POLITECNICO DI TORINO

Corso di Laurea Magistrale In

Ingegneria Gestionale

Tesi di Laurea Magistrale

Tecniche di analisi di dati a supporto della manutenzione predittiva. Caso di studio: Tensionamento della cinghia di un robot industriale

Relatrice

Prof. Tania Cerquitelli

Co-Relatore

Dott. Francesco Ventura

Candidato

Luca Tamassia

Marzo 2019

Sommario

1. Introduzione

Negli ultimi dieci anni la tecnologia ha compiuto enormi passi avanti in molti campi, dalla robotica, alle telecomunicazioni, all'informatica. Questa incredibile accelerazione ha portato alla creazione di interi nuovi mercati, aggiunto valore a quelli esistenti e sviluppato modelli di business fino ad ora sconosciuti.

Uno sviluppo repentino che è stato fondamentale per entrare nel periodo che viene universalmente considerato come la quarta rivoluzione industriale o rivoluzione digitale, (fig. 1.1) rappresentata dall'introduzione di tecnologie digitali in settori prettamente manifatturieri, come l'automotive, l'aereospaziale ed il metalmeccanico.

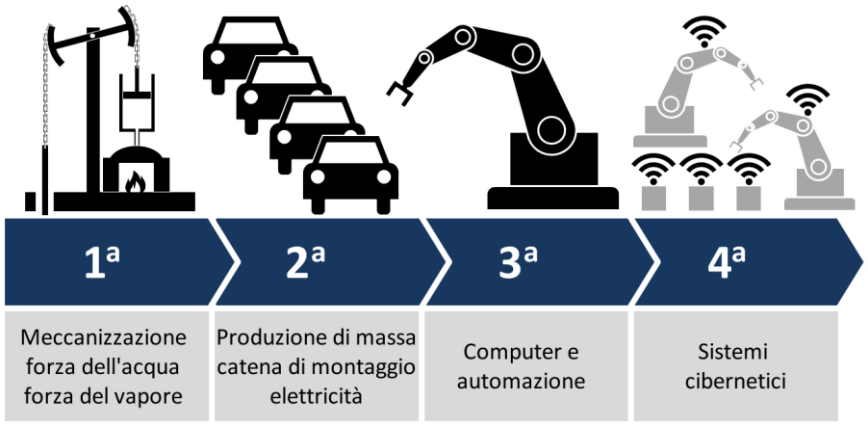


Fig. 1.1: Rivoluzioni Industriali (Wikipedia)

Si è potuti arrivare ad una tale rivoluzione grazie molteplici tecnologie abilitanti, che sono a loro volta variegate, come il cloud computing o i Big Data Analytics, e che si inseriscono nei vari contesti produttivi in maniera unica a seconda dell'applicazione desiderata.

In particolare, questa tesi vuole comprendere il ruolo dell'analisi di dati e della machine learning in contesti industriali classici come quello metalmeccanico e più nello specifico, nella manutenzione predittiva di parti meccaniche di robot. Grazie alla collaborazione con un'importante azienda di robotica è stato possibile analizzare il comportamento di una cinghia di trasmissione di uno dei loro bracci robotici, con lo scopo di predire il degrado che palesa la necessatà di un intervento di manutenzione. Quindi, l'elaborato s'inserisce nel filone di applicazione di una tecnologia digitale all'ambito industriale, tipico dell'Industria 4.0.

Questa tesi si articola in una descrizione dell'Industria 4.0, cos'è, come si sviluppa, quali sono gli elementi principali che la costituiscono, le sfide, le problematiche e gli impatti economici che sta avendo nel mondo, con un piccolo focus sulla situazione italiana. Segue una parte sulle tecnologie utilizzate in contesti produttivi 4.0, inserendo alcuni casi rilevanti di utilizzo/studio, un'introduzione sulla natura e gli obiettivi dell'analisi dei Big Data, con uno sguardo al mondo della quarta rivoluzione industriale, un'analisi dello stato dell'arte e delle conoscenze base per approcciarsi al mondo dei Big Data, gli step da seguire, le problematiche più comuni, le tecniche più usate, con particolare attenzione alla classificazione, che sarà il metodo impiegato per analizzare i dati utilizzati per questa tesi. Successivamente si riporta uno spaccato sul machine learning, il suo legame con il data mining e come viene suddiviso in varie categorie a seconda della sua applicazione, alcuni esempi, ed infine un breve riassunto delle problematiche più frequenti. Conclude la parte bibliografica, l'introduzione al caso di studio, con una panoramica sulla Manutenzione Predittiva e al tema del Belt Tensioning, con introduzione al problema del braccio robot.

Entrando nel merito, si passa alla disamina dei dati veri e propri, con una prima parte incentrata sull'analisi esplorativa dei dati, su come sono strutturati e sulle loro caratteristiche principali. A seguire un'introduzione al percorso seguito per arrivare alle conclusioni: una prima fase di pre-processing dei dati e poi la descrizione dettagliata dei due filoni di analisi condotti per questo progetto di tesi, uno riguardante un modello generale di predizione dei dati, l'altro riguardante un modello evolutivo che prenda in esame solo porzioni del dataset.

Il modello generale esplora le possibili porzioni di dataset necessarie per addestrare un modello che abbia valori di richiamo e precisione sufficientemente buoni per validarlo; in pratica cerca il numero minimo di dati necessari per creare un modello.

Il modello evolutivo invece, si prefigge come obiettivo di creare un modello che sia in grado di individuare un degrado delle prestazioni della macchina, dividendo il dataset iniziale in classi per l'addestramento dell'algoritmo e gruppi di test.

Per concludere si sono aggiunte delle considerazioni finali sui risultati ottenuti e si sono proposti spunti per future ricerche ed eventuali ampliamenti dell'argomento trattato.

2. Industria 4.0

L'Industria 4.0 nasce come termine in Germania, durante la Fiera di Hannover del 2011, dove alcuni industriali tedeschi suggerirono al governo alcune misure da intraprendere per stimolare l'industria manufatturiera nazionale; verrà poi concretizzato un piano, come il nome appunto di Industria 4.0 alla fine del 2013, con l'obiettivo di ammodernare il sistema produttivo tedesco tramite investimenti in infrastrutture, istruzione, energia e ricerca e sviluppo, seguito poi da molti altri paesi come gli Stati Uniti e l'Italia. (Maci, 2018).

La definizione più tecnica di questo nuovo termine può essere intesa come un aumento dell'automazione industriale, che vede l'integrazione di nuove tecnologie atte a migliorare le condizioni di lavoro, ad aumentare le performance di produttività e qualità dei prodotti e creare nuovi modelli di business (Wikipedia, 2018). Può, però, essere riassunta anche nel passaggio dal modo di produrre classico, a cui si è stati abituati fino ad oggi, ad uno digitale, con le fabbriche, intese, sia come linee produttive che come singole macchine, che si connettono fra loro e con le altre unità di azienda. Partendo dalla rivoluzione informatica degli anni '70, la 4.0, ne prende le basi (i computer, internet, la digitalizzazione) le esalta, le implementa nelle forme più creative per risolvere i più disparati problemi che hanno sempre afflitto la produzione industriale (Marr, 2018).

L'importanza di questo nuovo fenomeno globale non è solo dovuta al miglioramento del settore manifatturiero, perché non bisogna dimenticare l'effetto "disruptive" che ha avuto, ha e avrà sui posti di lavoro dell'intero mercato, dando vita ad un vero e proprio sconvolgimento della società, non solo economico o di performance; come è stato sottolineato dalla ricerca **"**The Future of the Jobs**",** divulgata al World Economic Forum 2016, si creeranno 2 milioni di nuovi posti di lavoro, ma contemporaneamente se ne perderanno quasi 7, con un saldo fortemente negativo per l'occupazione; cambieranno quindi le competenze richieste dal mercato del lavoro e per usufruire efficacemente di queste evoluzioni, occorre che ogni paese prepari il terreno alle imprese e le incentivi ad accogliere tutte le novità (Bucceri, 2017).

Come è stato già in parte anticipato, le novità portate dall'Industria 4.0 praticamente risiedono nella connessione delle macchine, che diventano "intelligenti", nella raccolta e condivisione dei dati che loro stesse producono o che provengono dall'esterno dell'azienda e correlate con le attività produttive interne, combinando varie nuove tecnologie per migliorare l'efficienza e diminuire gli sprechi; in questa condivisione d'informazione risiede il vero potere della quarta rivoluzione industriale(Marr, 2018).

Secondo un report di McKinsey, le aziende, per rimanere competitive ed aggiornate e per rispondere allo sviluppo delle tecnologie emergenti ed a nuovi modelli di business, che combinati

possono dare valore all'organizzazione, vorranno investire in quelle dimensioni tecnologiche, che fondamentalmente si possono sintetizzare in 4 filoni principali (McKinsey, 2017):

- Raccolta dati
- Analisi e business intelligence
- Interazioni uomo macchina
- Conversione del mondo fisico

Inoltre, su questi assi di ricerca e sviluppo, possono essere individuate in particolare nove (fig. 2.1) tecnologie che abilitano l'Industria 4.0 e che permettono alle aziende di entrare nella nuova era della competizione industriale; queste fanno da cardini alla connessione tra ambiente digitale ed industriale di cui si è parlato nell'introduzione (Scalabre, 2018):

- Advanced manufacturing solution: sistemi avanzati atti alla produzione, interconnessi che danno sia flessibilità che performance. Esempi sono i sistemi di movimentazione dei materiali automatici e la robotica avanzata.
- Additive manufacturing: sistemi di produzione additiva che aumentano l'efficienza dell'uso dei materiali (Wikipedia, 2018).
- Augmented reality: visione tramite realtà aumentata in aiuto alle attività quotidiane degli operatori (dal manovrare una gru ad un'operazione chirurgica).
- Simulation: simulazione tra macchine interconnesse con il fine di ottimizzare i processi produttivi.
- Horizontal e verticalintegration: integrazione e scambio di informazioni su ogni asse della catena di produzione tra tutti gli attori coinvolti.
- Industrial internet: uso di internet per la connessione non solo intra-aziendale ma anche con l'esterno.
- Cloud: utilizzo di svariate tecnologie cloud come lo storage online delle informazioni, cloud computing, e servizi esterni di analisi dati, ecc. sono incluse le tecniche di gestione di enormi quantità di dati attraverso sistemi aperti.
- Cyber-security: l'aumento delle connessioni sia interne che esterne abilitano lo sviluppo del settore della sicurezza delle informazioni.
- Big Data Analytics: tecniche di gestione e di analisi di enormi quantità di dati attraverso vari sistemi che consentono previsioni o predizioni.

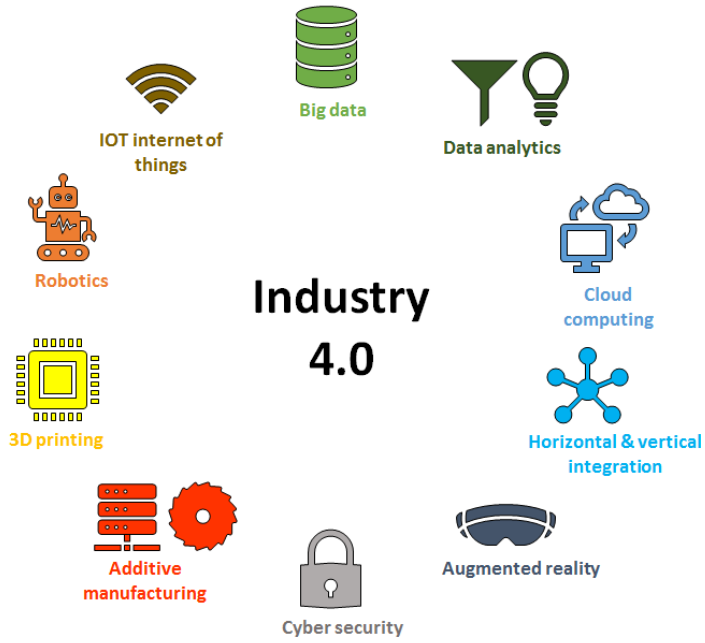


Fig. 2.1 le 9 tecnologie abilitanti

Un'altra interpretazione è stata data da Mueller , nel suo tentativo di riportare un'architettura generale, comune a tutti i settori industriali, che identifica, non solo le tecnologie chiave, ma qualcosa di diverso, qualcosa che integra queste tecnologie in un sistema più ampio; sono 3 i fattori fondamentali individuati per lo sviluppo dell'industria 4.0: l'Internet of Thing (IoT), Cyber-Physical System (sistemi fisico-digitali, CPS) and Smart Factory, a cui si possono ricollegare le nove tecnologie sopra citate, ma che vengono collegate in maniera differente; IoT è la dimensione che riconosce e collega gli oggetti fisici all'ambiente esterno e ai database, permettendo così all'informazione fisica di entrare nel mondo virtuale, tramite l'uso di sensori, attuatori, telefonini e tutti quei dispositivi che permettono la traduzione di segnali o informazioni fisiche in info digitali (Mueller et al, 2017).

Altri autori affiancano all'IoT anche l'Internet of Service (IoS), che sta ad indicare quei sistemi che offrono e combinano diversi servizi, di un solo fornitore come nel caso di Amazon Web Service, o di più fornitori, creando un prodotto con valore aggiunto; le fabbriche del futuro potrebbero quindi offrire non solo i prodotti in sé per sé, ma anche tecnologie produttive offerte tramite IoS (Hermann et al., 2015). Questi due fattori creano le fondamenta per i CPS, che

permettono di chiudere il cerchio portando dal mondo virtuale le informazioni operative e di controllo per i processi fisici, a livello locale, globale, orizzontalmente e verticalmente rispetto alla catena del valore; importante sottolineare che non è semplice raccolta di dati, ma è la struttura di analisi di questi ultimi che fornisce valore aggiunto, inviando le elaborazioni ai processi autonomi che si autogestiscono (Mueller et al, 2017).
La fusione tra mondo reale e cibernetico, crea e abilita le Smart Factories che hanno un metodo innovativo di approcciarsi alla produzione: queste nuove fabbriche sono quelle che aiutano "consapevolmente del contesto" le persone e le macchine nei loro lavori; ciò significa che ci sono sistemi che lavorano in background aggregando dati fisici e virtuali in relazione al contesto in cui sono raccolti, comunicando ed interagendo con l'ambiente esterno; in pratica permette la comunicazione tra CPS diversi, utilizzando l'IoT, così da poter soddisfare il proprio obiettivo di essere di supporto ai lavori svolti in fabbrica (Hermann et al., 2015)

Come si può ben comprendere, essendo un nuovo campo di studio, anche se è uno degli argomenti più dibattuti nella letteratura odierna, non si possiede una visione unanime su come definire o sviluppare progetti di Industria 4.0 od ancora implementarli nelle strutture aziendali esistenti; quindi, sempre grazie al lavoro di Hermann, sono stati definiti dei principi guida dell'industria 4.0, che rappresentano le caratteristiche degli elementi sottolineati nell'architettura appena descritta sopra:

- Interoperabilità: è la capacità dei CPS di comunicare tra loro
- Virtualizzazione: i CPS monitorano processi fisici tramite sensori, a loro volta legali a modelli simulativi per tener traccia dell'evoluzione del processo produttivo e se è in controllo
- Decentralizzazione: significa che il sistema attribuisce dei "TAG" alle macchine per indicare quali lavori sono necessari negli step successivi, senza aver bisogno di un controllo centralizzato
- Orientamento al servizio: è il cambio di direzione che avviene nelle fabbriche, non solo più orientate al singolo prodotto, ma anche ai servizi che possono offrire e a quelli dei CPS utilizzati per rendere "intelligenti" le aziende
- Capacità in tempo reale: collezione dei dati in tempo reale
- Modularità: sistemi modulari che si possono adattare facilmente a nuove richieste o necessità della fabbrica, regolandosi in base al contesto sempre in evoluzione della moderna Industria 4.0

Indipendentemente da come si vuole classificare la quarta rivoluzione industriale o, anche più importante, da dove approcciarsi, bisogna adottare una prospettiva digitale in ogni ambito, ed è fondamentale per le aziende ammodernarsi seguendo una profonda trasformazione digitale della

struttura aziendale ma soprattutto dei processi produttivi e di gestione-controllo, (McKinsey, 2015).

Questo pone una serie di sfide importanti che ogni azienda deve assicurarsi di affrontare per poter continuare a competere nel proprio mercato di riferimento (fig. 2.2): sicuramente è di cruciale importanza mettere al centro del proprio cambiamento i dati, che devono essere intesi come un asset centrale e così come deve crescere l'importanza della loro gestione da un punto di vista strategico, cosa condividere e con chi e se la condivisione porta valore aggiunto (McKinsey, 2015).

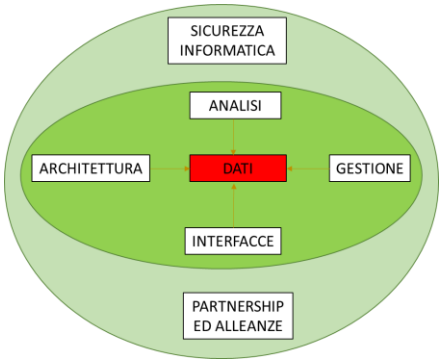


Fig. 2.2 filoni di sviluppo per le aziende del futuro

Dati che devono inoltre, essere acquisiti in tempo reale per poterli sfruttare al meglio, da ogni tipo di fonte possibile, che siano i sensori, i controlli o la rete di comunicazione interna aziendale; ancora devono essere organizzate in una struttura di acquisizione dati precisa e non ridondante perché questi dati saranno alla base per l'estrapolazione di informazioni da parte dei CPS, che se suddivisi in appropriati livelli può sviluppare una complessa rete di distribuzione di informazioni a diversi soggetti (alle macchine per migliorare la loro efficienza, fino ai top manager per prendere decisioni su come indirizzare la produzione) (Lee *et al*., 2015); infine per quanto concerne il trattamento dei dati, lo sviluppo di interfacce per acquisire facilmente e velocemente dati da sempre nuove fonti, può migliorare di molto l'efficienza delle analisi oltre a permettere un'ottima scalabilità per future implementazioni (McKinsey, 2015).

Un altro punto che è vitale da comprendere ed utilizzare in maniera accorta è la creazione di partnership ed alleanze; questo sia per una questione legata sempre all'acquisizione dati, se ci si integra verticalmente nella propria catena si facilita la collezione delle informazioni legate al processo produttivo, sia per migliorare il controllo della qualità del prodotto e dell'efficienza che derivano dalla maggiore comprensione della catena del valore fornita dai dati; in questo modo tutti i servizi offerti dai vari CPS integrati fra loro sono esposti e possono essere indirizzati nei

processi di business ad elementi della produzione, dal controllo qualità, alla logistica, all'ingegnerizzazione etc. (Almanda-Lobo, 2015).

Ultimo macro-argomento, che il progredire dell'Industria 4.0 porta come terreno di sviluppo, è legato alla sicurezza informatica: l'aumento della mole di dati provenienti dalle fabbriche e non solo più dai reparti amministrativi, l'incremento della loro importanza strategica per le finalità di business e il raggiungimento degli obbiettivi aziendali ha indotto grande attenzione verso la loro protezione. La questione sicurezza assume ancora più importanza in relazione alla necessità di aprire i CPS a sistemi esterni per meglio integrare i dati in proprio possesso, alla comparsa dell'IoT e alla comunicazione di dati dalle macchine verso una rete non più esclusivamente interna alla fabbrica.

Le tecnologie operative adottate fino ad oggi non avevano bisogno di connessioni ad internet per poter funzionare, cosa adesso non più vera, e che comporta un problema di ammodernamento delle architetture di comunicazione tra le macchine e di protezione dei dati così trasmessi; bisogna quindi andare a migliorare la sicurezza a livello di rete, scegliere quale topologia fisica e quali protocolli utilizzare; c'è poi da considerare la sicurezza a livello di sistema per proteggere il software che controlla e gestisce il flusso di dati; infine si dovrebbe migliorare il sistema che regola l'accesso dei manager dei dati, migliorando la sicurezza dal punto di vista dell'identificazione dell'utente che li sta effettivamente utilizzando (Gylchrist, 2016).

Questi tre indirizzi di sviluppo per il futuro portano tuttavia con sé alcune problematiche che vanno oltre la mera comprensione ed implementazione di nuove tecnologie. Si crea la necessità di ripensare completamente il proprio business e la propria struttura organizzativa con il confine tra produzione e servizi che diventa sempre più labile e il servirsi sempre più frequentemente di corsi interdisciplinari, per tenere aggiornati i propri dipendenti su come interagire con la clientela. Infatti, i prodotti usciti dalle fabbriche continueranno a rimanere "connessi" così che gli utenti finali, oltre ad avere prodotti sempre più personalizzati, potranno godere di maggiori servizi di assistenza e le aziende ottenere dati utili a migliorare l'efficienza dei propri processi produttivi, facendo de facto entrare i clienti nella catena del valore (Botticini, 2016). In pratica le aziende tradizionali del settore secondario devono diventare aziende digitali, è la mancanza di strategie in questo campo che costringe le firme manufatturiere a ripensarsi quasi completamente per inserirsi in questo nuovo contesto aperto, dovuto anche a mancati investimenti in ricerca e sviluppo (Schröder, 2017).

Da ultima la problematica, di carattere quasi filosofico, di definire l'autonomia e il grado di autogestione che le macchine e i robot hanno, all'interno del processo produttivo e quanto questo influenzi la forza lavoro operante nelle fabbriche (Foidl, 2015).

2.1 impatti economici

La domanda può sorgere è quindi, quali sono gli impatti economici dell'Industria 4.0 e che effetti abbia sul mercato del lavoro. Già si è detto del saldo negativo in perdita di posti di lavoro, ma ci sarà anche un aumento nel divario tra le fasce di lavoratori, secondo lo schema *low skill-low pay e high skill-high pay* (Botticini, 2016).Ci sono poi autori nella letteratura che invece prevedono perdita di posti di lavoro, non solo tra quelli con una routine fissa, ma anche tra quelli con necessità di *high skill* che richiedono task non abitudinari e ragionamento, con un conseguente aumento della competizione tra i lavoratori (Brynjolfsson et McAfee, 2014); altri ancora come la Boston Consulting group, in un report esclusivo, dipinge uno scenario più roseo, con le creazione di 100'000 nuovi lavori a supporto di tutte le nuove funzionalità dell'Industria 4.0 nei prossimi dieci anni, con grande attenzione alle persone con competenze di programmazione ed IT in generale (Bonekamp et Sure, 2015); interessante ancora, citare uno studio che oltre a sottolineare di nuovo la perdita di lavoro tra quelli più semplici e ripetitivi, si avranno cambiamenti anche nei reparti manageriali, poiché l'aumento di automazione e decentralizzazione della decisione porterà ad un rimpicciolimento della gerarchia aziendale e meno necessità di competenze di management tradizionali, con invece un bisogno crescente di manager con capacità trasversali, di controllo e di improvvisazione (Bonekamp et Sure, 2015). La letteratura trovata si può quindi riassumere con un aumento della pressione sul mercato del lavoro, perdite tra i lavori di bassa manovalanza, necessità per i lavoratori di avere competenze riguardanti il mondo informatico, essere flessibili ed avere buone capacità di lavorare in gruppo; tutto ciò comporterà un bisogno di ripensare anche il welfare e il sistema di tasse dei vari stati per compensare l'aumento delle disuguaglianze e le minori entrate (Bonekamp et Sure, 2015).

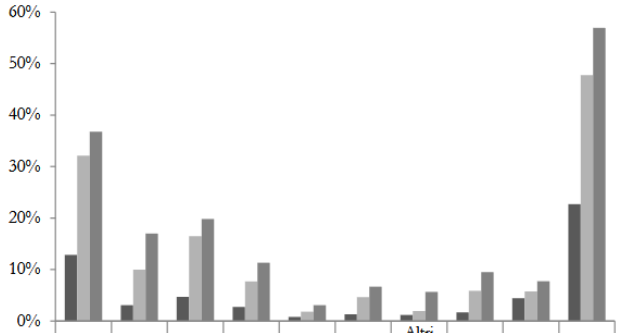
Proprio gli stati svolgono un ruolo fondamentale, non solo nel sostenere i lavoratori, ma anche per stimolare l'industria ad investire in questa nuova rivoluzione industriale (basti ricordare che il termine stesso Industria 4.0 è nato per spiegare al governo tedesco il nuovo modo di concepire la produzione manufatturiera).

Infatti per rimanere al passo coi tempi si rendono necessari investimenti di portata notevole se non quasi impossibile in stati come l'Italia e la Germania; questi sono paesi dove la maggior parte dell'industria è concentrata nelle piccole e medie imprese (PMI), che possono trovare difficile sostenere le spese necessarie per ammodernarsi, cosa non vera, per esempio in posti come gli Stati Uniti, dove a farla da padrona sono le grandi aziende e le corporates che possono

permettersi grandi investimenti per poter rimanere competitive. Ecco quindi che entrano due fattori chiave in gioco: lo Stato ed il mondo del credito.

Il primo può agire tramite agevolazioni alle aziende che adottano certe tecnologie o accorgimenti nei propri impianti produttivi, oppure con lo stanziamento di fondi per specifici progetti d'Industria 4.0, come ha fatto la Germania con 450 milioni di Euro; ancora con l'istituzione di Fondi Centrali di Garanzia, che permettono alle PMI di accedere al credito bancario senza dover firmare onerose polizze assicurative o fidejussioni, avendo come garante del debito stipulato con la banca lo Stato stesso, la garanzia pubblica, in pratica, sostituisce le costose garanzie normalmente richieste per ottenere un finanziamento**.**

Se si va ad osservare ancora più nel dettaglio il caso italiano, sono molte le iniziative messe in campo dal Ministero dello Sviluppo Economico, che vanno ad impattare tutte le categorie dimensionali delle aziende, da agevolazioni per export, super/iper ammortamento, fondi, detassazioni etc. In (*fig. 2.3*) si può comprendere quanto queste forme di aiuto statali siano importanti per lo sviluppo delle aziende nel nuovo contesto produttivo che si sta delineando.



	Iper e Super-ammortamento	Credito d'imposta R&S, Patent box	Nuova Sabatini	Fondi di garanzia	Start-up e PMI innovative	Export e internazionalizzazione	Altri incentivi nazionali per R&S&I	Incentivi regionali per R&S&I	Altri incentivi	Almeno un incentivo
Imprese tradizionali	12,8%	3,1%	4,7%	2,8%	0,8%	1,3%	1,2%	1,7%	4,4%	22,7%
Imprese tradizionali con interventi 4.0 programmati	32,1%	10,0%	16,5%	7,7%	1,8%	4,7%	1,9%	5,9%	5,7%	47,8%
Imprese 4.0	36,8%	17,0%	19,8%	11,3%	3,1%	6,7%	5,7%	9,5%	7,7%	56,9%

Fig. 2.3 agevolazioni e percentuali utilizzo per tipologia azienda (Ministero Italiano Sviluppo Economico)

Tuttavia, per poter accedere a questi finanziamenti sono richieste conoscenze specifiche delle procedure e capacità organizzative, spesso fuori dalla portata di una piccola impresa, tanto che dovrebbero essere ripensati con un iter più trasparente per facilitarne l'accesso.

Importanti inoltre le collaborazioni tra stati, come quella che avviene tra Francia, Italia e Germania dal 2017, con un accordo a tre per promuovere insieme la digitalizzazione del settore manufatturiero e supportare tutta l'Unione Europea verso questo obiettivo.

Il comparto finances invece agisce in maniera differente a seconda di chi stanno per finanziare: i consumatori di prodotti innovativi, facenti parte di settori tecnologicamente avanzati, hanno già contratti di lungo termine con varie banche ed è relativamente facile per loro ottenere nuovo credito così da poter fare investimenti progressivi per rimpiazzare le vecchie linee produttive con quelle nuove; diverso il discorso per le aziende appena create che mettono a punto i loro servizi ed prodotti *smart* da inserire nelle nuove fabbriche e che sono spesso compagnie IT giovani con cashflow negativo e con prodotti e modelli di business che non sono ben compresi dalle banche, risultando perciò rischiosi; per questo fanno la loro comparsa i grandi fondi d'investimento, per finanziare queste start up, e che, al contrario delle banche, diventano veri e propri azionisti, portando oltre al capitale reti di conoscenze e know-how specifici (Schröder, 2017). Caso particolare è quello italiano, dove le PMI sono ancora molto legate alle banche a differenza di altri paesi dell'eurozona e dove gli investimenti in capitali sociali di nuove aziende fanno fatica ad avvenire, il mercato finanziario è relativamente piccolo; questa problematica è stata solo in parte risolta dall'utilizzo delle Garanzie Pubbliche (fig. 2.4) (Botticini, 2016).

Fig. 3 – Garanzie pubbliche (mld €)

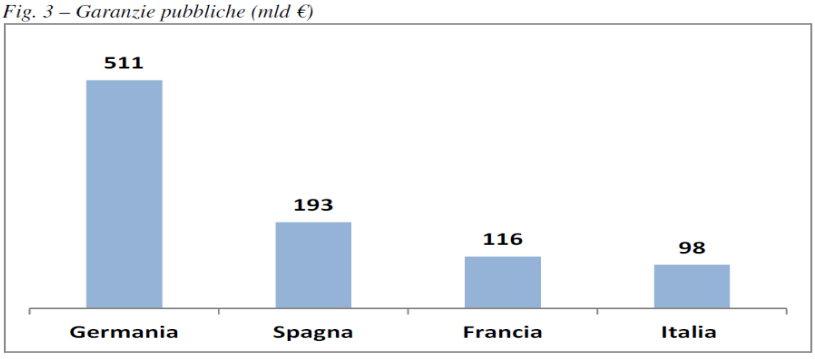


Fig. 2.4 utilizzo fondi garanzie pubbliche (mld €) (Eurostat)

2.2 Casi applicativi Industria 4.0

L'Industria 4.0 sta, quindi, rimodellando il mondo in molte delle sue parti, dal mercato del lavoro, al modo di pensare e di risolvere i problemi, all'economia, ma soprattutto la fabbrica e tutto ciò che vi è connesso.

