POLITECNICO DI TORINO

Corso di Laurea Magistrale in Ingegneria Gestionale

Progettazione e sviluppo di una metodologia semi-supervisionata per caratterizzare i cicli di produzione nel contesto dell'Industria 4.0

Relatore
Prof.Tania CERQUITELLI

Candidato
Ylenia FERLAZZO

Correlatori
Dott.Paolo BETHAZ
Dott.Riccardo CALLÀ

Luglio 2020

Ringraziamenti

Un ringraziamento speciale va alla mia famiglia, per avermi dato la possibilità di raggiungere questo traguardo e per avermi supportato.
A Fabio, per essere stato sempre presente, soprattutto nei momenti complicati.
Alla Professoressa Cerquitelli, Paolo Bethaz e Riccardo Callà per la disponibilità, la pazienza, il supporto e gli utili consigli.
A tutti coloro che sono stati presenti, in un modo o nell'altro, e mi hanno accompagnato lungo questo percorso straordinario.

Indice

Elenco delle tabelle

Elenco delle figure

Capitolo 1

Introduzione

Nell'era dell'Industria 4.0 si assiste ad un processo di trasformazione digitale che ha permesso di realizzare un ambiente manifatturiero sempre più connesso, integrando dati, persone, sistemi e risorse industriali. Gli obiettivi finali che si intende raggiungere con l'Industria 4.0 sono: l'ottimizzazione della produzione, ottenuta riducendo le inefficienze e aumentando il livello di personalizzazione, assistenza e interazione con il cliente finale; il miglioramento delle condizioni di lavoro; la creazione di nuovi modelli di business. Alla base delle innovazioni tecnologiche portanti di questa rivoluzione vi è il sistema cyber-fisico il quale, poichè mette in comunicazione le diverse parti di uno stabilimento, rappresenta il ponte che collega ambiente fisico e digitale. L'interconnessione è realizzata mediante sensori posizionati sulle macchine, grazie ai quali si raccoglie un'enorme quantità di dati che, per la loro dimensione ed eterogeneità, vengono definiti Big Data. Una delle applicazioni principali dei Big Data nel contesto dell'Industria 4.0 è la manutenzione predittiva. Grazie al monitoraggio continuo di macchinari e attrezzature, la strategia di manutenzione predittiva permette di effettuare gli interventi in base alle reali condizioni dei macchinari: ciò consente di supportare il processo di *decision making*, aumentare la produttività e ridurre gli sprechi. Le analisi esposte in questo elaborato sono realizzate sulla base di dati riguardo alla corrente assorbita dal motore di un braccio robotico in ciascun ciclo di produzione. L'obiettivo delle analisi è offrire un supporto all'attività di manutenzione, monitorando e prevedendo il corretto livello di tensione della cinghia di trasmissione del motore. Infatti, un errato tensionamento di questo organo produce problematiche che hanno delle ripercussioni sulla bontà della lavorazione. Se la tensione è superiore al livello normale si possono verificare dei surriscaldamenti che rischiano di danneggiare la cinghia stessa; se la tensione è inferiore a quella normale, invece, si possono determinare slittamenti che conducono ad un'usura prematura dell'organo.

I cicli di produzione sono forniti da un'azienda italiana leader nel settore metalmeccanico e, per ognuno di essi, è noto il livello di tensione della cinghia.

Nell'elaborato viene presentata una metodologia semi-supervisionata per etichettare

correttamente i cicli di lavorazione del robot in questione. Dopo aver preparato opportunamente i dati, sono applicati tre diversi algoritmi di Clustering. Inizialmente si selezionano i migliori parametri di input per ciascuna tecnica utilizzando dei metodi data-driven che sfruttano l'analisi della distribuzione dei dati; in seguito, si validano i risultati ottenuti confrontandoli con le etichette reali di appartenenza di ciascun ciclo di produzione.

L'elaborato è organizzato in sei capitoli, di cui il primo è una breve introduzione alle tematiche esposte nel seguito.

- Capitolo 2: in questo capitolo viene fatta una panoramica sul concetto di Industria 4.0, descrivendo brevemente i pilastri tecnologici che supportano le innovazioni apportate dalla rivoluzione. In seguito, è presentato un quadro sulla diffusione dei principi dell'Industria 4.0 nei principali leader mondiali nel settore manifatturiero, descrivendo più approfonditamente gli interventi adottati dall'Italia per allinearsi alla rivoluzione.

- Capitolo 3: in questo capitolo si introduce il tema della *data analytics* e si espongono i vantaggi che ottengono le aziende leader in questo settore. Inoltre, si descriverà la tipica *pipeline* che segue la maggior parte dei processi di analisi dei dati, che prende il nome di Knowledge Discovery from Data o KDD e il cui obiettivo finale è l'estrazione della conoscenza. In particolare, si farà una panoramica sulle principali tecniche utilizzate per ricavare informazioni dai dati, ovvero Clustering, Classificazione e Concept Drift.

- Capitolo 4: in questo capitolo sono descritti più dettagliatamente gli algoritmi di Clustering, a partire dalla logica di funzionamento e selezione dei migliori parametri di input fino alla valutazione delle loro performance.

- Capitolo 5: questo capitolo contiene innanzitutto una descrizione del caso di studio, delineando più approfonditamente la tematica, la composizione dei dataset utilizzati e le operazioni di preparazione dei dati alle successive analisi. In seguito, sono presentati i risultati delle tecniche applicate.

- Capitolo 6: l'elaborato si conclude con un breve capitolo in cui sono discussi i principali risultati ottenuti dalle analisi e sono esposti dei possibili sviluppi futuri di questa metodologia.

Capitolo 2

Industria 4.0

2.1 Origini e storia dell'Industria 4.0

2.1.1 Origini del termine Industria 4.0

Il termine "Industria 4.0" è stato utilizzato per la prima volta in Germania nel 2011 durante la Fiera di Hannover. In questa occasione, un gruppo di lavoro composto da rappresentanti dell'Università, della politica e del mondo del business ha annunciato un progetto per lo sviluppo del settore manifatturiero tedesco, lo "Zukunftsprojekt Industrie 4.0". Il governo ha supportato quest'idea e l'ha resa parte integrante del progetto "Hi-Tech Strategy 2020 for Germany", puntando a promuovere la Germania come leader nell'innovazione tecnologica. Nell'aprile 2013, sempre alla Fiera di Hannover, è stato diffuso il report finale contenente un piano per gli investimenti futuri su un ampio spettro di aree, come infrastrutture, scuole, sistemi energetici, enti di ricerca e aziende. L'obiettivo del piano era rimodernare il sistema produttivo tedesco e rendere il settore secondario competitivo a livello globale. In seguito, il modello tedesco ha ispirato numerose iniziative europee e il termine Industria 4.0 si è diffuso anche a livello internazionale. Tuttavia, l'Industria 4.0 è un concetto ancora in movimento, in cui gli standard implementativi e le definizioni sono in corso di evoluzione. [1]

2.1.2 Storia dell'Industria 4.0

In campo tecnico-scientifico si è già assistito a tre rivoluzioni industriali. [2].

La Prima Rivoluzione Industriale, avvenuta alla fine del 18° secolo, si è caratterizzata per l'introduzione in campo tecnologico della macchina a vapore. La principale problematica riscontrata nella produzione di beni era legata al vincolo energetico: infatti, fino a poco tempo prima, si disponeva solamente del lavoro manuale degli operai. Questo periodo si contraddistingue per l'utilizzo dell'energia idrica legata ai mulini ad acqua e a vento e all'impiego del carbone. I settori che hanno beneficiato principalmente delle

innovazioni apportate e delle nuove risorse energetiche utilizzate sono stati l'industria tessile e meccanica, l'industria del trattamento dell'acciaio, l'estrazione mineraria e il settore dei trasporti.

La Seconda Rivoluzione Industriale, avvenuta all'inizio del 20° secolo tra il 1890 e il 1960, è vista come il periodo in cui l'elettricità e il petrolio hanno contribuito alla nascita dei nastri trasportatori e alla diffusione della produzione di massa, a cui sono seguite gli esempi di produzione fordista (da Henry Ford) e Di Frederick Taylor. Tra i settori maggiormente interessati nel corso della rivoluzione rientrano le industrie chimiche ed elettrotecniche.

La Terza Rivoluzione Industriale, invece, inizia negli anni '70 ed è legata allo sviluppo delle conoscenze nate nel contesto militare durante le guerre mondiali, alle sempre più stabili condizioni economiche dei Paesi occidentali e alla scoperta dell'energia atomica. In questo contesto, si assiste all'introduzione di tecnologie attinenti al campo dell'elettronica (i transistor) e dell'informatica (i personal computer e la rete Internet) e all'automazione dei processi produttivi.

Oggi si assiste alla Quarta Rivoluzione Industriale, caratterizzata dall'ascesa di robot autonomi, sistemi cyber-fisici, Internet of Things ed altri elementi che verranno discussi più approfonditamente nel seguito. Nel corso della Quarta Rivoluzione Industriale si darà vita alle cosiddette "Smart Factory", ovvero un nuovo modello di fabbrica concepito come rete di "oggetti intelligenti" in cui la presenza umana è ridotta al minimo, grazie al massiccio ricorso all'automazione basata sull'impiego dei sistemi cyber–fisici (CPS). Le macchine di una "Smart Factory" sono in grado di auto-ottimizzarsi, di auto-configurarsi e di sfruttare l'intelligenza artificiale per svolgere compiti complessi consentendo allo stesso tempo di ridurre i costi e aumentare la qualità del prodotto.

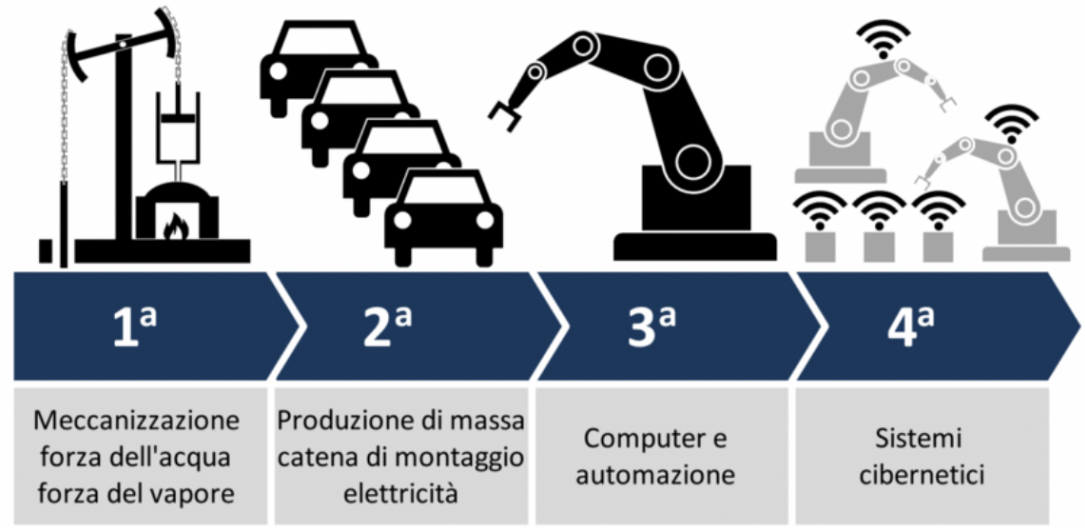


Figura 2.1: Le quattro rivoluzioni industriali. Fonte: AllAboutLean.com

2.2 Tecnologie abilitanti dell'industria 4.0

Come già sottolineato, l'Industria 4.0 è un concetto ancora in corso di evoluzione non recepito allo stesso modo da tutti i paesi mondiali, sia per quanto riguarda gli standard implementativi sia per le aree in cui effeture nuovi investimenti. In linea generale, l'Industria 4.0 si riferisce alle modalità organizzative della produzione di beni e servizi sfruttando l'integrazione tra macchinari e processi. Tale risultato non potrebbe essere raggiunto senza l'adozione di tecnologie in grado di permettere l'interconnessione tra i vari elementi del processo produttivo, in primis il sistema cyber-fisico. Da uno studio condotto da Boston Consulting Group, emerge che l'Industria 4.0 si fonda sulla convergenza e sull'applicazione di nove tecnologie abilitanti, elencate in Figura 2.2.

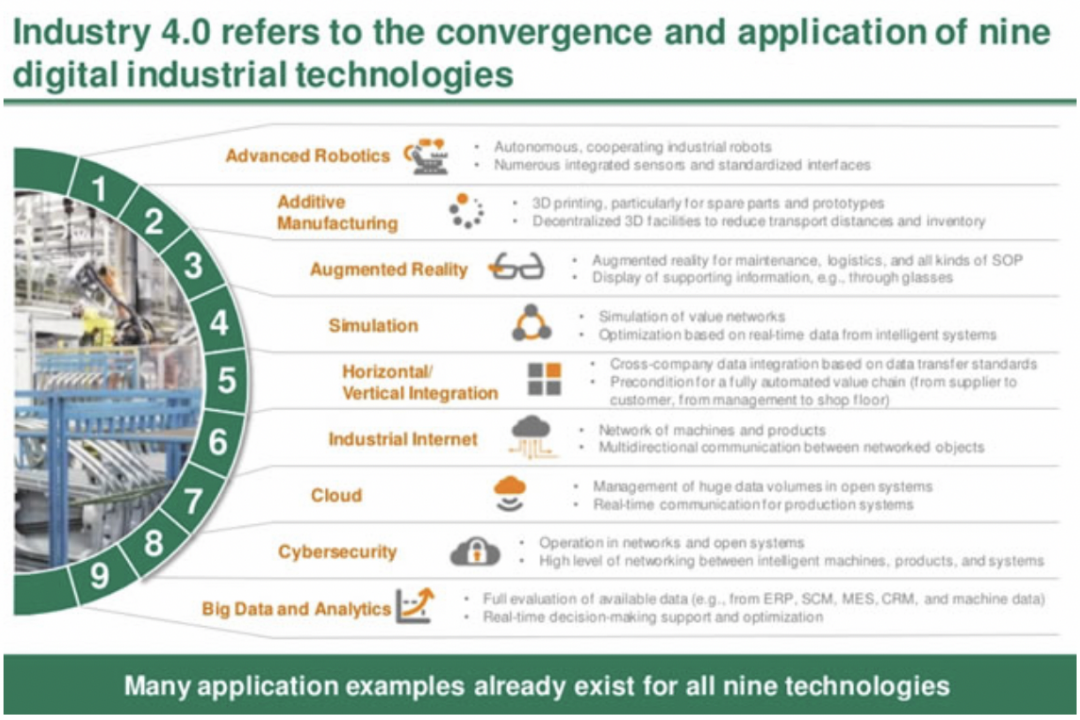


Figura 2.2: Le tecnologie abilitanti dell'Industria 4.0. Fonte: Boston Consulting Group

2.2.1 Sistema Cyber-Fisico

Il sistema cyber-fisico consente di combinare ambiente fisico e mondo digitale. Secondo Edward A. Lee, professore dell'Università della California, Berkeley, il sistema cyber-fisico o CPS è definito come "l'integrazione di calcolo e processi fisici. I computer e le reti controllano i processi, generalmente con circuiti di feedback in cui i processi fisici influenzano i calcoli e viceversa". Sostanzialmente, un sistema cyber-fisico è costituito da una combinazione di oggetti, componenti fisici e sistemi intelligenti che si collegano attraverso

la rete internet e permettono di far comunicare le diverse parti di un sistema.

La prima generazione dei sistemi cyber-fisici è caratterizzata dai RFID (Radio-Frequency Identification), una tecnologia basata sulla propagazione di onde elettromagnetiche in grado di consentire l'identificazione univoca e automatica di oggetti distanti, sia statici che in movimento. Le informazioni raccolte da questi dispositivi necessitano di un servizio esterno centralizzato per poter essere elaborate ed immagazzinate. [3]

La seconda generazione dei sistemi cyber-fisici è costituita da sensori e attuatori. In questa fase tali oggetti sono dotati di funzionalità quali la capacità di immagazzinamento ed elaborazione, ma in misura limitata.

I sistemi cyber-fisici di terza generazione, invece, sono equipaggiati con sensori e attuatori altamente innovativi in grado di elaborare, analizzare e immagazzinare i dati raccolti. Inoltre, i sistemi cyber-fisici sono dotati di un indirizzo IP (Internet Protocol) univoco: ciò significa che possono collegarsi alla rete Internet e, grazie ad essa, permettere alle parti connesse di comunicare tra loro.

Una volta definiti gli elementi costitutivi di un sistema cyber-fisico, in letteratura sono stati proposte delle linee guida su come dovrebbe essere strutturato un CPS al fine di facilitarne l'implementazione. [4]. Generalmente, un sistema cyber-fisico deve essere in grado di supportare due principali funzionalità:

- Acquisizione di dati real-time durante il processo produttivo e invio di feedback all'intero sistema sul funzionamento e sulle condizioni di lavoro;
- Capacità di elaborare, analizzare e immagazzinare i dati raccolti dal sistema fisico.

La struttura gerarchica proposta è chiamata anche architettura 5C ed è costituita da 5 livelli. Figura 2.3.

Alla base della piramide si trova il blocco di **Smart Connection**, il cui compito consiste nell'acquisire correttamente e in modo affidabile i dati dai macchinari. A questo scopo, è necessario disporre di una rete di sensori opportunamente selezionati e configurati.

Il secondo livello è chiamato **Data-to-Information Conversion** ed è in questa fase che avviene la trasformazione dei dati in informazioni significative utili per il business. In particolare, sono utilizzati algoritmi per valutare lo stato di salute delle attrezzature e per stimare la vita utile rimante dei macchinari.

Le informazioni raccolte da ogni macchinario nel blocco Data-to-Information Conversion sono utilizzate nel gradino successivo della piramide che prende il nome di **Cyber Level**.

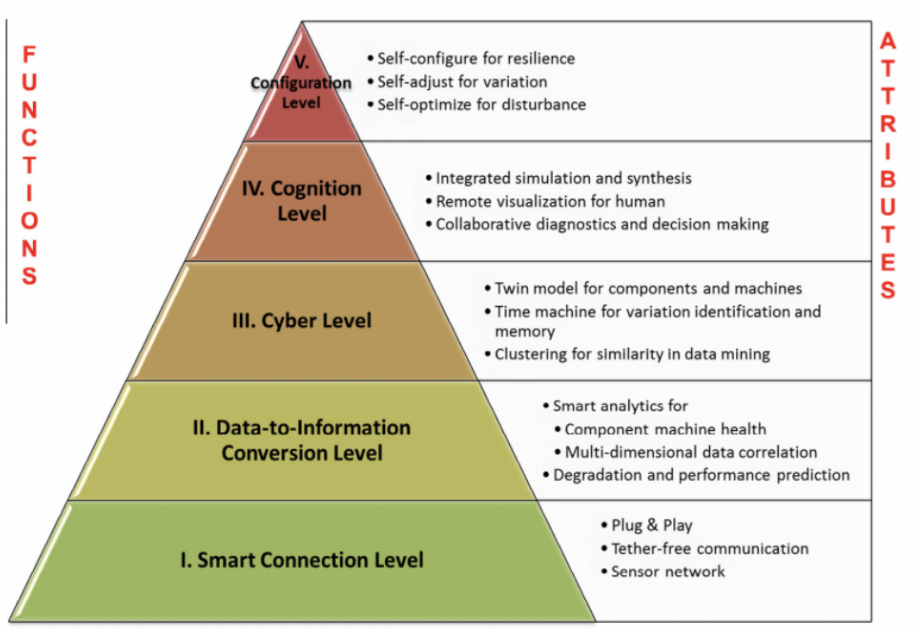


Figura 2.3: Architettura 5C per l'implementazione di un sistema cyber-fisico [4]

Dopo avere generato una grande quantità di informazioni, vengono effettuate delle analisi più approfondite per aggiungere ulteriore conoscenza sulle condizioni dei macchinari. In particolare, le analisi eseguite riguardano il confronto tra le prestazioni di un macchinario con gli altri e il paragone tra le performance attuali e storiche delle attrezzature allo scopo di predire il loro comportamento futuro.

Una volta raccolti i risultati delle analisi, questi possono essere utilizzati per monitorare il sistema con un'ottica a più ampio spettro nel blocco chiamato **Cognition**. Infatti, una volta ottenuta una panoramica sulle condizioni di macchinari e attrezzature è possibile ottimizzare i compiti più prioritari. Tuttavia, i risultati devono essere presentati appropriatamente agli utenti tramite info-grafici allo scopo di facilitare il processo decisionale.

L'ultimo livello della piramide chiamato **Configuration Level** rappresenta i feedback inviati dall'ambiente informatico a quello fisico e funge da controllo di supervisione sulle macchine auto-configurate e auto-adattive. In questa fase si interviene sulle decisioni correttive o preventive prese nello stadio precedente di Cognition.

Nella Figura 2.4 sono schematizzate per ciascun livello dell'architettura 5C le applicazioni e le tecniche utilizzate.

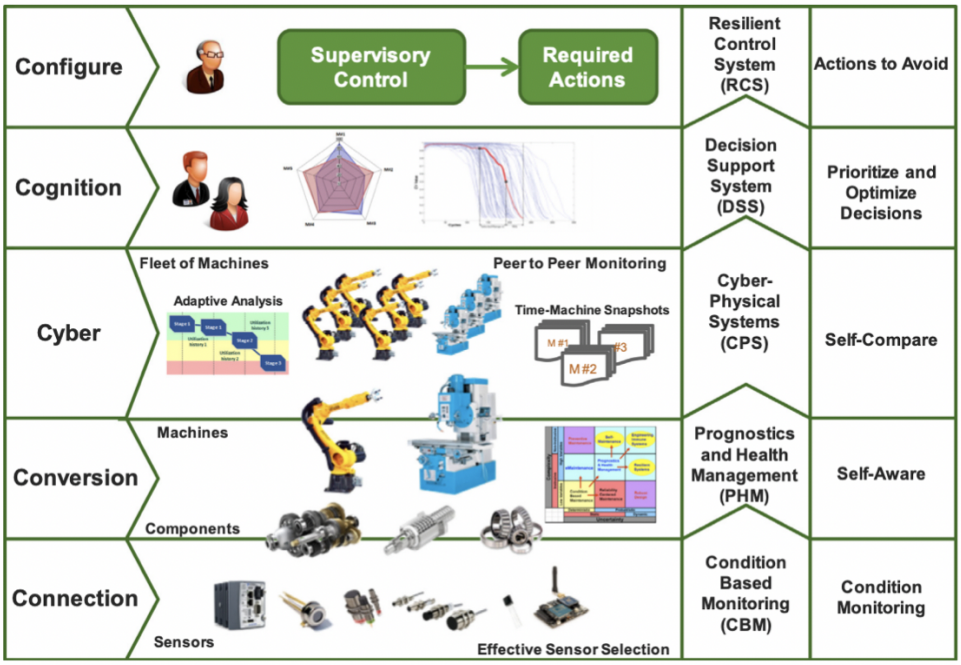


Figura 2.4: Applicazioni e tecniche associate a ciascun livello dell'architettura 5C [4]

I sistemi cyber-fisici sono elementi alla base dell'Industria 4.0 e consentono di usufruire di funzionalità aggiuntive che aumentano la produttività e riducono le inefficienze. Tra queste rientrano:

- Monitoraggio dell'integrità delle condizioni di un impianto o macchinario;
- Diagnosi sullo stato di salute di un macchinario;
- Servizi in remoto di monitoraggio, controllo e diagnosi;
- Tracciabilità dei processi.

2.2.2 Advanced Robotics

Uno dei pilastri dell'Industria 4.0 è l'automazione, ottenuta dai cosiddetti robot collaborativi o *cobot*. I robot sono in grado di svolgere compiti in modo affidabile, diventando sempre più autonomi, flessibili e versatili. In passato, i robot industriali erano così ingombranti dal punto di vista dello spazio fisico occupato tanto che veniva dedicata loro un'intera isola di lavoro; oggi, invece, sono perfettamente integrati nell'area di impiego. Gli ultimi sviluppi tecnologici hanno come protagonisti robot sempre più intelligenti e in grado di realizzare connessioni grazie ai sensori applicati sulle macchine. I robot possono essere applicati in numerose aree aziendali, a partire dalla manifattura, fino alla logistica e al management. Inoltre, i robot possono essere controllati da remoto: se dovesse verificarsi qualche criticità non è più necessario essere fisicamente presenti per riconfigurarlo, ma è

possibile fornire istruzioni sul funzionamento anche da un altro luogo inviando i parametri per il settaggio tramite lo smartphone. [5]

2.2.3 Additive Manufacturing

L'Additive Manufacturing (o stampa 3D) consente la produzione di un oggetto fisico attraverso la sovrapposizione (layer-by-layer) di strati di materiale sulla base dei dati di un progetto in 3D. In passato, le tecnologie additive erano utilizzate solamente nella fase di prototipazione; oggi, invece, grazie all'innovazione tecnologica che ha reso il processo di stampa 3D sempre più rapido, questa modalità di produzione è impiegata anche in altre fasi del processo. A differenza del processo di produzione sottrattiva, nel quale le macchine CNC (Controllo Numerico Centralizzato) producono un pezzo asportando materiale, la stampa 3D avviene senza l'impiego di attrezzature convenzionali: con l'Additive Manufactuting è sufficiente inviare al sistema un file digitale di un progetto in 3D. Rispetto alla produzione tradizionale, l'Additive Manufacturing consente di:

- Realizzare un prodotto fisico personalizzato, senza vincoli di tipo geometrico né legati ai macchinari convenzionali, che richiedono grandi costi fissi;
- Adattarsi ai cambiamenti di progetto rapidamente, senza vincoli produttivi e senza sostenere costi aggiuntivi.

Questa tipologia di produzione presenta, quindi, grande flessibilità, consentendo all'impresa di garantire un elevato grado di personalizzazione senza sostenere costi di produzione enormi: ciò permette all'azienda di rispondere alle esigenze del cliente rapidamente. Inoltre, la produzione 3D consente di ridurre ulteriormente i costi di produzione in quanto vi è un impiego minore di materiali e l'assenza di sprechi rispetto al processo di produzione sottrattiva. In questa modalità di realizzazione infatti, la lavorazione inizia da un blocco più grande e si asporta del materiale - tramite processi di foratura, alesatura, truciolatura - fino a quando si ottiene il pezzo desiderato. Infine, la produzione additiva permette di ottenere un processo sostenibile, sia dal punto di vista dell'impiego energetico, sia dal punto di vista delle risorse utilizzate. Infatti, i materiali impiegati nel processo di stampa 3D sono tipicamente polimeri, metalli e ceramiche: i polimeri possono essere facilmente riciclati, mentre metalli e ceramiche generano delle polveri che possono essere impiegate in altre produzioni.

Tuttavia, le tecnologie additive non consentono lo sfruttamento di economie di scala tipiche, invece, della produzione in serie: ciò significa che hanno un costo marginale costante all'aumentare della quantità prodotta – e non decrescente come nel caso della produzione in serie. Si evince che in presenza di un mercato in cui i prodotti hanno una geometria fissa, possono essere realizzati con un unico setup e devono essere fabbricati in grandi volumi è meglio usare la produzione in serie con l'impiego delle attrezzature

tradizionali; invece, in presenza di un mercato con prodotti dotati di geometrie complesse da produrre in piccole quantità è conveniente l'Additive Manufacturing. [6]

2.2.4 Realtà aumentata

Per realtà aumentata (o Augmented Reality AR) si intende l'arricchimento della percezione umana della realtà circostante sulla base di informazioni manipolate e convogliate elettronicamente, che non sarebbero percepibili con i cinque sensi. A differenza della realtà virtuale, che crea un ambiente totalmente artificiale, la realtà aumentata utilizza l'ambiente esistente e vi frappone ulteriori informazioni che consentono agli operatori di essere più interattivi con il contesto circostante. La realtà aumentata può costituire un valido aiuto nelle attività di manutenzione, nella logistica e in alcune SOP (Standard Operating Procedures). Il manutentore, infatti, tramite tablet, può accedere ad un livello di informazioni aggiuntivo mentre svolge le operazioni. In particolare, le informazioni necessarie per svolgere le attività di manutenzione appariranno sul tablet sovrapponendosi all'immagine dell'oggetto da manutenere, dandogli supporto durante il processo, anche sfruttando video tutorial. Tali sistemi consentono inoltre di individuare quali sono gli strumenti necessari per compiere le operazioni, nonchè le parti di ricambio disponibili a magazzino. Questi sistemi si interfacciano anche con le macchine e i processi aziendali consentendo all'operatore di visualizzare direttamente sul tablet i dati raccolti dal sistema. Inoltre, questi sono visualizzati attraverso una grafica semplice ed efficace che contiene tutte le informazioni necessarie all'addetto per accedere in ogni istante ai dati sullo stato del processo. [7]

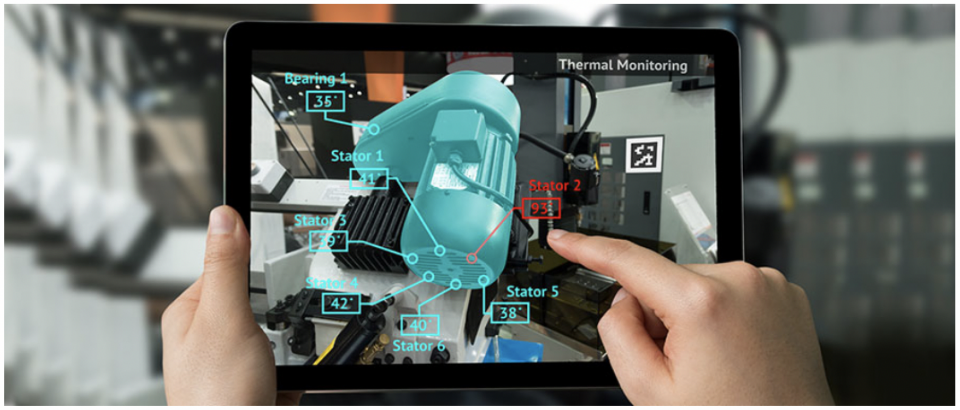


Figura 2.5: Realtà aumentata [8]

2.2.5 Simulazione

Per simulazione si intende un modello della realtà che consente di valutare e prevedere lo svolgersi dinamico di una serie di eventi susseguenti all'imposizione di certe condizioni

da parte dell'analista o dell'utente. Le simulazioni sono uno strumento sperimentale di analisi molto potente che utilizza i dati raccolti in tempo reale per rispecchiare il mondo fisico in ambiente virtuale, includendo macchinari, prodotti e operatori. La simulazione, infatti, altro non è che la trasposizione in termini logico-matematica-procedurali di un "modello concettuale" della realtà. Essa dunque è assimilabile ad una sorta di laboratorio virtuale che consente agli operatori di testare le configurazioni per una nuova linea di prodotto prima di realizzarlo fisicamente: ciò permette di ottenere un abbattimento dei costi, in quanto non è necessario ricorrere ad esperimenti reali modificando i setup delle macchine e, quindi, fermando la produzione per la durata dei test. Infine, poiché la simulazione si avvale dati raccolti in tempo reale, la loro accuratezza è un tema cruciale per lo sfruttamento a pieno dei vantaggi che questa nuova tecnologia può offrire. [5] [7]

2.2.6 Integrazione verticale e orizzontale

Nell'Industria 4.0, l'integrazione verticale e orizzontale tra macchine e internet, macchine e operatori, macchina e macchina lungo la catena del valore rappresenta la base per creare un sistema informativo.

L'integrazione orizzontale consiste nell'integrazione dei sistemi IT attraverso i vari processi di produzione e di pianificazione aziendale; in altre parole, riguarda la digitalizzazione lungo tutta la catena di fornitura, dai fornitori ai clienti, quindi è fondamentale disporre di un solido sistema di scambio dati. I vantaggi che derivano da un buon coordinamento lungo tutta la catena di fornitura riguardano: ottimizzazioni della produzione, della produttività e della soddisfazione dei dipendenti; miglioramento del servizio offerto al consumatore, in quanto la collaborazione efficace con gli attori a monte e a valle della catena di fornitura aumenta la velocità delle operazioni. L'integrazione orizzontale permette di creare degli ecosistemi di valore, basati su informazioni rilevanti. Tuttavia, raggiungere una perfetta integrazione lungo tutta la *supply chain* è una sfida non indifferente.

L'integrazione verticale, invece, riguarda l'integrazione dei sistemi IT ai vari livelli gerarchici di produzione nelle apparecchiature di produzione e automazione. Attraverso l'integrazione verticale, i dati e le informazioni raccolte nel processo produttivo tramite sensori e sistemi di controllo possono essere trasferiti ad un livello più alto di management il quale, dopo avere effettuato delle analisi sui dati ricevuti, è in grado di prendere decisioni sia strategiche che tecniche. L'impresa Industria 4.0 integrata verticalmente detiene un vantaggio competitivo in quanto riesce a rispondere in modo appropriato ai rapidi cambiamenti del mercato e a fronteggiare nuove opportunità. [9]

2.2.7 Industrial Internet

Il termine Industrial Internet è sinonimo di IIoT (Industrial Internet of Things). Con IoT (Internet of Things) si intende un gruppo di oggetti interconnessi tra loro in grado di

comunicare un grande volume di dati ad altri oggetti o al sistema all'interno dell'ambiente. Esempi di oggetti di questo tipo possono essere lucchetti per biciclette intelligenti oppure uno smartwatch. Con IIoT, invece, si intende l'uso delle tecnologie messe a disposizione dall'IoT applicate all'industria manifatturiera [10]. I dispositivi dell'IIoT, inoltre, sono dotati di performance maggiori rispetto a quelli dell'IoT in riferimento sia al numero di connessioni che questi sono in grado di garantire sia all'autonomia, in termini di batteria, poiché essendo posizionati su grandi macchinari risulta difficile effettuare operazioni di sostituzione. Concludendo, i due termini si riferiscono a due obiettivi diversi: infatti, l'IoT ha il focus sul prodotto, mentre l'IIoT si concentra sul sistema di produzione.

Il concetto di Industrial Internet è stato introdotto da General Electric (GE). Secondo la definizione fornita da GE, l'Industrial Internet è costituito da due componenti: la connessione alla rete Internet di sensori e attuatori di macchine industriali; la successiva connessione ad altre importanti reti industriali. L'Industrial Internet abilita le aziende, tramite l'uso dei sensori, dei software e di altre tecnologie, ad utilizzare i dati raccolti dall'ambiente di produzione e altre fonti. Da questa definizione si evince l'importanza dall'analisi dei dati. I dati raccolti dai sensori sono archiviati istantaneamente nei Cloud: questi verranno utilizzati per migliorare il processo produttivo (eliminando gli sprechi e riducendo i costi), per aumentare la sicurezza dei lavoratori in ogni reparto, per supportare il processo di *decision making* e per generare valore aggiunto al servizio offerto. [11] [12]

2.2.8 Cloud

Le attività di *decision making* necessitano dell'elaborazione di una grande quantità di dati raccolta durante il processo produttivo. Fino a poco tempo fa, questi dati erano elaborati da risorse di calcolo e raccolti in server per database. Ciò causava inefficienze nello scambio e nella condivisione delle informazioni e nell'allocazione delle risorse. Il Cloud Computing, invece, è una tecnologia che offre elevate performance contenendo i costi in quanto offre la possibilità di immagazzinare ed elaborare i dati in server su Cloud che possono essere pubblici o privati. [13] Nei Cloud pubblici l'hardware, software e altre infrastrutture di supporto sono di proprietà del provider, cioè del fornitore del servizio, e da esso gestiti; tuttavia, le stesse risorse sono condivise tra più aziende. Nei Cloud privati, l'hardware, software e altre infrastrutture possono essere fisicamente localizzate nel data center di una organizzazione oppure ospitate presso un provider di servizi di terze parti; in questo caso, l'hardware e il software sono strettamente dedicate ad una organizzazione: ciò garantisce da un lato un certo grado di personalizzazione del servizio e dall'altro una maggiore sicurezza delle informazioni. [14]

In generale, il Cloud offre vantaggi sia a livello economico che a livello tecnico, quali:

- Abbattimento dei costi fissi iniziali, legati agli investimenti sull'hardware e software.
- Maggiore flessibilità e scalabilità: in caso di necessità di maggiore spazio, è sufficiente adeguare le condizioni contrattuali alle proprie esigenze. Nel caso di Cloud privati, la flessibilità è maggiore in quanto è possibile personalizzare il servizio alle richieste specifiche di un'azienda.
- Maggiore attenzione al core business, in quanto, poiché il Cloud è gestito dal provider, vengono liberate le risorse allocate alla gestione dell'infrastruttura.
- Accesso al Cloud da remoto e indipendenza dalle periferiche: l'accesso alle informazioni avviene online mediante Web Browser, smartphone e tablet.
- Sicurezza del sistema: è possibile mettere in piedi un sistema di protezione dei dati ed effettuare backup periodici.

Tra gli svantaggi di questa tecnologia si evidenziano:

- Dipendenza da Internet: in caso di assenza della rete Internet si è impossibilitati ad accedere alle informazioni.
- Sicurezza informatica e violazione della privacy. Questo rischio è attenuato nel caso di Cloud privati nei quali le risorse non sono condivise con altri utenti.

2.2.9 Cybersecurity

L'avvento dell'Industria 4.0 ha portato sistemi di produzione data-driven che sfruttano i sistemi cyber-fisici e gli strumenti messi a disposizione dall'IoT. Grazie a sensori wireless, sistemi machine-to-machine, RFID e molto altro, l'Industria 4.0 è in grado di gestire una grande mole di dati, sviluppando sistemi di interconnessione e migliorando le comunicazioni tra dispositivi digitali e ambiente fisico. A tal fine, è necessario mettere in piedi una struttura Cloud come base digitale per immagazzinare i dati e permettere agli utenti di accedervi da qualunque luogo geografico. Tuttavia, questa innovazione porta con sé nuove sfide, in particolare riguardo alla protezione dei dati.

Con il termine cybersecurity si intende la protezione delle informazioni di un business e delle conoscenze riguardo una materia o un sistema contro abusi, accessi non autorizzati e furti. Con l'incremento di connessioni tra oggetti portate dalla diffusione dell'Industria 4.0 sono aumentati gli attacchi informatici, che causano danni come corruzione di dati, violazioni della privacy, perdita della reputazione e della credibilità aziendale: per questi motivi, molte aziende vittime di attacchi non dichiarano la violazione per non ammettere la vulnerabilità dei propri sistemi. Per fronteggiare questo rischio, molte aziende hanno

rafforzato ulteriormente le protezioni di sicurezza e fatto investimenti per finanziare lo sviluppo di nuove strategie e tecnologie da applicare in tema di cybersecurity. Sono state condotte ricerche e previsioni sulla futura diffusione dei sistemi internet-based e tutte concordano sul fatto che la loro applicazione e diffusione è destinata ad aumentare, come si vede in Figura 2.6. Tuttavia, all'aumentare dei dispositivi connessi aumenta anche il rischio di subire un cyber-attacco. Di conseguenza, è necessario conoscere i rischi possibili al fine di prendere tutte le precauzioni necessarie.

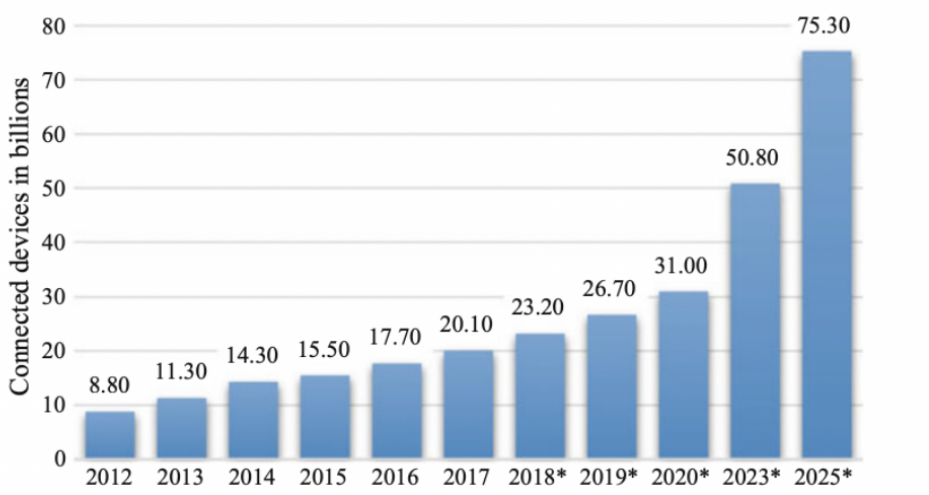


Figura 2.6: Il numero di dispositivi connessi (Internet of Things) nel mondo dal 2012 al 2025 [15]

Non esiste un'architettura universale per l'IoT, ma in generale si evidenziano quattro componenti base:

- *Perception*: è detto anche "livello sensitivo" ed è composto da oggetti fisici e dispositivi di rilevamento, come sensori e RFID.
- *Network*: rappresenta l'infrastruttura per supportare le connessioni wireless tra i sensori e il sistema informativo.
- *Service*: componente che fornisce i servizi necessari agli utenti o alle applicazioni e funge da collegamento con il database.
- *Application*: comprende i metodi di interazione tra utenti e applicazioni.

La Figura 2.7 mostra i componenti e le minacce in termini di sicurezza informatica suddivise per i livelli architetturali elencati precedentemente.

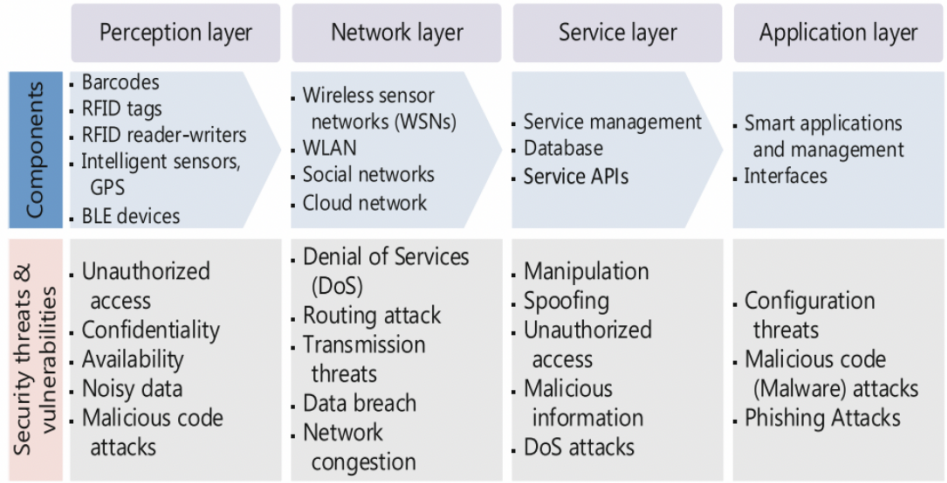


Figura 2.7: Minacce alla sicurezza e vulnerabiltà per ogni livello architetturale [15]

In linea generale, per garantire la sicurezza dell'intero sistema IoT bisogna rispettare i seguenti principi:

- **Confidenzialità**: capacità di nascondere le informazioni a persone non autorizzate.
- **Integrità**: protezione delle informazioni contro modifiche non autorizzate o non intenzionali.
- **Disponibilità**: possibilità di un utente o di un dispositivo di accedere alle informazioni in qualunque momento.
- **Autenticità**: permesso di eseguire determinate operazioni all'interno della rete solo alle entità autorizzate.
- **Nonrepudiation**: possibilità di tracciare il controllo.
- **Privacy**: la misura con cui un'entità interagisce con l'ambiente circostante e condivide le informazioni.

Alla luce di ciò, risulta fondamentale rispettare le misure e le precauzioni il più possibile per ridurre il rischio di attacchi informatici. Tra queste *best practise* rientrano l'uso dei firewall, l'accesso alla rete tramite VPN (Virtual Private Network), privilegi di accesso alle informazioni separatamente per ruolo aziendale, applicazione di password robuste e il loro aggiornamento, l'erogazione di corsi in materia di cyber security ai dipendenti e molto altro. [15]

2.2.10 Big Data & Analytics

Con l'espressione Big Data si intende una grande collezione di dati contenente informazioni rilevanti. Tali dati sono tipicamente non strutturati ed eterogenei e, di conseguenza, non possono essere manipolati ricorrendo ad approcci tradizionali. Sono necessari strumenti in grado di raccogliere, immagazzinare, gestire dati ed effettuare analisi in tempo reale. Secondo Gartner, i Big Data possono essere descritti da tre dimensioni, ovvero *volume*, *velocità* e *varietà*; accanto a queste, IBM ha aggiunto la dimensione di *veracità*. Infine, il modello che prende il nome di "5 V dei Big Data" può essere arricchito con la dimensione di *valore* (da notare che, secondo uno studio condotto da Deloitte, il valore è dato dalla somma delle quattro dimensioni identificate da IBM e della "viabilità"). [16] [17] [18]

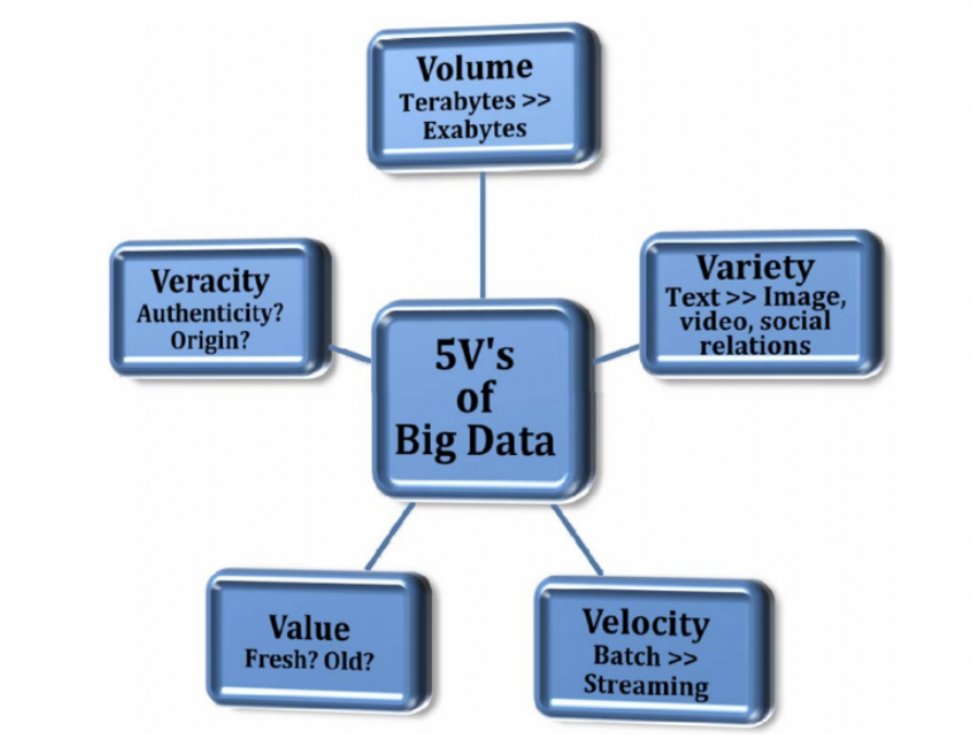


Figura 2.8: Le 5V dei Big Data [16]

- **Volume**: si riferisce alla grande quantità di dati che le organizzazioni intendono sfruttare per migliorare i processi di *decision making*. Quando si parla di "grande volume di dati" si intende una quantità dell'ordine di terabyte o petabyte, anche se non esiste una soglia convenzionale per la definizione di Big Data. Tuttavia, avere a che fare con una tale mole di dati rende necessario disporre di adeguati strumenti hardware e software.

