8 [40].

Figura 2.8: Albero decisionale [40]



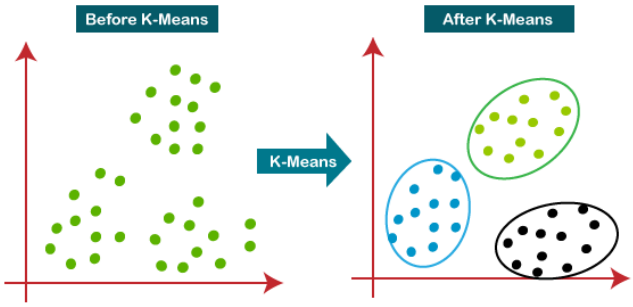
2.5.2. Apprendimento non supervisionato

Nell'apprendimento non supervisionato, invece, non vengono distinti dati di input e dati di output, ma l'algoritmo esplora tutti i dati e crea modelli cercando di individuare quali elementi possono essere predittivi di altri. Fa parte dell'apprendimento non supervisionato il modello di clustering, che ha l'obiettivo di raggruppare i dati che hanno molte similarità e differenziarli dai gruppi di dati che hanno attributi diversi [37].

Alcune delle tipologie di metodi di clustering sono le seguenti:

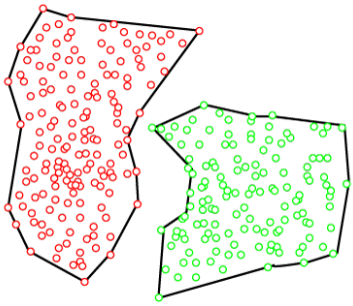
- Partizionamento: è un esempio di clustering non gerarchico, il cui algoritmo più utilizzato è il k-means clustering. Il dataset viene suddiviso in k gruppi. Il centro del cluster viene stabilito in modo che la distanza tra i punti di un cluster e il baricentro di un altro cluster sia minima, come mostrato in figura 2.9 [41].

Figura 2.9: partizionamento con metodo k-means [41]



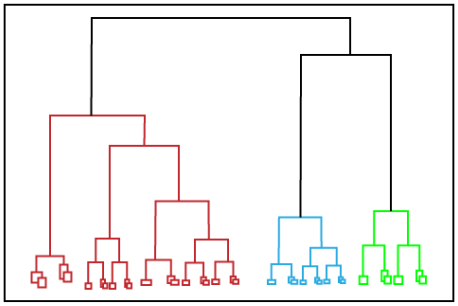
- Clustering basato sulla densità (DBSCAN): raggruppa le aree ad alta densità in cluster finché le aree ad alta densità possono essere collegate. I cluster sono suddivisi tra loro da aree meno dense, come in figura 2.10.

Figura 2.10: clustering basato sulla densità [37]



- Clustering gerarchico: in questo caso i dati vengono suddivisi in cluster e si crea una struttura ad albero detta dendrogramma, come in figura 2.11.

Figura 2.11: clustering gerarchico [37]



- Fuzzy clustering: è un metodo in cui i dati possono appartenere a uno o più cluster. Ogni dataset ha dei coefficienti di partecipazione che dipendono dal grado di partecipazione ad un cluster [37].

CAPITOLO 3

Casi studio

3.1. Metodologia proposta

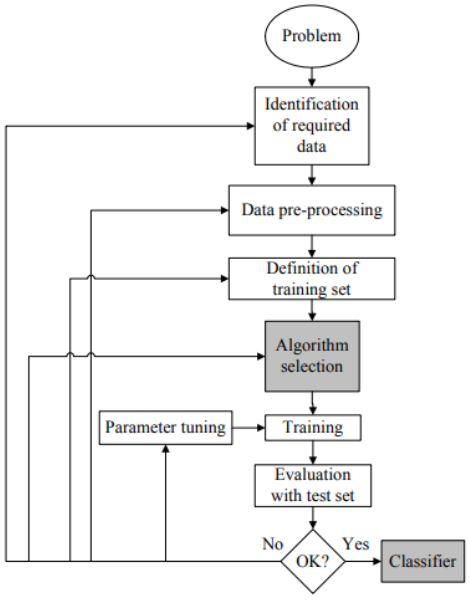
Il metodo che verrà applicato in seguito per eseguire la manutenzione predittiva e, quindi, il controllo della qualità è un algoritmo di classificazione binaria. Il modello di classificazione prevede la probabilità che un'osservazione appartenga ad una specifica classe e successivamente la classifica sulla base di quanto ottenuto. Il risultato è un modello predittivo [4] [8].

3.1.1. Training set e test set

La creazione del modello consiste nella suddivisione dei dati secondo un criterio stabilito dall'utente: dopo aver raccolto i dati, devono essere riconosciuti i possibili difetti e suddivisi in classi. La variabile di risposta deve essere una variabile binaria, ovvero deve poter essere categorizzata in due gruppi. Le variabili predittive, invece, possono essere continue o categoriali e nell'esecuzione dell'analisi può essere utilizzata una combinazione delle due. I dati devono essere suddivisi in un set di addestramento (training set) e in un test set, che serviranno rispettivamente per la creazione del modello e per la sua successiva valutazione. Generalmente, quando il dataset è di grandi dimensioni (numero di osservazioni > 5000) i software, come Minitab, suddividono automaticamente i dati in training set e test set. Le osservazioni facenti parte del training set vengono utilizzate per creare un classificatore, che deve poi potersi adattare anche alle osservazioni contenute nel test set [13].

L'algoritmo si sviluppa seguendo la sequenza di fasi illustrata in figura 3.1.

Figura 3.1: fasi di un algoritmo di apprendimento supervisionato [16]



Identificati i dati più rilevanti ai fini dello studio, è necessario eseguire un'operazione di preelaborazione degli stessi, ovvero gestione dei dati mancanti o anomali. Successivamente, si definisce il training set e si seleziona l'algoritmo che permette di lavorare in modo efficace con il dataset. Si esegue l'addestramento dei dati e si valuta il modello attraverso il confronto con un test set. Se il modello è soddisfacente si conclude lo studio, altrimenti si torna al passo precedente e si esegue nuovamente il procedimento da quel punto in poi [16].

3.1.2. Matrice di confusione e indicatori di performance

La fase di valutazione del modello creato avviene attraverso l'utilizzo di alcuni strumenti, come la matrice di confusione. Per ogni osservazione si può ricavare uno dei seguenti quattro esiti:

- TN (true negative): evento negativo previsto correttamente;
- TP (true positive): evento positivo previsto correttamente;
- FN (false negative): evento positivo, erroneamente previsto come negativo, detto anche errore di II tipo;
- FP (false positive): evento negativo, previsto erroneamente come positivo, detto anche errore di I tipo.

Da questi viene creata la matrice di confusione (figura 3.2), che riporta sulle righe i valori che il modello aveva previsto e sulle colonne i valori effettivi [38].

Figura 3.2: matrice di confusione [38]

	Actual Positive	**Actual Negative**
Predicted Positive	True Positive	False Positive
Predicted Negative	False Negative	True Negative

Il successo del modello è misurato dagli indicatori di performance, elencati in tabella 3.1 [4] [16].

Tabella 3.1: indicatori di performance

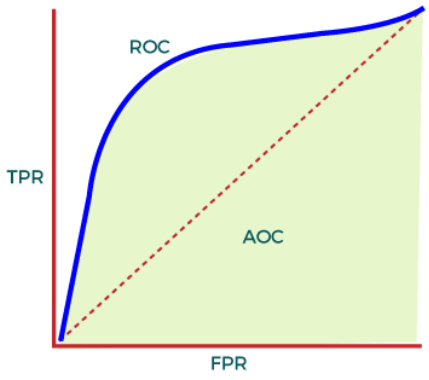
INDICATORE	**FORMULA**	**DESCRIZIONE**
Accuratezza	$\frac{TN + TP}{TN + TP + FN + FP}$	Indica quanto spesso il modello prevede l'output corretto.
Tasso di errore	$\frac{FP + FN}{TN + TP + FN + FP}$	Indica quanto spesso il modello prevede l'output errato ed è il complementare dell'indice di accuratezza.