Si vengono così a creare differenze tangibili tra fabbrica tradizionale e quella 4.0; dove prima l'obiettivo di ogni azienda, per poter essere competitiva nel mercato di riferimento, era un'economia di scala (minimizzazione costi), così come l'utilizzo dei dati era relegato a migliorare la performance della linea produttiva ed avere meno tempi di inattività possibili (classica gestione della produzione, in cui si guarda al passato per capire se il presente è in linea e non si pensa al futuro) ora, con l'Industria 4.0, l'obiettivo è la creazione di un unico grande ecosistema che interagisce al suo interno e con l'esterno (gli utenti nei vari punti della catena del valore), e che si autogestisce con il solo scopo di migliorare l'oggetto finale e la sua produzione. Ne consegue che l'attenzione si sta spostando verso la connessione di ogni aspetto legato al prodotto, che siano le macchine per la produzione dello stesso, i sensori che monitorano il processo e le performance dei macchinari, i sistemi dei fornitori e dei distributori legati all'azienda, ed il prodotto finale in mano al consumatore,

Un esempio di questo aumento dei servizi connessi al prodotto è ben rappresentato dalla Rolls Royce; questa azienda del settore automotive fornisce anche motori per gli aerei della Boeing e negli ultimi anni ha iniziato ad offrire un servizio di manutenzione in collaborazione con Microsoft: invece di vendere i motori, li affitta, l'azienda che li compra dà l'accesso ai sistemi che gestiscono questi componenti nei loro aerei, così la Rolls può offrire un servizio di manutenzione predittiva, di miglioramento dei percorsi e consumo del carburante, su come far arrivare in orario gli aerei; grazie ai dati che riceve dalle varie compagnie che producono aerei, inoltre può capire come migliorare il proprio prodotto ricevendo dati su come vengono utilizzati, gli stress a cui sono sottoposti, le condizioni etc. Questo è l'esempio perfetto di creazione di un Sistema connesso ai fini di migliorare l'intera filiera del valore (CorCom, 2016).

Comunque, i casi applicativi sono molteplici e non è obiettivo di questo elaborato discernerli tutti; nondimeno può risultare utile per la comprensione del nuovo mondo che si sta delineando e per introdurre l'elemento al centro di questa tesi, elencare e spiegare alcuni esempi di come le nuove tecnologie possono essere implementate nelle fabbriche e di come cambiano il settore manufatturiero.

2.2.1 Realtà aumentata

Uno dei temi principali su cui i ricercatori della quarta rivoluzione industriale si stanno soffermando è l'utilizzo della realtà aumentata in contesti manufatturieri o nel settore edile.
La realtà aumentata può accrescere l'efficienza produttiva in quasi ogni contesto; molti studi sostengono che possa arrivare ad un miglioramento anche del 32% sulle performance di produzione di una fabbrica (Lena, 2017). Una ricerca della Toshiba sui trend futuri sottolinea come l'82% delle aziende pensano che la RA sarà utile ai loro processi (Reynolds, 2018). Con l'aiuto di occhiali appositi, gli operatori non dovranno più spendere ore lavorative per imparare manuali di assemblaggio o di manutenzione (fig. 2.5), ma potranno iniziare direttamente a lavorare sulle macchine o sulla catena produttiva (Lena, 2017). La realtà aumentata permette di elaborare una grande quantità di dati e di far visualizzare ai tecnici solo quelle importanti, come segnalare quali parti si sono guastate, o come congiungere delle parti della macchina o come ripararle velocemente per far ripartire la produzione.
Per esempio, Amazon sta pensando di dotare i propri dipendenti di questo tipo di tecnologia per poter meglio individuare i prodotti; inoltre questa funzionalità può essere integrata nel prodotto finale per aiutare i consumatori a sfruttare appieno ciò per cui hanno pagato, come nel caso di robot per la chirurgia medica, che permettono ai medici di avere immagini 3D del corpo del paziente e provare in anticipo interventi complessi.

Fig. 2.5 Augmented Reality forindustrial Training (Wikipedia)

2.2.2 Cloud Computing

Il cloud computing può essere riassunto come la capacità di immagazzinare dati ed eseguire programmi che non sono direttamente sulla macchina da cui si sta operando, ma, attraverso una connessione internet, su un computer differente. I network attached storage (NAS) dedicati, che sono server nell'ufficio o comunque interni alla fabbrica, non contano come cloud computing; per essere un servizio cloud, questo deve utilizzare l'infrastruttura di Internet per comunicare. Si vengono così creare vari modi d'integrazione di questa tecnologia, abilitata negli ultimi anni grazie al miglioramento delle reti di telecomunicazioni, e si può anche assistere alla nascita di nuovi modelli di business (fig. 2.6).

Il cloud computing può essere dato direttamente all'utente finale: ciò comporta che il prodotto acquistato rimane connesso al sistema dell'azienda che l'ha creato, che così può continuare a monitorarne il ciclo vitale e dare suggerimenti agli utenti su come sfruttarlo al meglio (vedi caso Rolls Royce); quello che viene offerto è soprattutto un servizio, tanto migliore tanto sono più i clienti che lo sfruttano, poiché i suggerimenti si basano su maggior volumi di dati (Schmidt, 2015).

Oltre a servizi che si differenziano per il contesto in cui il cloud viene applicato, questo metodo mette a disposizione un fattore importantissimo per molte aziende ma che alcune, essendo lontane dal proprio core business, non saprebbero come sviluppare adeguatamente: i web server. Sono ormai moltissime le aziende che invece di creare i propri servizi online su una piattaforma unica, si appoggiano a servizi di cloud messi a disposizione dai colossi del mondo digitale. Esempio più conosciuto è quello degli Amazon Web Service (AWS), un pacchetto di tool e servizi messo a disposizione dal gigante tech americano ad altre aziende, in modo che queste possano creare le proprie applicazioni senza dover sviluppare da zero il prodotto od investire grandi quantità di denaro in server e potenza di calcolo; Amazon ha creato un intero nuovo mercato, fornendo tutto il necessario ai propri clienti per creare il loro servizio web (Griffith, 2016).

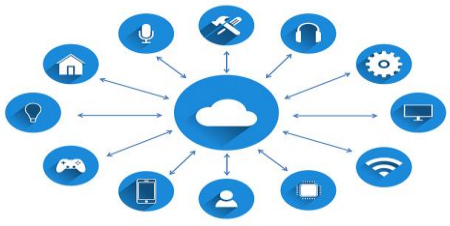


Fig. 2.6 Possibilità cloud Computing (Tumisu, Pixabay)

2.2.3 Produzione Avanzata

Le advanced manufacturing solutions (AMS), ovvero soluzioni di produzione avanzata, rientrano tra quelle tecnologie che permettono la gestione di grandi volumi di prodotto, mantenendo però una personalizzazione di questi ultimi come la crescente domanda richiede; questo risultato è ottenuto dalla combinazione di più tecnologie avanzate, IoT, Big Data, Cloud Computing, intelligenze artificiali, che permettono di avere una produzione flessibile, controlli automatici e diminuzione dei costi.

La partenza è il design, da qui inizia l'ottimizzazione del processo, grazie anche alle visuali 3D, i manager possono vedere come si svilupperanno fisicamente i loro impianti e capire come organizzarli al meglio; segue la fase di esecuzione, che può essere facilmente controllata con dati provenienti in tempo reale dal sito di costruzione; infine c'è la necessità di collaborazione tra i vari stakeholder per ottimizzare sia il modello che la realizzazione del progetto. Un esempio di applicazione si può trovare nell'industria automobilistica, con le nuove linee produttive che si adattano ai vari optional richiesti dai clienti. Anche Boeing nei suoi aerei e nei suoi impianti produttivi sta perseguendo un obiettivo di questo genere (Vinovski, 2018).

3. Analisi di Big Data

Si è quindi arrivati a descrivere quello che non è un caso applicativo come gli altri, ma è il fulcro di questo elaborato e del suo obiettivo: l'Analisi di Big Data. Senza questa metodologia, tutte le casistiche viste sopra non si sarebbero potute sviluppare, perché il fulcro di tutta la quarta rivoluzione industriale sono i dati e la capacità di analizzarli per carpire informazioni fino ad oggi nascoste, così da migliorare le proprie performance, che siano le vendite, la qualità del prodotto o l'efficienza del processo produttivo. I Big data, una mole enorme di dati provenienti da ogni fonte della catena produttiva, combinati insieme ed analizzati danno accesso ad un mondo che non era conosciuto, permettendo la nascita non solo di nuovi modelli di business ma di mercati completamente nuovi.

L'analisi di dati non è una tecnica nuova al mondo tecnologico, ma grazie al grande incremento della potenza di calcolo disponibile e all'aumento di informazioni utilizzabili per le analisi, questa pratica ha ottenuto sempre più rilevanza nei contesti applicativi, sia industriali che finanziari che medici, i campi di applicazione sono praticamente infiniti, ovunque ci sia la possibilità di raccogliere dati (Lee, 2014).

L'obiettivo delle analisi è determinare pattern che siano validi, non noti a priori, potenzialmente utili e comprensibili, dunque percorsi logici nel marasma di dati; trovare queste associazioni e comprenderle può portare a grandi miglioramenti nelle performance produttive (ridurre i costi, velocità e precisione) e di vendite, ma anche nel design del prodotto stesso e nell'esperienza del cliente; l'analisi di dati e trend non è più solo utile ai mercati finanziari per capire quale titolo salirà e quale scenderà, ma ottiene una connotazione di evoluzione, si vuole aspirare a qualcosa di nuovo e migliore (Gylchrist, 2016).

Nel nuovo scenario dell'Industria 4.0, processori e sensori vengono posizionati sulle macchine e nei prodotti per ricevere in tempo reale i dati sull'andamento del processo e sullo stato dei robot, generando un flusso di dati enorme appunto i Big Data. Così tanti e di vario tipo (strutturati, operativi, di vendite etc), che non possono essere gestiti dai database tradizionali, ed è difficile trattarli e renderli coerenti. Le nuove metodologie di analisi riescono in questa confusione ad estrarre, in tempi ragionevoli, informazioni utili al management ed agire in conseguenza dei dati raccolti o al reparto di manutenzione per effettuare aggiustamenti o riparazioni prima che occorra il guasto vero e proprio (Zhou, 2015). Con la comparsa dei CPS, l'integrazione di ancora più fonti di dati, dei fornitori, delle comunicazioni tra le macchine, dei rivenditori e dei clienti, renderà sempre più importante il ruolo dei data Analyst e i data scientist anche in contesti manufatturieri oltre a quelli digitali e finanziari (Zhou, 2015).

3.1 Stato dell'Arte

L'argomento analisi di Big Data di per sé è molto amplio e non è l'obiettivo di questa tesi elencare tutti metodi, tool, accorgimenti e possibilità che questa tecnologia mette a disposizione. Tuttavia, è bene cercare di riassumere quanto meno i punti salienti (fig. 3.1), da dove partire, le tecniche maggiormente usate e le problematiche più facilmente riscontrabili.

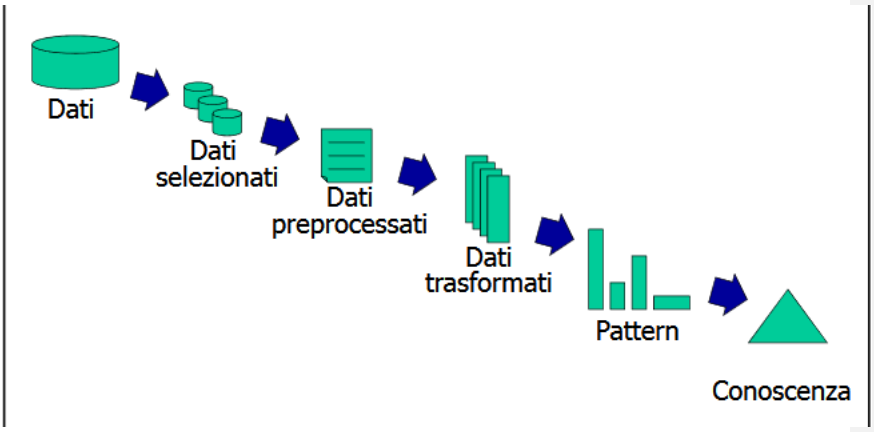


Fig. 3.1 Step per analisi dei dati (Baralis, politecnico di Torino)

Al centro di tutto risiedono ovviamente i dati, non vanno però considerati tutti, ma è di fondamentale importanza la loro raccolta e come viene fatta. Quindi per prima cosa va pensato quale sia l'obiettivo che si vuole ottenere analizzando i dati, se è per manutenzione, per migliorare il prodotto, velocizzare la produzione, o ancora incrementare l'efficienza di un macchinario; l'essenza del data analytics è la ricerca di un'ipotesi che poi si cerca di provare analizzando i dati non strutturati che sono in proprio possesso, scoprendo informazioni prima nascoste. Compreso questo vanno scelti e raccolti solo i dati utili a questo fine, in modo da non averne di inutili, che renderebbero solo più lungo il processo di analisi, oltre a rischiare di ingenerare qualche errore di calcolo. Nella scelta va anche considerata la tipologia del dato che si vuole tenere, possono essere tradizionali (numerici), di testo, date, righe intere organizzate in file json, sensori, coordinate spaziali e molte altre ancora (O'Neil, 2013).

Si effettua poi una fase esplorativa tramite grafici, tabelle, statistiche riassuntive, per carpire eventuali problemi nei dati e come questi siano strutturati; si plottano le variabili tramite le serie temporali, si analizzano le distribuzioni e le varie statistiche (O'Neil, 2013).

Segue poi una delle fasi più importanti e più lunga per un data analyst, la pulizia, perché i dati raccolti non sono mai "puliti". È un processo fondamentale per non avere impurità nell'analisi finale, come degli outliers, punti che non c'entrano con gli altri, ma poiché sono stati considerati, possono portare a pessime interpretazioni delle informazioni ottenute. Questa fase consiste nell'identificare gli errori o i dati inconsistenti e correggerli, ove possibile, così da avere dati di alta qualità. Per farlo bisogna guardare al database di cui si è in possesso, verificarne l'accuratezza, la presenza di duplicati, se i dati sono coerenti ed espressi tutti con la stessa sintassi. Spesso viene eseguita anche una standardizzazione dei dati, per equipararli, con, esempio, tecniche di min-max, oppure tramite i valori della distribuzione statistica della curva normale; si devono inoltre completare gli eventuali dati mancanti e si può realizzare ciò sostituendo la media della colonna con i dati mancanti, con la mediana o altre tecniche, purché mantengano coerenti i dati e non li snaturino, per questo sono migliori database bilanciati. Dopo tutte queste procedure si può procedere alla validazione dei dati e proseguire allo step successivo (SunTec, 2016).

Si passa poi ad una fase di trasformazione dei dati, che inizia già con la normalizzazione degli stessi e prosegue con una selezione di quelli (delle colonne di dati) che si vogliono portare avanti nell'analisi, oppure semplicemente per ridurre il numero di variabili in gioco ed analizzarle poco alla volta, combinandole anche in maniera diversa per poi cercare d'interpretare i risultati incrociandoli. È possibile eseguire anche campionamenti di vario tipo, per ridurre, invece che le variabili, il numero di dati da esaminare, o ancora discretizzare una variabile in diverse fasce; per esempio la durata dei viaggi se si possiedono le corse in bicicletta di un'azienda di bike sharing, si possono creare fasce in base al tempo trascorso per completare il tragitto dell'utente. Da citare infine sono le tecniche che sintetizzano decine e decine di variabili riassumendole in poche, perdendo le informazioni reali si possono però poi rappresentare dati che prima avevano centinaia di colonne, in grafici con 3 assi; esempio sono le tecniche*:* principal component analysis*,* abbreviata in PCA o la decomposizione in valori singolari (SVD).

Ultimi due step dell'analisi dati: decidere il tipo di estrazione dei dati e scegliere l'algoritmo che si vuole applicare per estrapolare le informazioni, sia con la finalità di descrivere meglio i dati di partenza sia se l'obiettivo è fare predizioni di qualche tipo su dati nuovi, partendo da quelli che si possiedono. Il lavoro del data Analyst si conclude quando ha estrapolato la conoscenza e l'ha resa comunicabile, comprensibile ai manager che in base a queste informazioni dovranno poi prendere le decisioni.

Ovviamente l'analisi dei dati porta con sé alcune criticità rilevanti a priori, che ostacolano gli addetti ai lavori. Come già accennato, uno dei problemi più importanti è l'individuazione del rumore, sono dati di disturbo che distorcono le analisi, che possono capitare in caso di errori nelle misurazioni, o perché si è considerata la variabile sbagliata; esistono i cosiddetti Silos di dati,

ovvero memorie, o veri e propri database, in cui vengono immagazzinate moltissime informazioni, ma che non hanno nessun legame fra loro, e creano solo un agglomerato di dati confusi da cui non è possibile estrarre informazione (Piesync, 2018); si è già parlato di dati mancanti ed inconsistenti; degna di sottolineatura è anche la problematica legata alla potenza di calcolo e la scalabilità dei dati, che diventa molto difficile con grandi masse di dati. Nelle aziende si presentano inoltre ulteriori problematiche: la mancanza di know-how nel metodo di raccolta e trattamento dati, c'è una grossa penuria di data Analyst nel mercato del lavoro, sia in America che in Europa; infine hanno grandi problemi di privacy e di come trattare e mettere al sicuro i dati sensibili che raccolgono dagli utenti e dai propri dipendenti (Bi-Survey, 2018). Sono da menzionare, ancora, le difficoltà che si incontrano quando si implementano gli algoritmi, che hanno delle limitazioni proprie (alcuni adatti a lavorare su grandi volumi di dati, altri solo in parallelo su più macchine etc.) e che ogni volta devono essere adeguati ai dati che si sta trattando in quel momento.

Quando si riesce a limitare i problemi generati da tutti questi fattori, si può passare alla fase di analisi, che per i Big Data, spesso consiste nel predire un qualcosa di sconosciuto, o nel cercare conferma di una propria ipotesi tramite metodi di data mining.

3.1.1 tecniche di analisi: Regole di Associazione

Una delle prime tecniche di data mining che si possono utilizzare sono le regole di associazione: trattasi di un algoritmo matematico basato sulle regole che intercorrono tra le variabili del database che vengono analizzate, in pratica una ricerca di correlazioni tra gli elementi forniti. Tecnica inizialmente usata nei settori di marketing delle aziende, per capire le preferenze di acquisto dei clienti, poi estesasi ad altri campi di studio, come la medicina o analisi di trend sul web.

La definizione originale delle regole di associazione è: dato un gruppo I di *n* variabili binarie chiamate oggetti (le variabili, o meglio le colonne del database) e un gruppo D di *m* transazioni (le righe del DB), una regola di associazione è definita come $X \Rightarrow Y$, dove $X, Y \subseteq I$, questa procedura viene ripetuta per ogni coppia possibile di variabili (Hahsler., 2015).

A livello pratico significa che quando si ha X si ha anche Y, un esempio classico spesso citato è sui carrelli della spesa di un supermercato, si è rilevato che il 2% delle transazioni contenevano sia pannolini che birra e che nel 30% dei casi chi comprava pannolini comprava anche la bevanda alcolica.

Questa metodologia di analisi si basa sull'impostazione di due parametri principali (ci sono poi una serie di altri fattori che si possono variare): il primo è il valore del supporto, che indica il

valore soglia per cui entrambe le variabili si presentano insieme nel gruppo delle transazioni; il secondo valore è la confidenza, quante volte si presenta Y dato X. Nelle varie regole che si creano si considerano solo quelle con il valore del Lift maggiore di 1, che sta a significare correlazione positiva tra le due variabili analizzate.

Ecco alcuni esempi tratti da un progetto personale di analisi effettuato nel corso di Business Intelligence, ricavati da un database di Bike Sharing tramite l'algoritmo FP-Growth: in (fig. 3.2) si possono vedere alcuni esempi di regole di associazione, come la prima che ha un'occorrenza nel database dell'1% e che dice che gli uomini che partono da quella determinata stazione sono per il 57% compresi nella fascia di età che va dai 32 ai 46 anni.

No.	Premises	Conclusion	Support	Confidence	Lift
393	member_gender = Male, start_station_name = San Francisco Ferry Building (Harry Bridges Plaz	member_birth_year = [1972.6 - 1985.8]	0.010	0.572	1.274
389	member_gender = Male, end_station_name = San Francisco Ferry Building (Harry Bridges Plaza	member_birth_year = [1972.6 - 1985.8]	0.011	0.557	1.241
378	member_gender = Male, end_station_name = The Embarcadero at Sansome St	member_birth_year = [1972.6 - 1985.8]	0.011	0.534	1.189
356	start_hour = range16 [15 - 16]	member_birth_year = [1972.6 - 1985.8]	0.038	0.494	1.101

Fig. 3.2 esempi di regole di associazione

3.1.2 tecniche di analisi: Clustering

Il Clustering è una metodologia di analisi molto importante che si pone l'obiettivo di riunire in gruppi i dati che hanno similarità tra loro, assegnando ad ogni dato del database un'etichetta che identifica il gruppo. È una tecnica utilizzata in diversi campi, dal machine learning, all'analisi di immagini, compressione dati, computer grafica ed altri ancora, è quasi sterminato il campo di applicazioni possibili per questa metodologia.

Tuttavia, non esiste un unico algoritmo di clustering, univoco per ogni situazione, bensì una serie di tecniche, che variano per come calcolano la distanza tra i punti, dalla tipologia dei dati che si cerca di raggruppare; è un processo che va per tentativi, per trovare il metodo migliore che rappresenti i dati e che permetta una maggiore comprensione degli stessi.

Il parametro più importante da indicare è il numero di gruppi in cui si vuole spezzare il database, indicato generalmente con K; questo si ricava dall'analisi del MeanSqaureError (MSE) al variare di K stesso; si deve poi tipicamente indicare come deve essere calcolata la distanza tra i punti (di base si usa la distanza euclidea).

Le tipologie di algoritmi di clustering si possono riassumere in: clustering gerarchico e clustering per partizione. Nel primo i dati possono appartenere a più cluster, poiché si vengono a creare delle gerarchie di gruppi (tipicamente rappresentati come un albero); a sua volta questa tipologia di analisi può partire da un unico grande gruppo che contiene tutti i dati e tramite un

processo iterativo suddivide i punti in raggruppamenti concentrici successivi, oppure considera ogni punto come un cluster e procede per aggregazione di dati. Nella seconda tipologia, invece, i dati sono ben separati ed ogni punto appartiene ad uno ed un solo cluster. Ci sono poi altre distinzioni per catalogare gli algoritmi di clustering, come se sono inclusivi od esclusivi, fuzzy o non-fuzzy, omogenei o eterogenei. I cluster risultanti possono quindi variare molto loro stessi, ben separati, basati su un centro comune, gruppi creati per densità di dati etc.

Tra le principali tecniche di clustering si vuole qua ricordare: Il K-Means; uno dei principali algoritmi di questa categoria, è del tipo per partizione, dove ogni cluster è associato ad un centroide, un valore che fa da centro (la media dei punti tipicamente), scelti inizialmente come casuali, gli altri punti vengono assegnati al cluster del centroide a cui sono più vicini. In sé non è un algoritmo complesso e viene valutato tramite il calcolo del Sum of SquareError (SSE). In (fig. 3.3) si può osservare una rappresentazione di cluster ottenuti tramite K-Means, sempre sull'esempio del Bike Sharing, ottenuta diminuendo i numeri di attributi tramite la tecnica di riduzione SVD, non è l'esempio di cluster meglio separarti, ma rende un'idea di un caso applicativo reale.

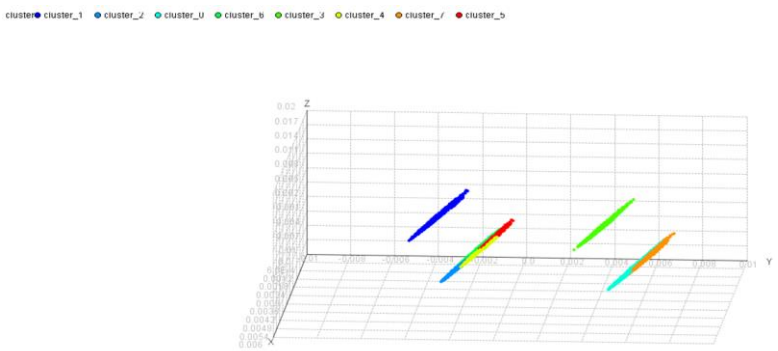


Fig. 3.3 Cluster su Bike Sharing, tramite tecnica SVD

Un'altra tecnica di clustering è il DBSCAN; i cluster, con questo metodo, vengono identificati con aree a maggior densità di punti, calcolando sempre le distanze da un centro (questa volta rappresentato da un punto vero e proprio dell'insieme di dati), ma aggiungendo una soglia sul numero di punti che devono essere intorno ad esso per far parte di quel gruppo. La densità EPS è proprio il numero soglia di punti che devono essere intorno al centro perché questo sia considerato core point; i border point non superano EPS ma sono vicini ad un core point; infine ci sono i punti di rumore che non sono nessuno dei due tipi elencati sopra. Spesso il DBSCAN è proprio utilizzato allo scopo di individuare il rumore nei dati per poi eliminarlo, viene utilizzato come mezzo

di pulizia dati. Per poter essere efficace al massimo deve lavorare su gruppi con densità simile (non devono esserci gruppi densamente popolati e altri scarsamente).

3.1.3 tecniche di analisi: metodi di predizione

Il clustering divide in gruppi i dati eseguendo una sorta di predizione con il fine di scoprire la natura degli stessi e le regole di associazione cercano di prevedere come si abbinano le variabili fra di loro. Esistono metodi volti alla predizione pura che non scoprono informazioni sui dati, ma predicono il comportamento di dati nuovi, non ancora valutati, sulla base di quelli già esistenti, (tipicamente si cerca una funzione che dato un insieme X ci porti a trovare Y, una funzione di approssimazione). Questo è il caso di una tecnica che prende lo stesso nome di una statistica, la regressione. Lo scopo di questo metodo applicato al data mining è predire una variabile di tipo continuo, come il prezzo di una casa, è quindi utilizzato quando i possibili valori che il fattore può assumere sono infiniti; differisce dal metodo statistico per prima cosa perché le variabili possono assumere si qualsiasi valore ma non continuo come in termini matematici, esempio quando si predice l'età di una persona; in secondo luogo la regressione in data mining non cerca di approssimare i dati ad una retta e non ne cerca l'appartenenza ad un semipiano specifico (Moschese, 2004). Esempi di algoritmi utilizzati sono la regressione lineare, quella multipla, non lineare o ancora quella locale pesata.

A fianco della regressione si trova sempre un altro metodo per la predizione, che predice, invece di un valore possibilmente infinito, un valore discreto con dimensionalità piccola, è il caso della classificazione. L'input sarà formato da un insieme di dati, suddivisi in *n* variabili, tra cui una chiamata classi; si trova un modello per questo fattore "classi" come funzione dei valori degli altri attributi; l'obiettivo è allenare il modello su dati già classificati (training set), per poi predire nuovi dati senza questa etichetta e deve essere in grado di gestire grandi quantità di dati. Un problema di regressione può essere trasmutato in uno di classificazione (discretizzando gli intervalli) e viceversa (Moschese, 2004). Gli utilizzi della classificazione sono molteplici: dalla manutenzione di macchinari, al classificare le transazioni delle carte di credito, le strutture delle proteine, o verificare l'affidabilità creditizia di un cliente di una banca.

Ci sono vari algoritmi per classificare raggruppati per famiglie di tecniche, ma prima di poterli usare devono tutti essere validati; cioè devono essere ricercati i migliori parametri possibili da utilizzare, diversi a seconda del metodo scelto; normalmente si esegue la validazione K-Folds, che è una variante della Cross-Validation. In questo tipo di validazione il gruppo di dati usare per allenare l'algoritmo viene diviso in tanti sottogruppi tanto è il valore di K (valore base 10) e ogni

sottogruppo viene predetto in base agli altri K-1 piccoli gruppi; all'aumentare di K la precisione aumenta ma anche il tempo di esecuzione, per questo si sceglie tipicamente il valore di 10.

Prima di validare un modello si può eseguire un campionamento (stratificato, casuale o deciso dall'utente) per tenere una porzione di dati da predire dopo aver eseguito la validazione e di solito si parte da dati normalizzati; su questi dati verrà poi fatta la predizione e calcolata la matrice di confusione, che serve a capire quanto è "buono il modello", se ha predetto i dati nella classe a cui effettivamente appartenevano. Infine, ogni algoritmo può essere valutato tramite vari indicatori, che si ricavano anche dalla matrice di confusione stessa; questo processo può essere fatto tramite varie stime come: l'accuratezza, precisione (la percentuale di oggetti correttamente assegnati ad una classe, su tutti gli elementi assegnati alla stessa), richiamo (la percentuale di oggetti correttamente assegnati ad una classe, su gli elementi appartenenti realmente alla classe stessa), scalabilità (es. grandezza necessaria per il training set), robustezza a rumori e dati mancanti, interpretabilità del modello (se è comprensibile come vengono fatte le scelte di classificazione, esempio la leggibilità di un albero di decisione.

Esaminando gli algoritmi per creare un albero di decisione, di base vengono divisi i record del database per gli attributi seguendo criteri dettati dall'algoritmo scelto e dal tipo di dato che si sta trattando, se nominale, ordinale o continuo; l'algoritmo continua a dividere gli attributi fino a quando non raggiunge la massima profondità impostata dall'utente o ogni foglia è formata da elementi di una sola classe; l'attributo radice, il miglior attributo per rappresentare i dati, viene scelto in base a criteri che variano dall'algoritmo scelto (Sanjeevi, 2017).Uno "split" (una divisione di un attributo) è buono se è omogeneo, cioè con dati appartenenti quasi esclusivamente ad una sola classe, l'impurità dei nodi viene verificata tramite diversi indici, come ad esempio quello di GINI. Un esempio di albero di decisione si può vedere in (fig. 3.4), dove si possono vedere nodi foglia non particolarmente omogenei, che significa che l'algoritmo si è fermato perché ha raggiunto la profondità massima data dall'utente.