- **Varietà**: con questa dimensione si intende sottolineare l'eterogeneità dei dati e per questo motivo la loro gestione risulta complessa. I dati possono essere di diverso

tipo, in forma strutturata, semi-strutturata e non strutturata e provenire da fonti differenti. Grazie alla diffusione di dispositivi intelligenti, sensori e social network, le organizzazioni hanno a disposizione dati quali testi, tweets, audio, video, file di log e molto altro.

- **Velocità**: la creazione di dati in tempo reale rende necessario accelerare il più possibile il processo di raccolta, immagazzinamento e analisi. In certi tipi di business, analizzare velocemente i dati raccolti in tempo reale può generare valore in più per l'azienda.
- **Veracità**: si riferisce al livello di affidabilità associato a certi tipi di dati. Infatti, avere dati di qualità è un requisito fondamentale per poter estrarre informazioni significative ed effettivamente utili; per questo motivo, è necessario pulire i dati effettuando preprocessing o filtrarli per eliminare informazioni irrilevanti.
- **Valore**: una volta raccolti i dati, è necessario capire come estrarre valore dalla conoscenza generata dalle analisi.

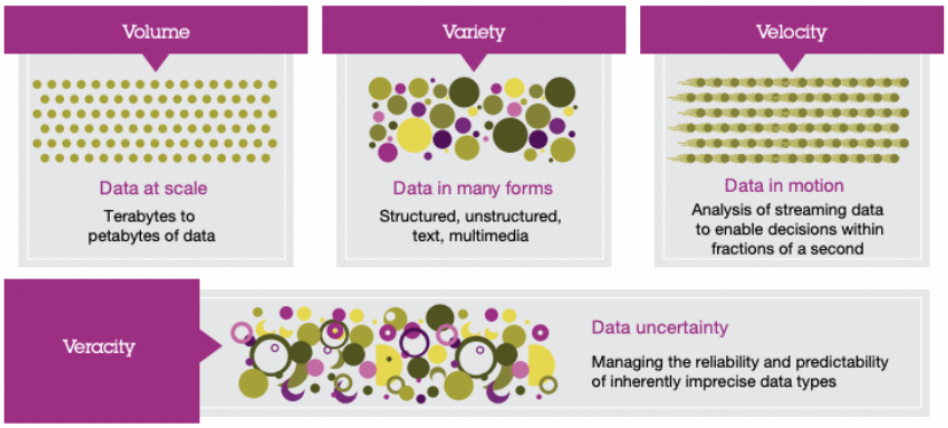


Figura 2.9: Le quattro dimensioni dei Big Data. Fonte: www.ibm.com

Concludendo, i Big Data sono una combinazione di queste caratteristiche in grado di generare un vantaggio competitivo alle organizzazioni, migliorando le performance e il processo di *decision making*.

2.3 Principi di progettazione dell'Industria 4.0

Industria 4.0 è un'espressione utilizzata per indicare tecnologie e concetti per l'organizzazione della catena del valore. All'interno delle Smart Factories, il sistema cyber-fisico monitora i processi, crea una copia virtuale della realtà, arricchendola di contenuti, e in questo modo aiuta la decentralizzazione delle decisioni. Grazie all'IoT, inoltre, il sistema

cyber-fisico comunica e coopera con i dispositivi e con gli operatori in tempo reale. Da questo risultato, si evincono dei principi di progettazione che devono essere implementati per dare vita ad una Smart Factory.[19]

- **Interoperabilità**: per dare vita alle fabbriche intelligenti è necessario garantire la connessione tra operatori, macchine e standard di processi di lavoro e ciò è possibile grazie agli strumenti messi a disposizione dall'IoT. L'interoperabilità riguarda anche la capacità di far dialogare tra loro standard diversi in modo tale da sfruttare a pieno i dati provenienti da fonti diverse.
- **Virtualizzazione**: si intende la capacità di monitorare il processo fisico. Grazie ai sensori apposti sui macchinari, il sistema cyber-fisico è in grado di creare una copia virtuale dello stabilimento, arricchendola con informazioni aggiuntive. Senza l'interconnessione tra le parti non sarebbe possibile unire le informazioni del mondo digitale e del mondo fisico.
- **Decentralizzazione**: data la numerosità dei prodotti e delle loro esigenze, è difficile effettuare un controllo centralizzato. Il sistema cyber-fisico è in grado di prendere delle decisioni senza coinvolgere altri attori.
- **Capacità in tempo reale**: allo scopo di gestire l'enorme quantità di dati, è necessario raccoglierli ed analizzarli in tempo reale.
- **Orientamento ai servizi**: è correlato al fatto che la produzione deve essere adattata alla domanda del cliente, garantendo maggiore valore aggiunto a prodotti e servizi. Inoltre, l'orientamento al servizio è legato alla necessità dei produttori di sviluppare nuovi servizi basati sull'interpretazione dei dati raccolti.
- **Modularità**: la flessibilità è un requisito fondamentale che consente di rispondere tempestivamente alle esigenze del cliente e rimanere competitivo sul mercato. I sistemi in grado di adattarsi ai requisiti in base alle condizioni del mercato permettono di fronteggiare abilmente fluttuazioni stagionali oppure cambiamenti nelle caratteristiche del prodotto.

2.4 Panoramica sull'Industria 4.0 nel mondo

La trasformazione delle aziende manifatturiere da tradizionali a Smart Factories è un processo in via di sviluppo, in cui rappresentanti di Governo, industrie e società devono far fronte ad una serie di incertezze; in particolare, i Paesi devono ripensare le loro strategie industriali. Con il lancio del progetto Industrie 4.0, la Germania è stata il primo Paese ad aumentare la digitalizzazione all'interno della catena del valore e l'interconnessione dei processi. Sulla scia dell'esempio tedesco, ciascun paese ha sviluppato la propria strategia

per fronteggiare questa trasformazione. Gli Stati Uniti hanno lanciato nel 2011 il progetto Advanced Manifacturing Partnership; l'Italia nel 2012 ha fondato il Cluster Tecnologico Italiano con l'obiettivo di sensibilizzare le aziende su questi temi e nel 2017 ha presentato il Piano Nazionale Industria 4.0; il Giappone nel 2016 ha presentato il piano Japan's Society 5.0, puntando a rivoluzionare non solo la sfera produttiva, ma anche quella sociale. I Paesi attivi con piani e iniziative nazionali su questo tema sono mostrati in Figura 2.10. [20] [21]

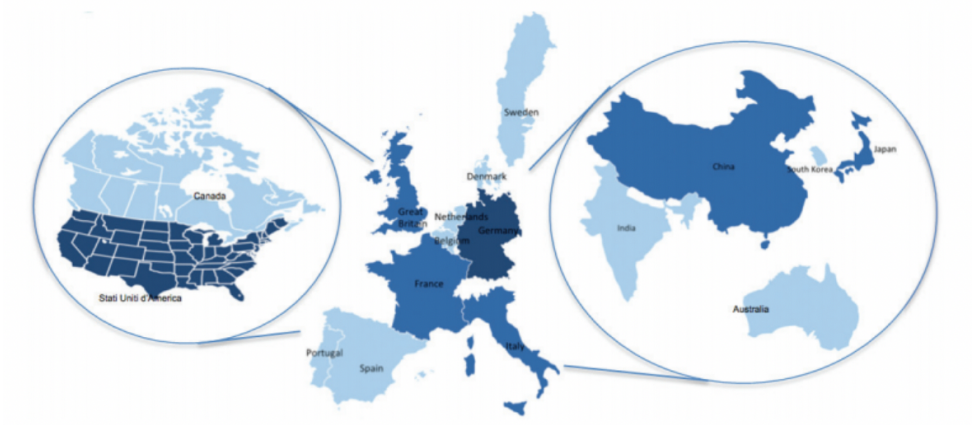


Figura 2.10: Le principali economie interessate all'Industria 4.0 [20]

In linea generale, si identificano due modelli, europeo e statunitense, che si differenziano sul focus a cui i due approcci puntano l'attenzione.

Il modello europeo mira a creare uno standard di riferimento che tutte le aziende possano adottare per implementare le tecnologie dell'Industria 4.0: l'obiettivo è raggiungere livelli di efficienza e produttività sempre maggiori allo scopo di creare delle fabbriche intelligenti.

Il modello statunitense, invece, punta a migliorare sempre di più i servizi e i prodotti intelligenti ed ha come tecnologia di riferimento l'IoT, mettendo in primo piano l'importanza della connessione tra prodotto e cliente finale.

Nonostante le differenze, entrambi i modelli riconoscono l'importanza del sistema cyber-fisico, in quanto esso consente l'integrazione tra macchine, oggetti, persone, intese sia come operatori che come consumatori, non solo all'interno o ai confini dell'impresa, ma nella società. I piani messi in azione dai governi si possono analizzare sulla base di tre parametri:

- **Governance e attori coinvolti** per promuovere la diffusione di queste tematiche.
- **Aspetti tecnologici** ritenuti più significativi.
- **Modalità di supporto** per favorire l'implementazione delle tecnologie dell'Industria 4.0.

Dal punto di della Governance e degli attori coinvolti, si evidenzia che nei Paesi europei vi è una forte collaborazione tra settore pubblico e privato che vede da un lato Governo,

università, centri di ricerca e dall'altro aziende fornitrici di tecnologie (nel caso della Germania in particolare Bosch). Il modello statunitense, invece, si caratterizza per una presenza più contenuta del Governo a favore di gruppi privati del settore ICT, imprese fornitrici di tecnologie, università e centri di ricerca privati.

Per quanto riguarda gli aspetti tecnologici, in Italia, data l'eterogeneità delle aziende produttive diffuse sul territorio, si è posta l'attenzione sull'adozione delle tecnologie abilitanti dell'Industria 4.0, in misura maggiore Additive Manufacturing, Big Data, IoT e realtà aumentata. Sulla scia del modello italiano, anche il piano francese si è focalizzato sull'adozione delle tecnologie abilitanti, principalmente allo scopo di ottenere maggiori prestazioni dal punto di vista della sostenibilità ambientale. La Germania, invece, punta a creare uno standard di riferimento per la creazione delle fabbriche intelligenti, con particolare attenzione sul sistema cyber-fisico e sull'IoT. Gli Stati Uniti, come già anticipato, si focalizzano non tanto sull'ottimizzazione del processo produttivo sfruttando le tecnologie abilitanti dell'Industria 4.0, quanto all'integrazione lungo tutta la catena del valore delle aree di business resa possibile grazie alle piattaforme Cloud e all'interconnessione tra azienda e consumatori.

Dal punto di vista delle modalità di supporto, infine, in Italia e Francia il settore pubblico offre finanziamenti di oltre 10 miliardi di euro. In Germania, invece, l'impegno pubblico si attesta sul miliardo di euro, mentre negli Stati Uniti sul mezzo miliardo di dollari.

	Governance e attori coinvolti	Aspetti Tecnologici	Modalità di supporto
Italia	Forte presenza del settore pubblico, per mezzo di Governo, Ministeri, università, poli di ricerca. Più contenuto l'intervento del settore privato.	Le tecnologie digitali a cui si punta maggiormente sono: Additive Manufacturing, Big Data, IoT e Realtà aumentata.	Iniziative fiscali che comprendono super-rammortamento e iper-ammortamento, credito d'imposta e detrazioni fiscali. In totale, il Governo ha stanziato 13 Mrd di euro.
Germania	Compresenza di attori del settore pubblico e di imprese fornitrici di tecnologie come Bosch.	Particolare enfasi su sistema cyber-fisico e tecnologie dell'IoT. Meno attenzione su Big Data, realtà aumentata e robot collaborativi.	Finanziamenti ad aziende, centri di ricerca, anche a fondo perduto per le imprese che svolgono attività di ricerca e sviluppo. Impegno pubblico per 1 Mrd di euro.
Stati Uniti	Forte presenza dei gruppi privati ICT, imprese fornitrici di servizi e centri di ricerca. Minore presenza del Governo.	Focus sulla creazione di una piattaforma Cloud e sui Big Data allo scopo di integrare le diverse aree di business. Meno attenzione alle singole tecnologie.	Impegno pubblico per mezzo miliardo di dollari.

Tabella 2.1: Una panoramica schematica dei piani messi in atto da Italia, Germania e Stati Uniti

2.4.1 Il Piano Nazionale Industria 4.0 in Italia

Il Piano Nazionale Industria 4.0 messo in atto nel 2017 aveva come obiettivo il raggiungimento entro la fine dell'anno di 10 miliardi di investimenti privati e 11 miliardi di euro in Ricerca, Sviluppo e Innovazione con focus sulle tecnologie dell'Industria 4.0. Per raggiungere questi obiettivi, il Piano aveva previsto le seguenti misure: [22] [23]

- **Iper e Super ammortamento**: l'obiettivo di questo provvedimento è incentivare gli investimenti delle imprese private in beni strumentali, materiali e immateriali (software e sistemi IT), funzionali alla trasformazione tecnologica e digitale dei processi. L'iper-ammortamento consiste nella supervalutazione al 250% dell'investimento in beni, dispositivi e tecnologie abilitanti dell'Industria 4.0, mentre il super-ammortamento

prevede la supervalutazione del 140% degli investimenti sui nuovi strumenti acquistati oppure in leasing.

- **Nuova Sabatini**: è un'agevolazione messa a disposizione con l'obiettivo di facilitare alle imprese l'accesso al credito aumentando così la competitività delle aziende italiane. Possono beneficiarne tutte le PMI che devono sostenere le spese necessarie per l'acquisto di macchinari, attrezzature, beni strumentali a uso produttivo, software e tecnologie digitali. Il finanziamento necessario deve essere compreso tra 20.000 e 2.000.000 euro, non deve superare i 5 anni e ha un tasso di interesse compreso tra 2.75% e 3.57% (nel caso di investimenti in tecnologie di Industria 4.0).

- **Credito d'imposta Ricerca e Sviluppo**: viene garantito un credito d'imposta al 50% a tutte le aziende che investono in Ricerca e Sviluppo; il beneficio riconosciuto arriva fino ad un massimo annuale di 20 milioni di euro. L'obiettivo della misura è favorire la spesa delle aziende in Ricerca Sviluppo in modo tale da aumentare la competitività delle imprese italiane.

- **Patent Box**: è un regime di tassazione agevolata sui redditi derivanti dall'uso di beni immateriali, quali brevetti industriali, marchi registrati, *know how* e software protetto da copyright e consiste in una riduzione del 50% dell'IRES e dell'IRAP. Questo provvedimento ha la funzione di rendere il mercato maggiormente attrattivo per investimenti sia nazionali che esteri e favorire la spesa in Ricerca e Sviluppo.

- **Startup e PMI innovative**: consiste in una serie di provvedimenti volti a sostenere le imprese innovative in tutte le fasi del loro ciclo di vita e diffondere una nuova cultura imprenditoriale incentrata sulla collaborazione e sull'internazionalizzazione. Le imprese in oggetto godono di semplificazioni in merito alla sfera amministrativa, all'esonero dalla disciplina fallimentare in caso di insuccesso e agevolazioni fiscali con detrazioni fino al 30%.

- **Fondo di garanzia**: è un provvedimento volto a sostenere le imprese che hanno difficoltà ad accedere al credito bancario perché non dispongono delle garanzie. Le imprese possono usufruire di una garanzia pubblica sul finanziamento fino all'80% per investimenti sia di breve che medio-lungo termine.

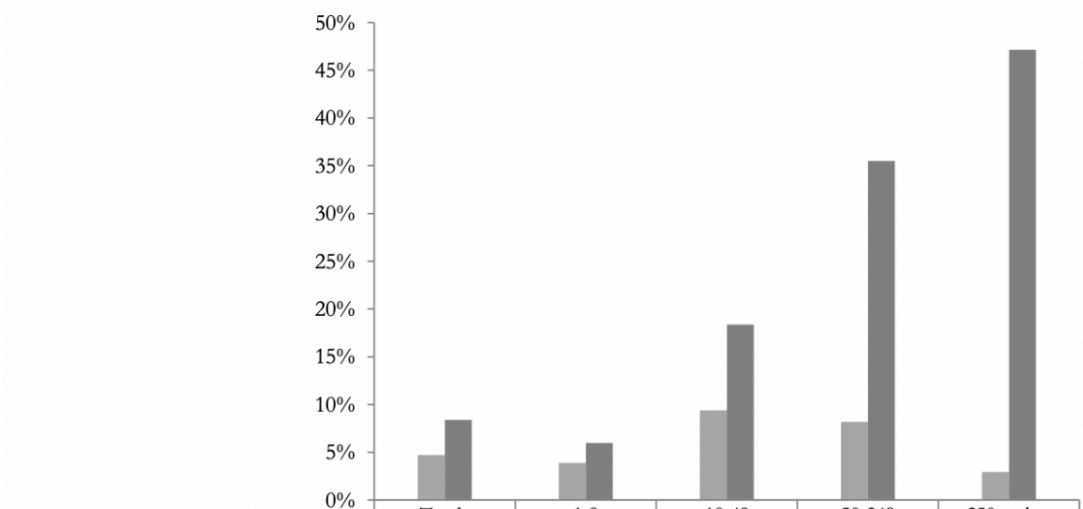
- **ACE (Aiuto alla Crescita Economica)**: è un'iniziativa volta a potenziare il capitale in impresa al fine di ottenere strutture finanziarie più solide e, dunque, più competitive. Si tratta di deduzioni dal reddito complessivo di importo pari al rendimento nozionale del nuovo capitale proprio.

- **IRES, IRI e contabilità per la cassa**: consiste in una riduzione della pressione fiscale dal 27% al 24%, uniformandola all'aliquota media europea.

- **Salario di produttività**: fornisce una tassazione al 10% per i premi salariali fino a un massimo di 3.000 euro. L'idea alla base del provvedimento è favorire l'incremento della produttività e l'efficienza dei lavoratori aumentando il loro salario.

Nel luglio 2018, il Ministero dello Sviluppo Economico ha pubblicato un report contenente i risultati sugli investimenti delle imprese a seguito delle misure adottate. Le indagini sono state condotte nel periodo tra ottobre 2017 e febbraio 2018 coinvolgendo un campione rappresentativo costituito da 23.700 imprese italiane operanti nel settore della produzione di beni e fornitura di servizi di tutte le dimensioni. [24]

In prima battuta, è stata analizzata la diffusione delle tecnologie abilitanti dell'Industria 4.0. Sul totale del campione considerato, il 4,7% delle imprese è costituito da aziende tradizionali che hanno in programma di effettuare investimenti in tecnologie attinenti all'Industria 4.0 nel prossimo triennio; l'8,4% è costituito da imprese 4.0 già consolidate; di conseguenza, si evince che l'86,9% delle aziende del territorio sono tradizionali e non hanno in programma di effettuare investimenti in Industria 4.0. Inoltre, per le imprese 4.0 si evidenzia un maggiore utilizzo di queste tecnologie all'aumentare della dimensione. Figura 2.11.



	Totale	1-9	10-49	50-249	250 e oltre
Imprese tradizionali con interventi 4.0 programmati	4,7%	3,9%	9,4%	8,2%	2,9%
Imprese 4.0	8,4%	6,0%	18,4%	35,5%	47,1%

Figura 2.11: Diffusione delle tecnologie 4.0, dettaglio per classe dimensionale. Valori percentuali. [23]

Tra le imprese 4.0, le tecnologie maggiormente adottate sono risultate essere IoT, cybersecurity, integrazione orizzontale delle informazioni e gestione dei dati su Cloud. Gli investimenti delle imprese, dunque, sono focalizzate sulle tecnologie basate sui dati piuttosto che quelle produttive. Figura 2.12

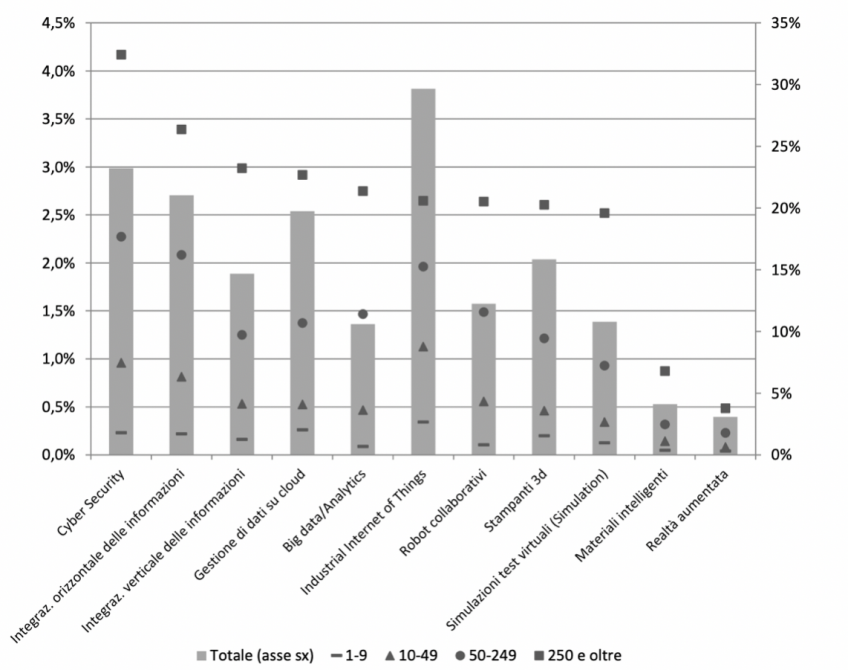


Figura 2.12: Diffusione delle tecnologie 4.0 per classe dimensionale (totale asse sinistro, classi dimensionali sull'asse destro) [23]

Le misure adottate dal Piano Nazionale Industria 4.0 complessivamente hanno generato investimenti di oltre 10 miliardi di euro per la componente di iper-ammortamento sui beni materiali e 3 miliardi per la componente di super-ammortamento sui beni immateriali, coinvolgendo più di un milione di imprese. Inoltre, grazie agli incentivi sulle spese di Ricerca e Sviluppo, le imprese italiane hanno investito più di 8 miliardi di euro. Queste misure hanno agevolato in prevalenza le imprese medio-grandi (64%). Nonostante il Piano Nazionale Industria 4.0 prevedesse delle agevolazioni per le startup innovative e PMI, su questo fronte non sono stati registrati ottimi risultati. Infatti, gli investimenti sono diminuiti del 38%, passando dalle 92 operazioni del 2016, alle 57 nel 2017.

Dopo il 2017, tuttavia, si è registrato un calo negli ordini di beni di macchine e utensili che è diventato più consistente nel 2019. Figura 2.13 [25] [26].

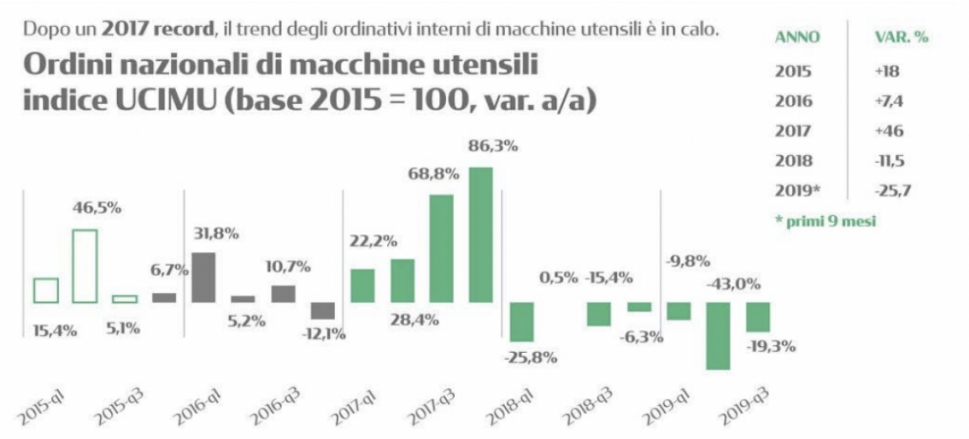


Figura 2.13: Ordini nazionali di macchine utensili nel periodo 2015-2019

Con la nuova Legge di Bilancio del 2020 è stato introdotto il nuovo piano Transizione 4.0, con il quale sono state sostituite le agevolazioni dell'iper-ammortamento e del super-ammortamento con un nuovo credito di imposta per gli investimenti in beni materiali, con un'aliquota differenziata in base alla tipologia di strumento acquistato.

La Nuova Sabatini è stata nuovamente implementata e rifinanziata, aumentando l'importo massimo finanziabile, così come il Fondo di Garanzia.

La stessa normativa ha introdotto un nuovo credito d'imposta per sostenere la competitività delle imprese e le loro spese in innovazione tecnologica e design e per favorire i processi di transizione digitale e la sostenibilità ambientale in sostituzione del precedente credito d'imposta in Ricerca, Sviluppo e Innovazione. Inoltre, è stato prorogato il credito d'imposta per le spese di formazione del personale finalizzate all'acquisizione o al consolidamento delle competenze nelle tecnologie rilevanti dell'Industria 4.0.

Per questi pilastri portanti della normativa, il Ministero dello Sviluppo Economico ha messo a disposizione 7 miliardi di euro. L'obiettivo del piano triennale 2020-2022 Transizione 4.0 rimane sostenere le imprese che puntano a investire nel processo di trasformazione tecnologica e digitale della produzione. [27]

2.5 Benefici dell'Industria 4.0

In generale, l'Industria 4.0 ha come obiettivo essenziale l'ottimizzazione della produzione, rendendola più veloce, efficiente e incentrata sul cliente. Di seguito, una panoramica dei principali vantaggi. [19]

- **Miglioramento della produttività attraverso ottimizzazione e automazione**: ciò consente di risparmiare sui costi, ridurre gli sprechi, aumentare la redditività, prevedere errori o ritardi, ottimizzare l'uso del macchinario in base alle sue condizioni.

- **Sfruttare dati in tempo reale per soddisfare le esigenze del cliente**: il miglioramento della produttività non si riduce solamente al processo produttivo, ma riguarda l'intera *supply chain*. L'Industria 4.0 è incentrata sulle esigenze del cliente e, poiché il consumatore finale si aspetta prodotti di buona qualità nei tempi desiderati, l'impresa deve essere in grado di migliorare non solo il processo di fabbricazione, ma di ottimizzare l'intera *value chain*. Questo risultato si ottiene allineando le informazioni tra cliente, catena di approvvigionamento e condizioni di funzionamento dei macchinari: maggiore sarà il numero di dati raccolti tempestivamente, maggiore sarà il valore lungo la *value chain*.

- **Continuità delle funzioni aziendali grazie al monitoraggio e alla manutenzione predittiva**: grazie al monitoraggio continuo dello stato di salute dei macchinari messo a disposizione dal sistema cyber-fisico, si possono effettuare interventi di manutenzione predittiva in modo tale da evitare guasti. Il fermo macchina dovuto al guasto, infatti, interrompe la catena di produzione generando danni non solo relativamente al costo dei pezzi di ricambio e al lavoro di manutenzione, ma anche alla reputazione dell'azienda, che aumenterà il tempo di risposta al cliente che, di conseguenza, potrebbe annullare l'ordine.

- **Prodotti di migliore qualità**: tra le aspettative del consumatore rientra, come già anticipato, un servizio di risposta rapido ed efficiente. Tuttavia, ciò non significa che il cliente sia disposto a sacrificare la qualità del prodotto, anzi il mercato diventa sempre più esigente anche da questo punto di vista. A tal fine, il monitoraggio garantito dal sistema cyber-fisico permette di controllare la qualità dei prodotti in tempo reale mentre la presenza sempre maggiore dei robot nel processo produttivo riduce gli errori di lavorazione.

- **Migliori condizioni di lavoro e sostenibilità**: il monitoraggio dei macchinari offre delle garanzie circa il loro stato di salute, ma non solo. Anche gli operatori beneficiano dell'introduzione del sistema cyber-fisico in quanto, grazie ad esso, è possibile rilevare parametri in tempo reale quali la temperatura, umidità, presenza di gas e radiazioni: in questo modo, si possono migliorare le condizioni di lavoro.

- **Personalizzazione**: il comportamento e le esigenze dei consumatori sono cambiati, in particolare tra le loro aspettative rientrano prodotti di qualità, servizi di consegna tempestivi ma anche un certo grado di personalizzazione. Ciò diventa possibile se le tecnologie e i processi dell'Industria 4.0 vengono implementati correttamente: infatti, eventuali modifiche nella catena di montaggio sono molto più facili da applicare.

- **Maggiore flessibilità**: grazie alle innovazioni introdotte dall'Industria 4.0, ci si può adattare alle diverse condizioni di mercato più facilmente. Ciò anche grazie allo sfruttamento di nuove tecnologie come Big Data, Artificial Intelligence, sistemi

cyber-fisici che consentono di prevedere la domanda stagionale e di adattare ad essa la produzione.

- **Sviluppo di nuove capacità innovative per nuovi modelli di business**: l'industria 4.0 apre le porte a nuovi tipi di mercati e, quindi, nuovi fonti di guadagno. In questo modo si riesce ad essere competitivi con le altre aziende e proporre il proprio prodotto o servizio in maniera più accattivante per consumatori.

2.5.1 Benefici Industria 4.0 per area aziendale

Con l'Industria 4.0 si rivoluziona il concetto di fabbrica e il modo di approcciarsi ad essa in quanto si verifica uno stravolgimento della cultura alla base dell'industria, che rende l'organizzazione sempre più cosciente ed informata su quello che succede in ogni fase del ciclo produttivo. Tuttavia, affinché si ottengano i vantaggi precedentemente elencati, è fondamentale non solo fare investimenti nelle nuove tecnologie abilitanti, ma anche integrarle efficacemente con le quelle già esistenti: questo consente all'azienda di migliorare le proprie performance ed ottenere dei benefici in ciascun settore di attività.

Area	Benefici
Produzione	- Ottimizzazione la produzione in base a diversi criteri - Rilevamento dei parametri della produzione e modifica in tempo reale - Gestione efficace delle risorse energetiche
Logistica interna	- Movimentazione automatica delle merci e del loro tracking in azienda - Gestione efficace delle merci in ingresso - Gestione automatizzata dei magazzini
Acquisti	- Automatizzazione del processo d'acquisto - Transazione condizionata allo stato della merce - Certificazione della merce acquistata
Manutenzione	- Automazione dello scheduling di unloading - Automazione del carico/scarico nel sistema gestionale - Maggiore coordinamento tra trasportatore e magazzino interno - Modellazione dei comportamenti degli attori all'interno della supply chain
Distribuzione e vendita	- Acquisizione dei dati di acquisto/vendita in tempo reale - Automatizzazione nel processo di fatturazione
Assistenza post-vendita	- Acquisizione dei dati sull'uso del prodotto - Diminuzione dei costi di assistenza e marketing - Possibilità di offrire assistenza post-vendita da remoto - Maggiore personalizzazione del servizio e update del prodotto

Tabella 2.2: Benefici dell'Industria 4.0 per area aziendale

2.5.2 Manutenzione

L'attività di manutenzione è parte del processo di creazione del valore di un prodotto o servizio in quanto contribuisce a migliorare la qualità e la produttività e a ridurre i costi. Una manutenzione efficace, infatti, aumenta la vita utile del soggetto coinvolto e permette di mantenere elevati livelli di qualità del prodotto lavorato. Contrariamente, se l'attività di manutenzione non è effettuata con scrupolosità, i guasti alle attrezzature sono più frequenti: ciò determina il fermo della macchina, l'interruzione del processo produttivo e, di conseguenza, una diminuzione della produttività. Quando i componenti non sono appropriatamente manutenuti, la loro vita utile si riduce e, quindi, necessitano di essere sostituiti più frequentemente: ciò genera un aumento dei costi.

In letteratura, diversi autori si sono occupati di descrivere le possibili strategie di manutenzione. Secondo Bateman, esistono tre tipologie di programmi: la manutenzione *reattiva* e la manutenzione preventiva e predittiva, tecniche appartengono alla categoria di manutenzione *proattiva*. Accanto a queste strategie, Weil ha introdotto un'ulteriore categoria di manutenzione, chiamata *aggressiva*, in cui spicca la tecnica TMP (Total Productive Maintenance).[28]

Manutenzione reattiva

Nella strategia manutentiva **reattiva** o **correttiva**, il macchinario continua ad essere operativo fin quando non si verifica un guasto oppure fin quando il progredire di un'anomalia non costringa a fermare il processo produttivo. In questi casi, i componenti danneggiati possono subire degli interventi di manutenzione temporanei per poi procedere con la riparazione definitiva in seguito. Questa tecnica permette di minimizzare le spese relative al personale adibito alle operazioni di manutenzione, in quanto il costo di manutenzione è nullo fin tanto che la macchina funziona, e sull'acquisto strumenti che monitorano le attrezzature durante il loro funzionamento.

Tuttavia, la manutenzione reattiva rende la produzione di un'impresa oscillante e intermittente a causa dei possibili fermi macchina, generando una consistente perdita di ricavi; inoltre, il costo della manutenzione aumenta in quanto è necessario sostituire i componenti più frequentemente; infine, poiché non viene rilevata l'usura del macchinario, aumenta il rischio di prodotti scartati oppure di eseguire lavorazioni al di fuori dei limiti di tolleranza.

La strategia reattiva è consigliata per macchinari facilmente riparabili oppure quando si opera in un contesto in cui un eventuale fermo della linea produttiva non genera gravi danni al ciclo produttivo. È il caso di avarie alle singole macchine, il cui ruolo e lavorazioni possono essere facilmente sostituite da una macchina gemella.

Manutenzione preventiva

La manutenzione **preventiva** è una strategia appartenente alla categoria della manutenzione *proattiva*. L'obiettivo delle strategie proattive è ridurre la probabilità di un guasto inaspettato dei macchinari mediante monitoraggio del deterioramento dei componenti e piccoli interventi manutentivi allo scopo di ripristinare le condizioni delle attrezzature coinvolte. La manutenzione preventiva prevede di effettuare interventi quali lubrificazione, pulizia e sostituzione dei componenti in intervalli di tempo predeterminati oppure dopo un certo periodo di uso della macchina. Infatti, gli interventi sono programmati sulla base della probabilità che il componente arrivi a rottura all'interno di uno specifico intervallo. La manutenzione preventiva consente di evitare fermo macchina imprevisti e di estendere la vita utile di un macchinario/componente. Inoltre, la possibilità di programmare un intervento manutentivo consente una migliore organizzazione del lavoro, gestendo l'interruzione nella maniera più opportuna e conveniente. D'altro canto, applicare questa strategia significa programmare l'intervento e accettare di fermare il processo produttivo.

La strategia di manutenzione preventiva risulta efficace quando il guasto si manifesta ad intervalli regolari ed è consigliata nei casi in cui una possibile interruzione del processo produttivo genera danni consistenti al ciclo di produzione, alla salute delle persone, agli ambienti e agli impianti.

TPM – Total Productive Maintenance

Il **Total Productive Maintenance o TPM** è una strategia appartenente alla categoria della manutenzione *aggressiva*, in cui l'obiettivo è migliorare il funzionamento complessivo dell'attrezzatura mediante il coinvolgimento della progettazione di attrezzature nuove o già esistenti.

Il TPM nasce negli stabilimenti giapponesi allo scopo di supportare il processo di implementazione della filosofia *just-in-time*. Le attività del TPM si focalizzano sull'eliminazione di sei principali fonti di "perdite", ovvero: guasti ai macchinari, tempo di set-up e regolazione dei parametri, inattività e arresti minori del ciclo produttivo, riduzione della velocità del processo e scarti di prodotto finito.

Il TMP può essere visto come una partnership formata da membri di funzioni aziendali diverse. Infatti, si costituiscono dei piccoli team in cui i lavoratori del reparto produzione aiutano quelli del reparto di manutenzione nell'eseguire gli interventi. In questo modo, i lavoratori del reparto produzione giocano un ruolo attivo nel rilevamento delle anomalie e contribuiscono ad aumentare l'efficienza degli impianti e a mantenerli in buone condizioni. In questi piccoli team possono essere coinvolti anche membri del reparto progettazione: infatti, grazie all'esperienza dei lavoratori dell'area produzione, gli ingegneri possono progettare e sviluppare componenti in grado di offrire migliori performance e garantire elevati livelli di qualità. Il TPM basato sui team prevede due tipologie di attività: il

team di prevenzione e il team di miglioramento. Il primo ha lo scopo di garantire delle performance più elevate di macchinari e impianti intervenendo sul processo di progettazione in modo tale che i componenti installati siano facili da manutenere e da adoperare. Il secondo riguarda il miglioramento del piano di manutenzione vero e proprio, identificando e correggendo le condizioni critiche per le operazioni di manutenzione; grazie ad uno scheduling efficace delle attività manutentive, è possibile migliorare la disponibilità delle attrezzature e diminuire il costo e il tempo di riparazione.

Manutenzione predittiva

La manutenzione predittiva è la seconda strategia appartenente alla categoria della manutenzione *proattiva*; l'obiettivo della manutenzione proattiva è ridurre la probabilità di un guasto inaspettato dei macchinari controllando continuamente lo stato di usura dei componenti ed effettuando piccoli interventi manutentivi allo scopo di mantenere in buono stato le attrezzature coinvolte.

Utilizzando la strategia di manutenzione preventiva, si rischia di intervenire sul componente oppure di sostituirlo quando quest'ultimo è ancora in grado di lavorare e di garantire un buon livello di qualità: ciò si traduce in sprechi sia di tempo che di denaro, in quanto l'intervento del manutentore genera, tipicamente, l'interruzione della produzione e quindi una perdita dei ricavi. Nella manutenzione predittiva, invece, si effettuano operazioni di manutenzione in base alle condizioni in tempo reale di un determinato componente. Infatti, poiché gli strumenti forniti dall'IoT permettono di misurare diversi parametri di un componente, per esempio la temperatura, la corrente, le vibrazioni, si interviene con le attività manutentive per ripristinare le condizioni di salute quando questi mostrano comportamento anomalo. Quindi, a differenza della manutenzione preventiva, si agisce in base alle condizioni del macchinario e non a intervalli di tempo regolari.

Questa nuova modalità di approcciarsi alle attività di manutenzione richiede un cambiamento di mentalità e degli investimenti iniziali consistenti. Tuttavia, è in grado di determinare un abbattimento del 25%-30% dei costi di manutenzione grazie all'identificazione precisa del possibile guasto e, di conseguenza, del migliore servizio di assistenza; una riduzione del 35%-40% dei tempi vuoti: poiché gli interventi manutentivi sono effettuati sulla base delle condizioni del macchinario, non è necessario fermare la produzione ad intervalli di tempo regolari e, dunque, si aumenta il periodo di attività della macchina; infine, si ottengono incrementi della produttività della linea del 20%-25%. [29]

Per implementare la manutenzione predittiva e beneficiarne dei vantaggi, un'azienda deve organizzare tre macro-aree di lavoro:

- **Raccolta dei dati dai macchinari** in modo da avere informazioni sul loro stato di funzionamento. In questa fase è fondamentale ottenere dati accurati e di qualità per poter costruire successivamente un modello di predizione efficace. A questo scopo,

è necessario disporre di strumenti di misurazione, sistemi di interconnessione tra i sensori e computer che permettano di far fluire i dati ai sistemi di analisi dei dati. Inoltre, per poter costruire un modello di predizione è necessario avere conoscenza sui tipi di guasto, le cause scatenanti e i campanelli d'allarme che segnalano quando il macchinario non sta funzionando correttamente; per fare ciò, si raccoglie una quantità ingente di dati per disporre di quante più informazioni possibili. Il processo di raccolta dei dati storici può impiegare settimane o mesi, a seconda della complessità del macchinario da studiare.

- **Pulitura ed elaborazione**: si eliminano i dati ridondanti che non aggiungono contenuto informativo rilevante. Questa fase è necessaria da un lato per ottenere un modello accurato e, dall'altro, serve per snellire e velocizzare il processo di elaborazione.
- **Generazione del modello predittivo** sfruttando gli algoritmi di machine learning e data mining. [29]

Una volta generato il modello e ottenuti i risultati, questi devono essere interpretati in base al contesto da un esperto di dominio. Quando il modello viene validato si otterrà un processo automatico in cui si implementeranno le fasi descritte precedentemente. Grazie ad esso, si sarà in grado di determinare se le performance del macchinario non sono ottimali e se sono necessario interventi manutentivi. Inoltre, la manutenzione predittiva può suggerire indicazioni sul miglior intervento in base al componente soggetto alla riparazione da effettuare.

Concludendo, la manutenzione è un'attività critica nel ciclo di vita di un'impresa e una strategia adatta al contesto consente di migliorare l'efficienza e la produttività dell'impianto. La manutenzione predittiva può essere la soluzione migliore in determinati contesti in quanto permette di ridurre i fermi di produzione non programmati, i costi delle attrezzature e di manutenzione e di aumentare i ricavi e la sicurezza del personale coinvolto. Tuttavia, è necessario effettuare degli investimenti iniziali per le attrezzature necessarie e per formare le competenze dei dipendenti che devono approcciarsi a questa nuova realtà.

Strategia	Descrizione	Costo di implementazione	Vantaggi	Conseguenze
Reattiva	Intervento quando si verifica il guasto	Basso	Ideale per macchinari con bassa priorità e non critici	Può generare costi elevati in caso di gravi danni.
Preventiva	Interventi programmati in intervalli temporali regolari	Medio	Ideale in caso di scarsa esperienza nelle operazioni di manutenzione	Se non correttamente ottimizzata, possono verificarsi guasti.
TPM	Interventi programmati con la collaborazione di tutte le funzioni aziendali	Medio	Responsabilizzazione degli operatori; riduzione degli sprechi e dei costi di manutenzione; attrezzature agevoli da usare.	Se il personale non è correttamente formato, si possono verificare guasti
Predittiva	Interventi in base alle condizioni dei macchinari monitorate in tempo reale	Alto	Monitoraggio in tempo reale; fornire le cause a seguito del verificarsi di anomalie.	La fase di set-up può risultare costosa.

Tabella 2.3: Differenze tra le quattro tipologie di strategie manutentive

Capitolo 3

Stato dell'arte

3.1 Data analytics

La raccolta di dati in tempo reale dal processo produttivo e la loro analisi è un'attività fondamentale nell'era dell'Industria 4.0 in quanto è in grado di garantire vantaggi competitivi all'impresa.

L'espressione *data analytics* si riferisce all'analisi di set di dati con l'obiettivo di ricavare informazioni rilevanti per il business che facilitano il processo di *decision making*. Ad esempio, le imprese possono sfruttare le tecniche di *data analytics* per aumentare il fatturato: talvolta è sufficiente analizzare accuratamente i dati delle vendite per identificare nuove opportunità di business, migliorare le proprie performance ed individuare strategie più efficaci per conquistare una quota di mercato maggiore; inoltre, l'analisi dei dati può essere utile per prevedere la domanda futura dei consumatori, basandosi sul loro comportamento e le loro preferenze, oppure per migliorare il rapporto con i clienti proponendo vendite mirate. [30] [31]

Secondo l'Osservatorio di Big Data Analytics e Business Intelligence 2019 del Politecnico di Milano, il valore del mercato Analytics si aggira intorno ai 1.7 miliardi di euro, con un incremento rispetto all'anno precedente del 23%: ciò riflette una sempre maggiore consapevolezza dell'importanza della *data analytics* nella definizione delle strategie aziendali. [32] Nel fatturato le principali voci di spesa relative agli strumenti di *data analytics* si distribuiscono come mostrato in Figura 3.1. In particolare, le spese riguardo al software sono composte per il 53% da strumenti per la visualizzazione dei risultati delle analisi, mentre il restante 47% è costituito da tools necessari per gestione dei dati raccolti.

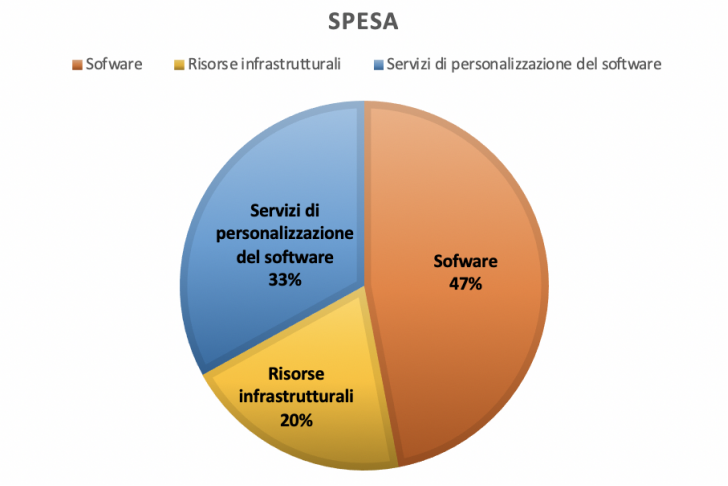


Figura 3.1: Distribuzione delle voci di spesa del mercato dell'Analytics

L'uso delle tecniche di *data analytics* offre vantaggi ad aziende appartenenti ad un ampio range di settori. Ad esempio, le banche possono sfruttare le analisi dei dati per monitorare i prelievi così da prevedere frodi o furti di identità; gli operatori di rete mobile esaminano i dati degli abbonati per prevedere l'abbandono della compagnia; le imprese manifatturiere sfruttano i dati raccolti in tempo reale dai macchinari per valutare eventuali interventi di manutenzione; le società di e-commerce e i fornitori di servizi di marketing analizzano i click sulle pagine Web per individuare i potenziali acquirenti di un determinato prodotto / servizio e per proporre advertising mirato alle preferenze dei consumatori; infine, le organizzazioni sanitarie raccolgono dati sui pazienti per valutare l'efficacia dei trattamenti su diverse patologie [30]. I settori maggiormente interessati secondo lo studio dell'Osservatorio di Big Data Analytics e Business Intelligence sono riportati in Figura 3.2.

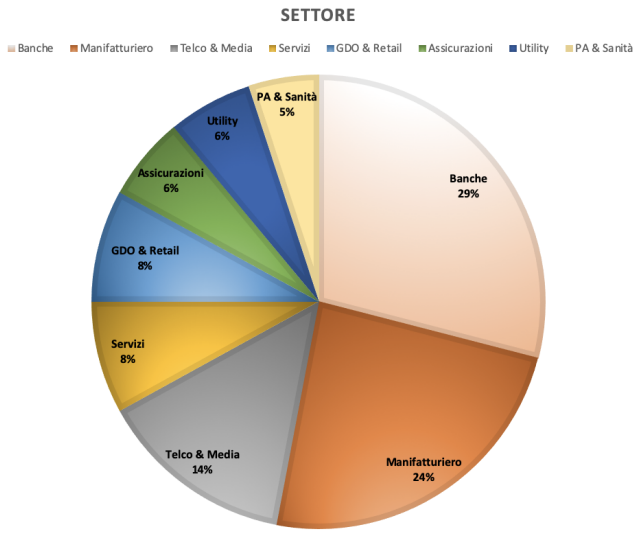


Figura 3.2: Investimenti nel mercato Analytics per settore

3.2 Data Mining

Le tecniche di *data analytics* fanno parte dell'ambito del *data mining*, ovvero una metodologia utilizzata per scoprire informazioni utili e pattern nascosti a partire da grandi collezioni di dati; inoltre, grazie alle tecniche di *data mining* è possibile prevedere il valore futuro di determinate osservazioni. Occorre sottolineare che non tutte le attività di reperimento delle informazioni rientrano nel campo del *data mining*. Ad esempio, la Web Search non appartiene a quest'area, mentre raggruppare un insieme di documenti restituiti dal motore di ricerca in base ad informazioni di contesto fa parte della disciplina.

Con lo sviluppo tecnologico, le analisi dei dati tradizionali hanno riscontrato numerose difficoltà legate alle caratteristiche dei nuovi dataset, in particolare riferimento a [33]:

- **Scalabilità**: attualmente, la dimensione dei dataset più comuni è dell'ordine di gigabytes, terabytes e petabytes; tuttavia, le tecniche di analisi classiche non sono in grado di trattare queste collezioni di dati.

- **Elevata dimensionalità dei dati**: le tecniche analisi dei dati tradizionali lavorano bene con dataset formati da poche features. Attualmente, è sempre più comune dover trattare dataset formati da centinaia o migliaia di attributi. Le tecniche tradizionali non sono in grado di gestire questo aspetto in quanto per molti algoritmi la complessità computazionale cresce rapidamente all'aumentare del numero delle caratteristiche.

- **Complessità ed eterogeneità dei dati**: fino a qualche decennio fa, i dataset contenevano attributi dello stesso tipo, tipicamente continui o categorici. Oggi, invece, i dati sono complessi e provengono da diverse fonti: le tecniche tradizionali non sono in grado di gestire questo tipo di attributi.

- **Analisi non tradizionali**: le analisi classiche basati sul paradigma "ipotesi e test" richiedono un processo estremamente laborioso, in quanto per rispondere adeguatamente alle richieste bisogna generare e valutare migliaia di ipotesi: da qui, la necessità di automatizzare il processo di valutazione delle ipotesi e test.

Il *data mining* nasce allo scopo di fronteggiare le sfide elencate precedentemente e si caratterizza come un insieme di strumenti efficienti e scalabili in grado di gestire diversi tipi di dati. La base su cui si fonda il *data mining* è costituita da un lato da tecniche di Statistica, come stima e verifica di ipotesi, e dall'altro da Intelligenza Artificiale, riconoscimento di pattern e *Machine Learning*. Inoltre, altre aree giocano un ruolo fondamentale nella definizione di *data mining* come nozioni di gestione dei database e tecniche di *parallel computing*.

Gli obiettivi del *data mining* si dividono principalmente in due categorie di analisi, descrittiva e predittiva.

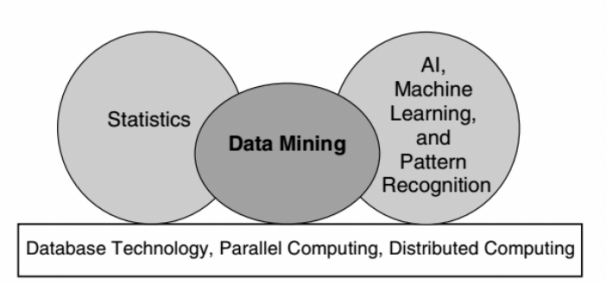


Figura 3.3: Il data mining come incontro di numerose discipline [33]

L'analisi descrittiva o esplorativa è un approccio *data driven*: non si ha conoscenza a priori sull'obiettivo specifico che si intende raggiungere, ma si esplorano i dati allo scopo di estrarre informazioni implicite, precedentemente sconosciute e potenzialmente utili. Ad esempio, le tecniche esplorative sono utili per identificare segmenti di clienti con caratteristiche comuni.

Le tecniche predittive hanno lo scopo di prevedere valori sconosciuti o futuri di altre variabili sfruttando quelle note. In questo caso, dunque, è necessario avere conoscenza a priori sul dato, e, quindi, avere a disposizione serie storiche che permettono di descrivere il fenomeno oggetto di studio. Inoltre, per poter costruire un modello predittivo è necessario conoscere l'obiettivo specifico. Esempi di applicazione riguardano la previsione del comportamento futuro di un cliente oppure di un macchinario, in particolare se si guasterà entro un determinato intervallo di tempo o se continuerà a funzionare correttamente [33].

3.3 Machine Learning

Con l'espressione **machine learning** o apprendimento automatico si intende la scienza che si occupa del problema di "costruire programmi per computer che migliorino automaticamente con l'esperienza" [34]. Negli ultimi anni, grazie allo sviluppo di tecniche di *machine learning* sempre più efficaci, è stato possibile implementare queste metodologie nell'ambito dell'Industria 4.0 al fine di rendere sempre più innovative e intelligenti le Smart Factories: durante il processo produttivo, le tecniche di *machine learning* permettono di verificare la presenza di oggetti e le loro caratteristiche oppure eventuali anomalie. Queste informazioni sono impiegate nell'ambito del controllo qualità, aiutando le fabbriche ad aumentare l'efficienza e ridurre gli sprechi. Inoltre, grazie ai sensori posizionati sui macchinari, è possibile implementare soluzioni di manutenzione predittiva così da intervenire sul macchinario solo quando realmente necessario e prima che si verifichi un guasto.

Altre applicazioni legate all'apprendimento automatico interessano il settore Finance per la prevenzione delle frodi o furti di identità, analizzando i comportamenti degli utenti e l'utilizzo delle carte di credito; il settore della sicurezza informatica, per creare filtri

anti-spam sempre più intelligenti in grado di intercettare e-mail potenzialmente sospette e di eliminarle prima che siano visualizzate nella casella di posta; l'ambito medico, in cui le tecniche di *machine learning* permettono di diagnosticare tumori e altre patologie con una tempestività sempre maggiore; inoltre, il *machine learning* è impiegato nei sistemi di raccomandazione utilizzati sui siti di e-commerce come Amazon o piattaforme come Spotify per mostrare advertising mirata alle ricerche dei consumatori; grazie all'apprendimento supervisionato, sono possibili realtà come il riconoscimento di immagini e suoni e sistemi di self-driving, che sfruttano queste tecniche per conoscere l'ambiente circostante (con i dati raccolti i sensori o dal GPS) e adattare di conseguenza il loro comportamento [35].

Sostanzialmente, l'obiettivo degli algoritmi di *machine learning* è quello di analizzare dati al fine di ottenere informazioni utili, senza utilizzare modelli matematici noti a priori.

È possibile individuare due principali tipologie di tecniche di *machine learning*:

- **Tecniche di apprendimento supervisionato**, che si pongono l'obiettivo di adattare modelli matematici alla classificazione di dati di cui sono note le classi di appartenenza;
- **Tecniche di apprendimento non supervisionato**, che cercano di individuare pattern e strutture ricorrenti in collezioni di dati di cui non sono note le classi di appartenenza.

In altre parole, le tecniche di apprendimento supervisionato servono a costruire modelli predittivi partendo dalla conoscenza di dati di input e dei rispettivi output, o classi di appartenenza. Tali modelli, una volta creati, vengono utilizzati per classificare dati di cui non si conoscono le classi di appartenenza. Le tecniche di apprendimento non supervisionato permettono, invece, di individuare nei dati pattern nascosti o strutture intrinseche.

3.4 Knowledge Discovery Process

L'obiettivo ultimo del *data analytics* è l'estrazione della conoscenza e le tecniche di *data mining* sono parte integrante di questo processo che viene chiamato **Knowledge Discovery from Data (KDD)**. Il KDD è un processo iterativo e interattivo che coinvolge tecniche di estrazione della conoscenza adeguate in base all'obiettivo dell'analisi. Ogni algoritmo utilizzato nel work flow dovrà essere configurato opportunamente dall'analista.

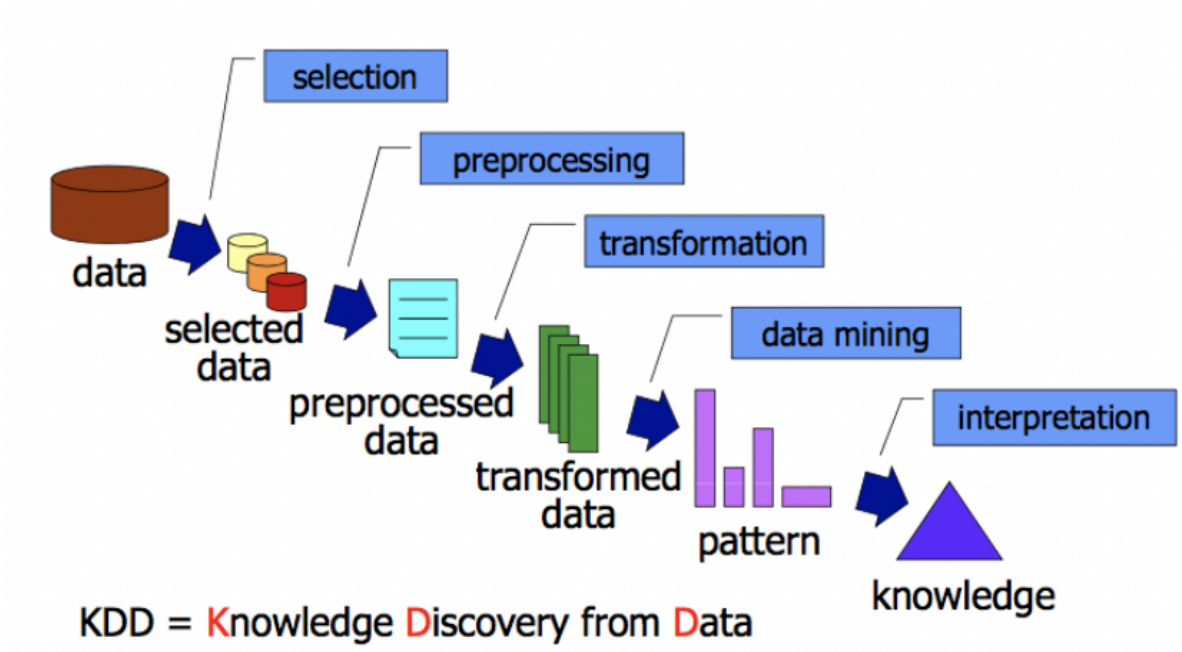


Figura 3.4: Knowledge Discovery from Data (KDD)

Innanzitutto, è necessario consolidare e approfondire la conoscenza del dominio di applicazione e la definizione degli obiettivi dell'utente finale. A questo proposito, risulta fondamentale la presenza di un esperto di dominio, ovvero una figura che ha una solida conoscenza del contesto di applicazione che avrà il compito di selezionare gli algoritmi che rispettano le leggi fisiche e modellano correttamente gli eventi oggetto di studio. Inoltre, l'esperto di dominio sarà in grado di interpretare i risultati delle analisi e di individuare le cause sottostanti ad eventuali problematiche.

3.4.1 Data Selection

Il dataset a disposizione è costituito generalmente da un'ingente quantità di dati eterogenei, quali dati testuali, dati numerici, immagini, video e molto altro. Elaborare una tale mole di dati sarebbe problematico dal punto di vista computazionale. Dunque, è fondamentale stabilire quali tipi di dati sono utili a rispondere adeguatamente al problema che si vuole risolvere e, a questo scopo, si seleziona una porzione rilevante del dataset.

3.4.2 Preprocessing

Nella fase di preprocessing vengono effettuate operazioni di base necessarie per filtrare i dati e prepararli per le analisi successive.

In primo luogo, viene eseguita la procedura di *data cleaning* allo scopo di trattare nel modo più appropriato:

- **Missing values**: talvolta, accade che i dati siano mancanti a causa di inefficienze o malfunzionamenti durante il processo di data-entry. In questi casi, si può intervenire eliminando i record con dati mancanti oppure effettuando delle stime dei valori assenti, ad esempio sostituendo il valore medio.
- **Dati rumorosi e outliers**: i dati rumorosi hanno delle caratteristiche tali da renderli significativamente differenti rispetto alla maggior parte dei valori del dataset. La presenza di queste anomalie può essere dovuta a errori casuali, limitazioni tecnologiche oppure malfunzionamenti della strumentazione utilizzata per misurare le variabili rilevanti per l'analisi. Per ovviare a questa problematica si possono effettuare delle analisi di clustering o di regressione per individuare i valori estremi.

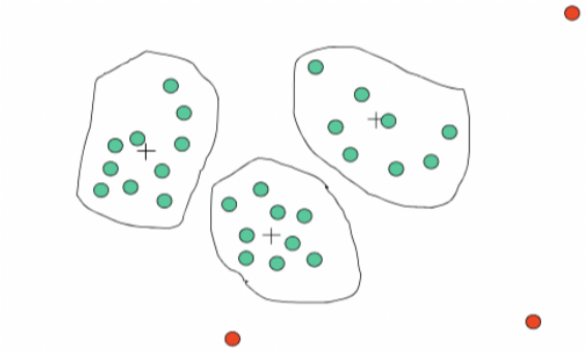


Figura 3.5: Identificazione degli outliers tramite clustering [36]

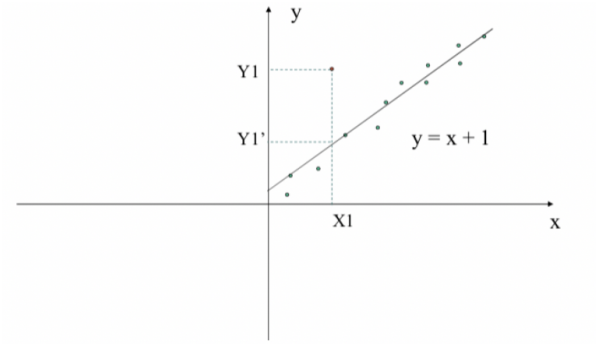


Figura 3.6: Identificazione degli outliers tramite regressione lineare [36]

In Figura 3.5 è possibile vedere come il clustering (in particolare la tipologia di partizionamento non completo) sia grado di identificare gruppi di dati con caratteristiche simili e di isolare valori estremamente diversi (i punti in rosso). In Figura 5.5, invece, identificazione degli outliers viene effettuata con la regressione lineare nella quale si osserva che i punti rumorosi sono decisamente distanziati dalla retta che approssima i dati.

A seguito della procedura di *data cleaning*, si esegue l'operazione di *data integration*; talvolta, per poter affrontare correttamente il problema, è necessario procedere ad un'integrazione dei dati a disposizione: per esempio, se il fenomeno da studiare è il livello di inquinamento dell'aria, potrebbe essere interessante collegare questi dati a quelli riferiti alle condizioni metereologiche.

3.4.3 Data Transformation

Per migliorare i risultati delle analisi, è utile operare una *data transformation*, ovvero un cambiamento della scala di riferimento. In questa fase si possono eseguire:

- **Normalizzazione:** si tratta di un tipo di trasformazione di dati fondamentale soprattutto quando si utilizzano algoritmi di clustering. Esistono diverse tecniche e non è possibile determinare a priori quale tipologia di normalizzazione sia ottimale utilizzare. Le tecniche più diffuse sono:

  - **Min-max:** i dati vengono riportati in un intervallo prefissato, tipicamente tra 0 e 1. Questa tipologia di normalizzazione è particolarmente indicata quando non si conosce la distribuzione dei dati oppure quando si ha conoscenza che la distribuzione non è gaussiana. Per ottenere la normalizzazione min-max, la trasformazione da effettuare è:

    $$z_i = \frac{x_i - min(x)}{max(x) - min(x)}$$

    dove x_i è il valore che si vuole standardizzare, $min(x)$ e $max(x)$ sono rispettivamente valore minimo e valore massimo assunti dal campione.

  - **Z-score:** la standardizzazione dei dati è il processo di ridimensionamento di uno o più attributi in modo che abbiano un valore medio di 0 e una deviazione standard di 1. Secondo questo criterio, quindi, le caratteristiche avranno le proprietà distribuzione normale standardizzata. Questa tipologia di normalizzazione è particolarmente indicata perché facilita la gestione dei valori anomali. Il valore z-score z_i è calcolato come segue:

    $$z_i = \frac{x_i - \mu}{\sigma}$$

    dove x_i è il valore che si vuole standardizzare, μ e σ sono rispettivamente media e deviazione standard dei campioni.

- **Aggregazione:** è un'operazione che consiste nel cambiare la granularità dei dati allo scopo di renderli più stabili e, al tempo stesso, di ridurre lo spazio occupato e il tempo di esecuzione degli algoritmi. Per esempio, le vendite giornaliere possono essere aggregate in vendite mensili. Tuttavia, bisogna essere cauti e non esagerare con l'aggregazione in quanto, naturalmente, si ha una perdita di informazioni potenzialmente interessanti.

Infine, l'ultima fase del processo di preparazione dei dati si conclude con le operazioni di *data reduction*. Infatti, il dataset può avere dimensioni dell'ordine dei terabytes e la sua analisi integrale potrebbe richiedere tempi di calcolo elevati. A questo scopo, si utilizzano delle tecniche per ridurre la dimensione del dataset, mantenendo al minimo la perdita di informazioni. La riduzione della dimensione può essere effettuata sia rispetto alle righe, cioè il numero di record, che alle colonne, cioè alle variabili note del problema.

- **Sampling o campionamento**: quando si effettua una riduzione del dato rispetto al numero di righe si parla di sampling o campionamento. Effettuare un campionamento significa identificare un sottoinsieme rappresentativo dei dati di partenza, cioè che mantenga le stesse caratteristiche della collezione iniziale. Esistono diverse strategie di sampling, tra cui:

  - **Sampling randomico**: secondo questa tipologia, ogni elemento ha la stessa probabilità di essere estratto. Può essere senza rimpiazzamento, se un elemento, una volta estratto, non viene rimesso nella collezione dei dati di partenza. In questo caso gli oggetti possono essere selezionati più volte. Esiste anche il sampling randomico con rimpiazzamento, se un elemento, una volta estratto, viene incluso nuovamente nella popolazione di dati iniziale.
  - **Sampling stratificato**: il sampling randomico presenta lo svantaggio di non rappresentare adeguatamente la classe di dati meno frequente; il sampling stratificato, invece, consiste nell'estrarre un campione proporzionale delle classi di interesse nella base dati di partenza: con questa strategia anche le classi meno numerose sono considerate opportunamente.

- **Feature selection**: quando si effettua una riduzione del dato rispetto al numero di attributi di parla di feature selection. Questa operazione si esegue allo scopo di eliminare attributi ridondanti, le cui informazioni sono già fornite da altre variabili, e attributi irrilevanti, che non contengono informazioni di valore. Inoltre, la rimozione degli attributi viene eseguita anche per ridurre il tempo di elaborazione e aumentare l'interpretabilità del risultato. Selezionare il sottoinsieme di caratteristiche utili richiede un approccio sistematico e l'intervento dell'esperto di dominio. Una possibile strategia di azione consiste nel calcolare, utilizzando tecniche diverse, l'indice di correlazione tra tutte le coppie di attributi e, dopo aver selezionato una soglia accettabile, si scartano tutti gli attributi che presentano un indice di correlazione superiore alla soglia impostata.

- **Discretizzazione**: si esegue una riduzione del numero di valori ammissibili da una variabile continua in un insieme di intervalli: in questo modo si riduce la cardinalità del dominio. Anche in questo caso, non esiste una modalità definitiva per determinare la migliore tipologia di discretizzazione, ma si utilizza la strategia che consente di produrre il migliore risultato degli algoritmi di *data mining*. Le tecniche di discretizzazione più utilizzate prevedono di suddividere il dominio dell'attributo in N intervalli della stessa ampiezza oppure della stessa frequenza; in alternativa, si può usare il clustering monodimensionale, soluzione che si adatta bene a dati sparsi e outliers.

Per ridurre la dimensionalità di un dataset, infine, esistono altri approcci che sfruttano tecniche di algebra lineare, utilizzate soprattutto quando gli attributi sono di tipo continuo. Tra le più diffuse si evidenziano la Principal Component Analysis (PCA) e il Singular Value Decomposition (SVD). Queste metodologie consentono di semplificare il lavoro di manipolazione delle caratteristiche e di migliorare i risultati dei classificatori.

3.5 Knowledge Extraction

Le ultime fasi del processo del KDD consistono nell'applicazione delle tecniche di *data mining*. A seconda dell'obiettivo dell'analisi, la conoscenza da estrarre sarà diversa e, quindi, si dovrà selezionare l'algoritmo che meglio risponda alla domanda iniziale. Il processo si conclude quando, una volta estratte le informazioni ricercate, il data analyst è in grado di comunicarle appropriatamente ai manager così che questi possano tradurre la conoscenza acquisita in azioni utili per il business e prendere decisioni più consapevoli.

3.5.1 Tecniche di analisi: Regole di associazione

Tra le tecniche di *data mining* utilizzate per effettuare analisi esplorative, che quindi non richiedono conoscenza a priori sui dati, rientrano le **regole di associazione**. Per la loro definizione è fondamentale chiarire dei concetti basilari, in primis quello di *itemset*, cioè una insieme di oggetti che appartiene alla collezione di dati; in secondo luogo, è rilevante il concetto di *transazione*, cioè un sottoinsieme di elementi frequenti e non ordinati dell'itemset. Per questo motivo, il database è chiamato *transazionale*. Un esempio di base dati transazionale è il carrello della spesa.

L'obiettivo delle regole di associazione è estrarre le correlazioni frequenti all'interno di una base dati transazionale. L'uso delle regole di associazione ha trovato particolare diffusione nell'ambito della *market basket analysis* (MBA), con l'obiettivo di rappresentare le abitudini di acquisto dei consumatori allo scopo di trovare relazioni tra i prodotti comprati. Ad esempio, è stato osservato che i clienti che acquistano i pannolini, probabilmente acquisteranno anche la birra. Le informazioni ricavate dalle regole di associazione possono essere utilizzate, ad esempio, per organizzare il layout degli scaffali di un negozio posizionando gli elementi vicini oppure offrendo in promozione gli oggetti acquistati congiuntamente più di frequente.

Una regola di associazione r si presenta nella forma:

$$r : A \implies B$$

dove A è l'elemento che costituisce il corpo della regola, mentre l'elemento B è detto testa della regola; corpo e testa possono avere numerosità variabile. Per misurare la robustezza (o forza) della regola, è possibile calcolare metriche quali:

- **Supporto**: data la regola di associazione $r : A \implies B$, il supporto misura la frazione di transazioni che contengono sia A che B.

$$sup = \frac{\#A, B}{|T|}$$

  dove $|T|$ è la cardinalità della base dati transazionale.

- **Confidenza**: data la regola di associazione $r : A \implies B$, la confidenza misura la probabilità che in una transazione ci sia B, dato che è già presente A. Questa metrica valuta la forza dell'implicazione.

$$conf = \frac{sup(A, B)}{sup(A)}$$

- **Lift**: indica come l'occorrenza di un evento fa aumentare le occorrenze dell'altro.
  - Se il $lift = 1$, la regola di associazione è irrilevante e, dunque, da scartare in quanto indica che gli eventi sono indipendenti.
  - Se il $lift > 1$, gli eventi sono correlati positivamente, cioè la probabilità che in una transazione ci sia B, dato che è già presente A, è maggiore della probabilità che sia presente B.
  - Se il $lift < 1$, gli eventi sono correlati negativamente e la regola diventa $r : A \implies \neg B$

Le regole di associazione estratte devono soddisfare i requisiti su supporto e confidenza, ovvero:

$$support \geq minsup$$

$$confidence \geq minconf$$

Dove $minsup$ e $minconf$ sono delle soglie impostate dall'utente a priori. Non esistono delle regole generali per trovare la migliore configurazione dei parametri, ma bisogna fare un trade-off in base ai dati. In particolare, se si imposta un supporto troppo basso si troverebbe un numero eccessivamente elevato di regole; viceversa, se si seleziona un valore di supporto troppo alto si troverebbero regole di associazioni ovvie.

Tra le tecniche più diffuse utilizzate per estrarre le regole di associazione si riportano l'algoritmo Apriori, un metodo iterativo che ad ogni $k - esima$ iterazione estrae itemset lunghi k. Tuttavia, quando il dataset è complesso e ha tanti attributi l'algoritmo Apriori è troppo oneroso dal punto di vista della complessità computazionale. In questi casi, si utilizza l'algoritmo FP-growth, un altro metodo più rapido ed efficiente per estrarre i pattern più frequenti.

3.5.2 Tecniche di analisi: Cluster Analysis

Tra le tecniche di *data mining* utilizzate nella fase di estrazione della conoscenza figura la *cluster analysis*. L'obiettivo è quello di raggruppare - o segmentare - collezioni di oggetti in sottoinsiemi, chiamati cluster, tali per cui oggetti nello stesso cluster abbiano caratteristiche simili tra loro, differentemente da elementi assegnati a cluster diversi. In questo senso, la distanza tra punti dello stesso cluster deve essere minimizzata mentre la distanza tra punti appartenenti a cluster diversi deve essere massimizzata. Figura 3.7.

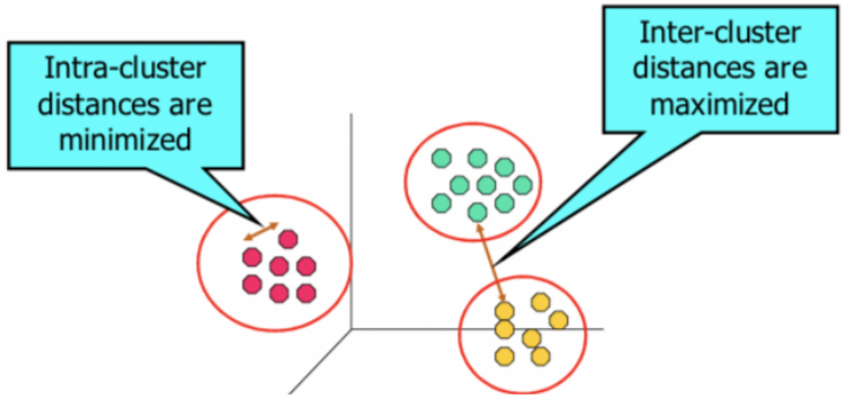


Figura 3.7: Distanza intra-cluster e inter-cluster [33]

Nella cluster analysis è fondamentale la definizione della misura del grado di somiglianza tra gli oggetti; infatti, la formazione di una partizione di oggetti piuttosto che di un'altra può dipendere molto dal tipo di misura utilizzata. Quando si devono trattare dati di tipo numerico la funzione che calcola la somiglianza prende il nome di **distanza**: in questo caso la funzione di distanza deve essere minimizzata, in quanto tanto più gli oggetti sono vicini tanto più hanno caratteristiche simili, cioè appartengono allo stesso cluster. Esempi di funzioni di distanza sono la funzione di distanza Euclidea oppure la distanza di Minkowski, ovvero la generalizzazione della distanza Euclidea. Quando si hanno a disposizione dati di tipo testuale la funzione che calcola la somiglianza prende il nome di **similarità**: in questi casi la similarità deve essere massimizzata, in quanto tanto più due testi sono simili tanto più sono vicini, cioè appartengono allo stesso cluster.

Il clustering può essere eseguito per raggiungere molteplici obiettivi, per esempio per effettuare statistiche descrittive che verifichino che i dati siano divisibili in sottogruppi distinti, ciascuno dei quali con proprietà diverse. Si immagini una collezione in cui sono memorizzati i dati dei clienti di una banca: eseguendo la cluster analysis, si ottiene un partizionamento del dataset di partenza che consente di individuare clienti dalle caratteristiche simili e, quindi, di applicare delle campagne promozionali mirate alle diverse tipologie di consumatori.

Con il termine clustering si intende un insieme di gruppi in cui il database è stato suddiviso. Esistono diverse tipologie di algoritmi di clustering e l'uso di una tecnica piuttosto che di un'altra dipende dall'obiettivo dell'analisi. Vi è una distinzione fondamentale tra:

- **Clustering partizionale**, in cui si generano delle partizioni indipendenti della collezione.
- **Clustering gerarchico**, in cui la soluzione fornita dall'algoritmo è un insieme di cluster nidificati, come in un albero gerarchico.

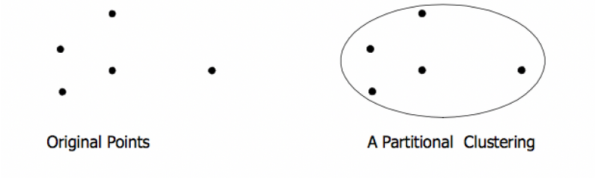


Figura 3.8: Clustering partizionale [33]

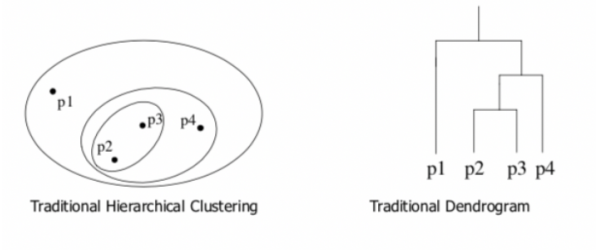


Figura 3.9: Clustering gerarchico [33]

Oltre queste tipologie, tuttavia, esistono altre distinzioni tra insiemi di cluster, tra cui:

- **Partizionamento esclusivo o non esclusivo**: nella soluzione non esclusiva i punti possono appartenere a più cluster; questa tecnica è utile per rappresentare i punti di confine oppure più tipi di classi.
- **Partizionamento completo o parziale**: nella soluzione parziale alcuni punti potrebbero non appartenere a nessun cluster, mentre nel tipo completo tutti i punti appartengono ad un solo cluster.
- **Partizionamento fuzzy o non fuzzy**: in un fuzzy clustering un punto appartiene a tutti i cluster presenti con un peso che varia tra 0 e 1; la somma dei pesi per ciascun punto deve essere pari a 1.
- **Partizionamento eterogeneo o omogeneo**: nel clustering eterogeneo i cluster possono avere forma, dimensione e densità molto differenti tra loro, fatto che non si verifica nel partizionamento omogeneo.

La nozione di cluster, tuttavia, può essere ambigua. Si immagini di avere venti punti in uno spazio: questi possono essere raggruppati in diversi modi; infatti, si può ottenere una partizione in due gruppi oppure da quattro o sei. Non si sa a priori qual è il modo migliore di raggruppare i dati, dipende dal tipo di visualizzazione che si desidera ottenere.

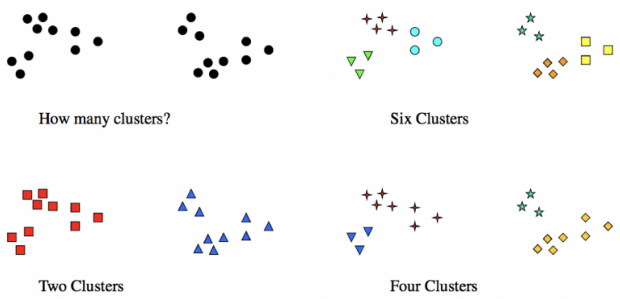


Figura 3.10: Tre modi diversi per effettuare clustering sullo stesso set di dati [33]

Esistono diversi algoritmi di clustering, tra i più diffusi e popolari si citano il K-means, DB Scan e Hierarchical Clustering, di cui, tuttavia, si tratterà più diffusamente nel capitolo successivo.

3.5.3 Tecniche di analisi: Classificazione

Tra le tecniche appartenenti all'ambito del *data mining* figura la classificazione, i cui algoritmi sfruttano metodi di apprendimento automatico supervisionato con l'obiettivo di fare predizioni sui dati. In particolare, viene fornito come input un insieme di dati già etichettati, cioè con l'indicazione della classe di appartenenza dei punti: grazie a queste informazioni, viene creato un modello di classificazione che sarà utilizzato sui dati non etichettati allo scopo di predire la classe di appartenenza.

Normalmente, il dataset viene suddiviso in *train set*, la porzione di dati che verrà utilizzata per costruire il modello di predizione, e *test set*, la restante frazione di dataset che verrà sfruttata per testare il modello precedentemente costruito. Prima di essere rilasciato in produzione, il modello deve essere validato: quando si lavora con dataset di dimensioni contenute, l'approccio utilizzato più diffuso è la *K-Fold Cross Validation*, dove K è un parametro impostato dall'utente. Il processo di *Cross Validation* prevede di dividere il dataset in K partizionamenti. Tra questi, $K-1$ set di dati sono utilizzati per la fase di training mentre il set rimanente è tenuto da parte per la fase di test. L'algoritmo è addestrato e testato K volte; ad ogni iterazione un nuovo set viene usato per la fase di test mentre i restanti sono impiegati per il training. Infine, il risultato della *K-Fold Cross Validation* è la media dei risultati ottenuti su ciascun set.

Tuttavia, quando la dimensione del dataset è eccessivamente elevata, questo approccio è sostituito da un partizionamento fisso in cui si divide la collezione di dati di partenza in *train set* e *test set* utilizzando delle proporzioni (tipicamente si utilizzano 2/3 dei dati per il *train* e il restante 1/3 come *test*).

Un esempio di applicazione delle tecniche di classificazione riguarda la previsione dell'uso fraudolento delle carte di credito, a partire dalle informazioni riguardo ai loro possessori

e dalla conoscenza delle precedenti transazioni etichettate come illecite. Si costruisce un modello di classificazione utilizzato per individuare i comportamenti fraudolenti delle future transazioni relativamente a una specifica carta di credito.

Generalmente, non si può scegliere a priori l'algoritmo da utilizzare, ma è necessario valutarli in base a parametri quali:

- **Accuratezza**: metrica che rappresenta la capacità del modello di predire correttamente le etichette e si calcola come

$$Accuratezza = \frac{\#(Etichette\,predette\,correttamente)}{\#(Etichette\,totali)}$$

  L'accuratezza è complessiva del modello; tuttavia, quando si ha un dataset sbilanciato nella distribuzione delle classi, l'accuratezza non è una metrica affidabile ed è necessario ricorrere ad altre misure della qualità della previsione quali precisione e richiamo.

- **Efficienza**: si parla di efficienza in termini di tempo di esecuzione sia di training che di test. Generalmente, quasi tutti gli algoritmi hanno tempi di predizione abbastanza contenuti ma tempi di training variabili. Quando si sceglie un algoritmo bisogna decidere se questo aspetto è un parametro da valutare oppure se è trascurabile.

- **Scalabilità**: si valuta se i tempi di esecuzione dell'algoritmo crescono all'aumentare del numero di attributi o della dimensione del *training set*.

- **Robustezza**: si intende stabilire se il classificatore è robusto al rumore; in altre parole, si vuole valutare se, in presenza di rumore, il classificatore è in grado di effettuare correttamente la predizione.

- **Interpretabilità**: si intende valutare se chi osserva l'output del modello è in grado di interpretare e spiegare le decisioni prese dall'algoritmo.

Come anticipato precedentemente, per valutare la bontà del modello di classificazione generato quando il dataset è sbilanciato nella distribuzione delle classi è utile osservare i valori assunti da **precisione** e **richiamo** per ciascuna classe. La prima metrica misura il rapporto tra il numero di elementi correttamente assegnati alla classe e il numero totale di elementi assegnati a tale classe; la seconda metrica, invece, misura il rapporto tra il numero di elementi correttamente assegnati alla classe e il numero totale di elementi appartenenti a tale classe.

Esistono molteplici tecniche di classificazione, tra cui le più diffuse sono:

- **Albero delle decisioni**: si tratta di un modello predittivo, in cui ogni nodo rappresenta un test sugli attributi; ogni arco verso un nodo figlio rappresenta un percorso in

base all'esito del test precedente; ogni nodo foglia rappresenta il valore predetto per la variabile obiettivo. Talvolta, è utile selezionare un criterio di arresto, detto anche *pruning*, al fine di determinare la massima profondità dell'albero. Infatti, una crescita eccessiva dell'albero potrebbe generare problemi di *overfitting*, cioè la creazione di un modello troppo specifico in cui l'errore diminuisce nel *training set*, ma aumenta nel *test set*. Figura 3.11.

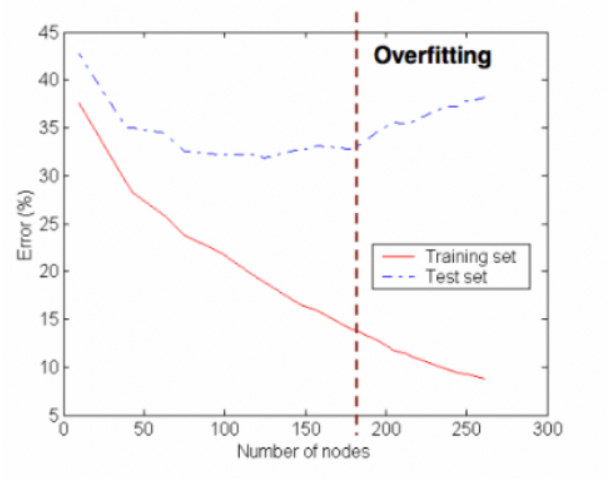


Figura 3.11: Overfitting [33]

L'albero delle decisioni è un algoritmo poco oneroso dal punto di vista computazione ed è in grado di fornire una soluzione rapidamente; tendenzialmente, l'output è di facile interpretazione e i livelli di accuratezza sono abbastanza elevati. Tuttavia, questa tecnica non è indicata quando il dataset presenta valori mancanti.

- **Random Forest**: si tratta di un classificatore ottenuto dall'aggregazione di molteplici alberi delle decisioni allo scopo di aumentare l'accuratezza e di contenere l'overfitting. Tuttavia, all'aumentare della dimensione del *training set* le sue prestazioni calano.

- **Classificazione Bayesiana**: si tratta di un classificatore che prevede di calcolare la probabilità che un certo punto appartenga alla classe. È un classificatore efficiente in termini di tempi di esecuzione, di buona interpretabilità e robusto alla presenza di dati rumorosi. Tuttavia, il punto di debolezza principale di questa tecnica riguarda l'assunzione dell'ipotesi Naive, cioè che le variabili siano statisticamente indipendenti, senza la quale la generazione del modello risulta più difficoltosa; rilasciando l'ipotesi Naive, l'accuratezza del modello diminuisce significativamente.

- **Reti Neurali**: è una tecnica che intende imitare le capacità di apprendimento umano attraverso un'architettura molto simile al sistema nervoso. Il sistema nervoso è costituito da miliardi di neuroni che ricevono degli input dagli organi sensoriali, elaborano le informazioni e successivamente decidono come rispondere all'input.

Secondo lo stesso principio, le reti neurali ricevono una serie di informazioni in ingresso, le inviano ad uno o più "livelli nascosti" di elaborazione ed infine forniscono la predizione dell'etichetta di classe. Questa tecnica consente di costruire dei modelli molto accurati e robusti al rumore; tuttavia i modelli prodotti sono "black box" e, quindi, risultano di difficile interpretazione.

- **Support Vector Machine o SVM**: l'obiettivo dell'SVM è trovare l'iperpiano che, nello spazio delle features, separa i data-point delle classi nel miglior modo possibile, ovvero garantendo la maggiore distanza possibile tra le classi. I vettori di supporto, o support vector, sono i data-point più vicini all'iperpiano di separazione. Lo svantaggio di questa tecnica riguarda la scarsa interpretabilità del risultato quando il dataset è molto compresso.
- **K-Nearest Neighbors o K-NN**: si tratta di uno degli algoritmi di apprendimento supervisionato più semplici, sia dal punto di vista concettuale che implementativo. Il funzionamento dell'algoritmo è basato sulla somiglianza delle caratteristiche, solitamente calcolata con la distanza Euclidea: più un elemento è vicino a un data-point, più il KNN li considererà simili. L'algoritmo prevede di fissare un parametro K, scelto arbitrariamente, che identifica il numero di data-points più vicini: a questo punto, l'algoritmo valuta le K minime distanze dal punto da etichettare e lo assegna alla classe che ottiene il maggior numero di queste distanze. Ovviamente, tra le principali problematiche dell'algoritmo rientra proprio la scelta del K iniziale, in base al quale si avranno tempi di classificazione variabili.

3.5.4 Tecniche di analisi: Concept Drift

In un contesto come quello dell'Industria 4.0 in cui si raccolgono dati in tempo reale dai macchinari, frequentemente accade che questi subiscano delle modifiche nel tempo. Ciò comporta che i modelli predittivi che assumono una relazione statica tra le variabili in gioco abbiano delle performance scarse e degradanti e non consentano di ricavare informazioni significative e utili per il business. Secondo la mappatura statica, infatti, si presuppone che il modello costruito sui dati storici sia altrettanto valido in futuro sui dati nuovi e che le relazioni tra i dati di input e di output non cambino. Ciò è vero per molte applicazioni reali, ma non per tutte. Nel campo dell'apprendimento automatico il problema delle relazioni mutevoli tra i dati prende il nome di *concept drift*. Questo fenomeno porta con sé due importanti sfide: in prima battuta, bisogna essere in grado di rilevare quando si verifica il concept drift; in secondo luogo è necessario riuscire a mantenere il modello aggiornato ai nuovi dati senza dover ricostruire il modello da zero. [37]

La letteratura classifica diversi tipi di concept drift in base al tempo oppure alle previsioni. Dal punto di vista temporale, esistono quattro tipologie:

- **Sudden drift**: si verifica quando si individua precisamente l'istante temporale in cui i dati cambiano bruscamente.
- **Gradual drift**: accade quando i dati cambiano nel tempo, ma si alternano dati nuovi e dati vecchi. Tuttavia, al trascorrere del tempo la classe nuova domina sulla prima, che si degrada fino a sparire del tutto.
- **Incremental drift**: si verifica quando i dati cambiano gradualmente nel tempo, passando da un tipo all'altro.
- **Re-occurring drift**: si riferisce al caso in cui, in seguito al cambiamento dei dati ad un'altra tipologia, dopo un certo periodo di tempo si ripresenti la classe precedente. [38]

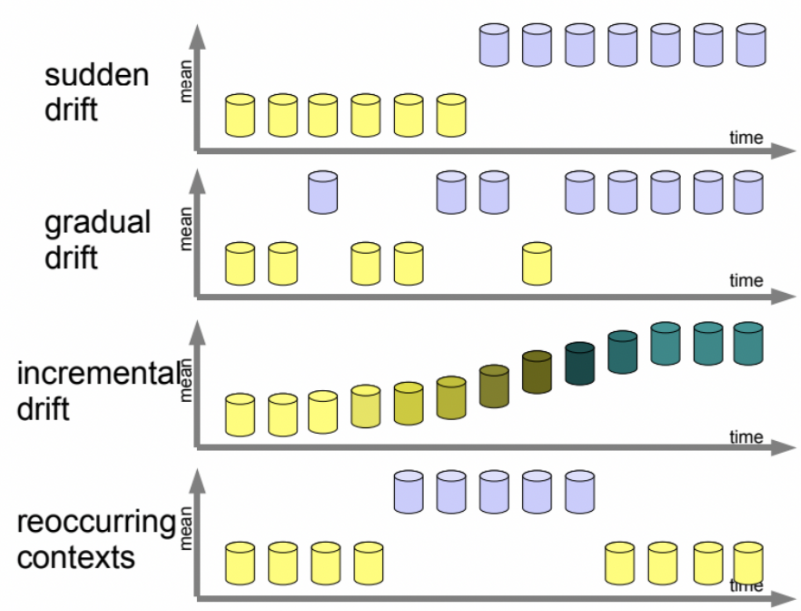


Figura 3.12: Le quattro tipologie di Concept Drift [39]

Dal punto di vista predittivo, esistono due tipologie di concept drift:

- **Real drift**: si riferisce alla probabilità che ci siano cambiamenti nella classe precedente. Per esempio, il real drift si verifica quando uno studente, normalmente interessato a leggere notizie sportive, per qualche motivo modifica le sue abitudini ed inizia a leggere notizie politiche.
- **Virtual drift**: si riferisce al cambiamento nella distribuzione dei dati in entrata. Per esempio, quando sta per iniziare un importante incontro sportivo, tutte le notizie relative a questo ambito si concentrano sull'evento attuale. Non cambia l'interesse per lo sport, ma la distribuzione delle notizie, che si concentrano solo sullo sport che sta per iniziare [39].

3.5.5 Anomaly detection

In un contesto come quello dell'Industria 4.0 in cui si raccolgono dati in tempo reale allo scopo di supportare attività come la manutenzione in base alle condizioni di salute dei macchinari, risulta di fondamentale importanza rilevare eventuali irregolarità. Le anomalie sostanzialmente sono pattern nascosti nei dati che si discostano dal comportamento tipico di una data variabile e la loro tempestiva identificazione spesso fornisce informazioni strategiche da applicare in diversi domini. Per esempio, un utilizzo anomalo della carta di credito potrebbe indicare una potenziale frode in corso; oppure la presenza di un elevato traffico in rete potrebbe riguardare un accesso non autorizzato. Poichè queste applicazioni necessitano di interventi risolutivi tempestivi per limitare eventuali danni, è essenziale che le tecniche utilizzate per rilevare le anomalie siano efficaci e di rapida esecuzione. Tra le possibili strategie da adottare a questo scopo vi è l'Isolation Forest, un metodo che costruisce un insieme di alberi a partire da un dataset di train. Sostanzialmente, l'Isolation Forest costruisce un modello in cui si traccia un confine tra i dati, separando gli elementi che mostrano un comportamento normale da quelli che risultano significativamente diversi. Il funzionamento del metodo è dovuto all'assunzione secondo cui le anomalie sono poche e differenti: infatti, l'Isolation Forest definisce le anomalie come quelle istanze che presentano un percorso mediamente breve nei rami degli alberi costruiti. Con questo algoritmo è sufficiente definire pochi parametri di input: il numero degli alberi da costruire, la dimensione del sotto-campione e la percentuale di outliers presenti nella collezione dati di partenza. É stato dimostrato che l'Isolation Forest è in grado di convergere rapidamente, permettendo di individuare le anomalie anche con una piccola quantità di alberi e di elementi nel sotto-campione.[40]

3.6 Interpretazione dei risultati ed estrazione della conoscenza

Il *data mining* crea pattern e modelli che possono costituire un valido supporto alle decisioni aziendali. L'ultima fase del KDD consiste nell'interpretazione dei risultati ottenuti dal processo attraverso grafici che visualizzano l'output delle analisi. Tuttavia, occorre valutare i modelli, cioè capire in quale misura questi possono essere utili per il business. Dunque, alla luce dei risultati ottenuti, è possibile rivedere una o più fasi dell'intero processo. Una volta stabilita la rilevanza del risultato, la conoscenza estratta deve essere consolidata, incorporata nel sistema informativa o nella documentazione relativa alle aree interessate.

Capitolo 4

Metodologia utilizzata

In questa tesi è proposta una metodologia in cui si ridefinisce il processo del KDD generale descritto nel capitolo precedente e si applicherà ad un caso di studio nell'ambito dell'Industria 4.0 a supporto della fase di manutenzione predittiva. Si utilizzerà come pipeline generale la metodologia mostrata in Figura 4.1.

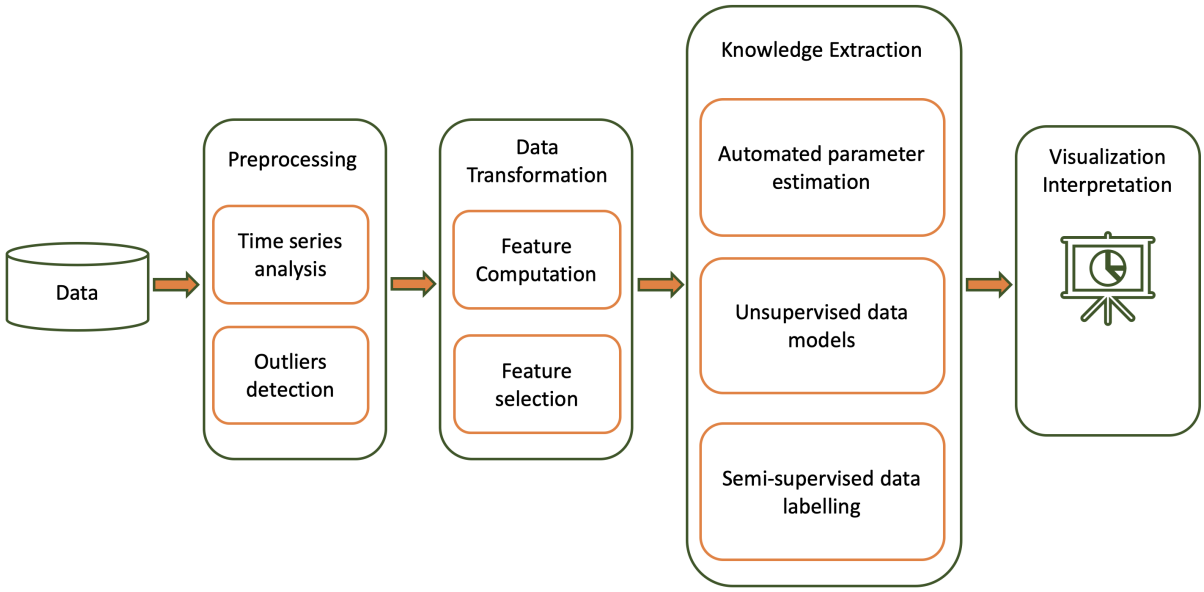


Figura 4.1: Metodologia semi-supervisionata implementata

Innanzitutto, si esegue un'analisi esplorativa dei dati a disposizione allo scopo di osservare i trend delle variabili in gioco ed individuare eventuali comportamenti o misurazioni anomale. Dopo aver filtrato i dati, si opera una *data trasformation* per ridurre la dimensionalità del dataset in cui si calcolano delle misure statistiche che descrivono i dati oggetto di studio. Inoltre, per diminuire ulteriormente la complessità del problema, si effettua una *feature selection* in cui si eliminano gli attributi maggiormente correlati che risulterebbero ridondanti.

Una volta completate le attività di preparazione dei dati, si passa alla fase di *Knowledge Extraction*. Tra le tecniche di data-mining tipicamente utilizzate nella fase di estrazione della conoscenza presentate nel capitolo precedente, in questa sede si impiega la cluster analysis, di cui verranno descritte più dettagliatamente gli algoritmi principali. Dopo aver preventivamente individuato i parametri migliori come input di ciascun algoritmo tramite tecniche basate sulla distribuzione dei dati, si costruisce un modello per suddividere i cicli di produzione. Infine, si valida il modello ottenuto confrontando le partizioni generate da ciascun algoritmo con la reale distribuzione dei dati.

4.1 K-Means

Il K-Means è una tecnica di clustering in grado di identificare K cluster, dove K è un valore arbitrario specificato a priori dall'utente. Si tratta di una tecnica **prototype-based** o **center-based** in cui i cluster generati sono rappresentati da un punto medio. L'algoritmo di clustering prende il nome di K-Means se i cluster generati sono descritti dal centroide, cioè un punto rappresentativo ottenuto dalla media di tutti i punti del cluster e che non necessariamente coincide con un data-point; l'algoritmo prende il nome di K-medoid se i cluster sono descritti dal medoide, cioè un punto rappresentativo appartenente alla collezione di dati. In questa sede, ci si soffermerà sul K-Means, uno dei primi e più diffusamente utilizzati algoritmi di clustering.

Il K-Means è un algoritmo di clustering di tipo partizionale, completo e omogeneo rispetto alla cardinalità, in quanto i cluster formati sono bilanciati, alla forma, proprietà che rende i cluster tipicamente di forma globulare, e alla densità, cioè il numero medio di punti presenti in una determinata area è costante.

La tecnica del K-Means è semplice ed intuitiva. L'utente sceglie arbitrariamente la quantità di cluster desiderati K, parametro che coincide con il numero di centroidi iniziali selezionati in modo casuale. Successivamente, ciascun punto nella collezione dei dati è assegnato al centroide più vicino e, alla fine del processo di allocazione, si sono formati K partizionamenti. Tuttavia, i centroidi iniziali di ciascun cluster non sono più rappresentativi: per questo motivo, si ricalcolano i centroidi sulla base dei cluster ottenuti al passaggio precedente e si ripete nuovamente il processo. Si esegue iterativamente la procedura di assegnazione e aggiornamento dei centroidi fin quando questi ultimi rimangono immutati.

Il K-Means è formalmente descritto dai seguenti step di esecuzione:

Algorithm 1 K-Means

1: Selezionare in modo casuale 4 punti come centroidi iniziali
2: Formare K cluster assegnando ciascun punto al centroide più vicino
3: Ricalcolare il centroide di ogni cluster
4: Ripetere i passi 2 e 3 finchè i centroidi non cambiano

Una fase fondamentale dell'algoritmo è l'assegnazione di ogni data-point al centroide più vicino. A questo scopo, è necessario utilizzare una funzione per calcolare la distanza. Per i dati in uno spazio Euclideo di solito si utilizza la distanza Euclidea, mentre per i dati di tipo testuale è più appropriata la similarità del coseno. Tuttavia, esistono anche altre funzioni di distanza da utilizzare a seconda della tipologia dei dati. Inoltre, si evince che, poiché il clustering è basato sul concetto di similarità e distanza, è fondamentale eseguire la normalizzazione dei dati nella fase di preprocessing.

Una delle problematiche del K-Means riguarda la scelta casuale dei K centroidi iniziali. Infatti, se si esegue una nuova sessione dell'algoritmo lasciando invariato il parametro K, i centroidi iniziali saranno diversi dato che questi sono scelti in maniera randomica: poiché il risultato finale dipende dalla scelta dei centroidi iniziali, la soluzione ottenuta sarà diversa da quella di partenza. È, dunque, necessario trovare una metrica che permetta di discriminare quali soluzioni siano migliori o peggiori.

Si consideri, per esempio, il caso in cui sia scelta come misura di distanza tra i punti la distanza Euclidea e come funzione obiettivo, che determina la qualità del clustering, lo scarto quadratico medio o SSE. In altre parole, si calcola l'errore di ogni punto, cioè la sua distanza dal centroide, e in seguito si sommano gli errori quadratici. Dati gli output prodotti da due sessioni del K-Means con lo stesso valore di K si sceglie il risultato che produce il valore di SSE minore in quanto ciò è indice di una maggiore coesione dei cluster. La definizione formale dell'SSE è la seguente:

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist(c_i, x_i)^2$$

dove $dist$ è la distanza Euclidea tra due oggetti nello spazio Euclideo, x_i è un punto generico nel cluster C_i e c_i è il punto rappresentativo del cluster.

Tipicamente, l'andamento dell'SSE è decrescente in K: man mano che il numero di cluster aumenta questi sono sempre più coesi e, dunque, l'SSE è minore. L'SSE può anche essere utile per la scelta iniziale del parametro K: il valore di K ottimale si ottiene selezionando il valore in corrispondenza del punto di massima decrescenza del SSE. Figura 4.2.

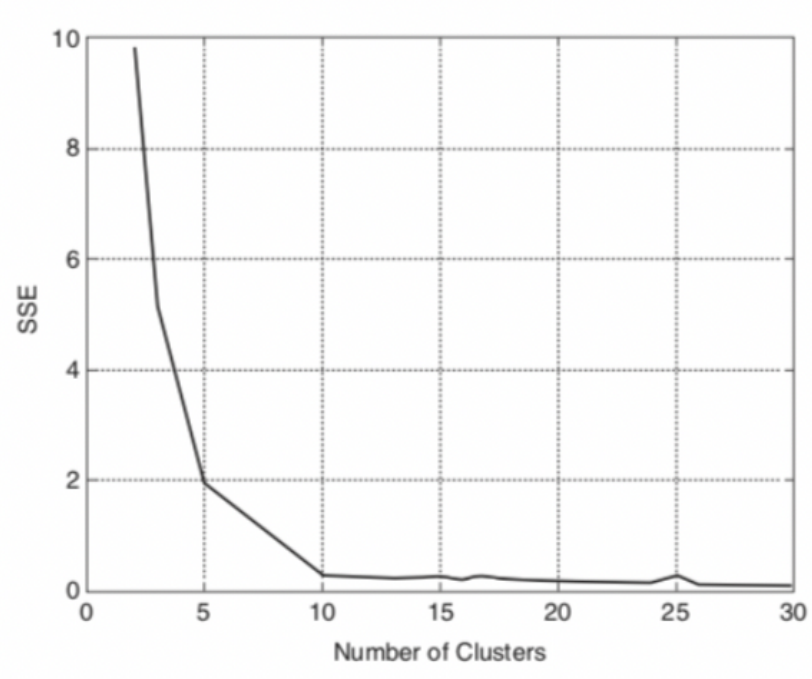


Figura 4.2: Andamento dell'SSE al variare di K [33]

La complessità computazionale dell'algoritmo è O($I \times K \times m \times n$) dove I è il numero di iterazioni, K è il numero di cluster, m è il numero di data-point e n è il numero degli attributi. Quindi, poiché il K-Means è lineare in m, il tempo di esecuzione richiesto dall'algoritmo è contenuto.

Talvolta, il K-Means può generare dei cluster vuoti se, durante la fase di assegnazione, nessun punto viene assegnato ad un centroide. Tuttavia, questa soluzione può determinare un valore di SSE molto elevato. Per ovviare a questa problematica si può individuare un centroide alternativo, utilizzando come strategia quella di scegliere il punto che contribuisce in larga misura alla definizione dell'SSE oppure selezionare come centroide un punto a partire dal cluster caratterizzato dal valore maggiore di SSE. Inoltre, la bontà del risultato del clustering è influenzata dalla presenza di outliers, che incide sul valore del centroide "spostandolo" e, di conseguenza, abbassando l'SSE. L'individuazione e l'eventuale rimozione degli outliers compete alla fase di preprocessing e la loro gestione dipende dal dominio di applicazione.

In generale, il K-Means soffre di limitazioni che non consentono all'algoritmo di raggiungere delle buone prestazioni quando i cluster naturali che si generano hanno numerosità o densità differente, quando la forma finale del cluster non è di tipo globulare e, infine, quando i dati contengono outliers. Per migliorare i risultati della sessione di clustering è indicato normalizzare i dati e individuare ed eliminare gli outliers in fase di preprocessing. Inoltre, in fase

di postprocessing sono suggeriti i seguenti accorgimenti: eliminare i piccoli cluster in quanto potrebbero rappresentare outliers; dividere i partizionamenti che presentano un valore elevato di SSE; unire cluster i cui centroidi sono vicini e che presentano valori di SSE bassi.

Il K-Means ha numerose varianti e, tra queste, rientra il Bisecting K-Means, una tecnica che consiste nell'applicare iterativamente il K-Means con $K = 2$ fin quando non si ottiene il numero di cluster desiderati. Il primo passo prevede di dividere l'intera collezione di dati in due partizioni; successivamente, si seleziona uno dei cluster generati e lo si divide nuovamente in due; si continua con questo procedimento fin quando non sono stati prodotti K cluster. Esistono diversi modi per scegliere quale cluster suddividere, per esempio si può decidere di selezionare il cluster più grande oppure quello che presenta il valore di SSE maggiore. Ovviamente, scelte diverse comportano cluster diversi. Generalmente, il Bisecting K-Means permette di ottenere delle migliori partizioni in quanto parte delle operazioni del postprocessing del K-Means sono incluse in questo algoritmo. Infine, questa tecnica è meno sensibile ai problemi indotti dalla scelta dei centroidi iniziali caratterizzante del K-Means [33].

4.2 Agglomerative Hierarchical Clustering

Le tecniche di clustering gerarchico sono la seconda categoria importante delle metodologie di clustering e, come il K-Means, questi approcci sono ancora molto diffusi.

In prima battuta, occorre distinguere due approcci basilari per generare un clustering gerarchico:

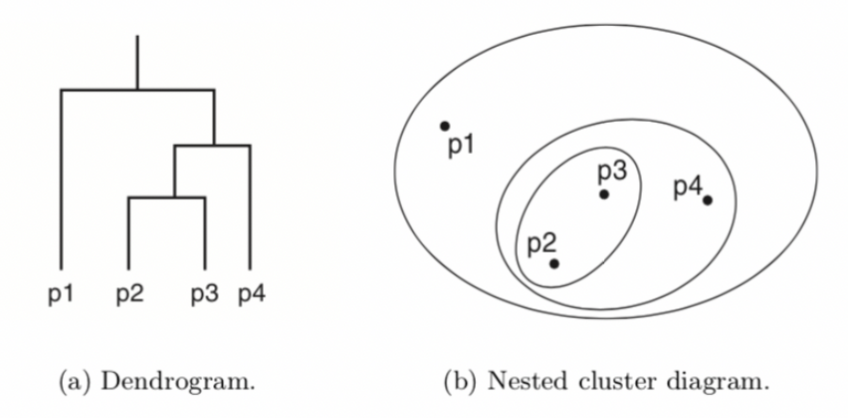
- **Agglomerativo**: si tratta di un metodo di tipo "bottom up" in cui ogni punto della collezione dati rappresenta un cluster. In seguito, i cluster più vicini vengono aggregati fino al caso limite in cui l'intera collezione dati rappresenta un unico cluster. Poiché questa tecnica è la più diffusa, ci si focalizzerà su questa tipologia.

- **Divisivo**: si tratta di un metodo di tipo "top down" in cui l'insieme di dati di partenza rappresenta un unico grande cluster. In seguito, il cluster viene diviso fino a quando si sono formati dei cluster in cui ognuno contiene un solo punto. In questo caso, è necessario stabilire a priori quali cluster sono da splittare e con quale criterio.

Questi approcci richiedono la definizione di una misura di distanza tra cluster e, normalmente, si utilizza una matrice di similarità o delle distanze o *proximity matrix*.

Tipicamente, le tecniche di clustering gerarchico producono un insieme di cluster annidati che spesso vengono rappresentati con il **dendrogramma**, un diagramma simile ad un albero che mostra la sequenza di fusioni tra i cluster. Il dendrogramma mostra le relazioni tra cluster sia dal punto di vista agglomerativo, se si legge il grafico dal basso

verso l'alto, sia dal punto di vista divisivo, nel caso in cui il diagramma venga letto dall'alto verso il basso. Inoltre, i cluster possono essere rappresentati con il nested diagram. Per maggiore chiarezza, la Figura 4.3 mostra un esempio di cluster gerarchico rappresentato nelle due versioni di diagrammi.



(a) Dendrogram. (b) Nested cluster diagram.

Figura 4.3: Clustering gerarchico di quattro punti visti come dendrogramma e come nested cluster [33]

Come già anticipato, la tipologia agglomerativa è la tecnica di clustering gerarchico più comune. L'algoritmo di base è molto semplice: inizialmente si crea un cluster per ogni punto, dopodiché si calcola la *proximity matrix*; in base alle distanze individuate, si fondono i due cluster più vicini e si aggiorna la matrice. Si continua iterativamente a fondere i cluster più vicini fin quando si ottiene un unico cluster contenente tutti gli elementi della collezione dati. L'Agglomerative Hierarchical Clustering è formalmente descritto dai seguenti step di esecuzione:

Algorithm 2 Agglomerative Hierachical Clustering

1: Creare un cluster per ogni elemento della collezione dati
2: Calcolare la *proximity matrix*
3: Fondere i due cluster più vicini
4: Aggiornare la *proximity matrix*
5: Ripetere i passi 3 e 4 fin quando rimane un unico cluster

Com'è facilmente intuibile, un aspetto particolarmente rilevante è il calcolo della distanza o similarità tra due cluster e i diversi modi di calcolarla distinguono le tecniche di clustering gerarchico. Tra i criteri più diffusi:

- **MIN o single link**: costruisce la *proximity matrix* individuando la minima distanza tra due elementi del cluster. Questo criterio permette di gestire anche cluster di forma non sferoidale, ma è sensibile alla presenza di outliers e dati rumorosi.
- **MAX o complete link**: costruisce la *proximity matrix* individuando la massima distanza tra due elementi dei cluster. Questo criterio è meno sensibile alla presenza

di outliers e dati rumorosi rispetto al criterio basato sulla minima distanza; tuttavia, tende a separare cluster di grandi dimensioni ed è maggiormente performante nel caso di cluster globulari.

- **Groupe Average**: costruisce la *proximity matrix* calcolando la media delle similarità tra tutte le coppie di punti dei cluster. Questo criterio rappresenta un compromesso tra MIN e MAX e risulta meno sensibile alla presenza del rumore; tuttavia, è indicato nel caso di cluster sferici.

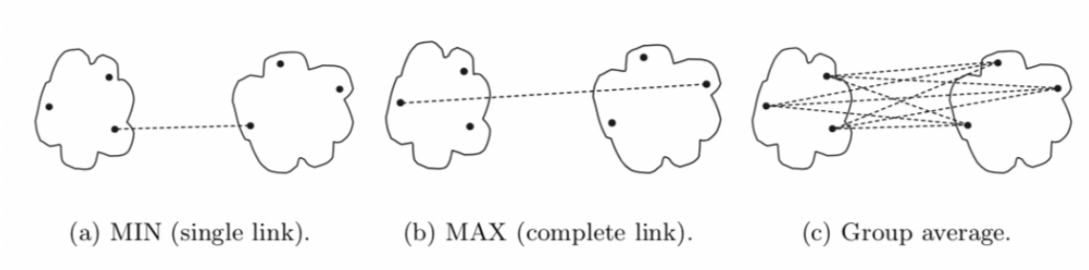


Figura 4.4: Definizioni di prossimità dei cluster [33]

Se, invece, si intende rappresentare i cluster mediante un centroide esistono si ricorre ad altre tecniche di calcolo della prossimità tra punti. In questi casi, infatti, nella *proximity matrix* la distanza si calcola tra i centroidi. Un metodo alternativo che assume che i cluster siano rappresentati dai centroidi è il *Ward Method* in cui si misura la vicinanza tra i due cluster in termini di incremento dell'SSE che risulterebbe a seguito della fusione i due cluster. Come nel K-Means, il *Ward Method* ha come funzione obiettivo la minimizzazione della somma degli scarti quadratici dei punti dal centroide. Questo metodo è meno suscettibile rispetto alla presenza del rumore, ma è indicato preferibilmente con cluster sferici.

A differenza del K-Means, queste tecniche non richiedono a priori la definizione di K: il numero di cluster desiderato può essere ottenuto eseguendo "un taglio" all'altezza opportuna nel dendrogramma. Inoltre, è possibile identificare una tassonomia, cioè classificazione gerarchica dei punti, in quanto elementi più simili saranno fusi più velocemente di quelli più distanti. Tuttavia, in queste tecniche manca una funzione di ottimizzazione generale e l'aggregazione o divisione dei punti non è un'operazione reversibile.

Lo spazio occupato dalla *proximity matrix* è $\frac{1}{2}m^2$, dove m è il numero di data-points, dunque lo spazio occupato è $O(N^2)$ quando il numero di punti è N. Il tempo di esecuzione dell'algoritmo, invece, è $O(N^3)$ in quanto sono necessari N passi per la costruzione del dendrogramma; inoltre, ad ogni passo, la matrice di prossimità deve essere aggiornata. Tuttavia, per alcuni approcci la complessità può essere ridotta a $O(N^2 \log(N))$. In alcune applicazioni, questi livelli di complessità computazionale sono esagerate.

Un altro aspetto critico da sottolineare riguarda la gestione dei cluster di dimensione diversa. Allo scopo di trattare equamente i cluster di numerosità differente sono stati proposti due approcci: l'approccio ponderato, in cui ai punti appartenenti a cluster differenti è assegnato un peso proporzionale alla loro dimensione, e l'approccio non ponderato, in cui ai punti appartenenti a cluster differenti è assegnato lo stesso peso.

Infine, gli outliers costituiscono un problema per negli approcci di clustering gerarchico, più accentuato nei casi in cui i gruppi sono rappresentati dal centroide. Ciò accade perché la presenza di outliers e dati rumorosi determina un incremento dell'SSE e, quindi, uno spostamento dei centroidi. Inoltre, la presenza di outliers favorisce la formazione di cluster singleton o di piccoli gruppi di dati rumorosi che non si aggregano a nessun altro cluster, se non nelle fasi finali del processo [33].

4.3 DBSCAN

Il DBSCAN è un algoritmo di clustering *density-based*, cioè in grado di individuare regioni dello spazio ad alta densità, separate dalle zone a bassa densità. Nonostante non esistano molte definizioni di densità, in questo contesto si farà riferimento al concetto di densità *center-based* caratterizzante dell'algoritmo del DBSCAN. La densità di ogni punto è definita come il numero di elementi presenti all'interno di una circonferenza di raggio *Epsilon*, includendo il punto stesso.

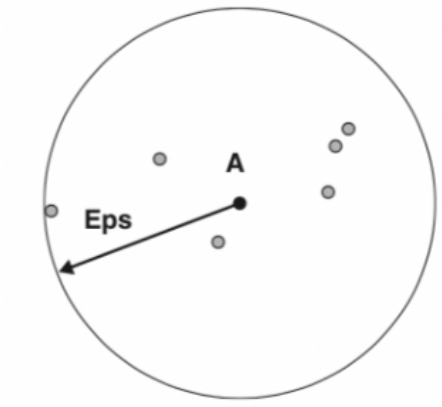


Figura 4.5: Densità center-based [33]

L'approccio *center-based* consente di classificare i punti in tre categorie:

- **Core Point**: un punto è definito *core* se, all'interno di una circonferenza di raggio *Eps*, sono presenti almeno un numero di punti pari a *MinPoints*, dove *Eps* e *MinPts*

sono parametri selezionati arbitrariamente dall'utente. Nell'esempio in Figura 4.6, A è un esempio di *core point* in una configurazione con *Minpoints* pari a 7.

- **Border Point**: un punto è definito *border* se è localizzato nelle vicinanze di un punto *core*, ma non lo è. Un esempio di *border point* in Figura 4.6 è il punto B.
- **Noise Point**: un punto è definito *noise* se non si trova vicino né ad un *core* né ad un *border*. Un esempio di *noise point* in Figura 4.6 è il punto C.

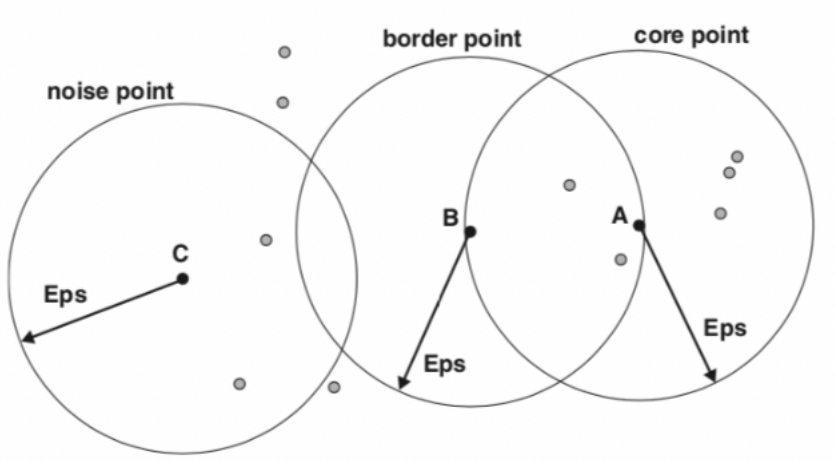


Figura 4.6: Core, Border e Noise Points [33]

Una volta chiariti questi concetti, è possibile spiegare il funzionamento dell'algoritmo e la modalità di assegnazione dei punti ai cluster. In particolare, due *core points* abbastanza vicini, cioè all'interno di un raggio *Eps* l'uno dall'altro, sono assegnati allo stesso cluster; secondo lo stesso principio, ciascun *border point* è associato al cluster a cui il *core point* vicino è assegnato; i *noise points* sono scartati. Il DBSCAN è formalmente descritto dai seguenti passi di esecuzione:

Algorithm 3 DBSCAN

1: Etichettare tutti i punti come *core*, *border* o *noise points*
2: Eliminare i *noise points*
3: Assegnare al cluster c_i i punti che abbiano almeno una distanza minore di *Eps* da almeno uno degli altri punti *core* assegnati al cluster
4: Assegnare i punti *border* al cluster a cui sono associati i corrispondenti punti *core*

Il DBSCAN è un algoritmo molto performante, robusto alla presenza di rumore e capace di generare cluster di forma e dimensione differente. In aggiunta, l'algoritmo è in grado di identificare cluster che il K-Means non sarebbe in grado di rilevare. Tuttavia, il DBSCAN è limitato quando i cluster hanno una densità variabile. Inoltre, è problematico quando i dati hanno elevata dimensionalità in quanto risulta complessa la definizione

della densità. Infine, il DBSCAN è dispendioso in termini di complessità computazionale, in quanto l'identificazione dei punti più vicini richiede il calcolo delle distanze a coppie tra gli elementi. Infatti, nei casi peggiori quando i dati hanno elevata dimensionalità, la complessità dell'algoritmo è $O(m^2)$, dove m è il numero di punti; quando, invece, il dataset ha poche caratteristiche la complessità computazionale diminuisce a $O(m \log(m))$.

Come il K-Means, anche il DBSCAN necessita della definizione a priori da parte dell'utente dei parametri *Minpoints* e *Eps*. L'approccio tradizionale consiste nell'osservare il grafico chiamato $k - dist$, in cui si visualizzano i punti ordinati in ordine crescente in base alla distanza dal loro $k - esimo$ punto, dove k è specificato dall'utente. Il principio su cui si basa il $k - dist$ consiste nel fatto che per i *core points* i $k - esimi$ punti più vicini saranno indicativamente alla stessa distanza, mentre i *noise points* avranno il $k - esimo$ punto più lontano. Di conseguenza, quando nella curva si verifica un repentino cambio della distanza questo segnala la separazione tra *core* e *noise points*. Quindi, il punto in corrispondenza del quale si verifica il cambio della pendenza coincide con un valore adatto di *Eps*. Di conseguenza, selezionando l'*Eps* così identificato e assumendo *Minpoints* pari a k, i punti che nel $k - dist$ hanno un'ordinata minore di *Eps* sono etichettati come *core*, mentre gli altri sono etichettati come *border* o *noise points*.

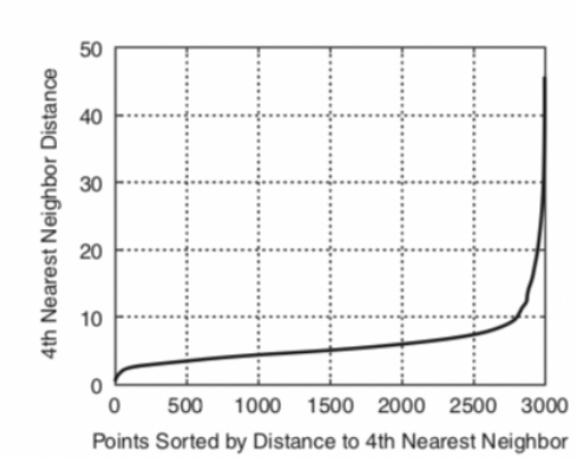


Figura 4.7: K-dist plot [33]

Il valore di *Eps*, dunque, si può evincere dal grafico del $k - dist$, ma in questo modo dipende dal parametro k. Fortunatamente, il valore di *Eps* non cambia molto al variare di k in quanto la curva rimane simile per valori sensati di k. Tuttavia, se k è troppo basso piccoli gruppi di dati rumorosi potrebbero essere etichettati erroneamente come cluster; viceversa, se k è troppo elevato è possibile che cluster di piccola dimensione vengano etichettati come rumore [33].

4.4 Valutazione della bontà del clustering

Per le tecniche di classificazione supervisionata esistono diverse misure per valutare la bontà dei risultati basate sul confronto tra le etichette reali del *test set* e quelle calcolate dall'algoritmo, come l'accuratezza, la precisione e il richiamo. Per il clustering non esistono delle metriche universali utilizzate per valutare la bontà dei partizionamenti, ma dipendono dall'algoritmo utilizzato. Per esempio, per il K-Means si può sfruttare l'SSE, ma questo approccio non è adeguato al DBSCAN in quanto i cluster risultanti tipicamente non sono di forma globulare. Ci si potrebbe chiedere come mai sia necessario valutare i risultati del clustering, dato che quest'analisi è condotta tipicamente a scopo esplorativo del dataset. Tuttavia, la fase di *clustering validation* risulta necessaria in quanto, generalmente, a seguito di algoritmi di clustering diversi risultano partizionamenti differenti. La fase di *clustering validation* risulta problematica sotto diversi aspetti. In primo luogo, determinare la *clustering tendency* dei dati, cioè l'abilità di distinguere se esiste una struttura non casuale nei dati. Inoltre, è difficile determinare il numero corretto di cluster, valutare se i risultati del clustering si adattano bene ai dati di partenza senza ricorrere a informazioni esterne, e, infine, stabilire quale tra due sessioni di clustering è migliore.

Le misure di valutazione della bontà del clustering sono tipicamente classificate in tre categorie:

- **Non supervisionate**: misurano la bontà del clustering senza ricorrere a informazioni esterne. Per questo motivo, poiché queste metriche sfruttano solamente le informazioni contenute all'interno del dataset sono anche dette misure interne; un esempio è l'SSE. Inoltre, le misure interne non supervisionate si dividono ulteriormente in misure di coesione del cluster, che determinano quanto gli elementi del cluster sono compatti, e misure di separazione dei cluster, che stabiliscono quanto i cluster sono tra loro separati.

- **Supervisionate**: misurano la bontà del clustering ricorrendo a informazioni al di fuori del dataset e, per questo motivo, sono anche dette misure esterne. Queste metriche valutano quanto bene le etichette di cluster corrispondono alle etichette reali; un esempio è l'entropia oppure l'Adjusted Rand Index.

- **Relative**: possono essere sia supervisionate che non e sono utilizzate per comparare due diversi clustering o due cluster.

Soffermandosi sulle misure interne, la coesione di un cluster può essere definita come la somma dei pesi dei collegamenti tra i punti nel grafico delle prossimità; la separazione tra i cluster può essere misurata dalla somma dei pesi dei collegamenti tra i punti in un cluster e l'altro. Nella Figura 4.8 (a) è mostrata la coesione, nella 4.8 (b) la separazione.

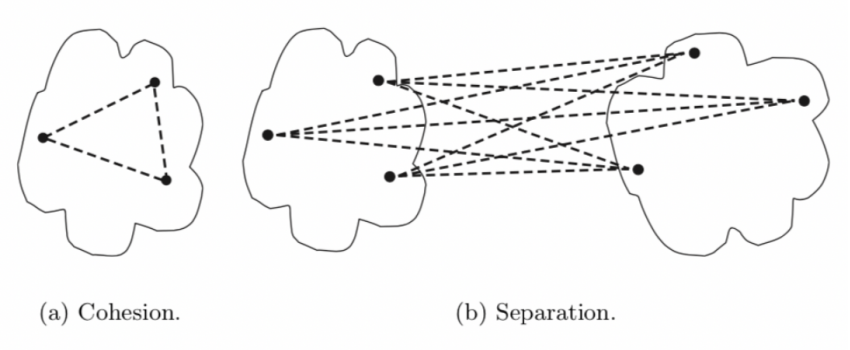


Figura 4.8: Coesione e separazione tra i cluster con rappresentazione basata sui grafi [33]

Le misure di coesione e separazione possono essere utilizzate per la valutazione dei singoli cluster. Per esempio, un cluster con un valore maggiore di coesione può essere considerato migliore di uno che ne presenta un basso valore. Oppure, in base alla coesione, si può decidere se suddividere o aggregare i cluster: in particolare se la coesione di un cluster è bassa, si può decidere di scinderlo in più sotto-cluster; viceversa, se due cluster sono molto compatti ma non molto separati, si può optare di unirli in unico cluster.

Una metrica molto popolare per valutare la bontà dei cluster che sfrutta i concetti di coesione e separazione è la **Silhouette**. Per ogni punto $i \in C_i$, si calcola la media della distanza tra tutti i punti del proprio cluster come:

$$a_i = avg_{j \in C}(dist(i, j))$$

Successivamente, per ogni punto $i \in C_i$ si calcola la distanza media tra i e tutti gli altri punti $j \in C_j$ con $C_i \neq C_j$, come:

$$b_i = \min_{C_j \neq C_i} (avg(dist(i, j))).$$

Infine, per ogni elemento i, la Silhouette si calcola come:

$$s_i = \frac{b_i - a_i}{max(a_i,\ b_i)}$$

Il valore della Silhouette può variare tra -1 e 1. È auspicabile un valore di Silhouette positivo, in quanto indica che il punto è correttamente abbinato al proprio cluster e scarsamente abbinato agli altri cluster; invece, un valore di Silhouette negativo indica il contrario e, quindi, si può dedurre che l'assegnazione dei punti nei cluster non è stata performante.

Poichè maggiore l'indice di Silhouette, migliori sono le performance della sessione di clustering, si può sfruttare il calcolo della Silhouette per diversi valori di K per individuare il numero di partizioni ottimali. Ovviamente, questo metodo suggerisce di selezionare K

in corrispondenza del quale si riscontra il valore maggiore di Silhouette [33].

Un'altra metrica utilizzata per valutare la bontà dei risultati del clustering è il **Rand Index**, appartenente alla categoria delle misure esterne in quanto si sfruttano informazioni al di fuori del dataset. L'indice misura la similarità tra due tipi di etichette, quelle assegnate dal clustering e quelle reali di appartenenza dei punti.
Il valore del Rand Index varia tra -1 e 1; i valori prossimi ad 1 indicano una buona etichettatura dei dati da parte del clustering, coerentemente con le etichette reali; i valori negativi o prossimi allo zero, al contrario, sono indice di performance scarse. Il Rand Index RI è aggiustato e modificato con la relazione:

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]}$$

dove $E[RI]$ è il valore atteso del Rand Index. In questo modo, i valori del Rand Index aggiustato negativi o prossimi allo zero indicano un'etichettatura casuale indipendentemente dal numero di cluster e di campioni. Inoltre, per l'uso di questo indice non è necessario fare alcuna assunzione sulla struttura dei cluster: infatti, il Rand Index può essere impiegato sia nel caso di cluster globulari, come quelli risultanti dal K-Means, sia nell'eventualità di cluster di forma diversa. Il principale limite del Rand Index è legato alla necessità di conoscere a priori l'etichetta reale di appartenenza dei punti, che nella pratica non è quasi mai disponibile. [41]

4.5 Strumenti utilizzati

Gli strumenti utilizzati nell'ambito di questo progetto di tesi sono sostanzialmente legati all'analisi e alla manipolazione di Big Data. Tutte le attività di *data analytics* sono state effettuate con Python, un linguaggio di programmazione di alto livello, che possiede numerose librerie ottimizzate per la gestione e l'analisi dei dati. L'ambiente di sviluppo utilizzato a questo scopo è Jupyter Notebook, un'applicazione client open-source basata sul web che si attiva con un browser standard. Jupyter Notebook permette la creazione di documenti web costituiti da una lista ordinata di celle di input e output, all'interno delle quali è possibile scrivere codici, visualizzare i dati, eseguire i calcoli ed esaminare immediatamente i risultati corrispondenti. Il principale vantaggio di questo strumento riguarda la possibilità di scrivere script ed eseguirli in tempo reale in celle indipendenti: in questo modo si possono testare blocchi di codice separatamente. Per utilizzare Jupyter Notebook è stata sfruttata l'applicazione GUI Anaconda, che include oltre Python altri pacchetti software utili per la data science, machine learning, analisi statistiche, predictive analytics e molto altro.

Di seguito, una breve panoramica sulle principali librerie di Python utilizzate nell'ambito di questa tesi:

- **Pandas**: è una libreria che fornisce strutture di dati flessibili progettate allo scopo di rendere l'interazione con dati etichettati o relazionali più facile e intuitiva. La maggior parte delle applicazioni sono gestite da due strutture portanti di Pandas: la Series, ovvero un vettore monodimensionale indicizzabile, e il DataFrame, cioè un array a due dimensioni la cui forma ricorda una tabella, che garantisce la possibilità indicizzare i dati e manipolarli in maniera rapida ed efficace. Pandas è stata utilizzata in fase di caricamento e salvataggio di file in diversi formati, principalmente JSON, CSV ed Excel. Inoltre, la libreria è stata sfruttata per effettuare analisi e calcoli numerici, oltre che per la visualizzazione immediata ed intuitiva di dati e risultati. [42]

- **Matplotlib** e **Seaborn**: nell'ambito della *data science* risulta di fondamentale importanza comunicare i risultati delle analisi in maniera efficace e immediata. A questo scopo, sono state utilizzate le librerie Matplotlib e Seaborn per creare grafici sia statici che dinamici. Soprattutto nella fase di *data exploration*, gli strumenti messi a disposizione hanno permesso di analizzare e plottare *time series* e mettere in evidenza le principali caratteristiche delle variabili in gioco. [43]

- **Scikit-learn**: è una libreria di Python utilizzata per l'analisi dei dati e il *machine learning*. Contiene algoritmi di *data mining* come classificazione, regressione e clustering, ma dispone anche di metodi utili alle operazioni di preprocessing e *data reduction*. Inoltre, sono a disposizione algoritmi di *model selection*, utili per scegliere i parametri per migliorare le prestazioni degli algoritmi, come la GridSearch, e valutare i loro risultati tramite le metriche più opportune, come la Silhouette e l'Adjusted Rand Score. [44]

Capitolo 5

Risultati sperimentali

5.1 Caso di studio

L'obiettivo di questa tesi è presentare una metodologia semi-supervisionata allo scopo di caratterizzare i cicli di produzione in base al livello di tensione della cinghia di trasmissione.

Le cinghie sono degli strumenti che consentono di trasmettere potenza, anche di consistente entità, tra due alberi meccanici in modo uniforme e a basso rumore, oltre che di assorbire eventuali urti e variazioni di carico, che potrebbero arrecare danni al motore. Il sistema di cinghie è in grado di correggere l'eventuale disallineamento assiale tra i due alberi con un'elevata tolleranza, richiede attività di manutenzione minime e non necessita di essere lubrificato. Inoltre, il costo delle cinghie aumenta meno che proporzionalmente al crescere della distanza tra i due alberi, a differenza di quanto accade per gli ingranaggi. Tuttavia, questi organi garantiscono minore resistenza e rigidità rispetto ad altri sistemi di trasmissione della potenza come ingranaggi e catene, ma i recenti sviluppi in questo ambito hanno consentito di impiegare le cinghie in applicazioni in cui l'uso degli ingranaggi era quasi esclusivo. [45]

Esistono diverse tipologie di cinghie da trasmissione sul mercato:

- Cinghia piatta: si tratta della tipologia più diffusa in passato, impiegata soprattutto in pompe e generatori. Tipicamente le cinghie piatte sono sottili, di forma rettangolare e costituite di materiali tessili o sintetici, come cuoio oppure nylon, dunque il loro costo è contenuto. Inoltre, sono facili da montare, trasmettono potenza ad elevate velocità ed hanno un comportamento elastico, caratteristica che permette di resistere ad urti di elevata entità. Tuttavia, sono particolarmente indicate per la trasmissione della potenza su alberi posti molto lontani l'uno dall'altro, mentre per distanze più contenute risultano più adeguate le cinghie trapezoidali.

- Cinghia trapezoidale: è la tipologia più diffusa attualmente sul mercato in quanto risulta meno ingombrante della cinghia piatta e garantisce il migliore compromesso

tra trazione, velocità, tensione sui supporti e durata. La forma trapezoidale permette una migliore adesione e un ottimo livello di attrito e, quindi, la trasmissione di grande potenza. Tipicamente, queste cinghie sono costituite da un'anima di nylon ricoperta da strati di gomma.

- Cinghia dentata: sono impiegate in applicazioni in cui è necessario garantire un'elevata precisione di lavorazione. I denti che costituiscono la cinghia, tipicamente di gomma, sono ricoperti da uno strato di nylon che offre elevata resistenza e permette di mantenere allineati gli organi meccanici collegati, con assenza di slittamenti.

Risulta fondamentale analizzare e monitorare livello di tensione della cinghia per garantire un buon funzionamento del braccio robotico. Infatti, se la cinghia fosse sovratensionata si verificherebbero surriscaldamenti con il rischio di danneggiare sia la cinghia che i cuscinetti; invece, se la tensione fosse troppo bassa potrebbero avvenire slittamenti che potrebbero determinare l'usura prematura della stessa. [46].

Il braccio robotico utilizzato per ricavare i dati utili per effettuare le analisi esposte presenta problematiche inerenti questa tematica. I dati oggetto di studio sono stati forniti da un'azienda italiana leader nel settore metalmeccanico, in particolare per quanto riguarda la realizzazione di sistemi di produzione e processi di automazione. Per eseguire le analisi sono stati esaminati due dataset, *Gray* e *White*, entrambi in formato JSON Mimosa. In ciascuna collezione dati sono contenuti i dati relativi alla corrente assorbita dal motore in ogni ciclo di produzione; ognuno di essi è associato univocamente ad un timestamp ed ha una durata di circa 24 secondi approssimativamente, periodo durante il quale rilevano 11967 misurazioni della corrente. Inoltre, l'azienda in questione ha fornito delle statistiche calcolate sui dati della corrente, tra cui media, minimo, massimo, deviazione standard, curtosi e asimmetria, nonchè informazioni sulla posizione del motore in ogni istante. Tuttavia, come già anticipato, le analisi condotte tengono conto unicamente dei dati della corrente, tralasciando, quindi, quelli sulla posizione. Le rilevazioni sono state effettuate per entrambi i dataset nei giorni compresi tra il 24/02/2020 e il 04/03/2020.

La prima parte di questa tesi è incentrata sull'analisi esplorativa di entrambi i dataset allo scopo di osservare i trend delle principali variabili in gioco ed eventuali anomalie. In secondo luogo si eseguiranno le attività di preprocessing e transformation dei dati grezzi raccolti dai sensori con l'obiettivo di estrarre le principali caratteristiche che descrivono ciascun ciclo di produzione. Infine, si procederà alla fase di Knowledge Extraction in cui verranno applicati le tecniche di clustering descritte nel capitolo precedente.

5.2 Analisi esplorativa

I due dataset forniti, *Gray* e *White*, contengono dati riguardo alla corrente assorbita dal motore di un braccio robotico in ogni ciclo di produzione e, per ognuno di essi, è nota la classe di appartenenza. La conoscenza delle etichette reali è un aspetto fondamentale in un approccio semi-supervisionato in quanto è necessaria in fase di validazione del modello per valutare se le partizioni generate dagli algoritmi di clustering sono coerenti con le classi di appartenenza. Sono state fatte le seguenti assunzioni circa le etichette fornite:

- Etichetta 0: tensionamento della cinghia normale;
- Etichetta 10: sovratensionamento della cinghia;
- Etichetta 15: sottotensionamento della cinghia.

Nelle Tabelle 5.1 e 5.3 rispettivamente per *Gray* e per *White* si può osservare più dettagliatamente la composizione di ciascuna collezione dati in termini di numero di cicli appartenenti a ogni etichetta. In particolare, si nota che la classe 10 è la più numerosa in entrambi i dataset e rappresenta circa il 57% delle misurazioni rilevate.

Nelle Tabelle 5.2 e 5.4 si osserva che sia in *Gray* che in *White* le misurazioni non sono state effettuate consecutivamente: infatti, sono assenti le rilevazioni della corrente a partire dal 28/02/2020 fino al 02/02/2020. Inoltre, queste date rappresentano il confine tra i cicli di produzione con etichetta 10 e quelli con etichetta 15. In questa zona si potrà osservare un picco nei valori delle features fornite.

Etichetta	Numero Cicli	% Dataset
0	1287	21.4%
10	3419	56,8%
15	1313	21,8%

Tabella 5.1: Distribuzione etichette nel dataset *Gray*

Data	Ora primo ciclo	Ora ultimo Ciclo	Numero Cicli	Etichette
24/02/2020	13.51.25	23.58.28	746	0
25/02/2020	08.23.04	23.57.58	541 1157	0 10
26/02/2020	00.58.36	23.57.49	1702	10
27/02/2020	00.59.15	08.27.33	544	10
03/03/2020	16.54.06	23.57.41	528	15
04/03/2020	00.59.07	11.32.29	786	15

Tabella 5.2: Informazioni sui giorni dei cicli macchina *Gray*

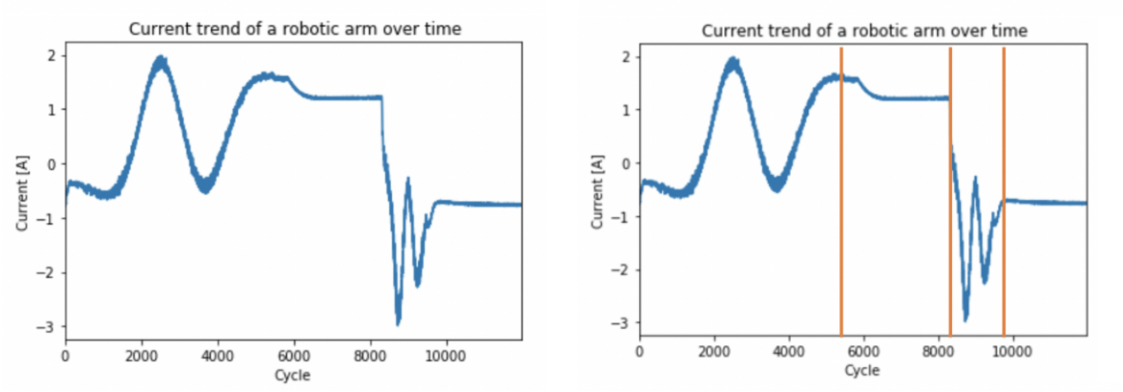
Etichetta	Numero Cicli	% Dataset
0	1216	21.2%
10	3275	57,2%
15	1238	21,6%

Tabella 5.3: Distribuzione etichette nel dataset *White*

Data	Ora primo ciclo	Ora ultimo ciclo	Numero cicli	Etichette
24/02/2020	14.49.13	23.58.50	678	0
25/02/2020	01.56.06	23.58.28	539 1088	0 10
26/02/2020	01.56.32	23.58.19	1626	10
27/02/2020	01.57.11	09.24.41	546	10
03/03/2020	17.51.45	23.58.42	454	15
04/03/2020	01.56.45	12.30.08	784	15

Tabella 5.4: Informazioni sui giorni dei cicli macchina *White*

In entrambi i dataset, l'andamento della corrente in ciascun ciclo di produzione è rappresentato dal grafico in Figura 5.1. Come si evince dalla Figura 5.1(b), il segnale si può suddividere in quattro zone. La prima è caratterizzata da oscillazioni della corrente tra -0.5A e 2A, a cui segue una seconda fase di assestamento del segnale intorno al valore di 1A. Nella terza zona si verifica un drastico crollo della corrente che raggiunge il valore minimo pari a -3A per poi fluttuare nelle rilevazioni immediatamente successive. Infine, l'ultima zona è caratterizzata da un assestamento della corrente intorno al valore di 1A. Le quattro aree evidenziate corrispondono a movimenti ben distinti del motore del braccio robotico.



(a) *Andamento del segnale* **(b)** *Aree del segnale*

Figura 5.1: Segnale della corrente in un braccio robotico in ogni ciclo di produzione

A partire dalle misurazioni della corrente in ciascun ciclo sono state calcolate le misure statistiche di media, minimo, massimo, deviazione standard, curtosi e simmetria e sono state confrontate con i valori delle stesse grandezze fornite dai proprietari dei dati. Si farà dapprima riferimento al dataset *Gray* e in seguito a *White*.

5.2.1 Gray

In primo luogo è stato valutato l'andamento del valore medio della corrente in ciascun ciclo di produzione. Dal confronto tra le Figure 5.2 e 5.3 si evince che nelle due fonti il dominio è differente: infatti, nei valori della media fornita la corrente varia tra [0.21A, 0.27A], mentre nella media calcolata questa appartiene all'intervallo [0.18A, 0.31A]. Tale differenza nel range del valore medio è dovuta a due cicli di produzione iniziali nella giornata del 24/02/2020, alle ore 14.28.07 e 15.00.49, in cui le misurazioni della corrente risultano anomale. Per questo motivo, si è deciso di escludere dalle analisi successive i cicli in questione.

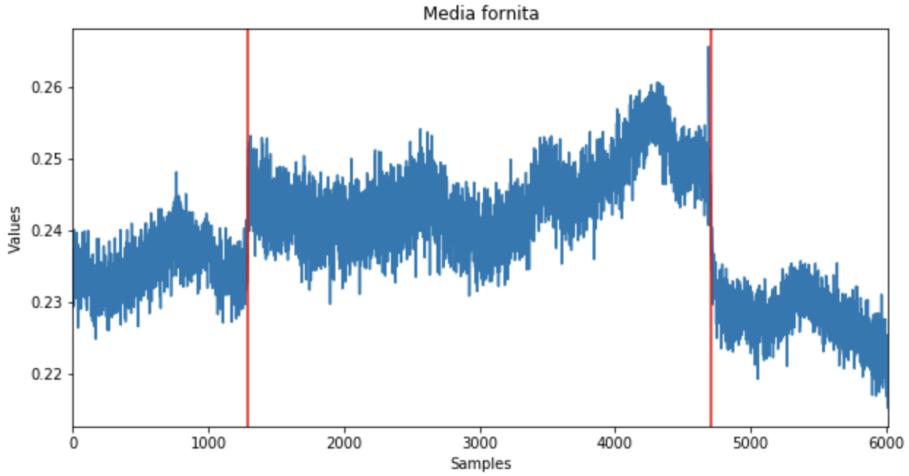


Figura 5.2: Media fornita dataset *Gray*

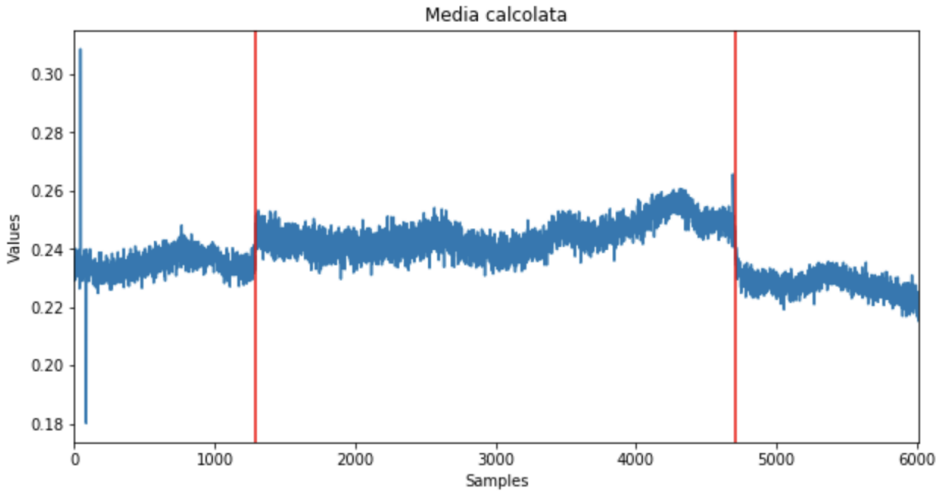


Figura 5.3: Media calcolata dataset *Gray*

Le linee verticali in rosso rappresentano i confini tra le classi che, come si può osservare, risultano ben separate tra di loro. Un tratto del grafico da evidenziare è il picco verso l'alto che si verifica nella zona di passaggio dalla classe 10 alla classe 15. Come sottolineato in precedenza, in questa zona si verifica un'interruzione delle rilevazioni della corrente nel periodo dal 28/02/2020 fino al 02/03/2020. Figura 5.4.

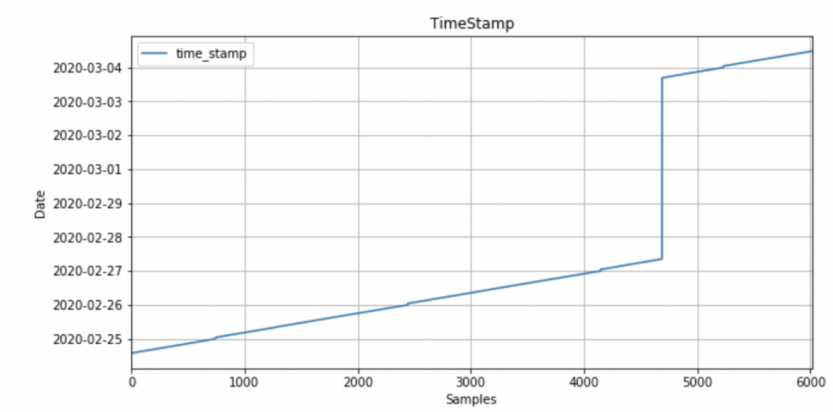


Figura 5.4: Date rilevazione corrente

In aggiunta, è stata condotta una regressione lineare allo scopo di evidenziare il trend complessivo della media della corrente. Come si può osservare dalla Figura 5.5, l'inclinazione della retta che approssima i dati è lievemente negativa. Inoltre, dalla Figura 5.5 si nota la presenza dei due outliers sopraccitati, che saranno esclusi dalle analisi successive.

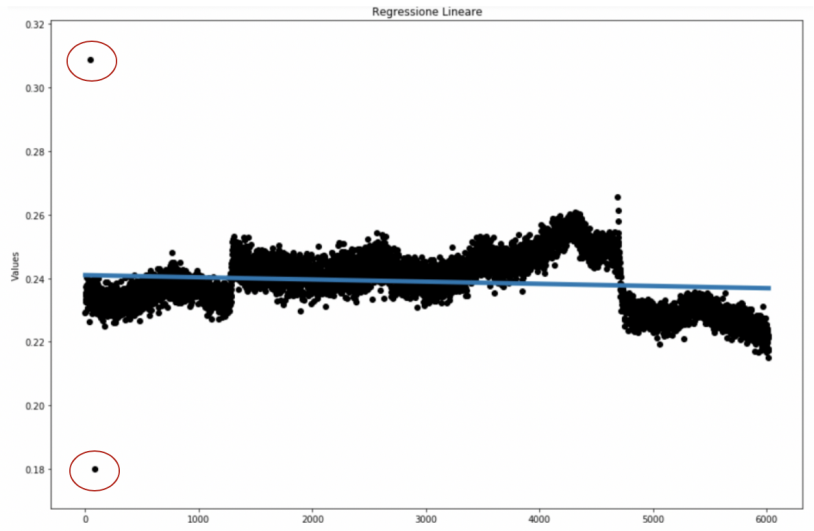


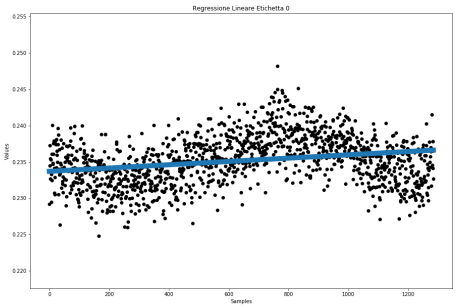
Figura 5.5: Regressione dataset *Gray*

La tecnica della regressione lineare è stata applicata separatamente per etichetta per distinguere l'andamento della media nel tempo nelle classi a disposizione. Come si può osservare dalla Figura 5.6 è emerso che:

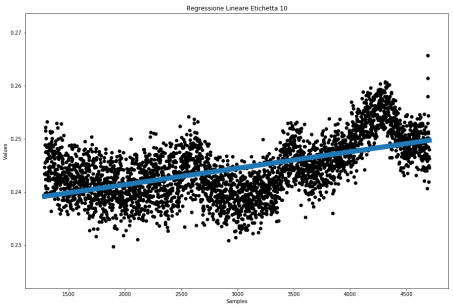
- nella classe 0, in cui la media varia tra 0.225 A e 0.245 A, si possono osservare delle regolari oscillazioni dei valori della corrente. La retta che approssima i dati ha una pendenza leggermente positiva.
- nella classe 10 la media varia tra 0.23 A e 0.26 A e, anche in questa etichetta, si

verificano delle fluttuazioni regolari. La retta che approssima i dati ha pendenza positiva, maggiore rispetto a quella riscontrata nell'etichetta 0.

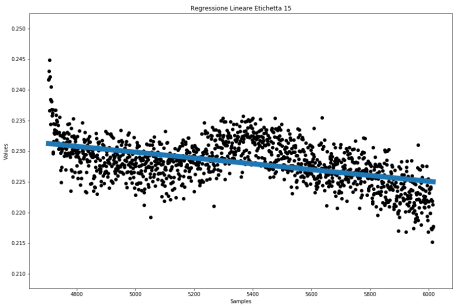
- nella classe 15, in cui la media varia tra 0.245 A e 0.215 A, si evidenzia la presenza di fluttuazioni meno regolari rispetto a quelle presenti negli andamenti delle medie delle etichette 0 e 10. La retta che approssima i dati ha una pendenza negativa.



(a) *Regressione classe 0*



(b) *Regressione classe 10*



(c) *Regressione classe 15*

Figura 5.6: Dataset *Gray*. Regressione lineare separatamente per etichetta

Oltre le analisi sull'andamento della media, sono stati eseguiti i plot delle altre statistiche messe a disposizione, ovvero deviazione standard, minimo, massimo, asimmetria e curtosi, le quali non aggiungono ulteriori informazioni oltre quelle emerse dal trend della media, cioè la presenza di valori anomali nei cicli iniziali e il picco che si verifica al confine tra le etichette 10 e 15.

5.2.2 White

Anche per questo dataset è stata rappresentata ed analizzata la distribuzione della media della corrente in ciascun ciclo di produzione. A differenza del dataset *Gray*, non si evidenziano discrepanze tra la media fornita e la media calcolata a partire dai dati e i valori ammissibili variano tra 0.27A e 0.31A. Anche per il dataset *White*, i confini delle

etichette 0, 10 e 15 sono evidenziati mediante delle linee verticali rosse e, come si può notare dalla Figura 5.7, le classi risultano ben separate tra loro. Come nel dataset *Gray*, si osserva un insieme di valori mediamente più elevati in prossimità dei cicli di produzione con etichetta 15, per poi diminuire bruscamente. Per maggiore chiarezza, in Figura 5.8 si riporta un ingrandimento della media nella zona interessata. Si ipotizza che questo comportamento sia dovuto all'interruzione delle misurazioni della corrente nei giorni tra il 28/02/2020 e il 02/03/2020. Figura 5.4.

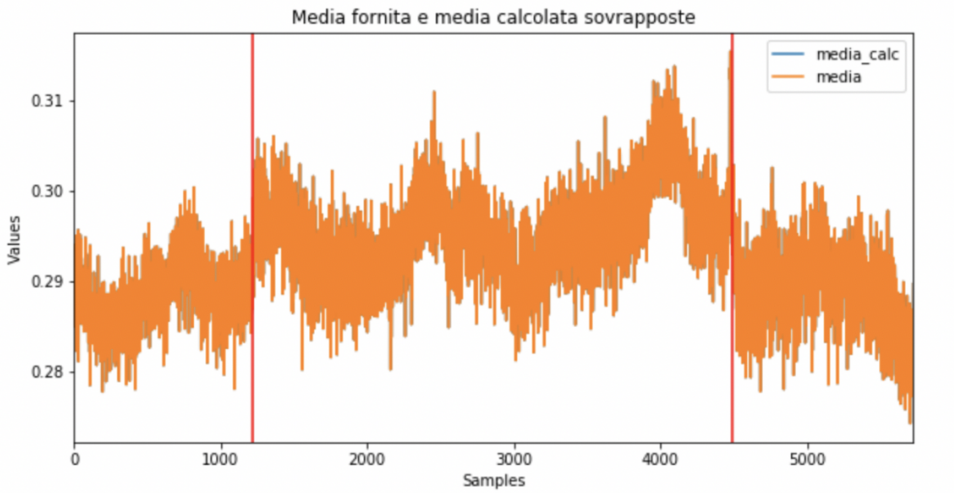


Figura 5.7: Media fornita e media calcolata sovrapposte dataset *White*

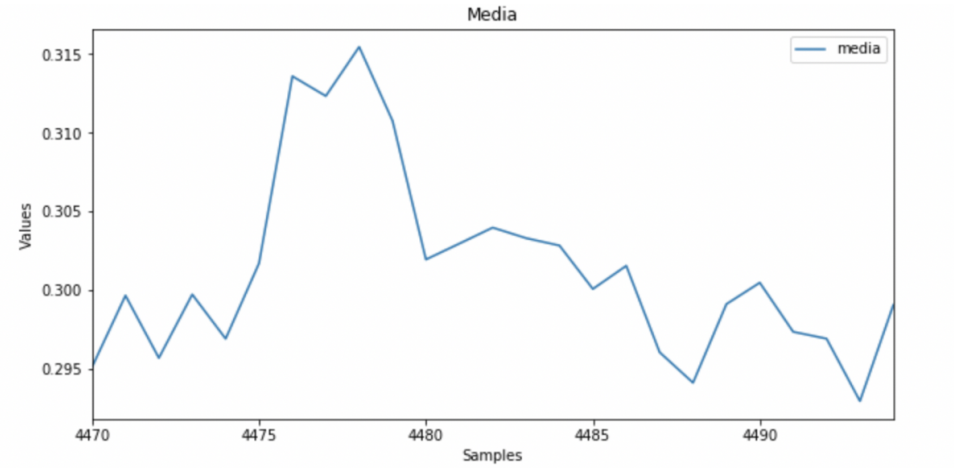


Figura 5.8: Media calcolata nei cicli di produzione al termine del 27/02/2020

Complessivamente, l'andamento della media della corrente nel dataset *White* può essere approssimato con una retta di regressione lineare che, come si può osservare dalla Figura 5.9 ha pendenza leggermente positiva. Tale comportamento è differente dall'andamento complessivo della media nel dataset *Gray*, la quale, tuttavia, era influenzata dalla presenza dei due cicli di produzione anomali.

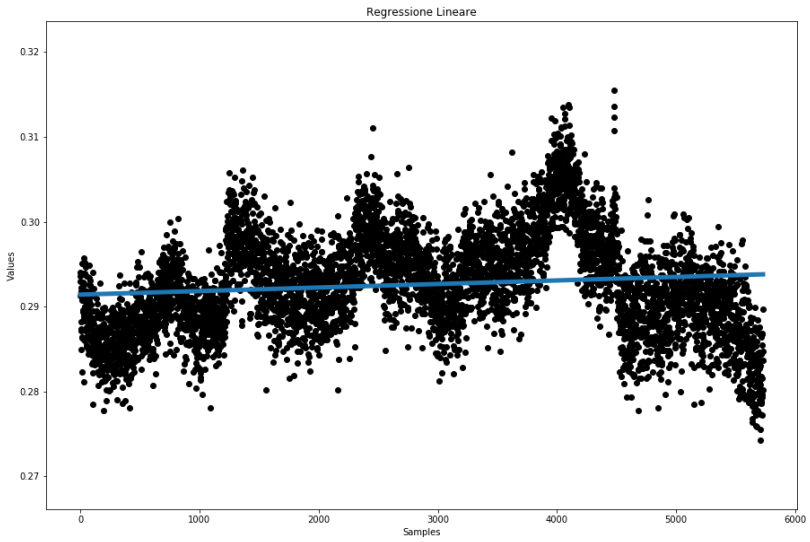


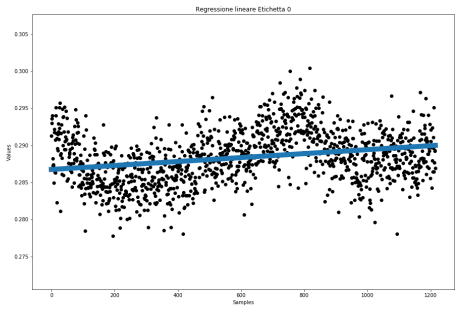
Figura 5.9: Regressione dataset *White*

In aggiunta, l'analisi di regressione lineare è stata condotta separatamente per etichetta per distinguere i comportamenti nel tempo delle classi a disposizione. Come si può osservare dalla Figura 5.10, è emerso che:

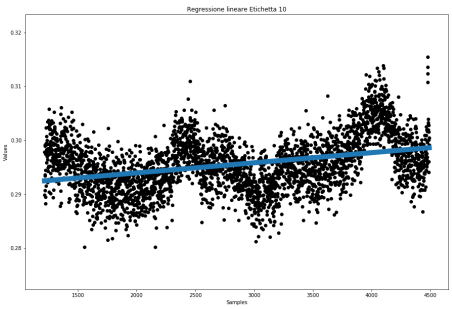
- nella classe 0, in cui la media varia tra 0.28 A e 0.3 A, si possono notare delle fluttuazioni dei valori della corrente. La retta che approssima i dati ha una pendenza lievemente positiva.

- nella classe 10 la media varia tra 0.28 A e 0.315 A e, anche in questo caso, si osservano delle oscillazioni regolari. La retta che approssima i dati è inclinata positivamente e la pendenza è maggiore rispetto a quella riscontrata nell'etichetta 0.

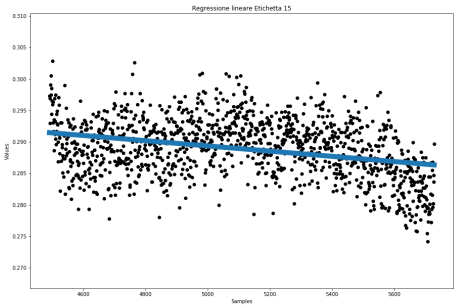
- nella classe 15, in cui la media varia tra 0.27 A e 0.3 A, non si distinguono fluttuazioni evidenti come quelle presenti negli andamenti delle medie delle etichette 0 e 10. La retta che approssima i dati ha una pendenza negativa, come nell'etichetta 15 del dataset *Gray*.



(a) *Regressione classe 0*



(b) *Regressione classe 10*



(c) *Regressione classe 15*

Figura 5.10: Dataset *White*. Regressione lineare separatamente per etichetta

5.3 Preprocessing e Data Transformation

Dopo aver analizzato i dati della corrente, si procede con la preparazione dei dati per le analisi successive. In primo luogo, in fase di Preprocessing sono stati esclusi da tutte le analisi che verranno descritte in seguito i due cicli di produzione anomali individuati nel dataset *Gray*. Nella fase di Data Transformation si opera una trasformazione delle time series grezze raccolte dai sensori in un insieme di variabili indipendenti dal tempo. Con questa strategia si ottiene una significativa riduzione della dimensionalità di entrambi i dataset, pur preservandone il contenuto informativo. Ciascun ciclo è ripartito in numero di suddivisioni stabilito dall'utente nel dominio del tempo, così da catturare la variabilità intra-ciclo; nel caso in esame, è stato scelto di dividere ogni ciclo di produzione in 24 split. Per ciascuna suddivisione sono state calcolate 14 misure statistiche, quali:

- **Media**;
- **Minimo, massimo**;
- **Deviazione standard**;
- **Primo quartile, mediana, terzo quartile**;

- **Curtosi**: si tratta di un indice statistico relativo alla forma della distribuzione che riflette la concentrazione dei dati attorno alla propria media. Se l'indice è maggiore di zero i dati collocano molto vicini alla media e la curva della distribuzione apparirà di forma allungata e appuntita; al contrario, se la curtosi assume valori minori di zero si otterrà una curva appiattita dovuta a dati più sparsi rispetto alla media.

- **Skewness**: si tratta di un indice di simmetria di una distribuzione. Un valore maggiore di zero indica che i dati si collocano per lo più a sinistra rispetto alla propria media; un valore minore di zero, invece, indica che i dati si collocano per lo più a destra; infine, un valore nullo indica che i valori sono equamente distribuiti intorno alla propria media.

- **Root Meen Squared Error o RMSE**, ovvero la radice quadrata della somma degli errori medi, calcolati come differenza tra valore previsto e valore osservato. Per costruzione, l'RMSE è sempre positivo e, tanto più è prossimo allo zero, tanto migliore è l'indice.

- **Somma dei valori assoluti**;

- **Numero di elementi oltre la media**;

- **Energia assoluta**;

- **Mean absolute change**, ovvero la differenza numerica tra ciascuna coppia di valori consecutivi nel segnale.

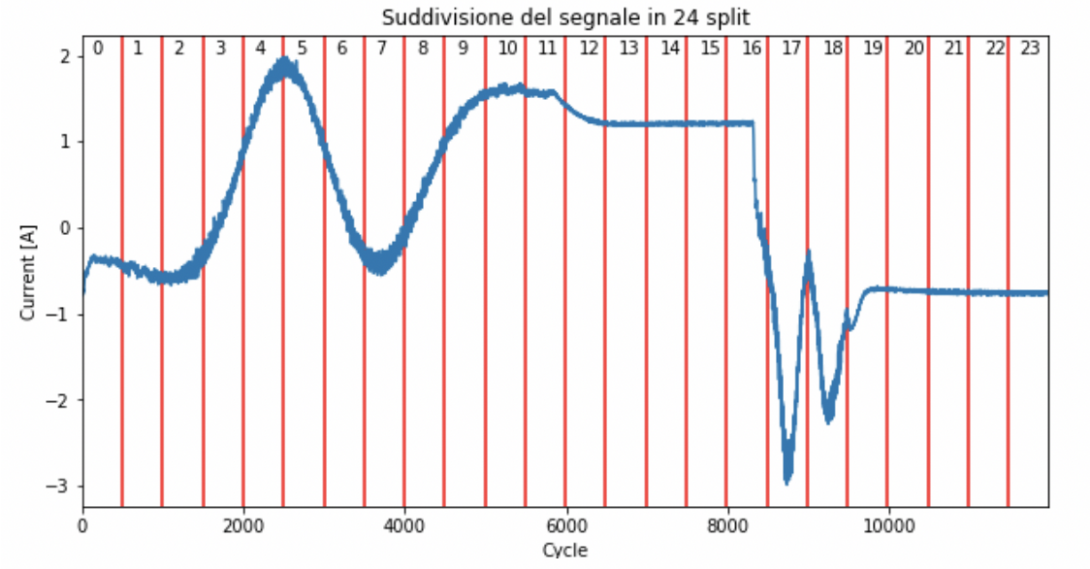


Figura 5.11: Suddivisione del segnale in 24 split

Poichè le caratteristiche vengono calcolate per ciascuna suddivisione, con questa strategia si genera una quantità numerosa di attributi, nel caso in esame un totale di $24 \times 14 = 336$, che, in alcune applicazioni, può influenzare negativamente le performance delle analisi successive. Di conseguenza, dato che alcune caratteristiche potrebbero essere altamente correlate con altre senza aggiungere alcun contenuto informativo interessante, si opera una *feature selection* al fine di sfruttare nelle analisi solamente gli attributi più rilevanti. A questo scopo, per ciascuna caratteristica si calcola il coefficiente *Mean Absolute Correlation* o anche MAC, ovvero la correlazione a coppie tra tutte le features [46]. Si è stabilito di escludere dalle analisi tutte le variabili che presentavano un valore di MAC superiore di una soglia impostata, in questo caso pari a 0.5. Al termine di questa procedura si sono ottenuti 299 attributi per il dataset *Gray* e 249 per il dataset *White*: ciò significa che su 336 attributi sia in *Gray* che *White* ne sono risultati ridondanti rispettivamente 37 e 87.

Per maggiore chiarezza, è stata eseguita una riduzione della dimensionalità di ciascun dataset mediante la rappresentazione in PCA (Principal Component Analysis) per mostrare graficamente come si distribuiscono i cicli nelle etichette 0, 10 e 15.

L'analisi delle componenti principali (PCA) è una tecnica ampiamente utilizzata in diversi campi, come la riduzione della dimensionalità e l'estrazione delle caratteristiche. Infatti, la PCA permette di individuare le variabili caratterizzate dalla massima varianza nei dati e di proiettarle su un nuovo sistema cartesiano con dimensioni uguali o inferiori a quello originale. Utilizzando una trasformazione lineare, il set di dati originale, che potrebbe essere composto da numerose variabili, è rappresentato solo con poche caratteristiche (dette componenti principali), che costituiscono l'output della PCA. In questo contesto è stato scelto di ridurre la dimensionalità del dataset in esame a tre componenti principali per meglio osservare la distribuzione dei cicli di produzione nelle tre etichette.

Come si può osservare dalla Figura 5.12, nel dataset *Gray* le classi 0 e 10 risultano compatte e vicine tra loro, mentre la classe 15 si colloca più distanziata dalle precedenti; d'altra parte, questa disposizione si poteva presupporre dall'analisi dell'andamento della media, nella quale è stato osservato come la classe 15 avesse dei valori mediamente inferiori rispetto alle due precedenti. Anche nel dataset *White* si evince questa particolarità, ma il distacco della classe 15 è minore e meno evidente, come mostrato in Figura 5.13.

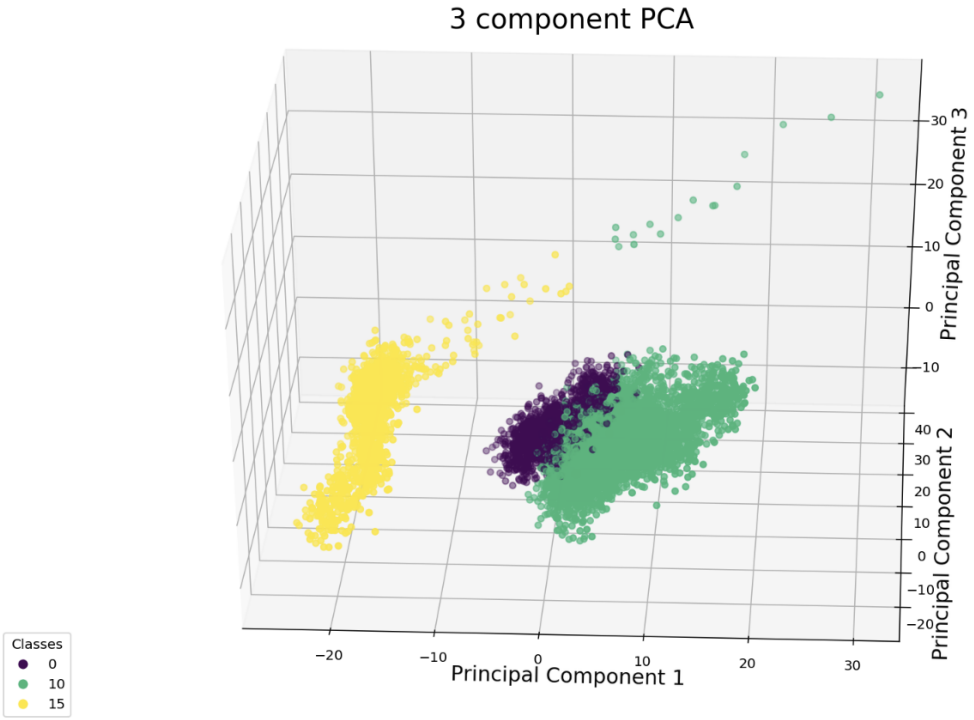


Figura 5.12: Rappresentazione in PCA dataset *Gray*

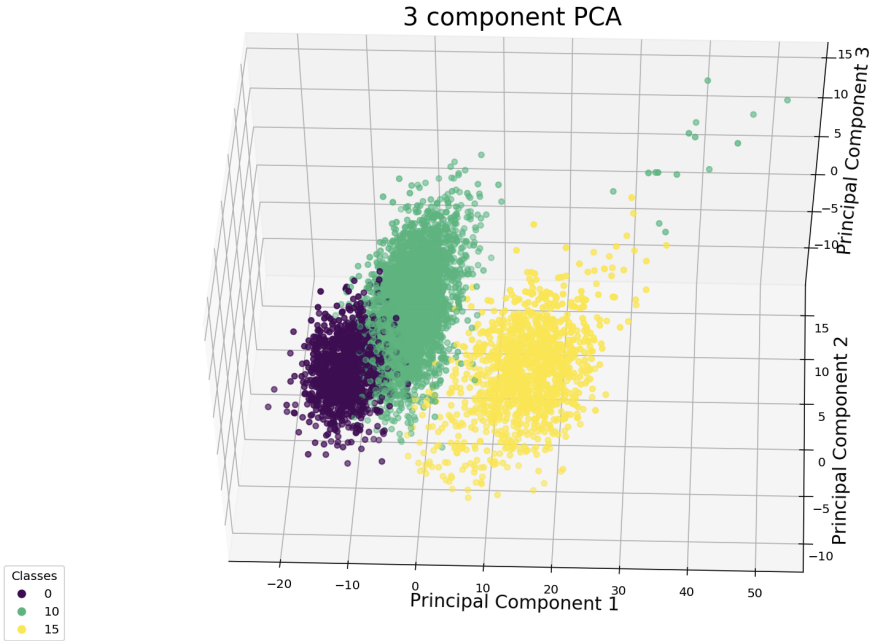


Figura 5.13: Rappresentazione in PCA dataset *White*

5.4 Cluster Analysis

Lo scopo di quest'analisi consiste nell'individuare gruppi di cicli di produzione dotati di caratteristiche simili, coerenti e ben separati a partire dai dati grezzi raccolti. In questa sede, si testeranno le performance degli algoritmi di clustering descritti nel capitolo precedente, ovvero K-Means, Agglomerative Hierarchical Clustering e DBSCAN. Prima di applicare gli algoritmi, tuttavia, è necessario normalizzare i dati; in questo caso di studio è stata utilizzata la normalizzazione Z-score. In prima battuta si individueranno i parametri di input ottimali per ciascun algoritmo utilizzando delle tecniche che sfruttano la distribuzione dei dati. In seguito, si verificherà la bontà delle partizioni ottenute confrontandole con le etichette reali di ogni ciclo. Verranno dapprima presentati i risultati del clustering per il dataset *Gray* e, in seguito, quelli relativi a *White*.

5.4.1 K-Means

Gray

Il K-Means è uno degli algoritmi di clustering più diffusi e rapidi dal punto di vista del tempo di esecuzione, ma, come già anticipato, prevede la scelta a priori del parametro K pari al numero di partizioni desiderate. Allo scopo di selezionare il miglior parametro di input, è stata calcolata e rappresentata la Silhouette al variare di K. Considerata la notevole separazione tra le tre classi, com'era possibile aspettarsi, la Silhouette ha un andamento decrescente dopo $K = 3$ indicando un peggioramento nella coesione/separazione dei cluster per valori di $K > 3$. Per questo valore di K, la Silhouette è pari a 0.2409. Figura 5.14.

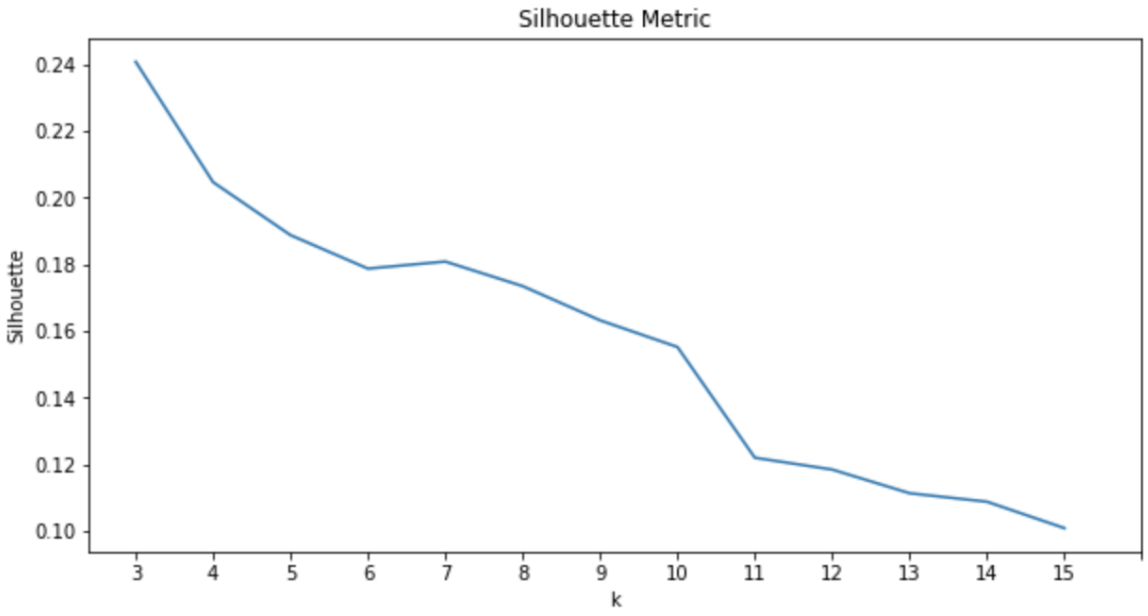


Figura 5.14: Silhouette dataset *Gray* al variare di K

Con il parametro K così individuato, il K-Means restituisce tre partizioni. Confrontando la rappresentazione in PCA dei punti con le etichette di cluster evidenziate in Figura 5.15 con la distribuzione dei punti con le etichette reali in Figura 5.12, emerge che il K-Means rispetta piuttosto fedelmente la suddivisione originale dei dati. Inoltre, a conferma di ciò è stato calcolato l'Adjusted Rand Index, il quale risulta pari a 0.6145; si ricorda che tanto più il valore dell'indice è prossimo all'unità, migliore è l'attività di etichettatura dell'algoritmo.

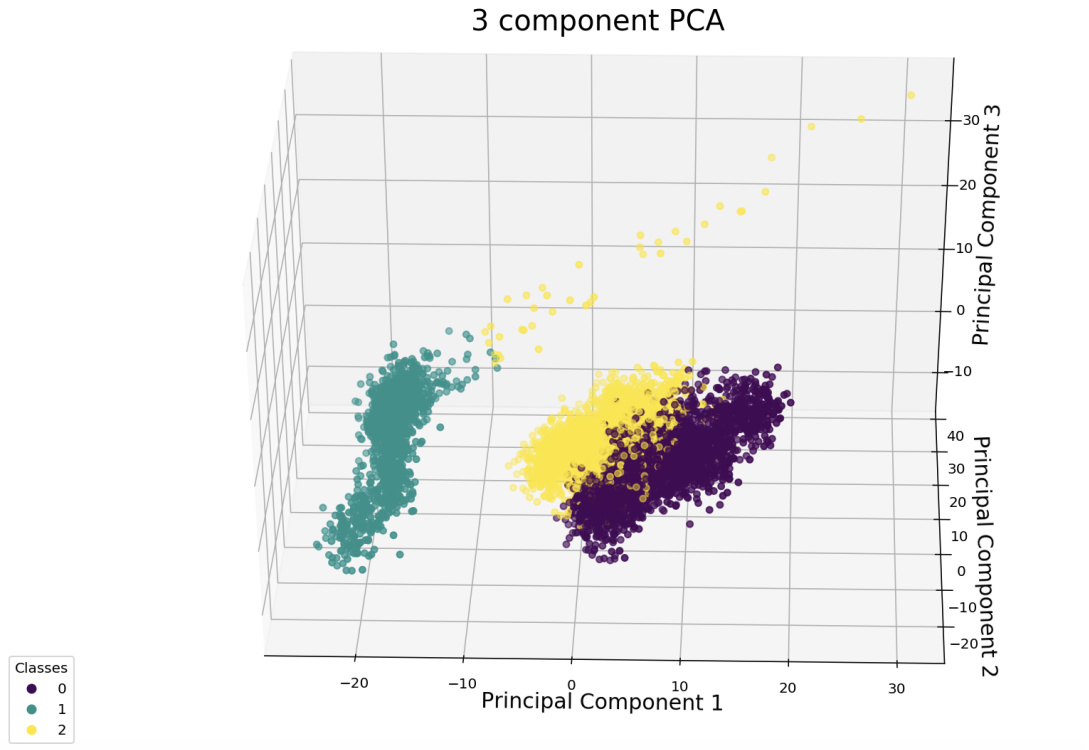


Figura 5.15: Rappresentazione PCA dataset *Gray* etichette K-Means

Per maggiore chiarezza, il grafico in Figura 5.16 mostra come le etichette reali dei punti si distribuiscono nei tre cluster risultanti. Da quanto mostrato dal grafico, si evince che il cluster 2 contiene tutti gli elementi della classe 0; il cluster 1 contiene tutti gli elementi dell'etichetta 15, di cui una minima frazione si colloca nella cluster 2, insieme alla classe 0; infine, la classe 10 si distribuisce per la maggior parte nel cluster 0 e parzialmente nel cluster 2.

Per un ulteriore approfondimento, a partire dalle etichette ottenute dal K-Means è stato applicato il Decision Tree Classifier allo scopo di identificare gli attributi che contribuiscono maggiormente alla formazione di ciascun cluster. L'albero risultante ha un'accuratezza del 96.17% e i quattro attributi più rilevanti ai fini della composizione dei cluster sono risultati: third_quartile_18, kurtosis_16, third_quartil_13, std_17. Le caratteristiche ottenute appartengono quasi tutte al tratto del segnale in cui la corrente subisce un consistente calo,

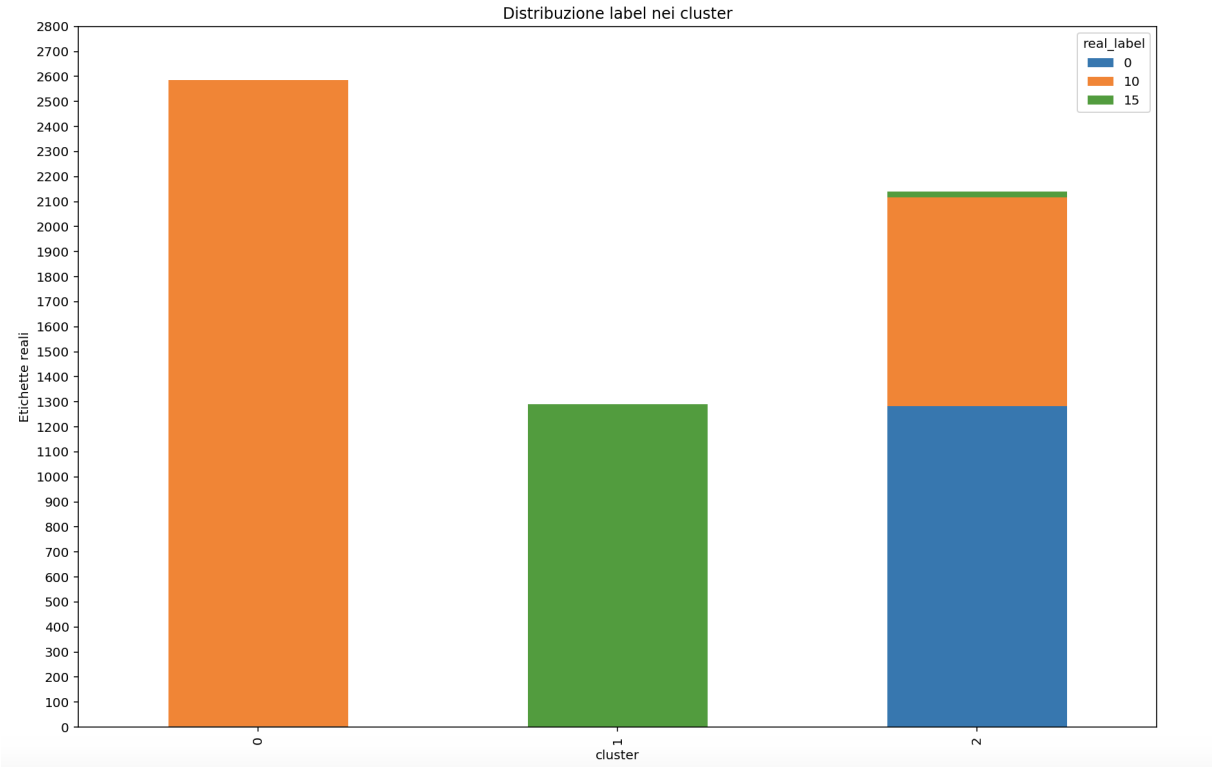


Figura 5.16: Dataset *Gray*. Distribuzione delle etichette nei cluster, K-Means con K = 3

cui segue una zona oscillante. Sono stati rappresentati i valori assunti dai centroidi dei tre cluster in un radar chart per verificare come le caratteristiche sopraccitate si posizionano.

Da quanto emerge dalla Figura 5.17, si può notare come la differenza di valore assunto da ciascuna caratteristica nei tre diversi cluster (messi in evidenza dai colori) renda ciascuna feature rilevante al fine della suddivisione dei punti nei tre gruppi. Ad esempio, third_quartile_18 assume valori molto differenti nel cluster 1 rispetto al cluster 0 e 2, che invece presentano valori molto simili. Dunque, tale feature contribuisce a separare gli elementi del cluster 1 da quelli in 0 e 2.

In alternativa, per selezionare il numero di partizioni ottimale come parametro di input del K-Means si può utilizzare l'Elbow Method, in cui si riporta il valore dell'SSE per diversi valori di K. Quando nella curva dell'SSE si nota un "gomito", ovvero un punto di flesso, il valore in corrispondenza sulle ascisse coincide con il parametro K da selezionare. Nel caso in esame, per implementare l'Elbow Method è stata utilizzata la libreria di Python Yellowbrick, che contiene una serie di strumenti visivi e diagnostici che permettono di rendere più immediati i risultati degli algoritmi della libreria Scikit-learn. Il metodo *kelbow_visualizer()* permette di calcolare l'SSE e di individuare immediatamente il punto di flesso. In questo caso, il punto di flesso è posizionato in corrispondenza di $K = 4$. Figura 5.18.

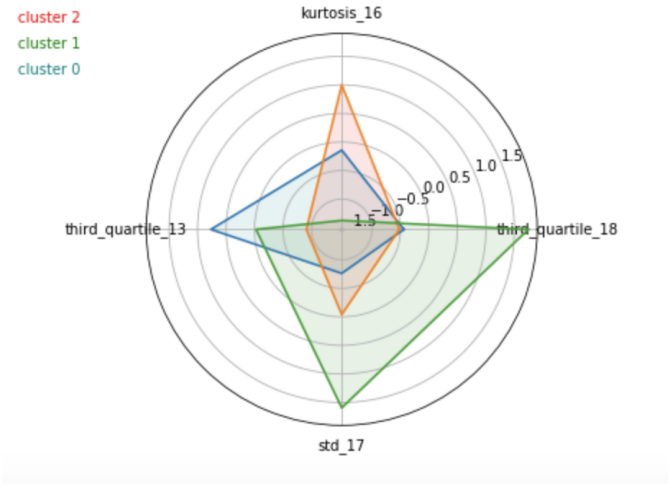


Figura 5.17: Radar Chart dataset *Gray*, valore dei centroidi negli attributi più rilevanti

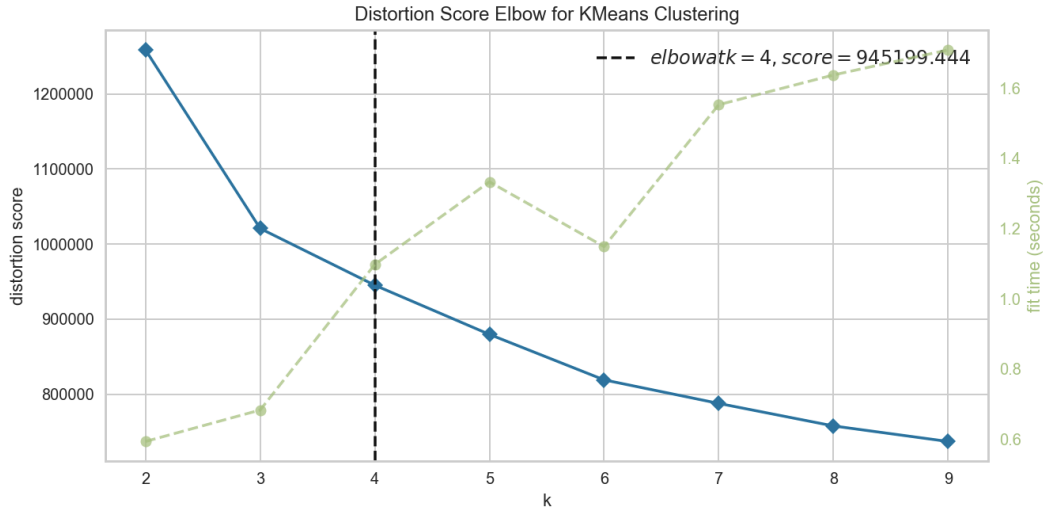


Figura 5.18: Elbow Method dataset *Gray*

Ripetendo le medesime analisi si ottengono quattro partizioni, in cui tuttavia le metriche di valutazione della bontà del cluster interne ed esterne, Silhouette e Adjusted Rand Index, hanno subito un calo, rispettivamente a 0.2047 e 0.4346. Infatti, anche la distribuzione delle etichette nei cluster è peggiorata notevolmente: i cicli di produzione etichettati con 10 sono stati ripartiti in tre cluster; i cicli etichettati con 0 sono stati accorpati ad una parte della classe 10; la classe 15, invece, è stata assegnata quasi integralmente ad un cluster separato, ad eccezione di pochi punti distribuiti in altri cluster. Figura 5.19.

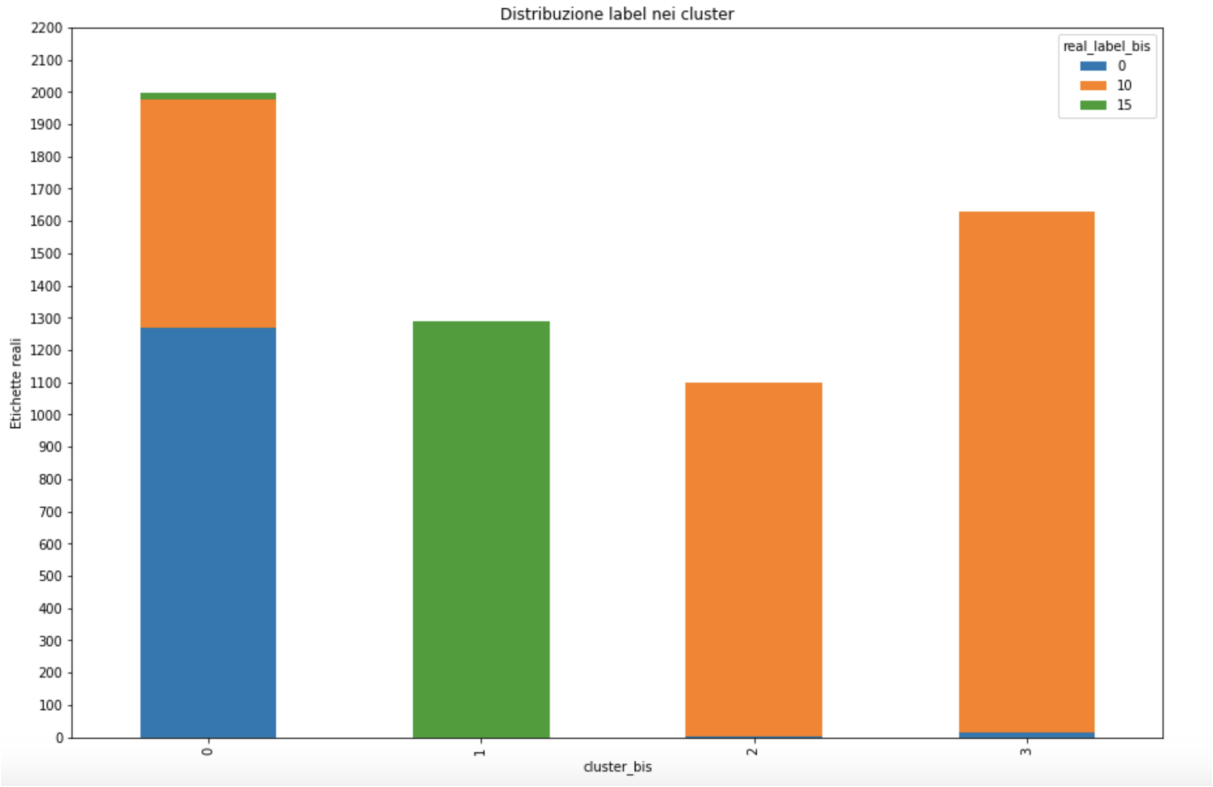


Figura 5.19: Dataset *Gray*. Distribuzione delle etichette nei cluster, K-Means con K = 4

White

Come nel caso precedente, è stata calcolata e rappresentata la Silhouette al variare di K per individuare il numero di partizioni ideale come input del K-Means. Anche per il dataset *White* valgono le stesse considerazioni sull'andamento della Silhouette, mostrata in Figura 5.20. L'andamento è decrescente per valori di $K > 3$, ad indicare un peggioramento nella coesione/separazione dei cluster. Per questo motivo, $K = 3$ è selezionato come parametro di input del K-Means.

Eseguendo una sessione del K-Means con il parametro K così individuato si generano tre gruppi di dati. Anche in questo caso, il K-Means rispetta la partizione originale dei dati e ciò si può affermare facendo il confronto tra la rappresentazione PCA con le etichette reali in Figura 5.13 e lo stesso grafico con le label assegnate dal clustering mostrato in Figura 5.21. Ciò emerge anche osservando il valore l'Adjusted Rand Index che risulta prossimo all'unità, più precisamente pari a 0.9556. Inoltre, guardando lo Stacked Bar in Figura 5.22 si vede che i tre cluster separano in modo quasi impeccabile le etichette reali dei cicli di produzione, ad eccezione di pochi punti delle classi 10 e 15.

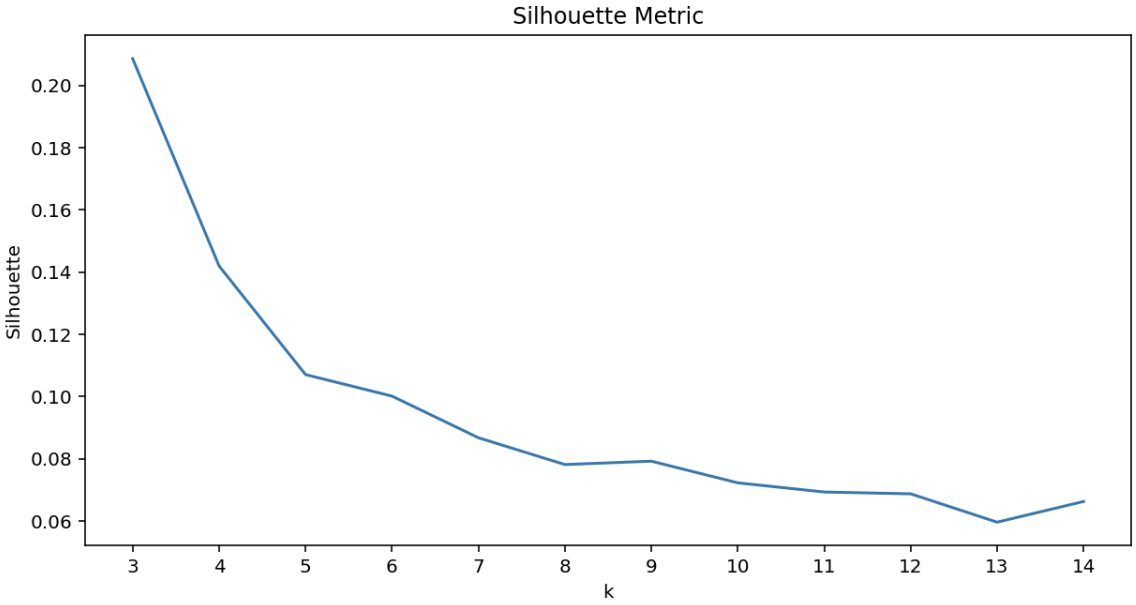


Figura 5.20: Silhouette dataset *White* al variare di K

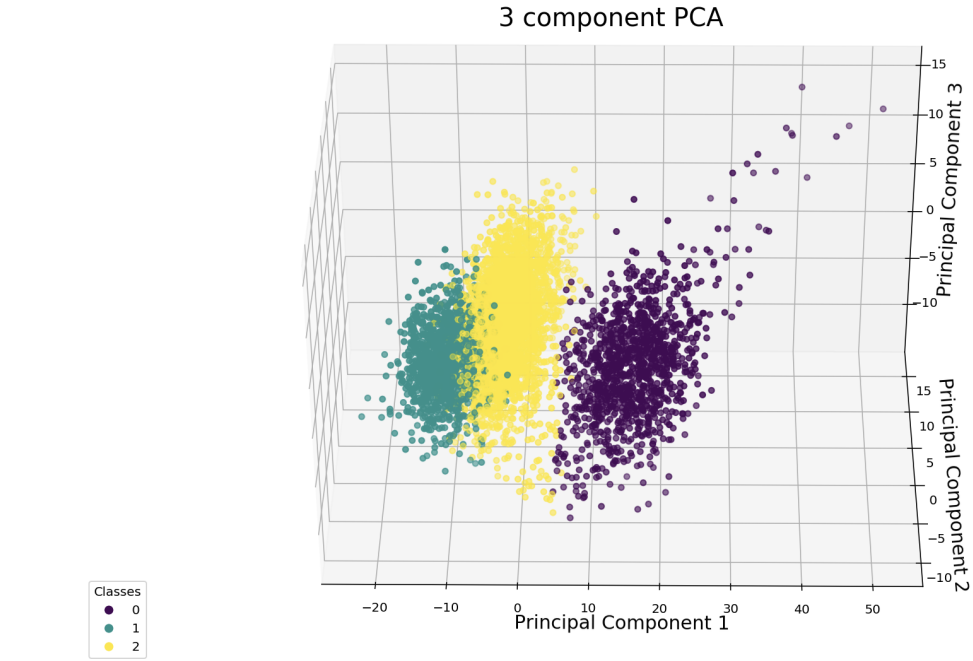


Figura 5.21: Rappresentazione PCA dataset *White* etichette K-Means

Anche per il dataset *White* è stata applicato il Decision Tree Classifier sulle etichette ottenute dal K-Means per individuare gli attributi che contribuiscono maggiormente alla formazione dei cluster. Rispetto al dataset precedente non coincidono nè le variabili che determinano la suddivisione dei punti nei tre cluster nè le porzioni rilevanti di segnale.

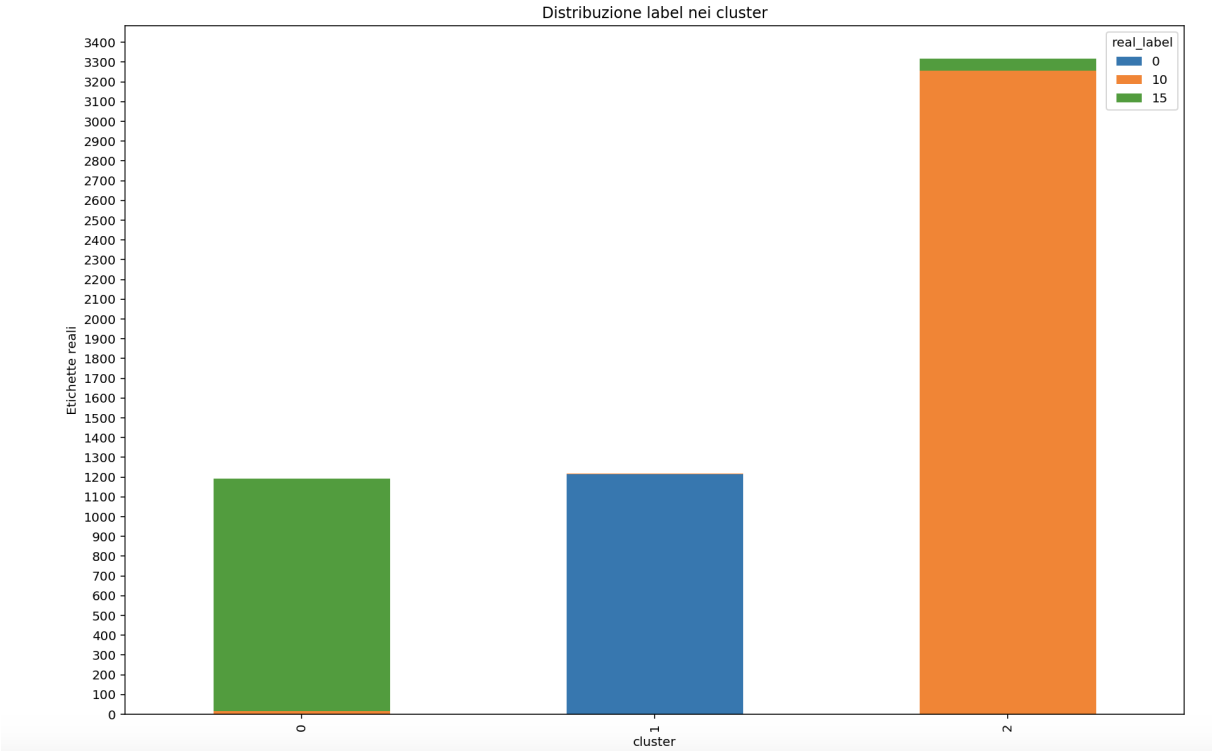


Figura 5.22: Dataset *White*. Distribuzione delle etichette nei cluster, K-Means con K = 3

La Figura 5.23 contiene un Radar Chart in cui sono stati riportati i valori assunti dagli attributi più rilevanti nei tre centroidi, ovvero mean_abs_change_8, std_12, skewnes_0, third_quartile_23. Ad esempio, si può notare come la feature mean_abs_change_8 assuma un valore decisamente minore nel cluster 1 e maggiore per i cluster 0 e 2.

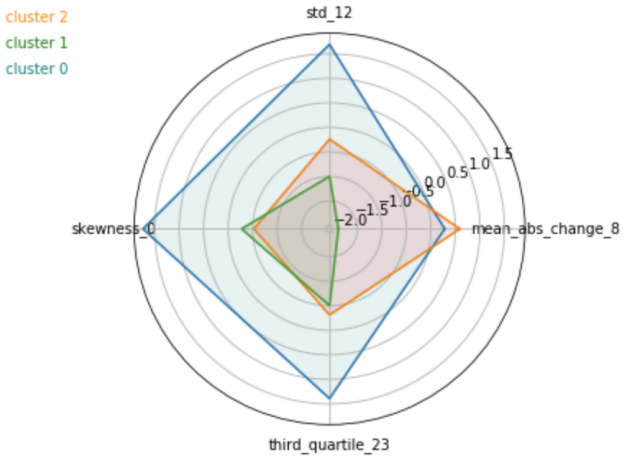


Figura 5.23: Radar Chart dataset *White*, valore dei centroidi negli attributi più rilevanti

Come per il dataset *Gray*, è stato applicato l'Elbow Method come alternativa al metodo della Silhouette per individuare il numero di partizioni ottimale. É stata ottenuta una curva simile a quella in Figura 5.18 e che ha fornito come valore ideale $K = 4$. Eseguendo nuovamente le stesse analisi precedenti si ottengono quattro partizioni, in cui tuttavia la bontà della sessione di clustering è peggiorata: infatti, sia la Silhouette che Adjusted Rand Index hanno subito un calo, rispettivamente a 0.1418 e 0.6116. I cicli di produzione etichettati con 10 sono stati ripartiti in due cluster, mentre i cicli etichettati con 0 e 15 sono stati assegnati quasi integralmente a due cluster separati, ad eccezione di una frazione di punti della classe 15 che l'algoritmo ha accorpato al cluster 1, contenente per lo più classe 10. Figura 5.24.

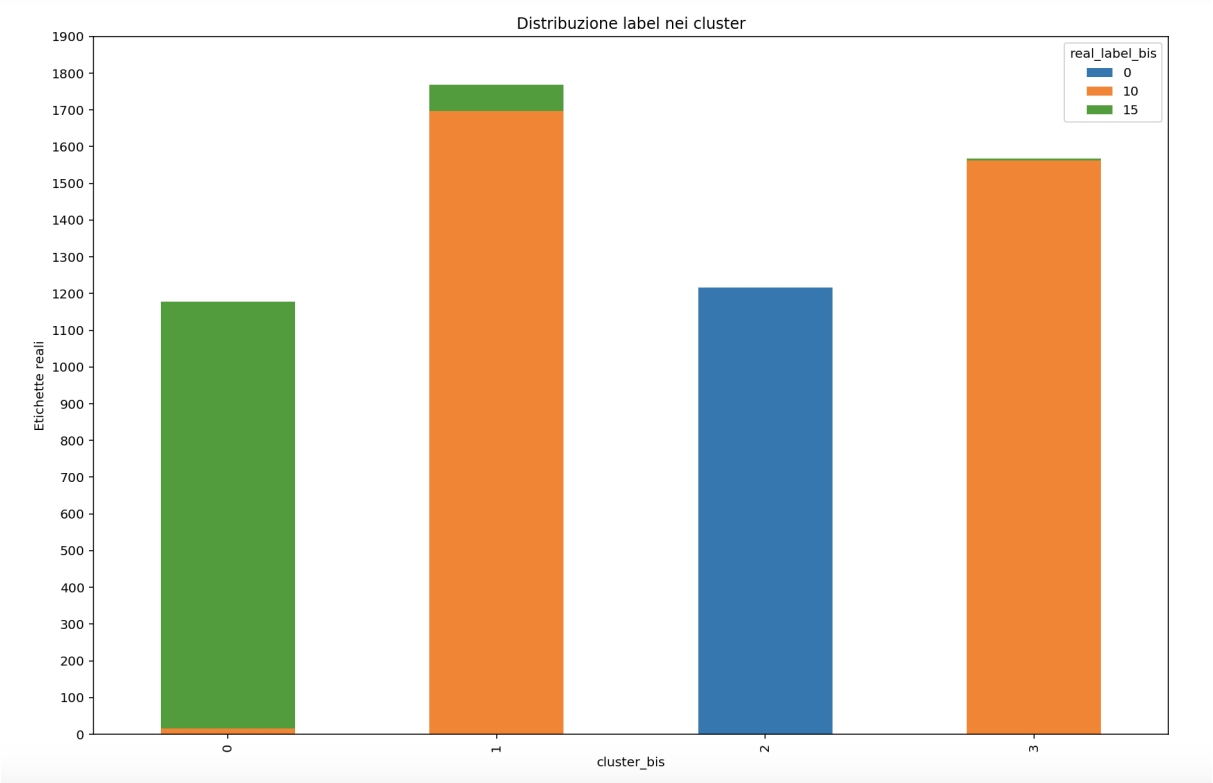


Figura 5.24: Dataset *White*. Distribuzione delle etichette nei cluster, K-Means con K = 4

5.4.2 DBSCAN

Gray

Il DBSCAN è un algoritmo di clustering che si basa sul concetto di densità e fornisce una soluzione partizionale e parziale. Tale algoritmo riceve come parametri di input *Eps* e *MinPoints*. Ciascun cluster formato dal DBSCAN è costituito da un insieme di punti *core* e dai punti *border* a loro associati; un punto è definito *core* se, all'interno di una circonferenza con centro il punto *core* e raggio *Eps*, sono presenti almeno un numero di punti pari a *MinPoints*; un punto è definito *border* se è localizzato nelle vicinanze di un punto *core*; infine, un punto è definito *noise* se non è in prossimità nè di un *core* nè di un *border point*. Com'è già stato anticipato, *Eps* e *MinPoints* devono essere impostati a priori dall'utente. Per identificare i parametri di input ottimali è stato rappresentato il grafico del *k-dist* al variare di *MinPoints*. Sono stati testati valori di *MinPoints* da 5 a 10 e si è constatato che la curva rimane pressocchè invariata per valori di *MinPoints* maggiori di 8. Il valore di *Eps* si trova in corrispondenza del ginocchio della curva, in questo caso per $Eps = 11$, come si può osservare dalla curva in Figura 5.25. Eseguendo il DBSCAN con i valori individuati si ottengono tre cluster. Per validare le partizioni generate, si confrontano le etichette ottenute dal DBSCAN con quelle originarie e si ottiene un valore di Adjusted Rand Score pari a 0.8010. Ciascun cluster separa i cicli di produzione appartenenti ad etichette differenti: ciò massimizza l'indice di purezza del cluster. Poichè il DBSCAN è un algoritmo di tipo partizionale non completo, alcuni punti sono etichettati come *noise point*: tutti questi sono racchiusi all'interno del cluster -1, che contiene una frazione delle classi 0,10 e 15, coinvolgendo complessivamente l'11% dei record del dataset.

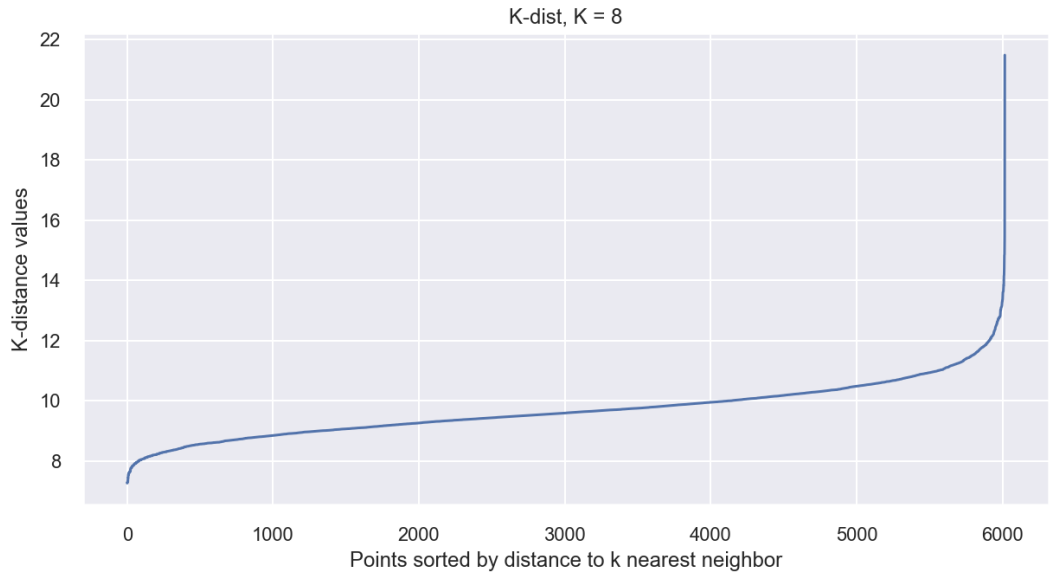


Figura 5.25: K-dist dataset *Gray* con k = 8

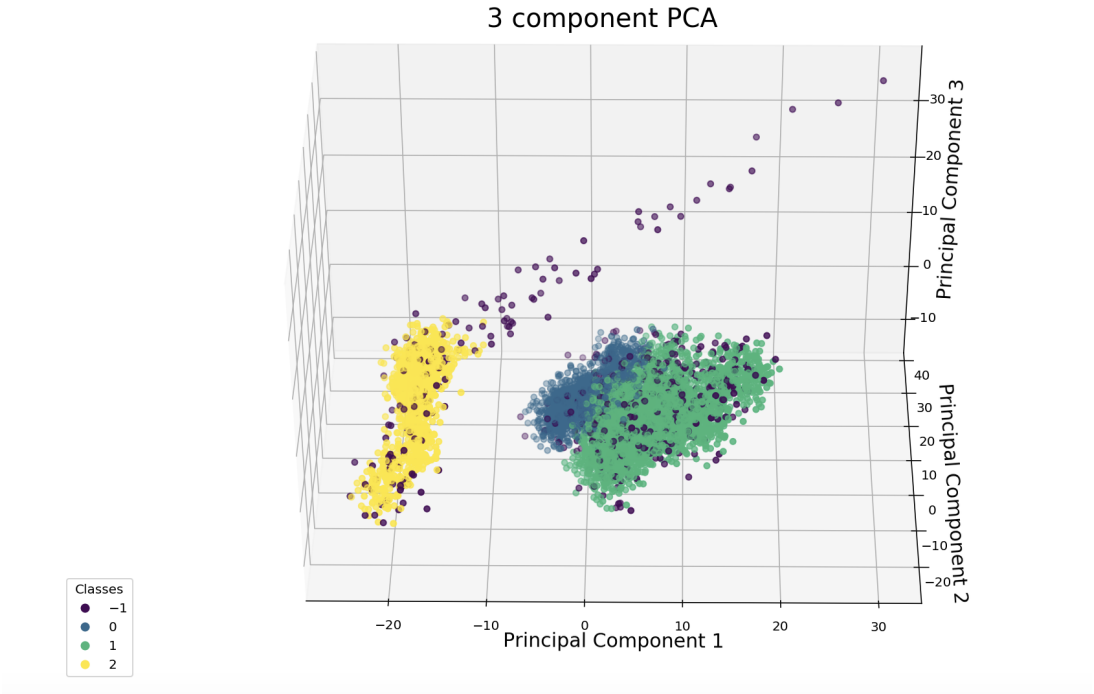


Figura 5.26: Rappresentazione PCA dataset *Gray* etichette DBSCAN

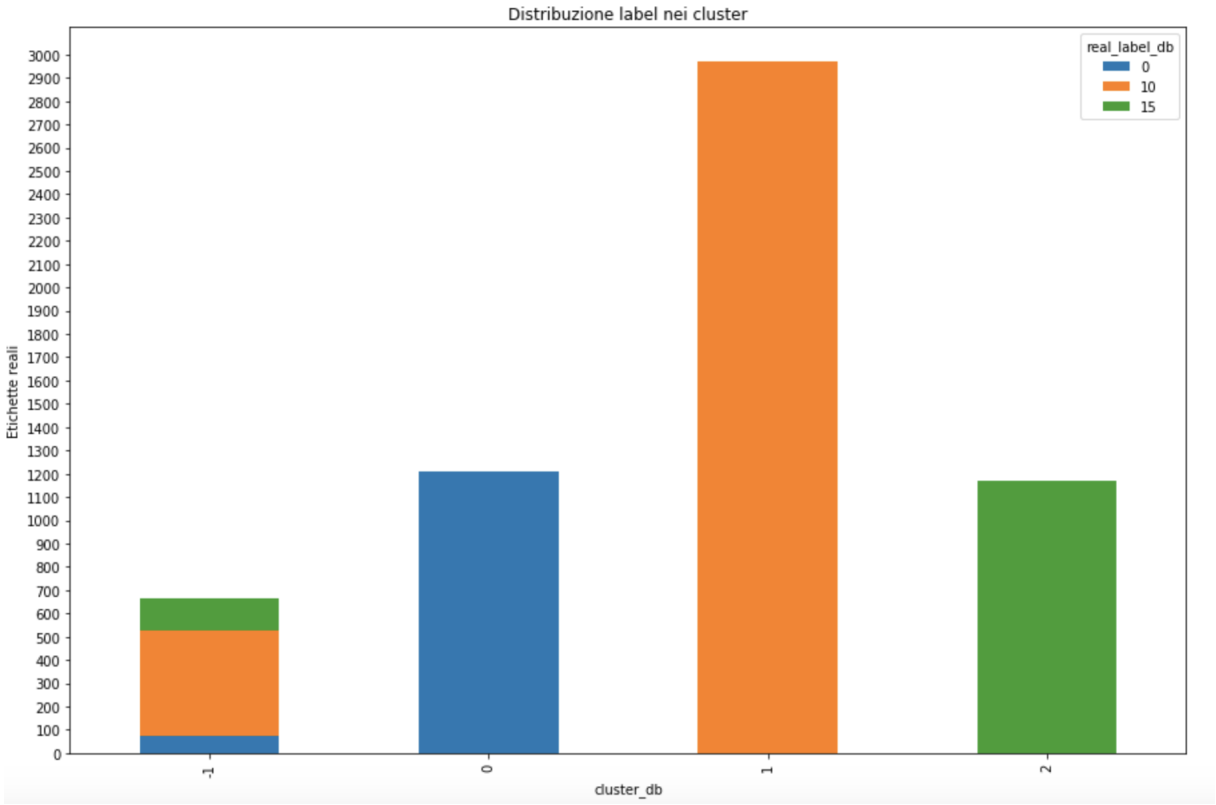


Figura 5.27: Dataset *Gray*. Distribuzione delle etichette reali nei cluster DBSCAN

White

Secondo lo stesso principio utilizzato per il dataset *Gray*, è stato rappresentato il grafico del *k-dist* testando valori di *MinPts* tra 5 e 10 ed infine è stato stabilito di rappresentare il grafico del *k-dist* per $k = 7$. A partire da questa curva, molto simile a quella precedentemente ottenuta per il dataset *Gray*, è stato individuato il gomito in corrispondenza di $Eps = 12$. Una volta ricavati i parametri di input per il DBSCAN, è possibile eseguire l'algoritmo e ottenere tre partizionamenti. Analogamente alla collezione dati precedente, ogni cluster contiene una sola tipologia di etichetta; in aggiunta, il DBSCAN ha costruito il cluster -1 in cui sono inseriti tutti i valori identificati dall'algoritmo come *noise points*, per un totale di 886 cicli di produzione, che corrispondono circa al 16% dei record totali del dataset.

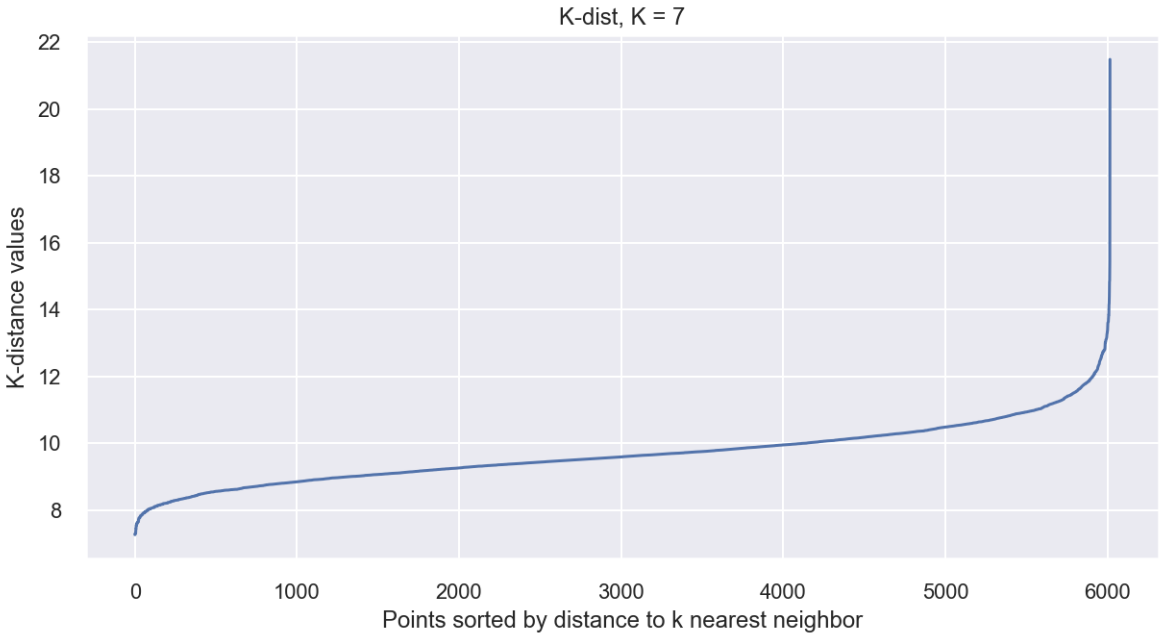


Figura 5.28: K-dist dataset *White* con k=7

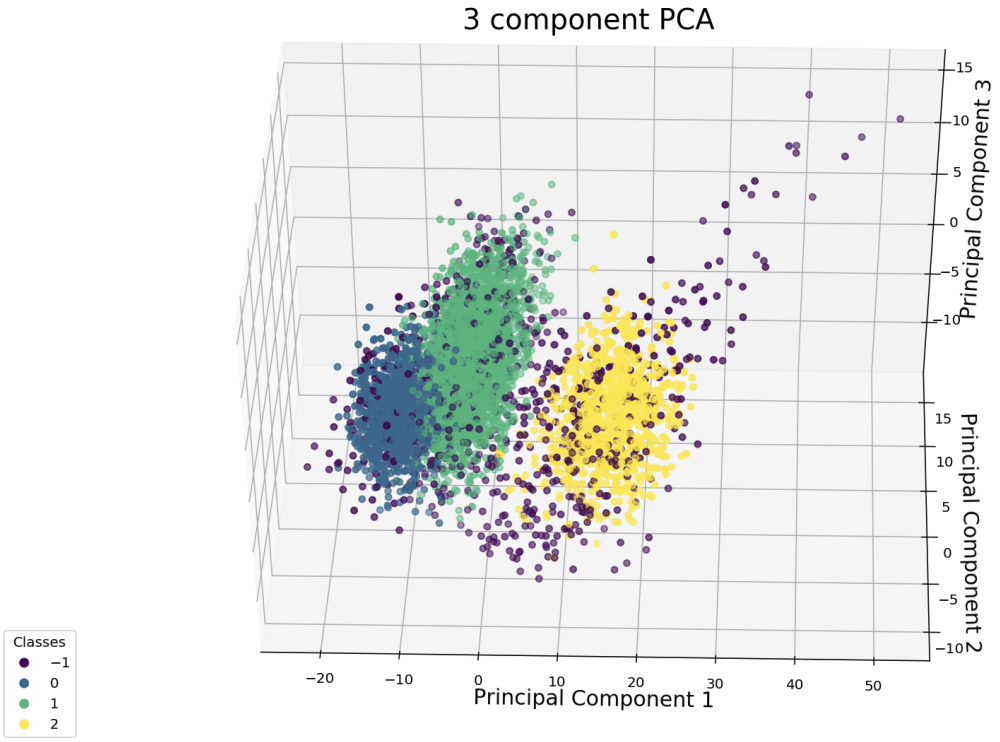


Figura 5.29: Rappresentazione PCA dataset *White* etichette DBSCAN

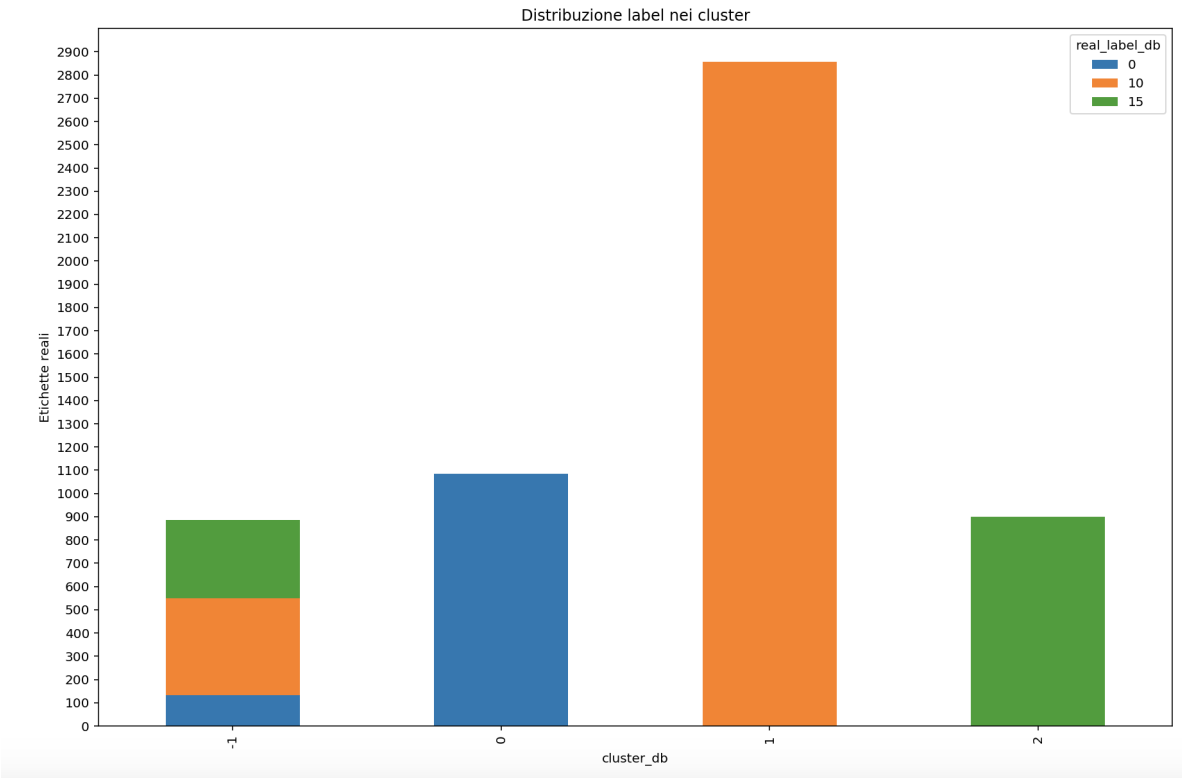


Figura 5.30: Dataset *White*. Distribuzione delle etichette reali nei cluster DBSCAN

5.4.3 Agglomerative Hierarchical Clustering

Gray

L'Agglomerative Hierarchical Clustering è un algoritmo di clustering gerarchico che fornisce come soluzione un insieme di cluster annidati. Come per il K-Means, è necessario definire a priori il numero di partizionamenti che si desidera ottenere. La scelta del numero ottimale di cluster può essere effettuata osservando il dendrogramma, un diagramma che rappresenta dal basso verso l'alto la sequenza di fusione dei cluster. In letteratura esistono più strategie per aggregare i cluster, tra le quali:

- Ward: unisce i cluster che presentano la minima varianza intra-cluster;
- Complete: utilizza la distanza massima tra punti appartenenti a gruppi diversi per fondere due cluster;
- Single: utilizza la minima distanza tra punti appartenenti a gruppi diversi per fondere due cluster;
- Average: utilizza la media distanza tra punti appartenenti a gruppi diversi per fondere due cluster.

Chiaramente, è necessario stabilire a priori una misura per calcolare la distanza tra i punti e, nel caso in esame, è stato scelto di utilizzare la distanza Euclidea. Inoltre, sono stati valutati diversi criteri di fusione dei cluster ed infine è stato deciso di utilizzare il metodo *ward*. L'albero ottenuto è rappresentato in Figura 5.31.

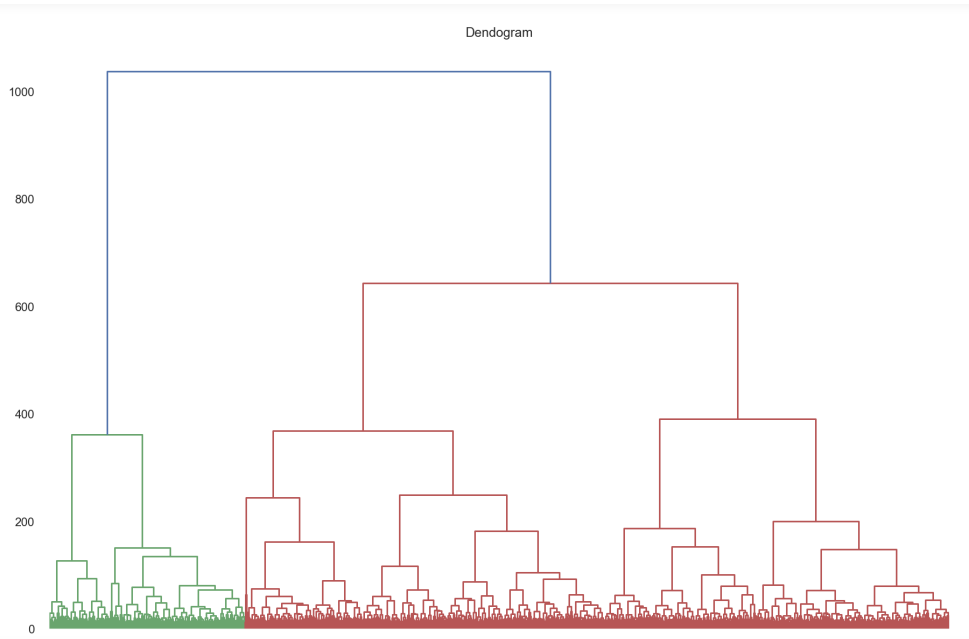


Figura 5.31: Dendrogramma dataset *Gray*

Nel dendrogramma le linee orizzontali indicano gli accoppiamenti tra i cluster e il tratto verticale che unisce due linee orizzontali rappresenta la distanza che deve essere colmata per fonderli. Quindi, minore è il tratto verticale di un accoppiamento, più vicini sono i cluster. Per selezionare il numero ottimale di cluster da ottenere si può tracciare una linea orizzontale in corrispondenza del tratto verticale più alto in quanto in questo modo si otterranno dei cluster ben separati. Osservando il dendrogramma si è ipotizzato di effettuare un taglio in modo da ottenere tre partizioni. A conferma di ciò, è stato calcolata la Silhouette media al variare del numero di partizioni e ne è stato rappresentato l'andamento. É emerso che il valore maggiore, pari a 0.222, si ottiene proprio con $n\_cluster = 3$. Per valori di 3, la Silhouette diminuisce e ciò indica un peggioramento della separazione/coesione dei cluster che si determinerebbero. Dalla rappresentazione in PCA in Figura 5.32 si può notare che effettivamente le partizioni ottenute con l'Agglomerative Clustering rispecchiano la reale suddivisione dei cicli di produzione mostrate in Figura 5.12. Per validare le partizioni generate si è osservata la suddivisione delle etichette reali dei cicli di produzione nei tre cluster e, calcolando l'Adjusted Rand Score, si ottiene un valore pari a 0.5629. Si può notare che la classe 10 è racchiusa in parte in un cluster che contiene solamente cicli di produzione di questa categoria, mentre la restante frazione è stata accorpata alla classe 0 nel cluster 0; alla classe 15, invece, è stato dedicato interamente il cluster 1. Figura 5.33.

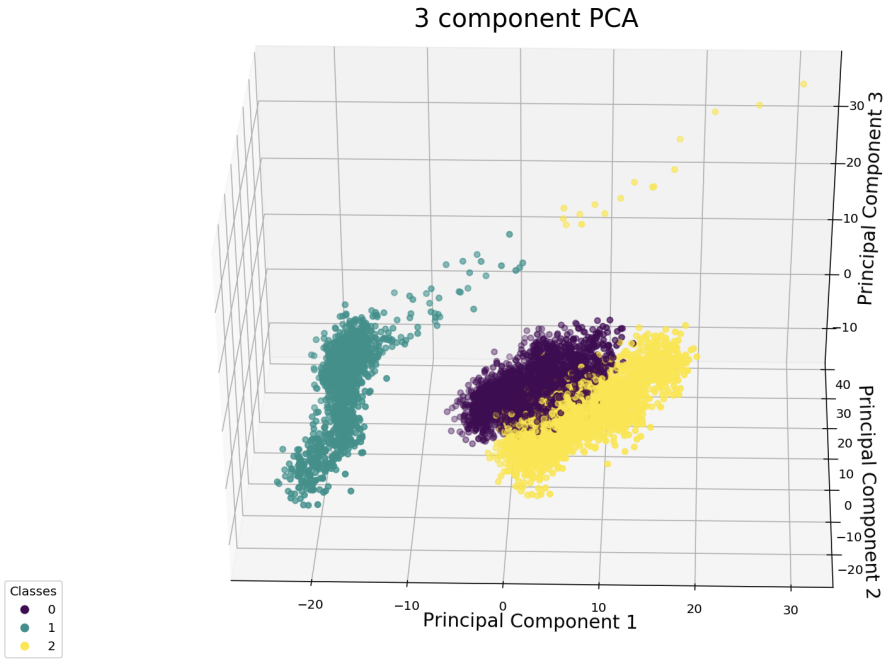


Figura 5.32: Rappresentazione in PCA dataset *Gray*, Agglomerative Hierarchical Clustering con $n\_cluster = 3$

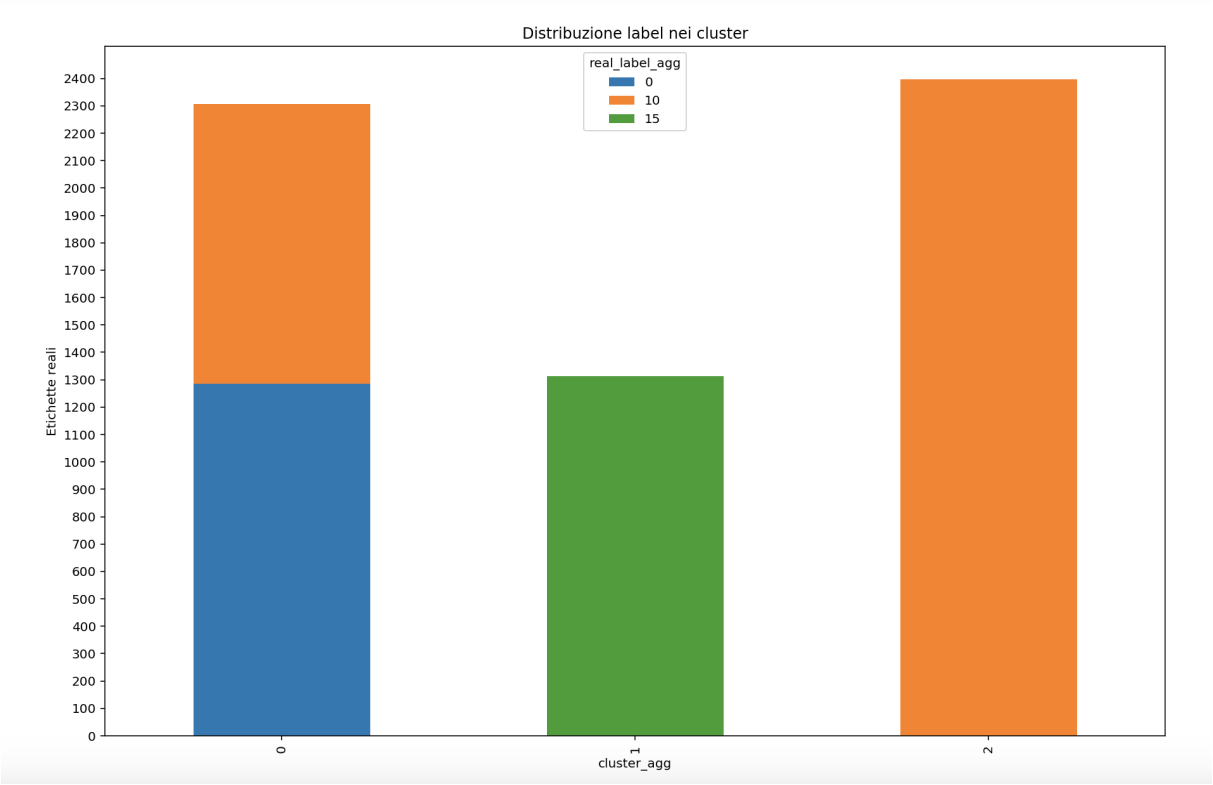


Figura 5.33: Dataset *Gray*. Distribuzione delle etichette nei cluster, Agglomerative Hierarchical Clustering con $n\_cluster = 3$

White

Analogamente al dataset precedente, è stato rappresentato il dendrogramma (Figura 5.34) per il dataset *White*, utilizzando come misura di distanza quella Euclidea e linkage la strategia di default, ovvero *ward*.

Anche da questo diagramma si intuisce che potrebbe essere effettuato un taglio in modo tale da determinare la formazione di tre cluster. É stato rappresentato l'andamento della Silhouette al variare del parametro $n\_cluster$ per determinare il numero di partizioni ideale da impostare come input dell'algoritmo; anche questa strategia suggerisce come miglior numero di partizionamenti 3, per gli stessi ragionamenti esposti in precedenza. Per questo valore di $n\_cluster$ il valore della Silhouette è pari a 0.2070. Eseguendo una sessione dell'Agglomerative Hierarchical Clustering con i parametri individuati si ottengono tre partizionamenti che, come si osserva dal confronto tra la Figura 5.35 e la Figura 5.13, rispecchiano quasi perfettamente la reale suddivisione dei cicli di produzione; infatti, l'Adjusted Rand Score è pari a 0.9915. Questo fatto si evince dallo Stacked Bar in Figura 5.36, in cui si nota che ciascun cluster contiene un'unica etichetta, ad eccezione del cluster 1 a cui si accorpano pochi cicli della classe 10.

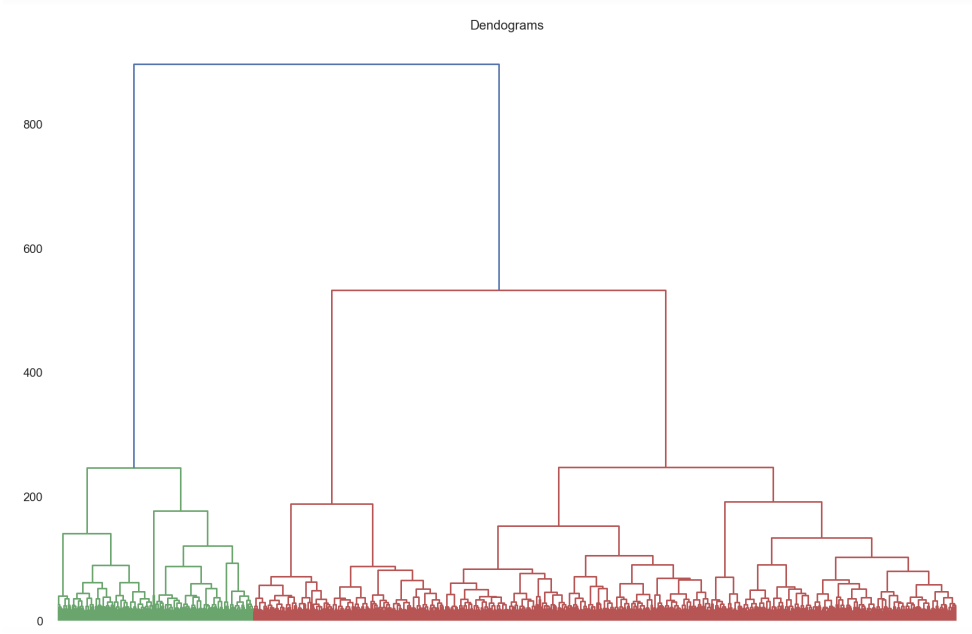


Figura 5.34: Dendrogramma dataset *White*

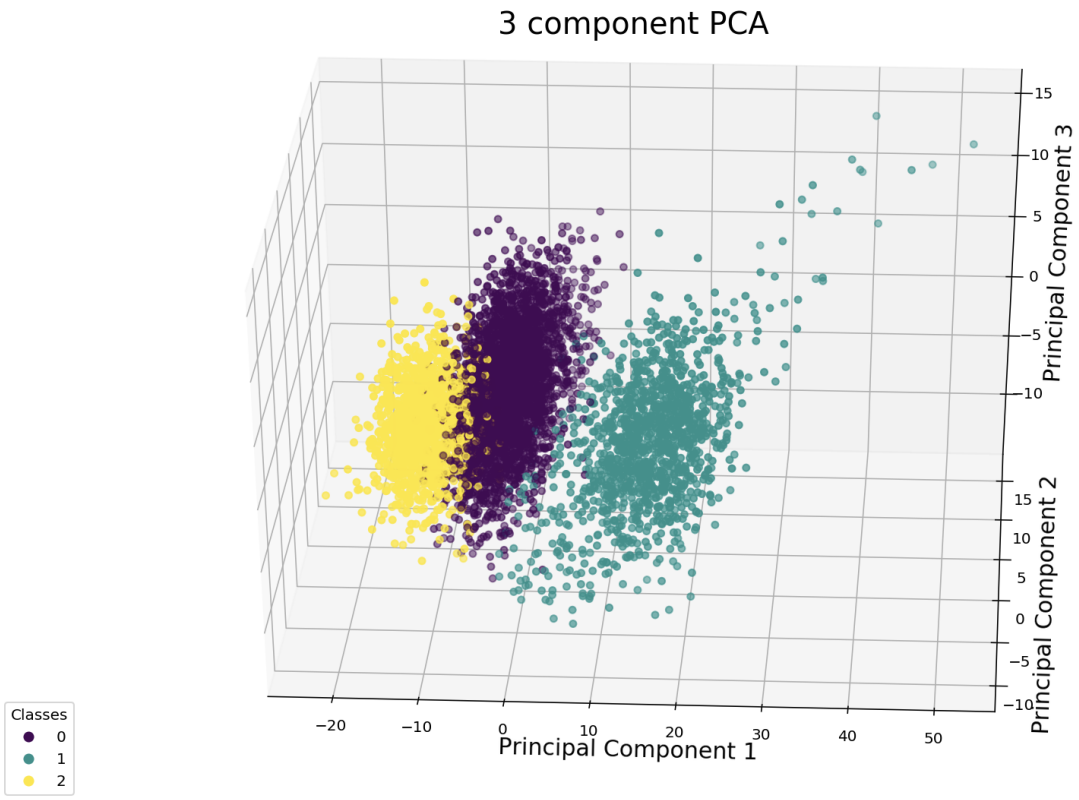


Figura 5.35: Rappresentazione in PCA dataset *White*, Agglomerative Hierarchical Clustering con $n\_cluster = 3$

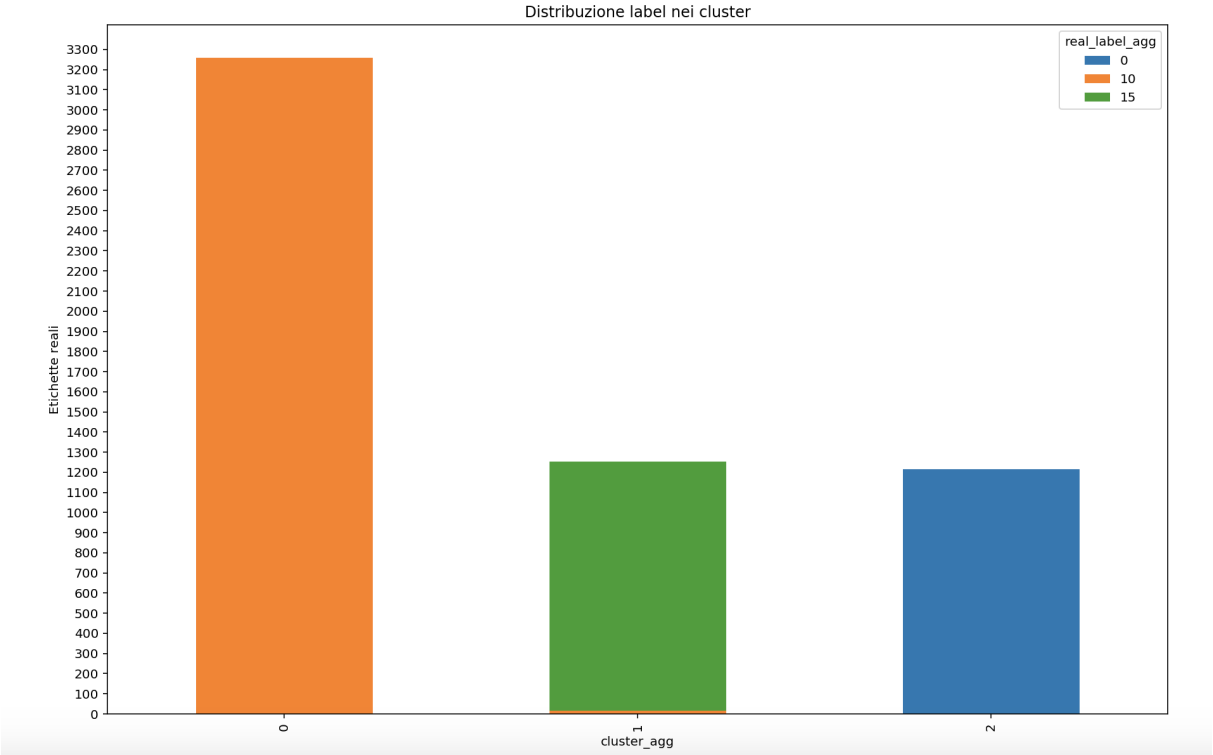


Figura 5.36: Dataset *White*. Distribuzione delle etichette nei cluster, Agglomerative Hierarchical Clustering con $n\_cluster = 3$

5.4.4 Confronto tra algoritmi

Concludendo, le partizioni ottenute con K-Means, DBSCAN e Agglomerative Hierarchical Clustering sono state in grado di di suddividere correttamente i cicli di produzione se confrontate con le etichette di appartenenza reali dei dati. Di seguito, nella Tabelle 5.5 e 5.6 si riportano per ciascun algoritmo i parametri utilizzati come input e i valori dell'Adjusted Rand Score.

Algoritmo	Parametri	Adjusted Rand Score
K-Means	K = 3	0.6145
DBSCAN	$Eps = 11$ $MinPts = 8$	0.8010
Agglomerative Hierarchical Clustering	n_cluster = 3	0.5629

Tabella 5.5: Riepilogo per il dataset *Gray*

Algoritmo	Parametri	Adjusted Rand Score
K-Means	K = 3	0.9556
DBSCAN	$Eps = 12$ $MinPts = 7$	0.7578
Agglomerative Hierarchical Clustering	n_cluster = 3	0.9915

Tabella 5.6: Riepilogo per il dataset *White*

5.5 Cluster analysis con dati rumorosi

Come si è potuto notare per le analisi presentate nelle sezioni precedenti, i partizionamenti ottenuti sono risultati particolarmente efficaci ai fini della caratterizzazione dei cicli di produzione. Tuttavia, dopo un confronto con esperti del dominio, è emerso che i dati forniti sono stati creati ad hoc modificando manualmente il livello di tensione della cinghia. Per testare la robustezza degli algoritmi di clustering utilizzati è stata simulata la costruzione di un due dataset *Gray* e *White* "sporchi", in cui è stato aggiunto artificialmente una percentuale di rumore ai dati per riprodurre un caso di studio più realistico in cui le classi sono state avvicinate per simulare tensionamenti meno drastici della cinghia. A questo scopo, la procedura utilizzata è stata la seguente:

- Step 1: Sono date le classi {0, 10, 15}. Inizialmente, si fissa una classe di riferimento (ad esempio la classe 0) e si considera l'insieme di tutti i segnali (o cicli di produzione) che la compongono:

$$c_0 = \{x_0(t), x_1(t), ..., x_n(t)\}$$

dove n è il numero di segnali compongono la classe 0.

- Step 2: Per ogni classe j si calcola la funzione media dei valori che assumono i segnali di ciascuna classe ad ogni istante t:

$$Avg_j(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t)$$

- Step 3: Dopo avere calcolato le funzioni medie di tutte le classi per ogni istante t, si calcola lo scarto delle medie tra la classe di riferimento e le altre. Quindi, assumendo come riferimento la classe 0, si calcoleranno gli scarti:

$$S_{0-10} = Avg_0(t) - Avg_{10}(t)$$

$$S_{0-15} = Avg_0(t) - Avg_{15}(t)$$

- Step 4: Si considerano i segnali appartenenti alle classi da allineare, ovvero:

$$c_{10} = \{x_0(t), x_1(t), ..., x_m(t)\}$$

$$c_{15} = \{x_0(t), x_1(t), ..., x_p(t)\}$$

dove m e p rappresentano il numero di segnali che compongono rispettivamente le classi 10 e 15.

- Step 5: Su ogni segnale $x_i(t) \in c_j$ con $j = \{10,15\}$ si effettuano le seguenti operazioni:

  a. Per ogni segnale $x_i(t) \in c_j$ con $j = \{10,15\}$ si effettua l'allineamento aggiungendo il rumore come segue: $x_i(t) = x_i(t) + S_{0-j}$

  b. Si definisce la funzione di probabilità della distribuzione uniforme per l'aggiunta del rumore

$$p(x) = \frac{1}{b-a}$$

  dove $b = (-0.05 * x_i(t))$ e $a = (0.05 * x_i(t))$. Il parametro 0.05 è stato scelto a partire dai dati.

  c. Infine, si calcola $x_i(t) = x_i(t) + p(x)$

A seguito dell'introduzione del rumore tramite la procedura descritta, prima di procedere con le analisi successive sono state effettuate le stesse operazioni di preparazione dei dati eseguite sui dati costruiti ad hoc. Innanzitutto, è stato suddiviso ciascun ciclo di produzione in 24 split e, per ognuno di essi, sono state calcolate 14 misure statistiche, ottenendo in totale 336 attributi. Dopodichè, è stato calcolato l'indice di correlazione tra tutte le coppie di attributi ma, poichè gli attributi generati hanno presentato un indice inferiore alla soglia impostata del MAC, durante l'operazione di *feature selection* non è stato rimosso alcun attributo in nessuno dei due dataset.

Analogamente ai dati originari, per maggiore chiarezza è stata eseguita una riduzione della dimensionalità di ciascun dataset mediante la rappresentazione in PCA (Principal Component Analysis) per mostrare graficamente come si distribuiscono i cicli di produzione a seguito dell'introduzione del rumore. Figure 5.37 e 5.38. Come si può osservare facendo il confronto con le rappresentazioni in PCA dei dati originari nelle Figure 5.12 e 5.13, emerge immediatamente che le tre etichette non sono più nettamente separate e distinguibili.

Anche i due dataset *Gray with noise* e *White with noise* sono stati preventivamente normalizzati e, dopo avere individuato i parametri ideali a partire dalla distribuzione dei dati, è stata eseguita una sessione di ciascun algoritmo. In seguito, si esporranno i risultati ottenuti dapprima per il dataset *Gray with noise* e in seguito quelli per *White with noise*.

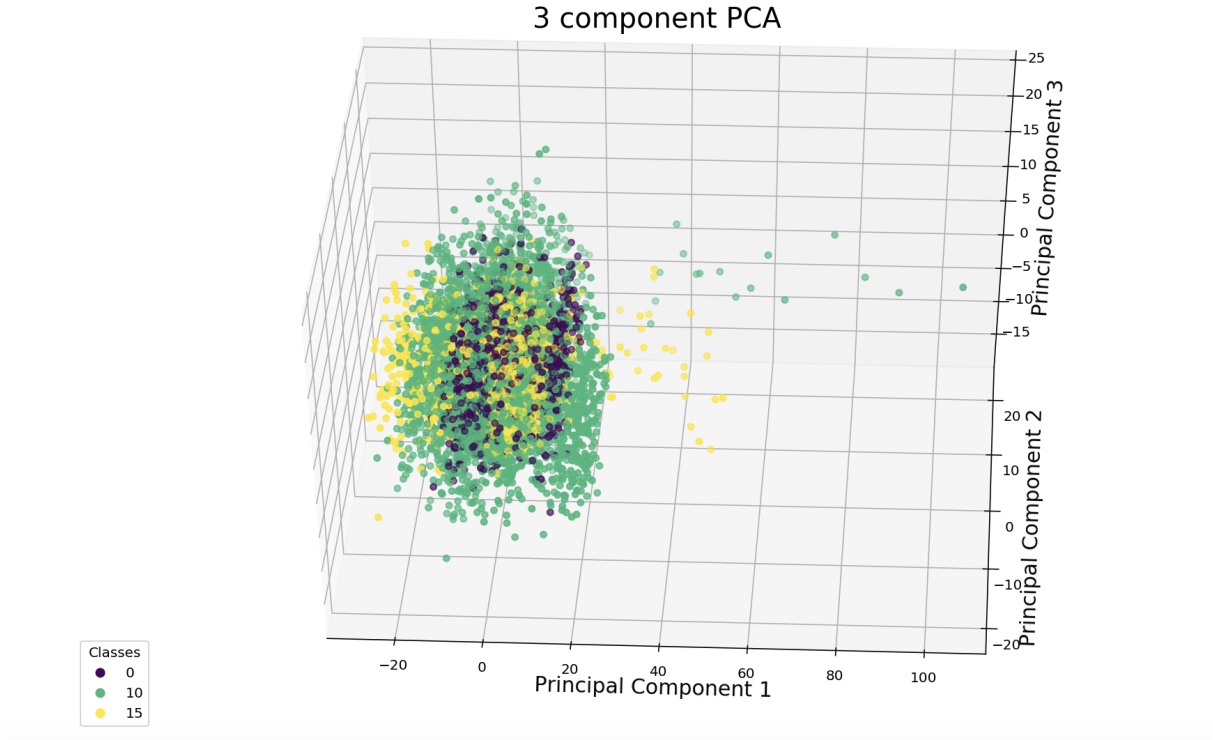


Figura 5.37: PCA dataset *Gray with noise*

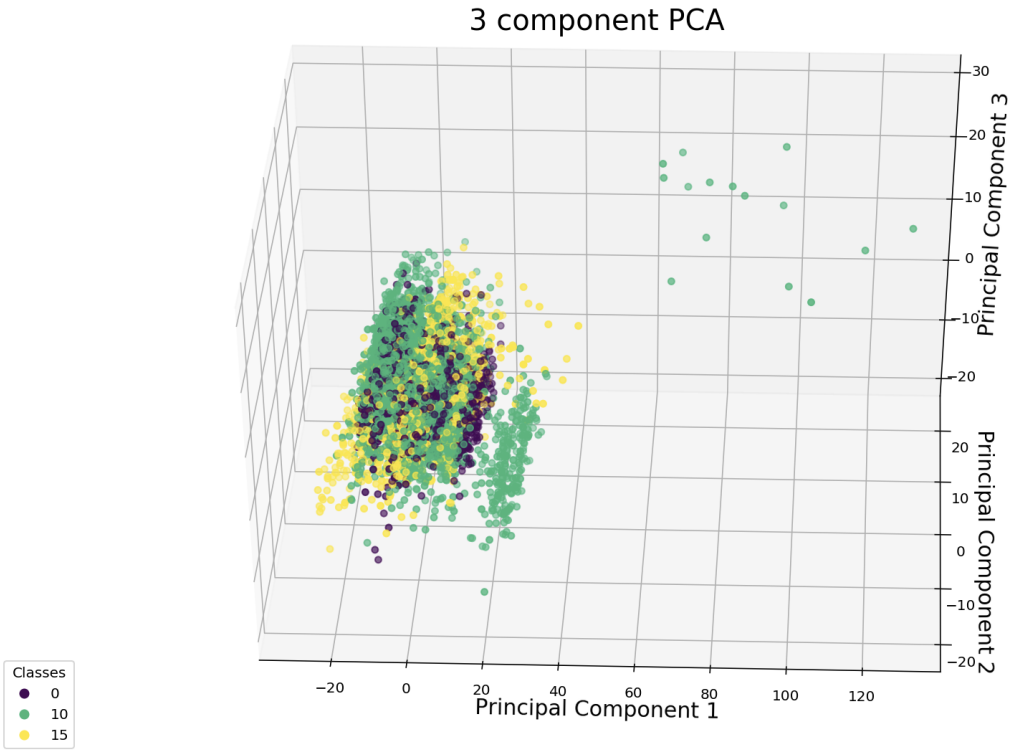


Figura 5.38: PCA dataset *White with noise*

5.5.1 K-Means

Gray with noise

Prima di procedere con l'esecuzione dell'algoritmo, è necessario individuare il numero di partizionamenti ottimali da fornire come input al K-Means. A questo scopo è stato rappresentato l'andamento della Silhouette media al variare del parametro K. Come nelle analisi precedenti, è stato selezionato come numero di partizionamenti ideale $K = 3$ in quanto per valori maggiori la Silhouette è decrescente e ciò indica un peggioramento della separazione/coesione dei gruppi. Il valore della Silhouette in corrispondenza di $K = 3$ è 0.1385. A primo impatto si può osservare come i valori della Silhouette siano sensibilmente inferiori rispetto al caso originale.

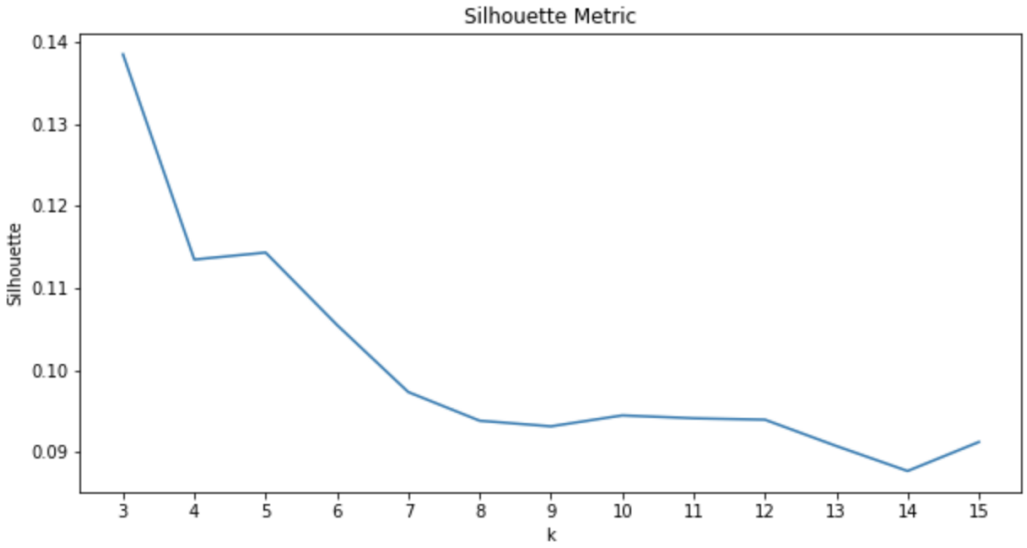


Figura 5.39: Silhouette dataset *Gray with noise* al variare di K

Eseguendo una sessione del K-Means così configurato si ottengono tre cluster separati tra loro che, tuttavia, non rispettano la suddivisione originale dei cicli nelle etichette. Infatti, anche l'Adjusted Rand Score è decisamente peggiorato e si assesta circa a 0.0095; si ricorda che, secondo la logica implementativa, un valore dell'indice prossimo allo zero indica un'attività di etichettatura dei dati casuale. Inoltre, come si può notare dallo Stacked Bar in Figura 5.41 si nota che i cluster contengono una frazione di ciascuna etichetta reale. Ad esempio, il cluster 0 contiene il 23.2% di dati etichettati con 0, il 67.7% di cicli appartenenti alla classe 10 mentre il restante 10% è costituito da punti etichettati con 15. Si può concludere che, anche se i cluster ottenuti sono separati tra loro, questi non rispettano la suddivisione originale dei cicli di produzione.

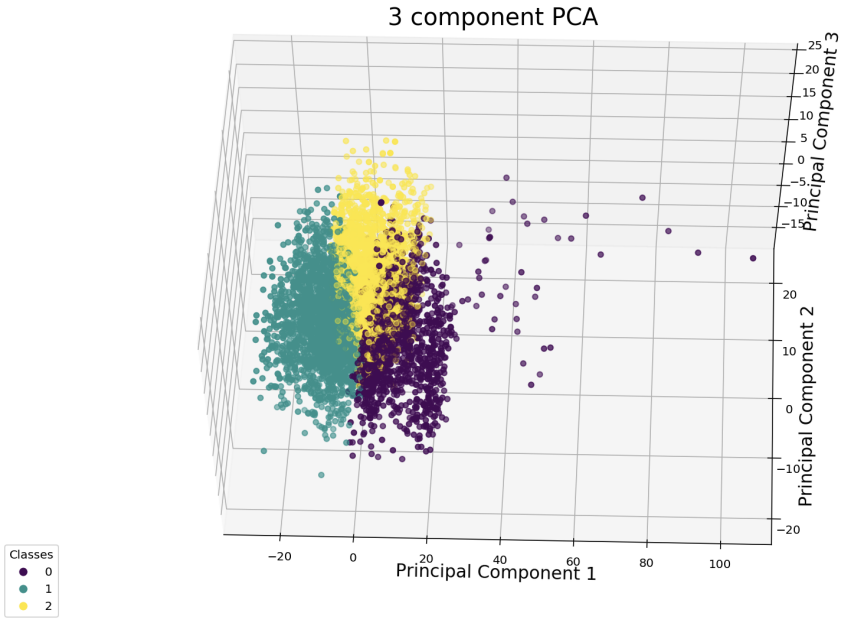


Figura 5.40: Rappresentazione PCA dataset *Gray with noise* K-Means con K = 3

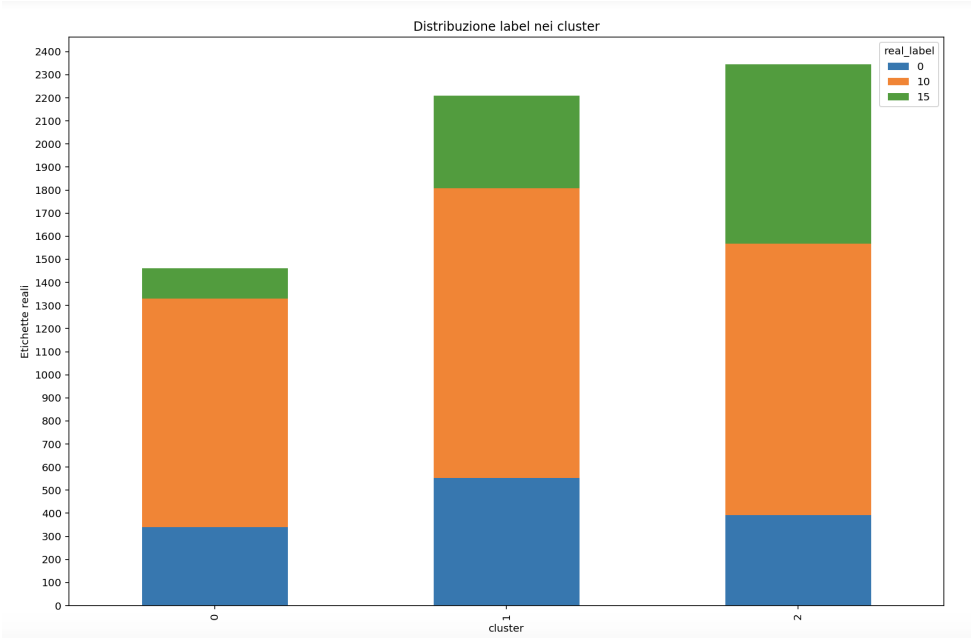


Figura 5.41: Dataset *Gray with noise*. Distribuzione delle etichette nei cluster, K-Means con $K = 3$

Inoltre, per trovare il numero di partizionamenti ottimale da fornire in input al K-Means è stato utilizzato l'Elbow Method. Come nel caso del dataset originario, secondo questo criterio il valore ottimale di gruppi è pari a quattro. Ripetendo l'analisi con $K = 4$ e confrontando la composizione dei cluster con le etichette reali dei punti si nota che ciascun

cluster è riempito con una frazione di ogni etichetta, come nel caso precedente con $K = 3$; a conferma di ciò, l'Adjusted Rand Score è rimasto invariato.

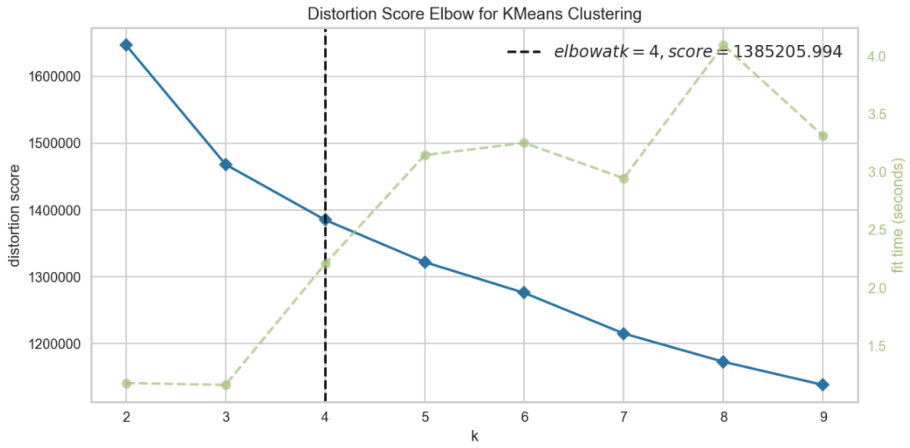


Figura 5.42: Elbow Method dataset *Gray with noise*

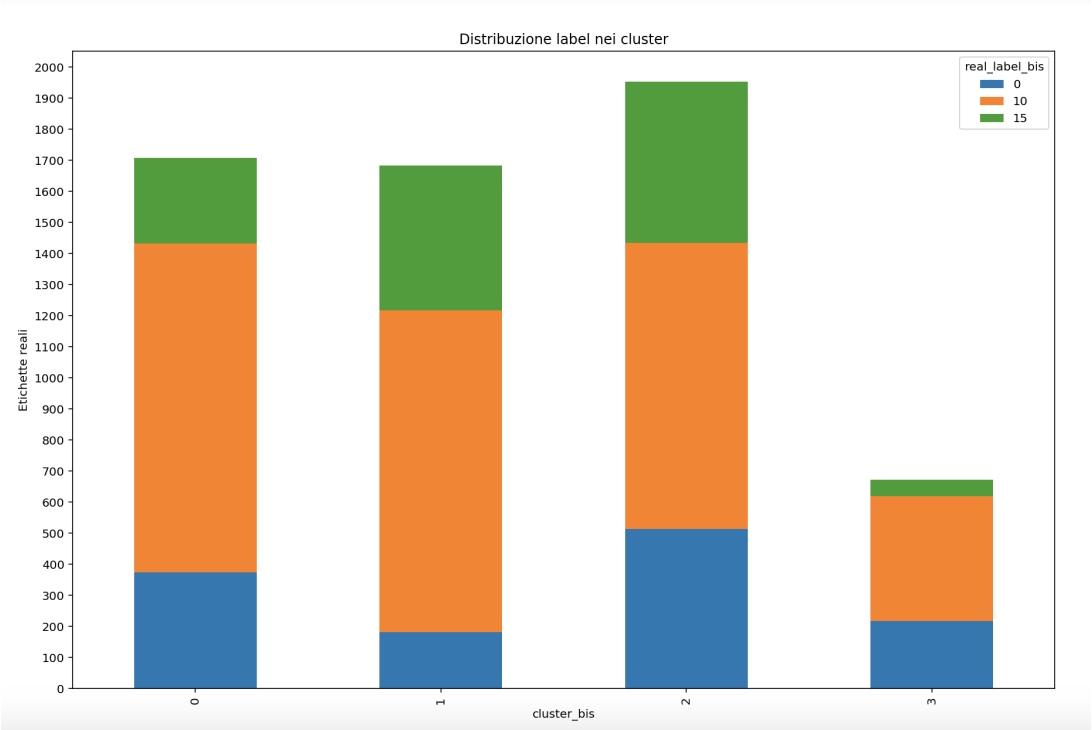


Figura 5.43: Dataset *Gray with noise*. Distribuzione delle etichette nei cluster, K-Means con $K = 4$

White with noise

Si utilizza nuovamente l'analisi dell'andamento della Silhouette media per determinare il parametro K ideale come input del K-Means. Come nel dataset *Gray with noise*, si verifica un sensibile calo della Silhouette massima rispetto al dataset di partenza a 0.127, valore che si riscontra anche in questo caso per $K = 3$. Figura 5.44.

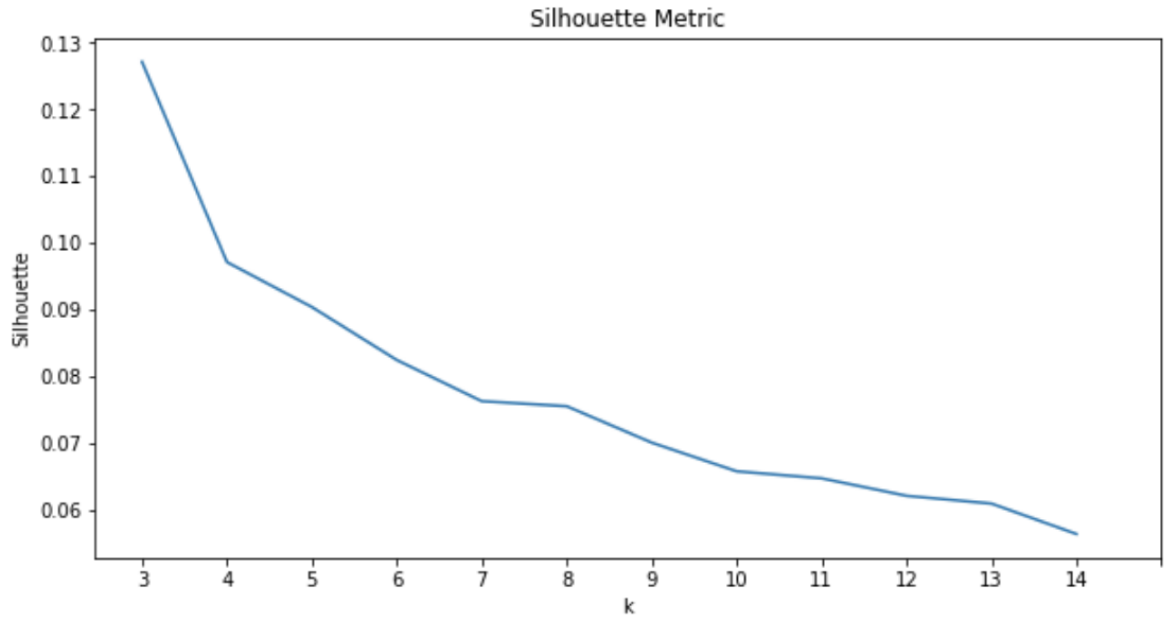


Figura 5.44: Silhouette dataset *White with noise* al variare di K

I gruppi risultanti si rivelano separati e distinguibili (Figura 5.45) ma la composizione dei cluster non riflette la reale distribuzione dei cicli di produzione. Inoltre, in questo caso i cluster formati sono sbilanciati: si sono ottenuti tre cluster, ma questi hanno dimensioni diverse tra loro. Infatti, i cluster 0 e 1 contengono circa 2700 cicli di produzione ciascuno, mentre il cluster 2 ne contiene solamente 300. Figura 5.46. Questo è confermato anche dall'Adjusted Rand score che, oltre ad essere prossimo allo zero ad indicare che l'etichettatura dei dati è stata casuale, è anche negativo.

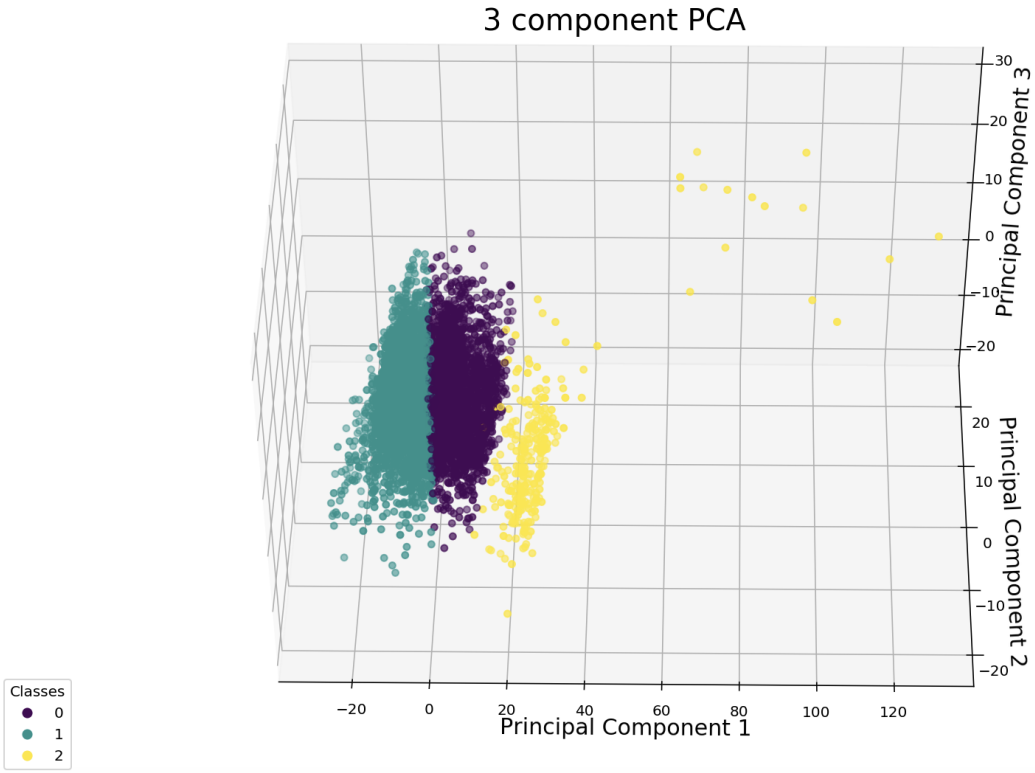


Figura 5.45: Rappresentazione PCA dataset *White with noise* K-Means con K = 3

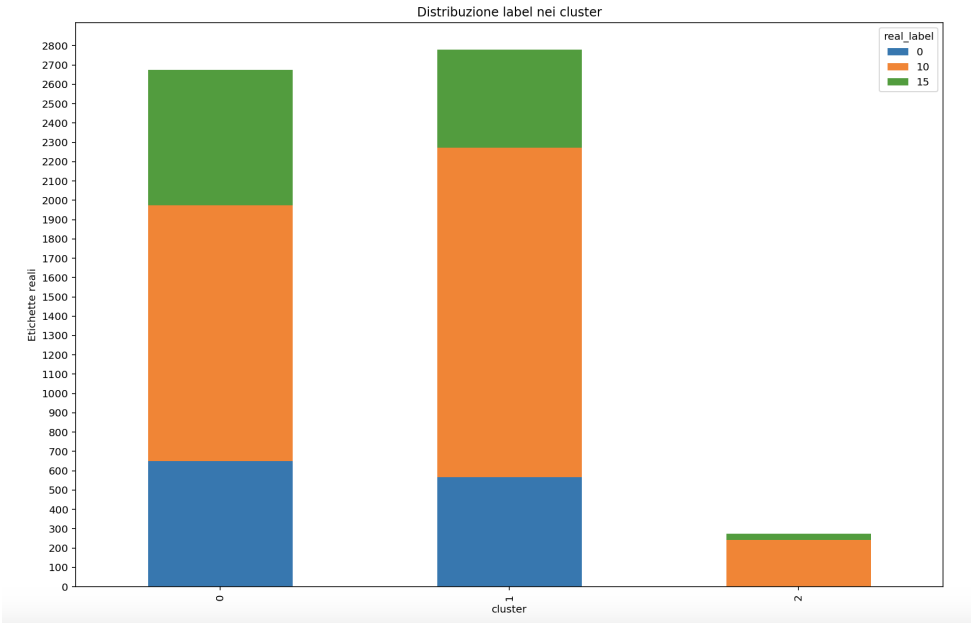


Figura 5.46: Dataset *White with noise*. Distribuzione delle etichette nei cluster, K-Means con $K = 3$

Se si utilizzasse come criterio l'Elbow Method per selezionare il parametro K ideale come input del K-Means si individuerebbe il gomito della curva in corrispodenza di $K = 5$. Figura 5.47.

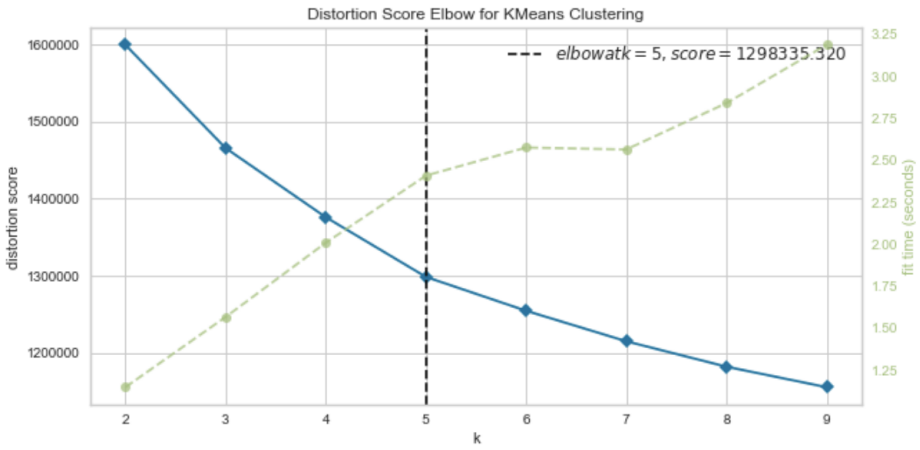


Figura 5.47: Elbow Method dataset *White with noise*

Tuttavia, anche variando K, si ottengono dei cinque cluster separati che non riflettono la suddivisione reale dei cicli di produzione. Anche in questo caso, i gruppi sono sbilanciati nella dimensione: il cluster 3 contiene una quantità decisamente inferiore di cicli di produzione rispetto agli altri gruppi. Figura 5.48.

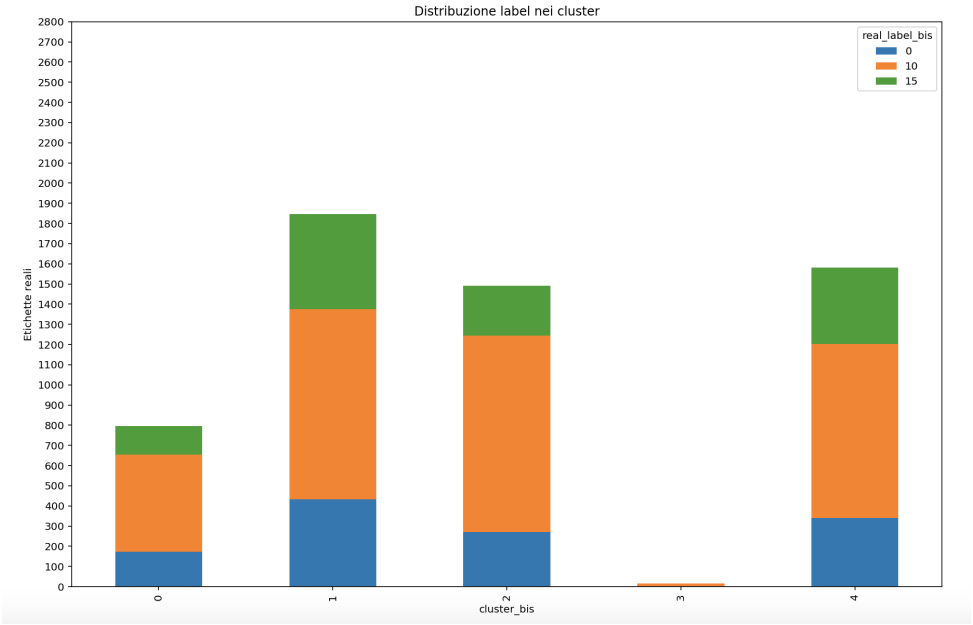


Figura 5.48: Dataset *White with noise*. Distribuzione delle etichette nei cluster, K-Means con $K = 5$

5.5.2 DBSCAN

Gray with noise

Per selezionare i parametri di input del DBSCAN *Eps* e *MinPts* è stato rappresentato il grafico del *k_dist* per valori di *MinPts* compresi tra 3 e 10. Dopo alcuni tentativi si è deciso di selezionare come parametro $MinPts = 5$ in quanto per questo valore il *k_dist* rimane immutato. Dal grafico in Figura 5.49 si può ricavare il parametro *Eps* in corrispondenza del ginocchio della curva che si attesta intorno a 15.

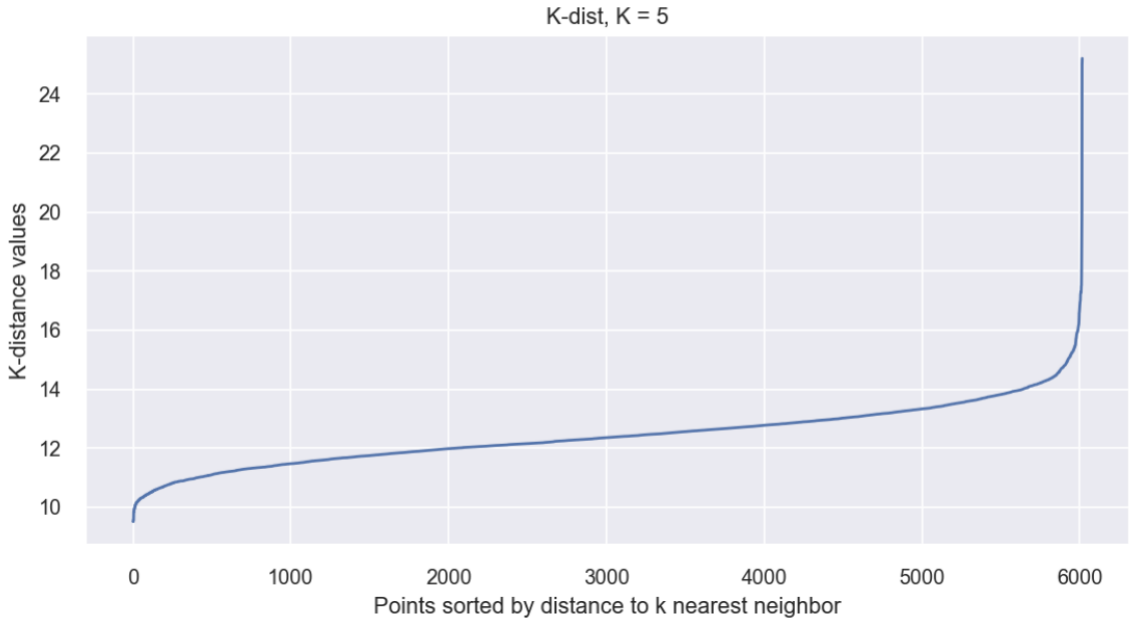


Figura 5.49: K-dist dataset *Gray with noise* con k = 5

Con questi parametri si identificano quattro cluster, compreso il cluster -1 contenente i valori identificati come outliers. Tuttavia, si può osservare che la generazione dei cluster con questa configurazione non è stata particolarmente efficiente: infatti, con i parametri individuati la quasi totalità dei punti è stata assegnata al cluster 0 e solamente una decina di punti complessivamente occupano i cluster 1 e 2. Dalla rappresentazione in PCA, infatti, si nota una massa di punti in blu corrispondenti al cluster 0, mentre gli altri sono quasi invisibili. Di conseguenza, il cluster 0 contiene tutti i punti del dataset, senza distinguere le etichette reali di appartenenza dei cicli di produzione; il valore dell'Adjusted Rand Score riflette anche in questo caso l'assegnazione casuale dei punti ai cluster attestandosi intorno allo zero.

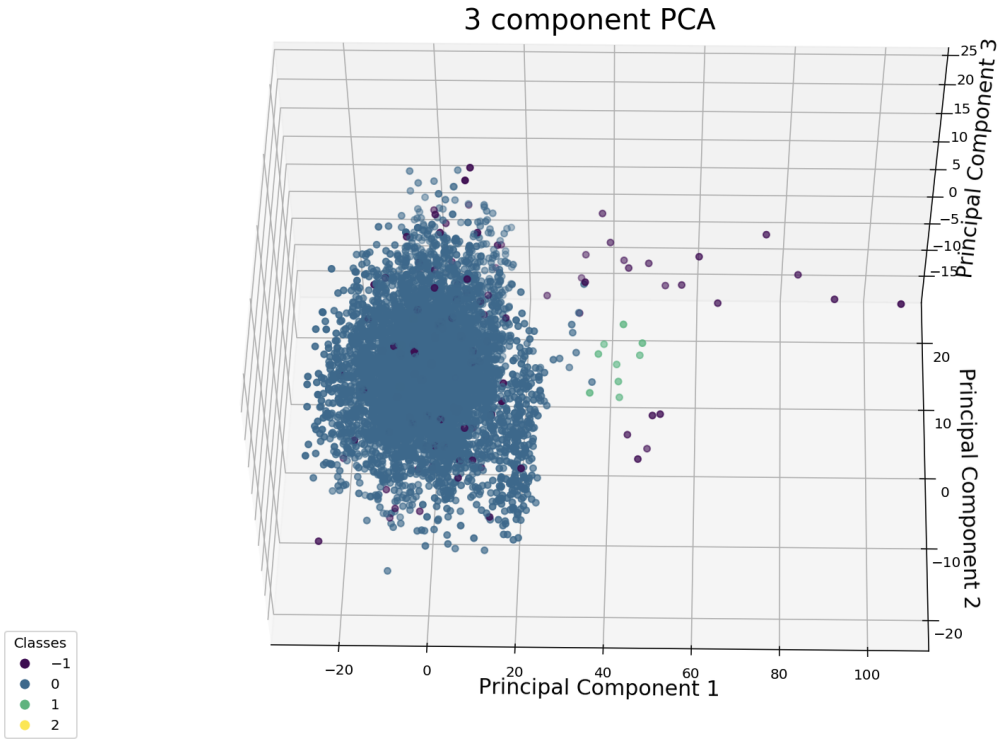


Figura 5.50: Rappresentazione PCA dataset *Gray with noise*

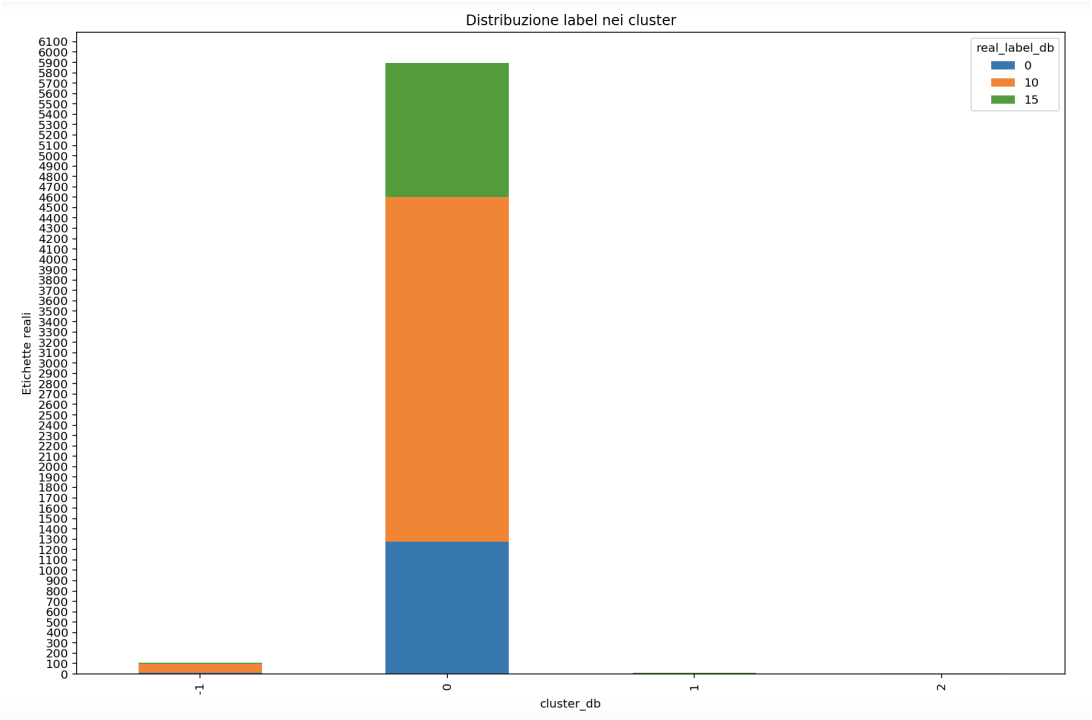


Figura 5.51: Dataset *Gray with noise*. Distribuzione delle etichette nei cluster DBSCAN

White with noise

Analogamente al dataset *Gray with noise*, è stato rappresentato il *k-dist* al variare di *MinPts*. Per $Minpts = 8$ è stato identificato come *Eps* ottimale 15.5.

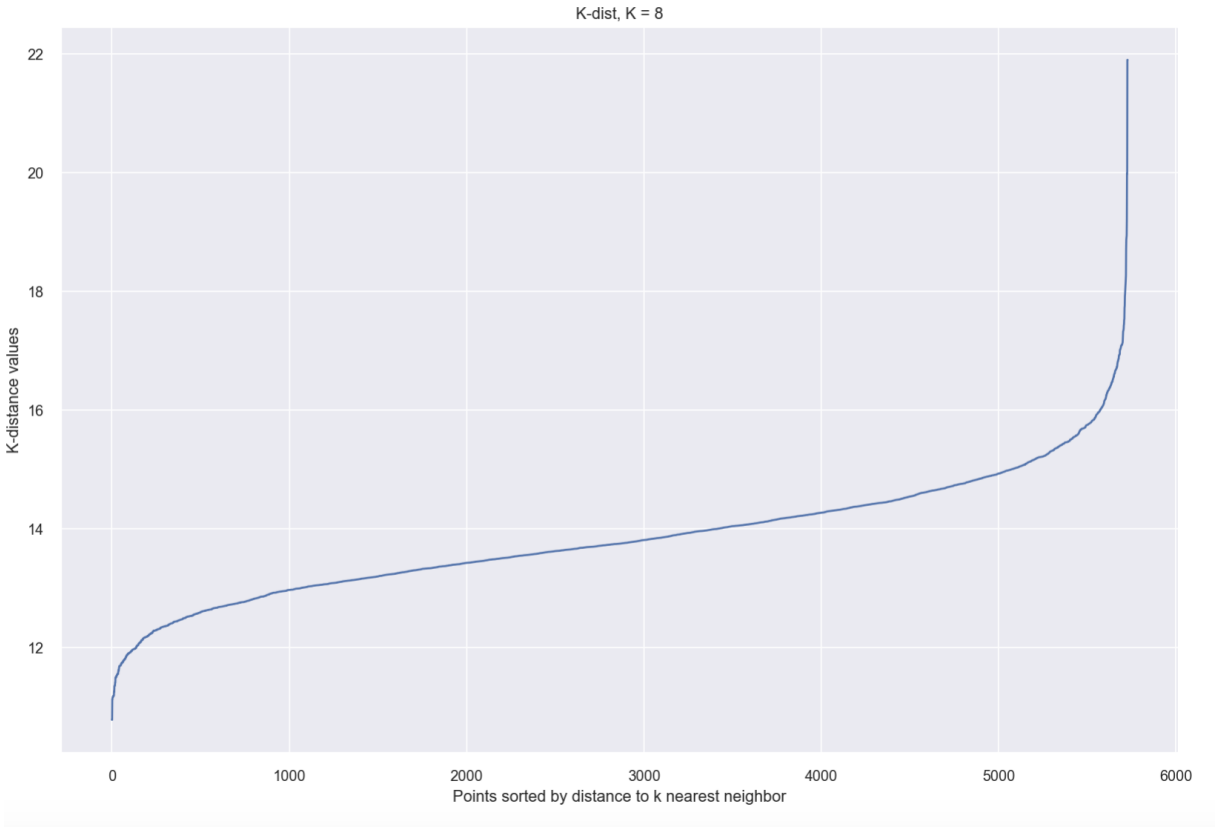


Figura 5.52: K-dist dataset *White with noise* con k = 8

Per questi parametri di input, tuttavia, si verifica lo stesso comportamento osservato per il dataset *Gray with noise*, ovvero vengono creati quattro cluster (compreso il cluster -1 contenente *noise point*), ma la quasi totalità dei punti è assegnata al cluster 0, mentre i cluster 1 e 2 contengono solamente poche decine di cicli di produzione. Anche in questo caso, l'Adjusted Rand Score risulta prossimo allo zero.

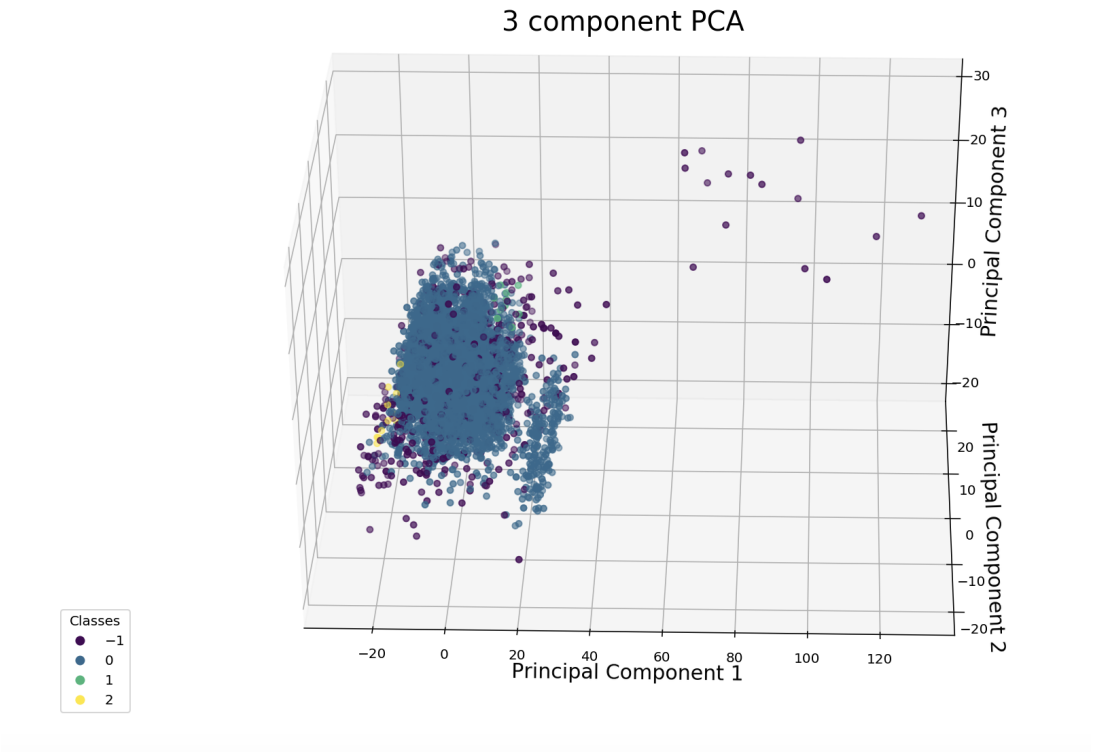


Figura 5.53: Rappresentazione PCA dataset *White with noise* etichette DBSCAN

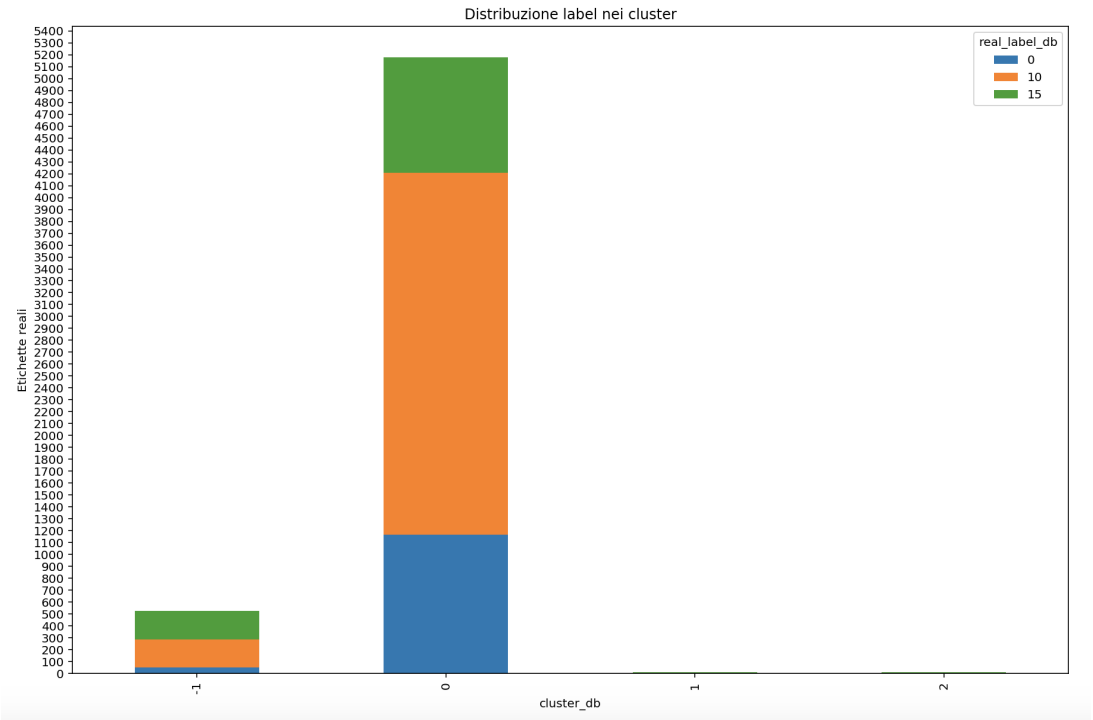


Figura 5.54: Dataset *White with noise*. Distribuzione delle etichette nei cluster DBSCAN

5.5.3 Agglomerative Hierarchical Clustering

Gray with noise

Come nel caso del dataset *Gray*, è stata utilizzata come misura della distanza quella Euclidea e come linkage *ward*. Per selezionare parametro *n_cluster* ottimale è stato dapprima osservato il dendrogramma in Figura 5.55; si è ipotizzato di effettuare un taglio che permettesse di ottenere tre cluster e, per conferma, è stato osservato l'andamento della Silhouette al variare di *n_cluster*. Poichè la Silhouette è risultata massima con un valore pari a 0.1298 in corrispondenza di $n\_cluster = 3$ e decrescente per valori di $n\_cluster > 3$, è stato applicato l'Agglomerative Hierarchical Clustestering con i parametri così identificati. Osservando la rappresentazione in PCA dei gruppi individuati in Figura 5.56, si notano due gruppi di dati coesi e numerosi, mentre il terzo cluster risulta di dimensione decisamente inferiore ai precedenti.

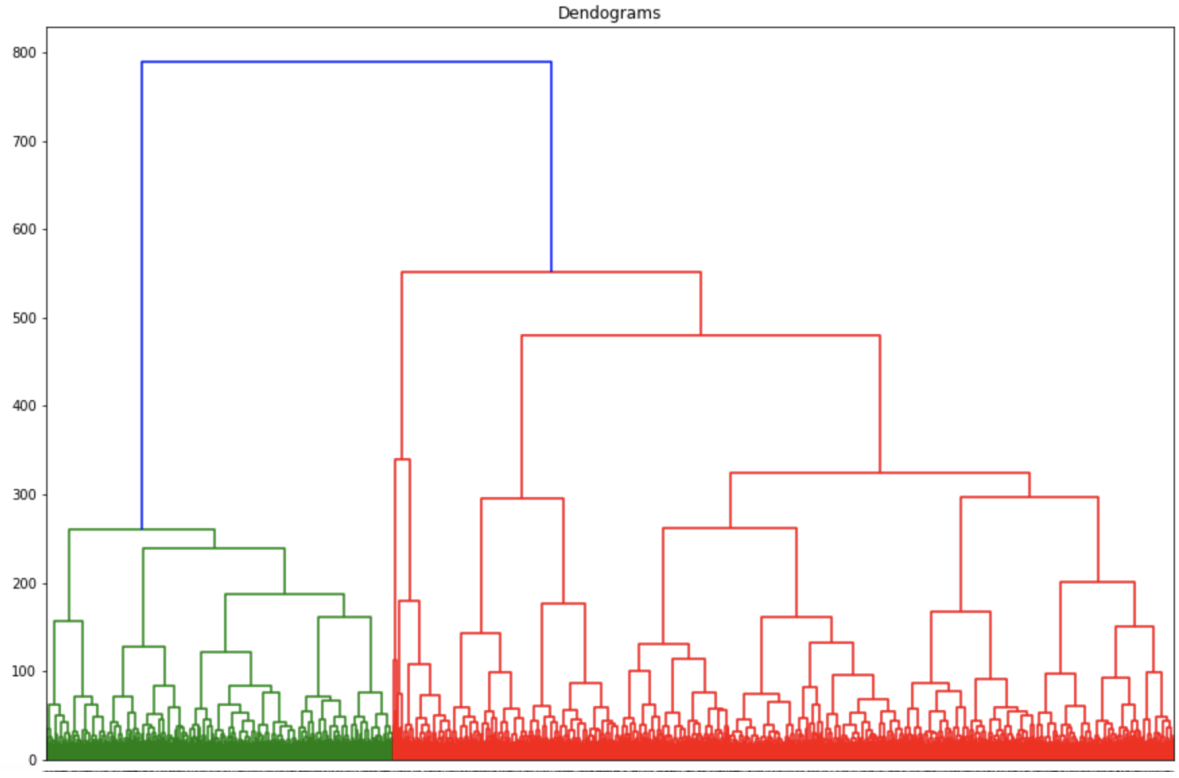


Figura 5.55: Dendrogramma dataset *Gray with noise*

Dunque, è stato verificato se i dati contenuti nei tre cluster effettivamente rispettano le etichette di appartenenza dei cicli. Tuttavia, come si vede dallo Staked Bar in Figura 5.57, anche in questo caso i cluster sono formati da percentuali variabili di etichette diverse e non si nota un cluster in cui prevale una classe piuttosto che un'altra, ad eccezione del cluster 2 costituito per lo più da cicli etichettati con 10.

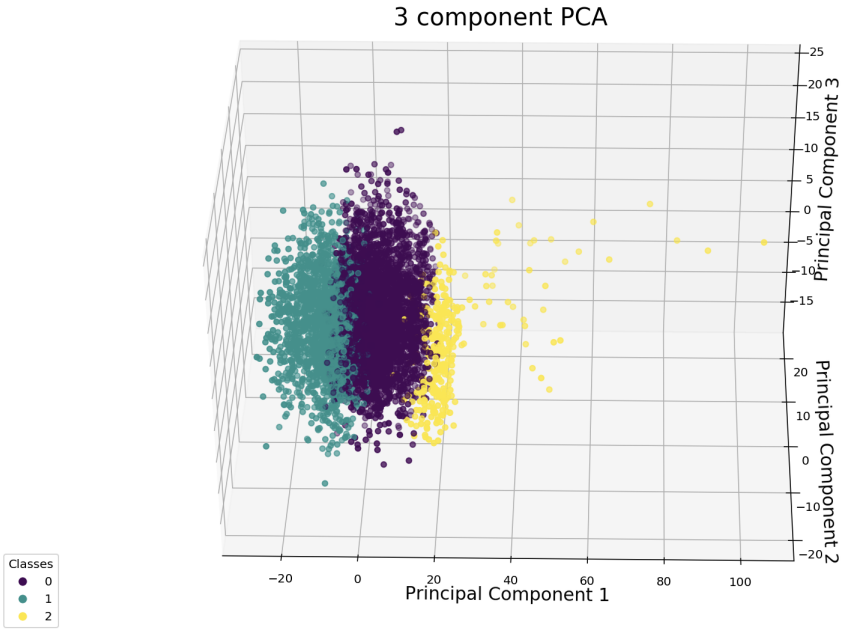


Figura 5.56: Rappresentazione PCA dataset *Gray with noise*, Agglomerative Hierarchical Clustering con $n\_cluster = 3$

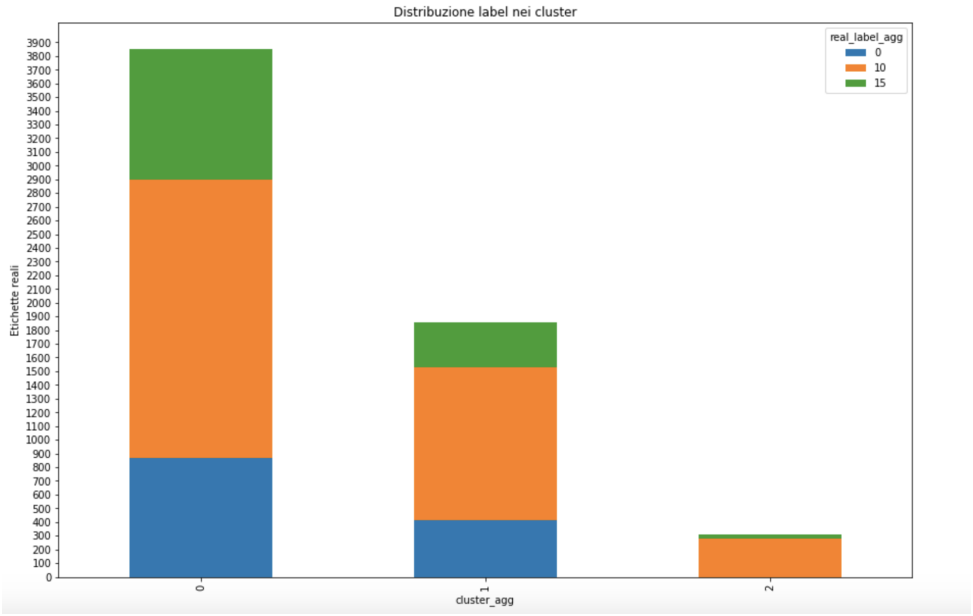


Figura 5.57: Dataset *Gray with noise*. Distribuzione delle etichette nei cluster, Agglomerative Hierarchical Clustering

White with noise

Seguendo gli stessi passi già ripercorsi per il dataset precedente, si osserva il dendrogramma per il dataset *White with noise* in Figura 5.58 e si è ipotizzato di tracciare un taglio in

modo da ottenere tre cluster. Inoltre, è stato osservato l'andamento della Silhouette al variare del numero di cluster, che è risultata decrescente per $n\_cluster > 3$. Poichè un peggioramento della Silhouette indica una degradazione della coesione/separazione dei gruppi, è stato eseguito l'algoritmo con il numero di cluster così individuato e tutti gli altri parametri pari al valore di default. Anche in *White with noise*, sono emersi tre cluster sproporzionati nella dimensione in quanto si notano due gruppi numerosi e un terzo più contenuto con una minore quantità di cicli di produzione. Osservando la composizione dei cluster con l'aiuto delle etichette reali dei cicli di produzione si può notare che il cluster 2, di dimensione decisamente inferiore rispetto agli altri due, è formato principalmente da cicli di produzione con etichetta 10 e solo pochi punti con etichetta 15. Gli altri due cluster, invece, contengono punti di tutte e tre le etichette. Figura 5.60

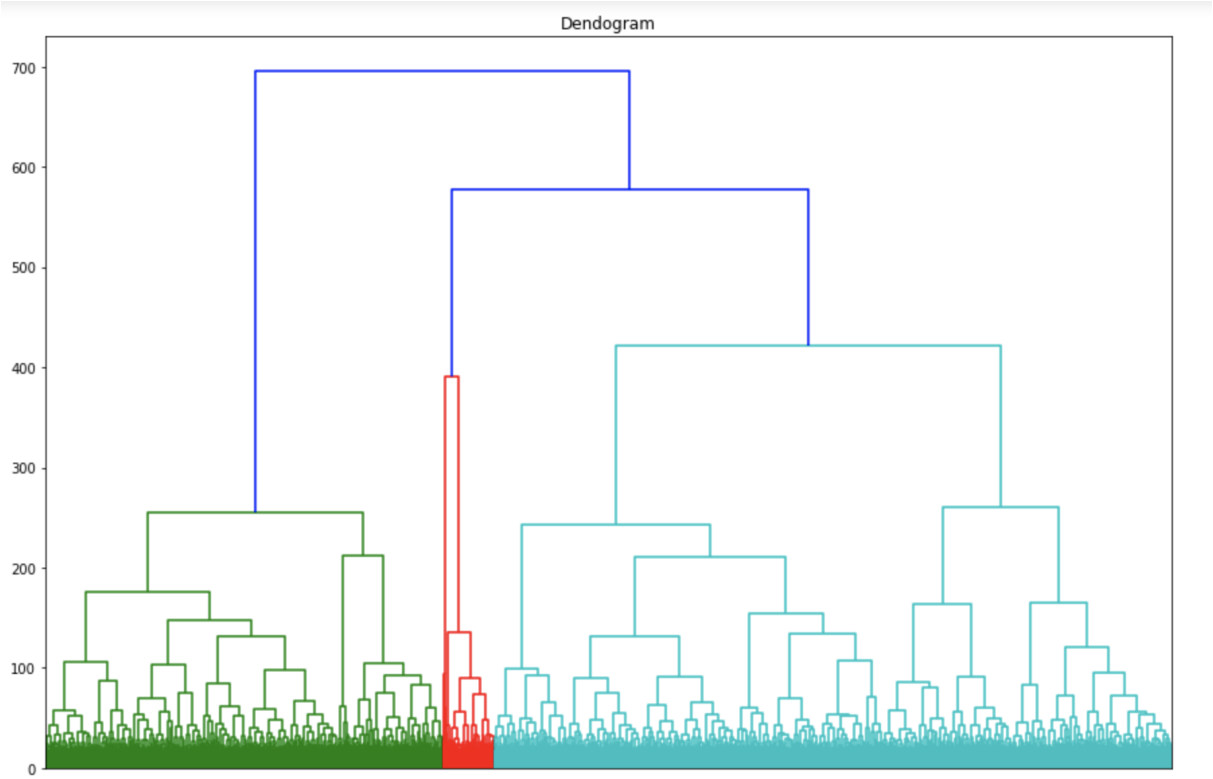


Figura 5.58: Dendrogramma dataset *White with noise*

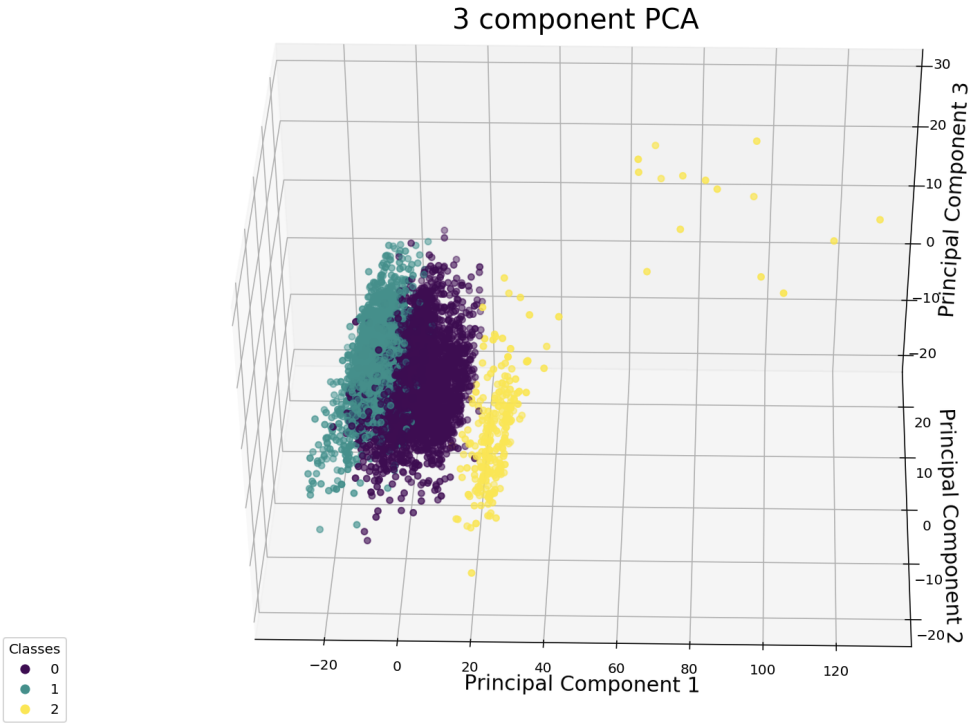


Figura 5.59: Rappresentazione PCA dataset *White with noise*, Agglomerative Hierarchical Clustering $n\_cluster = 3$

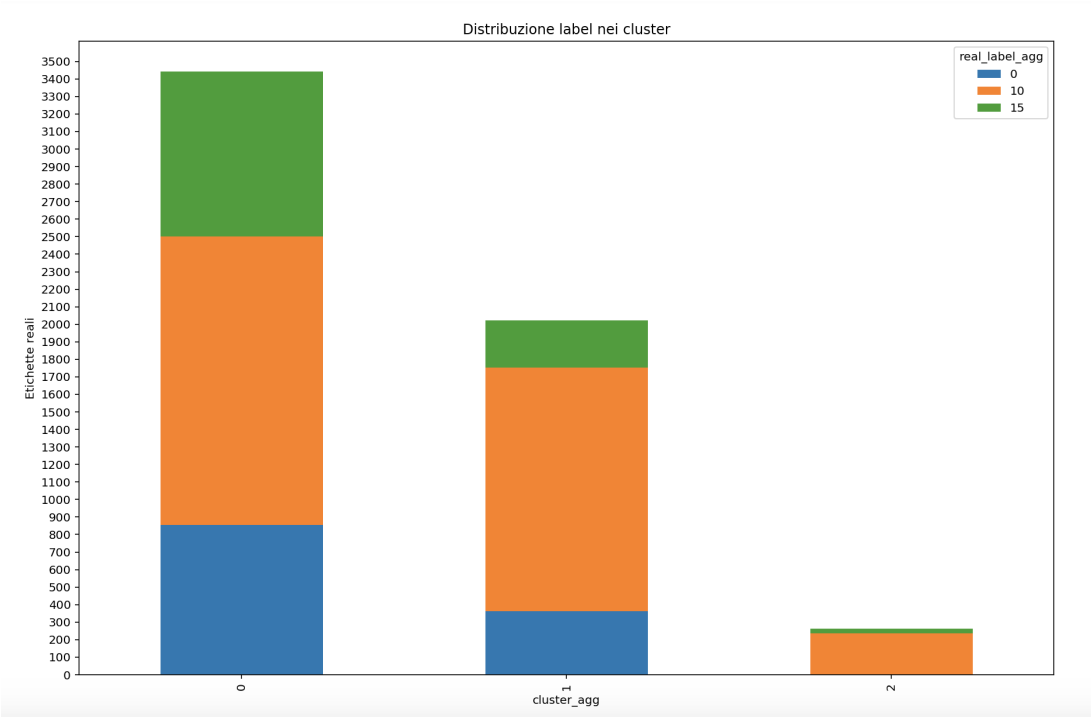


Figura 5.60: Dataset *White with noise*. Distribuzione delle etichette nei cluster, Agglomerative Hierarchical Clustering

5.5.4 Confronto tra algoritmi

Concludendo, le partizioni ottenute con K-Means, DBSCAN e Agglomerative Hierarchical Clustering non sono risultate efficienti allo scopo di suddividere correttamente i cicli di produzione nei dataset con aggiunta di rumore. Di seguito, nelle Tabelle 5.7 e 5.8 si riportano per ciascun algoritmo i parametri utilizzati come input e i valori dell'Adjusted Rand Score.

Algoritmo	Parametri	Adjusted Rand Score
K-Means	K = 3	0.0095
DBSCAN	$Eps = 15$ $MinPts = 5$	-0.0083
Agglomerative Hierarchical Clustering	n_cluster = 3	0.0027

Tabella 5.7: Riepilogo per il dataset *Gray with noise*

Algoritmo	Parametri	Adjusted Rand Score
K-Means	K = 3	-0.0056
DBSCAN	$Eps = 15.5$ $MinPts = 8$	0.0358
Agglomerative Hierarchical Clustering	n_cluster = 3	-0.0222

Tabella 5.8: Riepilogo per il dataset *White with noise*

Capitolo 6

Conclusioni e sviluppi futuri

L'obiettivo del presente lavoro è stato quello di progettare e sviluppare una metodologia semi-supervisionata allo scopo di caratterizzare i cicli di produzione nel contesto dell'Industria 4.0 in base al livello di tensione della cinghia di trasmissione. In tal modo è possibile offrire un supporto agli esperti di dominio in fase di manutenzione predittiva, in particolare durante l'attività di etichettatura dei cicli di produzione, consentendo di svolgere tale operazione automaticamente e non più manualmente. Per estrarre delle conoscenze utili da applicare al contesto sono stati utilizzati tre algoritmi di clustering, K-Means, DBSCAN, Agglomerative Hierarchical Clustering. Prima di costruire i modelli, è stata analizzata la distribuzione dei dati per ricavare i parametri ottimali da fornire come input a ciascun algoritmo. Infine, con l'aiuto delle etichette reali dei cicli di produzione, le partizioni ottenute sono state validate.

Il *framework* è stato applicato a due dataset contenenti dati circa la corrente assorbita dal motore di un braccio robotico. Il primo è stato fornito da un'azienda italiana leader nel settore metalmeccanico che, tuttavia, a seguito di un confronto con gli esperti di dominio, è risultato essere stato realizzato ad hoc modificando manualmente il livello di tensione della cinghia; il secondo, invece, è stato costruito sul dataset iniziale introducendo artificialmente una percentuale di rumore per testare la metodologia su un caso più realistico.

A seguito delle analisi, è emerso che nel dataset fornito inizialmente la metodologia è stata particolarmente efficace nel distinguere cicli di produzione con livelli di tensione della cinghia diversi: ciò è accaduto perchè le tre tipologie di cicli sono caratterizzate da valori di corrente decisamente differenti tra loro e, dunque, i gruppi sono risultati ben separati.

Questo comportamento, invece, non si è verificato nel dataset riprodotto artificialmente. In tal caso, la metodologia proposta non è stata in grado di individuare delle partizioni che rispettassero la reale suddivisione dei cicli di produzione. Ciò è accaduto perchè la procedura utilizzata per introdurre il rumore ha reso i cicli di produzione con diversi livelli di tensione della cinghia meno distinguibili rispetto al caso originario. Gli algoritmi di clustering sono configurati analizzando la distribuzione dei dati della corrente e privilegiando

la formazione di gruppi di dati coesi e separati; poichè nel dataset artificiale le tre classi sono coese ma non più ben separate tra loro, nessuna tra le tecniche utilizzate ha restituito delle partizioni che rispecchiassero la reale etichettatura dei cicli di produzione.

Si conclude che se i cicli di produzione nel caso reale sono abbastanza distinguibili, questa metodologia può essere applicata ed offrire un valido supporto per l'attività di manutenzione predittiva; infatti, è possibile individuare quando i cicli di produzione presentano valori di corrente tali da indicare un mal tensionamento della cinghia.

Futuri sviluppi possono riguardare l'applicazione della metodologia presentata su dei casi di studio realistici, avendo preventivamente raccolto a sufficienza i dati della corrente per avere informazioni sui valori associati ai diversi livelli di tensione. Inoltre, l'approccio può essere applicato sui risultati di un classificatore one-class utilizzato in un contesto di *outlier detection* e rilevamento del *concept drift*: in questo modo è possibile scoprire nuove tipologie di cicli di produzione, arricchendo così la conoscenza degli esperti di dominio sul contesto di applicazione.

Bibliografia

[1] Dr. Johannes Helbig Prof. Dr. Henning Kagermann Prof. Dr. Wolfgang Wahlster. «Recommendations for implementing the strategic initiative INDUSTRIE 4.0». In: (apr. 2013) (cit. a p. 3).

[2] *Rivoluzione industriale*. URL: https://it.wikipedia.org/wiki/Rivoluzione_industriale (cit. a p. 3).

[3] Mario Hermann, Tobias Pentek e Boris Otto. «Design Principles for Industrie 4.0 Scenarios: A Literature Review». In: (gen. 2015). DOI: 10.13140/RG.2.2.29269.22248 (cit. a p. 6).

[4] Jay Lee, Behrad Bagheri e Hung-An Kao. «A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems». In: *SME Manufacturing Letters* 3 (dic. 2014). DOI: 10.1016/j.mfglet.2014.12.001 (cit. alle pp. 6–8).

[5] Mohd Bahrin, Fauzi Othman, Nor Azli e Muhamad Talib. «Industry 4.0: A review on industrial automation and robotic». In: *Jurnal Teknologi* 78 (giu. 2016). DOI: 10.11113/jt.v78.9285 (cit. alle pp. 9, 11).

[6] Gallinaro Silvana. «Dai modelli lineari di business alla piattaforma di progettazione e manifattura. Gli effetti delle tecnologie additive sulla logica di creazione del valore delle imprese manifatturiere». In: (feb. 2019) (cit. a p. 10).

[7] Leonello Trivelli Gloria Cervelli Simona Pira. «Industria senza Slogan». In: (2017) (cit. alle pp. 10, 11).

[8] Arlapp. URL: www.airlapp.com (cit. a p. 10).

[9] K. Chukalov. «Horizontal and vertical integration, as a requirement for cyber-phisical system in the context of industry 4.0». In: () (cit. a p. 11).

[10] Bill Hallaq Hugh Boyes. «The industrial internet of things (IIoT): An analysis framework». In: (2018) (cit. a p. 12).

[11] D. Floyer. Defining e Sizing the Industrian Internet. URL: http://wikibon.org/wiki/v/Defining_and_Sizing_the_Industrial_Internet (cit. a p. 12).

[12] Industrial Internet Insights Report for 2015. URL: https://www.ge.com/digital/sites/default/files/industrial-internet-insights-report.pdf (cit. a p. 12).

[13] Dirk Schaefer Lane Thamesa. «Software-Defined Cloud Manufacturing for Industry 4.0». In: (2016) (cit. a p. 12).

[14] Azure Microsoft. URL: https://azure.microsoft.com/it-it/overview/what-are-private-public-hybrid-clouds/ (cit. a p. 12).

[15] Alp Ustundag e Emre Cevikcan. *Industry 4.0: Managing The Digital Transformation*. Gen. 2018. ISBN: 978-3-319-57869-9. DOI: 10.1007/978-3-319-57870-5 (cit. alle pp. 14, 15).

[16] Okyay Kaynak e Shen Yin. «Big Data for Modern Industry: Challenges and Trends [Point of View]». In: *Proceedings of the IEEE* 103 (feb. 2015), pp. 143–146. DOI: 10.1109/JPROC.2015.2388958 (cit. a p. 16).

[17] M Schroeck, R Shockley, J Smart, Dolores Romero Morales e P Tufano. «Analytics: the real-world use of big data: How innovative enterprises extract value from uncertain data, Executive Report». In: (gen. 2012) (cit. a p. 16).

[18] Deloitte. URL: www.deloitte.com (cit. a p. 16).

[19] i-scoop. *Industry 4.0: the fourth industrial revolution – guide to Industrie 4.0*. URL: i-scoop.eu/industry-4-0 (cit. alle pp. 18, 25).

[20] Massimo Zanardini Luca Franzoni. *Industria 4.0 in Italia e nel mondo. I Governi rilanciano il manifatturiero*. URL: www.iqconsulting.it (cit. a p. 19).

[21] World Economic Forum. *www.weforum.org* (cit. a p. 19).

[22] Economy Up. URL: https://www.economyup.it/innovazione/cos-e-l-industria-40-e-perche-e-importante-saperla-affrontare/ (cit. a p. 21).

[23] Ministero dello Sviluppo Economico. URL: https://www.sviluppoeconomico.gov.it/images/stories/documenti/PIANO-NAZIONALE-INDUSTRIA-40_ITA.pdf (cit. alle pp. 21, 23, 24).

[24] Ministero dello Sviluppo Economico. *La diffusione delle imprese 4.0 e le politiche: evidenze 2017*. 2018 (cit. a p. 23).

[25] Ministero dello Sviluppo Economico. «Transizione 4.0: una nuova politiche industriale». In: () (cit. a p. 24).

[26] «Startup, pochi nuovi investimenti (-38%). Il 2017 è un anno di transizione». In: (2018) (cit. a p. 24).

[27] Camera dei Deputati. «Industria 4.0». In: (2020) (cit. a p. 25).

[28] Laura Swanson. «Linking maintenance strategies to performance». In: (2001) (cit. a p. 28).

[29] Michele Zubani. 2018. URL: `https://www.toolsforsmartminds.com/` (cit. alle pp. 30, 31).

[30] Search Data Management. URL: `https://searchdatamanagement.techtarget.com/definition/data-analytics` (cit. alle pp. 33, 34).

[31] zerounoweb. URL: `https://www.zerounoweb.it/analytics/big-data/come-fare-big-data-analysis-e-ottenere-valore-per-le-aziende/` (cit. a p. 33).

[32] zerounoweb. 2019. URL: `https://www.zerounoweb.it/analytics/e-lora-della-data-science-strategica-e-strutturata/` (cit. a p. 33).

[33] Kumar Tan Steinbach. *Introduction to Data Mining* (cit. alle pp. 35, 36, 44–46, 48, 55–61, 63, 64).

[34] Tom M. Mitchel. «Machine Learning». In: (1997) (cit. a p. 36).

[35] URL: `https://www.ai4business.it/intelligenza-artificiale/machine-learning/machine-learning-cosa-e-applicazioni/` (cit. a p. 37).

[36] Jiawei Han e Micheline Kamber. *Data mining : concepts and techniques*. San Francisco [u.a.]: Kaufmann, 2005 (cit. a p. 39).

[37] Lior Rokach Oded Maimon. *Data Mining and Knowledge Discovery Handbook* (cit. a p. 49).

[38] Nayer Wanas, Dina Said, Nabila Khodeir, Magda Fayek e Ahmed Gaffer. «Detection and handling of different types of concept drift in news recommendation systems». In: (feb. 2019) (cit. a p. 50).

[39] Indre Zliobaite. «Learning under Concept Drift: an Overview». In: (2010) (cit. a p. 50).

[40] Fei Tony Liu, Kai Ming Ting e Zhi-Hua Zhou. «Isolation-based anomaly detection». English. In: *ACM Transactions on Knowledge Discovery from Data* 6.1 (2012), pp. 1–39. ISSN: 1556-4681. DOI: `10.1145/2133360.2133363` (cit. a p. 51).

[41] *Clustering*. URL: `https://scikit-learn.org/stable/modules/clustering.html` (cit. a p. 64).

[42] Wes McKinney et al. «Data structures for statistical computing in python». In: *Proceedings of the 9th Python in Science Conference*. Vol. 445. Austin, TX. 2010, pp. 51–56 (cit. a p. 65).

[43] J. D. Hunter. «Matplotlib: A 2D graphics environment». In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: `10.1109/MCSE.2007.55` (cit. a p. 65).

[44] F. Pedregosa et al. «Scikit-learn: Machine Learning in Python». In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. a p. 65).

[45] *Cinghie di trasmissione: guida rapida dalle tipologie al tensionamento*. URL: `https://www.ilprogettistaindustriale.it/cinghie-di-trasmissione/` (cit. a p. 66).

[46] Francesco Ventura Stefano Proto Daniele Apiletti Tania Cerquitelli Simone Panicucci Elena Baralis Enrico Macii Alberto Macii. «A new unsupervised predictive-model self-assessment approach that SCALEs». In: (2019) (cit. alle pp. 67, 78).