Sensitività	$\frac{TP}{TP + FN}$	Indica la probabilità che l'evento di interesse sia previsto correttamente rispetto al totale di eventi positivi.
Precisione	$\frac{TP}{TP + FP}$	Indica la percentuale di output positivi previsti correttamente, rispetto a quelli totali positivi previsti.
Punteggio F	$\frac{2 * sensitività * precisione}{sensitività + precisione}$	Raggiunge il massimo se sensitività e precisione sono uguali e serve per confrontare i due indicatori.

3.1.3. Curva ROC

Per esaminare la performance del modello di classificazione, si può utilizzare anche la curva operativa caratteristica del ricevitore (ROC), ovvero uno strumento grafico che valuta le prestazioni del modello a differenti livelli di soglia (figura 3.3). Sull'asse delle ordinate riporta i veri positivi (true positive), mentre sull'asse delle ascisse i falsi positivi (false positive) [35].
I valori su ogni asse variano da 0 a 1 e alcuni punti della curva sono particolarmente rilevanti. Il punto in basso a sinistra (0, 0) indica che il classificatore non commette mai errori (falsi positivi), ma allo stesso tempo non ottiene veri positivi. La strategia opposta (1, 1) indica che si emettono incondizionatamente classificazioni positive. La classificazione perfetta si ottiene in corrispondenza del punto (0, 1).
La curva ROC è un grafico bidimensionale, tuttavia per valutare un classificatore è utile ridurre la valutazione ad un valore scalare. Il metodo più immediato per farlo è il calcolo dell'area sotto la curva (AUC). Un classificatore perfetto ha un'AUC pari a 1. Al contrario, quando le variabili predittive non sono ben collegate all'evento, il classificatore genera una curva che è la bisettrice del primo quadrante del piano cartesiano ed è detto classificatore casuale. In generale, più la curva si trova verso l'alto e verso sinistra, più il modello è buono [9].

Figura 3.3: curva ROC [35]



3.2. Manutenzione predittiva di una fresatrice

Lo studio che è stato eseguito in seguito riguarda la manutenzione predittiva di una fresatrice. I dati sono raccolti in un dataset reso pubblico da Matzka, S. (2020), costituito da 10000 rilevazioni e 14 caratteristiche rilevate per ognuna di esse, descritte di seguito [26].

1. UID: identificativo univoco che va da 1 a 10000.
2. ID del prodotto: costituito da una lettera L, M o H rispettivamente per le varianti di bassa, media o alta qualità e da un numero di serie specifico della variante.
3. Tipo: tipo di prodotto L, M o H dalla colonna 2.
4. Temperatura dell'aria [K]: valori numerici intorno ai 300 K che seguono una distribuzione normale con deviazione standard di 2 K.
5. Temperatura del processo [K]: valori numerici ricavati aggiungendo 10 K alla temperatura dell'aria, che seguono una distribuzione normale con deviazione standard di 1 K.
6. Velocità di rotazione [rpm]: calcolato a partire da una potenza di 2860 W, sovrapposta ad un disturbo normalmente distribuito.
7. Coppia [Nm]: valori normalmente distribuiti intorno ai 40 Nm con una deviazione standard di 10 Nm e nessun valore negativo.
8. Usura dell'utensile [min]: le varianti di qualità H/M/L aggiungono 5/3/2 minuti di usura all'utensile utilizzato nel processo.
9. Guasto macchina: variabile binaria che indica se la macchina è soggetta a guasto per uno dei cinque motivi elencati di seguito. Anch'essi sono rappresentati da variabili binarie, che assumono valore 1 in caso di fallimento del processo e 0 altrimenti. Se almeno una delle seguenti cause di fallimento assume valore 1, il valore della variabile "guasto macchina" è pari a 1, altrimenti è uguale a 0.
    - Guasto da usura utensile (TWF): l'utensile viene sostituito o si guasta ad un tempo d'usura compreso tra 200 e 240 minuti.
    - Guasto per dissipazione di calore (HDF): se la differenza tra temperatura dell'aria e quella del processo è minore di 8,6 K e la velocità di rotazione degli utensili è inferiore a 1380 giri/min avviene il guasto.
    - Mancanza di potenza (PWF): se la potenza, data dal prodotto tra coppia e velocità di rotazione, è inferiore a 3500 W o superiore a 9000 W, il processo fallisce.
    - Fallimento da sovraccarico (OSF): se il prodotto tra usura dell'utensile e coppia è maggiore di 11000 minNm per la variante L di prodotti (12000 per M e 13000 per H), il processo fallisce per sovraccarico.
    - Fallimenti casuali (RNF): fallimenti indipendenti dai parametri del processo [45].

3.2.1. Risultati dell'analisi con classificazione CART

Il dataset è stato studiato tramite l'utilizzo del software Minitab, attraverso l'impostazione di un algoritmo di classificazione binaria [39]. La risposta binaria è la variabile "guasto macchina", che dipende dalle variabili continue "temperatura dell'aria", "temperatura del processo", "velocità di rotazione", "coppia" e "usura dell'utensile" e da quella categoriale "tipo".

La probabilità preventiva è la probabilità che un'osservazione ricada in un certo gruppo ed in questo caso si assume sia uguale per tutti i livelli. Per la suddivisione dei nodi nel diagramma ad albero si utilizza il metodo di Gini, che ha lo scopo di minimizzare le impurità nei nodi successivi [19]. Il criterio di selezione dell'albero ottimale è quello entro 1 errore standard del costo minimo di errata classificazione. Infine, il 70% dei dati costituisce il training set e il restante 30% il test set e la suddivisione delle righe avviene in maniera casuale. Tutti i metodi utilizzati sono elencati in figura 3.4.

Figura 3.4: metodologia applicata

Method

Prior probabilities	Same for all classes
Node splitting	Gini
Optimal tree	Within 1 standard error of minimum misclassification cost
Model validation	70/30% training/test sets
Rows used	10000

La tabella in figura 3.5 mostra i risultati sia per il training set che per il test set ottenuti dall'indagine. La probabilità di guasto della macchina è molto bassa, poco più del 3% sia per il training test, che per il test set.

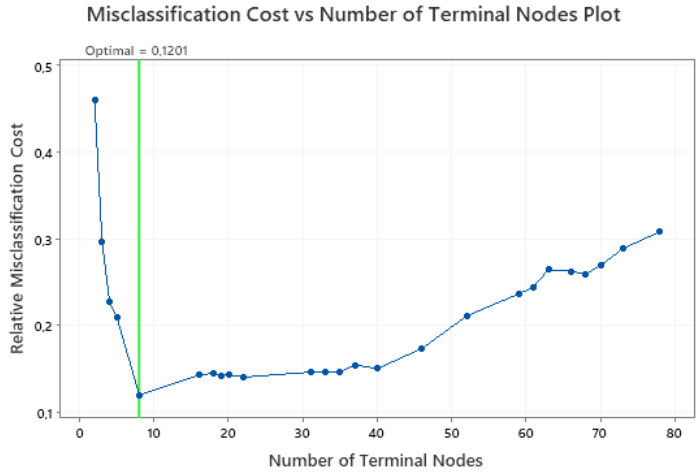
Figura 3.5: risposta binaria ottenuta

Binary Response Information

		Training		Test	
Variable	Class	Count	%	Count	%
Machine failure	1 (Event)	242	3,5	97	3,2
	0	6745	96,5	2916	96,8
	All	6987	100,0	3013	100,0

Il diagramma ad albero con 8 nodi è quello che ha minor costo di errata classificazione, ovvero pari a 0,1201 come si vede dal grafico dei costi di errata classificazione rispetto al numero di nodi terminali (figura 3.6).