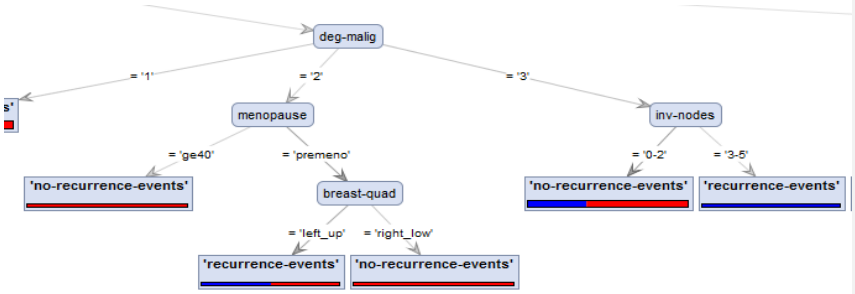


Fig. 3.4 Esempio albero di decisione

I vantaggi sono la facilità di costruzione del modello e il costo basso dal punto di vista del calcolo, ottimo a classificare dati sconosciuti, accuratezza, simile ad altre tecniche per dataset semplici, tuttavia soffre in questo ultimo parametro se ci sono dati mancanti nei record.

Seguono, poi, gli algoritmi di Random Forest; è un metodo dove il computer costruisce una moltitudine di alberi di decisioni tra loro indipendenti, per scegliere quello che meglio si adatta ai dati; tramite una sostituzione continua del campione su cui viene fatto il training di un algoritmo di albero di decisione, viene selezionato di volta in volta un sottogruppo di variabili casuale, per evitare che pochi attributi molto importanti vengano ripetutamente utilizzati nei vari alberi, rendendo così i modelli fortemente correlati tra loro, ed infine viene eseguita la predizione, anche su dati mai predetti.

L'algoritmo del K-Nearest Neighbor, abbreviato in K-NN, richiede tre cose principali per addestrare l'algoritmo di classificazione, un set di record, la metrica da usare per calcolare la distanza tra i punti (standard è quella euclidea) ed il numero di punti K che devono essere trovati vicini per essere classificati uguali. Quando verrà usato il modello per predire dati nuovi, verrà calcolata la distanza di ogni punto da quelli utilizzati nel training e verranno assegnati al gruppo a cui sono più vicini. Il problema di questa tecnica risiede nello scegliere un K adatto, né troppo piccolo (troppo sensibile al rumore) né troppo grande (include anche elementi di altre classi); inoltre non viene creato un vero e proprio modello di classificazione come l'albero di decisione ed è dispendioso classificare dati sconosciuti.

Altra tecnica di classificazione importante da citare è quella che si ispira al connessioni del cervello umano, le reti neurali; i neuroni diventano le unità di elaborazione e le sinapsi le reti di comunicazione; si hanno i nodi di partenza che ricevono un set di dati ciascuno ed un proprio set di pesi, che verranno usati sui punti di input, eseguiti i calcoli in un nodo si passa a quello successivo fino a quando non si arriva ad un output finale che viene confrontato con quello previsto per

calcolare l'errore del modello. È un metodo molto robusto e con alta accuratezza, ha problemi però nel tempo di training (scala male con molti dati usati per il training), inoltre non crea un modello interpretabile, quindi non possono essere usate le conoscenze del dominio per capire lo schema utilizzato per predire e viene utilizzata per riconoscimento del parlato o delle immagini, traduzioni, filtri per social network (van Gerven, 2017).

Infine, per quanto concerne la classificazione ma più in generale i metodi per prevedere un output, è interessante citare la Survival Analysis; tecnica particolarmente usata in campo medico, questa viene utilizzata per predire il momento in cui accade uno o più eventi dati, come possono essere la morte di un paziente, il ripresentarsi di una malattia cronica, il presentarsi di un guasto in una macchina, o di un evento storico in contesti di analisi sociali. Per poter eseguire questo tipo di analisi sono necessarie due cose: sapere precisamente quale evento si vuole modellare e un dataset che contengono un'informazione temporale su questo accadimento che si desidera predire, con questi input si può calcolare la funzione di probabilità su dati sconosciuti per capire quando l'evento capiterà. Il problema più grande di questa tecnica risiede nei dati censurati, perché non gli è mai accaduto l'evento che si sta cercando al momento dei calcoli oppure gli è capitato prima delle analisi; inoltre anche è da trattare anche il troncamento, cioè quelle osservazioni che non sono state portate a termine perché si sono perse le tracce del soggetto in esame (Miller, 2011).

3.2 Machine Learning

Un argomento strettamente connesso al data mining è il Machine Learning (ML). Un insieme di metodi, inizialmente sviluppati parallelamente alle intelligenze artificiali, ha visto poi allargarsi i propri campi di studio, in particolare al data mining. Il ML è un termine coniato nel 1959 da Arthur Samuel, è lo studio di modelli statistici atti a migliorare le performance di un algoritmo dopo varie iterazioni, in pratica ricerca l'ottimizzazione di un processo cognitivo digitale (Koza, 1996); si può considerare come una scienza che studia algoritmi di apprendimento per computer, che anche se la maggior parte sono in circolazione da decenni, ha trovato nuova linfa nei miglioramenti della potenza di calcolo e nella necessità di analizzare i Big Data.

Come per la classificazione, dove appunto ha trovato molte applicazioni, l'algoritmo si allena su un dataset di training, dove impara a riconoscere i pattern, per poi predire nuovi dati anche se non è stato specificatamente programmato per eseguire quel task e si perde la necessità dell'esperto di dominio dei dati, l'algoritmo impara da solo cosa è necessario sapere (Koza, 1996). Il data mining è diventato un campo di ricerca nell'ambito più vasto del machine learning perché si possono così esplorare i dati in maniera autonoma (nel caso per esempio delle regole di associazione) e prevedere con maggiore precisione (classificazione e regressione). La differenza più grande tra i due argomenti rimane però, che il machine learning si basa su conoscenze acquisite, il data mining puro sullo scoprire nuove conoscenze, anche se entrambi usano metodi dell'altro campo per i propri processi, per esempio il ML usa il data mining in fase pulizia ed esplorazione iniziale dei dati.

Il processo legato al ML è partire da dati sconosciuti, di cui non si conosce neanche la distribuzione degli stessi, produrre un modello generale in grado di comprendere ed infine predire nuovi dati sconosciuti con un certo grado di accuratezza; essendo inoltre il futuro incerto si parla più di limiti probabilistici che di dati certi (Bishop, 2006). Inoltre, è bene tenere a mente che gli algoritmi giocano un ruolo secondario, perché se i dati sono scorretti o hanno parametri sbagliati o ancora non hanno un dato da predire, l'algoritmo darà un output povero in termini di accuratezza e pieno di rumore; quindi è necessario porre grande attenzione nella fase di modellazione dei dati e nello stilare l'ipotesi, uno schema semplice con i dati corretti lavora meglio di uno complesso (Engel, 2016).L'aspetto più importante del ML è la ripetitività, perché più i modelli sono eseguiti sui dati, più sono in grado di adattarsi in modo totalmente autonomo. I computer imparano dalle precedenti elaborazioni in modo tale da dare risultati e prendere decisioni che siano a mano a mano più affidabili; l'algoritmo testa più e più volte i dati in maniera autonoma e apprende i

comportamenti senza la necessità di un intervento umano sulle istruzioni da eseguire nei vari casi che il computer potrebbe dover affrontare (Vance, 2018).

È ovvio, però, che questa tecnologia non è esente da problematiche, una delle più importanti è la presenza di qualche tipo di Bias cognitivo, che dipende dal campo applicativo, cioè valori ed errori che dipendono non dall'algoritmo in sé ma di che scrive il codice o di chi raccoglie i dati o come si fa; se per esempio, se si crea un algoritmo per assumere personale in base ai dati pubblici americani su istruzione ed assunzioni si creano problemi razziali e di genere, perché i dati raccolti sono relativi per lo più a uomini bianchi perché sono la maggioranza, ma non necessariamente migliori (Reuters, 2018). Si hanno inoltre, problemi sul reperimento di sufficienti informazioni per il training set, infrastrutture vecchie e non adatte a gestire la mole di dati necessaria, problemi di privacy per alcune tipologie di dati, comprendere come alcuni modelli lavorino realmente, come le reti neurali.

Due, i principali filoni di ricerca ed applicazione del machine learning, già identificati dallo stesso Arthur Samuel nel suo elaborato degli anni '50, che permettono di distinguere l'apprendimento automatico in base al fatto che il computer riceva dati completi da utilizzare come indicazione per eseguire il task necessario (apprendimento supervisionato) oppure che si lasci lavorare il software senza alcun "aiuto" (apprendimento non supervisionato) (Boldrini, 2018)

Nel primo caso vengono forniti al modello sia i dati veri e propri che i risultati desiderati per ciascuno di essi e si lascia poi al computer trovare la regola o funzione generale che lega le informazioni all'output voluto; inoltre, si può valutare precisamente l'accuratezza del modello, a condizione di conoscere l'obiettivo ricercato. Una volta scoperta la regola matematica che sta dietro, si può facilmente riapplicare a casistiche simili quante volte si vuole (Geitgey, 2014). Ci sono svariati algoritmi che si possono adattare a questa tipologia di problemi, i fattori da tenere in considerazione quando se ne sceglie uno sono: la dimensionalità dei dati e la loro eterogeneità, la quantità di dati necessari per il training, la presenza di rumore e ridondanza, e altre difficoltà.

Invece quando si parla di apprendimento non supervisionato, si tratta di problemi dove non si ha la classe di appartenenza dei dati di training, bensì è l'algoritmo che deve capire le relazioni tra gli elementi senza nessun intervento umano, come nel caso dei motori di ricerca (es. Google). Ci sono anche qui svariati algoritmi possibili, ma tutti ricercano legami ed informazioni nascoste tra i dati, analizzando similarità e differenze tra i punti per trovare una struttura che li colleghi. L'apprendimento non supervisionato viene soprattutto utilizzato in casi di clustering e reti neurali e nei casi in cui si debba analizzare la densità dei dati, o ancora per determinare variabili latenti, come il metodo dei momenti, dove parametri sconosciuti vengono legati ai momenti dei dati (variabili statistiche come la media o la covarianza) (Anandkumar, 2014).

Si possono poi citare altre tipologie di machine learning oltre a queste due principali: una è l'apprendimento per rinforzo dove si decide come due agenti devono agire in un ambiente per massimizzare una funzione obiettivo; è il problema tipico della Teoria dei Giochi ben spiegato dal dilemma del prigioniero, dove due persone devono decidere se accusare l'altro del crimine commesso in base ai vari pay-off che possono ottenere. Differisce dalle due categorie sopra descritte perché non è presente la corretta assegnazione input/output, l'obiettivo è il miglioramento delle performance cercando un equilibrio tra tempo di esplorazione di soluzioni sconosciute e sfruttamento della conoscenza pregressa (Busoniu, 2010).C'è poi una via di mezzo tra l'apprendimento supervisionato e quello non, che corrisponde al semi-supervisionato; con questo metodo si usa un modello che sta a metà, dove al calcolatore viene fornito un dataset incompleto per l'addestramento; alcuni di questi input possiedono i rispettivi output (come nell'apprendimento supervisionato), altri invece no (come in quello non supervisionato). L'obiettivo, alla fine, non cambia: identificare regole e funzioni che legano i dati per risolvere il problema posto, oltre a trovare strutture di dati atti a raggiungere gli obiettivi prefissati. (Boldrini, 2018).

Infine, spesso per classificare le sottocategorie di Machine Learning, si usano gli algoritmi stessi che vengono utilizzati, seguendo così un approccio più pratico per la classificazione del ML: si identificano le tecniche e i metodi di cui si è già parlato a lungo nella sezione dedicata al data mining e in parte riprese in quelle seguenti (reti neurali, clustering etc), a cui si possono poi aggiungere i modelli probabilistici (come la rete di Bayes che rappresenta in un grafo le variabili e le dipendenze a loro connesse) (Boldrini, 2018), le investigazione di anomalie, i rumori nei dati o devianze come può essere una frode bancaria ; i metodi da usare possono essere sia supervisionati che non (Hodge, 2004); sempre con gli stessi due metodi si possono trovare algoritmi di apprendimento delle caratteristiche, usati spesso come fase di pre-processing per analisi più avanzate come la classificazione. Questi metodi mantengono le informazioni iniziali contenut2 negli input ma le rendono utili e comprensibili per ulteriori approfondimenti (Bengio, 2013).

Queste sono solo alcune delle varie tecniche, modelli ed algoritmi utilizzati nel campo del Machine Learning, ma bastano per far comprendere quanto questo sia vario ed ampio, così come sono le sue possibili applicazioni. Quindi per un'azienda è lecito chiedersi, ma quando sono da utilizzare tecnologie di machine learning nel proprio ambito lavorativo? Quali sono le applicazioni più comuni oggi?

Nella maggioranza dei casi la risposta è no, non serve dotarsi di applicativi di ML, l'analisi di Big data è spesso sufficiente, così da non dover investire in tecnologie estremamente complesse come gli algoritmi di apprendimento automatico. Questi sono utili quando si sa precisamente cosa si vuole ricavare dai dati, ma si è incerti su quali siano le variabili importanti da usare come input per decidere, si usano quindi tecniche che imparano autonomamente cosa sia importane e cosa no

e possa estrarre informazioni utili. Casi tipici sono quelli dove si hanno anche centinaia di variabili che possono influenzare il sistema in esame, come il caso del sistema di raffreddamento dei data center di Google, dove algoritmi di machine learning sono stati usati per districarsi attraverso 120 variabili che erano tenute in considerazione dal sistema (McClelland, 2018). Di questa stessa tipologia di applicazione è la guida autonoma, dove ci sono tantissime variabili che il computer deve tenere in considerazione, provenienti dai più svariati sensori, gps, pressione sui freni, giri del motore, velocità telecamere per rilevare oggetti, persone e altri veicoli, oltre ad essere in grado di fare predizioni su eventi sconosciuti che possono accadere durante un viaggio in macchina, grazie anche alla raccolta da svariate macchine test già su strada (Romano, 2017).

Ci sono poi applicazioni meno legate al mondo reale, ma per cui il machine learning torna utile: esempi di questo sono Amazon e Netflix. In particolare, quest'ultima ha avuto la grande intuizione di utilizzare queste tecniche per il suggerimento agli utenti di film adatti a loro. Tale algoritmo, per elaborare una raccomandazione, non si limita a considerare i contenuti consumati in passato dallo spettatore stesso, ma implementa funzionalità di incrociamento di pattern tra spettatori simili per offrire raccomandazioni molto più personalizzate (fig. 3.5), informazioni sull'area geografica di proveniente del cliente, quanto tempo questo passi a leggere il riassunto di un contenuto, la velocità di scroll della pagine e molte altre informazioni provenienti sia dal sito che dall'ambiente reale dell'utente (Rayna,Striukova, 2016).

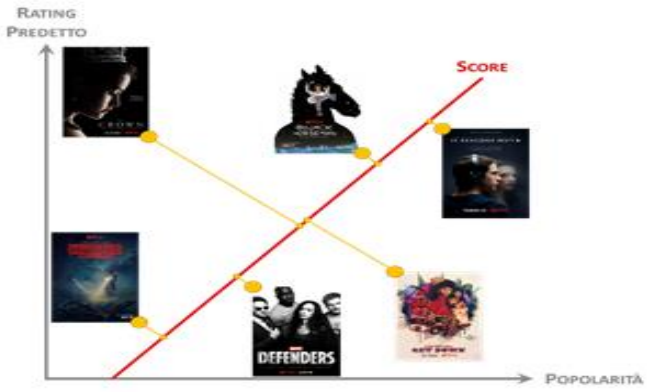


Fig. 3.5 Spazio definito dalla funzione di rating

Amazon, dall'altro canto, fa una cosa similare per poter suggerire ai propri clienti prodotti adatti a loro, perché altri utenti simili li hanno acquistati o perché inerenti alla loro cronologia ed al genere di cliente. Si hanno poi applicazioni curiose in campo della sicurezza informatica, non tanto per identificare le minacce comuni, ma per adattare le difese all'ambiente circostante e ad

attacchi portati fisicamente da persone esterne o da interni, in modo da proteggere il sistema da attacchi generati tramite ingegneria sociale, o per furti d'identità, per filtrare le mail ed altro ancora (Engel, 2016). Interessante da annoverare tra i possibili utilizzi del ML, c'è l'applicazione in campo economico-digitale, come il progetto di Akvelon per prevedere il valore di un dato mercato tramite le news provenienti da tutto il mondo; in particolare questo caso è stato realizzato sulla predizione del valore del Bitcoin, utilizzando esclusivamente software e tool open source che però è ancora in sviluppo, necessita di metodi di validazione e miglioramento dell'accuratezza.

Ovviamente non si esauriscono qui i possibili usi di una tecnologia tanto potente, però questa piccola esplorazione serve a far capire la flessibilità d'uso del Machine Learning e l'adattabilità a vari contesti, così che non ci si stupisca che si cerchi d'implementarla anche in contesti industriali e di manutenzione predittiva come si è proposti di fare nel progetto collegato a questa tesi.

4. Introduzione alla tematica di studio

4.1 Manutenzione Predittiva

La manutenzione predittiva (MP) è sicuramente una delle applicazioni più legate alla produzione manufatturiera di IoT e Machine Learning. L'obiettivo di questa pratica è stabilire un programma per effettuare correzioni ai macchinari ed ai robot della linea produttiva prima che questi si rompano effettivamente, così da non avere stop improvvisi nella produzione che possono portare a gravi danni economici (Mobley, 2002); in questo modo si può aggiustare il workflow del processo e adattarlo alle necessità, anche in base agli ordini che arrivano alla fabbrica ed avere il minor tempo possibile dedicato alla riparazione della linea; inoltre questo permette di comprendere meglio anche la vita residua dei macchinari sulla linea, così da eventualmente predisporre la sostituzione con i minori disagi possibili alla produzione.

Questa tecnica differisce dalle precedenti per vari aspetti; si discosta totalmente da manutenzioni di tipo correttivo, che si limitano semplicemente a riparare i macchinari dopo che il guasto è avvenuto cioè il metodo più semplice e con costi di solo intervento più bassi, che è, tuttavia, l'obiettivo opposto a quello che si prefigge la manutenzione predittiva, che ha come cardine prevenire i guasti. È poi differente dalla manutenzione preventiva perché questa si basa su analisi statistiche sulla media e l'aspettativa di vita, ed è inoltre effettuata regolarmente nel tempo, a date stabilite, invece la MP guarda all'effettivo stato di usura della macchina ed avverte prima che il guasto effettivamente si verifichi all'interno del processo. Infine la MP tramite IoT è differente dalla manutenzione predittiva manuale, poiché in questa sono degli operatori che vanno fisicamente a raccogliere dati a campione sui macchinari, riportati poi su fogli di calcolo per capire quando è necessario un intervento; invece la MP di nuova concezione è un processo non solo automatizzato, ma che connette fra loro le varie macchine e che prende in considerazione molti più parametri di quelli che potrebbe fare un persona e trova connessioni nei dati raccolti impossibili da scoprire altrimenti.

Quindi gli effetti sull'ecosistema della fabbrica, sia in termini di tempi che di costi che di processo produttivo, sono molteplici e variegati: porta ad un risparmio sui costi di manutenzione generali rispetto alla manutenzione preventiva di circa il 25-30%, grazie all'esatta individuazione del guasto e alla scelta del tipo di assistenza più adatta; i tempi di attività della macchina si possono allungare fino al 35-40% in più grazie a questa tecnica, poiché si riducono i tempi vuoti e si allungano i periodi tra una manutenzione e la sua successiva; infine la produttività può aumentare fino al 12% rispetto ad una strategia preventiva ed addirittura fino al 40% rispetto ad una reattiva;

queste risultanze stanno portando sempre più la manutenzione predittiva ad essere uno standard produttivo necessario per tutte le aziende competitive (Zubani, 2018).

Le applicazioni che necessitano di questa tipologia di controllo sono quelle critiche, cioè i macchinari più delicati o che non possono essere facilmente sostituiti, che in caso di rottura provocherebbero gravi danni alla produzione; inoltre devono essere monitorabili senza aumentare troppo i costi, perché altrimenti si perderebbe il senso di questa operazione (Goriveau, 2016). La MP, come si è già accennato, riduce i costi generali di manutenzione ma ha costi molto più elevati di quella preventiva per la sua diretta applicazione, sia in termini di costi di monitoraggio che di capacità ed esperienza necessarie perché sia sfruttata al meglio, perché, oltre ad essere specifica per ogni macchina, deve essere riadattata di volta in volta, e questo spesso porta le aziende a doversi avvalere di piccoli contractor specializzati (Fiix, 2018).

Per effettuare manutenzione predittiva serve, per prima cosa, l'individuazione di alcuni parametri chiave da tenere sotto controllo, tramite vari sensori posizionati sulla macchina scelta, con i quali si può misurare la corrente assorbita dal macchinario, la temperatura dei componenti critici (sia tramite termometri che immagini termiche), la velocità e l'accelerazione del motore, analisi di vibrazioni o del livello dell'olio, praticamente tutte le variabili che possono influire sul corretto funzionamento della macchina (Fiix, 2018). Tutto questo viene fatto in tempo reale, quindi in ogni istante si è a conoscenza dell'esatto stato della macchina, delle sue performance, se è in linea con i parametri prefissati. Si passa poi alla fase di raccolta dati in un sistema centralizzato, il computerized maintenance management system (CMMS), che si occupa della gestione della manutenzione, legando i dati al corretto oggetto monitorato e spesso questa fase può richiedere tempi molto lunghi (da settimane ad anche mesi) in base alla complessità della macchina che si sta analizzando; spesso su questo sistema, vengono implementati gli algoritmi necessari di data mining e ML, così che possano effettuare le analisi e le valutazioni dei dati; questi devono essere, alle prime iterazioni, addestrati su dati di partenza, anche per questo spesso ci si appoggia a terzi per fare MP, poiché queste aziende possono allenare i propri algoritmi su macchine simili ma in diversi contesti, rendendo così più breve e preciso il processo di apprendimento dell'algoritmo (ad esempio IBM, o ancora Amazon); una volta eseguito questo passaggio, vengono usate le varie tecniche di analisi, il sistema può, quindi, decidere quali parti del processo necessitino di interventi, che spesso vengono effettuati con i macchinari ancora in funzione, per ridurre al minimo l'impatto sulla produzione (Peng, 2012). Questo ultimo passaggio è particolarmente critico, poiché la rappresentazione dei dati è difficile e necessita di esperti del dominio per poter essere efficace nella misura più elevata possibile; non è solo un segnalare il bisogno di riparazioni della macchina, ma si specifica anche che tipo di intervento è necessario e quali siano stati campanelli di allarme che hanno fatto scattare il segnale di riparazione. La struttura dei passaggi per fare

manutenzione predittiva è schematizzata in fig. 4.1, dove si possono vedere riassunti i 4 passaggi appena descritti per implementare al meglio un sistema di manutenzione predittiva IoT.

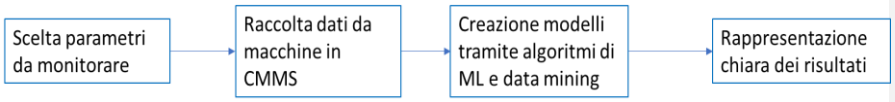


Fig. 4.1 passaggi per manutenzione predittiva

E' dunque possibile controllare molte variabili diverse tra loro e per farlo si possono usare varie tecnologie di tipo non distruttivo ovvero che non compromettono l'integrità dell'elemento processato, come lo studio delle vibrazioni, adatta a macchinari che lavorano ad alte velocità, e che è utile per valutare con grande precisione lo stato di usura dei componenti ed ha un buon livello di automazione, anche se risulta una delle tecniche più costose da implementare e necessita sempre di una persona fisica per interpretare i risultati alla luce della teoria sulle vibrazioni (Yung, 2006). L'ispezione visiva da remoto è tra le prime ad essere stata applicata; si osserva il pezzo con una luce sufficientemente potente ed è poco costosa. Analisi acustica tramite ultrasuoni permette di percepire le frizioni e lo "stress" posto sui macchinari sottoposti a importanti rotazioni; si usano gli ultrasuoni perché possono esserci onde non udibili da orecchio umano e il sistema li separa da quelle a bassa frequenza o dal rumore tipico della macchina in funzione; un sistema così sofisticato che può distinguere/rilevare, tramite opportune analisi di quale tipo di malfunzionamento si tratti, rottura, mancanza di lubrificante etc (Kennedy, 2006). Per il monitoraggio dei componenti c'è un'altra tecnologia a disposizione: l'analisi con infrarossi, che prevede, basandosi sui campioni di olio, non solo lo stato di lubrificazione, ma anche la qualità del materiale usato e se è adatto al macchinario permettendo di individuare problemi sia meccanici che elettrici, in tipologie di macchinari diverse. Infine, le analisi sull'olio possono essere le più efficaci, ma sono a lungo termine e possono volerci anni per raggiungere livelli alti di efficienza (Robin, 2006).

La manutenzione predittiva permette di effettuare diverse tipologie di analisi (Della Mura, 2018):

- **Trend**: se ci sono o meno, se il processo è in controllo (come nelle carte di controllo utilizzate in Ingegneria della Qualità, che utilizzano però metodi puramente statistici, è di tipo preventivo non predittivo)
- **Relazioni**: le cause principali che portano ad un guasto ed eventuali connessioni fra loro
- **Segmentazione:** analisi di clustering sui dati
- **Associazioni**: applicazione delle regole di associazione, per capire la correlazione tra i dati

- **Anomalie**: individuazione di anomalie, o punti fuori controllo e comprendere se sono errori nella misurazione o se c'è stato uno spostamento dello stato del processo

Grazie quindi alla grande flessibilità di tecniche utilizzabili per il monitoraggio, si possono acquisire dati in diverse realtà produttive, rendendo molto ampio il range applicativo della manutenzione predittiva; i casi più comuni sono: sulle linee ferroviarie, per prevenire incidenti, come nel settore aerospaziale, dove si stima che quasi il 61% delle compagnie aeree investiranno in programmi di MP nel 2019 (Aliperto 2018), in quello dell'energia, nelle centrali elettriche come negli impianti di estrazione di petrolio e gas, nei mercati specifici come quello degli ascensori e, come già detto in quello manifatturiero,. Si ha poi la creazione di sistemi di manutenzione predittiva generali, applicabili a varie casistiche, come il SIMAP (Intelligent System for Predictive Maintenance) che agisce come CMMS, raccogliendo i dati da vari sensori per poi elaborarli e verificare che la macchina operi in condizioni normali per la produzione; è un sistema che è stato applicato con grande successo in impianti eolici, creando un calendario di manutenzione ottimale in base ai reali bisogni di riparazioni, alla vita operativa della turbina eolica così come ad altri criteri economici e tecnici (Garcia, 2006).

Ovviamente non tutti i settori e neanche tutte le fabbriche sono pronte ad utilizzare una tecnologia come questa, poiché necessita di grandi investimenti, non solo nel sistema stesso e nella sua gestione, ma anche nell'ammodernamento dei macchinari utilizzati nel processo produttivo, che devono essere all'avanguardia per poter essere correttamente monitorati dai sensori. Tuttavia è una tecnologia che ben si adatta ad un problema come quello studiato in questa tesi, il citato braccio robot, poiché si parla di macchinari altamente tecnologici, progettati per aver molti sensori di serie e che necessitano di alta precisione per continuare a lavorare secondo gli standard dettati; ecco quindi che entra in gioco un sistema di manutenzione predittiva atto a mantenere i livelli produttivi della macchina alti, cercando di capire quando il braccio robot necessita di interventi alla cinghia di trasmissione. Questo è l'argomento della prossima sezione, che introduce alla problematica specifica del caso di studio.

4.2 Belt Tensioning

La cinghia (belt in inglese) è un organo per la trasmissione di potenza meccanica tra due alberi, tramite l'uso di pulegge innestate su questi ultimi. Tipicamente si ha un lato della cinghia molto teso (in tensione) ed uno poco teso (ramo condotto).

Le cinghie permettono una trasmissione della potenza costante e a basso rumore, oltre a permettere l'assorbimento di urti e variazioni improvvise del carico, che potrebbero andare a danneggiare i motori ed i supporti. Tuttavia, per la natura dei materiali utilizzati (tipicamente cuoio, fibre tessili o fili di nylon), presentano una rigidezza e una resistenza inferiori rispetto alle catene o agli ingranaggi, ma i recenti sviluppi nel campo delle cinghie di trasmissione permettono l'uso di queste ultime in applicazioni che storicamente erano esclusivamente per gli ingranaggi. Una cinghia di trasmissione richiede una manutenzione minima e nessuna lubrificazione, in più, ha elevata tolleranza al disallineamento ed efficienza (in media del 95%, fino al 98%), e il suo costo rimane contenuto, senza grandi aumenti in funzione della distanza fra i due alberi, al contrario delle trasmissioni ad ingranaggi. Il campo di temperatura varia usualmente fra -30 e 80° C. La trasmissione però, può variare considerevolmente alla comparsa di stati di usura dell'apparato, a causa dello slittamento e della distensione della cinghia (Crivelli, 2013).

La variazione negativa della tensione porta ad un peggioramento della trasmissione della potenza, quindi comporta minori performance della macchina e maggiori consumi di corrente, oltre a causare problemi come denti cavi o recisione dei denti in particolari tipi di cinghie; una tensione troppo elevata invece provoca usura dell'organo di trasmissione, è quindi necessario porre molta attenzione al problema. Lo slittamento, d'altro canto, anche se di piccola entità, porta ad avere sfregamento contro la calotta di protezione, il coperchio, con conseguente fusione della plastica. Il contatto con altre parti causa l'abrasione della cinghia: i bordi si sfrangeranno, assottigliando così la cinghia. La velocità del degrado dipende dal disallineamento e dal materiale contro cui sfrega; alternativamente, la cinghia di distribuzione può tagliarsi longitudinalmente, causando ovviamente l'arresto del motore (Gates, 2017). Lo spostamento laterale della cinghia può anche arrivare a un punto tale da causare, insieme a un'errata tensione (eccessiva o insufficiente), il surriscaldamento e il danneggiamento sia della cinghia sia della superficie di scorrimento in metallo del tenditore o della puleggia.

La potenza trasmessa si calcola tramite la differenza tra le tensioni della cinghia moltiplicate la velocità della stessa; le tensioni sono legate da una relazione che dipende sia dal coefficiente di attrito con il materiale che forma le pulegge, sia dall'angolo al centro (misurato in radianti) che formano con queste ultime.