Figura 3.6: grafico dei costi di errata classificazione rispetto al numero di nodi terminali



L'albero decisionale CART è un albero binario che rappresenta gli 8 nodi terminali (figura 3.7). Il colore blu indica i guasti, mentre il colore rosso indica le macchine non soggette a guasto. Al nodo 1 si osserva che il numero di guasti è di 242 sui 6987 totali, pari al 3,5%. Le 6987 rilevazioni vengono suddivise nei nodi 2 e 5 in base alla velocità di rotazione. Da questi due nodi si osserva che la percentuale di guasto è significativamente più elevata a velocità rotazionale minore di 1386,5 rpm (nodo 2), rispetto a quella rilevata nel nodo 5. In particolare, è pari al 15,1% nel nodo 2 e all'1,4% nel nodo 5. Entrambi i nodi appena studiati vengono a loro volta suddivisi in altri nodi, in base ai valori delle variabili "temperatura dell'aria", "usura dell'utensile" e "coppia", fino ad arrivare ai nodi terminali. Il percorso per arrivare ai nodi terminali è quello che genera i gruppi più puri. I nodi terminali sono nodi che non possono più essere suddivisi ulteriormente.

Dovendo classificare un'osservazione di cui si conoscono i valori per ogni variabile, il diagramma permette di ricavare il nodo terminale in cui ricade l'osservazione e, in base a questo, determinare la probabilità che l'evento si verifichi o meno. Se ad esempio un'osservazione assume i valori in figura 3.8, percorrendo l'albero si va dal nodo 1 al nodo 5, essendo la velocità di rotazione pari a 1400 rpm. Dal nodo 5 si raggiunge il nodo terminale 8, poiché l'usura dell'utensile è pari a 230 min. In questo modo si può dedurre che con una probabilità del 91,2% la macchina non è soggetta a guasto.

Il grafico di importanza delle variabili (figura 3.9) indica la variabile più rilevante assegnandole un'importanza pari al 100%, in questo caso è la variabile "coppia". Tutte le altre variabili hanno un'importanza espressa in percentuale, misurata rispetto alla variabile più rilevante. In questo caso, oltre alla variabile "coppia", in ordine di importanza ci sono "velocità di rotazione" (71,9%), "usura dell'utensile" (28,3%), "temperatura dell'aria" (18,5%). Invece "temperatura del processo" e "tipo" possono essere considerate irrilevanti.

Figura 3.7: diagramma ad albero con 8 nodi terminali

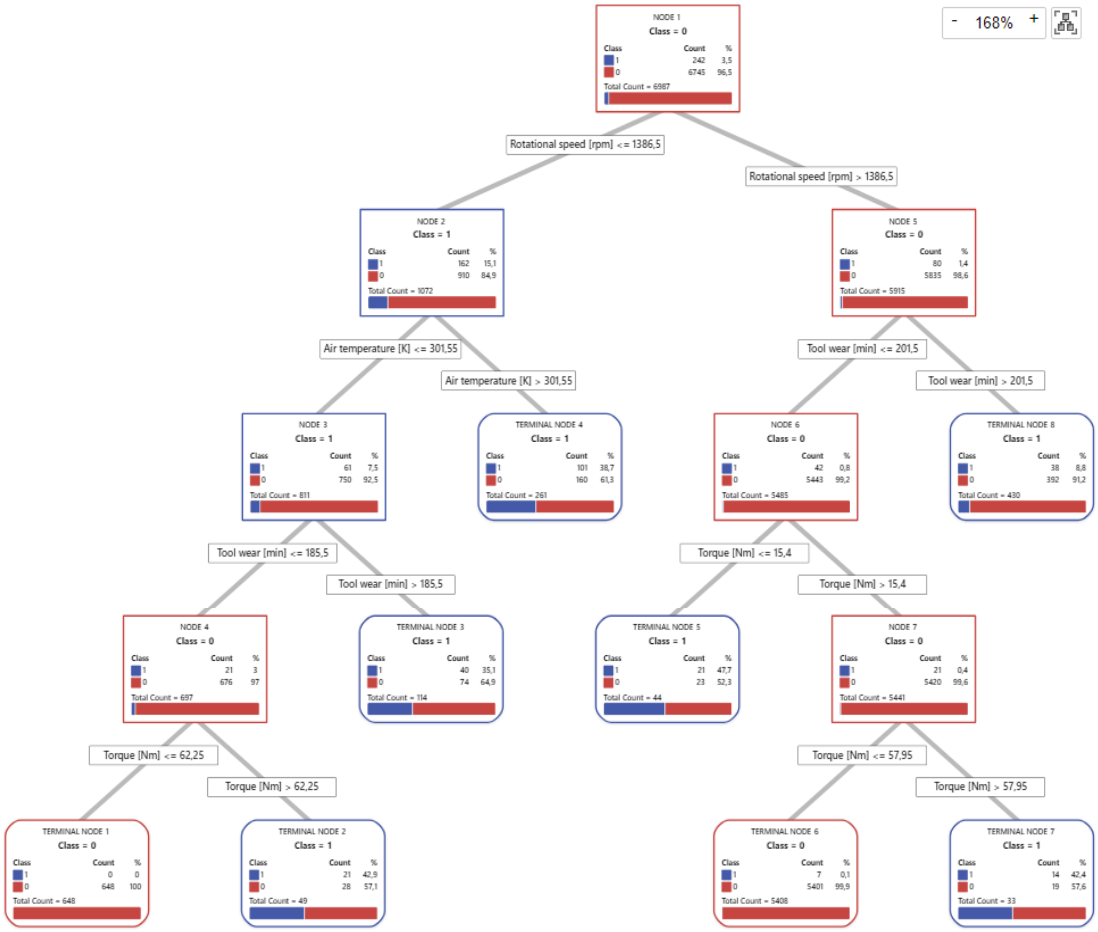


Figura 3.8: previsione di un guasto macchina

Prediction for Machine failure

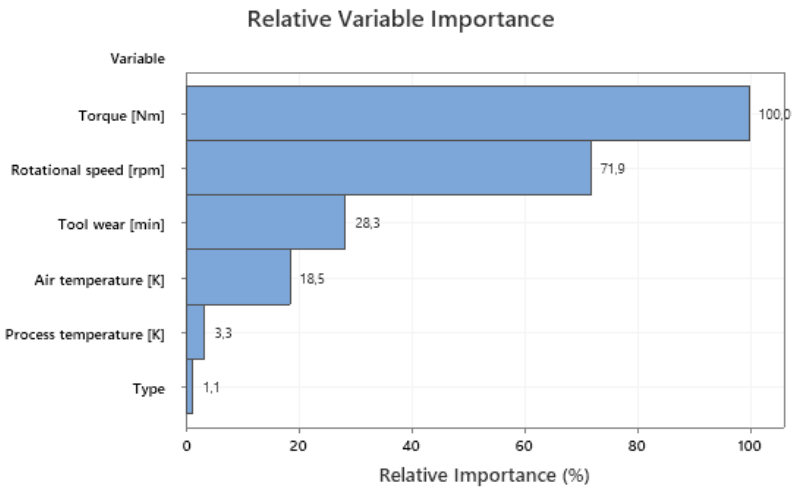
Settings

Air temperature [K] = 305; Process temperature [K] = 315; Rotational speed [rpm] = 1400; Torque [Nm] = 20; Tool wear [min] = 230; Type = M

Prediction

Obs	Terminal Node ID	Class	Prob (Class = 1)	Prob (Class = 0)
1	8	1	0,0883721	0,911628

Figura 3.9: grafico di importanza delle variabili



La matrice di confusione (figura 3.10) indica che l'albero esegue delle previsioni molto buone, perché i tassi di veri positivi e veri negativi sono elevati sia per il training set che per il test set, mentre i tassi di errore sono molto bassi. L'errore di I tipo (falsi positivi) è più frequente di quello di II tipo, ma complessivamente la percentuale di correttezza è intorno al 90% sia per il training set che per il test set.

Dalla matrice di confusione si ricavano gli indicatori di accuratezza, sensitività, precisione e punteggio F, che sono molto elevati, mentre quello di errore è molto basso (Tabella 3.2). Questo conferma che il modello creato esegue previsioni molto buone.

Figura 3.10: matrice di confusione

Confusion Matrix

	Predicted Class (Training)				Predicted Class (Test)			
Actual Class	Count	1	0	%Correct	Count	1	0	%Correct
1 (Event)	242	235	7	97,1	97	95	2	97,9
0	6745	696	6049	89,7	2916	290	2626	90,1
All	6987	931	6056	89,9	3013	385	2628	90,3

Statistics	Training (%)	Test (%)
True positive rate (sensitivity or power)	97,1	97,9
False positive rate (type I error)	10,3	9,9
False negative rate (type II error)	2,9	2,1
True negative rate (specificity)	89,7	90,1

Tabella 3.2: indicatori di performance per il training set e per il test set

	Training set	**Test set**
Accuratezza	93,4%	94%
Tasso di errore	6,6%	6%
Sensitività	97,1%	97,9%
Precisione	90,4%	90,8%
Punteggio F	93,6%	94,2%

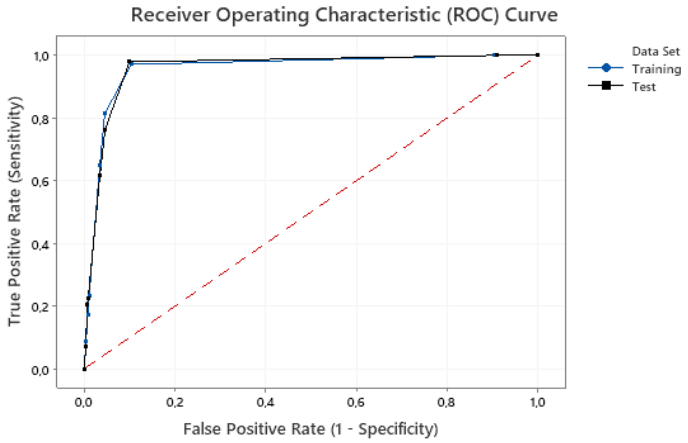
La curva ROC (figura 3.11) indica se i dati sono classificati bene. Sull'asse y c'è il tasso di veri positivi, mentre sull'asse x quello di falsi positivi. Quando il modello consente di classificare correttamente tutti i dati, l'area sottesa alla curva ROC è pari a 1. In questo caso è pari a 0,9568 per il training set e 0,9588 per il test set, quindi in entrambi i casi molto buono.

Figura 3.11: curva ROC

Misclassification

Input Misclassification Cost	Predicted Class 1	Predicted Class 0
Actual Class 1		1,00
Actual Class 0	1,00	

	Training				Test			
Actual Class	Count	Misclassed	% Error	Cost	Count	Misclassed	% Error	Cost
1 (Event)	242	7	2,9	0,0289	97	2	2,1	0,0206
0	6745	696	10,3	0,1032	2916	290	9,9	0,0995
All	6987	703	10,1	0,0661	3013	292	9,7	0,0600



Area Under Curve: Training = 0,9568; Test = 0,9588

3.2.2. Risultati dell'analisi con regressione logistica binaria

Il dataset viene studiato in questo sottoparagrafo con il metodo della regressione logistica, descritto in figura 3.12.

Figura 3.12: metodologia applicata

Method

Link function	Logit
Categorical predictor coding	(1; 0)
Rows used	10000
Test set fraction	30,0%

Response Information

Variable	Value	Training Count	Test Count
Machine failure	1	246	93 (Event)
	0	6754	2907
	Total	7000	3000

L’equazione di regressione che si ottiene dal modello è quella riportata in figura 3.13, dove si vede che Y’ può essere rappresentata da tre funzioni, ovvero una per ogni livello della variabile categoriale “tipo”.

Figura 3.13: equazione di regressione

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Type

H Y' = -54,06 + 0,6764 Air temperature [K] - 0,5912 Process temperature [K] + 0,01094 Rotational speed [rpm] + 0,2689 Torque [Nm] + 0,01462 Tool wear [min]

L Y' = -53,44 + 0,6764 Air temperature [K] - 0,5912 Process temperature [K] + 0,01094 Rotational speed [rpm] + 0,2689 Torque [Nm] + 0,01462 Tool wear [min]

M Y' = -53,88 + 0,6764 Air temperature [K] - 0,5912 Process temperature [K] + 0,01094 Rotational speed [rpm] + 0,2689 Torque [Nm] + 0,01462 Tool wear [min]

Come mostra la figura 3.14, che riporta i coefficienti riferiti all’equazione Y’ per il tipo H, i p-value di tutti i termini sono molto bassi, eccetto per il tipo M, questo dimostra che l’associazione statistica tra la variabile di risposta e i termini che hanno p-value basso è significativa. Per il tipo M il p-value è molto elevato; pertanto, non c’è evidenza statistica che ci sia correlazione tra il guasto della macchina e le varianti di prodotto con qualità media (M).

Figura 3.14: tabella dei coefficienti

Coefficients

Term	Coef	SE Coef	Z-Value	P-Value	VIF
Constant	-54,1	16,9	-3,19	0,001	
Air temperature [K]	0,6764	0,0828	8,17	0,000	4,68
Process temperature [K]	-0,591	0,110	-5,39	0,000	4,59
Rotational speed [rpm]	0,010936	0,000616	17,75	0,000	4,97
Torque [Nm]	0,2689	0,0132	20,38	0,000	5,07
Tool wear [min]	0,01462	0,00136	10,77	0,000	1,11
Type					
L	0,617	0,314	1,96	0,050	3,73
M	0,177	0,338	0,52	0,601	3,73

I rapporti di probabilità servono per comprendere l’effetto delle variabili predittive. Per quanto riguarda le variabili continue, se i rapporti di probabilità sono maggiori di 1 è probabile che l’evento avvenga all’aumentare della variabile predittiva, se sono uguali a 1 la variabile predittiva non influenza la probabilità che l’evento avvenga, mentre se sono minori di 1 è probabile che l’evento non avvenga all’aumentare della variabile predittiva. In particolare, come si vede in figura 3.15 nel caso in questione all’aumentare di temperatura dell’aria, velocità rotazionale, coppia e usura dell’utensile la probabilità che avvenga la rottura della macchina aumenta, anche

se per velocità rotazionale e usura dell'utensile aumenta di poco. Invece, all'aumentare della temperatura del processo è probabile che l'evento non avvenga.

Per le variabili categoriali, invece, se il rapporto di probabilità è maggiore di 1 è più probabile che l'evento avvenga al livello A piuttosto che al livello B, se è minore di 1 il contrario, mentre se è uguale ad 1 la variabile predittiva non influenza la probabilità dell'evento. In questo caso, è più probabile che la rottura della macchina avvenga quando il prodotto è di tipo L anziché H, M anziché H e L anziché M, quindi l'evento si verifica più probabilmente se la variante del prodotto ha qualità bassa (L) e meno probabilmente se ha qualità elevata (H) [46].