Le cinghie possono essere di varie tipologie, che vengono anche utilizzate per classificare le possibili applicazioni:

- Cinghie piatte: sono cinghie poco spesse ma larghe, di forma rettangolare, tipicamente sono prodotte con fibre tessili o sintetiche; adatte a trasmettere potenza su lunghe distanze, poco costose e semplici da installare; tuttavia sono molto ingombranti e per questo sono state sostituite da motori elettrici o da cinghie a V.
- Cinghie dentate: è formata da una serie di denti nella parte interna e sull'esterno da una fascia continua, tenuti insieme con un'anima in acciaio; sono necessarie quando si deve trasmettere grandi quantitativi di potenza in estrema silenziosità e senza slittamenti e perdite di potenza.
- Cinghie trapezoidali o a V: ottimo compromesso tra potenza trasmessa, spazio occupato e velocità, inoltre risolvono i problemi di slittamento ed allineamento; hanno per l'appunto la forma di un trapezio, con la base minore rivolta verso l'interno, per fornire migliore aderenza alla puleggia e minor rumore; tipicamente sono fatte con un'anima di fili di nylon.

Il Robot che è stato utilizzato per le analisi esposte in questa tesi e da cui si sono ricavati i dati di corrente e posizione del motore, presenta problematiche inerenti a questa tematica. Infatti, si è studiato l'impatto sulle performance della macchina in termini di assorbimento della corrente in caso di maggiori o minori tensioni della cinghia, grazie all'inserimento di pulegge per aumentare o diminuire la tensione, così da contrastare eventuali segni di degrado nel macchinario.

Posto che il focus della tesi è l'analisi delle pulegge da inserire per gestire la tensione dell'organo di trasmissione si vuole qui approfondire la tematica delle difficoltà della tensione di una cinghia in termini generali. Infatti, al crescere della tensione, cresce anche lo stress (carico) sulla cinghia e sui cuscinetti. La cintura ideale è quella non scivola sotto carichi elevati alla tensione minima. Le tensioni dovrebbero, inoltre, essere commisurate al tipo di cinghia, alle dimensioni, alla velocità e alle caratteristiche della puleggia. Per esempio, le cinghie dentate hanno bisogno di una tensione sufficiente a mantenere la cinghia a contatto con la puleggia, in quelle trapezoidali serve per avere perfetta aderenza alle ruote collegate agli alberi di trasmissione. Soprattutto per le cinghie a V la corretta impostazione di questo importantissimo parametro è fondamentale nelle fasi di installazione dell'organo di trasmissione, da questo passaggio si determina il tempo di funzionamento futuro della cinghia; troppo poca tensione porta a slittamento, surriscaldamento e maggiore usura sui vari componenti (cinghie e pulegge); troppa, invece, è causa di sforzi eccessivi su cinghie, cuscinetti e alberi; tuttavia, esiste un'ampia gamma di tensioni per le quali la cintura continua a funzionare in modo soddisfacente. L'intento è quello di trovare questa rosa adeguata, indipendentemente dalla V-Belt scelta (IBT Inc, 2018).

Quindi si può ben comprendere come sia importante capire lo stato di una cinghia all'interno di un macchinario, in particolare in quelli ad alta precisione come il braccio robot qui considerato,

e si può fare tramite l'analisi delle correnti assorbite dalla macchina, poiché abbiamo visto che un peggioramento delle condizioni della cinghia porta a maggiori consumi di elettricità.

5. Analisi esplorativa

Tutte le analisi e le osservazioni che verranno esposte qui di seguito, sono state ricavate tramite l'utilizzo di script Python creati con l'ambiente di sviluppo Jet Brains PyCharm. La versione del linguaggio utilizzata è la 3.7; inoltre per effettuare tutte le ricerche necessarie sono state implementate varie librerie, messe a disposizione tramite open source dai singoli sviluppatori; questo è l'elenco completo delle librerie che sono state utilizzate nei vari passaggi del progetto:

- import matplotlib.pyplot as plt
- from math import sqrt
- import math
- import decimal
- from scipy.stats import kurtosis, skew
- import numpy as np
- import pandas as pd
- import csv
- from pymongo import MongoClient
- from sklearn import tree
- from IPython.display import Image
- from sklearn.externals.six import StringIO
- from sklearn.tree import export_graphviz
- from sklearn.model_selection import GridSearchCV
- from sklearn.model_selection import train_test_split
- from sklearn.tree import DecisionTreeClassifier
- from sklearn.metrics import accuracy_score
- from sklearn.metrics import confusion_matrix
- from sklearn import metrics
- from sklearn.metrics import silhouette_samples, silhouette_score
- from sklearn.cluster import KMeans
- from sklearn.ensemble import RandomForestClassifier
- from sklearn.neighbors import KNeighborsClassifier
- from sklearn.metrics import classification_report, confusion_matrix
- from sklearn.metrics import mean_squared_error

L'utilizzo di alcune di queste funzionalità verrà spiegato più in dettaglio ad ogni sezione specifica di analisi; in ogni caso si sono rivelati fondamentali gli import di alcune librerie, che sono state utilizzate in molti dei passaggi di questo progetto: le librerie numpy e math per accedere a funzioni matematiche già scritte; Panda per poter meglio gestire ed analizzare i dati, essendo una libreria fondamentale per rappresentare i dati in matrici ed eseguire in maniera chiara e facile varie operazioni, grazie alla grande varietà di funzioni disponibili (es. campionamento, estrazione di solo alcune colonne o di righe); infatti Panda è considerata una delle librerie fondamentali per chi fa analisi di dati tramite il linguaggio Python grazie alla sua versatilità e ottima esposizione dei dati, purtroppo non scala molto bene se cresce troppo il numero di record da analizzare; infine le librerie csv e matplotlib per poter salvare alcuni dei risultati ottenuti (tabelle nel primo caso e grafici nel secondo).

Per quanto concerne i dati veri e propri invece, questi sono stati ricevuti in formato json, per facilitare la gestione della grande quantità di dati fornita. Per la lettura di questo particolare formato è stato scelto di usare il database non relazionale MongoDB e le relative librerie; è stato utilizzato con brevi query per poter inizialmente capire la struttura e l'organizzazione dei dati, oltre vari dettagli che vengono ora descritti.

I file sono organizzati come un dizionario di Python, dove ogni elemento del dizionario rappresenta un ciclo macchina analizzato dai sensori disponibili; ogni ciclo ha come struttura la seguente:

```
{id,
robotId,
StartTime,
CycleTime,
RobotProgram,
sensorData: { Current: { UoM,
                         values: []
                       },
              Position: { UoM,
                          values: []
                        }
            },
numWashers,
currentFeatures: { Max,
                   Min,
                   Mean,
```

```
                Peak2Peak,
                Var,
            Std,
                Rms,
            Kurtosis,
                Skewness,
                Rms_LP,
                Rms_HP,
                Rms_LHP
            }
}
```

Tra i dati rilevanti in questo schema si nota: un robot Id che identifica su quale macchina sono state fatte le misurazioni, uno StartTime che indica data e ora (con i dettagli fino a i secondi) a cui sono state effettuate le rilevazioni, cioè i dati effettivamente misurati dai sensori, che sono corrente e posizione; sempre riguardo al tempo si aggiunge che ogni ciclo ha una durata di circa 24 secondi. È importante sottolineare che nelle analisi sono stati utilizzati i soli dati della corrente poiché quelli della posizione non sono le effettive posizioni del braccio robot ma quelle inviate dal motore come comando alla macchina. Dato fondamentale per lo sviluppo di questo progetto è quello contenuto in numWashers, che non è altro che il numero di rondelle che si possono inserire per aumentare o diminuire la tensione della cinghia, ha come valore massimo di tensione pari alla categoria 0 e minimo a quella corrispondente a 4; è molto importante poiché nelle analisi sarà utilizzato come label per i modelli di classificazione utilizzati, in pratica è il parametro che si è cercato di predire. È bene precisare che il valore 0 è un caso estremo non realistico in ambito industriale, quindi viene escluso da analisi con risvolto più pratico, al contrario del valore 4 che rappresenta il tensionamento minimo, che può essere considerato quando la macchina necessita di interventi sul numero di rondelle per ripristinare lo stato iniziale. Infine, si ha current Features, che ha al suo interno tutta una serie di variabili statistiche, che vengono però ricalcolate nella prima parte di analisi per controllare la coerenza con i dati della corrente.

I cicli di monitoraggio effettuati sul braccio robot sono stati raccolti su più giorni nel mese di ottobre 2018, precisamente 17 giorni compresi tra il 10 ed il 30 del mese in questione (vedere tab. 5.1), per un totale su tutti i giorni di 23833 cicli; quasi tutte le giornate (più del 75% dei giorni) hanno i primi cicli che vengono inizializzati intorno a mezzanotte, a parte il 4, il primo che inizia nel tardo pomeriggio, e 3 durante la mattinata, anche se ad orari diversi; inoltre a parte due cicli, quelli del 11, 18 e 27 ottobre che finiscono ad orari particolari, tutti gli altri finiscono praticamente

a mezzanotte del giorno successivo. È importante sottolineare questo fatto, poiché possono esserci diverse partenze in base alla temperatura esterna ed al fatto che l'ambiente sia o meno condizionato, cioè può esserci un fattore che influenza l'iniziale assorbimento di corrente in caso faccia più o meno freddo, come si vedrà in grafici successivi.

Data	Ora primo ciclo	Ora ultimo ciclo	Numero di cicli	Rondelle
10/10/2018	16.57.02	23.59.49	531	4
11/10/2018	00.00.37	17.48.26	1298	4, 3
12/10/2018	09.02.58	23.57.11	1123	1
13/10/2018	00.00.23	23.59.26	1805	1
14/10/2018	00.01.02	23.58.30	1803	1
15/10/2018	00.00.05	23.59.23	1764	1, 2
16/10/2018	00.00.59	23.58.25	1731	2, 0
17/10/2018	00.00.48	23.59.54	1730	0, 3
18/10/2018	00.00.42	08.54.22	670	3
22/10/2018	12.55.32	23.59.32	821	3
23/10/2018	00.00.20	23.59.24	1433	3
24/10/2018	00.00.11	23.59.15	1805	3
25/10/2018	00.00.03	23.59.24	1762	3
26/10/2018	00.00.12	23.59.43	1739	3, 2
27/10/2018	00.00.31	11.30.31	868	2
29/10/2018	08.32.57	23.59.29	1150	2
30/10/2018	00.00.16	23.59.26	1803	2

Tab. 5.1 informazioni sui giorni dei cicli macchina

Sempre dalla tabella 5.1, si può notare che non tutti i giorni hanno test eseguiti su un solo numero di rondelle, ma possono avere anche due serie di misurazioni diverse; come nel caso del giorno 11, in cui vengono raccolti dati sia per la categoria 4 che 3; analogamente per i giorni 15, 16, 17 e 26 (20% dei giorni disponibili). Questo influisce sui dati disponibili per ogni gruppo disponibile per le analisi, si è quindi deciso di riassumere in tabella 5.2 il numero di cicli a seconda del gruppo di rondelle.

NumWashers	Numero di cicli	% sul dataset totale
0	2392	10.03650401
1	5367	22.51919607
2	6212	26.06470021
3	8707	36.53337809
4	1155	4.846221625

Tab. 5.2 divisione dei cicli in base al gruppo di appartenenza

È facile capire da questi numeri che si hanno molti meno dati per i raggruppamenti 0 e 4, che insieme rappresentano neanche il 15% del dataset totali, in particolare il gruppo 4 ha poco più di un migliaio di record, questo è dovuto al fatto che il segnale uscente dalla macchina impostata alla classe 4 dovrebbe avere un comportamento molto simile al gruppo 2, almeno questo dovrebbe essere lo standard; si hanno invece moltissimi record a disposizione per il gruppo 3, che occupa oltre il 35% dei dati disponibili.

Quella descritta fino adesso è la macro-organizzazione dei cicli, ma come è stato già detto i due dati importanti sono numWashers che è già stata descritta precedentemente e la corrente. La rilevazione di quest'ultima variabile è stata appunto realizzata su cicli di circa 24 secondi ed è stato salvato il valore assorbito dal robot ogni 2 millisecondi, che si traduce in approssimativamente 22'000 osservazioni per ogni record del database; si è quindi voluto dare una rappresentazione della corrente utilizzata dalla macchina e questo è stato realizzato in maniera esplorativa, prendendo come esempi il quarto ed ultimo giorno dei test. Sono stati scelti questi perché partono alla stessa ora ed hanno un numero quasi uguale di cicli. Si è quindi proceduto a mettere su uno stesso grafico (separato per giorno) la corrente usata dal robot (asse delle ordinate) nel primo ciclo di ogni ora (andamento nel tempo sull'asse delle ascisse). Per esempio, per il giorno 13 (fig. 5.1), la prima parte di grafico rappresenta il primo ciclo delle 16 PM, la seconda parte il primo delle 17 PM e così via fino alla fine del giorno; analogamente per il giorno 30 (fig. 5.2).

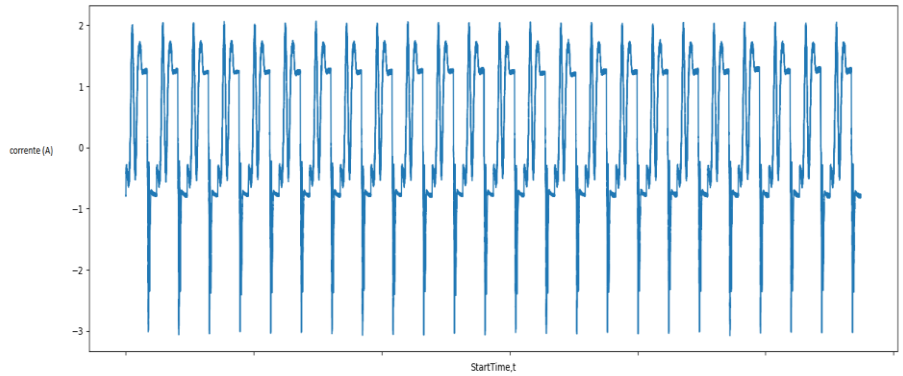


Fig. 5.1 ciclo corrente campionata ad ogni ora per il giorno 13/10/18

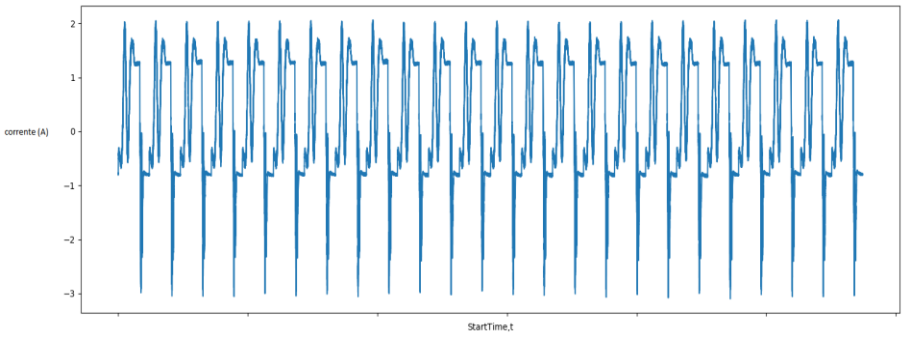


Fig. 5.2 ciclo corrente campionata ad ogni ora per il giorno 30/10/18

Da queste semplici rappresentazioni non si possono notare macro-differenze importanti tra i due giorni, con massimi e minimi molto simili tra i vari cicli ed andamento anche similare, nonostante partano con diverso numero di rondelle (rispettivamente 1 e 2). Inoltre, non si riescono a rilevare trend importanti all'interno del segnale, come ad esempio un degradamento del macchinario che si potrebbe associare a maggiori valori di corrente assorbiti. Se si guarda ai singoli cicli hanno una partenza con un valore negativo, per poi andare rapidamente a valori positivi, per infine raggiungere un picco negativo e ritornare a stabilizzarsi intorno ai valori iniziali; i valori negativi corrispondono a i momenti in cui la macchina compie movimenti per tornare alla posizione che viene considerata come origine, quindi quando il braccio meccanico torna indietro allo stato base.

Per approfondire maggiormente la conoscenza del dataset, sono state eseguite ulteriori analisi esplorative: prima tramite un campionamento stratificato non proporzionale per rappresentare la

media e la deviazione standard della corrente dei singoli gruppi, quindi non si tiene conto della reale percentuale degli strati all'interno della popolazione, poiché essendo una semplice fase di comprensione iniziale dei dati si è ritenuto fosse sufficiente un'analisi non troppo complicata; dopo con un'analisi eseguita su alcuni giorni scelti in maniera casuale.

La prima parte consiste nella rappresentazione di alcune statistiche calcolate su un campione; per la raccolta di quest'ultimo si è scelto di estrapolare 10 elementi casuali da ognuna delle etichette possibili, per un totale di 50 record totali che sono stati ordinati per data e non per gruppo di appartenenza.

Nel primo grafico, fig. 5.3, si può osservare la media dei singoli cicli estratti, differenziati per il gruppo a cui fanno capo.

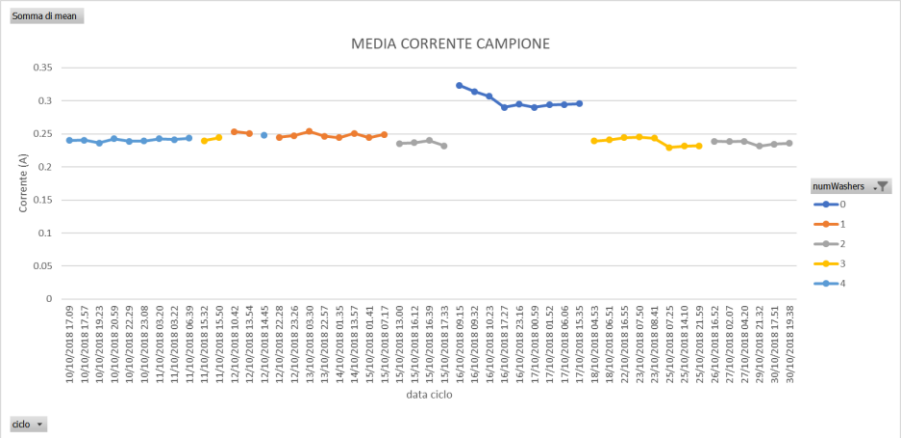


Fig. 5.3 media corrente del campione

Il gruppo 0 ha un comportamento molto differente da tutti gli altri come è facile notare per quanto riguarda la media della corrente assorbita, si vede inoltre che il gruppo 1 ha mediamente valori più alti dei gruppi rimanenti; per i gruppi 2, 3 e 4 si notano alcune piccole differenze per la terza etichetta, ma sono troppo piccole per poter generalizzare.

Nel secondo grafico, fig. 5.4, si osserva invece della media, la deviazione standard, sempre divisa per gruppi.

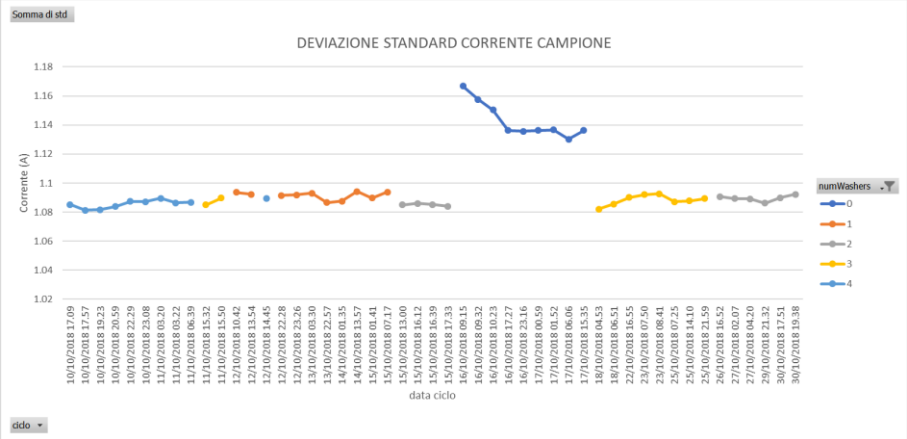


Fig. 5.4 Deviazione standard della corrente del campione

Come per la media il gruppo 0 continua ad avere un comportamento nei valori nettamente diversi dalle altre etichette, che invece si assomigliano molto tra loro; interessanti sono i trend crescenti nei valori di deviazione standard che si possono notare nei gruppi 2 e 3.

Nel terzo grafico, fig. 5.5, invece si vedono a confronto gli andamenti di massimo, minimo e media dei cicli del campione considerato, non più separati per numero di rondelle. Non emergono importanti differenze o trend, né crescenti né decrescenti, se non tra i primi dieci campioni che hanno minimi e soprattutto massimi leggermente superiori di quelli successivi; questi dati, infatti, corrispondono al gruppo 0, che continua a dimostrare una differenza sostanziale tra sé e tutte le altre categorie.

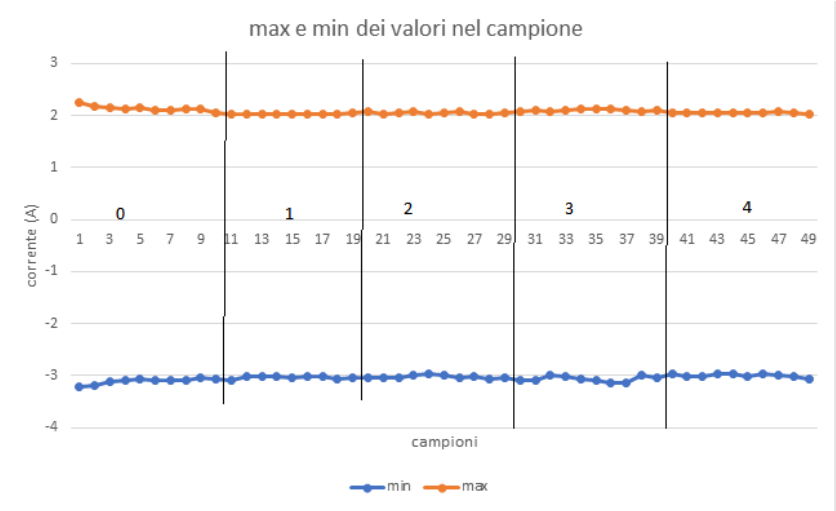


Fig. 5.5 Min e max del campione

Infine, sono stati creati i grafici di tutte le medie per ogni giorno, per cercare di osservare trend che segnalino eventuali degradazioni del macchinario. Sono stati scelti alcuni giorni di esempio, questi sono stati riportati in figura 5.6.a e 5.6.b.

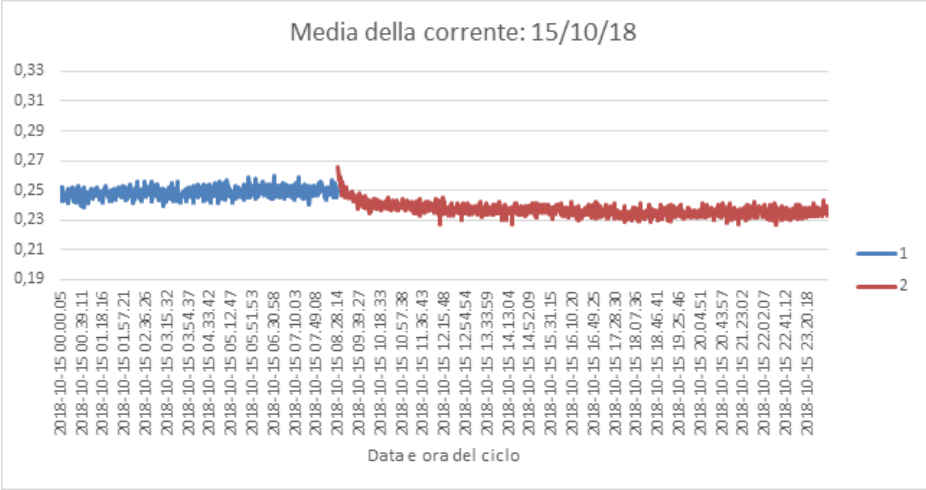


Fig. 5.6.a media corrente giorno 15 ottobre 2018

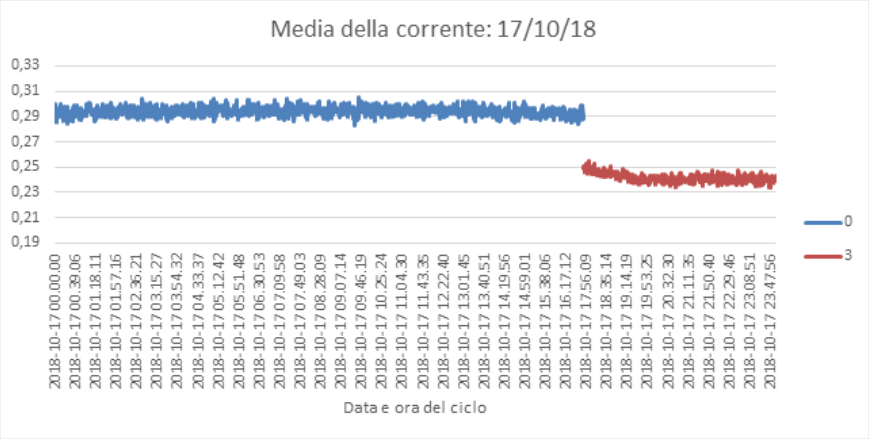


Fig. 5.6.b media corrente 17 giorno ottobre 2018

Nel primo grafico si hanno elementi dal gruppo 1 e 2, nel secondo invece di 0 e 3. Per prima cosa si notano le differenze macroscopiche tra i gruppi nella stessa figura; infatti, oltre al già citato comportamento anomalo della categoria 0, si osserva anche che il gruppo 1 ha valori mediamente più del 2 nel giorno in questione, trend che si può vedere anche in altre giornate anche se meno marcatamente. In secondo luogo, si vuole porre l'attenzione su un fenomeno importante che verrà poi ripreso nelle analisi successive, cioè che quando si ferma la macchina per il tempo di setup necessario, per esempio nel primo grafico per impostare la macchina con due rondelle invece che una, i primi cicli assorbono più corrente di quando è a regime, questo è dovuto al fattore della temperatura; poiché la macchina si è raffreddata nel periodo di fermo, avrà bisogno di alcuni cicli macchina per poter tornare alle condizioni ottimali di utilizzo. Questo fatto verrà poi utilizzato successivamente per fare una piccola pulizia dei dati in una specifica analisi.

6. Analisi del caso di studio

In base alle risultanze esplorative si è proceduto con la fase di analisi vera e propria; questa è stata eseguita in vari passaggi che sono stati riassunti in fig. 6.1: per prima cosa si è effettuata una fase di pre-processing divisa in due parti distinte, una di calcolo di vari indici statistici ed una di selezione delle features maggiormente significative per la comprensione dei dati originali; dopo ciò, si è voluto spezzare il lavoro di analisi in due filoni distinti:

il primo è inerente allo studio del caso tenendo in considerazione tutti i gruppi disponibili, sui quali vengono addestrati i modelli di classificazione e valutati per diverse percentuali di dati nel training e test set; questa sezione ha come obiettivo scoprire il numero minimo di cicli su cui è necessario fare il training per ogni classificatore, fattore estremamente importante per una fabbrica, poiché indica dopo quanto tempo è utilizzabile il modello per effettuare le analisi di manutenzione predittiva.

nel secondo invece, si selezionano solo alcune classi di numWashers sui quali vengono addestrati i modelli; si predicono poi i dati provenienti dai gruppi lasciati fuori per verificare se gli algoritmi sono in grado di individuare un degrado, questo tramite l'utilizzo di vari indici di dispersione.

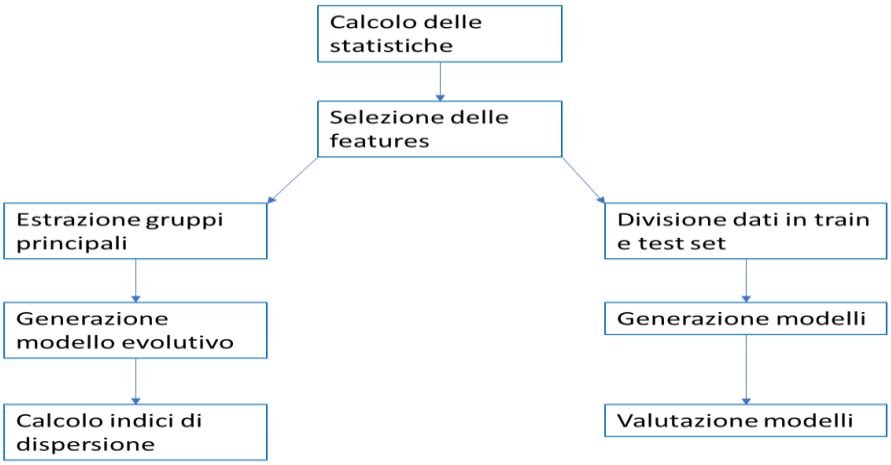


Fig. 6.1 architettura analisi

6.1 Pre-Processing

A seguito dell'analisi esplorativa, sono emersi due fatti rilevanti: il primo, che la mole di dati originali non permetteva analisi rapide, in quanto le funzioni messe a disposizione dalle librerie di Mongo DB richiedono molta memoria e potenza di calcolo, perché ogni volta che si esegue una query lo script salva sulla RAM tutto il file json di circa 6 GB; la seconda osservazione importante è che, come si evince dai vari grafici, le differenze nette tra i vari gruppi non si notano tanto sul segnale puro della corrente, bensì sulle features (cioè le variabili statistiche) calcolate, come media, minimo, massimo etc. Si è quindi deciso di passare ad una fase di lavorazione dei dati divisa in due parti: una di trasformazione ed una di riduzione di variabili.

Nella prima parte si sono ricalcolate le statistiche che erano state fornite insieme ad ogni ciclo con alcune modifiche:

Media	Skewness
Deviazione standard e varianza	Root mean square (RMS)
Massimo e minimo	RMS low – pass
Primo quartile, mediana, terzo quartile	RMS high – pass
Kurtosis	RMS band – pass

Non viene più calcolata la variabile peak2peak, ma vengono aggiunti i quartili al 25%, 50% e 75%. Si sono mantenute le nuove variabili calcolate poiché sono emerse incongruenze con quelle fornite insieme ai dati originali. Riguardo agli indici meno comuni si ricorda che: la kurtosis studia la forma della distribuzione dei dati, in particolare quella delle code; la skewness (skew abbreviata) riguarda la simmetria della distribuzione di una variabile casuale rispetto alla sua media; il RMS è sempre la deviazione standard ma calcolata tenendo conto della media nelle osservazioni. Questo indice di dispersione viene poi calcolato applicando vari filtri di dati utilizzati specificatamente per l'analisi di segnali di correnti: il filtro passo basso (LP) che permette il passaggio dei segnali elettrici solo sotto una soglia determinata; il suo complementare è il filtro passo alto (HP); infine si è calcolato l'RMS (Root Mean Square) per il filtro passa banda (BP), che è una combinazione dei due filtri appena descritti, che filtra esclusivamente i segnali che sono all'interno di una specifico range.

Tutti questi indici tuttavia, non sono stati utilizzati solo sul segnale intero iniziale, bensì si è eseguito un ulteriore passaggio di elaborazione per comprendere meglio quali parti del segnale

fossero realmente importanti: infatti ogni ciclo è stato spezzato in 24 sezioni temporali che, come è stato accennato precedentemente, corrispondono a blocchi di circa un secondo (si ha una rilevazione dell'amperometro ogni 2 millisecondi per circa 12'000 segnali). Ogni segmento comprende quindi 495 elementi, divisi come segue:

```
{'key': (0, 495)}, {'key': (495, 990)}, {'key': (990, 1485)}, {'key': (1485, 1980)}, {'key': (1980, 2475)}, {'key': (2475, 2970)}, {'key': (2970, 3465)}, {'key': (3465, 3960)},{'key': (3960, 4455)}, {'key': (4455, 4950)}, {'key': (4950, 5445)}, {'key': (5445, 5940)}, {'key': (5940, 6435)}, {'key': (6435, 6930)}, {'key': (6930, 7425)}, {'key': (7425, 7920)},{'key': (7920, 8415)}, {'key': (8415, 8910)}, {'key': (8910, 9405)}, {'key': (9405, 9900)}, {'key': (9990, 10395)}, {'key': (10395, 10890)}, {'key': (10890, 11385)}, {'key': (11385, 11880)}
```

Infine, per ognuna di queste partizioni sono state calcolate le statistiche sopra elencate, portando così il numero di variabili a 352, se si contano anche le colonne corrispondenti al numero di rondelle e a quella data più ora. Tuttavia, un tale numero di features può portare al problema dell'overfitting, considerando anche il fatto che molti dei segmenti creati hanno segnale poco caratteristico e simile a quello di tutti gli altri (vedi fig.5.1 i segnali hanno oscillazioni molto simili), indipendentemente dal gruppo di appartenenza.

Questo porta alla seconda fase di elaborazioni preliminari sui record che consiste in una selezione delle features più significative. Provati diversi metodi per fare questa operazione, alla fine si è scelto di utilizzare la matrice di correlazione che serve per stabilire quanto siano simili due variabili procedendo perciò per coppie ed il range di valori ammissibili è [-1; 1]. Si è quindi creata una matrice 351x351 (escludendo ovviamente la variabile della data) per proseguire poi con il calcolo della media per righe di questa tabella prendendo il valore assoluto delle varie celle, dove però non sono stati tenuti in considerazione gli elementi sulla diagonale maggiore della matrice poiché rappresentano la correlazione di una variabile con sé stessa.

L'operazione appena descritta permette quindi di individuare le variabili che contengono poca informazione e di conseguenza eliminarle, o per meglio dire l'informazione contenuta in quella variabile è già spiegata da altre variabili maggiormente significative e quindi le prime sono ridondanti. In figura 6.1 si visualizza il grafico di Pareto per i valori della distribuzione, poiché inserire la tabella completa era di difficile comprensione, si è voluto mostrare la distribuzione dei valori; come passo per i vari range si è scelto il valore 0,08, perché permette di finire quasi esattamente al valore massimo di 0,65, partendo dal valor minimo di circa 0,02.

Caso interessante il numWashers ha una correlazione di poco superiore a 0,5 (0,54 circa), per questo motivo per effettuare le analisi successive è stata appositamente reinserita nelle matrici dei dati, essendo questa feature l'etichetta di classe, quindi fondamentale per proseguire.

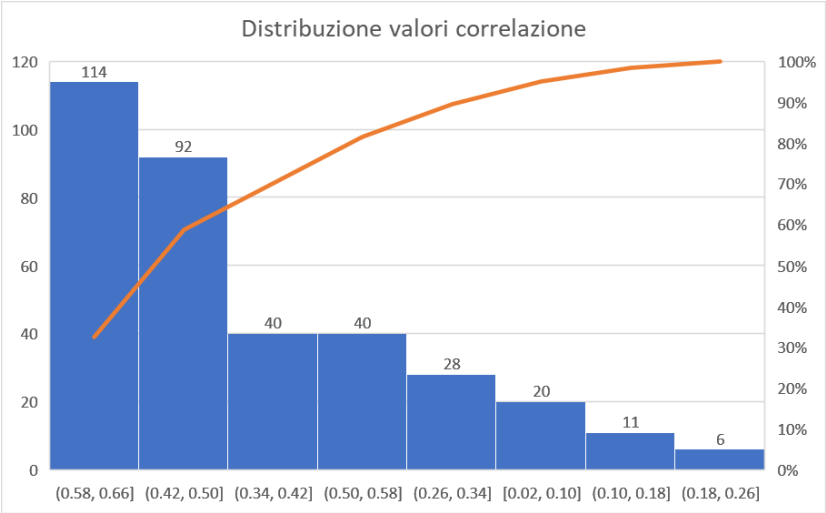


Fig. 6.2 distribuzione di Pareto per i valori di correlazione

Inizialmente la soglia di eliminazione era posizionata al valore 0,8 per la correlazione media, tuttavia, dato che il valore massimo trovato, si aggira intorno a 0,65, si è scelto di abbassarlo a due possibili soglie: la prima più alta a 0,5, la seconda più bassa a 0,3. Si può comprendere la scelta di questi valori sempre da figura 6.2; la soglia più alta permette una prima scrematura ma senza perdere troppe variabili (se ne perdono circa il 40%); la seconda soglia invece elemina quasi tutte le statistiche calcolate, infatti la cumulata dei valori sopra a 0,3 corrisponde a circa l'80%. Quindi con il valore 0,5 rimangono 198 colonne a cui si aggiungono numWashers e data dei cicli; nel secondo in caso il numero di colonne cala drasticamente a 48 features disponibili. È interessante notare che per questa matrice di dati ridotta si può facilmente vedere quali sono le variabili statistiche maggiormente presenti, in particolare si nota la presenta di molte skewness e kurtosis dei vari segmenti del segnale, per quanto riguarda invece le variabili calcolate sul segnale intero risultano con una correlazione abbastanza bassa solo il minimo, RMS_LP e RMS_BP. È stata creata una terza matrice, combinando la matrice con valore 0,5 con un metodo di estrazione di features basato sull'albero di decisione; grazie a questo metodo sono state estratte le 4 features che l'algoritmo ritiene più significative per spiegare i dati e sono state salvate in una matrice a parte per poter eseguire alcuni test di confronto con la matrice base. Le 4 variabili identificate sono:

skew(1980. 2475) RMS_LP(8910. 9405) kurtosis(7920. 8415) skew(2475. 2970)

Sono tutte variabili inerenti a segmenti e non al segnale generale, di cui due skewness di segmenti adiacenti; interessante notare che sono tutte variabili con valori di correlazione compresi tra 0,3 e 0,45, nessuna di quelle selezionate è tra le prime dieci features per valori più bassi rispetto alla correlazione. Nella figura 6.3 si osserva un grafico in due dimensioni (2D) delle prime due variabili più importanti, la skew (1980. 2475) e il RMS_LP(8910. 9405).

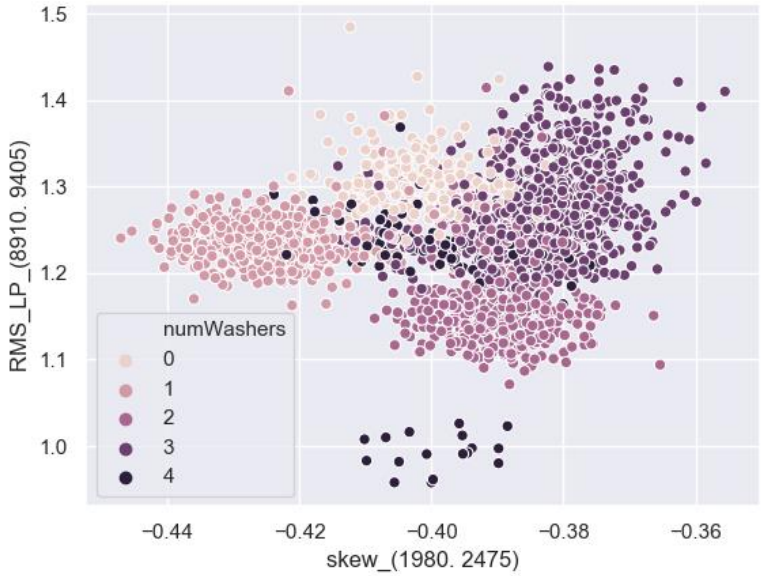


Fig. 6.3 rappresentazione grafica prime due variabili più importanti

Grazie a questa figura si evidenzia la buona separazione tra i primi 4 gruppi, da 0 a 3 rondelle, per poi notare la minor coesione del gruppo 4, che in parte si perde tra gli altri gruppi ed una parte è rappresentato da outlier esterni a tutti gli altri punti. La difficoltà nella rappresentazione di questo raggruppamento si ritroverà anche nelle analisi successive.

Infine si è voluto creare un modello aggregato dei dati tramite il metodo della PCA; si ricorda che quest'ultimo è un metodo di aggregazione delle variabili iniziali, eseguendo calcoli per riassumere in poche features la moltitudine che si ha all'inizio; questo si fa per ottenere una più facile rappresentazione, anche se in questo modo si perdono le informazioni relative a quali variabili sono contenute in quelle nuove; nel caso in esame si è scelto di riassumere le variabili iniziali sempre della matrice a 0,5 in 3 variabili, così da poter essere facilmente visualizzabili su grafici

2D/3D. In figura 6.4 si osserva per l'appunto, un grafico due D avente sull'asse X la variabile denominata semplicemente 0, e sull'asse Y la 2.

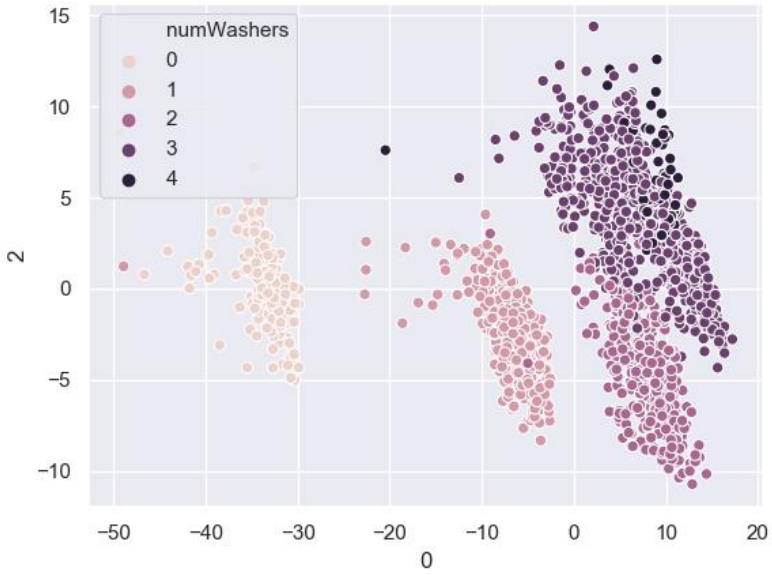


Fig. 6.4 rappresentazione grafica di due variabili ottenute tramite metodo PCA

Si mostra solo il grafico inerente a queste due variabili perché è quello che meglio permette di notare una divisione abbastanza netta dei cluster in analisi, come per il grafico relativo alle features più importanti; infatti tramite la lettura della legenda si nota che tutti i gruppi a parte il 4, sono abbastanza ben separati, a parte per alcuni outlier visibili vicini ai gruppi 0 ed 1. In fig. 6.5 invece, si può osservare il grafico 3D delle 3 variabili risultanti della PCA. Anche in questa figura si può osservare la netta distinzione tra i gruppi, in particolare tra 0 e tutti gli altri e che la classe 4 si mischia con maggiore frequenza con il gruppo 3.

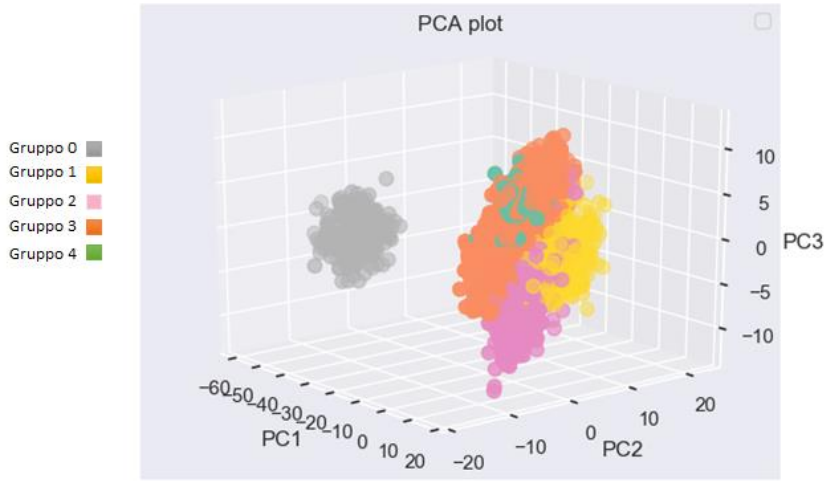


Fig. 6.5 grafico 3D delle variabili della PCA

Come matrice base per tutte le analisi che sono state effettuate si è scelto di utilizzare la matrice tagliata al valore di 0,5 per la correlazione, tutte le altre sono state utilizzate sono in alcuni casi come confronto e non verranno riportate in questo elaborato poiché non hanno portato a risultati significativamente diversi; è stato scelto questo valore poiché permette l'eliminazione di un buon numero di features dai dati iniziali (circa il 43% viene escluso con questo valore) ma non esclude variabili che anche se non vicine allo zero, possiedono un buon grado di informazione al loro interno, come invece accade scegliendo la matrice tagliata a 0,3 o quella con le features più importanti.

6.2 Modello Generale

Il primo passo che segue il preprocessing del dataset è la costruzione dei modelli da utilizzare per la manutenzione predittiva del braccio robot in oggetto. La tecnica per effettuare questo processo è, come ampiamente descritto, la classificazione. L'obiettivo dei classificatori impiegati è quello di utilizzare il numero di rondelle come etichetta identificativa della classe, per poi predire gruppi di test prelevati dai dati in possesso, quindi con una propria etichetta, ma che vengono considerati come dati sprovvisti di questa variabile, ovvero considerati come dati nuovi. Lo scopo dei classificatori è quello di comprendere se il macchinario sta funzionando esattamente come dovrebbe (quando l'etichetta predetta corrisponde a quella effettiva) o se sta assumendo un comportamento che devia dallo standard creato dal modello (ovvero quando la predizione si discosta dai valori precedentemente analizzati). In questo caso è necessario intervenire sul numero di rondelle ed indagare sul difetto che provoca il funzionamento anomalo.

Durante i diversi tentativi di analisi effettuate è stato identificato come modello generale l'insieme di classificatori che, preso in input un train set con dati sui cicli aventi tutte le configurazioni possibili di "numWashers" (da 0 a 4), eseguono predizioni sui dati rimanenti. La qualità dei modelli è stata valutata confrontando i risultati delle predizioni con i dati appartenenti a questo dataset di test (i dati non compresi nell'addestramento) per capire se i vari cicli sono stati predetti nella classe corretta. Il vero obbiettivo finale del modello generale tuttavia, non era quello di creare un modello di predizione fine a sé stesso, ma quello di cercare la quantità minima di dati necessari per crearne uno "buono", cioè in grado di predire correttamente i dati in input. Questo è molto importante in casi applicativi reali poiché nella realtà non è possibile sapere quanti dati si possano avere a disposizione né se questi siano esaustivi delle casistiche possibili (di norma eventualità rara) tanto che si considera di avere un buon modello quanto questo è in grado di adattarsi a situazioni sconosciute.

Tra gli algoritmi di classificazione esistenti (di cui è stata fornita una descrizione nel capitolo terzo di questa tesi) è stato scelto di utilizzare due classificatori messi a disposizione dallo stato dell'arte: l'albero di decisione e il K-Nearest Neighbor per elaborare il modello generale. Il motivo di questa decisione è stata dettata necessità di avere per le analisi un modello interpretabile, come è l'albero di decisioni, ma anche un metro di confronto che utilizzasse una base diversa per il calcolo predittivo, da qui la scelta del k-nn; inoltre entrambi sono algoritmi estremamente veloci nei calcoli, in particolare nella validazione e conseguente creazione del modello poi utilizzato

nelle predizioni; tutto questo senza compromettere la bontà dei risultati, infatti entrambi i classificatori permettono di avere buoni risultati nei parametri di valutazione.

Per quanto riguarda l'albero di decisione, non si è creato direttamente il modello di predizione, ma si è aggiunto il passaggio di validazione di quest'ultimo, grazie alle librerie messe a disposizione da Scikit-learn. La scelta dei parametri del L'Albero di Decisione è stata fatta tramite l'utilizzo di "GridSearchCV" con lo scopo di facilitare la ricerca dei parametri migliori da inserire nella programmazione dell'albero di decisione. Questa è una funzione molto utile poiché permette di utilizzare il metodo della cross-validation in maniera rapida ed efficace, basta passare due liste di parametri (i param_grid) ed infine impostare il parametro 'cv' al numero di sezioni che si vogliono creare (tramite un campionamento stratificato) e ognuna di queste, a turno, svolge il ruolo di test set e le altre come set di addestramento; il numero selezionato per le analisi di questa tesi è 10 e sarà utilizzato per tutte le analisi esposte da questo punto in poi. Invece, per quanto riguarda param_grid i due parametri da selezionare per l'albero di decisione sono i seguenti:

- max-features: numero massimo di variabili usate per cercare la miglior divisione possibile sui vari nodi degli alberi: può essere sia un valore intero che decimale, oppure assumere i valori standard "sqrt" e "log2", "auto"
- 'max_depth' (massima profondità): valore intero che indica il numero massimo di nodi che si possono attraversare partendo dal nodo radice per arrivare al nodo foglia più lontano

Tra i valori passati in queste due liste vengono scelti, alla fine della validazione del modello, i migliori parametri per ciascun insieme e questi verranno poi utilizzati per creare un modello di albero di decisione. Per il numero massimo di features, dopo numerosi test, la tecnica più efficiente scelta è stata quella di usare la radice quadrata. Per quanto riguarda il secondo parametro, la massima profondità, la validazione tendeva a selezionare sempre i valori più alti tra quelli proposti. Per questa ragione, sono stati cercati dei numeri sufficientemente alti in modo da ottenere prestazioni accettabili dall'algoritmo ma che comunque non compromettessero le sue performances; quindi la massima profondità degli alberi è stata impostata quasi sempre a 20. Per quanto riguarda il k-nn invece, si è scelto di non eseguire cross-validation poiché non si è ritenuta necessaria dati i valori sufficientemente buoni ottenuti senza validazione; per il modello generale si è scelto di usare k (il valore di punti vicini da considerare per classificare l'elemento che viene analizzato in quel momento) uguale a 10.

I risultati presentati in questa sezione sono solo quelli più significativi ai fini delle analisi effettuate; saranno descritti tramite le cosiddette matrici di confusione, ovvero tabelle ove su ogni riga

si trovano i valori reali di una classe e su ogni colonna invece, i corrispettivi valori predetti. In questo modo è possibile visualizzare i dati classificati correttamente sulla diagonale principale mentre nelle altre celle ci saranno i dati non predetti correttamente. Oltre a queste matrici, per ogni esperimento, sono riportati i valori di accuratezza, richiamo e precisione per ogni gruppo.

Tramite questi indici vengono quindi valutate le matrici di confusione di conseguenza i modelli ad esse connessi:

- L'accuratezza si calcola come rapporto tra numerosità di dati classificati correttamente e numero totale di dati presenti nel dataset usato per il test.

- Il richiamo e la precisione sono calcolati per ogni classe. Il primo indice è calcolato come il rapporto tra i cicli classificati correttamente in una data classe e quelli totali appartenenti al medesimo gruppo; il secondo valore, invece, è il rapporto tra cicli classificati in maniera corretta, sempre in una specifica classe, e il numero di dati assegnati a quest'ultima.

Vengono aggiunti questi due valori poiché l'accuratezza da sola non è sufficiente per la valutazione, in particolare in presenza di dataset non bilanciati nella composizione delle classi (gruppi con numerosità molto differenti).

Il primo esperimento è il più generico e classico; di norma corrisponde al primo tentativo che si tenta e che utilizza le percentuali di training set e test set standard delle funzioni per la divisione del dataset messe a disposizione da Scikit-Learn (train_test_split (x, y, test_size = testsize, random_state=0)), che corrispondono a 90% train e 10% test; questo significa che il gruppo usato come test ha 2384 elementi al suo interno. In tabella 6.1a e 6.1b si possono osservare le matrici di confusione, rispettivamente per albero di decisione e k-nn:

		predetti					somma		
	nW	0	1	2	3	4		richiamo	precisione
Reali	0	258	0	0	0	0	258	1.000	1.000
	1	0	557	0	0	0	557	1.000	1.000
	2	0	0	649	12	0	661	0.982	0.997
	3	0	0	2	795	10	807	0.985	0.979
	4	0	0	0	5	96	101	0.950	0.906
Somma		258	557	651	812	106	2384	accuratezza	0.9878356

Tab. 6.1a Matrice di confusione (train = 90% ottenuto con campionamento casuale), modello L'Albero di Decisione

		predetti					somma		
	nW	0	1	2	3	4		richiamo	precisione
Reali	0	258	0	0	0	0	258	1.000	1.000
	1	0	557	0	0	0	557	1.000	1.000
	2	0	0	661	0	0	661	1.000	0.998
	3	0	0	1	806	0	807	0.999	1.000
	4	0	0	0	0	101	101	1.000	1.000
Somma		258	557	662	806	101	2384	accuratezza	0.9995805

Tab. 6.1b Matrice di confusione (train = 90% ottenuto con campionamento casuale), modello k-nn

I risultati ottenuti, con accuratezza pari quasi ad uno per entrambi i modelli sembrano eccezionali, soprattutto considerando che è stato il primo esperimento, perché significa che i dati sono stati predetti quasi tutti correttamente (pochissimi elementi fuori dalla diagonale maggiore della matrice. Tuttavia, questo risultato nasconde alcune problematiche: la grande differenza di numerosità degli elementi tra il set di training e quello di test, in aggiunta al campionamento casuale e all'elevato numero di features, può aver portato alla creazione di un modello eccessivamente complesso che si riesce ad adattare a qualsiasi tipologia di dati. Si tratta, molto probabilmente, di un problema legato all'overfitting di dati, ovvero al sovradimensionamento dei dati di addestramento usati per il modello; di conseguenza questo tipo di modello è stato scartato.

Commentato [TLS1]: Lo lascerei

Il secondo esperimento, sempre con l'obiettivo di esplorare le possibilità di split del dataset in training e test, rovescia le percentuali usate nel caso precedente, utilizzando quindi solo il 10% dei dati per addestrare i classificatori ed il restante 90% per verificarne le capacità di analisi, portando a più di 21'000 gli elementi del test set; questo è stato fatto anche per verificare, in caso di drastico cambiamento dei risultati, la presenza del problema di overfitting. I risultati sono mostrati in tabella 6.2a e 6.2b, la prima sempre riguardante l'Albero di Decisione e la seconda per il k-nn:

		predetti					somma		
	nW	0	1	2	3	4		richiamo	precisione
reali	0	253	1047	116	608	129	2153	0.118	0.088
	1	766	2178	359	1217	311	4831	0.451	0.213
	2	551	2717	1131	954	238	5591	0.202	0.493
	3	1110	3932	618	1691	486	7837	0.216	0.366
	4	200	347	72	144	277	1040	0.266	0.192
somma		2880	10221	2296	4614	1441	21452	accuratezza	0.2577848

Tab. 6.2a Matrice di confusione (train = 10% ottenuto con campionamento casuale), modello L'Albero di Decisione

		predetti					somma		
	nW	0	1	2	3	4		richiamo	precisione
reali	0	301	167	1148	526	11	2153	0.140	0.098
	1	641	619	2520	1027	24	4831	0.128	0.370
	2	425	214	3938	989	25	5591	0.704	0.326
	3	1551	618	4046	1449	173	7837	0.185	0.336
	4	169	56	422	325	68	1040	0.065	0.226
somma		3087	1674	12074	4316	301	21452	accuratezza	0.2971751

Tab. 6.2b Matrice di confusione (train = 10% ottenuto con campionamento casuale), modello k-nn

Si nota subito che l'accuratezza precipita a meno del 30% per entrambi i classificatori, con risultati ancora peggiori nel l'Albero di Decisione e che, in generale, i modelli utilizzati hanno performance molto scadenti nell'assegnare le predizioni; evidenze che hanno portato a scartare questa percentuale e il modello annesso.

Gli esperimenti seguenti mirano ad ampliare ulteriormente la conoscenza su basse percentuali di dati processati per l'addestramento ma cambiando la tipologia di campionamento utilizzata. Il primo metodo di analisi è stato ottenuto grazie all'uso di una funzione apposita che fosse capace di mantenere l'ordine temporale dei cicli e di conseguenza estrarre una percentuale fissa di dati ordinati per comporre il dataset di training. La seconda tipologia invece, estrae un campione fisso ma non casualmente, bensì in maniera stratificata non proporzionale da ogni gruppo, senza però tenere conto dell'ordine temporale.

L'obiettivo, per la prima parte, è quindi quello di scoprire se un piccolo dataset, ordinato temporalmente, può offrire un buon modello per predire i dati futuri, appartenenti a finestre temporali successive. Nella seconda invece, si verifica se sia meglio utilizzare un sistema di scelta randomico o uno temporale. Le percentuali usate sono state diverse, in un range che spazia dal 5% al 30%; in questa tesi verranno presentati solo i risultati riguardanti il 10%, sia per una questione di coerenza e confronto con le analisi precedenti, sia per i discreti risultati ottenuti nelle due tipologie rispetto alle altre percentuali testate, anche se non differiscono di molto; in ogni caso il 10% è un buon compromesso per la corretta rappresentazione dei risultati ottenuti.

Sebbene questi due tipi di analisi possano assomigliarsi, a conti fatti hanno portato a risultati incredibilmente diversi. Nelle tabelle 6.3a e 6.3b si possono osservare le matrici di confusione riguardanti i due classificatori usati fino ad ora riguardanti il primo tipo di esperimento, quello temporale. Sono mostrate rispettivamente due matrici di confusione:

		predetti					somma		
	nW	0	1	2	3	4		richiamo	precisione
reali	0	279	798	489	403	184	2153	0.130	0.098
	1	847	1416	1261	960	347	4831	0.293	0.216
	2	459	1542	2459	696	435	5591	0.440	0.403
	3	1079	2551	1628	1918	661	7837	0.245	0.457
	4	176	239	260	219	146	1040	0.140	0.082
somma		2840	6546	6097	4196	1773	21452	accuratezza	0.28985642

Tabella 6.3a: Matrice di confusione (train = primo 10% di ogni gruppo), modello l'Albero di Decisione

		predetti					somma		
	nW	0	1	2	3	4		richiamo	precisione
reali	0	298	164	1164	514	13	2153	0.138	0.099
	1	633	624	2551	993	30	4831	0.129	0.364
	2	413	221	3967	960	30	5591	0.710	0.326
	3	1501	648	4060	1416	212	7837	0.181	0.337
	4	167	59	422	318	74	1040	0.071	0.206
somma		3012	1716	12164	4201	359	21452	accuratezza	0.29736155

Tabella 6.3b: Matrice di confusione (train = primo 10% di ogni gruppo), modello k-nn

Come si può facilmente notare i risultati ottenuti non sono soddisfacenti: per entrambi i modelli l'accuratezza è sotto il 30%; i valori di richiamo sono molto bassi per tutte le classi ad eccezione della 2 per il modello k-nn; anche la precisione è molto bassa, rasentando lo zero per la classe 0. In generale neanche gli altri test effettuati con diverse percentuali hanno portato a risultati soddisfacenti con il campionamento in ordine temporale; probabilmente questo è dovuto al fatto che solo il primo 10% in ordine di tempo non è sufficiente a modellare le casistiche che possono intercorrere durante l'attività standard del braccio robot.

Per questo motivo si è passati al secondo tipo di campionamento i cui risultati possono essere osservati in tabella 6.4a e 6.4b. Qui invece, si osservano le tabelle relative alla seconda tipologia di campionamento effettuato, con il 10% casuale stratificato (da ogni gruppo):

		predetti					somma		
	nW	0	1	2	3	4		richiamo	precisione
reali	0	646	582	339	457	129	2153	0.300	0.346
	1	474	2631	745	805	176	4831	0.545	0.590
	2	154	333	4354	709	41	5591	0.779	0.671
	3	435	651	897	5585	269	7837	0.713	0.709
	4	160	262	150	318	150	1040	0.144	0.196
somma		1869	4459	6485	7874	765	21452	accuratezza	0.62306545

Tabella 6.4a: Matrice di confusione (train = estrazione casuale del 10% dei cicli da ogni gruppo), modello l'Albero di Decisione

		predetti					somma		
	nW	0	1	2	3	4		richiamo	precisione
reali	0	415	402	972	364	0	2153	0.193	0.845
	1	56	2205	1803	767	0	4831	0.456	0.747
	2	2	42	5326	221	0	5591	0.953	0.563
	3	5	46	916	6870	0	7837	0.877	0.808
	4	13	257	443	279	48	1040	0.046	1.000
somma		491	2952	9460	8501	48	21452	accuratezza	0.69289577

Tabella 6.4b: Matrice di confusione (train = estrazione casuale del 10% dei cicli da ogni gruppo), modello k-nn

La seconda parte di esperimenti porta a risultati nettamente migliori: si può notare, infatti, che l'accuratezza del modello mostrato arriva al 69% nel caso dell'albero di decisione e poco sopra il 62% per il k-nn, che in ogni caso rappresentano un netto miglioramento rispetto al primo tipo di campionamento; gli unici valori bassi si hanno in corrispondenza dell'indice richiamo, ma solo dei gruppi più esterni (questo sta a significare che i cicli appartenenti alle classi 0 e 4 vengono predetti nelle altre classi). Da questi risultati si deduce che il problema non risiede tanto nella percentuale di dati che si estraggono, poiché il 10% nel caso 2 sembra essere più che sufficiente, ma il problema è relativo più al metodo di estrapolazione dei cicli per il training; sembra dai risultati, che un'estrazione casuale di dati, anche se piccola, sia preferibile, probabilmente perché questo permette di far conoscere al classificatore un maggior numero di casistiche possibili di diversi momenti temporali, anche se di numerosità bassa. In ogni caso, la buona valutazione di questa tipologia di prove non è sufficiente a risolvere tutte le problematiche iniziali: infatti, adottando un campionamento casuale c'è il rischio che il modello, dopo aver predetto alcune nuove serie di dati, non si riesca più ad adattare perché potrebbero mancare casistiche importanti.

Gli ultimi esperimenti riguardante il modello generale sono stati condotti sempre con l'obbiettivo di cercare la quantità minima di dati in modo da avere un modello che predica correttamente ma al tempo stesso mantenga l'informazione temporale sulla serie di cicli.

La sezione seguente è stata organizzata in 10 test differenti, con numerazione crescente a partire da 1, a cui sono associati altrettanti modelli di classificazione, separatamente per albero di decisione e k-nn. Questi esperimenti sono stati realizzati non considerando ogni volta l'intero dataset, ma solamente delle sue porzioni, caratterizzate da numerosità maggiori al progredire degli stessi:

il primo esperimento utilizza il 10% dei cicli, il secondo il 20% e così via fino ad utilizzare l'intero dataset. Un altro fattore da sottolineare di questa fase finale per il modello generale è il campionamento utilizzato, cioè una selezione di dati stratificata sulle classi in ordine di tempo; quindi il primo esperimento contiene il primo 10% dei dati da ogni gruppo, la restante parte dei dati viene invece ignorata. La ragione della scelta sta nel voler mantenere le proporzioni tra i gruppi ma avere casistiche da ogni classe. Inoltre, ad ogni esperimento, i dati estratti sono stati divisi in training e test, rispettivamente con percentuali corrispondenti a 70% e 30% ed infine sono state calcolate le usuali metriche per la valutazione, accuratezza, precisione e richiamo.

In tabella 6.5a e 6.5b, rispettivamente per l'Albero di Decisione e k-nn, si possono osservare i dati riassuntivi per i 10 esperimenti condotti, riportando i valori degli indici di valutazione più le informazioni sui cicli utilizzati, la loro divisione tra le varie classi e le percentuali assegnate ad ogni test:

esp.	#cicli totali	% c. usati	#cicli g0	#cicli g1	#cicli g2	#cicli g3	#cicli g 4	accuratezza
1	2381	10%	239	536	621	870	115	0.853556485
2	4765	20%	478	1073	1242	1741	231	0.867088608
3	7148	30%	717	1610	1863	2612	346	0.807172799
4	9530	40%	956	2146	2484	3482	462	0.790632646
5	11915	50%	1196	2683	3106	4353	577	0.735738255
6	14299	60%	1435	3220	3727	5224	693	0.731423247
7	16680	70%	1674	3756	4348	6094	808	0.690633114
8	19064	80%	1913	4293	4969	6965	924	0.718056284
9	21447	90%	2152	4830	5590	7836	1039	0.684169644
10	23833	100%	2392	5367	6212	8707	1155	0.711729344

Tab. 6.5a Riassunto informazioni sui 10 esperimenti finali con l'Albero di Decisione

Esp.	#cicli totali	% c. usati	#cicli g0	#cicli g1	#cicli g2	#cicli g3	#cicli g 4	accuratezza
1	2381	10%	239	536	621	870	115	0.928870293
2	4765	20%	478	1073	1242	1741	231	0.933659218
3	7148	30%	717	1610	1863	2612	346	0.895668374
4	9530	40%	956	2146	2484	3482	462	0.851101014
5	11915	50%	1196	2683	3106	4353	577	0.843959732
6	14299	60%	1435	3220	3727	5224	693	0.830887491
7	16680	70%	1674	3756	4348	6094	808	0.819852207
8	19064	80%	1913	4293	4969	6965	924	0.798461807
9	21447	90%	2152	4830	5590	7836	1039	0.793071307
10	23833	100%	2392	5367	6212	8707	1155	0.778973857

Tab. 6.5b Riassunto informazioni sui 10 esperimenti finali con il k-nn

I risultati portano al risultato che ci si aspettava, cioè si osserva un deterioramento dei valori di accuratezza al passare del tempo. Questo significa che probabilmente la macchina è soggetta ad un lento degrado nelle performance, poiché entrambi i classificatori, in particolare l'albero di decisione, passa da un valore al primo esperimento di oltre l'85% per arrivare, quando usa l'intero dataset, intorno al 70%. Le performance dei modelli, quindi, peggiorano con il passare del tempo predicono sempre meno correttamente i dati utilizzati nei test. Inoltre, questi esperimenti dimostrano che un modello addestrato con il 70% dei dati può dare risultati soddisfacenti nei valori di accuratezza.

Tuttavia, non basta questo indice per valutare i vari modelli, quindi si riportano in questa sezione finale del modello generale i grafici legati anche a richiamo e precisione per entrambi i classificatori; in figura 6.6a e 6.6b si possono osservare i valori di richiamo rispettivamente di l'Albero di Decisione e k-nn.

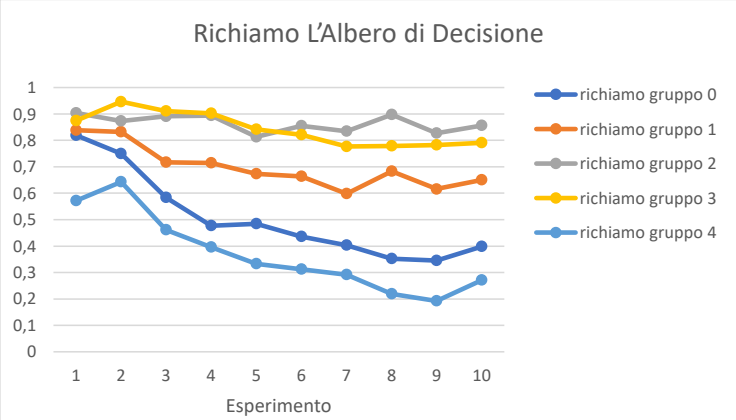


Fig. 6.6a Richiamo per ogni gruppo nei 10 esperimenti, modello l'Albero di Decisione

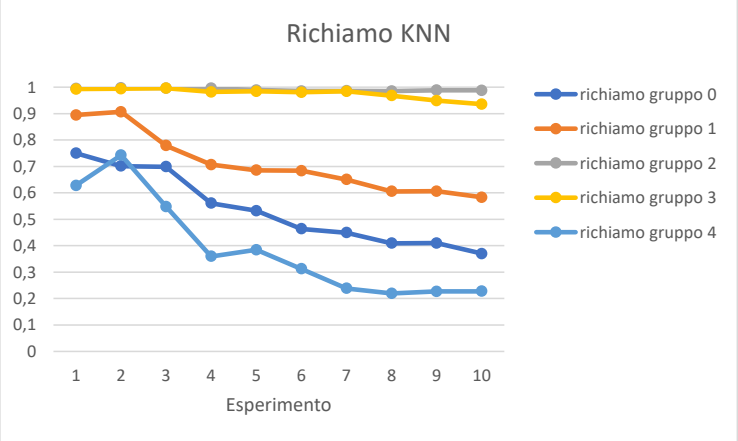


Fig. 6.6a Richiamo per ogni gruppo nei 10 esperimenti, modello k-nn

I valori del richiamo risultano discreti nel caso del l'Albero di Decisione, ma solo per i gruppi 'interni' quindi da 1 a 3 e crollano per le due classi estreme; situazione che diventa ancora più netta nel caso del secondo classificatore, dove si ottengono ottimi risultati solo sui gruppi 2 e 3 ed inizialmente 1, per poi avere un tracollo anche su questa classe come per 0 e 4. Valori bassi significano che i dati appartenenti alle altre classi vengono assegnate a quelle in questione ovvero i dati vengono scambiati per classi a cui in realtà non appartengono.

Invece, in figura 6.7a e 6.7b si possono vedere i risultati relativi alla precisione nei 10 esperimenti condotti, sempre rispettivamente per albero di decisione e k-nn.

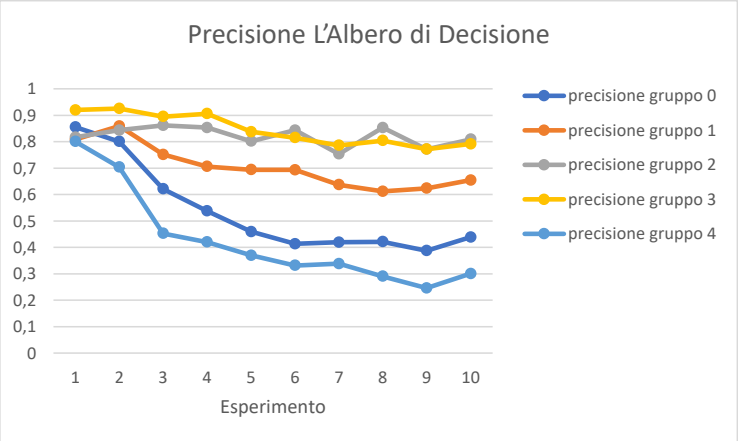


Fig. 6.7a precisione per ogni gruppo nei 10 esperimenti, modello l'Albero di Decisione

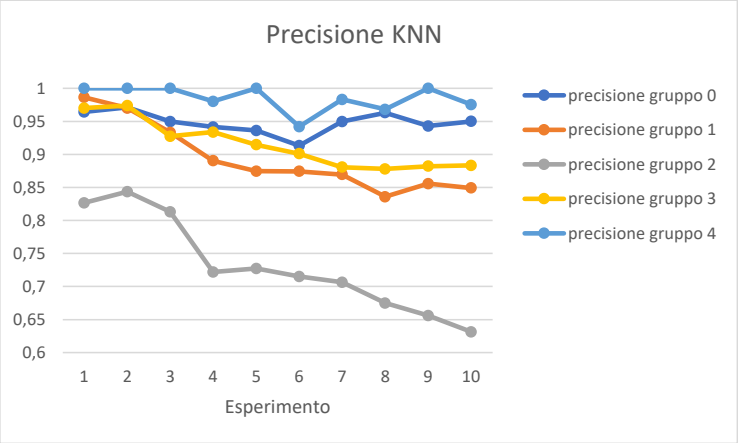


Fig. 6.7b precisione per ogni gruppo nei 10 esperimenti, modello k-nn

La precisione nel caso del l'Albero di Decisione segue il trend già visto per il richiamo nei vari gruppi, con valori alti per gruppi interni e bassi per gli esterni. L'indice, invece, vede un'inversione di tendenza rispetto al richiamo se si utilizza il secondo classificatore; qui i gruppi con i valori più alti risultano i più esterni, le classi 0 e 4, valori inizialmente alti ma poi si abbassano per i gruppi intermedi, cioè 1 e 3, ed infine il gruppo 2 che ha valori molto bassi che scendono fino al 60% se si considera l'ultimo esperimento con l'intero dataset. Valori così bassi significano che i dati appartenenti al gruppo in questione vengono erroneamente assegnati alle altre classi analizzate.

Commentato [TLS2]: Venendo assegnati? È questo in senso?

6.3 Modello Evolutivo

A seguito dei risultati del modello generale, si è reso necessario pensare una soluzione idonea a raggiungere l'obbiettivo di fare manutenzione predittiva; la risposta è stata la creazione di un modello evolutivo.

Questo consiste in un sistema che sia in grado di adattarsi alle condizioni mutevoli a cui viene sottoposto il braccio robot a cui il modello è collegato; infatti in ambiti industriali reali quello che si ha è una porzione di dati inziali su cui si esegue il training del modello, ma successivamente si ricevono dati senza etichetta reale, poiché la macchina, essendo soggetta ad usura, cambia i valori del segnale che si raccoglie. Un sistema di manutenzione predittiva si propone di riuscire ad individuare questo cambiamento nei nuovi dati ricevuti e segnalare per tempo la necessità di un intervento di aggiustamento nel numero di rondelle che regolano la cinghia di tensionamento.

Per realizzare un modello del genere si sono eseguiti alcuni passaggi che sono comuni a tutte le analisi che vengono presentate in questa sezione del progetto; per prima cosa si sono selezionati solo alcuni dei gruppi disponibili riguardanti il numero di rondelle, tenendo gli altri come gruppi esterni, quindi un insieme di dati "nuovi" di cui non si conosce l'etichetta a priori. In seguito, vengono utilizzati i classificatori già usati per il modello generale, ossia albero di decisione e K-NN ed infine i risultati sono stati verificati tramite vari indici di dispersione. Come per le analisi precedenti, si è scelto di usare la matrice tagliata al valore 0,5 di correlazione tra variabili ed utilizzare.

Il primo passo, quindi, consiste nella selezione dal dataset iniziale di sole alcune classi, che verranno intese come gruppi principali su cui poi addestrare i modelli di classificazione, che vengono poi testati sui gruppi non considerati inizialmente. I gruppi principali scelti sono le classi 1, 2 e 3 per la prima parte delle analisi, questa selezione viene poi cambiata per eseguire alcuni approfondimenti; la scelta iniziale è ricaduta su questi due valori principalmente per 3 motivi:

- il primo è dovuto alla centralità dei valori; si è voluto prendere le classi con i valori interni di numero di rondelle, escludendo i valori estremi di 0 rondelle e 4; in particolare 0, come si è già accennato, è un caso non industriale, viene quindi usato come test sono inizialmente per poi essere messo da parte.
- la seconda motivazione è dovuta alla numerosità dei dati; i gruppi 1, 2 e 3 formano l'85% dei dati disponibili, in particolare il gruppo 3 è molto ricco di informazioni, invece il gruppo 4 ha pochissimi record al suo attivo e la classe 0 poco di più; anche per questo si

è scelti ti tenerli come gruppi esterni questi ultimi, poiché ben rappresentano una casistica reale dove si hanno una serie di dati iniziali e poi poco alla volta si raccolgono nuovi dati e si classificano

- infine, i valori corrispondenti ad 1 e 2 sono i casi maggiormente interessanti ed in uso in ambito industriale, per quanto riguarda il robot analizzato, risulta quindi fondamentale approfondire la conoscenza di queste classi ed avere un modello in grado di capire se i nuovi dati si stanno discostando da questi due valori base

Si è deciso di ordinare i dati estratti per data, così da rimanere il più possibile aderenti ad una casistica reale in una fabbrica, tenendo in considerazione il fattore tempo.

Il secondo passaggio di queste analisi consiste nell'addestramento dei modelli di classificazione scelti; come per il modello generale si sono usati l'albero di decisione e il K-NN, quindi anche in questa fase si sono ricercati i migliori parametri possibili per il classificatore ad albero tramite sempre il metodo di validazione cross-validation con K-folds uguale a 10. Come per le analisi precedenti sono sempre due i parametri che vanno selezionati per questo modello: la massima profondità e il massimo numero di features da considerare per uno split. Per quest'ultimo fattore, i valori possibili sono uguali al modello generale:

max_features: ['auto', 'sqrt', 'log2'].

Per quanto riguarda la profondità dopo vari test i valori finali selezionati sono stati:

max_depth = [5, 10, 20, 40].

I valori che risultano da ogni analisi vengono descritti nella sezione di appartenenza poiché variano a seconda della matrice o dei gruppi selezionati. Per altri parametri che possono essere impostati sono stati usati quelli di default del pacchetto SciKitLearn. Per quanto riguarda invece il k-nn, il valore k è stato impostato, diversamente dal modello generale, ad un valore pari a 5, in modo tale da avere un modello maggiormente sensibile al rumore nei dati.

L'ultimo passaggio comune ai vari test eseguiti, riguarda i metodi di valutazione del modello; non avendo modo di effettuare un'analisi sui valori di richiamo e precisione come fatto nel modello generale, poiché si ricorda che i dati dei gruppi esterni vengono considerati come nuovi, quindi l'etichetta originale non viene considerata; per ovviare al problema si sono calcolati vari indici di dispersione:

- il primo indice, considerato anche come base di valutazione, è la silhouette che è stata calcolata ad ogni test eseguito

- L'MSE che è stato utilizzato dopo i primi risultati della silhouette, per avere un metodo di verifica
- Negli ultimi test si è affiancato anche l'utilizzo del MAAPE, dati i risultati poco soddisfacenti dell'MSE

La silhouette è un indice che misura la coesione dei cluster, andando a verificare, valore per valore quanto questo sia simile al suo gruppo di appartenenza (coesione) e quanto diverso dagli altri (separazione), grazie ad una misura della distanza dagli altri punti. In particolare, in questo lavoro si è scelto di usare la distanza euclidea; il range di valori ammissibili è da -1 a 1, quanto più vicino è ad 1 tanto migliore è stata la classificazione, se molti dati ottengono valori alti, significa che il modello funziona bene ed i gruppi sono ben separati; al contrario, valori bassi o anche negativi stanno a significare che i gruppi non sono coesi e che ci sono troppi o troppo pochi gruppi per rappresentare i dati.

Un altro indice che è stato scelto di calcolare in alcuni passaggi delle analisi è stato l'MSE (mean square error); classica misurazione di dispersione, calcola la media al quadrato dello scostamento tra valori osservati e valori stimati, calcolando il valore medio di ogni record separatamente per ogni gruppo (nel calcolo del valore medio ovviamente viene esclusa l'etichetta) che viene usato come valore stimato ed applicando in seguito le funzioni messe a disposizione sempre da SciKit-Learn per il calcolo dell'MSE vero e proprio, riga per riga del dataset.

Infine, per le analisi più approfondite, eseguite per la creazione di un modello evolutivo, si è scelto di affiancare agli indici già elencati anche il MAAPE (mean arctangent absolute percentage error), che è un'evoluzione del MAPE (mean absolute percentage error), il cui utilizzo specifico viene spiegato nella sezione apposita. Viene calcolato sempre un errore tra valori osservati e valori stimati, tuttavia in questo caso è reso in percentuale rispetto a quelli realmente ottenuti (osservati), si applica l'errore assoluto ed infine l'operatore dell'arcotangente. Quest'ultima operazione viene effettuata prima del calcolo del valore medio poiché in presenza di molti valori prossimi allo zero il MAPE produce risultati infiniti o indefiniti, come nel caso in esame dove il segnale inziale, e quindi tutti gli indici statistici da esso calcolati, si aggirano intorno allo zero, in particolare alcuni segmenti in cui è stata spezzata la lettura della corrente; grazie all'arcotangente questo problema viene ovviato, tuttavia la lettura dei dati diventa meno intuitiva, poiché i valori vengono tradotti in radianti. La formula utilizzata è la seguente:

$$\mathrm{MAAPE} = \frac{1}{N}\sum_{t=1}^{N}\mathrm{AAPE_t} = \frac{1}{N}\sum_{t=1}^{N}\arctan\left(\left|\frac{y_t - f_t}{y_t}\right|\right)$$

6.3.1 Prima fase

Le prime analisi effettuate si sono concentrate sullo studio della silhouette per l'insieme dei gruppi principali rispetto ai due gruppi esterni. Prima di procedere con la spiegazione esatta della procedura seguita si vuole sottolineare che i processi di predizione e di silhouette sono separati ed indipendenti tra loro e per questa ragione inizialmente i valori della matrice venivano normalizzati sono per la parte di predizione e mai per il calcolo dell'indice di dispersione; tuttavia, a seguito dei risultati non particolarmente brillanti così ottenuti, si è pensato di estendere il processo di normalizzazione all'intero processo, poiché la silhouette, essendo un calcolo di distanze euclidee, beneficia enormemente di dati normalizzati, infatti si sono subito notati valori migliori e più coerenti nella silhouette. Si vuole far ancora notare che si è scelta la normalizzazione per riga, per non perdere informazione riguardo a cicli notevolmente differenti da altri, cosa che sarebbe successa eseguendo quella per colonna.

Per questa fase si usa l'intero dataset, sia per i gruppi principali sia per gli esterni. Si sono presi i primi e sono stati utilizzati nella loro totalità per l'addestramento dei due modelli di classificazione, il K-nn e l'albero di decisione; in particolare per quest'ultimo i migliori parametri riscontrati sono stati: 'max_depth': 10, 'max_features': 'auto'. Si sono poi usati i due gruppi esterni come test in maniera separata ed indipendente, senza quindi che la predizione di un gruppo influenzasse l'altro; si è conseguentemente eliminata l'informazione su numWashers e si sono considerati questi dati come "nuovi", provenienti da una macchina normalmente funzionante che vanno quindi classificati secondo i modelli creati riguardanti le sole etichette conosciute.

Per semplificare la visualizzazione dei risultati si è scelto di mostrare nei grafici che seguiranno solo dei campioni dei dati, precisamente 1200 dati casuali per ogni gruppo di training che è pressoché il numero di quelli disponibili per la classe più piccola 4, che contiene 1155 record; inoltre questo permette di portare coerenza tra la grande disparità nella numerosità dei dati dei vari gruppi a disposizione. Si tiene a precisare in ogni caso, che solo la visualizzazione è stata campionata, il calcolo della silhouette è stato eseguito su tutti i dati di partenza (sia di training sia quelli usati per la predizione).

I primi risultati così ottenuti sono mostrati in fig. 6.8a e 6.8b, dove si possono osservare i valori di silhouette campionati per i tre gruppi principali selezionati per questa parte, prima della predizione del corrispettivo gruppo esterno e dopo, quando questo viene inserito nel dataframe inziale per effettuare il calcolo dell'indice; nella figura 6.6a si trovano i valori relativi al gruppo 0 per il modello l'Albero di Decisione, nella seconda quelli riguardanti il modello K-NN.

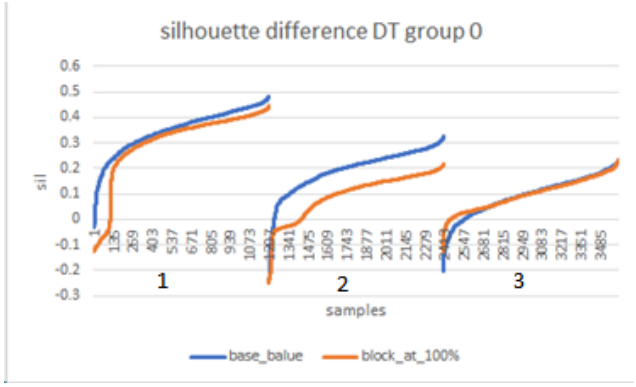


Fig. 6.8a differenza di silhouette tra gruppi principali gruppo 0 modello DT

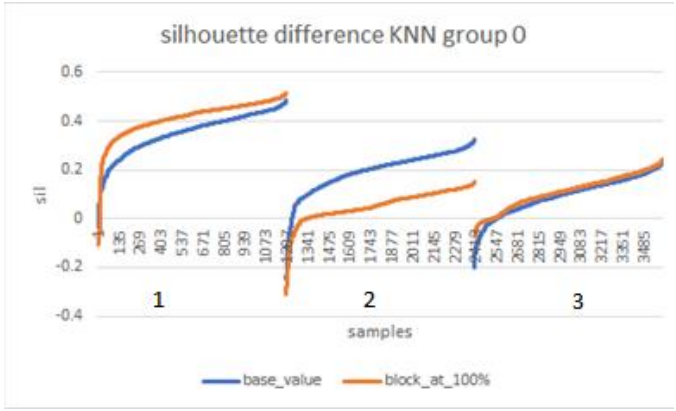


Fig. 6.8b differenza di silhouette tra gruppi principali gruppo 0 modello K-NN

Commentato [TLS3]: Mettere graifco gruppo 4

L'albero di decisione mostra una separazione abbastanza netta per il gruppo 1 ed in particolare per il gruppo 2, dove la forbice tra curva del valore base, quindi prima della predizione, e quella con l'intero blocco 0 predetto è molto ampia; molto più piccola invece la differenza nel gruppo 3, dove addirittura c'è un miglioramento della silhouette iniziale. Il k-nn invece, mostra un comportamento simile all'albero solamente nel gruppo 2 e con un andamento negativo decisamente più netto; nel gruppo 3 e soprattutto nell'1 invece, il valore della silhouette si ritrova al di sopra del valore considerato base. Questo fatto è probabilmente dovuto alla normalizzazione dei dati prima sia della predizione sia del calcolo della silhouette, poiché sia l'indice sia il classificatore

misurano distanze e quindi beneficiano fortemente di una normalizzazione. Tuttavia, questo non permette di visualizzare un peggioramento significativo e coerente della coesione dei gruppi al loro interno dopo l'inserimento di nuovi dati che si sa a priori essere molto diversi.

Analogamente per quanto fatto per il gruppo 0, anche per il gruppo 4 si mostrano in figura 6.9a e 6.9b i valori di silhouette campionati per i tre gruppi principali, pre e post predizione.

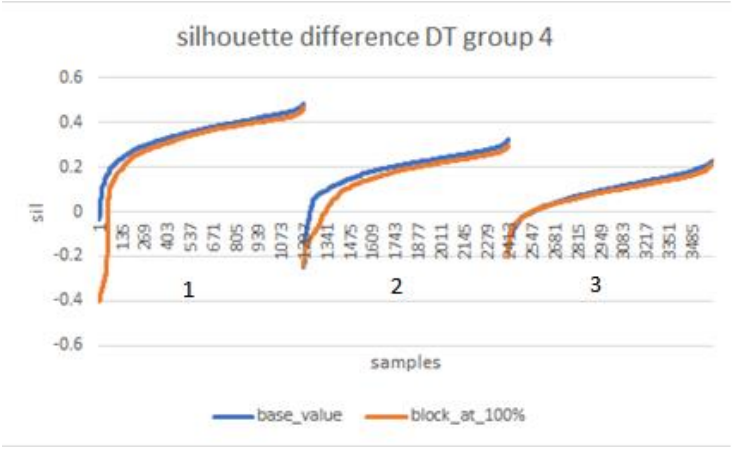


Fig. 6.9a differenza di silhouette tra gruppi principali gruppo 0 modello DT

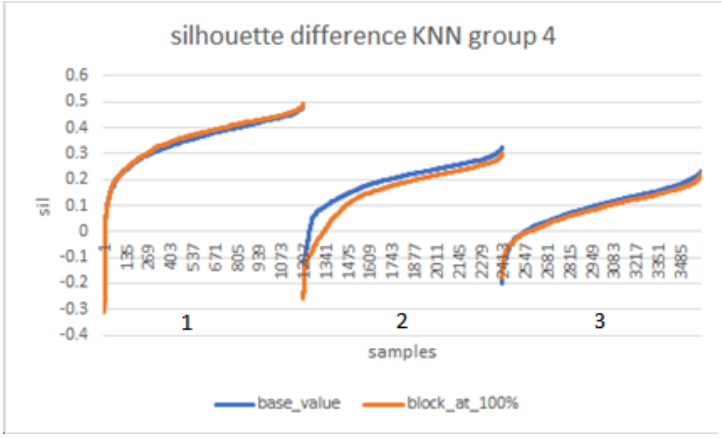


Fig. 6.9b differenza di silhouette tra gruppi principali gruppo 0 modello K-NN

Diversamente dal gruppo 0 i modelli sembrano comportarsi meglio con il gruppo 4; infatti, anche se nell'albero di decisione per i primi due gruppi la forbice non è ampia come per il primo gruppo qua affrontato, si può notare una piccola differenza in ogni gruppo di training, non solo nei primi due; inoltre questo risultato si può vedere anche per il classificatore del K-Nearest Neighbor, dove il gruppo 1 post predizione (linea arancione sempre) non sovrasta più il valore base ma quasi lo ricalca e come per il primo classificatore, la forbice nel gruppo 2 si assottiglia visibilmente, anche se rimane chiaro il peggioramento della coesione in questa classe e diventa leggermente più marcata anche nel gruppo 3.

Queste prime analisi quindi, portano a risultati contrastanti, sia in base ai classificatori usati sia in base al gruppo che si è scelto di predire, con il k-nn che permette di arrivare a valori di silhouette più elevati dopo la predizione rispetto al valore base ma solo per il gruppo 1, accentuando le differenze nel gruppo 2 usando 0 come test e nel gruppo 3 invece di 4.

In generale, si può comunque osservare un peggioramento del valore della silhouette dei vari gruppi osservati e per queste ragioni si è voluto indagare più a fondo la fluttuazione dell'indice di dispersione, a seconda non solo più del classificatore o del gruppo di test ma anche in base ad altre variazioni che verranno spiegate qui di seguito.

6.3.2 Seconda fase

La seconda fase di analisi si concentra quindi sulla ricerca del degrado in maniera più specifica e dettagliata. In particolare, si vuole individuare il peggioramento dello standard della macchina nel tempo, poiché l'obbiettivo è sempre trovare un modello evolutivo. Quindi è di fondamentale importanza che il modello scelto sia in grado di riconoscere tempestivamente e in maniera precisa il degrado.

Per raggiungere l'obbiettivo di comprendere l'evoluzione del segnale nel tempo, si è deciso di condurre diversi esperimenti su base temporale e di gruppi di training, per meglio comprendere l'andamento della macchina; in pratica si è deciso di dividere il dataset in 3 fasce temporali:

- La prima settimana, che va dal 10 ottobre al 15 ottobre
- Il secondo caso è rappresentato dalle prime due settimane assieme, quindi sempre dal 10 ottobre al 23 dello stesso mese
- Infine, il terzo caso, comprende il dataset nella sua interezza

In tabella 6.6 si possono osservare le numerosità delle varie classi dipendentemente dal caso in esame. È stato eliminato il caso del gruppo 0 perché, come già accennato, è una casistica non realistica in ambito industriale e la cui analisi risulta inutile per un modello come quello evolutivo che è strettamente connesso al caso industriale in cui viene utilizzato.

cicli per gruppo	dal 10 al 15 ott.	dal 10 al 23 ott.	dal 10 al 30 ott.
gruppo 1	5367	5367	5367
gruppo 2	1126	1733	6212
gruppo 3	674	4059	8707
gruppo 4	1155	1155	1155
Totale	8322	12314	21441

Tab. 6.6 Divisione dei numeri di cicli per i gruppi nei vari casi temporali

La numerosità del gruppo 3 è incentrata soprattutto nella seconda e terza settimana dei dati; è bene perciò ricordare questo fattore quando si analizzano i risultati del caso 1 per i gruppi in cui 3 è presente come training, poiché lo scarso numero di dati potrebbe essere la causa del risultato. Al contrario le classi 1 e 4 hanno i loro dati concentrati tutti nella prima settimana, quindi non

cambiano nei casi 2 e 3. Infine, il gruppo 2 ha un comportamento molto simile a 3, ma con i dati concentrati tutti nella parte finale del dataset, in ordine temporale.

Quindi si sono condotti vari esperimenti, non solo in base alle fasce temporali, ma anche variando il gruppo di addestramento del modello, utilizzando per ognuno il dataset diviso nei tre casi sopra citati, in maniera indipendente l'uno dall'altro, cioè separatamente. Il primo esperimento tiene in considerazione come gruppo di training sempre l'insieme delle classi 1, 2, e 3, poi diviso ulteriormente nei tre casi temporali già spiegati; il secondo esperimento invece, si focalizza su i due gruppi che sono considerati i valori ottimali a regime della macchina, poiché i più utilizzati ed importanti per il corretto funzionamento del braccio robot che sono le classi 1 e 2. Infine, si è utilizzato l'insieme 2 e 3 poiché 2 è il valore migliore possibile e 3 può essere considerato come un leggero peggioramento nel limite dell'accettabile. A questo punto come classi di test sono stato utilizzati 4, in tutte le combinazioni possibili di esperimenti e anche 3 per quanto riguarda l'analisi che utilizza come gruppo di training 1 e 2.

Infine, per affinare ulteriormente il modello nella sua capacità di individuare il degrado nel tempo, si è scelto di suddividere ogni gruppo di test in 5 sottogruppi di uguale dimensione, ordinati temporalmente; questi vengono predetti indipendentemente dagli altri, poiché il classificatore non viene più riaddestrato dopo la prima volta e viene così calcolata la silhouette. Per il calcolo di quest'ultimo indice tuttavia, viene utilizzato un dataframe a cui vengono incorporati i sottogruppi precedenti con le loro nuove etichette, uscite fuori dal modello di predizione, aggiungendo praticamente un blocco alla volta del gruppo di test, fino ad arrivare al 100% dei dati. In questo modo si vuole osservare se ci sia o meno un degrado nel valore base di coesione dei gruppi all'avanzare del tempo, cioè dopo aver integrato diversi sottogruppi.

6.3.2.1 Primo esperimento

Il primo esperimento è eseguito con l'addestramento dei classificatori con i gruppi 1, 2 e 3; il gruppo di test invece è solamente il gruppo 4.

I risultati sono ulteriormente divisi per casi, qui di seguito viene presentata una selezione dei risultati disponibili tra cui i migliori parametri per l'albero di decisione, il grafico rappresentante le curve di silhouette ad ogni sottogruppo (sempre estraendo un campione di 1200), il grafico della media ad ogni step temporale (quindi ad ogni sottogruppo) ed infine la tabella che riassume come ogni piccolo insieme è stato predetto nelle varie classi di training.

I migliori parametri dell'albero di decisione per l'esperimento 1, caso 1, cioè considerando solo i dati appartenenti alla prima settimana del dataset originale, sono 'max_depth': 10, 'max_features': 'auto', uguale in tutti e tre i casi seguenti; in tabella 6.7a, si possono osservare invece come sono stati predetti i vari sottogruppi per il modello albero di decisione, invece in 6.7b per il k-nn, ogni sottogruppo è composto da 231 elementi.

dT	block_at_20%	block_at_40%	block_at_60%	block_at_80%	block_at_100%
group 1	158	156	156	163	167
group 2	71	74	73	64	63
group 3	2	1	2	4	1

Tab. 6.7a predizioni gruppo di test, modello l'Albero di Decisione

knn	block_at_20%	block_at_40%	block_at_60%	block_at_80%	block_at_100%
group 1	230	227	228	226	225
group 2	1	4	3	5	6
group 3	0	0	0	0	0

Tab. 6.7b predizioni gruppo di test, modello k-nn

Il primo classificatore predice circa il 60% dei dati nel gruppo 1 ed il restante nel gruppo 2 (pochissimi dati nel 3, risultati non significativi); invece il k-nn predice la quasi totalità dei cicli nel gruppo 1 e i restanti nel gruppo 2. In fig. 6.10a si vedono invece le curve della silhouette ad ogni blocco del gruppo di test per l'albero di decisione. Si vede chiaramente un peggioramento dei valori di silhouette nel tempo per i primi due gruppi, effetto che ovviamente non si riscontra in 3 non essendoci una numerosità di dati predetti sufficienti a perturbare il valore di coesione del gruppo. Risultati analoghi per il modello k-nn in figura 6.10b, con ovviamente anche il gruppo 2 senza perturbazioni non avendo molti nuovi dati predetti al suo interno.

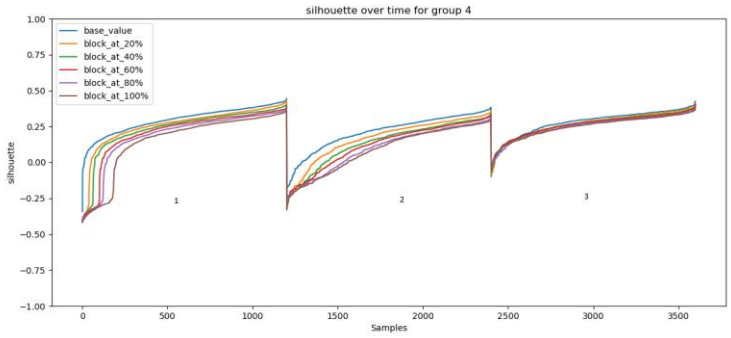


Fig. 6.10a curve di silhouette per i vari sottogruppi di test, modello l'Albero di Decisione

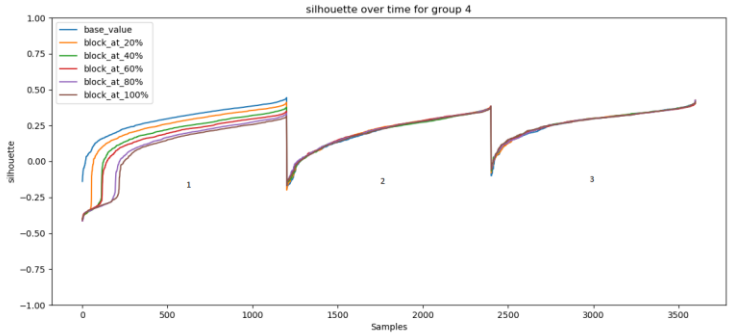


Fig. 6.10b curve di silhouette per i vari sottogruppi di test, modello k-nn

I risultati ottenuti per il gruppo 3 sono dovuti probabilmente ai pochi dati riguardanti questa classe nel dataframe di addestramento, poiché nella prima settimana si hanno pochi dati riguardanti questo gruppo.

Infatti, nel caso 2, prendendo invece in esame le prime due settimane, per cui il gruppo 3 aumenta la sua numerosità di 3500 cicli, ha forti differenze nei risultati di predizione per entrambi i modelli. In tabella 6.8a e 6.8b si osservano le predizioni rispettivamente per l'albero di decisione e il modello k-nn.

dT	block_at_20%	block_at_40%	block_at_60%	block_at_80%	block_at_100%
group 1	108	106	104	105	95
group 2	65	52	70	64	86
group 3	58	73	57	62	50

Tab. 6.8a predizioni gruppo di test, modello l'Albero di Decisione

knn	block_at_20%	block_at_40%	block_at_60%	block_at_80%	block_at_100%
group 1	132	133	129	114	103
group 2	42	58	59	68	76
group 3	57	40	43	49	52

Tab. 6.8b predizioni gruppo di test, modello k-nn

Si può notare una forte differenza dal caso 1; il primo classificatore divide abbastanza equamente i dati, anche se ne predice ancora una leggera maggioranza nel primo gruppo; comportamento simile ha il modello del k-nn. Per quanto riguarda i valori di silhouette invece, questa volta si possono osservare gli andamenti dei valori medi, in figura 6.11a ed in 6.11b.

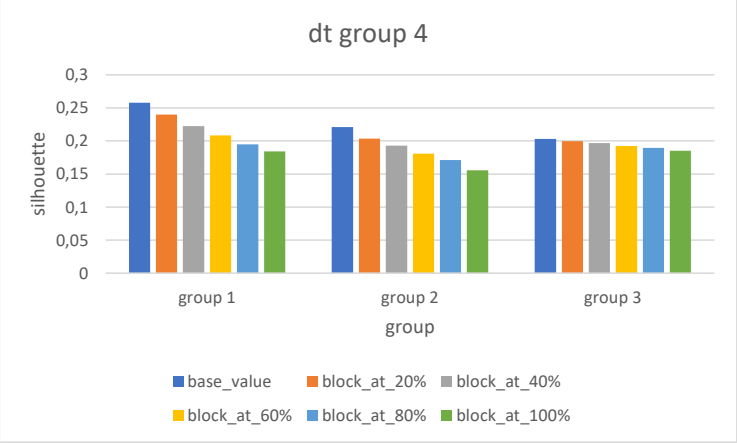


Fig. 6.11a valori di silhouette media dei sottogruppi, modello l'Albero di Decisione

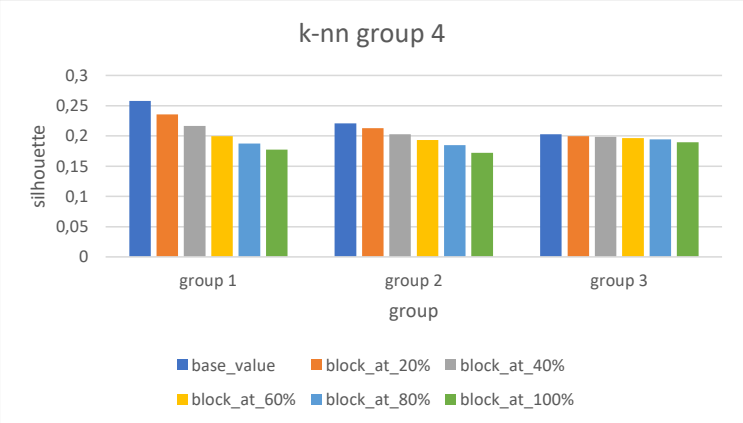


Fig. 6.11a valori di silhouette media dei sottogruppi, modello k-nn

Si può osservare molto bene in entrambi i casi un degrado nei valori di coesione dei gruppi, in particolare questo peggioramento è netto nelle prime due classi, che sono anche quelle considerate come valori centrali e più importanti. Tuttavia, per tutti i gruppi si hanno valori di silhouette media piuttosto bassi anche per il valore base, tutti inferiori a 0,3.

Il caso 3 presenta predizioni ancora più diversificate; come si può vedere in tabella 6.9a e 6.9b, in entrambi i modelli la maggior parte dei valori è stato predetto nel secondo gruppo. Addirittura, il secondo classificatore arriva a non predire quasi più nessun dato all'interno della classe 1.

dT	block_at_20%	block_at_40%	block_at_60%	block_at_80%	block_at_100%
group 1	51	54	59	58	31
group 2	108	108	100	98	124
group 3	72	69	62	75	76

Tab. 6.9a predizioni gruppo di test, modello l'Albero di Decisione

knn	block_at_20%	block_at_40%	block_at_60%	block_at_80%	block_at_100%
group 1	3	5	3	4	5
group 2	144	140	145	143	140
group 3	84	86	83	84	86

Tab. 6.9b predizioni gruppo di test, modello k-nn

Per quanto riguarda i grafici della silhouette, per una migliore comprensione anche in questo caso si è scelto di usare i valori medi, figure 6.12a e 6.12b; si può osservare un aumento dell'indice di coesione per il valore base dei primi due gruppi ed un netto peggioramento per il terzo che scende sotto il valore medio iniziale di 0,1. Il primo modello mostra la stessa tendenza presentata nel caso 2, quindi di degrado per ogni gruppo. Il k-nn invece, per il gruppo 1 crea una controtendenza, arrivando a migliorare i valori medi di silhouette nel tempo. Indipendentemente dal classificatore la terza classe mostra una lieve tendenza decrescente.

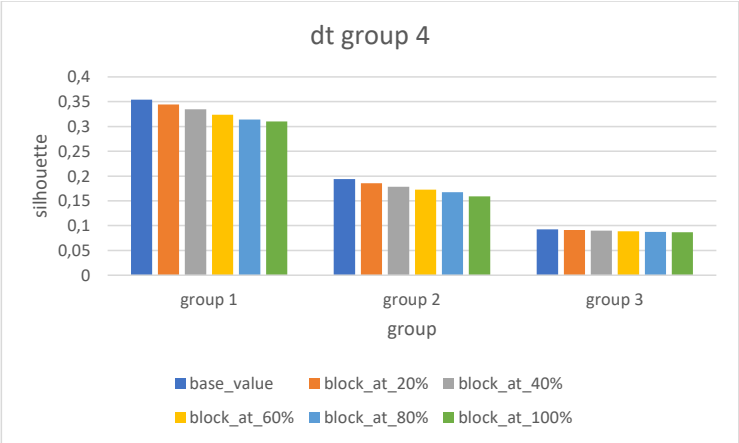


Fig. 6.12a valori di silhouette media dei sottogruppi, modello l'Albero di Decisione

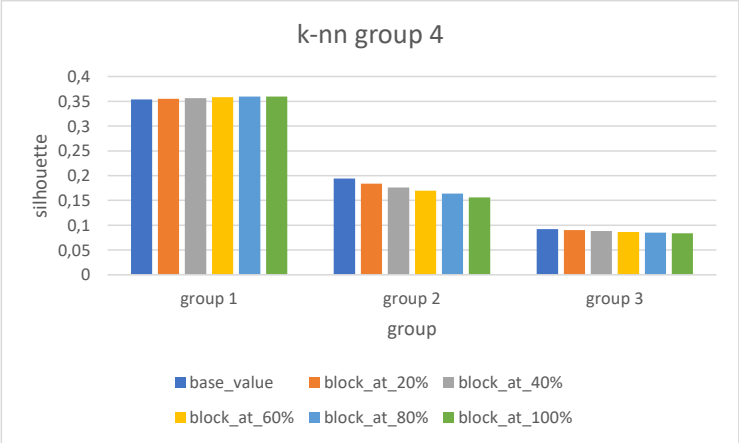


Fig. 6.12a valori di silhouette media dei sottogruppi, modello k-nn

In aggiunta alla silhouette si vuole riportare per questo caso anche il grafico riguardante il MSE; come per l'indice precedente, nel grafico in figura 6.13 si possono osservare 6 curve di valori campionati (numerosità sempre pari a 1200 per i motivi già citati), una per il valore base e poi una per ciascuno step progressivo, per il modello relativo all'albero decisionale.

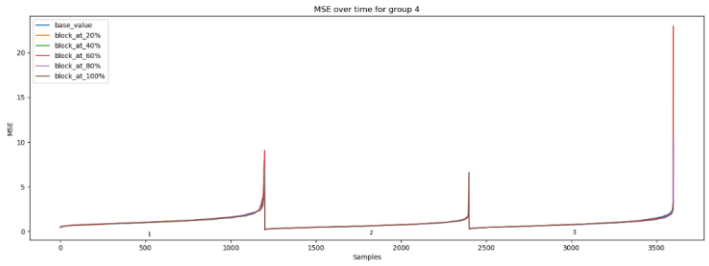


Fig. 6.13 valori di mse per sottogruppo, modello l'Albero di Decisione

Come si può facilmente comprendere i grafici sono illeggibili e di impossibile interpretazione poiché le curve sono sovrapposte per la maggior parte dei campioni; risultati molto simili sono stati trovati in tutti gli esperimenti indipendentemente dal caso e per questo non ne verranno riportati altri esempi, in quanto non utili ai fini delle analisi.

6.3.2.2 Secondo esperimento

Il secondo esperimento utilizza come gruppo di training 1 e 2 e punta ad esplorare appunto la coesione dei dati rispetto a queste classi che sono quelle obiettivo, cioè da raggiungere e mantenere in casi applicativi. Inoltre, come gruppi di test, oltre al già utilizzato 4 si aggiunge la classe 3, per cercare di capire quanto bene il modello si adatta a trovare un degrado anche se meno netto come il terzo gruppo rispetto alle classi 1 e 2.

I migliori parametri del modello dell'albero di decisione nel caso 1 sono: 'max_depth': 10, 'max_features': 'log2'; i sottogruppi di test sono pari sempre a 231 per il gruppo 4, costante per tutti i tre casi, essendo tutti i dati riguardanti questa classe concentrati nella prima settimana; per quanto riguarda il gruppo 3 invece il numero di elementi per sottogruppo è: 134 nel primo caso, 881 nel caso 2 e 1741 nel caso 3. Le predizioni per il gruppo 3 possono essere osservate in tabella 6.10a e 6.10b.

dT	block_at_20%	block_at_40%	block_at_60%	block_at_80%	block_at_100%
group 1	112	111	110	111	108
group 2	22	23	24	23	26

Tab. 6.10a predizioni gruppo di test 3, modello l'Albero di Decisione

knn	block_at_20%	block_at_40%	block_at_60%	block_at_80%	block_at_100%
group 1	131	134	133	134	134
group 2	2	0	1	0	0

Tab. 6.10b predizioni gruppo di test 3, modello k-nn

Per il gruppo 4 si hanno praticamente gli stessi numeri di elementi predetti per gruppo principale che si sono ottenuti nel caso 1 del primo esperimento e non si ritiene necessario ripeterli.

Per quanto riguarda invece i grafici della silhouette media, effettuando la predizione sulla classe numero 3 (fig. 6.14a e 6.14b) si ottengono valori simili in termini assoluti, ma il degrado del gruppo 2 è più netto rispetto al primo esperimento a parità di caso per il modello l'Albero di Decisione, per il k-nn invece non si notano grandi differenze.

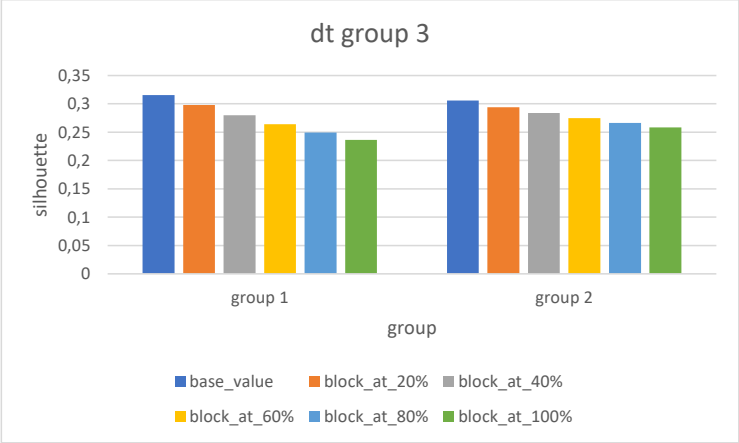


Fig. 6.14a valori di silhouette media dei sottogruppi, modello l'Albero di Decisione

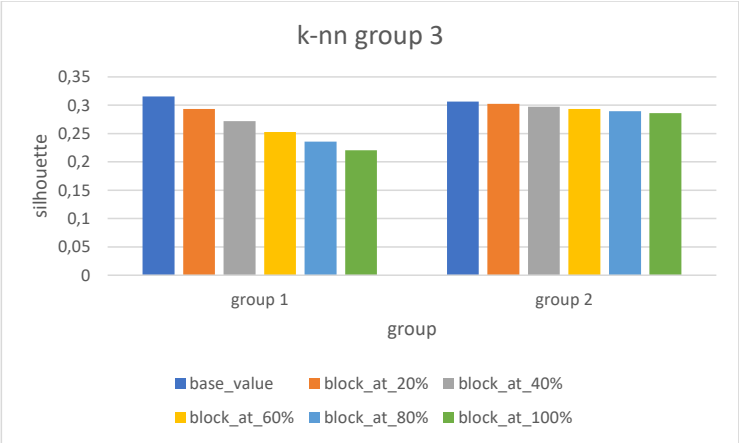


Fig. 6.14a valori di silhouette media dei sottogruppi, modello k-nn

Il gruppo 4 invece ottiene un effetto molto degradante sulla coesione dei gruppi 1 e 2, come si nota in figura 6.15 per il modello albero di decisione, dove la decrescita dei valori medi è molto più netta per entrambi le classi ma in particolare per 1, che arriva ad avere un valore medio minore di 0,15, partendo dal valore base intorno a 0,33. L'effetto di peggioramento è attenuato nel classificatore k-nn per quanto riguarda il gruppo 2 e rimane invece uguale per il primo, esattamente come per il gruppo 3.

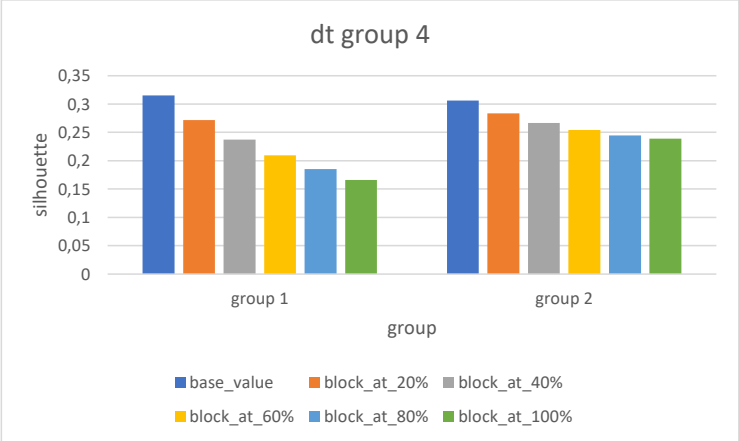


Fig. 6.15 valori di silhouette media dei sottogruppi, modello l'Albero di Decisione

Nel caso 2 cambiano i valori migliori per creare l'albero di decisione del modello, infatti varia la massima profondità che passa da 10 a 5, rimane invariato invece il massimo delle variabili. I risultati riguardanti il test del quarto gruppo sono praticamente identici sia per predizioni che per valori medi di silhouette, poiché cambia solo leggermente la numerosità della seconda classe usata nell'addestramento del modello, quindi non vengono qui riportati. Invece cambia ampliamente il numero di elementi del gruppo 3, portando ad 881 dati per ogni sottogruppo. In tabella 6.11a e 6.11b vengono riportate le predizioni.

dT	block_at_20%	block_at_40%	block_at_60%	block_at_80%	block_at_100%
group 1	597	601	588	597	598
group 2	214	210	223	214	213

Tab. 6.11a predizioni gruppo di test 3, modello l'Albero di Decisione

knn	block_at_20%	block_at_40%	block_at_60%	block_at_80%	block_at_100%
group 1	792	796	799	784	789
group 2	19	15	12	27	22

Tab. 6.11b predizioni gruppo di test 3, modello k-nn

Come si può osservare l'Albero di Decisione predice più del 70% dei dati nel gruppo 1, invece il k-nn la quasi totalità dei campioni viene assegnata alla prima classe principale. I valori della silhouette calcolati dopo l'utilizzo dell'albero di decisione vengono descritti in figura 6.16 per quanto riguarda i valori medi, dove si denota un netto degrado della coesione dei gruppi; non si riportano i valori medi riguardanti il k-nn poiché ha lo stesso effetto già visto nei casi precedenti di lasciare invariati i valori medi del primo gruppo e di migliorare leggermente la coesione del secondo.

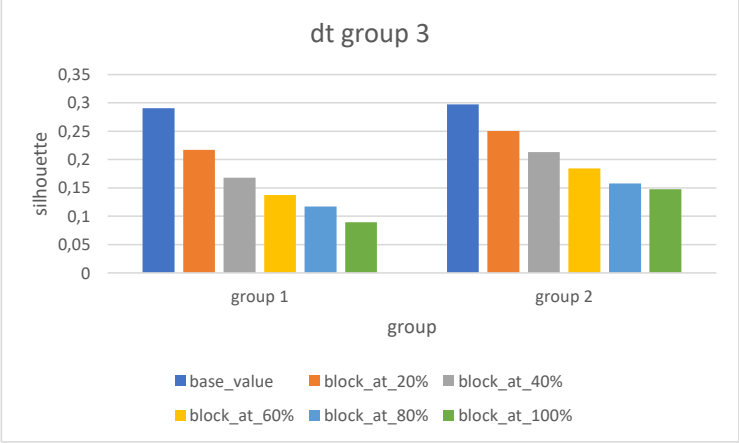


Fig. 6.16 valori di silhouette media dei sottogruppi, modello l'Albero di Decisione

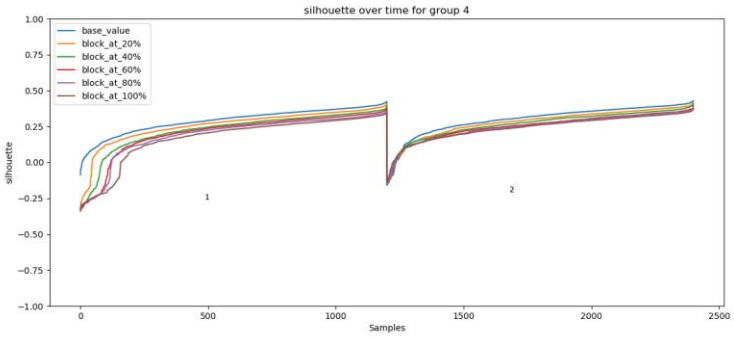


Fig. 6.17 curve di silhouette per i vari sottogruppi di test, modello l'Albero di Decisione

Invece in figura 6.17 si possono osservare le curve della silhouette riguardante sempre lo stesso modello, dove è netta la distinzione tra le curve per entrambe le classi principali e si osserva molto bene il peggioramento nel tempo dell'indice di dispersione. Per questo test si è voluto approfondire ulteriormente la conoscenza sul degrado del segnale rispetto al gruppo 4; per questo motivo si è aggiunto l'indice MAAPE. Rispetto alla figura 6.17 il MAAPE è il calcolo del valore medio in arcotangente della differenza percentuale tra la curva base, che viene identificata come valore effettivo e le singole curve di ogni step, considerate ad una ad una come valore predetto. In figura 6.18 si vedono i valori del MAAPE per i due gruppi ad ogni step per il modello del l'Albero di Decisione:

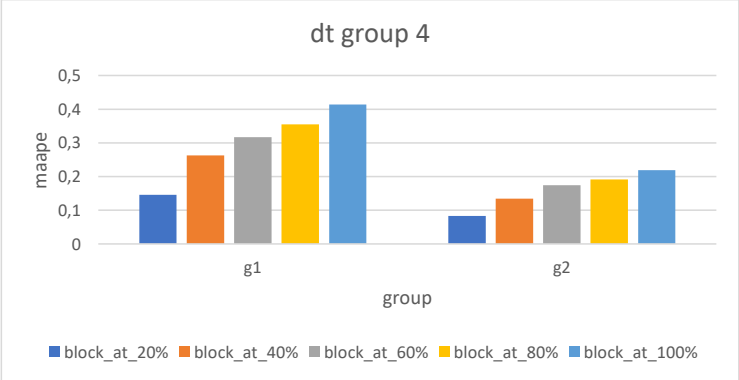


Fig. 6.18 maape per sottogruppo, modello l'Albero di Decisione

Si osserva che l'errore è molto più alto per il gruppo 1 e come ci si aspettava si osserva anche per questo indice un degrado, evidenziato dal crescere del valore del MAAPE.

Infine, per il caso 3 cambiano ancora i migliori parametri del l'Albero di Decisione in: 'max_depth': 5, 'max_features': 'auto'; questo risulta il caso più diverso affrontato fino ad ora, con i valori di predizione che cambiano nettamente sia per il gruppo 3, i dati sono riportati in tabella 6.12a e 6.12b, che per il gruppo 4 i cui dati invece si trovano in tabella 6.13a e 6.13b.

dT	block_at_20%	block_at_40%	block_at_60%	block_at_80%	block_at_100%
group 1	798	790	802	784	798
group 2	943	951	939	957	943

Tab. 6.12a predizioni gruppo di test 3, modello l'Albero di Decisione

knn	block_at_20%	block_at_40%	block_at_60%	block_at_80%	block_at_100%
group 1	675	744	791	645	751
group 2	1066	977	950	1096	990

Tab. 6.12b predizioni gruppo di test 3, modello k-nn

dT	block_at_20%	block_at_40%	block_at_60%	block_at_80%	block_at_100%
group 1	110	106	104	106	95
group 2	121	125	127	125	136

Tab. 6.13a predizioni gruppo di test 4, modello l'Albero di Decisione

Knn	block_at_20%	block_at_40%	block_at_60%	block_at_80%	block_at_100%
group 1	94	91	89	88	75
group 2	137	140	142	143	156

Tab. 6.13b predizioni gruppo di test 4, modello k-nn

Come si può notare il primo modello classifica in maniera abbastanza equivalente i dati tra i due gruppi principali per entrambi i gruppi di test utilizzati; invece, il k-nn, propende per entrambi i gruppi esterni a inserire più dati all'interno del gruppo 2. Questa tendenza si ripercuote ovviamente sui valori medi della silhouette, che vengono riportati in figura 6.19a e 6.19b per la classe 3. In questi grafici, in particolare quello del l'Albero di Decisione, si può osservare un vero e proprio tracollo della coesione dei gruppi all'avanzare del tempo, tanto che il gruppo 1 raggiunge quasi lo zero di valore della silhouette; in questo caso neanche il modello creato tramite il k-nn riesce a mitigare questo degrado.

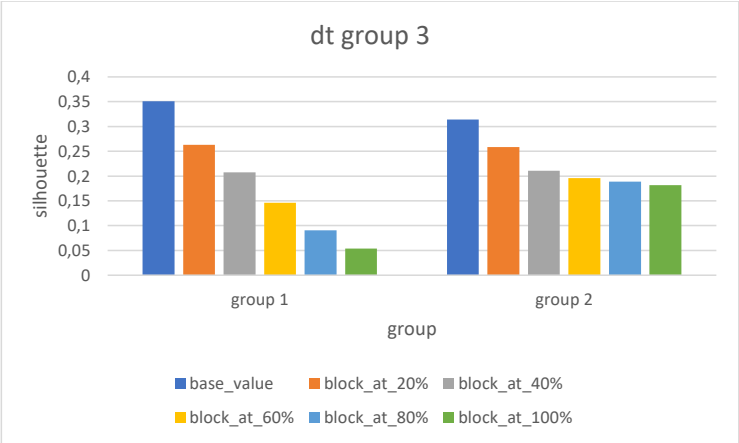


Fig. 6.19a valori di silhouette media dei sottogruppi, modello l'Albero di Decisione

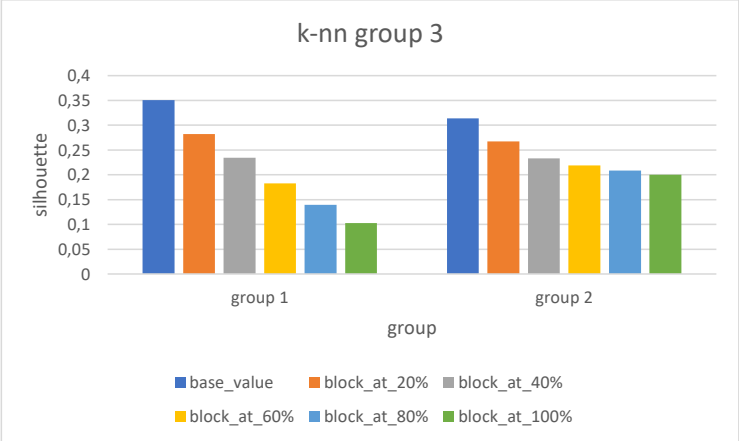


Fig. 6.19b valori di silhouette media dei sottogruppi, modello k-nn

Infine, in figura 6.20 si può osservare il peggioramento dei valori di coesione inerenti al gruppo 4 con l'utilizzo del modello dell'albero di decisione; in questo caso si ha un peggioramento decisamente meno marcato rispetto a quello visto per il gruppo 3, inoltre il classificatore k-nn porta a praticamente gli stessi risultati e per questo non è stato incluso.

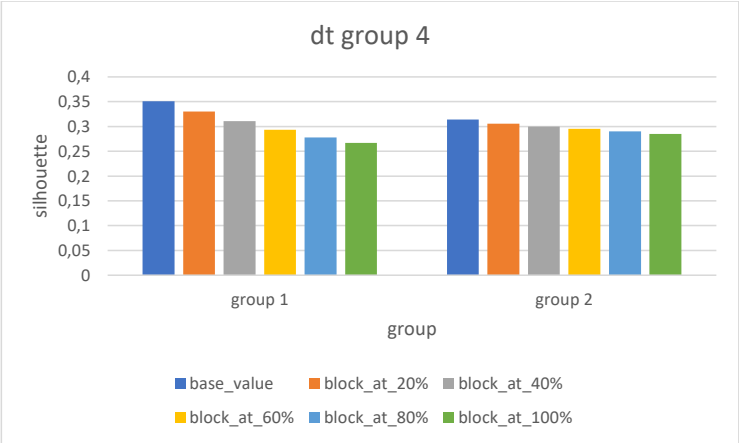


Fig. 6.20 valori di silhouette media dei sottogruppi, modello l'Albero di Decisione

6.3.2.3 Terzo esperimento

L'ultimo esperimento è stato scelto per esplorare meglio il comportamento del gruppo 2, insieme ai valori di tensione leggermente peggiori dati dal gruppo 3; considerato come valore migliore possibile, si vuole comprendere se il modello è in grado di individuare un degrado intenso come il gruppo 4 rispetto al più leggero gruppo 3 e segnalare il peggioramento correttamente. Il test è stato fatto solo sulla classe 4, con i soliti 231 elementi per ogni sottogruppo.

Per il primo caso, cioè la prima settimana, si ha sempre i migliori parametri uscenti dalla validazione dell'albero di decisione, che in questo caso risultano: 'max_depth': 10, 'max_features': 'log2'. Come sempre le predizioni sono riassunte nelle tabelle seguenti, 6.14a e 6.14b:

dt	block_at_20%	block_at_40%	block_at_60%	block_at_80%	block_at_100%
group 2	159	165	152	160	162
group 3	72	66	79	71	69

Tab. 6.14a predizioni gruppo di test 4, modello l'Albero di Decisione

knn	block_at_20%	block_at_40%	block_at_60%	block_at_80%	block_at_100%
group 2	161	66	163	70	141
group 3	70	165	68	161	90

Tab. 6.14b predizioni gruppo di test 4, modello k-nn

Come era prevedibile più del 30% dei dati viene predetto appartenente alla classe 3, che rappresenta già un valore di peggioramento dallo standard; tuttavia vengono predetti un numero sufficienti di dati da far peggiorare la silhouette, se si osserva figura 6.21, si nota il netto peggioramento dell'indice in questione sia per il gruppo 3 che per il 2. Essendo i grafici identici praticamente tra l'Albero di Decisione e k-nn, viene riportato solo il primo.

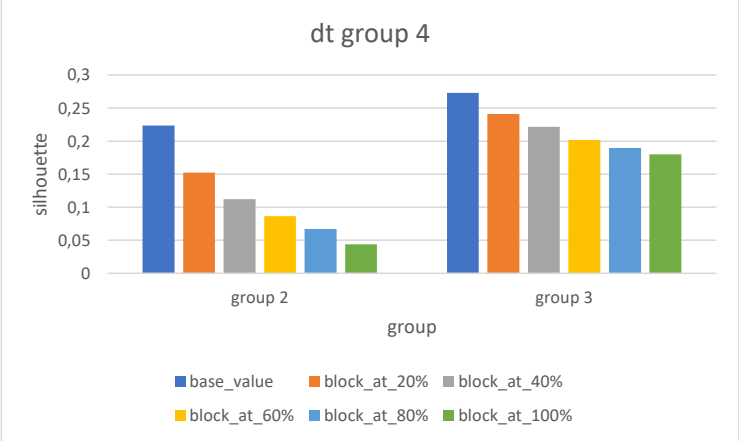


Fig. 6.21 valori di silhouette media dei sottogruppi, modello l'Albero di Decisione

Se si considerano le prime due settimane invece, si nota un totale cambiamento dei parametri migliori per l'albero, con la massima profondità che passa a 20 livelli e il numero massimo di variabili diventa 'auto'. Anche le predizioni cambiano radicalmente, vedasi tabella 6.15a e 6.15b:

dt	block_at_20%	block_at_40%	block_at_60%	block_at_80%	block_at_100%
group 2	57	41	51	50	33
group 3	174	190	180	181	198

Tab. 6.15a predizioni gruppo di test 4, modello l'Albero di Decisione

knn	block_at_20%	block_at_40%	block_at_60%	block_at_80%	block_at_100%
group 2	37	35	33	37	22
group 3	194	196	198	194	209

Tab. 6.15b predizioni gruppo di test 4, modello k-nn

Questo spostamento nei valori predetti influisce pesantemente sui valori medi della silhouette, vedere figura 6.22.

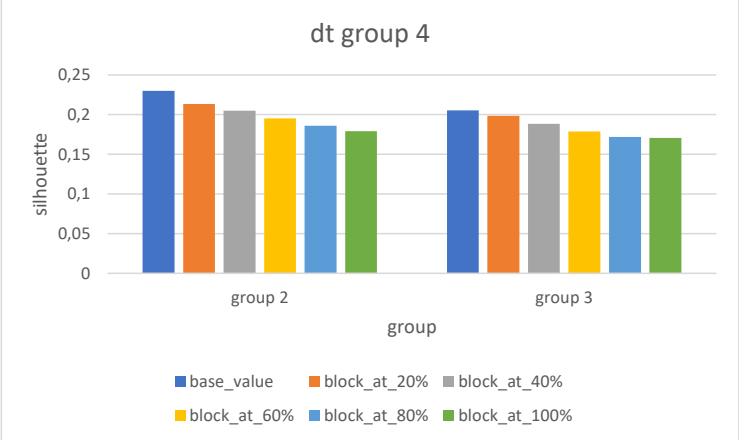


Fig. 6.22 valori di silhouette media dei sottogruppi, modello l'Albero di Decisione

I valori medi base sono più alti, in particolare per il gruppo 3 ed il degrado ai vari step è molto meno marcato rispetto al primo caso di questo esperimento; il grande aumento della numerosità del terzo gruppo porta ad una maggiore coesione e ad un effetto perturbativo dei sottogruppi di test inferiore. Si può osservare questo avvicinamento dei valori nelle curve di silhouette della figura 6.23, sempre ricavate dal modello con l'albero di decisione; qui, a parte un iniziale differenza nel gruppo 2, le curve tendono poi a convergere in entrambi i gruppi. Questi stessi effetti appena descritti nelle due figure si possono osservare uguali per il modello basato sul k-nn.

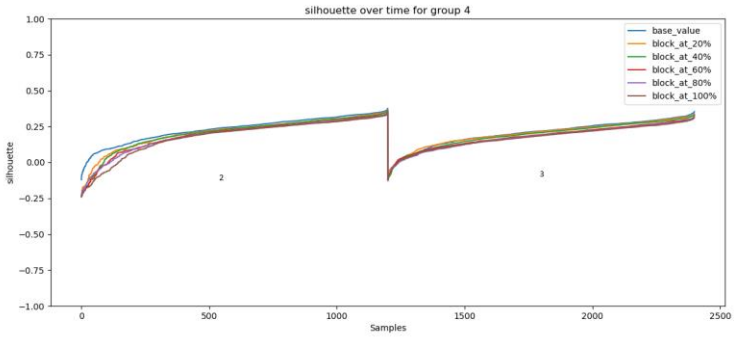


Fig. 6.23 curve di silhouette per i vari sottogruppi di test, modello l'Albero di Decisione

Infine, si osservano i risultati del caso 3 che tiene in considerazione l'intero dataset, dove i parametri non cambiano rispetto alla seconda casistica. In tab 6.16a e 6.16b si osservano le predizioni, che invertono la tendenza vista precedente, assegnando un numero leggermente più grande di dati alla seconda classe.

dt	block_at_20%	block_at_40%	block_at_60%	block_at_80%	block_at_100%
group 2	95	96	98	104	91
group 3	136	135	133	127	140

Tab. 6.16a predizioni gruppo di test 4, modello l'Albero di Decisione

knn	block_at_20%	block_at_40%	block_at_60%	block_at_80%	block_at_100%
group 2	76	72	90	92	63
group 3	155	159	141	139	168

Tab. 6.16b predizioni gruppo di test 4, modello k-nn

Questo ennesimo cambiamento nel trend delle predizioni provoca ovviamente diversi valori di silhouette, vedasi figura 6.24a e 6.24b:

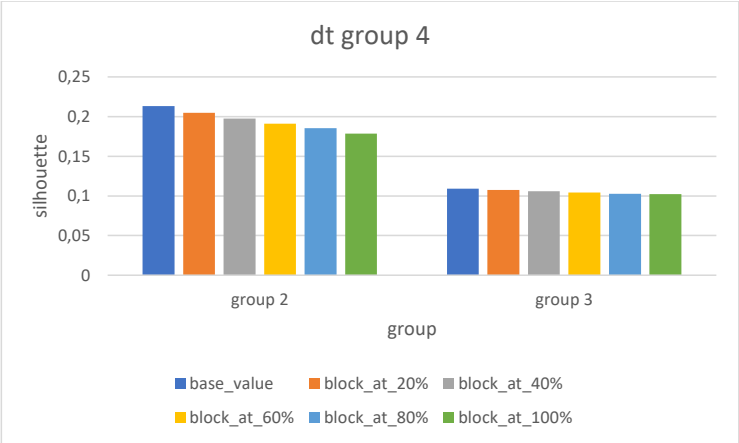


Fig. 6.24a valori di silhouette media dei sottogruppi, modello l'Albero di Decisione

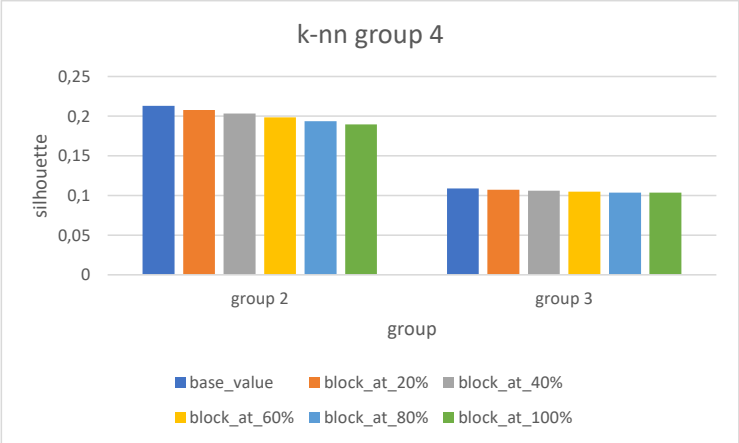


Fig. 6.24a valori di silhouette media dei sottogruppi, modello k-nn

Intanto si notano valori di coesione base decisamente inferiori nel gruppo 3, stabili per il secondo; tuttavia il degrado è bassissimo nel terzo gruppo, tanto che il modello basato sul k-nn praticamente non rileva nei valori medi di silhouette, invece continua ad essere presente il peggioramento per la classe 2.

7. Conclusioni

In conclusione, si sono ricavati diversi risultati interessanti dalle due tipologie di analisi condotte in questo elaborato. Per quanto riguarda il modello generale e la ricerca del numero minimo di dati per addestrare un modello funzionante che ottenga buoni risultati in termini di accuratezza, precisione e richiamo, si è visto che molto dipende dalla grandezza del dataset iniziale e dal tipo di campionamento che si utilizza. In casi di dataset molto ampi e strutturati su molti giorni di lavoro del macchinario, è consigliato utilizzare un'alta percentuale del dataset iniziale oppure piccole porzioni ma ricavate tramite un campionamento stratificato casuale eseguito su tutte le classi disponibili. Se invece, si vuole tener conto della serie storica dei dati, sicuramente è necessario usare un dataset relativamente piccolo per eseguire l'addestramento ed il test, come si è appurato nella serie di dieci esperimenti progressivi per il tempo, utilizzando sempre un campionamento stratificato. Queste evidenze sono vere per entrambi i classificatori utilizzati, l'albero di decisione ed il K-Nearest Neighbor, ma in particolare il secondo ha mostrato performance sempre leggermente migliori al l'Albero di Decisione, indipendentemente dal campionamento o dataset utilizzato, per brillare in maniera più netta nei dieci esperimenti finali. Il sistema preso in esame potrebbe, quindi, essere usato per l'individuazione di malfunzionamenti nella macchina, poiché è estremamente sensibile alle variazioni, come si è osservato usando l'intero dataset ed estraendo i dati in ordine temporale si sono ottenute pessime performance, segno che c'è stato un probabile degrado nel funzionamento del macchinario preso in esame.

Da questo punto nasce e si sviluppa il secondo filone di analisi, quelle relative al modello evolutivo, che punta ad individuare un modello in grado di evidenziare il progressivo degrado nelle performance del braccio robot. I risultati qui riportati sono quelli che maggiormente indicano la presenza di un peggioramento e quindi la capacità di entrambi i classificatori utilizzati, l'albero di decisione ed il K-Nearest Neighbor, d'individuare un degrado nella coesione delle classi analizzate, anche se sicuramente il primo modello è quello più performante da questo punto di vista. Questi risultati si sono verificati in quasi tutti gli esperimenti ed i casi condotti, ma desta particolare interesse quello riguardante il caso che utilizza i gruppi 1 e 2 come dati di training per il modello, poiché come è stato più volte sottolineato, sono le classi con maggiori applicazioni industriali e quindi considerate le più importanti. In questi casi entrambi i modelli si sono mostrati in grado di individuare il degrado nel tempo associato ad entrambi i gruppi di test, il 3 ed il 4, indipendentemente dalla finestra temporale considerata; è stato inoltre utilizzato un secondo metodo, il MAAPE, che mostra un progressivo aumento dell'errore nella silhouette, che quindi

affianca l'analisi dei valori medi come metodo in grado d'individuare il peggioramento delle performance. Interessante fatto da notare per questo esperimento è che, contrariamente alle supposizioni iniziali che il gruppo 2 e 4 fossero molto simili, i dati di quest'ultima classe vengono più spesso predetti nella classe 1 piuttosto che nella 2, soprattutto se si utilizzano le prime due finestre temporali descritte e solo nella terza i dati vengono distribuiti al 50% tra gruppo 1 e 2.

Future ricerche si potrebbero indirizzare alla creazione di un unico modello che parta da una piccola percentuale di dati per l'addestramento di un classificatore idoneo poi all'individuazione del degrado delle performance della macchina, tramite l'utilizzo della silhouette o di altri indici di coesione ancora da testare; inoltre sarebbe interessante approfondire l'utilizzo dei due classificatori e le discrepanze osservate tra i due modelli, visto che nel modello generale è il k-nn ad avere performance migliori per la creazione di un modello a partire da pochi dati, ma poi sia l'Albero di Decisione il classificatore migliore per individuare un degrado nel segnale della corrente e quindi nella coesione delle classi.

Inoltre si potrebbero strutturare i dati in maniera diversa con l'aggiunta di una nuova variabile come la temperatura, così da poterne correlare le variazioni alle varie statistiche calcolate in modo da raffinare ulteriormente i modelli utilizzati; infine sarebbe anche interessante organizzare i dati con un ulteriore approccio: dopo aver costruito un modello funzionante per l'individuazione di guasti, può essere un' opzione condurre una nuova serie di monitoraggi del braccio robot, così da avere dati utili per studi del tipo survival analysis, che sono una tecnica d'individuazione delle failures, usata in particolare nei campi medici e della manutenzione predittiva.

8. Bibliografia

- Almanda-Lobo F., *"The Industry 4.0 revolution and the future of Manufacturing Execution Systems (MES)",* Journal of Innovation Management JIM 3, 4 16-21, (2015)
- Anandkumar A., Ge R., Hsu D., Kakade S., Telgarsky M., *"Tensor Decompositions for Learning Latent Variable Models"*, Journal of Machine Learning Research (JMLR) 15: 2773–2832., (2014)
- Y. Bengio; A. Courville, P. Vincent,*"Representation Learning: A Review and New Perspectives"*, IEEE Trans. PAMI, special issue Learning Deep Architectures. 35: 1798–1828. arXiv:1206.5538. doi:10.1109/tpami.2013.50, (2013)
- Bishop C. M., "*Pattern Recognition and Machine Learning",* Springer, ISBN 0-387-31073-8, (2006)
- Bonekamp L., Sure M., *" Consequences of Industry 4.0 on Human Labourand Work Organisation"*, Journal of Business and Media Psychology 6, (2015)
- A. Botticini, A. Pasetto, Z. Rotondi, "*Sviluppo e prospettive dell'industria 4.0 in Italia e ruolo strategico del credito",* Università degli Studi di Urbino, (2016)
- Brynjolfsson E., & McAfee A., *"The Second Machine Age. Work, Progress, and Prosperity in a Time of Brilliant Technologies"*, New York: W.W. Norton & Company, (2014)
- Busoniu *L.,*Babuska R, *De Schut*ter T.,Ernst D., "*Reinforcement Learning and Dynamic ProgrammingusingFunctionApproximators*.", Taylor & Francis CRC Press. *ISBN 978-1-4398-2108-4*., (2010)
- Foidl H., Felderer M., *" Research Challenges of Industry 4.0 for Quality Management",* Part of the Lecture Notes in Business Information Processing book series (LNBIP, volume 245), (2 aprile, 2016)
- Garcia M., Sanz-Bobi M.A., del Pico J., *" SIMAP: Intelligent System for Predictive Maintenance Application to the health condition monitoring of a windturbine gearbox",* Computers in Industry Volume 57, Issue 6, Pages 552-568, (August 2006)
- van Gerven M and Bohte S, *"Editorial: ArtificialNeural Networks as Models ofNeural Information Processing.",*Front. Comput. Neurosci. 11:114. doi: 10.3389/fncom.2017.00114, (19 dicembre, 2017)
- Goriveau R., Medjaher K., Zerhouni N.,*"Fromprognostics and healthsystemsmanagementtopredictivemaintenance: monitoring and prognostics".* ISTE Ltd and John Wiley &Sons, Inc, (novembre, 2016)
- Gylchrist A., "*Industry 4.0: Th e Industrial Internet of Things",*Apress, (2016)
- Hahsler Michael*, "Introduction to arules – A computational environment for mining association rules and frequent item sets"* . Journal of Statistical Software, (2005)
- Hermann H., T. Pentek, B. Otto, ''*Design principles for industry 4.0 scenarios,''* Working Paper, Tech. Univ. Dortmund, Dortmund, Germany, (Feb. 2015)
- Hodge, V. J.; Austin, J., *"A Survey of Outlier Detection Methodologies Artificial Intelligence*", Review. 22 (2): 85–126. CiteSeerX 10.1.1.318.4023. doi:10.1007/s10462-004-4304-y, (2004)
- Kennedy S., *"New tools for PdM",* Plant Service, (2006*)*
- Koza, John R.; Bennett, Forrest H.; Andre, David; Keane, Martin A., "*Automated Design of Both theTopology and Sizingof Analog ElectricalCircuitsUsing Genetic Programming*.", Artificial Intelligence in Design '96. Springer, Dordrecht. pp. 151–170, (1996)

- J. Lee, B. Bagheri, and H.-A. Kao, *"Recent Advances and Trends of Cyber-Physical Systems and Big Data Analytics in Industrial informatics*", Proceeding of Int. Conference on Industrial Informatics (INDIN), (2014)
- J. Lee, B. Bagheri, and H.-A. Kao, *''A cyber-physical systems architecture for industry 4.0-based manufacturing systems",* Manuf. Lett., vol. 3, pp. 18–23, (gennaio, 2015)
- Miller RG Jr, *"Survival Analysis",* John Wiley &Sons, (25 gennaio, 2011)
- Mobley R. K.*,"An introduction to predictive maintenance",*(2nd ed.). Butterworth-Heinemann., (26 settembre, 2002)
- Egon Mueller, Xiao-Li Chen, Ralph Riedel, *"Challenges and Requirements for the Application of Industry 4.0: A Special Insight with the Usage of Cyber-Physical System",* Chin. J. Mech. Eng. (2017) 30:1050–1057, DOI 10.1007/s10033-017-0164-7
- Cathy O'Neil, Rachel Schutt, *"Doing Data Science: Straight Talk fromthe Frontline",* O'Reilly Media Inc, (9 ottobre, 2013) pp 21-51
- Peng, K., "Equipment Management in the Post-Maintenance Era: A New Alternative to Total Productive Maintenance (TPM)", CRC Press. pp. 132–136, (2012)
- Rayna, T., & Striukova, L. (2016, Aprile 25). *"360° Business Model Innovation: Toward an Integrated View of Business Model Innovation",*Research-Technology Management, 21-28. doi:10.1080/08956308.2016.1161401
- Robin L., *"Slick tricks in oil analysis" Plant Services*, (2006)
- Schmidt R., Möhring M., Härting RC., Reichstein C., Neumaier P., Jozinović P., „In*dustry 4.0 - Potentials forCreating smart Products: Empirical Research Results*.", In: Abramowicz W. (eds) Business Information Systems. BIS (2015), Lecture Notes in Business Information Processing, vol 208. Springer, Cham
- Schröder C.,*"The Challenges of Industry 4.0 for Small and Medium-sized Enterprises"*,THE FRIEDRICH-EBERT-STIFTUNG, (2017)
- Yung C., "*Vibration analysis: what does it mean?*", *Plant Services, (2006)*
- Zhou, T. Liu, and L. Zhou, *''Industry 4.0: Towards future industrial opportunities and challenges,''* in Proc. Int. Conf. Fuzzy Syst. Knowl. Discovery, (agosto, 2015), pp. 2147–2152.

9. Sitografia

- Aliperto D., "*Verso la vera Predictive Maintenance, come fondere IoT e Machine learning*", Internet 4 Things, (25 giugno, 2018), www.internet4things.it/industry-4-0
- Bucceri G., *"Che cos'è l'Industria 4.0 e perché è importante saperla affrontare",* Network Marketing, (12 novembre, 2017) http://www.networktomarketing.it/notizie
- Bi-Survey, *"Insufficient Skills Are Curbing the Big Data Boom*", (2018), www.bi-survey.com
- Boldrini N., *"Cos'è il Machine Learning, come funziona e quali sono le sue applicazioni"*, AI4BUSINESS, (19 dicembre,2018), www.ai4business.it/intelligenzaartificiale/machine-learning
- Crivelli D., *"Cinghie di trasmissione: guida rapida dalle tipologie al tensionamento",* il Progettista Industriale, (24 maggio, 2013), www.ilprogettistaindustriale.it
- Della Mura M., "*Big Data e IoT: la sfida della predictive maintenance*", Big Data 4 Innovation, (3 gennaio, 2018), www.bigdata4innovation.it/big-data/
- Engel G., *"3 Flavors of Machine Learning: Who, What & Where*", Dark Reading, (2 novembre, 2016), www.darkreading.com/threat-intelligence
- Fiix, "*Predictive Maintenance*", (visitato il 20 gennaio, 2018), www.fiixsoftware.com/maintenance-strategies
- Gates, *"Come prevenire e diagnosticare lo spostamento laterale della cinghia di distribuzione",* Gates TechZone, (24 febbraio, 2017), www.gatestechzone.com/it/notizie
- Geitgey A., *"Machine Learning is Fun!"*, Medium, (5 maggio, 2014), www.medium.com
- Griffith E., *"What is it cloud computing?"*, PcMag Australia, (3 maggio, 2016), www.au.pcmag.com/networking-communications-software/29902
- IBT Inc, *"How To Tension a V-Belt (& How Not To)"*, IBT Inc, (5 febbraio, 2018), www.ibt-inc.com
- Lena," Applications *of augmented reality: What are the potentials for the manufacturing industry?"*, Code-N, (23 ottobre, 2017), www.code-n.org/blog
- Maci L., "*Che cos'è l'Industria 4.0 e perché è importante saperla affrontare"*, Economy Up, (28 novembre, 2018), www.economy.it/innovazione
- McClelland C., "*Data Analytics vs. Machine LearningWhat's the difference? Which should your business use? And how does machine learning apply to IoT?",*Medium, adapated from book published by Leverege, (30 aprile, 2018), www.medium.com/iotforall
- McKinsey & Company*, "Industry 4.0 How to navigate digitization of the manufacturing sector",* (aprile, 2015), www.mckinsey.com/business-functions/operations/our-insights
- McKinsey & Company*, "Are you ready for the 4th industrial revolution?",* (2017), www.mckinsey.com/business-functions/operations
- Marr B., "*What is Industry 4.0? Here's A Super Easy Explanation for Anyone ",* Forbes, (2 settembre, 2018), www.forbes.com/sites/bernardmarr/2018/09/02
- Moschese G., *"Data Mining : tecniche di trasformazione dei dati (Parte quarta)*", Apoge Online, (11 novembre, 2004), http://www.apogeonline.com/webzine
- Piesync, "*Top 5 Problems with Big Data (and how to solve them)* ", Piesync Blog, (7 febbraio, 2018), www.piesync.com/blog
- Reynolds C., *"AR Goggles for Enterprise Use: Business Interest and Product Range Heat Up"*Computer Business Review, (8 maggio, 2018), www.cbronline.com
- Reuters, "Amazon ditched AI recruiting tool that favored men for technical jobs", The Guardian, (11 ottobre, 2018), www.theguardian.com/technology/2018/oct/10

- Romano C., *"Machine learning e guida autonoma: Apple sta facendo progressi",* Digital Day, (11 dicembre, 2012), www.dday.it/redazione
- Sanjeevi M, *"Chapter 4: DecisionTreesAlgorithms",* Medium, (6 ottobre, 2017), www. medium.com
- Scalabre O., „*Embracing Industry 4.0 and Rediscovering ",* Boston Consulting Group, (2018), www.bcg.com/it-it/capabilities/operations
- Suntec India Blog,"Clean Data in CRM: The Key to Generate Sales-Ready Leads and Boost Your Revenue Pool", (20 febbraio, 2016), www.suntecindia.com/blog
- *"Industria 4.0"*, (dicembre 2018), https://it.wikipedia.org/wiki/Industria_4.0
- Corcom, *"IoT per il trasporto aereo, asse Rolls Royce-Microsoft ",* Corriere Comunicazioni, (20 luglio, 2016), www.corrierecomunicazioni.it/digital-economy/cloud
- Vance R., "*Do You Know The Difference Between Data Analytics And AI Machine Learning?"*, Forbes, (1 agosto, 2018), www.forbes.com/sites/forbesagencycouncil
- Vinovski J., *"Beyond Industry 4.0: Getting To Tomorrow's Manufacturing Solutions Today ",* Forbes, (25 novembre, 2018), www.forbes.com/sites/jimvinoski
- Zubani M., *"Manutenzione Predittiva: tipi di manutenzioni a confronto e vantaggi"*, T4SM, (5 giugno, 2018), https://www.toolsforsmartminds.com/it/insight/blog

Legenda

Mean Arctangent Absolute Percentage Error (MAAPE)

Manutenzione Predittiva (MP)

Network attached storage (NAS)

Amazon Web Service (AWS)

Advanced manufacturing solutions (AMS)

l'Internet of Thing (IoT)

Cyber-Physical System (sistemi fisico-digitali, CPS)

'Internet of Service (IoS)

piccole e medie imprese (PMI)

principal component analysis (PCA)

Decomposizione in valori singolari (SVD)

Mean Sqaure Error (MSE)

Sum of Square Error (SSE)

K-Nearest Neighbor (K-NN)

Machine Learning (ML)

computerized maintenance management system (CMMS)

Intelligent System for Predictive Maintenance (SIMAP)

Root Mean Square (RMS RMS_LP e RMS_BP)

Ringraziamenti

Alla Prof. Cerquitelli e Al Dott. Ventura per l'aiuto accademico.
Alla mia famiglia per avermi permesso di seguire questo percorso e di arrivare fino a questo punto.
A tutti i miei amici per i consigli e il supporto in questi lunghi anni di Università.
Infine, a Erika, la persona che mi ha dato la spinta giusta a finire e senza la quale difficilmente avrei concluso così splendidamente come con lei al mio fianco.