Figura 3.15: rapporti di probabilità

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
Air temperature [K]	1,9668	(1,6722; 2,3132)
Process temperature [K]	0,5537	(0,4465; 0,6865)
Rotational speed [rpm]	1,0110	(1,0098; 1,0122)
Torque [Nm]	1,3085	(1,2751; 1,3428)
Tool wear [min]	1,0147	(1,0120; 1,0174)

Odds Ratios for Categorical Predictors

Level A	Level B	Odds Ratio	95% CI
Type			
L	H	1,8532	(1,0006; 3,4324)
M	H	1,1934	(0,6149; 2,3160)
M	L	0,6439	(0,4549; 0,9116)

Odds ratio for level A relative to level B

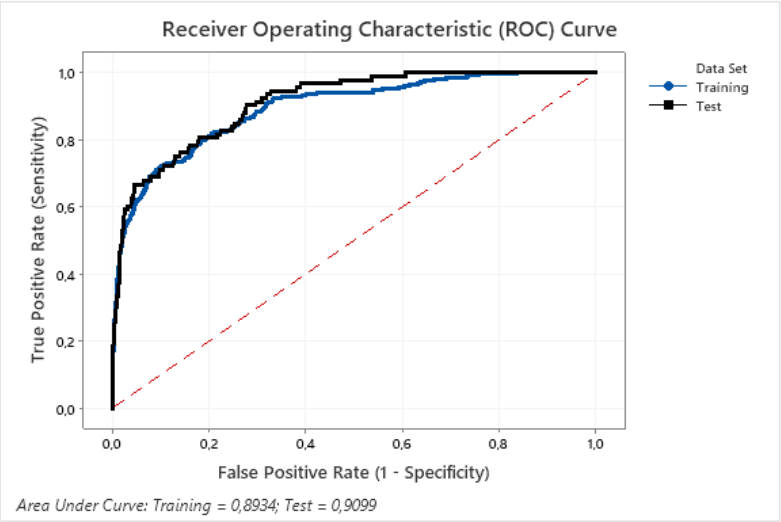
Il riassunto del modello in figura 3.16 riporta alcuni indicatori di performance. Sapendo che i valori di R^2 e di R^2 aggiustato possono variare da 0% a 100%, si può ritenere che siano entrambi molto bassi in quanto inferiori al 35%, mentre quello della parte di test è di poco superiore al 36%. Un indicatore più rilevante per il confronto con il modello studiato al sottoparagrafo 3.2.1 è l'AUC, di cui è riportata la curva ROC in figura 3.17. Pur essendo buono, è inferiore rispetto a quello del modello di classificazione CART; quindi, la performance di questo modello è peggiore.

Figura 3.16: indicatori di performance

Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC	AICc	BIC	Area Under ROC Curve	Test Deviance R-Sq	Test Area Under ROC Curve
34,66%	34,33%	1408,17	1408,19	1463,00	0,8934	36,23%	0,9099

Figura 3.17: curva ROC



3.2.3. Confronto tra i due metodi

Dalle analisi appena eseguite si conclude che il modello migliore è quello con il metodo di classificazione CART perché la sua curva ROC ha un'AUC maggiore. Inoltre, nella regressione logistica il coefficiente di determinazione R^2 è molto basso, indice che le capacità predittive del modello non sono buone, probabilmente perché le variabili che influenzano il guasto della macchina sono numerose.

Dal modello ottenuto con il metodo di classificazione CART si deduce che la probabilità di guasto è abbastanza bassa e le variabili che hanno maggior impatto sono: coppia, velocità di rotazione, usura dell'utensile e temperatura dell'aria. Inoltre, come si vede dagli indici e dalla curva ROC, il modello creato è molto attendibile.

3.3. Confronto tra diversi casi studio

Nel seguente paragrafo sono stati confrontati alcuni casi studio analizzati da diversi autori. Inoltre, è stato fatto un paragone anche con il caso studio al paragrafo 3.2.

3.3.1. Controllo qualità del calcestruzzo

Nello studio di Khodaparasti et al. (2023) vengono analizzati 1030 campioni di calcestruzzo e per ognuno di essi vengono rilevate nove caratteristiche, con l'obiettivo di determinarne la qualità in base alla sua resistenza a compressione. In realtà, il caso è stato studiato da diversi autori con l'utilizzo di diversi algoritmi di regressione, tra cui regressione lineare, vettoriale di supporto (SVR), con albero decisionale e con Random Forest. Dai valori di coefficiente di correlazione (R^2), errore medio assoluto (MAE) e radice dell'errore quadratico medio (RMSE), si è ottenuto che l'algoritmo migliore è quello di Random Forest, come mostra la figura 3.18. Pertanto, lo scopo dell'articolo è quello di creare un algoritmo migliorativo di Random Forest, obiettivo

raggiunto effettuando delle modifiche al metodo classico, ovvero nella fase di test al posto dell'indice di Gini sono stati considerati un coefficiente informativo e il peso degli alberi. Come si vede dalle figure 3.19 e 3.20, grazie a queste modifiche i parametri che indicano la performance del modello sono aumentati [15].

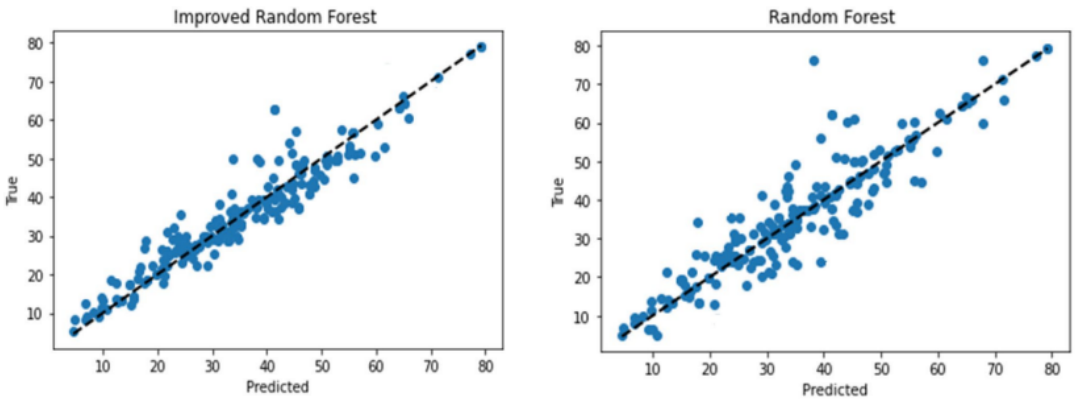
Figura 3.18: confronto tra diversi metodi di ML [15]

Algorithm	Performance parameters		
	R^2	MAE	RMSE
Linear regression	0.5809	8.2296	10.725
SVR(RBF Kernel)	0.4088	8.799	12.7373
Decision tree regressor	0.78	4.73	5.6
Random forest regressor	0.892	3.84	5.6

Figura 3.19: confronto tra Random Forest classico e modificato [15]

Algorithms	Performance parameters		
	R^2	MAE	RMSE
Random forest regressor	0.892	3.84	5.6
Improved random forest regressor	0.931	3.21	4.71

Figura 3.20: previsione della resistenza a compressione del calcestruzzo con i due metodi di Random Forest [15]



3.3.2. Ispezione di qualità di un separatore di batterie

In questo caso studio presentato da Huber et al. (2016) vengono raccolti 746 eventi per i quali possono verificarsi 8 tipologie di difetti. Lo studio consiste nell'applicazione della classificazione multiclasse per un'ispezione sulla qualità dei separatori di batterie e la tecnica di modellazione utilizzata è l'albero decisionale (figura 3.21). Per la valutazione del modello si utilizza una matrice di confusione (figura 3.22), dai cui valori si ricavano degli indici di performance molto buoni, superiori al 90% [13].

Figura 3.21: albero decisionale [13]

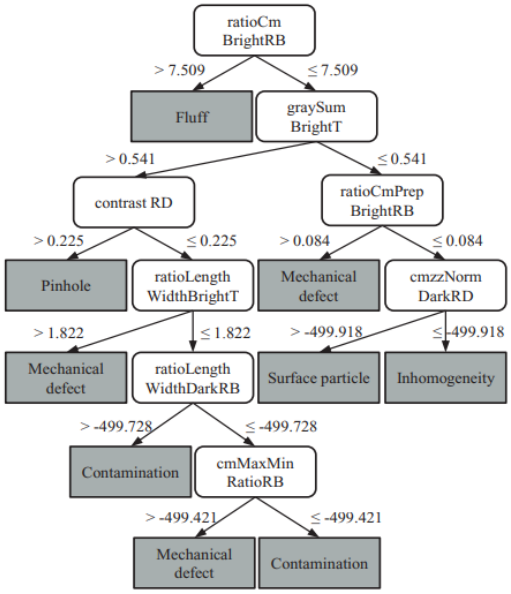


Figura 3.22: matrice di confusione con parametri e valori target [13]

Parameter	Criterion	Maximal depth	Confidence	Min. gain	Min. leaf size	Min. size for split	
Optimized setting	gain ratio	10	0,01	0	3	1	
	True mechanical	True contamination	True inhomogeneity	True particle	True fluff	True pinhole	Class precison
Predicted mechanical	72	4	1	0	1	0	92.31 %
Predicted contamination	2	96	0	0	0	4	94.12 %
Predicted inhomogeneity	0	0	21	2	0	0	91.30 %
Predicted particle	0	5	1	61	0	0	91.04 %
Predicted fluff	0	0	0	0	16	0	100.00 %
Predicted pinhole	0	1	0	0	0	40	97.56 %
Class sensitivity	97.30 %	90.57 %	91.30 %	96.83 %	94.12 %	90.91%	**accuracy: 93.58 %**

3.3.3. Qualità della saldatura

Il caso di Sumesh et al. (2015), invece, studia la qualità della saldatura tramite l'emissione di segnali acustici. Il test è stato eseguito su 20 campioni ed eseguite tutte le rilevazioni acustiche è stato impostato un albero decisionale. La fase di classificazione è stata eseguita utilizzando due diversi algoritmi, ovvero il J48 e l'algoritmo di Random Forest. Le matrici di confusione, mostrate in figura 3.23 e 3.24, permettono di calcolare l'efficienza di ciascuno di essi e si osserva che quella di Random Forest supera di quasi 20 punti percentuali quella dell'algoritmo J48 (figura 3.25) [30].

Figura 3.23: matrice di confusione algoritmo J48 [30]

a	b	c	Classified as
268	268	43	a=LF
34	779	51	b=Good
15	150	315	c=BT

Figura 3.24: matrice di confusione algoritmo di Random Forest [30]

a	b	c	Classified as
524	67	41	a=LF
18	761	26	b=Good
31	34	418	c=BT

Figura 3.25: efficienza dei due algoritmi utilizzati [30]

Algorithm	Classification Efficiency
J48	70.78
Random Forest Algorithm	88.69

3.3.4. Rilevamento anomalie macchinari automatizzati

Il caso di Bono et al. (2023) rileva le anomalie di macchinari automatizzati i cui parametri sono molto variabili, tramite un nuovo approccio basato sull'applicazione di un sensore soft e l'utilizzo di una rete neurale. La matrice di confusione permette di valutare l'accuratezza del modello ed in una tabella viene mostrato il confronto con altri metodi di risoluzione, ovvero analisi discriminante (DA), SVM, KNN, albero decisionale (EBT), rete neurale (NN). Si ricava che il metodo proposto è quello più accurato, con indici sempre superiori al 90%, come mostrato in figura 3.26 [4].

Figura 3.26: confronto indici con diversi metodi di risoluzione [4]

		DA	SVM	KNN	EBT	NN	PM
400	Accuracy	81,8%	97,5%	95,9%	94,8%	99,8%	98,4%
	Precision	87,1%	98,2%	96,9%	96,2%	100,0%	98,5%
	Recall	88,5%	98,4%	97,6%	96,9%	99,8%	99,5%
	F1score	87,8%	98,3%	97,3%	96,5%	99,9%	99,0%
500	Accuracy	39,7%	39,9%	40,2%	41,1%	57,4%	92,5%
	Precision	40,7%	41,1%	41,1%	41,5%	57,1%	92,5%
	Recall	89,9%	89,7%	90,2%	91,2%	96,8%	99,6%
	F1score	56,0%	56,3%	56,5%	57,1%	71,8%	95,9%
600	Accuracy	31,9%	31,4%	33,5%	35,1%	57,8%	92,2%
	Precision	34,8%	35,0%	36,9%	37,9%	56,4%	91,9%
	Recall	70,8%	69,8%	71,7%	73,4%	99,1%	99,9%
	F1score	46,7%	46,7%	48,7%	50,0%	71,9%	95,7%
800	Accuracy	82,9%	97,8%	95,3%	95,1%	99,8%	96,3%
	Precision	83,5%	98,4%	96,6%	97,1%	99,7%	95,5%
	Recall	95,9%	98,7%	97,1%	96,4%	100,0%	100,0%
	F1score	89,3%	98,6%	96,9%	96,8%	99,9%	97,7%

3.3.5. Confronto

In tabella 3.3 sono elencate le caratteristiche dei diversi casi confrontati e spiegati nei sottoparagrafi precedenti. Per i casi studio risolti con più metodi viene riportato in tabella soltanto il metodo che ha permesso di ottenere la performance migliore.

Tabella 3.3: confronto casi studio

Caso studio	Disponibilità dataset	Algoritmo	Metodo	Valutazione	Indicatori di performance
Paragrafo 3.2	Sì	Classificazione binaria	Albero decisionale	Matrice di confusione e curva ROC	Molto buoni
Controllo qualità calcestruzzo	Su richiesta	Regressione	Algoritmo migliorativo di Random Forest	R^2, MAE, RMSE	Molto buoni
Qualità separatore di batterie	No	Classificazione multiclasse	Albero decisionale	Matrice di confusione	Molto buoni
Qualità delle saldature	No	Classificazione multiclasse	Albero decisionale con algoritmo di Random Forest	Matrice di confusione	Buoni
Rilevamento anomalie macchinari automatizzati	No	Classificazione binaria	Sensore soft + rete neurale	Indicatori di performance	Buoni

Tutti i casi studio confrontati sono casi di apprendimento supervisionato, o di regressione o di classificazione. Il metodo di soluzione più utilizzato è l'albero decisionale, perché è il più semplice e permette di interpretare velocemente i dati e facilitare il processo decisionale. Tuttavia, all'aumentare del numero di nodi potrebbe diventare complesso e non essere il metodo con maggior efficienza. Inoltre, è considerato molto instabile perché una leggera variazione dei dati nel training set potrebbe generare un albero molto diverso [15]. Dall'analisi eseguita in questo paragrafo si osserva che non esiste un metodo migliore in assoluto, ma per ogni set di dati si deve individuare il metodo più adatto a seconda delle caratteristiche dei dati stessi. Spesso viene individuato a seguito di un confronto tra gli indicatori di performance, come accade nel caso studio di Bono et al. (2023), e talvolta il miglior metodo è dato da una combinazione o da un miglioramento di quelli testati, come nel caso di Khodaparasti et al. (2023). Per la fase di valutazione nei casi di classificazione in genere si utilizza la matrice di confusione, dalla quale si possono calcolare gli indicatori di performance del modello, riportati in tabella 3.3 come "molto buoni" se superiori al 90% e come "buoni" se compresi tra il 70% e il 90%. Nel caso della regressione, invece, gli indicatori utilizzati sono R^2, MAE e RMSE.

Conclusione

Questo lavoro di tesi aveva lo scopo di analizzare lo sviluppo e il cambiamento dei sistemi per il controllo della qualità nel corso del tempo, prestando maggior interesse ai metodi più recenti e attualmente in fase di studio. Si è osservato che il problema del controllo della qualità è sorto con l'aumento delle pretese dei consumatori e le aziende si sono dovute adattare alle loro richieste. Per questo motivo, nel corso del tempo, si è arrivati all'implementazione del metodo tradizionale del Six Sigma secondo l'approccio DMAIC, costituito da una sequenza di fasi piuttosto rigida che permette di tenere sotto controllo in tempo reale il processo produttivo e, di conseguenza, la qualità dei prodotti. Tuttavia, l'aumento della quantità di dati e lo sviluppo delle nuove tecnologie hanno reso necessario l'adattamento delle aziende a nuovi sistemi di controllo qualità. La diffusione di internet, cloud, tecnologie blockchain, machine learning e molti altri strumenti ha permesso di raggiungere obiettivi più ambiziosi, ovvero prevedere azioni future a partire da dati storici oppure ottenuti in tempo reale, accelerare i tempi di previsione, connettere tra loro le diverse funzioni aziendali e facilitare i processi decisionali. Attualmente è in rapida diffusione il machine learning, un metodo meno rigido e strutturato rispetto al Six Sigma, che attraverso gli algoritmi di regressione, classificazione e clustering ha la capacità di risolvere problemi di diverso tipo.

L'algoritmo di classificazione è stato applicato al caso studio al paragrafo 3.2, che aveva lo scopo di studiare la qualità di una fresatrice in funzione di diverse variabili. Dall'analisi, eseguita tramite classificazione CART e regressione logistica binaria, è emerso che coppia, velocità rotazionale, usura dell'utensile e temperatura dell'aria sono le variabili più influenti nella qualità della fresatrice. Da un confronto tra i due metodi si è osservato che la regressione logistica in questo caso fornisce risultati meno affidabili dell'albero decisionale, probabilmente a causa di un elevato numero di variabili predittive. Tuttavia, consente di esaminare separatamente l'effetto della variabile categoriale "tipo" e dai rapporti di probabilità si nota che il guasto della macchina avviene più probabilmente quando la variante di qualità del prodotto è bassa e meno probabilmente quando è alta. L'albero decisionale, invece, permette di classificare in modo immediato le osservazioni rilevate e definire rapidamente la probabilità di guasto di ogni rilevazione.

Il confronto tra diversi casi studio in cui sono applicati algoritmi di apprendimento supervisionato dimostra che non esiste una sequenza prestabilita di operazioni da svolgere per eseguire l'analisi, ma spesso diversi metodi vengono testati o combinati tra loro per ottenere la soluzione migliore. La valutazione in genere avviene attraverso un confronto degli indicatori di accuratezza, sensitività, precisione e punteggio F e permette di stabilire quale modello è il più valido. Anche la curva ROC è uno strumento importante in questa fase, infatti semplicemente confrontando l'area sotto la curva di diversi modelli si riesce ad individuare il migliore.

Infine, si osserva che nell'apprendimento supervisionato il metodo più utilizzato è quello dell'albero decisionale perché, se il numero di nodi non è troppo elevato, consente di ottenere risultati semplici da interpretare e può essere facilmente utilizzato per eseguire previsioni.
Potrebbe essere interessante, in eventuali sviluppi futuri di questo lavoro, impostare ulteriori modelli per la risoluzione del caso studio al paragrafo 3.2 per scoprire se esiste un metodo che permette di ottenere risultati migliori rispetto all'albero decisionale. Inoltre, in un'ottica futura di Industria 5.0, orientata alla sostenibilità socio-ambientale e alla sempre maggiore personalizzazione dei prodotti, sarebbe curioso scoprire come si evolveranno i sistemi per il controllo della qualità e se gli algoritmi di machine learning saranno uno strumento sufficiente per soddisfare le esigenze di questa nuova rivoluzione.

Bibliografia

[1] Albers, A., Gladysz, B., Pinner, T., Butenko, V., & Stürmlinger, T. (2016). Procedure for defining the system of objectives in the initial phase of an industry 4.0 project focusing on Intelligent Quality Control Systems. *Procedia CIRP*, *52*, 262–267. https://doi.org/10.1016/j.procir.2016.07.067

[2] Allen, T. T. (2019). In *Introduction to engineering statistics and Lean Six Sigma: Statistical Quality Control and design of experiments and systems* (pp. 12–20, 42, 63-67, 86, 96, 108-110, 130, 141-144, 162, 170-173, 180-184, 189, 279-280, 291, 398, 403, 407, 411). Springer.

[3] Al-Refaie, A., & Bata, N. (2010). Evaluating measurement and process capabilities by GR&R with four quality measures. *Measurement*, *43*(6), 842–851. https://doi.org/10.1016/j.measurement.2010.02.016

[4] Bono, F. M., Radicioni, L., & Cinquemani, S. (2023). A novel approach for quality control of automated production lines working under highly inconsistent conditions. *Engineering Applications of Artificial Intelligence*, *122*, 106149. https://doi.org/10.1016/j.engappai.2023.106149

[5] Bottani, E., Montanari, R., Volpi, A., & Tebaldi, L. (2023). Statistical Process Control of assembly lines in manufacturing. *Journal of Industrial Information Integration*, *32*, 100435. https://doi.org/10.1016/j.jii.2023.100435

[6] Bueno, A., Godinho Filho, M., & Frank, A. G. (2020). Smart production planning and control in the industry 4.0 context: A systematic literature review. *Computers & Industrial Engineering*, *149*, 106774. https://doi.org/10.1016/j.cie.2020.106774

[7] Dharmik, R. C., & Bawankar, B. U. (2023). Design challenges for machine/Deep Learning Algorithms. *Machine Learning for VLSI Chip Design*, 195–209. https://doi.org/10.1002/9781119910497.ch13

[8] Escobar, C. A., & Morales-Menendez, R. (2018a). Machine learning techniques for quality control in high conformance manufacturing environment. *Advances in Mechanical Engineering*, *10*(2), 168781401875551. https://doi.org/10.1177/1687814018755519

[9] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

[10] Forza, C. (2021). In *Lezioni di Qualità e Organizzazione* (pp. 41-42, 48, 64-70, 94, 101-106, 201–204). Progetto.

[11] Gu, J., Zhao, L., Yue, X., Arshad, N. I., & Mohamad, U. H. (2023). Multistage quality control in manufacturing process using blockchain with machine learning technique. *Information Processing & Management*, *60*(4), 103341. https://doi.org/10.1016/j.ipm.2023.103341

[12] Hakimi, S., Zahraee, S. M., & Mohd Rohani, J. (2018). Application of Six sigma DMAIC methodology in plain yogurt production process. *International Journal of Lean Six Sigma*, *9*(4), 562–578. https://doi.org/10.1108/ijlss-11-2016-0069

[13] Huber, J., Tammer, C., Krotil, S., Waidmann, S., Hao, X., Seidel, C., & Reinhart, G. (2016a). Method for classification of battery separator defects using optical inspection. *Procedia CIRP*, *57*, 585–590. https://doi.org/10.1016/j.procir.2016.11.101

[14] Isaksson, A. J., Harjunkoski, I., & Sand, G. (2018). The impact of digitalization on the future of Control and Operations. *Computers & Chemical Engineering*, *114*, 122–129. https://doi.org/10.1016/j.compchemeng.2017.10.037

[15] Khodaparasti, M., Alijamaat, A., & Pouraminian, M. (2023). Prediction of the concrete compressive strength using improved random forest algorithm. *Journal of Building Pathology and Rehabilitation*, *8*(2). https://doi.org/10.1007/s41024-023-00337-8

[16] Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review*, *26*(3), 159–190. https://doi.org/10.1007/s10462-007-9052-3

[17] Linderman, K., Schroeder, R. G., Zaheer, S., & Choo, A. S. (2002). Six sigma: A goal-theoretic perspective. *Journal of Operations Management*, *21*(2), 193–203. https://doi.org/10.1016/s0272-6963(02)00087-6

[18] Lindström, V., Persson, F., Viswanathan, A. P., & Rajendran, M. (2023). Data quality issues in production planning and Control – Linkages to smart PPC. *Computers in Industry*, *147*, 103871. https://doi.org/10.1016/j.compind.2023.103871

[19] Liu, H., Zhou, M., Lu, X. S., & Yao, C. (2018). Weighted gini index feature selection method for imbalanced data. *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)*. https://doi.org/10.1109/icnsc.2018.8361371

[20] Marangoni, G. (2018). *Mathematical Programming and Economic Analysis* (pp. 63-64, 81). Università della Svizzera italiana.

[21] Mikulová, P., & Plura, J. (2018). Comparison of approaches to gauge repeatability and reproducibility analysis. *MATEC Web of Conferences*, *183*, 03015. https://doi.org/10.1051/matecconf/201818303015

[22] Oluyisola, O. E., Bhalla, S., Sgarbossa, F., & Strandhagen, J. O. (2021a). Designing and developing smart production planning and control systems in the Industry 4.0 ERA: A methodology and case study. *Journal of Intelligent Manufacturing*, *33*(1), 311–332. https://doi.org/10.1007/s10845-021-01808-w

[23] Park, S., & Ha, C. (2020). Determination of optimal experimental design for ANOVA gauge R&R using Stochastic Programming. *Measurement*, *156*, 107612. https://doi.org/10.1016/j.measurement.2020.107612

[24] Pereira, R. B., Peruchi, R. S., de Paiva, A. P., da Costa, S. C., & Ferreira, J. R. (2016). Combining Scott-Knott and GR&R methods to identify special causes of variation. *Measurement*, *82*, 135–144. https://doi.org/10.1016/j.measurement.2015.12.033

[25] Ramasamy, A., & Chowdhury, S. (2020). Big Data Quality Dimensions: A systematic literature review. *Journal of Information Systems and Technology Management*. https://doi.org/10.4301/s1807-1775202017003

[26] S. Matzka, "Explainable Artificial Intelligence for Predictive Maintenance Applications," 2020 Third International Conference on Artificial Intelligence for Industries (AI4I), 2020, pp. 69-74, doi: 10.1109/AI4I49448.2020.00023

[27] Sahoo, S. (2019). Assessment of TPM and TQM practices on Business Performance: A multi-sector analysis. *Journal of Quality in Maintenance Engineering*, *25*(3), 412–434. https://doi.org/10.1108/jqme-06-2018-0048

[28] Sánchez-Rebull, M.-V., Ferrer-Rullan, R., Hernández-Lara, A.-B., & Niñerola, A. (2020). Six sigma for improving cash flow deficit: A case study in the food can manufacturing industry. *International Journal of Lean Six Sigma*, *11*(6), 1105–1126. https://doi.org/10.1108/ijlss-12-2018-0137

[29] Schroeder, R. G., Linderman, K., Liedtke, C., & Choo, A. S. (2007). Six sigma: Definition and underlying theory⋆. *Journal of Operations Management*, *26*(4), 536–554. https://doi.org/10.1016/j.jom.2007.06.007

[30] Sumesh, A., Rameshkumar, K., Mohandas, K., & Babu, R. S. (2015). Use of machine learning algorithms for Weld Quality Monitoring using acoustic signature. *Procedia Computer Science*, *50*, 316–322. https://doi.org/10.1016/j.procs.2015.04.042

[31] Thakur, V., Anthony Akerele, O., Brake, N., Wiscombe, M., Broderick, S., Campbell, E., & Randell, E. (2023). Use of a lean six sigma approach to investigate excessive quality control (QC) material use and resulting costs. *Clinical Biochemistry*, *112*, 53–60. https://doi.org/10.1016/j.clinbiochem.2022.12.001

[32] Usuga Cadavid, J. P., Lamouri, S., Grabot, B., Pellerin, R., & Fortin, A. (2020a). Machine learning applied in production planning and Control: A state-of-the-art in the era of industry 4.0. *Journal of Intelligent Manufacturing*, *31*(6), 1531–1558. https://doi.org/10.1007/s10845-019-01531-7

[33] Weckenmann, A., Akkasoglu, G., & Werner, T. (2015a). Quality management – history and trends. *The TQM Journal*, *27*(3), 281–293. https://doi.org/10.1108/tqm-11-2013-0125

[34] Zanobini, A., Sereni, B., Catelani, M., & Ciani, L. (2016). Repeatability and reproducibility techniques for the analysis of measurement systems. *Measurement*, *86*, 125–132. https://doi.org/10.1016/j.measurement.2016.02.041

Sitografia

[35] *AUC-Roc Curve in machine learning - javatpoint*. www.javatpoint.com. (n.d.-a). https://www.javatpoint.com/auc-roc-curve-in-machine-learning

[36] *Classification algorithm in Machine Learning - Javatpoint*. www.javatpoint.com. (n.d.-a). https://www.javatpoint.com/classification-algorithm-in-machine-learning

[37] *Clustering in machine learning - javatpoint*. www.javatpoint.com. (n.d.-b). https://www.javatpoint.com/clustering-in-machine-learning

[38] *Confusion matrix in machine learning - javatpoint*. www.javatpoint.com. (n.d.-c). https://www.javatpoint.com/confusion-matrix-in-machine-learning

[39] *Data Considerations for CART® Classification*. Minitab. (n.d.). https://support.minitab.com/en-us/minitab/20/help-and-how-to/statistical-modeling/predictive-analytics/how-to/cart-classification/before-you-start/data-considerations/

[40] *Decision tree algorithm in Machine Learning - Javatpoint*. www.javatpoint.com. (n.d.-c). https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

[41] *K-means clustering algorithm - javatpoint*. www.javatpoint.com. (n.d.-c). https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

[42] *K-Nearest Neighbor (KNN) algorithm for Machine Learning - Javatpoint*. www.javatpoint.com. (n.d.-b). https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

[43] *Linear regression in machine learning - javatpoint*. www.javatpoint.com. (n.d.). https://www.javatpoint.com/linear-regression-in-machine-learning

[44] *Logistic regression in machine learning - javatpoint*. www.javatpoint.com. (n.d.-c). https://www.javatpoint.com/logistic-regression-in-machine-learning

[45] Matzka, S. (2022, November 6). *Predictive maintenance dataset (AI4I 2020)*. Kaggle. https://www.kaggle.com/datasets/stephanmatzka/predictive-maintenance-dataset-ai4i-2020

[46] *Odds ratios for fit binary logistic model*. Minitab. (n.d.). https://support.minitab.com/en-us/minitab/21/help-and-how-to/statistical-modeling/regression/how-to/fit-binary-logistic-model/interpret-the-results/all-statistics-and-graphs/odds-ratios/

[47] *Support Vector Machine (SVM) algorithm - javatpoint*. www.javatpoint.com. (n.d.-e). https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm