Politecnico di Torino

Laurea Magistrale in Ingegneria Gestionale

TESI DI LAUREA

Studio e Analisi dei Big Data per lo sviluppo di un modello di manutenzione predittiva in ambito veicolare

Laureando:

Lorena Crapa

Relatore:

Ch.mo Prof.ssa Tania Cerquitelli

Co-relatori:

Daniele Apiletti

Simone Monaco

ANNO ACCADEMICO 2021/2022

Politecnico di Torino

Laurea Magistrale in Ingegneria Gestionale

TESI DI LAUREA

Studio e Analisi dei Big Data per lo sviluppo di un modello di manutenzione predittiva in ambito veicolare

Laureando:

Lorena Crapa

Relatore:

Ch.mo Prof.ssa Tania Cerquitelli

Co-relatori:

Daniele Apiletti

Simone Monaco

ANNO ACCADEMICO 2021/2022

Quella mossa che avevi paura di fare potrebbe essere proprio quella giusta.

Abstract - Italiano

Nell'epoca attuale, l'industria 4.0 ha portato il tema dell'efficientamento in primo piano e ha modificato e moltiplicato le esigenze manutentive di veicoli connessi. Grazie alla capacità di monitoraggio delle condizioni, alla possibilità di raccogliere e analizzare i Big Data prelevati dai sensori di bordo e al supporto di strumenti tecnologici, si è in grado di apprendere e fornire nuove informazioni elaborando i dati che l'IoT industriale fornisce.
Questo lavoro di tesi ripercorre gli aspetti fondamentali dell'Industria 4.0 e sfrutta diversi strumenti per eseguire l'analisi esplorativa e successiva caratterizzazione dei dati, necessaria per evitare incongruenze nei parametri raccolti e per consentire un tracciamento affidabile dei parametri di stato dei veicoli analizzati, fondamentale per l'allenamento di un modello di predictive maintenance.
Lo studio è stato applicato ad un caso reale: SEA Soluzioni Eco Ambientali s.r.l.
La SEA, con l'obiettivo di riduzione strutturale dei costi operativi e gestionali, ha fornito il proprio data lake per eseguire l'analisi dei dati raccolti dai sensori a bordo dei veicoli che eseguono la raccolta differenziata e le fatture degli interventi di manutenzione di quest'ultimi. L'esplorazione dei dati proposta considera sia la frequenza dei campioni che i valori dei parametri monitorati. Inoltre, è stato eseguito un confronto dei dati prelevati da sensori riforniti da differenti fornitori, per confrontare i valori raccolti da uno stesso veicolo appartenente ad una determinata flotta. In parallelo, è stata eseguita la categorizzazione degli interventi di manutenzione dei veicoli in base al tipo di guasto, in modo tale da poter comprendere quali sono i componenti dei veicoli maggiormente soggetti ad un guasto e con quale frequen-

za temporale avviene un intervento di manutenzione. Analisi promettente per lo sviluppo di un futuro modello di predictive maintenance.

Abstract - English

In the current era, Industry 4.0 has brought the issue of efficiency to the forefront and changed and multiplied the maintenance needs of connected vehicles. With condition monitoring capabilities, the ability to collect and analyze Big Data taken from on-board sensors, and the support of technological tools, one is able to learn and provide new information by processing the data that industrial IoT provides.
This thesis work traces the fundamental aspects of Industry 4.0 and leverages different tools to perform exploratory analysis and subsequent characterization of the data, which is necessary to avoid inconsistencies in the collected parameters and to enable reliable tracking of the state parameters of the analyzed vehicles, which is essential for training a predictive maintenance model.
The study was applied to a real case: SEA Soluzioni Eco Ambientali s.r.l. SEA, with the goal of structural reduction of operational and management costs, provided its data lake to perform analysis of data collected from on-board sensors of vehicles performing waste collection and maintenance invoices of the latter. The proposed data exploration considers both the frequency of samples and the values of monitored parameters. In addition, a comparison of data taken from sensors supplied by different suppliers was performed to compare the values collected from the same vehicle belonging to a given fleet. In parallel, categorization of vehicle maintenance work by type of failure was performed, so that it was possible to understand which vehicle components are most prone to failure and how often maintenance work occurs over time. Promising analysis for the development of a future predictive maintenance model.

Indice

I Introduzione **8**

I.1 Industria 4.0 9

I.1.1 Industria 4.0: storia ed evoluzione tecnologica 10

I.1.2 Le caratteristiche della Quarta Rivoluzione Industriale 12

I.1.3 I dati dell'industria 4.0 in Italia 12

I.2 Le tecnologie abilitanti dell'industria 4.0 13

I.2.1 Autonomous Robots 14

I.2.2 Additive Manufacturing 15

I.2.3 Augmented Reality 15

I.2.4 Simulation 16

I.2.5 Integration System 16

I.2.6 Internet of Things 17

I.2.7 Cloud Computing 19

I.2.8 Cybersecurity 20

I.2.9 Big Data 20

I.3 Impatto dell'introduzione delle tecnologie abilitanti dell'industria 4.0 22

I.3.1 Benefici delle tecnologie abilitanti 22

I.3.2 Criticità delle tecnologie abilitanti 23

I.4 Organizzazione della tesi 23

II Strumenti e Tecnologie **25**

II.1 Introduzione 25

II.2 Il Maching Learning 25
II.2.1 Storia del Machine Learning 26
II.2.2 I diversi apprendimenti di una macchina 27
II.2.3 Dall'apprendimento automatico al Data Mining 28
II.3 Il Data Mining 28
II.3.1 Knowledge Discovery in Database e Data mining 29
II.3.2 Metodi di Data Mining 31
II.3.3 Principali campi di applicazione 34
II.3.4 Strumenti per il data mining 35
II.3.5 Algoritmi di Data mining per la Manutenzione Predittiva . . . 39
II.4 La Manutenzione Predittiva 40
II.4.1 Strategie di manutenzione dei veicoli 40
II.4.2 Approcci PdM nel campo automobilistico 42

III Caso di studio: SEA - Soluzioni Eco Ambientali s.r.l. 45
III.1 Introduzione 45
III.2 Presentazione di SEA 46
III.3 Descrizione del caso di studio 48

IV Schema del Progetto 52
IV.1 Introduzione 52
IV.2 Strumenti di supporto 53
IV.3 Fornitori di dati 54
IV.4 Fatture di Manutenzione 55
IV.5 Feature e Ipotesi di Correlazione 57
IV.5.1 Caratteristiche del veicolo 58
IV.5.2 Caratteristiche contestuali del veicolo 59
IV.6 Modelli proposti per la Manutenzione Predittiva 60

V Analisi e Risultati Sperimentali 64
V.1 Introduzione 64

V.2 Analisi dei Fornitori 64
V.2.1 Analisi esplorativa 64
V.2.2 Confronto dei dati multi-provider 65
V.2.3 Analisi comparativa sulle caratteristiche monotoniche 67
V.2.4 Analisi nel tempo 69
V.3 Analisi delle Fatture di Manutenzione 73
V.4 Analisi delle Ipotesi di Correlazione 76
V.4.1 Algoritmo di Clustering 76
V.4.2 Analisi dei Turni di Servizio 78
V.4.3 Analisi dei Dipendenti 83
V.5 Analisi Predittiva 86
V.5.1 LSTM 86
V.5.2 XGBoost 88

Conclusioni e Sviluppi Futuri **91**

Bibliografia **94**

Elenco delle figure **105**

Elenco delle tabelle **108**

Capitolo I

Introduzione

Nell'epoca attuale, l'industria 4.0 ha portato il tema dell'efficientamento in primo piano e ha modificato e moltiplicato in aziende operanti in diversi settori, le esigenze manutentive.
Sono differenti gli approcci sviluppati in questo campo poiché sono diverse le tipologie di manutenzione che variano a seconda dei bisogni, dalla necessità improvvisa di una riparazione, al prefiggere un intervento programmato fino all'approccio migliorativo che permette di prevedere interventi non ricorsivi di miglioramento degli impianti, delle macchine e degli asset.
La gestione degli interventi di manutenzione non è semplice: gli interventi prefissati vanno calendarizzati, bisogna fronteggiare i guasti in maniera tempestiva e tenere traccia sia delle riparazioni che delle sostituzioni, monitoraggio e controllo sono eseguiti nelle scadenze previste in ambito sicurezza e vengono gestite le procedure inventariali.
Gli approcci tradizionali erano basati su documentazione cartacea caratterizzata dall'impiego di tempo, risorse e margini di errore troppo elevati per essere in linea con le esigenze della nuova era dell'industria.
Ragione per cui, sono necessarie nuove soluzioni software che gestiscano e coordinino le informazioni giuste al momento giusto riguardanti non soltanto gli interventi di manutenzione ma anche persone, scadenze e supporto decisionale.

Grazie alla capacità di monitoraggio delle condizioni, alla possibilità di raccogliere ed analizzare i big data e allo sviluppo di strumenti tecnologici, si è in grado di apprendere e fornire nuove informazioni elaborando i dati che l'IoT industriale fornisce, dando origine alla Manutenzione Predittiva.

In questo lavoro di tesi, dopo aver ripercorso gli aspetti fondamentali dell'Industria 4.0 e dell'IoT, le moderne tecnologie di Data Mining, ci si concentra su una particolare categoria di algoritmi di apprendimento automatico supervisionato, cioè gli algoritmi di Classificazione che, nel caso specifico, serviranno per realizzare in futuro un accurato modello di Manutenzione Predittiva.
Lo studio è applicato ad un caso reale: SEA Soluzioni Eco Ambientali s.r.l. La raccolta dati, proviene da sensori a bordo di una flotta di veicoli che eseguono la raccolta differenziata, sulla quale è stata eseguita l'esplorazione e successiva caratterizzazione, necessaria per evitare incogruenze nei parametri raccolti e per consentire una modellazione valida e guidata dai dati.
Contestualmente, sono state analizzate le fatture di manutenzione dei veicoli, seguita dalla loro categorizzazione in base al tipo di guasto di ciascun componente del veicolo.
Sono differenti i risultati sperimentali emersi dalla ricerca quali, l'identificazione dei comportamenti anomali nelle misurazioni, difformità legate alla raccolta dei dati. Queste problematiche sono state comunicate all'azienda al fine di portare grandi vantaggi: miglioramento della gestione dei dati e in particolare, la raccolta dati dai sensori della flotta e la riduzione strutturale dei costi operativi e gestionali.

I.1 Industria 4.0

Le origini della "Quarta rivoluzione industriale" risalgono al 2011 alla Fiera di Hannover, in Germania, come ipotesi di progetto.
Un gruppo di lavoro nel 2012 ha presentato al governo federale tedesco una serie di raccomandazioni per l'implementazione del Piano Industria 4.0. L'8 aprile 2013, sempre alla Fiera di Hannover, è stato diffuso il report finale con una previsione degli

investimenti necessari su infrastrutture, scuole, sistemi energetici, enti di ricerca e aziende per ammodernare il sistema produttivo tedesco e riportare la manifattura tedesca ai vertici mondiali rendendola competitiva a livello globale. Il modello è stato fonte di ispirazione per tutti gli altri Paesi [1].

L'ingegnere ed economista tedesco, Klaus Schwab, che fondò il World Economic Forum, nonchè autore del libro "The Fourth Industrial Revolution", ha coniato questo termine per descrivere la rivoluzione industriale definendola "...caratterizzata da una fusione di tecnologie che sta annullando i confini tra il fisico, il digitale e il biologico"[2].

Ma qual è la storia della rivoluzione industriale ? Di seguito, una breve panoramica storica per comprendere i progressi ottenuti negli anni, nel mondo degli affari e dell'industria.

I.1.1 Industria 4.0: storia ed evoluzione tecnologica

Nella figura I.1, si riassumono le quattro tappe evolutive della storia industriale:

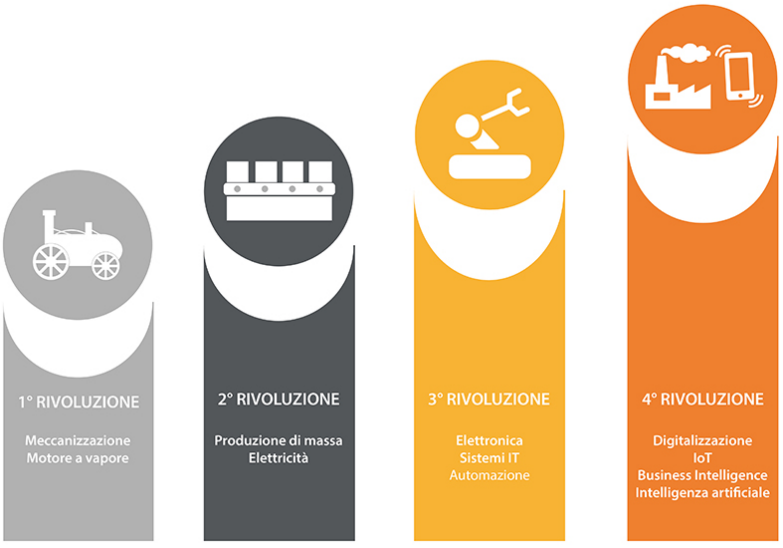

Fig. I.1: Tappe evolutive della storia industriale [3]

1784 - Industria 1.0

Rivoluziona il mondo della manifattura rispetto l'uso dell'energia: si parla dell'invenzione della macchina a vapore che consente alle fabbriche di abbandonare mulini per avviare una meccanizzazione della produzione all'insegna di una maggiore velocità e potenza.

1870 - Industria 2.0

È la seconda generazione energetica, che sfrutta prima l'elettricità e in seguito il petrolio, entrambi fondamentali per incrementare ulteriormente i livelli di meccanizzazione e di produzione. Grazie ad essi, la manifattura afferma gradualmente la catena di montaggio dando inizio ad una nuova era, cioè la produzione di massa.

1970 - Industria 3.0

Con la Terza rivoluzione industriale, l'ICT entra nel mondo delle fabbriche; infatti, discipline scientifiche come informatica ed elettronica automatizzano sia la produzione sia l'organizzazione. Quest'ultima permise la fondazione di nuovi processi che diversificarono le infrastrutture ed il lavoro dei dipendenti, portando grandi benefici alla qualità delle produzione.

2011 (e oltre) - Industry 4.0

In questa era, differenti tecnologie come robotica, sensoristica, connessione e programmazione, rivoluzionano la fabbricazione dei prodotti e l'organizzazione del lavoro.
In relazione all'uso esteso di Internet, come una piattaforma globale che permetta a oggetti di tutti i giorni di coordinarsi e comunicare tra di loro, nasce Internet of Things (IoT). L'IoT permette di estendere nuovi modelli produttivi sempre più interconnessi e automatizzati, migliorando tracciabilità e rintracciabilità di processi e sistemi e, conseguentemente, la gestione collettiva delle informazioni.
Infatti, si parla di Software di nuova generazione da un lato e Big Data Manage-

ment dall'altro, arrivando così ad una personalizzazione di massa. L'industria 4.0 è una continua evoluzione tecnologica che spazia differenti ambiti operativi associati all'Intelligenza Artificiale e tutte le derive del digitale, includendo la Blockchain.

I.1.2 Le caratteristiche della Quarta Rivoluzione Industriale

Schwab, nel suo articolo per il World Economic Forum [4], definisce fondamentali i seguenti punti: **la velocità, la portata e l'impatto dei sistemi**.
"*La velocità delle scoperte attuali non ha precedenti storici. Rispetto alle precedenti rivoluzioni industriali, la Quarta si sta evolvendo a un ritmo esponenziale piuttosto lineare. Inoltre, sta sconvolgendo quasi tutti i settori in ogni paese. E l'ampiezza e la profondità di questi cambiamenti annunciano la trasformazione di interi sistemi di produzione, gestione e governance.*"

I.1.3 I dati dell'industria 4.0 in Italia

Sono diverse le analisi eseguite dal 2019 al 2020 sugli investimenti delle imprese italiane nell'industria 4.0 e le sue tecnologie abilitanti.
I risultati emersi dagli studi sono positivi poiché crescenti, nonostante la crisi del 2020 dovuta alla pandemia causata dal Covid-19.

Industria 4.0 nel 2020 secondo Politecnico di Milano (Polimi)

L'Osservatorio Industria 4.0 del Polimi, riporta in un report i dati dell'Industria 4.0 in Italia [5], evidenziando come il volume d'affari del 2020, ebbe una crescita dell'8% rispetto il 2019 per un ammontare pari a 4,1 miliardi di euro. Risultato inferiore al 20%, ovvero sotto la previsione stimata nel 2019 ma pur sempre positivo. Il grafico I.2 evidenzia la crescita avvenuta nel 2020 dell'impiego delle applicazioni di industria 4.0 rispetto al 2019.

Per un totale di circa 1400 appplicazioni, si aggiudicano il primato le soluzioni di Industrial IoT, pari a un quarto del totale di 380 applicazioni (31% in più rispetto al 2019), spesso combinate con algoritmi di Analytics e Intelligenza Artificiale.

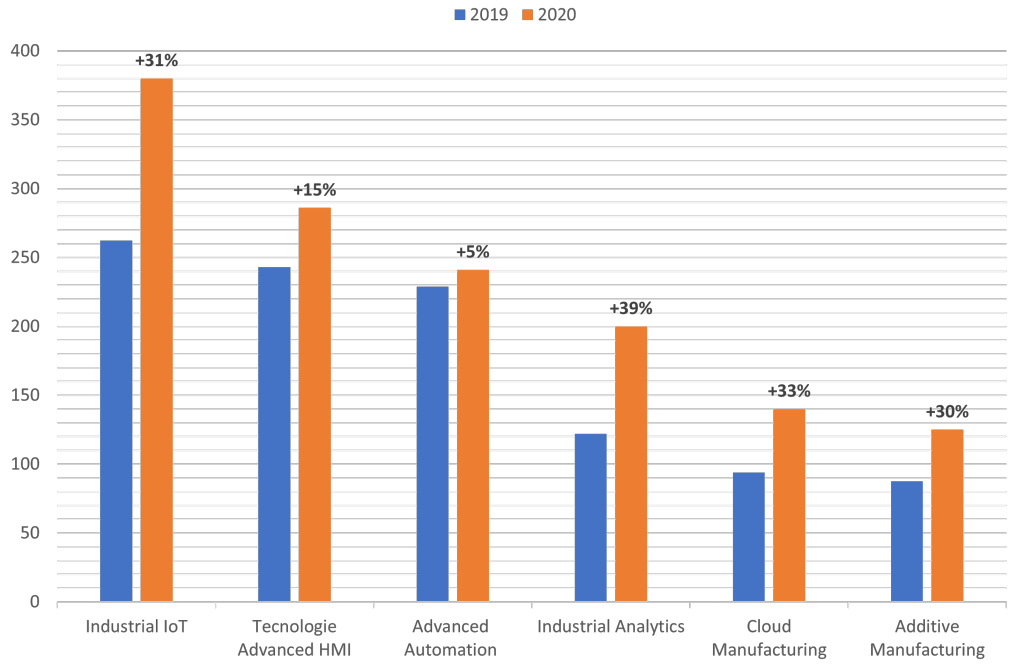

Fig. I.2: Confronto del numero di applicazioni impiegate nel 2019 e nel 2020

Mentre, risulta essere inferiore, rispetto la precedente, l'impiego di tecnologie come Additive Manufacturing (circa 125), nota anche come Stampa 3D e diffusa principalmente nei settori automotive e aerospaziale" [1].

Industria 4.0 nel 2020 secondo Deloitte

Anche aziende di servizi di consulenza e revisione come la Deloitte, hanno confermato che la tendenza positiva non si è arrestata nemmeno durante la pandemia: *"...durante la pandemia, malgrado la battuta d'arresto che ha investito tutta l'economia, il 63% delle aziende manifatturiere italiane, prese in esame, non solo non ha interrotto i flussi di capitali a supporto dei progetti di produzione intelligente, ma li sta accelerando dedicando in media il più di un terzo del proprio "factory budget"* [6].

I.2 Le tecnologie abilitanti dell'industria 4.0

Le tecnologie abilitanti dell'industria 4.0 (Figura I.3) rientrano tra le tecnologie "ad alta intensità di conoscenza e associate a elevata intensità di R&S, a cicli di

innovazione rapidi, a consistenti spese di investimento e a posti di lavoro altamente qualificati" [7].

Queste tecnologie sono fondamentali per l'industria 4.0 poiché aumentano l'efficienza della catena del sistema produttivo e la capacità di innovare processi, prodotti e servizi di differenti settori economici. Si parla dell'uso dell'Intelligenza Artificiale applicata agli oggetti, che crea una stretta connessione e collaborazione tra la realtà fisica e virtuale.

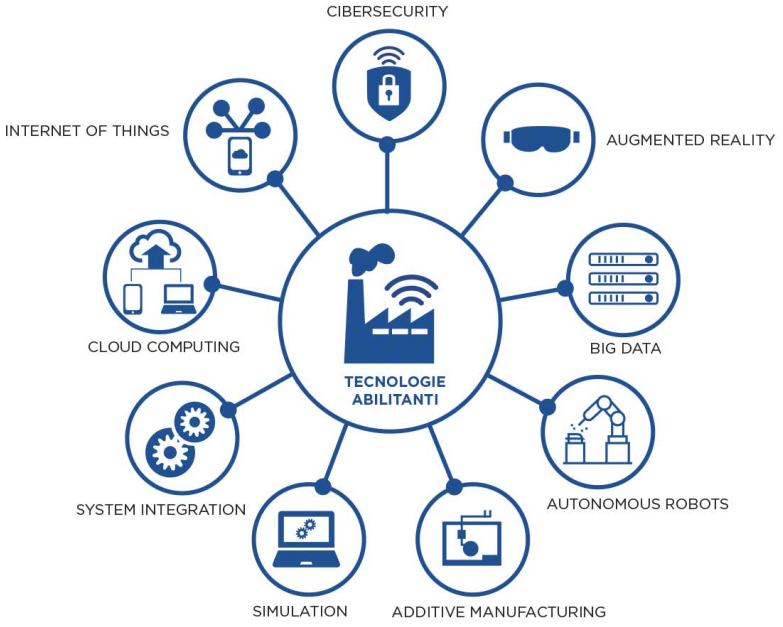

Fig. I.3: Le Tecnologie Abilitanti dell'Industria 4.0 [8]

I.2.1 Autonomous Robots

La componente principale è la "Robotica Collaborativa" ovvero la condivisione di uno spazio di lavoro tra uomo e robot in modo efficiente e sicuro grazie all'intelligenza artificiale delle macchine che sono in grado di percepire lo spazio circostante [9].

L'utilizzo della robotica nel processo di produzione risulta essere sempre più effi-

ciente ed in continua evoluzione. Infatti, i nuovi robot saranno più autosufficienti, autonomi ed interattivi con la forza lavoro umana, sostituendo task semplici e ripetitivi (standardizzabili) e comportando la nascita di nuove attività basate sulle soft skill, ovvero quelle "conoscenze" che garantiranno la capacità del problem solving.

I.2.2 Additive Manufacturing

Con Manifattura Additiva (o stampanti 3D) si intende la realizzazione di oggetti fisici tridimensionali che nascono da un modello digitale che sarebbe il loro design. Tradizionalmente, la manifattura si occupava della creazione di un determinato prodotto attraverso l'estrazione di materiale da un componente grezzo. Con l'avvento della manifattura additiva, la creazione di un componente avviene grazie il deposito stratificato di materiale.
Grazie a questa nuova tecnologia è possibile realizzare prodotti complessi con particolari caratteristiche geometriche, utilizzando un minor quantitativo di materiale, in tempi ridotti e costi relativi legati alle varianti più bassi.

I.2.3 Augmented Reality

La "Realtà Aumentata" è una tecnologia elevata che permette, tramite l'uso di particolari strumenti digitali, di interagire con l'ambiente esterno.
La realtà aumentata sfrutta elementi presenti nell'ambiente e sfrutta l'interazione con l'ambiente circostante per ottenere una conoscenza più approfondita e specifica di determinati elementi [10].
La realtà aumentata permette di incrementare e migliorare azioni svolte quotidianamente poiché determina un arrichimento delle percezioni attraverso elaborazioni digitali svolte dai software.

Infatti, grazie a questa nuova tecnologia si potrà:

- migliorare la conoscenza tramite una visualizzazione più accurata e dettagliata dell'ambiente

- incrementare le percezioni dei suoni
- comandare a distanza strumenti tecnologici come smartphone, tablet o altro.

I.2.4 Simulation

Le simulazioni sono usate in processi di produzione al fine di analizzare grandi quantità di dati reali in tempo reale in un modello virtuale controllato, dato dall'unione di impianti, prodotti e personale operativo. Attraverso l'attività di test si ha la possibilità di ottimizzare processi ancor prima della loro creazione comportando sia una riduzione dei tempi di installazione sia una migliore qualità del prodotto. Conseguentemente, si hanno grandi vantaggi nell'affrontare i rilevanti costi causati dal learning-by-doing.

I.2.5 Integration System

Con "Integration System" si intende l'integrazione di sistemi discreti utilizzando una varietà di tecniche come reti di computer, integrazione di applicazioni aziendali, gestione dei processi aziendali o programmazione manuale [11].
Si tratta di tecnologie che possono essere interconnesse sia verticalmente che orizzontalmente permettendo la creazione di sistemi per la condivisione di dati in real time.
L'integrazione verticale favorisce i rapporti tra azienda e tutti i membri della value chain (ad esempio fornitori, clienti) poiché determinata la condivisione di lavori standard.
Per quanto riguarda l'integrazione orizzontale, permette una migliore gestione delle informazioni che sono interne all'azienda (ad esempio la definizione del ciclo di vita di un prodotto).

I.2.6 Internet of Things

Il neologismo Internet of Things è stato utilizzato per la prima volta dal ricercatore britannico Kevin Ashton, che dopo essere stato leader nell'azienda Procter & Gamble, nel 1999 era al MIT (Massachussets Institute of Technology) e coniò questo termine per descrivere un sistema dove Internet è connesso al mondo fisico con una rete di sensori distribuiti [12].
Se dobbiamo, tuttavia, ripercorrere la storia dell'Internet of Things dagli inizi dobbiamo tornare ben più indietro, a una profezia scientifica fornita da Nikola Tesla, che nel 1926 si pronunciava così, a proposito dello sviluppo di una tecnologia wireless: "Quando la connessione senza fili sarà perfettamente applicata - sosteneva l'ingegnere e inventore serbo - tutto il pianeta sarà convertito in un enorme cervello, in cui tutte le parti rappresenteranno elementi di un insieme complessivo costantemente coordinato. Saremo in grado di comunicare con gli altri istantaneamente, senza preoccuparci della distanza fisica. Saremo in grado di vederci e sentirci perfettamente attraverso apparecchi incredibilmente più semplici degli attuali telefoni, che staranno dentro la tasca dei nostri cappotti" [13].
Oggi una definizione di Internet of Things potrebbe essere la seguente:
"è una specie di Internet degli strumenti quotidiani, un mega network in cui tutti gli oggetti che ci circondano sono connessi alla rete e interconnessi tra loro. In questo modo, raccolgono e si scambiano dati in tempo reale, dialogano tra loro e comunicano con il mondo esterno e ci forniscono informazioni alle quali prima non avevano e non avevamo accesso. Apparecchi e dispositivi con queste caratteristiche vengono definiti smart objects, oggetti smart, oggetti "intelligenti". [14].
L'Internet of Things trova numerevoli campi di applicazione, poiché quest'ultimo devolve grandi potenzialità ad oggetti usati ogni giorno. Le discipline in cui Internet of Things ha avuto maggior successo sono le seguenti:

- Industria automobilistica
- Avionica

- Industria manifatturiera
- Domotica
- Industria Biomedicale
- Telemetria

Particolare campo dell'Internet of Things: Le Smart Cars

Il campo delle Smart Car risulta essere di maggiore interesse in questo lavoro di tesi poiché le analisi coinvolgono il settore automobilistico. Sono diversi i sistemi disponibili ed esistono anche differenti tipi di connessione:

- **V2I "Vehicle to Infrastructure":** la tecnologia acquisisce i dati generati dal veicolo e fornisce informazioni sull'infrastruttura al conducente. La tecnologia V2I comunica informazioni su sicurezza, mobilità o condizioni ambientali.
- **V2V "Vehicle to Vehicle":** la tecnologia comunica informazioni su velocità e posizione dei veicoli circostanti attraverso uno scambio di informazioni wireless. L'obiettivo è evitare incidenti, alleggerire la congestione del traffico e avere un impatto positivo sull'ambiente.
- **V2C "Vehicle to Cloud":** la tecnologia scambia informazioni su e per le applicazioni del veicolo con un sistema cloud. In questo scenario, i dati sono raccolti localmente sul veicolo e poi sono inviati al cloud dal quale verranno estratti e analizzati. Nel nostro caso, i dati sono raccolti da sensori a bordo della flotta dei veicoli.
- **V2P "Vehicle to Pedestrian":** la tecnologia rileva le informazioni sul proprio ambiente e le comunica ad altri veicoli, infrastrutture e dispositivi mobili personali. Ciò consente al veicolo di comunicare con i pedoni e ha lo scopo di migliorare la sicurezza e la mobilità su strada.

- **V2X "Vehicle to Everythings":** la tecnologia interconnette tutti i tipi di veicoli e sistemi infrastrutturali con un altro. Questa connettività comprende auto, autostrade, navi, treni e aerei.

Applicazioni dell'Internet of Things

Le applicazioni dell'Internet of Things [15] sono differenti ma di seguito elenchiamo quelle di maggior successo.

- **Maggiore efficienza e produttività**: i macchinari delle linee di montaggio posseggono diversi sensori che consentono sia una estrema precisione nel calcolo della produttività di quello specifico macchinario sia l'individuazione di eventuali prodotti difettosi.
- **Manutenzione predittiva**: cambia il modo di intervenire sulla manutenzione grazie ai sensori e alla capacità di analisi tipica della IoT. Basata non più su tempi fissi o sulla risposta ai guasti ma sulla predizione e monitoraggio.
- **Gestione "data driven" delle scorte**: lo sviluppo di sistemi di riassortimento, grazie ai dispotivi smart, permettono la gestione delle scorte con meno sprechi e meno occupazione dello spazio.
- **Distribuzione ed intra-logistica**: l'integrazione tra dispositivi e umani o dispositivi e sistemi di movimentazione, ha portato benefici sia nel controllo di qualità e sicurezza del prodotto sia all'organizzazione di spedizioni, grazie alle informazioni ottenute in real time.

I.2.7 Cloud Computing

Il cloud computing indica, in informatica, un paradigma di erogazione di servizi offerti su richiesta da un fornitore a un cliente finale attraverso la rete internet (come l'archiviazione, l'elaborazione o la trasmissione dati), a partire da un insieme di risorse preesistenti, configurabili e disponibili in remoto sotto forma di architettura distribuita [16].

Nell'ambito di processi produttivi ed organizzativi, vista la crescita della domanda di personalizzazione del prodotto dei consumatori finali nell'Industria 4.0, il Cloud è un potente strumento che permetta la collaborazione tra più imprese che, una volta portate a termine il lavoro prefissato, consente lo sciogliemento delle organizzazioni temporanee di impresa (virtuali) per dar spazio a nuove opportunità di business.

I.2.8 Cybersecurity

Cybersecurity è un termine che più precisamente rappresenta una sottoclasse [17] della sicurezza informatica: con esso si enfatizzano spesso qualità di resilienza, robustezza e reattività che una tecnologia deve possedere per fronteggiare attacchi mirati a comprometterne il suo corretto funzionamento e le sue performance (attacchi cyber).
Qualsiasi dispositivo connesso ad internet potrebbe essere violato; pertanto, è importante per un'azienda la sicurezza delle informazioni fondamentali per il business.

I.2.9 Big Data

Con il termie Big Data si indica, genericamente, una raccolta di dati informativi così estesa in termini di volume, velocità e varietà da richiedere tecnologie e metodi analitici specifici per l'estrazione di valore o conoscenza [18].
Le tre V furono formulate per la prima volta dallo studioso Doug Laney nei primi anni del 2000 [19]:

- **Varietà**: i dati arrivano in modo disomogeneo, trattandosi di di diversi tipi di dati (foto, documenti, valori alfanumerici, video, audio e quant'altro);
- **Volume**: la grande quantità di dati provenienti da sorgenti differenti (social media, transazioni finanziarie, acquisti online);
- **Velocità**: riferita alla velocità con cui i dati viaggiano in real time ed alla conseguente necessità di utilizzarli in modo tempestivo.

I Big data sono un enorme mole di dati eterogenei, strutturati e non strutturati messi in relazione grazie a sofisticati metodi statistici e informatici di elaborazione, al fine di scoprire i legami tra fenomeni diversi (ad esempio correlazioni) e prevedere quelli futuri. I Big Data possono essere utilizzati per diversi scopi tra cui quello di misurare le prestazioni di un'organizzazione nonchè di un processo aziendale [20].

La gestione dei Big Data

L'utilizzo di piattaforme di differenti tipologie o di un qualsiasi dispositivo genera dati.
La gestione della grande quantità di informazioni si compone di differenti step che vanno dalla raccolta e memorizzazione alla loro successiva analisi [21].

- **Acquisizione**: avviene attraverso vari canali, ovvero tramite API (interfaccia di programmazione di un'applicazione) utilizzate appositamente per raccogliere dati quando si accede ad un sito, con software appositi per la raccolta di documenti, importando dati da database preesistenti oppure interpretando e estrapolando il flusso di dati che passano attraverso la rete.

- **Pulizia dei dati**: ne consegue che essi debbano essere bonificati ovvero ripuliti da tutte quelli che non rientrano nel formato richiesto per l'elaborazione.

- **Stoccaggio e archiviazione**: i Big Data devono essere stoccati ed archiviati ma qui entra in gioco il problema relativo all'enorme quantità di informazioni e per questo motivo, nel corso degli ultimi anni, sono stati studiati e realizzati sistemi in grado di immagazzinare dataset di grandi dimensioni.

- **Analisi e Modellazione**: avviene tramite lo sviluppo di algoritmi mirati ed alla successiva interpretazione affinchè le informazioni possano risultare utili ai fini delle performance aziendali.

I.3 Impatto dell'introduzione delle tecnologie abilitanti dell'industria 4.0

Sono diverse le aziende che hanno adottato le tecnologie abilitanti della Quarta Rivoluzione Industriale, poiché hanno portato grandi vantaggi all'interno dei propri asset registrando una crescita dell'efficienza produttiva. Tuttavia, conseguono importanti cambianti interni ad esse che risultano, spesso, difficili da fronteggiare.

I.3.1 Benefici delle tecnologie abilitanti

I benefici che le tecnologie abilitanti hanno portato alle aziende sono vari ma di seguito si elencano quelli più rilevanti.

- **Miglioramento della Customer Experience ed aumento fidelizzazione clienti.** Produzione basata su determinate specifiche su richiesta del cliente (on demand). Ciò è realizzabile grazie ad una produzione semplice, flessibile e digitalizzata. Garantire al cliente la diversità del prodotto secondo la logica "end to end integration" (ad esempio, la manutenzione predittiva su prodotti specifici dell'azienda)

- **Riduzione dei costi.** La tecnologia permette la semplificazione e la ridimensione del processo produttivo favorendo una produzione decentralizzata di alcune componenti (ad esempio, le stampanti 3D riducono i tempi, i costi di assistenza al cliente e riduce al minimo i pezzi di ricambio a magazzino).

- **Maggiore flessibilità organizzativa**. La risposta al mercato da parte dell'azienda risuta essere di gran lunga agevolata dai sistemi di integrazione poiché favoriscono una migliore comunicazione e collaborazione tra le diverse funzioni della catena di produzione, ciò comporta innovazione di prodotto e di servizio.

- **Nuovi modelli di business.** La tecnologia introduce nuovi business model che le aziende potranno sfruttare a loro vantaggio per la creazione del valore dell'impresa.

- **Miglioramento delle condizioni lavorative dei dipendenti.** Ci si riferisce a sicurezza, ergonomia e sforzo fisico. Sono diverse le imprese manifatturiere che aprono le strade allo smart manifacturing.

I.3.2 Criticità delle tecnologie abilitanti

Le problematiche affrontate dalle imprese sono legate soprattutto al cambiamento tecnologico; infatti, si parla di:

- **Velocità di cambiamento.** La curva di velocità di cambiamento si trova di gran lunga al di sopra della curva di apprendimento delle imprese che serve a perseguire l'evoluzione tecnologica.
- **Adattamento al cambiamento.** L'adozione di metodi e strumenti tecnologici dell'industria 4.0 richiede flessibilità dell'operatività, ragionamenti con mentalità più aperta, veloce, e globalizzata.
- **Acquisizione di nuove skill.** L'impiego di tecnologie abilitanti è sinonimo di ricerca di nuove competenze per raggiungere alti livelli di stimolazione e partecipazione dei dipendenti che si occupano dei processi produttivi.
- **Cambiamento della value chain.** Vuol dire elimiare tutte quelle attività aziendali che non sono di valore aggiunto all'impresa e ciò è possibile con la semplificazione di determinati processi.

Si deduce dai benefici e dalle problematiche sopra elencate che l'innovazione di un processo produttivo comporta differenti vantaggi ma è fondamentale la capacità delle imprese di allineare la loro organizzazione interna al cambiamento tecnologico che richiede l'industria 4.0.

I.4 Organizzazione della tesi

In questo capitolo, hanno avuto un ruolo centrale l'Industria 4.0 e le sue Tecnologie abilitanti, in particolare l'Internet of Things e Big Data. Il resto della tesi è

organizzata come segue:

Nel capitolo II, verrà presentato il Machine Learning branca dell'intelligenza artificiale che raccoglie metodi sviluppati in varie comunità scientifiche, sotto diversi nomi quali: statistica computazionale, riconoscimento di pattern, reti neurali artificiali, filtraggio adattivo, teoria dei sistemi dinamici, data mining, ecc...

Successivamente, particolare attenzione sarà rivolta al Data Mining che è l'insieme di tecniche e metodologie che hanno per oggetto l'estrazione di informazioni utili da grandi quantitdi dati e può essere applicato a differenti campi, quello di interesse in questo lavoro di tesi è la Manutenzione Predittiva.

Il Capito III propone una presentazione e descrizione del caso di studio, cioè dell'azienda italiana SEA Soluzioni Eco Ambientali s.r.l. evidenziando obiettivi ed eventuali punti di forza e debolezza dell'impresa.

Nel Capitolo IV, ci si addentrerà nel lavoro di ricerca vero e proprio poiché verrà illustrato lo schema di progetto, verranno presentati i vari fornitori di dati, l'organizzazione, digitalizzazione e categorizzazione delle fatture di manutenzione dei veicoli nonchè la gestione dei turni di servizi dei dipendenti e gli algoritmi di classificazione proposti.

Capitolo V mostrerà i risultati sperimentali emersi dalle analisi dei dati illustrati nel capitolo precedente (Capitolo IV).

L'ultimo capitolo, ovvero il capitolo conclusivo, conterrà una riflessione sugli eventuali vantaggi e svantaggi emersi dalle analisi dei dati, suggerimenti e accorgimenti sulla gestione dei dati dell'azienda, sviluppi futuri della ricerca e considerazioni personali.

Capitolo II

Strumenti e Tecnologie

II.1 Introduzione

Nel capitolo precedente (Capitolo I), sono stati esaminati il concetto di Industria 4.0 e il grande impatto che ha avuto oggigiorno, le tecnologie abilitanti di questa rivoluzione e in particolar modo sono state attenzionate IoT e Big Data.
Adesso, in questo capitolo, verrà introdotto il Machine Learning, branca dell'Intelligenza Artificiale, che si occupa dello studio e della costruzione di algoritmi utilizzando metodi di Data Mining come metodi di apprendimento non supervisionato. Questi algoritmi apprendono da un grande volume di dati con l'obiettivo di predire su questi, costruendo un modello basato su dei campioni. In questa sezione, il modello predittivo presentato sarà quello di Predictive maintenance (PdM).

II.2 Il Maching Learning

Il campo dell'apprendimento automatico ha ricevuto diverse definizioni formali in letteratura. Arthur Samuel, nel suo lavoro, ha definito l'apprendimento automatico come "un campo di studio che offre ai computer la capacità di apprendere senza essere programmati esplicitamente" [22]. Usando un lessico informatico, Tom Mitchell lo presentò come "Si dice che un programma per computer impari dall'esperienza (E) rispetto ad alcune classi di compiti (T) e misura delle prestazioni (P), se le prestazioni

in compiti T, come misurato da P, migliora con l'esperienza E" [23]. Ethem Alpaydin nel suo libro di testo ha definito l'apprendimento automatico come il campo della "Programmazione di computer per ottimizzare un criterio di prestazione utilizzando dati provenienti da esperienze passate" [24]. Queste varie definizioni condividono la nozione di "coaching computer" per eseguire in modo intelligente determinati compiti apprendendo l'ambiente circostante con ripetuti "esempi".

II.2.1 Storia del Machine Learning

L'apprendimento automatico è la tecnologia per lo sviluppo di algoritmi informatici in grado di emulare l'intelligenza umana. Attinge idee da diverse discipline come intelligenza artificiale, probabilità e statistica, informatica, teoria dell'informazione, psicologia, teoria del controllo e filosofia [23][24][25]. Questa tecnologia è stata applicata in campi così diversi come il riconoscimento di modelli [25], la visione artificiale [26], l'ingegneria dei veicoli spaziali [27], la finanza [28], l'intrattenimento [29][30], l'ecologia [31], la biologia computazionale [32][33] e applicazioni biomediche e mediche [34][35]. La proprietà più importante di questi algoritmi è la loro capacità distintiva di apprendere l'ambiente circostante dai dati di input [23][24]. Storicamente, l'inizio dell'apprendimento automatico potrebbe risalire al diciassettesimo secolo, allo sviluppo di macchine in grado di emulare la capacità umana di aggiungere e sottrarre da Pascal e Leibniz [36]. Nella storia moderna, Arthur Samuel di IBM ha coniato il termine "apprendimento automatico" e ha dimostrato che i computer possono essere programmati per imparare a giocare a dama [22]. Questo è stato seguito dallo sviluppo del perceptron di Rosenblatt come una delle prime architetture di rete neurale nel 1958 [37]. Mentre nel 1975 è stata raggiunta una svolta dallo sviluppo del perceptron multistrato (MLP) di Werbos [38]. A questo segue lo sviluppo di Alberi Decisionali di Quinlan nel 1986 [39] e di macchine vettoriali di supporto di Cortes e Vapnik [40]. Successivamente sono stati proposti algoritmi di apprendimento automatico che combinano più tipi di apprendimenti, tra cui Adaboost [41] e Random Forest [42]. Più recentemente, algoritmi di apprendimento distribuiti mul-

tistrato sono emersi sotto la nozione di deep learning [43]. Questi algoritmi sono in grado di apprendere buone rappresentazioni dei dati che facilitano l'estrazione di informazioni utili durante la creazione di classificatori o altri predittori [44].

II.2.2 I diversi apprendimenti di una macchina

L'apprendimento automatico può essere suddiviso in base alla natura dell'etichettatura dei dati. Come mostrato in II.1, vi sono tre tipi di apprendimento:

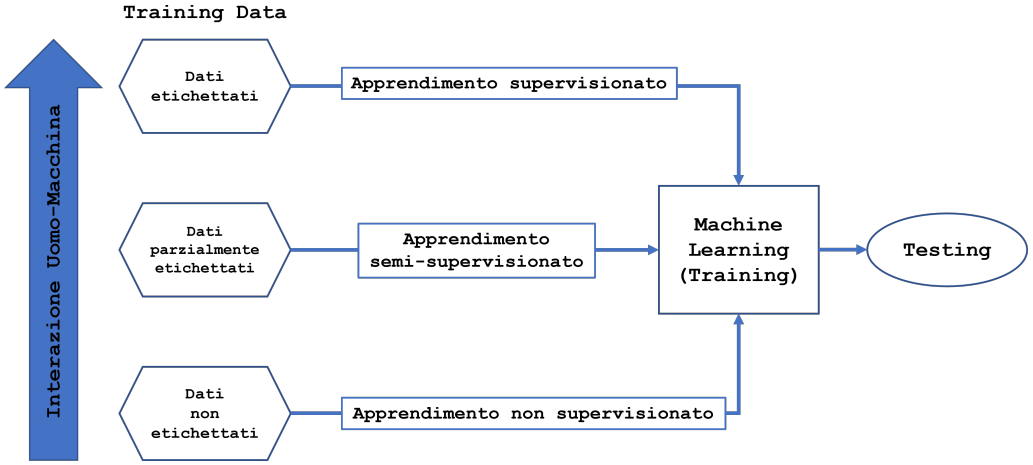

Fig. II.1: Categorie di algoritmi di apprendimento automatico in base alla natura dei dati di training

Apprendimento supervisionato

L'apprendimento supervisionato viene utilizzato per stimare una mappatura sconosciuta (input, output) da campioni noti (input, output), in cui l'output è etichettato (ad esempio, classificazione e regressione).

Apprendimento non supervisionato

Nell'apprendimento non supervisionato, al sistema di apprendimento vengono forniti solo campioni di input (ad esempio, raggruppamento e stima della funzione di densità di probabilità).

Apprendimento semi-supervisionato

L'apprendimento semi-supervisionato è una combinazione di supervisionato e non supervisionato in cui parte dei dati è parzialmente etichettata e la parte etichettata viene utilizzata per dedurre la parte non etichettata (ad esempio, sistemi di recupero di testo/immagine).

II.2.3 Dall'apprendimento automatico al Data Mining

La capacità di apprendere attraverso l'input dell'ambiente circostante è la chiave principale per sviluppare un'applicazione di apprendimento automatico di successo. In questo caso, l'apprendimento è definito come stima delle dipendenze dai dati [45]. I campi del data mining e dell'apprendimento automatico sono intrecciati. Il data mining utilizza algoritmi di machine learning per interrogare database di grandi dimensioni e scoprire proprietà prima sconosciute nei dati, mentre molti algoritmi di machine learning utilizzano metodi di data mining come metodi di apprendimento non supervisionato o come passi di preprocessing che aumentano l'accuratezza dell'apprendimento, prima di esaminare le attività desiderate.

II.3 Il Data Mining

Per Data Mining si intende l'individuazione di informazioni di varia natura (non risapute a priori) tramite estrapolazione mirata da grandi banche dati, singole o multiple (nel secondo caso, informazioni più accurate si ottengono incrociando i dati delle singole banche). Le tecniche e le strategie applicate alle operazioni di data mining sono per larga parte automatizzate, consistendo in specifici software e algoritmi adatti al singolo scopo. Ad oggi, in particolare, si utilizzano reti neurali, alberi decisionali, clustering e analisi delle associazioni. Le finalità del data mining sono applicabili ai più svariati campi: economico, scientifico, operativo, etc.

II.3.1 Knowledge Discovery in Database e Data mining

Il termine data mining è spesso usato come sinonimo del processo di estrazione di informazioni utili dai database. In [33], si esegue una distinzione tra quest'ultimo, chiamato KDD, e Data mining. Il termine data mining è stato utilizzato principalmente da statistici, analisti di dati e comunità di sistemi informativi di gestione (MIS). Ha anche guadagnato popolarità nel campo dei database. Ad esempio, già nel 1978 [39], il termine era usato per dimostrare come la regressione lineare generalizzata poteva essere utilizzata per risolvere problemi irrisolvibili sia dall'uomo che dalle tecniche statistiche tradizionali dell'epoca. Il termine KDD è stato coniato al primo workshop KDD nel 1989 [41] per sottolineare che "la conoscenza è il prodotto finale di una scoperta basata sui dati." Quindi, KDD si riferisce al processo generale di scoperta di conoscenze utili estratte dai dati mentre il data mining si riferisce a una fase particolare di questo processo, in cui sono applicati algoritmi specifici per estrarre modelli dai dati. I vari passaggi nel processo KDD sono approfondinti nel paragrafo II.3.1.

Il processo KDD

Il processo KDD è interattivo e iterativo e comprende numerosi passaggi (vedi Figura II.2) e molte decisioni prese dall'utente. Brachman e Anand(1996) danno una visione pratica del processo KDD sottolineando la natura interattiva del processo. Di seguito, si illustreranno a grandi linee alcuni dei suoi passaggi fondamentali:

1. **Sviluppare la comprensione del dominio dell'applicazione e delle relative conoscenze pregresse e identificare l'obiettivo del processo KDD dal punto di vista del cliente.**

2. **Creazione di un set di dati di destinazione**: selezionare un set di dati o concentrarsi su un sottoinsieme di variabili o campioni di dati, su cui eseguire la discovery.

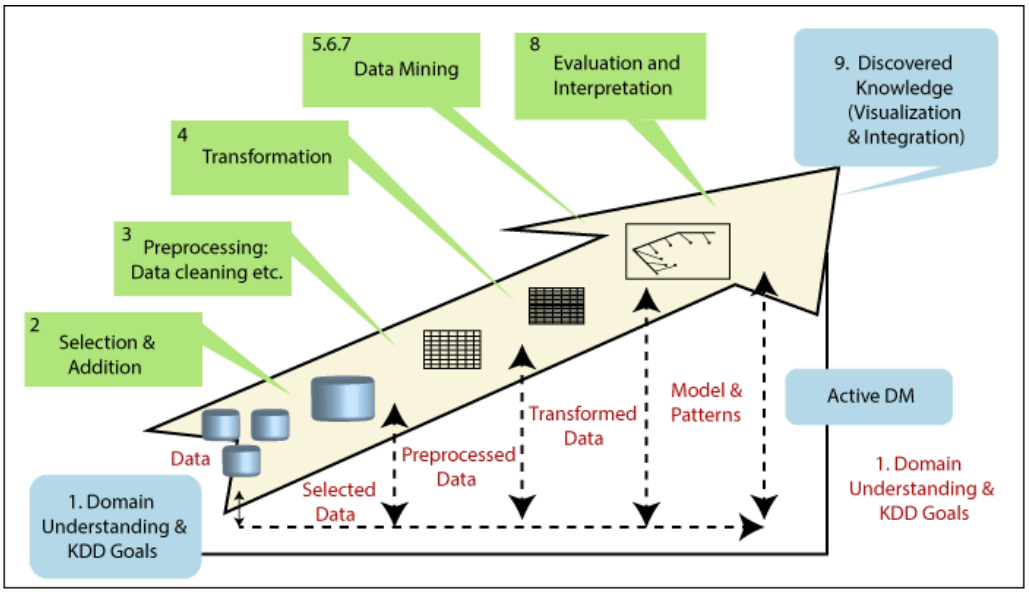

Fig. II.2: Processo KDD in data mining [46]

3. **Pulizia e preelaborazione dei dati**: operazioni di base come la rimozione o la considerazione del rumore, decidere le strategie per gestire i campi di dati mancanti, tenere conto delle informazioni sulla sequenza temporale e delle modifiche note.

4. **Riduzione e proiezione dei dati**: trovare funzionalità utili per rappresentare i dati in base all'obiettivo delle attività da svolgere. Utilizzo di metodi di riduzione o trasformazione della dimensionalità per ridurre il numero effettivo di variabili in esame o per trovare rappresentazioni invarianti per i dati.

5. **Abbinare gli obiettivi del processo KDD ad un particolare metodo di data mining**: sintesi, classificazione, regressione, clustering, ecc.

6. **Scelta degli algoritmi di data mining**: selezione dei metodi da utilizzare per la ricerca di modelli di dati. Ciò include la decisione di quali modelli e parametri possono essere appropriati (ad es. i modelli per i dati categoriali sono diversi dai modelli sui vettori rispetto ai reali) e la corrispondenza di un particolare metodo di data mining con i criteri generali del processo KDD (ad

esempio, l'utente finale può essere più interessato a comprendere il modello rispetto alle sue capacità predittive).

7. **Data mining**: ricerca di modelli di interesse in una particolare forma rappresentativa o in un insieme di tali rappresentazioni: regole di classificazione o alberi, regressione, clustering e così via.

8. **Interpretando i modelli estratti**, è possibile che si torni ad un qualsiasi passaggio descritto precedentemente e quindi che si facciano ulteriori iterazioni.

9. L'ultimo passaggio comporta la **visualizzazione dei modelli estratti**.

Il processo KDD implica un'iterazione significativa e può contenere loop tra due o più passaggi qualsiasi. Il flusso dei passaggi di base è illustrato nella Figura II.2. La maggior parte del lavoro basato su un processo KDD è concentrato sul passaggio 7, il data mining. Tuttavia, gli altri passaggi sono importanti per il successo nell'applicazione della pratica KDD. Dopo aver definito le nozioni di base e introdotto il processo KDD, ci concentriamo ora sulla componente di data mining, che ha ricevuto la maggiore attenzione in letteratura.

II.3.2 Metodi di Data Mining

Le tecniche di data mining possono essere suddivise in cinque classi di metodi elencate di seguito. Sebbene molte di queste tecniche siano state storicamente definite per funzionare su dati residenti in memoria (in genere letti da file flat), alcune di queste tecniche iniziano ad essere ridimensionate per funzionare sui database. Ad esempio nella classificazione si includono alberi decisionali [40] e in sintesi le regole di associazione.

Modellazione predittiva

L'obiettivo è prevedere alcuni campi in un database basato su altri campi. Se il campo previsto è una variabile numerica (continua) (come una misurazione fisica, ad esempio, l'altezza), il problema di previsione è un problema di regressione. Se il

campo è categoriale, allora è un problema di classificazione. Esiste un'ampia varietà di tecniche per la classificazione e la regressione. Il problema, in generale, è rappresentato dalla determinazione del valore più probabile della variabile da prevedere avendo anche altri dati in input, i dati di training (in cui viene fornita la variabile target per ogni osservazione) e un insieme di ipotesi che rappresentano le precedenti conoscenze del problema. La regressione lineare combinata con la trasformazione non lineare degli input potrebbe essere utilizzata per risolvere una vasta gamma di problematiche. Ma la trasformazione dello spazio di input non è semplice; infatti, richiede una buona conoscenza del problema. Nei problemi di classificazione questo tipo di trasformazione viene spesso definito "estrazione di caratteristiche". Nella classificazione l'obiettivo fondamentale è predire lo stato più probabile di una variabile categoriale (la classe) che si traduce in un problema di stima della densità. Esistono differenti tecniche per il calcolo della stima:

- **Stimatori di densità del kernel** [47] o **rappresentazioni grafiche della densità articolare** [48].
- **Metodi basati sullo spazio metrico**: definire una misura della distanza sui punti dati e cogliere il valore della classe in base alla vicinanza ai punti dati nel set di training. Ad esempio, il metodo K-nearest-neighbour [47].
- **Proiezione in regioni decisionali**: dividere lo spazio degli attributi in regioni decisionali e associare a ciascuna di esse una previsione.

Raggruppamento

Conosciuto anche come segmentazione, il clustering non specifica i campi da prevedere ma mira a separare gli elementi di dati in sottoinsiemi simili tra loro. poiché, a differenza della classificazione, non conosciamo il numero di "cluster" desiderati, gli algoritmi di clustering utilizzano tipicamente una ricerca in due fasi: un ciclo esterno su possibili numeri di cluster e un ciclo interno per adattarsi al miglior cluster possibile per un dato numero di cluster. Dato il numero K di cluster, i metodi di clustering possono essere suddivisi in tre classi:

1. **Metodi basati sulla distanza metrica**: viene definita una misura della distanza e l'obiettivo è quello di trovare la migliore partizione (K-way).

2. **Metodi basati su modelli**: viene ipotizzato un modello per ciascun cluster; infatti, l'obiettivo è trovare il modello che si adatta meglio a ciascun cluster. Se M_k è il modello ipotizzato per un determinato cluster, un modo per valutarne la bontà potrebbe essere il calcolo della probabilità:

$$Prob(M_k|D) = Prob(D|M_k) \times \frac{Prob(M_k)}{Prob(D)} \tag{II.1}$$

   $Prob(D)$ è una costante e quindi può essere ignorata ai fini del confronto, mentre $Prob(M_k)$ è la probabilità a priori assegnata ad un modello.

3. **Metodi basati sulla partizione**: fondamentalmente enumerare varie partizioni e quindi valutarle in base ad un criterio. Le due tecniche precedenti possono essere viste come casi speciali di questa classe. Molte tecniche nella letteratura sull'IA rientrano in questa categoria e utilizzano funzioni di punteggio ad hoc.

Riepilogo dei dati

A volte l'obiettivo è semplicemente quello di estrarre modelli compatti che descrivono sottoinsiemi di dati. Esistono due classi di metodi che rappresentano l'acquisizione di sezioni orizzontali (casi) o verticali (campi) dei dati. Nel primo, si vorrebbe produrre riassunti di sottoinsiemi: ad es. produrre statistiche sufficienti o condizioni logiche valide per i sottoinsiemi. Nel secondo, si vorrebbe prevedere le relazioni tra i campi. Questa classe di metodi si distingue da quella sopra in quanto, anzichè prevedere un campo specificato (ad es. classificazione) o raggruppare i casi (ad es. clustering), l'obiettivo è trovare relazioni tra i campi. Un metodo comune è chiamato regole di associazione [23]. Le associazioni sono regole che stabiliscono che determinate combinazioni di valori si verificano con altre con una determinata fre-

quenza e certezza. Un'applicazione di questo tipo è l'analisi del paniere di mercato in cui si vuole capire quali tipi di prodotti possono essere venduti insieme ad altri.

Modellazione delle dipendenze

La comprensione dei dati viene spesso ottenuta derivando una struttura causale all'interno dei dati. I modelli di causalità possono essere probabilistici (come nel derivare alcune affermazioni sulla distribuzione di probabilità che governa i dati) o possono essere deterministici come nel derivare dipendenze funzionali tra campi nei dati [41]. I metodi di stima della densità in generale rientrano in questa categoria, così come i metodi per la modellizzazione causale esplicita (ad esempio [35] e [37]).

Rilevamento di modifiche e deviazioni

Questi metodi tengono conto delle informazioni sulla sequenza, che si tratti di serie temporali o di qualche altro ordinamento (ad esempio il sequenziamento delle proteine nella mappatura del genoma). L'ordinamento delle osservazioni è la caratteristica distintiva di questa classe di metodi.

II.3.3 Principali campi di applicazione

È enorme il numero di campi di applicazione del data mining; per questo motivo, sono stati raggruppati in macrocategorie che elenchiamo di seguito:

- **Marketing:** è possibile individuare e clusterizzare le tipologie di clienti accumunati da differenti caratteristiche oppure analizzare comportamenti dei clienti di un determinato brand, in modo da poter adottare soluzioni strategiche. Infine, è possibile, con analisi di associazione, estrarre quali prodotti o servizi vengono acquistati insieme dagli acquirenti.

- **Economia e finanza:** è utile nel rilevamento di frodi o nel tracciamento di atti fraudolenti. Inoltre, è possibile prevedere l'andamento di indici azionari e le interazioni fra diversi mercati finanziari.

- **Scienza:** un grosso vantaggio è per settori come medicina e biologia, meteorologia e atronomia.
- **Tecnologie dell'informazione e della comunicazione (ICT).**
- **Statistica:** velocizzando l'esecuzione delle analisi si riesce a fornire validi modelli predittivi.
- **Industria:** ottimo supporto alla logistica e all'efficienza produttiva.

II.3.4 Strumenti per il data mining

Gli strumenti per il data mining sono differenti poiché, a seconda dei campi di applicazione (Sezione II.3.3), cambiano gli obiettivi di analisi. Le metodiche utilizzate sono varie e spesso possono anche integrarsi tra loro.

Decision Tree

Il Decision Tree (DT) è utilizzato per classificare le istanze di grandi quantità di dati [49][50][51][52]. è un algoritmo in cui qualsiasi percorso inizia dalla radice di un albero ed è descritto da una sequenza di separazione dei dati fino al raggiungimento di un risultato booleano nei nodi foglia che rappresentano le classificazioni [53][54][55][56]. Ogni nodo rappresenta le caratteristiche in una categoria da classificare e ogni sottoinsieme definisce un valore che può essere assunto dal nodo [57][58]. Grazie alla loro semplice analisi e alla loro precisione su più moduli di dati, gli alberi decisionali hanno trovato molti campi di implementazione [59].
La Figura II.3 illustra una struttura di DT.

Gli alberi decisionali sono utilizzati principalmente per scopi di raggruppamento e sono uno dei potenti metodi comunemente usati in vari campi, come l'apprendimento automatico, l'elaborazione di immagini e l'identificazione di modelli [61]. I DT sono un modello che legano una serie di test di base in modo efficiente e coeso, in cui una caratteristica numerica viene confrontata con un valore di soglia in ciascun test [62].

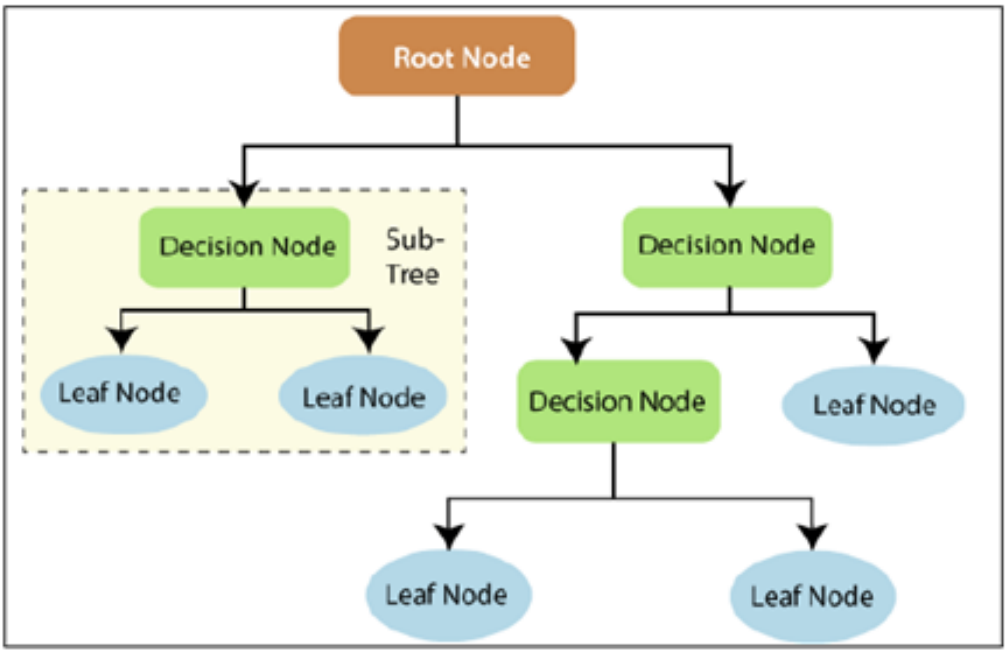

Fig. II.3: Decision Tree [60]

Le regole concettuali sono molto più facili da costruire rispetto ai pesi numerici nella rete neurale delle connessioni tra i nodi [63][64].

Random Forest

Random Forest (RF) è un algoritmo di classificazione basato sul modello dell'albero decisionale. Genera k diversi sottoinsiemi di dati di training partendo da un set di dati originale utilizzando un approccio di campionamento bootstrap; quindi, vengono creati k alberi decisionali [65]. Ciascun campione del set di dati di test è previsto da tutti gli alberi decisionali e il risultato finale della classificazione dipende dalle valutazioni dei questi alberi. Il set di dati di addestramento originale è formalizzato come:

$$S = (xi,_y j), i = 1, 2, ..., N; j = 1, 2, ..., M \tag{II.2}$$

dove x è un campione mentre y è una feature variable di S. Ciò vuol dire che il set di dati di training originale contiene N campioni, in ciascuno dei quali sono

presenti M feature variable. Il processo principale di costruzione dell'algoritmo RF è presentato in Figura II.4

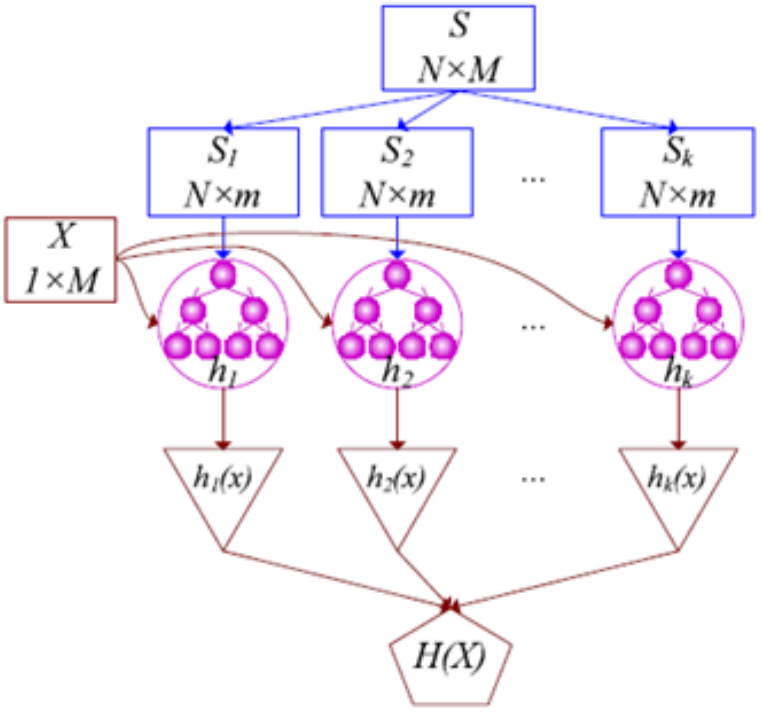

Fig. II.4: Processo di costruzione dell'algoritmo RF [66]

Reti Neurali

Una rete neurale artificiale (ANN), spesso chiamata semplicemente "rete neurale" (NN), è un modello matematico o un modello computazionale basato su reti neurali biologiche, in altre parole, è un'emulazione del sistema neurale biologico [67]. L'approccio è definito di connesssionismo, cioè basato su una rete composta da livelli di nodi che contengono un livello di input, uno o più livelli nascosti e un livello di output. Ciascun nodo, o neurone artificiale, si connette aad altri attraverso "sinapsi", che permettono il passaggio di informazione, generalmente mediata da una funzione di attivazione, o una soglia. Se l'output di qualsiasi singolo nodo è al di sopra del valore soglia specificato, tale nodo viene attivato, inviando i dati al successivo livello

della rete. In caso contrario, non viene passato alcun dato al livello successivo della rete.

Le reti neurali possono essere classificate a seconda della strategia di aggregazione dei differenti neuroni e dello scopo applicativo per il quale vengono ottimizzate. Qui di seguito, saranno riportati i tipi più comuni di reti neurali:

- **Il percettrone** è la più antica rete neurale, creata da Frank Rosenblatt nel 1958. Come mostrato in figura II.5, ha un singolo neurone ed è la forma più semplice di una rete neurale :

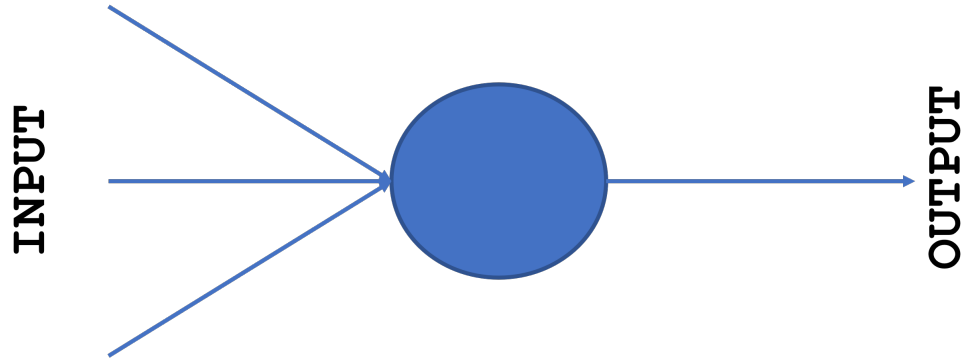

Fig. II.5: Modello Perceptron

- **Le reti neurali feedforward**, o percettroni multilvello (o MLP, multi-layer perceptron) sono formate da un livello di input, uno o più livelli nascosti e un livello di output. è importante notare che questo tipo di reti sono effettivamente formate da neuroni sigmoidi e non percettroni, poiché la maggior parte dei problemi del mondo reale è non lineare (Figura II.6). I dati vengono di norma passati in questi modelli per addestrarli e sono la base per la visione artificiale, l'NLP (natural language processing) e altre reti neurali.

- **Le reti neurali convolutive** (o CNN, convolutional neural network) sono simili alle reti feedforward, ma sono di solito utilizzate per il riconoscimento delle immagini, il riconoscimento dei modelli e/o la visione artificiale. Queste reti

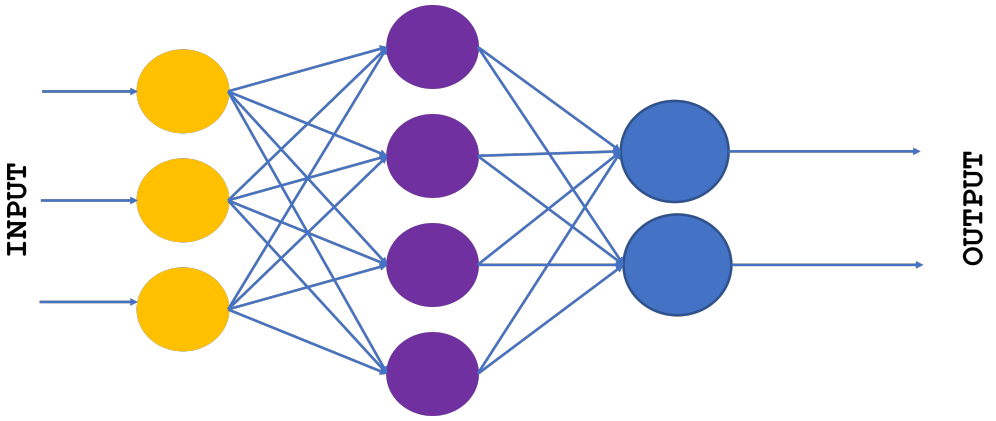

Fig. II.6: Rete Neurale Feedforward

sfruttano operazioni di convoluzione tra i pesi di ogni livello e i segnali entranti per produrre il segnale da trasmettere al livello successivo, per identificare i modelli all'interno di un'immagine.

- **Le reti neurali ricorrenti** (o RNN, recurrent neural network) includono neuroni collegati tra loro in un loop. Ci si avvale di questi algoritmi di apprendimento soprattutto quando si utilizzano dati di serie temporali per fare delle previsioni su risultati futuri, come ad esempio, le previsioni relative al mercato azionario o le previsioni di vendita o previsioni di guasto di una macchina.

II.3.5 Algoritmi di Data mining per la Manutenzione Predittiva

Il set di possibili algoritmi di Data Mining, sfruttati nella Manutenzione Predittiva (PdM), è vario. In questo lavoro di tesi, sono stati utilizzati sia le reti ricorrenti che il modello della regressione. Questi strumenti permettono di trovare una relazione tra più variabili associandogli un valore numerico che in questo caso corrisponderebbe alla RUL (Remaining Useful Life).
Nel paragrafo II.4, ci si addentra nel mondo della PdM e come quest'ultima si applichi al campo automobilistico.

II.4 La Manutenzione Predittiva

PdM è un approccio basato su vari tipi di informazioni che consentono ai responsabili della manutenzione, della qualità e delle decisioni operative di prevedere quando un asset necessita di manutenzione. La PdM è al centro del 21°secolo e si concentra non soltanto sui dati degli asset ma include altre tipi di informazioni quali l'ambiente circostante in cui opera l'asset, i processi e le risorse associate che interagiscono con l'asset. Fondamentalmente, PdM sfrutta varie tecniche analitiche per fornire una migliore visibilità dell'asset ai decisori. PdM raccoglie e analizza vari tipi di dati per fornire i seguenti vantaggi:

- Identificare i fattori che sono predittori chiave
- Determinare la probabilità di risultati predittivi
- Ottimizzare le decisioni applicando sistematicamente le conoscenze istituzionali

È importante comprendere i dati che vengono analizzati poiché la PdM si basa generalmente sulle caratteristiche di utilizzo e usura dell'asset, nonché su altre informazioni sulle condizioni dell'asset. I dati solitamente includono quanto segue:

- Dati di misurazione
- Dati di guasto

Inoltre, i dati disponibili possono trovarsi essere di diverso tipo:

- Dati storici: dati fissi che non cambiano nel tempo.
- Dati realtime: dati che cambiano nel tempo

II.4.1 Strategie di manutenzione dei veicoli

Una revisione della letteratura di [68], [69], [70], [71], [72], [73] e [74] rivela che esistono due categorie principali di strategie di manutenzione dei veicoli e alcune categorie possono includere sub-categorie:

1. Manutenzione correttiva/reattiva
2. Manutenzione preventiva
    - A Manutenzione a tempo
    - B Manutenzione basata sulle condizioni
        - i Manutenzione proattiva
        - ii Manutenzione predittiva

La manutenzione correttiva, viene eseguita dopo che un componente si è guastato. Cioè, in risposta ad un guasto che si è verificato, la strategia di manutenzione correttiva reagisce e lo risolve sostituendo o riparando i componenti rotti [69].
La manutenzione preventiva, iniziata negli anni '70 [72], mira a prevenire danni al veicolo mediante una rapida diagnosi dei guasti del veicolo [71]. Può essere suddiviso in due categorie: **Manutenzione basata sul tempo** e **Manutenzione basata sulle condizioni** (iniziata alla fine degli anni 2000 [72]). La manutenzione basata sul tempo (TBM) sostituisce i componenti preventivamente, in base ad un programma, per prevenire guasti imprevisti dei componenti [74]. Di conseguenza, i tempi di inattività sono programmati invece di essere completamente imprevisti. Il programma viene scelto in base al tempo previsto prima del guasto dell'apparecchiatura [74]. La manutenzione basata sulle condizioni (CBM) utilizza i dati diagnostici per decidere se è necessaria una riparazione nello stato attuale del veicolo [71][72]. In questo modo si evitano riparazioni non necessarie [74]. Il CBM consente una prognosi in tempo reale dell'apparecchiatura monitorando lo stato di salute di quest'ultima, che si ottiene analizzando i dati dei sensori che rappresentano lo stato fisico dell'apparecchiatura [74]. Inoltre, una revisione della letteratura di [72] e [73] rivela che le riparazioni sono programmate sui veicoli solo quando l'analisi dei sensori/parametri selezionati indica che è necessaria la manutenzione. Per questo motivo, il CBM può essere suddiviso in due sottocategorie: **Manutenzione proattiva** e **Manutenzione predittiva**. La manutenzione proattiva monitora l'apparecchiatura e tenta di identificare/definire i valori di soglia dei sensori/parametri target che identificano

al meglio le condizioni dell'apparecchiatura. Utilizzando queste soglie, l'obiettivo è prevedere futuri guasti del veicolo e i valori di soglia vengono adeguati quando necessario. La manutenzione predittiva è simile alla manutenzione proattiva, ma è più un approccio applicato in un ambiente di elaborazione distribuito. La manutenzione predittiva avviene dopo l'esecuzione della diagnosi dell'apparecchiatura. Leggendo i dati del sensore e altre informazioni diagnostiche, la manutenzione predittiva tenta di analizzare le condizioni (vita utile rimanente, modello di degrado, stato di salute, ecc.) dell'apparecchiatura e di prevedere quale tipo di manutenzione è necessaria per ridurre al minimo i costi e aumentare la durata dell'apparecchiatura. La manutenzione predittiva riduce i costi prevenendo guasti, manutenzione non programmata e tempi di fermo, e abbassa i costi assicurando che la sostituzione delle parti difettose avvenga solo quando necessario [75][76][74].

II.4.2 Approcci PdM nel campo automobilistico

È possibile utilizzare tre approcci per prevedere la manutenzione del veicolo:

1. Previsione della vita utile residua (RUL)
2. Rilevamento della deviazione
3. Classificazione supervisionata.

I tre apporcci possono utilizzare sia i dati storici che quelli in tempo reale come input. I metodi di rilevamento della deviazione si distinguono dagli altri metodi in quanto non forniscono informazioni specifiche su ciò che è guasto. Mentre, ad esempio, nell'approccio RUL, il componente guasto è intrinsecamente noto poiché questi modelli sono sviluppati per un componente specifico. Per quanto riguarda, l'approccio di classificazione può essere utilizzato sia per rilevare le deviazioni che per isolare l'effettivo componente guasto. Il rilevamento delle deviazioni è un problema più facile da risolvere poiché viene trascurata l'incertezza tra le deviazioni e i guasti. Per rendere i metodi di rilevamento delle deviazioni completi (e comparabili ai metodi RUL e Classification) devono essere combinati con dati storici di deviazioni legati

a guasti noti. Ciò introduce un'ulteriore incertezza nella previsione completa, che riduce l'accuratezza della previsione.

Sono differenti gli studi in letteratura dei tre approcci ma maggiore attenzione si presta al primo (RUL) poiché risulta essere quello trattato in questo lavoro di tesi.

Metodo RUL

La previsione RUL dei componenti critici è un compito non banale per diverse ragioni. I segnali dei sensori, ad esempio, sono generalmente oscurati dal rumore e quindi è molto difficile elaborare ed estrarre la rappresentazione informativa della RUL [77]. Un altro problema è l'incertezza di previsione dovuta alla variazione del tempo di fine vita che può differire per componenti realizzati dallo stesso produttore e funzionanti nelle stesse condizioni. Pertanto, i modelli proposti dovrebbero includere tali incertezze e rappresentarle in una forma probabilistica ([78]; [79]). I modelli di previsione RUL possono essere realizzati utilizzando due metodi diversi([80]):

1. **Metodi basati sulla fisica:** costruiscono modelli fisici delle componenti critiche desiderate per mezzo di modelli dello spazio degli stati ([81]) ed equazioni differenziali dinamiche ordinarie o parziali ([82]). Questi modelli richiedono un'ampia sperimentazione e verifica del modello ([83]), sono molto affidabili almeno fino a quando il sistema non viene aggiornato o modificato ([84]).

2. **Metodi basati sui dati:** possono essere utilizzati quando i primi principi del funzionamento del sistema sono complessi e lo sviluppo di un modello di fisica del guasto accurato non è fattibile ([85]; [86]).

I metodi di questa categoria utilizzano tecniche di riconoscimento di modelli e apprendimento automatico per caratterizzare il comportamento di degradazione dei componenti critici desiderati ([87]). Un modo per eseguire la previsione RUL basata sui dati è stimare prima lo stato di salute attuale del componente desiderato e quando il degrado supera la soglia di allarme, gli algoritmi iniziano a prevedere la RUL ([88]; [89]; [90]). In letteratura, sono stati proposti diversi modelli di regressione per affrontare problemi di predizione RUL basati sui dati come il modello

auto regressivo e le spline di regressione adattativa multivariata ([91], [92],[93], [94], [95], [96]). Uno svantaggio dell'utilizzo dei metodi di regressione è che quando la cronologia di degradazione dei componenti disponibile è incompleta, l'estrapolazione può portare a grandi errori ([97]). Ultimamente, la sperimentazione è stata eseguita su vari tipi di reti neurali e sistemi neurali fuzzy ([98], [99], [100], [101], [102], [103], [77], [104], [105], [106], [107]). Tuttavia, questi metodi generano modelli "black box" poiché difficile selezionare la struttura della rete ([108])

I metodi basati sulla somiglianza si sono dimostrati molto efficaci nell'esecuzione della previsione RUL. Questi metodi sono basati sulla regressione lineare per costruire modelli di degradazione offline, un approccio proposto è in [109]. Il metodo misura la somiglianza tra l'istanza di test e i modelli offline e l'istanza offline selezionata viene utilizzata per la previsione RUL. La densità di probabilità RUL dell'istanza di test viene stimata in base a più previsioni locali utilizzando il metodo di stima della densità del kernel. Il problema principale con questo metodo è la selezione manuale dei dati informativi del sensore.

In alternativa, invece di apprendere il degrado dai dati e prevedere il RUL, gli algoritmi di previsione RUL diretta apprendono la relazione tra i dati del sensore e la fine del ciclo di vita per prevedere la RUL. Per fare ciò, gli indicatori di salute vengono estratti dai segnali di monitoraggio grezzi, che possono essere originati da un singolo sensore o da più sensori aggregati per rappresentare l'evoluzione del degrado nel tempo.

Capitolo III

Caso di studio: SEA - Soluzioni Eco Ambientali s.r.l.

III.1 Introduzione

Nel Capito II sono stati descritti ed esaminati il contesto e le tecniche e metodologie che verranno utilizzate per raggiungere i risultati sperimentali di questo progetto. In particolare, ci si è soffermati sul Data Mining che, fornendo strumenti come algoritmi di classificazione, permette la progettazione di un modello di Manutenzione Predittiva.

Questo capitolo, volge l'attenzione al caso di studio, oggetto di questa tesi: SEA Soluzione Eco Ambientali s.r.l.

SEA è l'azienda che ha commissionato il progetto e, conseguentemente, fornito tutti i dati disponibili per la realizzazione di quest'ultimo. Il progetto riguarda l'applicazione di tecniche di Manutenzione Predittiva su dati provenienti da sensori a bordo di una flotta di veicoli (dell'azienda stessa) che eseguono la raccolta differenziata. L'obiettivo, per cui, è quello di riuscire a fornire a SEA uno strumento che possa ridurre i costi derivanti dagli interventi di manutenzioni dei veicoli.

III.2 Presentazione di SEA

La SEA - Soluzioni Eco Ambientali - è una società nata nel 2013 a Villanova Canavese (TO) specializzata in raccolta rifiuti urbani, elaborazione di progetti ecologici e in servizi d'igiene ambientale (nettezza urbana, pulizie mercatali etc.). L'azienda utilizza tecnologie di ultima generazione, così da offrire i migliori servizi disponibili in ambito ambientale coinvolgendo attivamente tutti gli attori del ciclo di produzione, raccolta e smaltimento di rifiuti urbani. Contraddistinta da un management giovane e coraggioso, che senza importanti investitori sospinge l'azienda verso risultati ambiziosi, fondandosi sui principi di sostenibilità, eccellenza operativa, innovazione tecnologica e resilienza.

Fig. III.1: Logo di SEA Soluzioni Eco Ambientali [110]

L'obiettivo di SEA è essere il punto di riferimento italiano nel settore dell'ecologia urbana, diventando il principale artefice del cambio di paradigma del rifiuto urbano, da scarto senza valore a risorsa centrale della sostenibilità ambientale, economica e sociale.

Inoltre, uno dei punti di forza dell'azienda è la collaborazione con enti del territorio per la ricerca delle soluzioni più all'avanguardia nel rispetto delle normative e nel miglioramento dell'ambiente e della società, nonché con aziende del settore, fornitori, società tecnologiche (italiane e straniere), università e centri di ricerca, per innovare costantemente i mezzi e le tecnologie.

Tra i diversi servizi di ecologia urbana vi sono:

- Raccolta rifiuti urbani e speciali

- Trasporti rifiuti
- Pulizie mercatali
- Spazzamenti manuali e meccanizzati
- Lavaggi e sanificazione stradali
- Lavaggio cassonetti
- Intermediazione rifiuti
- Gestione dei Centri di Raccolta
- Gestione di impianti di stoccaggio
- Selezione e trasferenza rifiuti
- Educazione e comunicazione ambientale

Azienda a forte espansione sia territoriale che di fatturato, attualmente attiva in numerose regioni d'Italia (vedi figura III.2) con 14 cantieri operativi, 300 mezzi e 250 dipendenti.

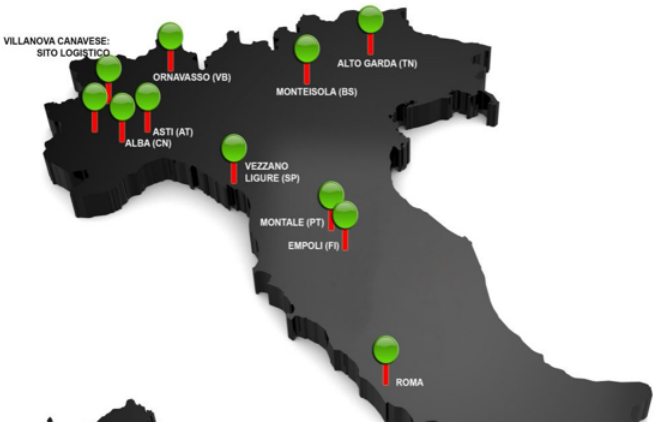

Fig. III.2: Cantieri attivi di SEA nelle regioni d'Italia [110]

III.3 Descrizione del caso di studio

Questa sezione, ha lo scopo di condurre un'analisi preliminare sui dati della flotta di raccolta rifiuti gestita da SEA. L'azienda agisce in collaborazione con l'amministrazione locale in molti comuni italiani, per gestire la raccolta, il trasporto e la selezione dei rifiuti urbani. A questo scopo, l'azienda impiega una flotta di veicoli. Essi sono dotati di sensori in grado di tracciare il veicolo durante i viaggi. Come detto precedentemente, ogni giorno, più di 300 veicoli operano in 14 diversi cantieri, situati in diverse città del nord e del centro Italia. Un guasto imprevisto di uno di questi veicoli richiederebbe una rapida riorganizzazione del carico di lavoro di tutti gli altri, per svolgere il servizio in modo ottimale. Quindi, un rilevamento precoce dei segnali che portano ad un possibile guasto è di primaria importanza, sia da un punto di vista economico che organizzativo.

Fornitori di dati

I fornitori di dati, che hanno permesso la raccolta di differenti informazioni sui veicoli, sono 13. I dati acquisiti ad oggi provengono prevalentemente da API proprietarie dei fornitori (software gestionali, produttori automezzi, sistemi di retrofit, etc..). Nella Tabella III.1 sono stati evidenziati 3 tipi di fornitori: **P1, P2 e P3** poichè raccolgono dati più promettenti per la realizzazione di un modello predittivo poiché gestiscono il tracking dei veicoli. Infatti, risultano essere installati sulla maggior parte di essi. Quest'ultimi, nel capitolo IV, verranno esaminati nel dettaglio.

ID	Fornitore	Descrizione
1	**P1**	Monitoraggio della mobilità di veicoli
2	**P2**	Fornitore di servizi come finanziamento, assicurazione e noleggio
3	**P3**	Monitoraggio in tempo reale del comportamento e posizione dei mezzi
4	P4	Fornitore di assistenza tempestiva alla flotta
5	P5	Fornitore di pneumatici
6	P6	Gestione del carburante
7	P7	Gestione programmata della manutenzione
8	P8	Fornitore di componenti elettriche
9	P9	Gestione da remoto della flotta
10	P10	Fornitore di macchine movimento terra
11	P11	Servizio di localizzazione
12	P12	Gestione dei rifornimenti
13	P13	Gestione delle timbrature

Tab. III.1: Fornitori di dati

Tipologia di veicoli

La flotta su cui eseguire un'analisi predittiva è eterogenea; infatti, è composta da 58 che si differenziano sia per la tipologia che per i fornitori di dati:

1. **Compattatore (13 unità)**: è un macchinario che può contenere e comprimere rifiuti misti in grande quantità.
    - P14 + P7 + P1 (5 unità)
    - P3 + P1 (8 unità)
2. **Mini-compattatore (22 unità)**: sono dei minicostipatori a caricamento posteriore, con scarico per ribaltamento della vasca, dotato di sistema di compattazione del tipo a monopala articolata.
    - P2 + P1 (11 unità)
    - P14 + P1 (11 unità)
3. **Vasca (23 unità)** : sono vasche semplici a caricamento posteriore, con scarico per ribaltamento tramite apposito cilindro a doppio effetto a singolo stelo.
    - P2 (9 unità)
    - P14 (14 unità)

Problema di SEA

La manutenzione della propria flotta e i consumi di carburante spesso rappresentano le voci di costo che gravano di più sul bilancio.
Ciò vuol dire che, se un'impresa possiede uno o più mezzi pesanti dovrebbe scegliere un piano di manutenzione preventiva per contenere i rischi, riducendo al minimo i costi e aumentando i profitti. Le manutenzioni eseguite periodicamente riguardano tutte le componenti essenziali per il corretto funzionamento dei mezzi pesanti come per esempio:

1. Sistema frenante

2. Cambio
3. Perdita olio (da differenziale)
4. Radiatore
5. Pala meccanica
6. Rastrelliera
7. Usura e snervamento gomme

I costi di manutenzione dei veicoli dell'azienda si dividono in **costi di manutenzione interna** e **costi di manutenzione esterna**. I primi sono calcolati come costo medio orario del lavoro poiché si trattano di interventi di manutenzione eseguite presso le officine di proprietà dell'azienda. Mentre, i secondi sono interventi di manutenzioni fatturati da officine esterne.

Obiettivo del progetto

Come già detto in precedenza, l'obiettivo è quello di analizzare i dati provenienti dai sensori a bordo della flotta di veicoli, cercare di capirne il comportamento e sulla base di ciò, riuscire a riorganizzare gli interventi di manutenzione per ridurre al massimo gli sprechi e aumentarne i vantaggi dell'azienda. Ciò vuol dire:

- **Incrementare i ricavi**: se il numero degli interventi di manutenzione hanno un calo, lo avranno anche i tempi operativi e ciò comporta una gestione più efficiente.
- **Riduzione dei costi di manodopera**: ridotti grazie al numero inferiore di interventi di sostituzione sui componenti dei veicoli.
- **Riduzione dei tempi di inattività persi**: un veicolo fermo, in manutenzione, provoca inefficienze che si ripercuoterebbero su attività della value chain.

- **Aumentare la sicurezza**: migliorano anche le condizioni lavorative dell'azienda, divenendo più sicure.
- **Aumentare l'efficienza del manutentore**: sarà in grado di fornire un servizio migliore grazie alla fornitura di una descrizione accurata della manutenzione da eseguire, informazioni della reperibilità delle parti di ricambio, strumenti e supporto necessario per correggere l'anomalia.
- **Creazione di un database**: possedere uno storico dei dati relative alle prestazioni e al comportamento dei veicoli migliora il grado di precisione delle previsioni future.

Capitolo IV

Schema del Progetto

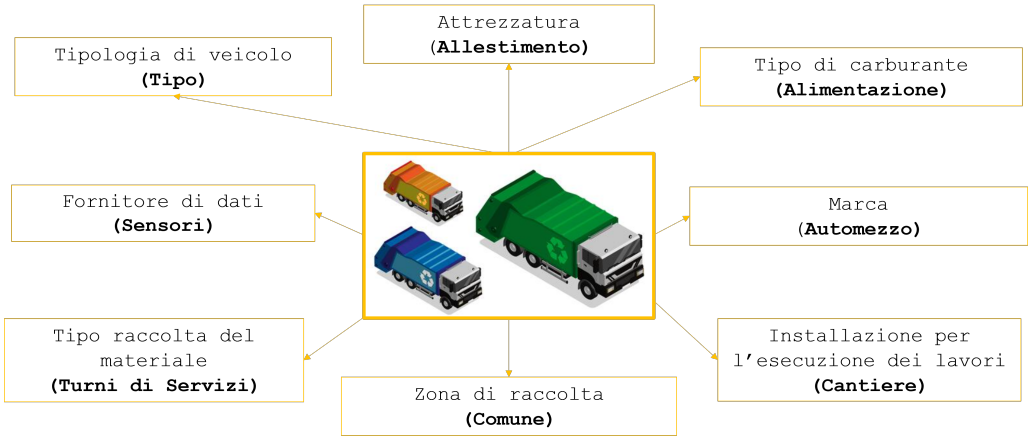

Fig. IV.1: Schema del progetto

IV.1 Introduzione

Lo scopo di questo capito è quello di presentare una panoramica del lavoro svolto. Infatti, verranno presentati gli strumenti utilizzati per l'esecuzione di questa ricerca; in seguito, saranno analizzati e confrontati tre diversi fornitori di dati (presentati nel capitolo III), ciascuno di essi possiede un sensore proprio. Un determinato parametro può essere specifico di un singolo fornitore oppure può essere monitorato da diversi fornitori (ognuno con le sue peculiarità).

Inoltre, sono stati digitalizzati e categorizzati tutti gli interventi di manutenzioni

(interni ed esterni all'azienda) eseguiti sui veicoli di interesse, utili per evidenziare comportamenti o anomalie notevoli. Il focus è sulla fase esplorativa dei fornitori di dati e sugli interventi di manutenzione, per i quali è stato pubblicato un articolo [111].
In realtà, al fine di progettare un modello predittivo più accurato, è stato eseguito uno studio sulle condizioni lavorative dei veicoli. Ciò implica che possano esserci delle possibili correlazioni tra guasti e il contesto che circonda il veicolo. Pertanto, parte del lavoro è stato dedicato all'analisi dei Turni di Servizio e dei Dipendenti alla guida dei veicoli.

IV.2 Strumenti di supporto

Prima di analizzare gli elementi del progetto, sono ilustrati gli strumenti adottati che supportano operazioni di Data Mining:

- **Jupyter Notebook** è un ambiente computazionale interattivo basato sul Web per la creazione di documenti notebook Jupyter. Supporta diversi linguaggi come Python (IPython), Julia, R, ecc. ed è utilizzato principalmente per l'analisi dei dati, la visualizzazione dei dati e altri calcoli esplorativi interattivi [112].

- **Python** è un linguaggio di programmazione, orientato agli oggetti e di alto livello con semantica dinamica. Python supporta diversi moduli e pacchetti:

    1. **Pandas** è un pacchetto Python open source ampiamente utilizzato per attività di data science/analisi dei dati e machine learning. È costruito su un altro pacchetto chiamato **Numpy**, che fornisce supporto per array multidimensionali [113].
    2. **Matplotlib** è una libreria multipiattaforma, di visualizzazione dei dati e di tracciamento grafico.
    3. Il modulo **sklearn** per algoritmi di tipo supervisionato e non supervisionato.

IV.3 Fornitori di dati

In questa sezione, verrà presentata un'analisi preliminare e caratterizzazione dei dati disponibili dei fornitori di dati. In particolare, sono stati selezionati 3 fornitori (P1, P2 e P3) ed un sottoinsieme di 40 veicoli, che sono i più rappresentativi e ben monitorati, ovvero quelli che possedevano maggiori rilevazioni nel tempo. La memorizzazione dei dati avviene per mezzo di un Data Lake di proprietà della SEA e gestito da T2D Transfer to Digital che, tramite API REST dei fornitori di dati, raccoglie il contenuto richiesto dal consumatore (la chiamata). I fornitori di servizi possono essere i produttori del veicolo oppure dei fornitori esterni che hanno aggiunto i loro sensori sia sul telaio che sugli allestamenti dei veicoli. Pertanto, le fonti di dati sono estremamente eterogenee e irregolari nella misurazione temporale, ciò comporta la presenza di una varietà di anomalie.
Tra tutte le fonti di dati disponibili dei fornitori P1, P2 e P3, importanti per il raggiungimento dell'obiettivo della ricerca sono i dati di tracciamento che descrivono i percorsi intrapresi da ogni veicolo.

P1, P2 e P3, hanno feature di tracciamento diverse: le uniche misure comuni sono la posizione **GPS**, il valore del **contachilometri**, il **tempo totale di attività del motore** e la **velocità media**. Alcuni provider includono anche altre misure come l'**accelerazione** e il **consumo stimato di carburante**.

Oltre alla variabilità delle feature tracciate, i diversi fornitori hanno anche politiche diverse sul campionamento dei dati, il che porta a tassi di misura giornalieri molto diversi. I set di dati grezzi raccolti dai fornitori sono delle serie temporali di un totale di:

- 117880 campioni di 38 feature per 35 veicoli per P1
- 68611 campioni di 68 feature per 8 veicoli per P2
- 91072 campioni di 17 feature per 15 veicoli per P3

IV.4 Fatture di Manutenzione

Come già detto precedemente, tra gli obiettivi della ricerca, oltre l'analisi esplorativa, si valuta la fattibilità di un futuro modello predittivo che miri a prevedere i guasti sulla base degli interventi storici di manutenzione. A questo scopo, abbiamo analizzato la lista delle fatture di tutte le riparazioni dei diversi componenti dei veicoli. Dal grafico IV.2, si ottiene un totale di 120 fatture associate a 36 veicoli. I diversi fornitori (P1, P2 e P3 sulle righe) forniscono informazioni sui 40 veicoli (sulle colonne).

La heatmap mostra il numero medio giornaliero di campioni, per ogni veicolo (colonna) e fornitore (riga). La prima riga riporta il numero di fatture per ogni veicolo, ciascuna corrispondente agli interventi di manutenzione.

Tra i 40 veicoli, i primi 17 condividono i dati di tracciamento da più fornitori allo stesso tempo, e sono descritti da più di 30 campioni al giorno per veicoli in alcuni casi, con una media inferiore a 5 campioni al giorno per veicolo. Il numero di interventi di manutenzione (fatture) varia da zero a 15 nei 9 mesi in analisi.

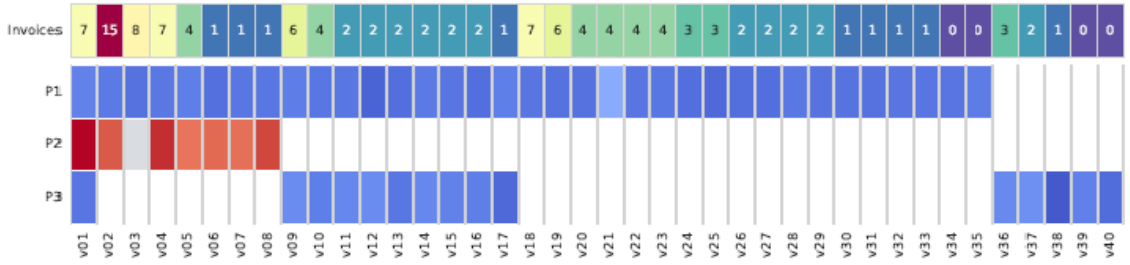

Fig. IV.2: Panoramica delle fatture e dell'attività giornaliera dei sensori sui veicoli associati [111]

Oltre ad aver analizzato la raccolta di fatture che indicano guasti di diversi componenti sui veicoli, con l'aiuto di esperti del settore, abbiamo escluso manualmente tutti gli interventi associati alle operazioni di manutenzione periodica da eseguire per legge (ad esempio, revisione, tagliando).

In seguito, con l'ausilio del questionario FMEA (Failure Mode Effect Analysis) dei meccanici, fornito dall'azienda, le fatture sono state opportunamente raggruppate in

categorie in base al tipo di guasto di un componente.
Analizzando le fatture categorizzate, si osserva un numero significativo di interventi della stessa categoria effettuati sullo stesso veicolo in un breve lasso di tempo (ad esempio, il giorno dopo). Ciò, può essere giustificato dal fatto che un solo intervento non fosse sufficiente per la riparazione di un determinato guasto. Al fine di un'analisi di manutenzione predittiva risultano essere irrilevanti; quindi, sono stati trascurati. La figura IV.3, mostra il numero risultante di guasti raggruppati per categoria dopo la pulizia dei dati. Per ogni riga, oltre al numero totale di guasti, è riportato anche il numero di seconde occorrenze, cioè quelle che non sono i primi guasti per il veicolo e la categoria. I guasti provenienti da questo secondo gruppo sono particolarmente interessanti quando si studia la possibilità di costruire un modello predittivo su questi dati, poiché possono dare una misura preliminare delle frequenze dei guasti. I processi di preprocessing e corretto raggruppamento delle fatture disponibili sono fondamentali poiché i dati raccolti saranno sfruttati per affrontare un'analisi di manutenzione predittiva, costruendo un modello predittivo in grado di stimare la vita residua di un veicolo (RUL) prima che sia necessario un intervento di manutenzione. La possibilità di individuare in anticipo malfunzionamenti o problemi ai componenti è infatti un aspetto fondamentale per le aziende automobilistiche, che possono così intervenire adeguatamente prima che si verifichi un guasto. Questo comporta un notevole risparmio economico per l'azienda, in quanto un modello predittivo correttamente implementato permetterebbe di evitare guasti durante la marcia del veicolo e di monitorare costantemente le condizioni dei vari componenti, programmando gli interventi di manutenzione solo quando sono realmente necessari.

Dalla Figura IV.3, si nota che per alcune categorie non vi sono abbastanza dati per poter costruire un modello predittivo robusto per la stima dei guasti futuri. Questo è causato dal periodo di tempo limitato di monitoraggio dei veicoli, che risulta essere breve per raccogliere abbastanza interventi di manutenzione per tutte le categorie.

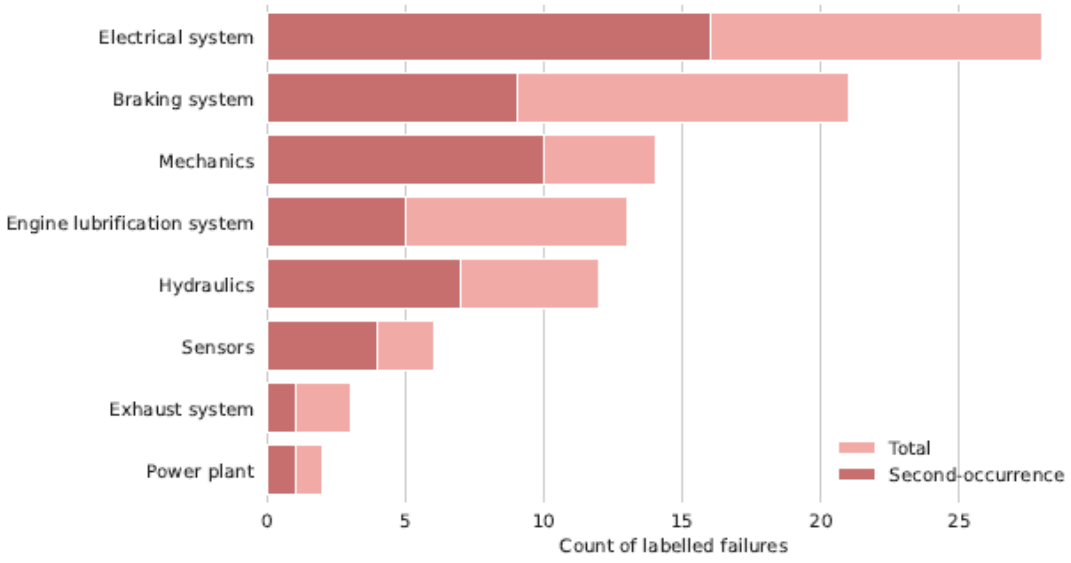

Fig. IV.3: Conteggio pulito dei fallimenti, separati per categoria [111].

IV.5 Feature e Ipotesi di Correlazione

Fin'ora, particolare attenzione hanno avuto le componenti meccaniche dei veicoli ma in realtà, per progettare un modello predittivo più accurato, è stato eseguito uno studio sulle possibili correlazioni tra guasti e la realtà che circonda il veicolo. Per tale motivo, le variabili di interesse sono sia correlate al veicolo stesso (vedi tabella IV.1) in modo da mappare lo stato di degrado di quest'ultimo, ma anche al contesto in cui il veicolo opera (vedi tabella IV.2).

CARATTERISTICHE DEL VEICOLO		
Variabile	**Descrizione**	**Feature**
Tipo veicolo	Tipologia di veicolo	Tipo
Allestimento	Configurazione del veicolo	Allestimento
Cantiere	Installazione per l'esecuzione dei lavori	Cantiere
Alimentazione	Tipo di carburante	Alimentazione
Automezzo	Marca	Automezzo

Tab. IV.1: Ipotesi di correlazioni tra Guasti e Caratteristiche del veicolo

CARATTERISTICHE CONTESTUALI DEL VEICOLO		
Variabile	**Descrizione**	**Feature**
Operatore	Stile di guida degli operatori	Età, Anni di Servizio
Servizio	Compattazione del tipo di materiale Comuni hanno diversa Altitudine e Area	Servizio, Comune

Tab. IV.2: Ipotesi di correlazioni tra Guasti e il Contesto in cui il veicolo vive

IV.5.1 Caratteristiche del veicolo

Algoritmo di Clustering

L'analisi del clustering è utilizzata per identificare un gruppo di oggetti correlati. Esistono differenti algoritmi di clustering classificabili come partizionali completi, non completi, di tipo gerarchico e basati su concetto di densità.
In questa tesi, l'algoritmo utilizzato è il **K-means**: algoritmo di tipo partizionale che fornisce soluzioni basati sul concetto di centroide:

- Ogni **cluster** è rappresentato dal proprio centroide.
- Ogni **punto** è assegnato al cluster con centroide più vicino; Ciascun punto troverà un gruppo sicuro al quale verrà assegnata un'etichetta di classe, quindi, è una soluzione completa.
- **K** rappresenta il numero di gruppi che vogliamo trovare.
  Si esegue più volte l'algoritmo, in modo da poter settare opportunamente K e trovare un buon partizionamento.

L'obiettivo dell'algoritmo di clustering è quello di raggruppare tutti quei veicoli che hanno eseguito interventi di manutenzione di una certa tipologia con una determinata frequenza e capire se questi abbiano una correlazione con le caratteristiche del veicolo (ad esempio, stessa tipologia, stesso allestimento, stesso cantiere, ecc.).

IV.5.2 Caratteristiche contestuali del veicolo

Turni di Servizio

Un buon coordinamento dei turni di servizio della raccolta e smaltimento rifiuti è uno degli aspetti fondamentali per l'azienda, sia per riuscire ad offrire un servizio efficiente per la comunità, sia per creare un vantaggio in termini economici.
L'organizzazione delle attività di raccolta di un determinato mezzo (ad esempio, l'ubicazione, il sottoutilizzo o eccessivo carico di lavoro), potrebbe impattare sulla frequenza dei guasti di quest'ultimo.
Inoltre, un altro aspetto importante risulta essere la gestione di alcuni tipi di dati, come il materiale raccolto e la pesatura, poiché potrebbero influenzare sia le parti meccaniche di un veicolo nonché il fatturato dell'azienda.
Per quanto riguarda la struttura dei dati, si dispone di un dataset suddiviso in cartelle a seconda del cantiere. Ciascuna di esse contiene due tipi di file:

- calendari dei servizi 2021
- calendari dei servizi 2022

La Tabella IV.3 mostra il numero di veicoli di interesse presenti nei dataset, divisi per cantiere e periodo di tempo.

TURNI DI SERVIZIO		
Cantiere	**Veicoli di interesse 2021**	**Veicoli di interesse 2022**
ASTI	0	0
BRA	-	18
EMPOLI	0	0
FIRENZE	2	4
LANGHE	14	16
PRATO	3	0
UDINE	3	4
VCO	24	-
VILLANOVA	4	-

Tab. IV.3: Cantiere cui appartengono i veicoli di interesse nel 2021 e nel 2022

I risultati sperimentali (capitolo V) derivanti dall'analisi esplorativa di questo set di dati verrà confrontata con i dati di tracciamento dei fornitori di dati; ciò

per comprendere la congruenza tra dati raccolti manualmente e dati provenienti dai sensori.

Dipendenti dell'azienda

Precedenti studi [114], hanno analizzato l'impatto dello stile di guida sul controllo tecnico del veicolo. Infatti, sono diversi i questionari condotti per raccogliere dati sullo stile di guida e sulle abitudini di manutenzione delle diverse tipologie di possessori di veicoli.
I risultati emersi sostengono che le condizioni dei veicoli cambiano in base alla classe di appartenenza di una determinata persona fisica; poiché, diversi fattori (ad esempio, età o data rilascio della patante) possono essere associati ad uno stile di guida più aggressivo che causa un maggiore sfruttamento del veicolo.
Pertanto, lo scopo dell'analisi dei dipendenti è quello di identificare la persona alla guida di un veicolo che potrebbe influenzare il periodo temporale con cui sono eseguiti gli interventi di manutenzioni.

Il dataset **DipendentList** è composto da 7639 records e 16 attributi e riporta informazioni personali sui dipendenti e sul rapporto di lavoro (Tabella IV.4).
I dati sono raccolti da: 1 Giugno 2021- 14 Giugno 2021 (ad eccezione del 2 Giugno).

Nel capito V sarà eseguita un'analisi esplorativa del dataset DipendentList e si verificherà se esiste una correlazione tra la variabile IdUtente del dataset DipendentList e le variabili che identificano gli utenti nei dati raccolti dai sensori a bordo dei veicoli.

IV.6 Modelli proposti per la Manutenzione Predittiva

I modelli proposti per la manutenzione predittiva sono due:

1. **LSTM** (Memoria A Lungo Termine) è un'architettura di rete neurale ricorrente artificiale (RNN) utilizzata nel campo del machine learning.

DipendentiList			
Attributi	**Formato**	**Tipologia**	**Unità di misura**
IdUtente	numero	categoria	
Cognome	stringa	categoria	
Nome	stringa	categoria	
CodiceFiscale	stringa	categoria	
DescrizioneRapporto	stringa	categoria	
DataAssunzione	stringa	tempo	data
Mansione	numero	categoria	
PartTime	stringa	categoria	
DataCessazione	stringa	tempo	data
TempoDeterminato	stringa	categoria	
Livello	stringa	categoria	
DescrizioneContratto	stringa	categoria	
CentroDiCosto	numero	categoria	
FondoTfr	numero	categoria	decimale
TrattenuteSindacali	stringa	categoria	
CessioneQuinto	stringa	categoria	

Tab. IV.4: Lista dei Dipendenti di SEA

Negli esperimenti dell'LSTM, sono stati utilizzati i dati del fornitore P1, in particolare uno specifico dataset di quest'ultimo.

2. **XGBoost** è un modello interpretabile a differenza delle reti neurali e quindi, permette di comprendere l'output prodotto dal modello in una determinata predizione, quali sono le regole interne che hanno portato il modello ad etichettare un record piuttosto che un altro. Negli esperimenti dell'XGBoost sono stati considerati i dati del fornitore P2, uno specifico dataset, cioè quello che presenta più sample per giorno per ogni truck, infatti, sono presenti circa 40 record per truck al giorno.

Time Series

La fase di preprocessing dei dati è pressoché uguale nei due modelli predittivi. Pertanto, presenteremo uno dei due approcci proposti, ovvero il modello LSTM.

Per migliorare la qualità dei risultati del modello, è stato eseguito un processo di Feature Engineering che consiste nell'aggregare giornalmente le informazioni a disposizione. Quindi, gli input del modello sono sample giornalieri su cui sono stati applicati varie statistiche aggregate (min. max, mean, std, median, count).

Le variabili estratte sono state filtrate per mezzo di una particolare tecnica di Feature Selection, ovvero, dopo aver generato la matrice di correlazione, raffigurata in IV.4, sono state analizzate le correlazioni tra ogni coppia di variabili. Se due variabili risultavano fortemente correlate tra loro, se ne mantiene solamente una, poiché troppe feature correlate, in ingresso al classificatore, potrebbero causare rumore.

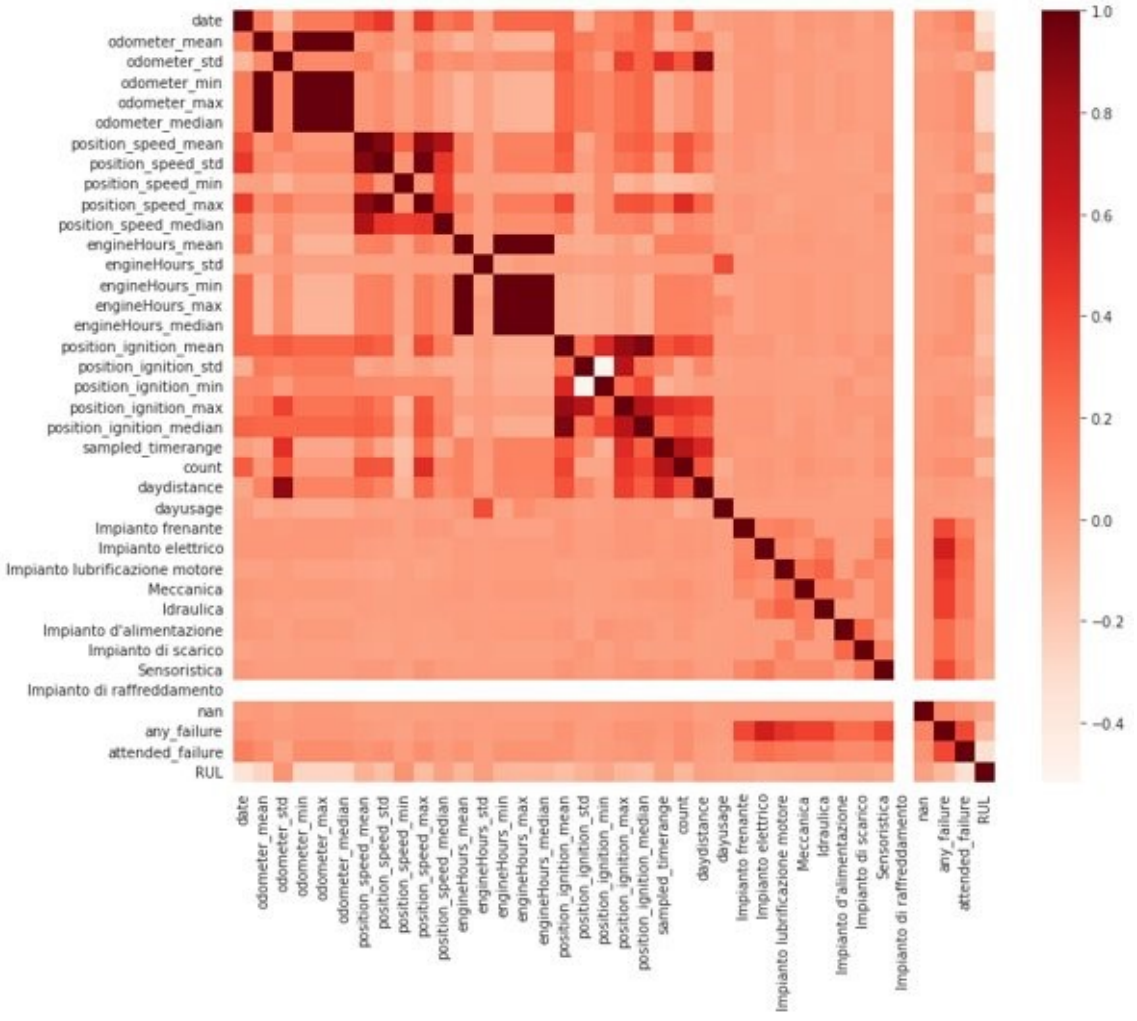

Fig. IV.4: Matrice di correlazione (feature di P1)

Dalla matrice di correlazione, si evince che alcune variabili siano fortemente

correlate tra loro, questo perché spesso la media e la mediana coincindono.

Infine, nel modello LSTM sono state aggiunte variabili come:

- **daydistance**: differenza tra il valore massimo e minimo dell'jupyter lab odometro
- **dayusage**: differenza tra il valore massimo e minimo del tempo di funzionamento del motore

Mentre, nell'XGBoost le variabili aggiunte sono:

- **HRTotalVehicleDistance**: differenza tra il valore max e min dell'odometro
- **EngineTotalFuelUsed**: differenza tra il valore max e min del carburante totale utilizzato dal motore

Capitolo V

Analisi e Risultati Sperimentali

V.1 Introduzione

Nel capitolo precedente, sono stati presentati i diversi elementi costitutivi del progetto. Adesso, in questo capitolo, verranno sottoposti alle analisi sperimentali, seguite dai relativi risultati.

V.2 Analisi dei Fornitori

In questa sezione, descriviamo il processo di analisi esplorativa dei fornitori di dati P1, P2 e P3, i quali saranno confrontati, evidenziando eventuali vantaggi e svantaggi [111].

V.2.1 Analisi esplorativa

Il dataset include le informazioni di 40 veicoli monitorati da maggio 2021 a gennaio 2022.
Anche se le API REST dei fornitori di dati sono state utilizzate direttamente per raccogliere i dati a intervalli regolari e sincronizzati durante l'intero periodo di tempo, i timestamp di alcuni campioni si riferiscono a periodi sensibilmente lontani dall'intervallo (maggio 2021-gennaio 2022). Inoltre, porzioni significative dei dati

originali sono duplicate, cioè diverse chiamate alle API del provider restituiscono gli stessi dati, anche con timestamp identici. Alcuni di questi dati ripetuti anomali sono associati a un periodo successivo senza dati raccolti.
Tuttavia, sembrerebbe dall'andamento generale dei veicoli che quest'ultimi erano in movimentato senza essere tracciati. Per questo motivo, abbiamo eliminato tutti i valori duplicati, poiché si è supposto che siano solo il risultato di irregolarità dei sistemi di aggiornamento dei fornitori.
Allo stesso modo, sono stati rimossi anche tutti i campioni al di fuori dell'intervallo di tempo considerato. Poiché ad ogni chiamata API, tutte le variabili sono restituite dal fornitore di dati, ma non tutte le variabili hanno effettivamente registrato un cambiamento nel loro valore, è stata utilizzata una fase di selezione delle feature per rimuovere quelle irrilevanti.

Quindi tra le operazioni di preprocessing sono:

1. **Rimozione dei record duplicati**
2. **Rimozione delle variabili non utili all'analisi**, che sono state eliminati secondo i seguenti criteri:
    - un rapporto di valori mancanti superiore al 50%
    - una varianza limitata (ad esempio, variabili i cui valori sono sempre costanti)

Come primo indicatore della cardinalità dell'informazione rimanente, abbiamo definito **il numero medio di campioni per veicolo al giorno** come misura sintetica e di alto livello. Di seguito, discutiamo i risultati dell'analisi esplorativa su questi dati.

V.2.2 Confronto dei dati multi-provider

Poiché ogni fornitore ha variabili diverse, è stata eseguita un'associazione semantica per identificare le feature correlate da diversi fornitori. Le feature risultanti con

le associazioni corrispondenti sono presentate nella Figura V.1, dove il diagramma di Venn mostra come i gruppi semantici sono condivisi tra i tre provider.

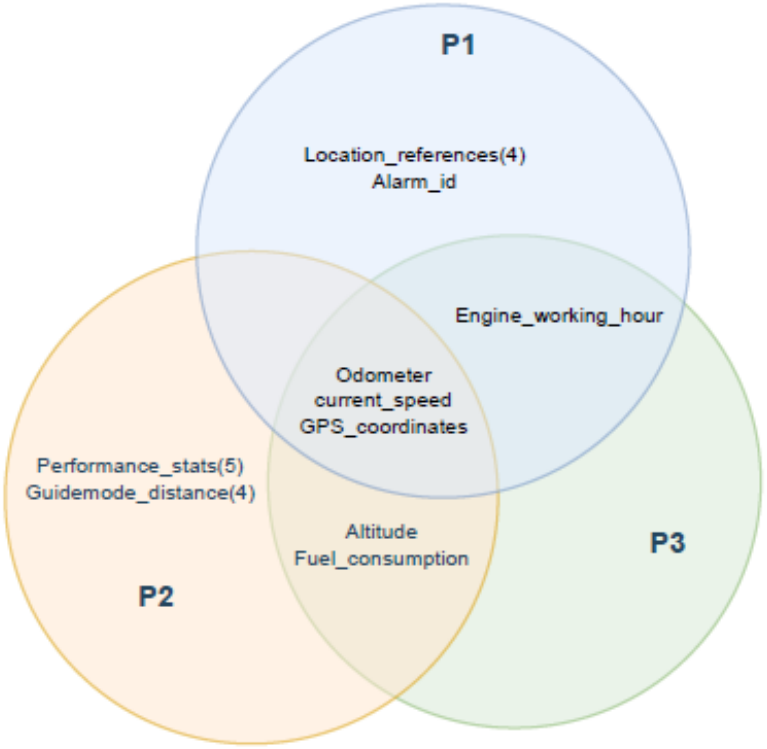

Fig. V.1: Diagramma di Venn dell'aggregazione semantica di alto livello delle caratteristiche di 3 fornitori di dati di tracciamento [111]

Le feature in comune a tutti i fornitori sono:

1. **Il contachilometri**
2. **La velocità**
3. **La posizione geografica dei veicoli**

Altre feature disponibili sono: il **tempo di funzionamento del motore** che è disponibile in 2 dei 3 fornitori.
L'odometro e il tempo di funzionamento del motore sono particolarmente interessanti per l'analisi poiché danno una misura dell'età" meccanica del veicolo, e nonostante

alcune possibili differenze sull'offset iniziale tra i diversi fornitori, si suppone che siano campionati in modo coerente tra tutti loro.
Anche altre variabili, come la velocità attuale, il consumo di carburante e le statistiche sulle prestazioni, sono potenzialmente significative per tracciare il degrado dei veicoli. Tuttavia, nonostante condividano la stessa semantica di alto livello, i loro valori sono così diversi che ogni fornitore traccia effettivamente una variabile per conto proprio, quindi non possono essere aggregate.

V.2.3 Analisi comparativa sulle caratteristiche monotoniche

Il contachilometri e il tempo di funzionamento del motore sono due variabili monotoniche condivise tra diversi fornitori di dati (P1 e P3). Entrambe le variabili sono costanti quando il veicolo è spento, mentre entrambe devono essere strettamente crescenti durante la guida.
I veicoli che condividono il contachilometri e il tempo di funzionamento del motore forniti da P1 e P3 sono pari a 10. In Figura V.2 riportiamo il confronto di tali variabili per 3 veicoli sovrapposti (ovvero **v13, v16, e v17**) nel periodo compreso tra agosto 2021 e novembre 2021, dopo una semplice conversione delle unità di misura (P1 utilizza chilometri e ore, mentre P3 metri e minuti).

P1 presenta un divario temporale nei campioni raccolti, di circa una settimana dal primo di settembre, per tutti i veicoli segnalati. Dato che l'ultima misura per ogni veicolo prima dello scarto e la prima dopo differiscono drasticamente, possiamo assumere che in questo periodo i veicoli erano ancora attivi. Questa supposizione è stata provata dalle misure della P3, poiché il fornitore stesso ha confermato che era associata a una mancanza di informazioni.
Inoltre, in questo periodo, come si evince dal grafico inferiore, vi è anche un drastico abbassamento del valore del tempo del motore. Questo può corrispondere a una sorta di reset interno dei sensori, che può aver fatto partire i valori registrati da una posizione di default, leggermente diversa da quella di qualsiasi veicolo. In questa situazione, una correzione di questo errore può essere ottenuta aggiungendo il primo

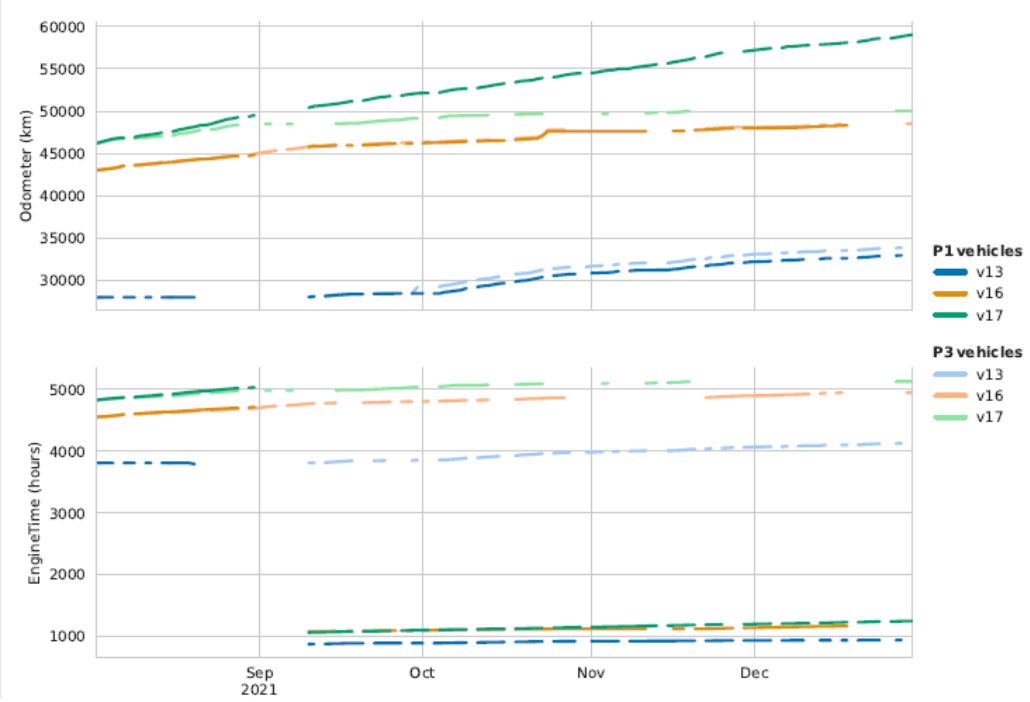

Fig. V.2: Confronto tra il contachilometri e il tempo di lavoro del motore per 3 veicoli monitorati da P1 e P3 [111]

valore di P3 del veicolo corrispondente a tutte le misure di P1. Con questa correzione, ogni coppia di curve dopo lo scarto diventa comparabile. Infine, il grafico superiore mostra anche un'altra anomalia riguardante il tracciamento dell'odometro di **v17**. Le tracce dei due fornitori hanno tendenze significativamente diverse che continuano a divergere all'aumentare del tempo. Il veicolo v17 ha viaggiato con una velocità media di 83 $km/giorno$ e 23 $km/giorno$, rispettivamente per P1 e P3. Tutte le distribuzioni di velocità media dei fornitori hanno un picco intorno ai 60 $km/giorno$, come mostrato in Fig. V.3, ma i valori medi calcolati sono sulla coda di entrambe le distribuzioni.

In questa situazione, come per altri veicoli monitorati da più di un fornitore di dati, si è giunti alla conclusione che le informazioni provenienti da diversi fornitori possono rivelare differenze significative nei valori riportati.

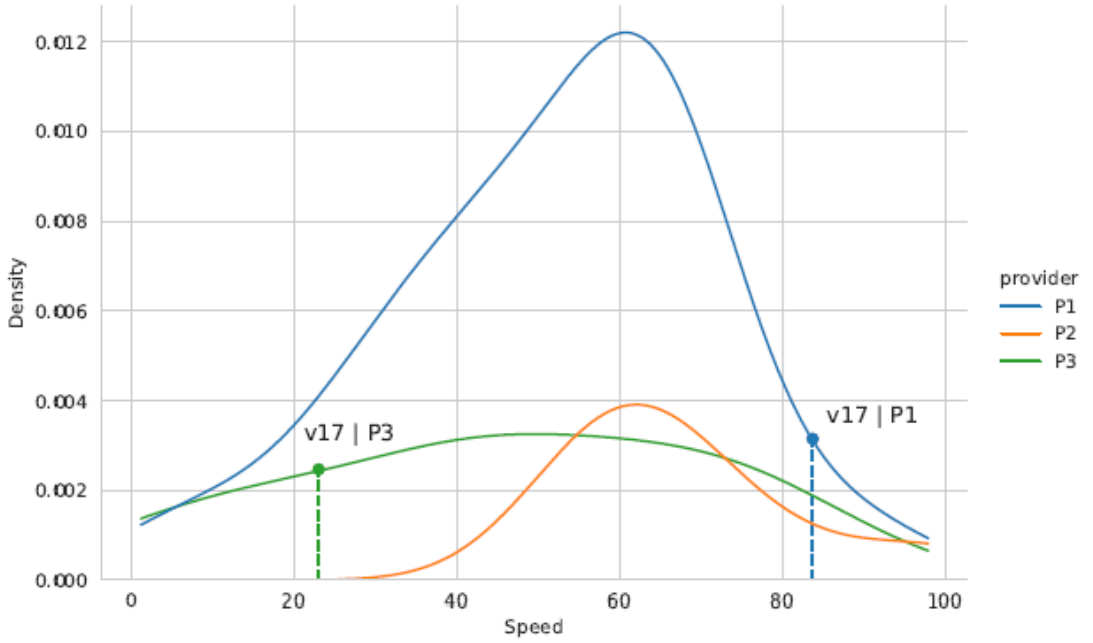

Fig. V.3: Distribuzione giornaliera della velocità per ogni provider. Due punti evidenziano la velocità media misurata per v17 [111]

V.2.4 Analisi nel tempo

Un'analisi importante da eseguire per comprendere efficacemente i dati raccolti è un'esplorazione del periodo di tempo in cui i veicoli sono monitorati. In quanto tempo e in che modo vengono raccolti i dati è, infatti, un'informazione utile per l'azienda che può evidenziare importanti differenze tra i vari fornitori. Di seguito, per ogni fornitore abbiamo riportato un grafico che rappresenta:

- Asse delle ascisse (X): il tempo di rilevazione
- Asse delle ordinate (Y): veicoli monitorati dai fornitori

In questo modo, per ogni veicolo possiamo facilmente analizzare la data di inizio e fine della raccolta ed eventuali anomalie nel periodo monitorato (ad esempio dati mancanti o cambiamenti nel tasso di dati giornalieri).

Fornitore P1

Il grafico della Figura V.4, mostra tutti i veicoli monitorati dal fornitore P1. Le diverse tonalità di colore dei punti rappresentano una diversa frequenza di campionamento in un dato giorno: **più scuro è il punto, più misurazioni sono state raccolte nello stesso giorno per quel veicolo.**
Questo grafico mostra che il numero di misurazioni giornaliere raccolte è quasi lo stesso per tutti i veicoli che sono stati monitorati durante il periodo da maggio 2021 a gennaio 2022. Le anomalie sono evidenziate dai rettangoli rossi nella figura. La prima anomalia che viene rilevata è comune a tutti i veicoli, ed è una mancanza di dati all'inizio del 2021/09 (settembre). Questo comportamento può essere spiegato da un periodo di chiusura dell'azienda, o da un fallimento generale nel processo di raccolta dati, che riguarda tutti i veicoli. Altre anomalie sono invece legate a veicoli specifici, e mostrano periodi piuttosto lunghi (anche di diversi mesi) di mancanza di dati. In generale, per capire il vero motivo della mancanza dei dati in queste situazioni, cioè se effettivamente il veicolo è stato fermo oppure si e verificato un problema nella raccolta dati, bisogna analizzare i valori cumulativi raccolti da variabili come il contachilometri. Se il valore di questa variabile nell'ultimo campione raccolto prima del periodo senza dati è uguale al valore del primo campione raccolto dopo il periodo senza dati, significa che il veicolo non ha effettivamente viaggiato in quel periodo, quindi nessuna informazione è stata persa.
Altrimenti, se i due valori sono diversi, significa che il veicolo si è mosso ma non abbiamo dati che hanno monitorato le sue operazioni durante quel periodo. Quindi c'è stato un errore nella fase di raccolta dei dati.

Fornitore P2

Un'analisi simile è stata eseguita per le misurazioni raccolte dal fornitore P2 e mostrata nella Figura V.5.
In questo caso, il numero di veicoli monitorati è molto più basso rispetto quelli di P1. Tuttavia, tutti i veicoli considerati da P2 sono monitorati anche da P1. Questo

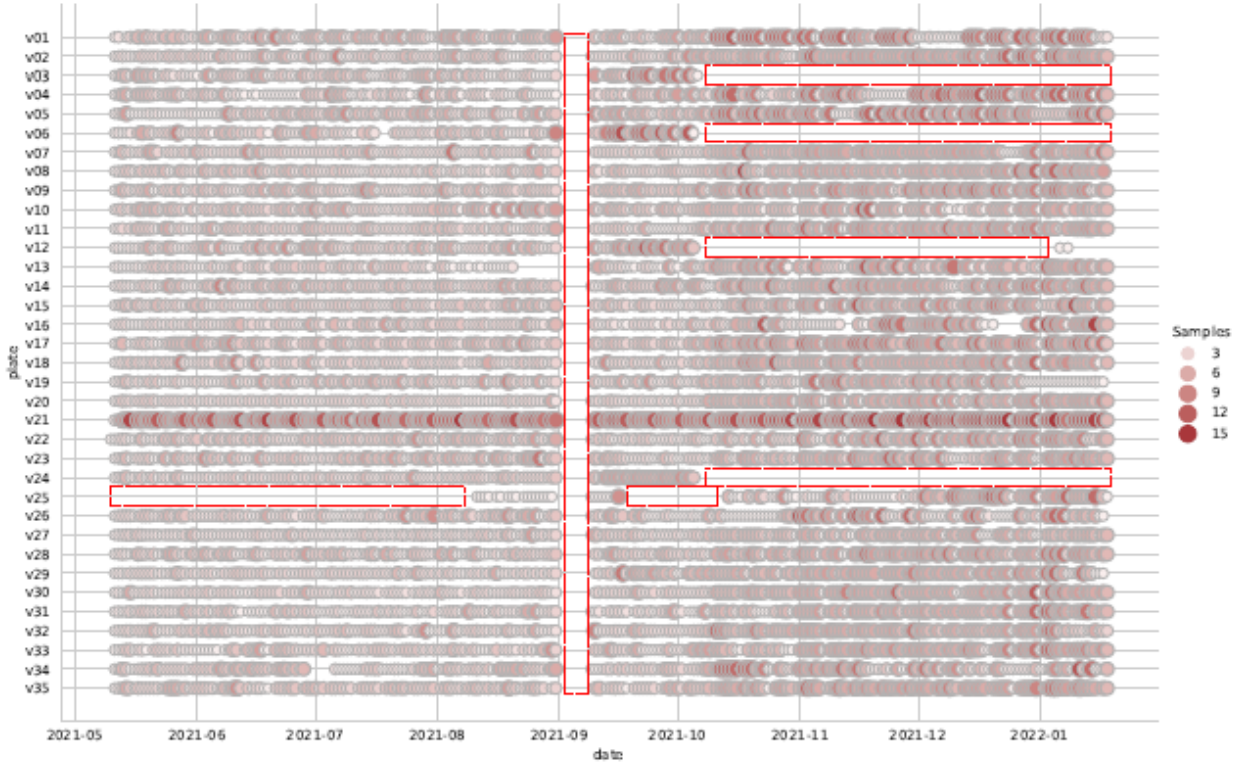

Fig. V.4: Analisi nel tempo per P1 [111]

è utile per fare un rapido controllo incrociato delle anomalie tra i due fornitori.

In particolare, possiamo facilmente notare che negli stessi periodi in cui mancavano i dati di P1 (sia il periodo comune a tutti i camion che quelli per i singoli camion), ora abbiamo le misure raccolte da P2. Questo ci porta a supporre che la precedente mancanza di dati fosse davvero dovuta a guasti legati a P1.

L'unico periodo in cui non abbiamo misurazioni in nessuno dei due fornitori è tra il 2021/12 e il 2022/01 per il veicolo v03. In questo caso, l'unico controllo che possiamo fare per verificare se il veicolo era davvero fermo è monitorare il valore del contachilometri prima e dopo la mancanza di dati.

Per quanto riguarda le misure raccolte da P2, si può notare un comportamento particolare nella frequenza di campionamento. Fino al 2021/12 c'è una frequenza di campionamento molto bassa, pari a circa una misurazione giornaliera per veicolo. Tuttavia, questo comportamento cambia improvvisamente negli ultimi mesi di raccolta, dove la frequenza di campionamento aumenta significativamente.

Questo sembra essere dovuto a un cambiamento nel modello di raccolta da dicembre in poi, probabilmente con l'obiettivo di raccogliere informazioni con un livello di dettaglio maggiore rispetto alle singole informazioni giornaliere raccolte in precedenza. L'uso dei veicoli non è cambiato da dicembre, ciò che è cambiato è la quantità di informazioni che descrivono il comportamento di un veicolo, con fino a 200 campioni al giorno.

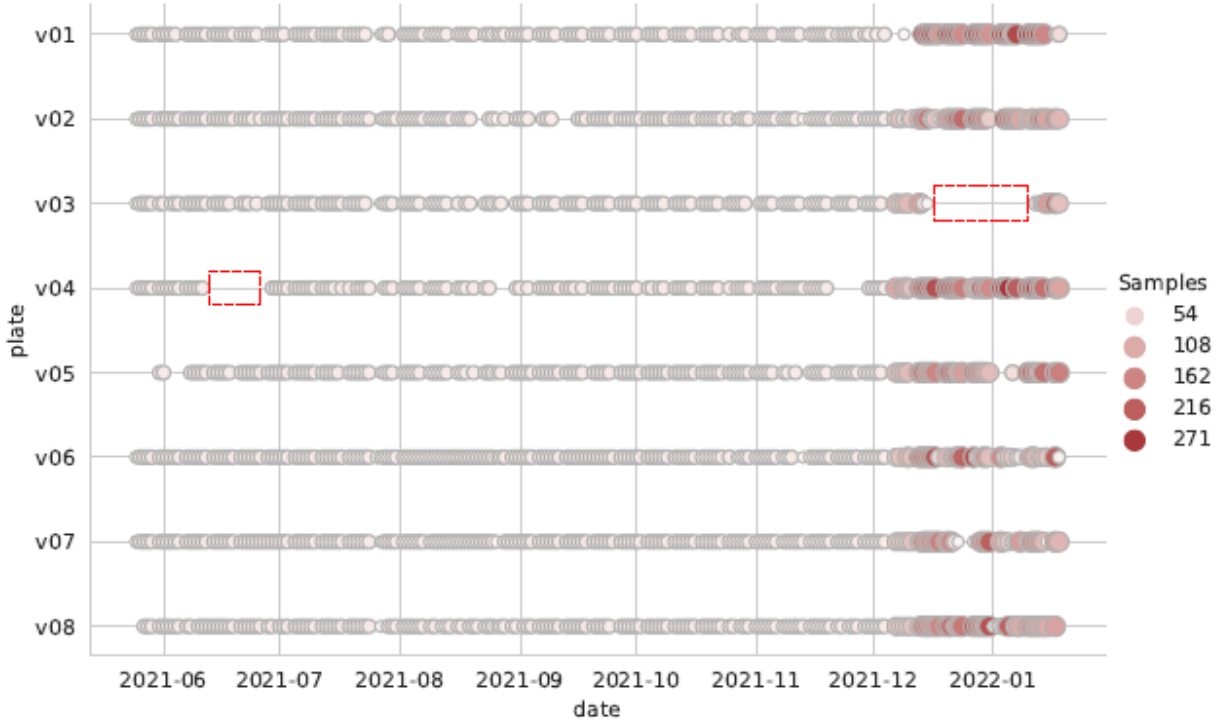

Fig. V.5: Analisi nel tempo per P2 [111]

Fornitore P3

La stessa analisi temporale è stata eseguita anche per i veicoli monitorati dal fornitore P3.

Cinque dei veicoli in Figura V.6 (da v36 a v40) sono monitorati solo da P3; quindi, è impossibile incrociarli con i dati degli altri provider per cercare di capire meglio le anomalie.

Per quanto riguarda, invece, i restanti veicoli, essi sono tutti monitorati da P1 o P2,

ciò permette di confrontare le misure raccolte. In particolare, le anomalie più evidenti nei dati raccolti da P3 sono per quei cinque veicoli dove non ci sono misurazioni nel periodo dal 2021/05 al 2021/10. Tuttavia, per gli stessi veicoli e negli stessi mesi, gli altri fornitori hanno raccolto misurazioni; quindi, si suppone che la mancanza di dati per la P3 sia un problema relativo solo a quel fornitore.

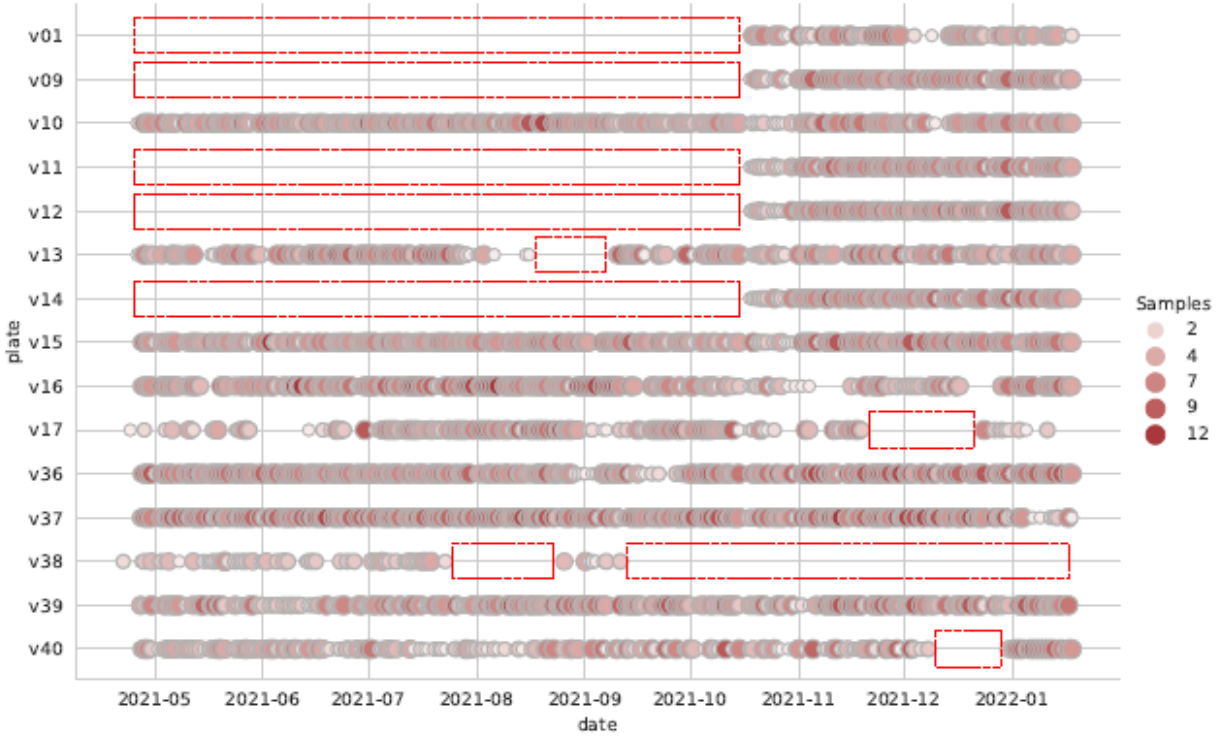

Fig. V.6: Analisi nel tempo per P3 [111]

V.3 Analisi delle Fatture di Manutenzione

Costruire un modello predittivo con accurency alta, è l'obiettivo principale; pertanto, sono state separate le etichette per categoria in modo tale da avere guasti omogenei.

Se facessimo un conteggio degli eventi che abbiamo per ogni veicolo, di fatto otteniamo un massimo di 4 eventi per ogni coppia targa-categoria, (Fig.V.7).

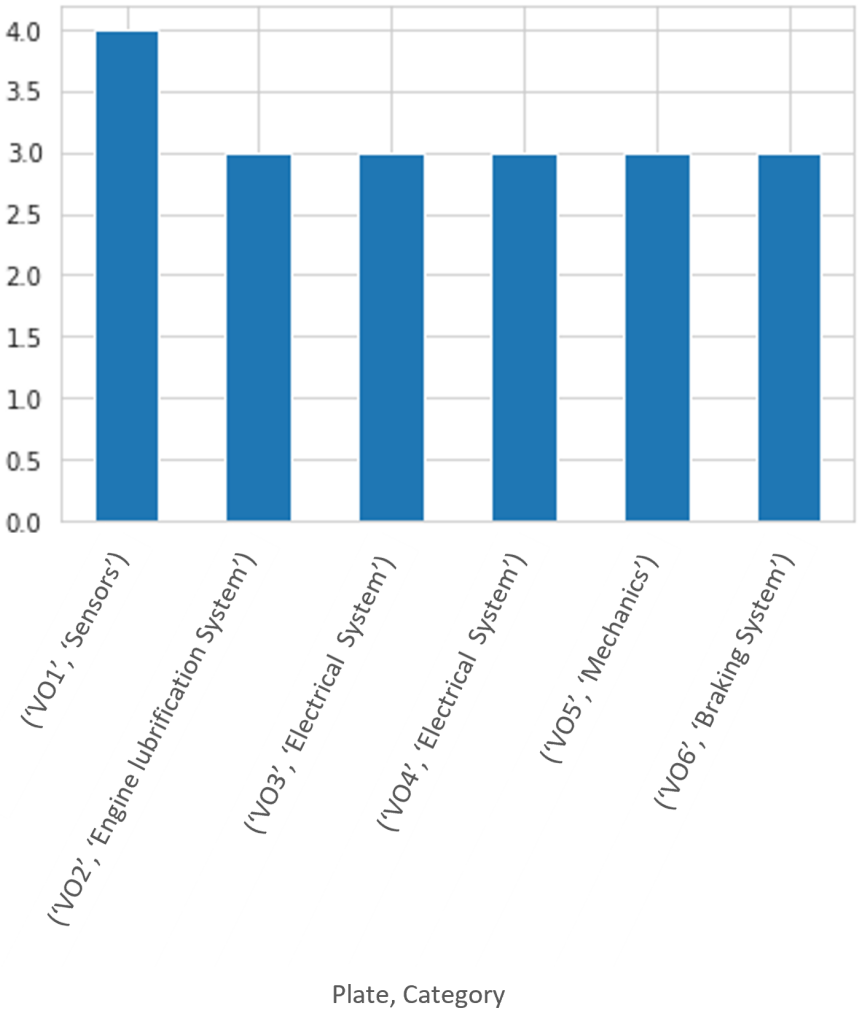

Fig. V.7: Numero massimo di eventi per ogni coppia targa-categoria

Gli interventi di manutenzione sono distribuiti in maniera abbastanza anomala nei dati a disposizione, cioè molti veicoli in poco tempo subiscono più interventi. Infatti, non tutte le coppie targa-categoria possono essere utilizzate poiché, come mostrato nella figura V.8, il veicolo V01 ha 4 interventi di manutenzione riguardanti la sensoristica ma risultano essere fatture molto ravvicinate.

Questo vuol dire che, se i dati risultanti vengono inseriti in una linea temporale, succede che quando si hanno diversi guasti molto ravvicinati, non possono essere utilizzati per costruire un modello predittivo perché il periodo in cui posso fare il training è disturbato dai suddetti guasti, come se ci fosse un'interferenza. Pertanto,

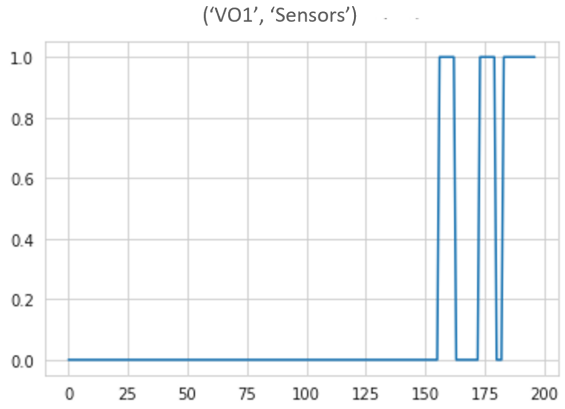

Fig. V.8: Frequenza dei guasti dell'impianto di sensoristica del V01

avere 4 guasti non è garanzia di qualità del training del modello predittivo.
In conclusione, interventi di manutenzione sulla stessa categoria che abbiano un periodo di tempo ravvicinato di 7 giorni (periodo che abbiamo considerato come threshold (soglia)), sono eliminati.

Tre etichette, invece, è un altro valore di riferimento per altri tipi di guasto e per altri mezzi (Figura V.9); quest'ultime coppie saranno utilizzate per allenare il modello predittivo.

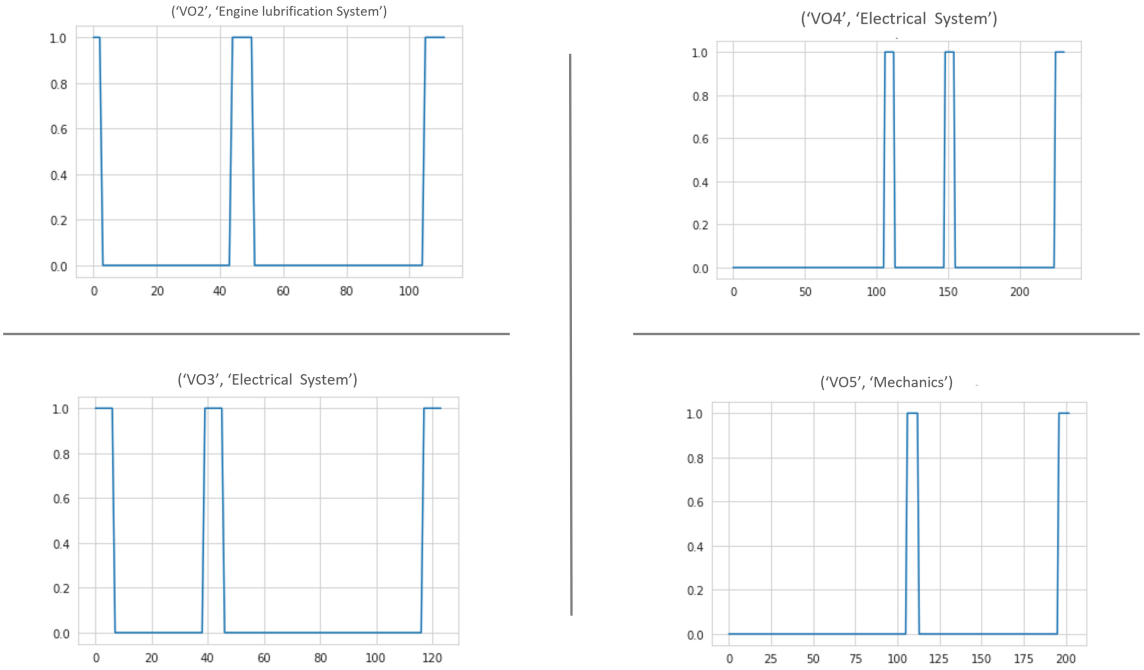

Fig. V.9: Frequenza di guasto di altri tipi di impianti dei veicoli V02, V03, V04, V05

V.4 Analisi delle Ipotesi di Correlazione

V.4.1 Algoritmo di Clustering

Gli interventi di manutenzione sono raccolti dal 2019 al 2022, calcoliamo per ciascun veicolo la differenza tra una fattura e l'altra (delta_data), cioè **Tempo che intercorre tra un guasto ed un altro**. Esempio riportato nella Tabella V.1:

Id	**Targa**	**Apertura_commessa**	**Delta_data**
46	V01	2019-11-24	-
48	V01	2019-11-30	6
132	V01	2020-07-20	226
328	V01	2021-03-10	240
438	V01	2021-07-10	122

Tab. V.1: Calcolo di delta_data del veicolo V01

L'approccio perseguito è data driven puro, infatti, per comprendere maggiormente i dati, calcoliamo le caratteristiche di un qualsiasi insieme di dati statistico:

- **Max**: massima distanza che intercorre tra due fatture
- **Min**: minima distanza che intercorre tra due fatture
- **Media**: la media dei delta_data
- **Std**: la variabilità di un insieme di dati dalla media

			Delta_data	
Targa	**min**	**max**	**mean**	**std**
V01	6.0	240.0	148.50	108.6
V02	1.0	167.0	59.75	53.87
V03	31.0	31.0	31.0	-
V04	31.0	122.0	66.16	44.0
V05	0.0	168.0	37.9	50.26

Tab. V.2: Calcolo delle statistiche di alcuni veicoli di esempio

La Tabella V.2 riporta i risultati statistici di alcuni veicoli. Notiamo che la **Std** risulta essere di notevole grandezza, a volte anche superiore alla Mean (veicolo

V05). Da ciò si deduce che il dataset sia sbilanciato a causa di veicoli che hanno eseguito interventi di manutenzione a distanza di pochi giorni oppure si può supporre che ci siano fatture mancanti e quindi, alcuni veicoli abbiano eseguito interventi di manutenzione a lunga distanza.

Feature in ingresso

- Numero massimo di interventi di manutenzione eseguiti su un tipo di componente per ogni veicolo
- Media di delta_data per ogni veicolo

Risultati dell'algoritmo di Clustering

Come già detto nel capitolo IV, l'algoritmo utilizzato è il K-means. Dopo aver eseguito l'algoritmo più volte, il numero K ottimale risulta essere pari a 4 (Figura V.10).

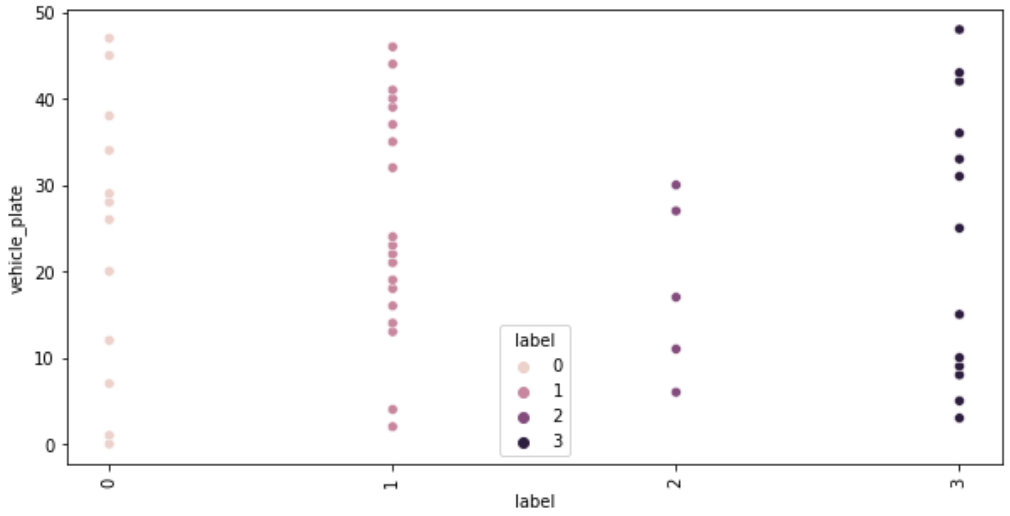

Fig. V.10: Gruppi di cluster per K=4

La tabella V.3, riassume il numero totale di veicoli appartenenti ad un cluster.

Associando ad ogni veicolo le proprie caratteristiche (TIPO, ALLESTIMENTO, CANTIERE, ALIMENTAZIONE, AUTOMEZZO), otteniamo un dataset di 49 sample (ovvero le targhe dei veicoli) e 6 feature (compresa la label).

CLUSTER			
CLUSTER_0	CLUSTER_1	CLUSTER_2	CLUSTER_3
12	19	5	13
TOTAL NUMBERS OF ITEMS: 49			

Tab. V.3: Numero di veicoli appartenenti a ciascun cluster

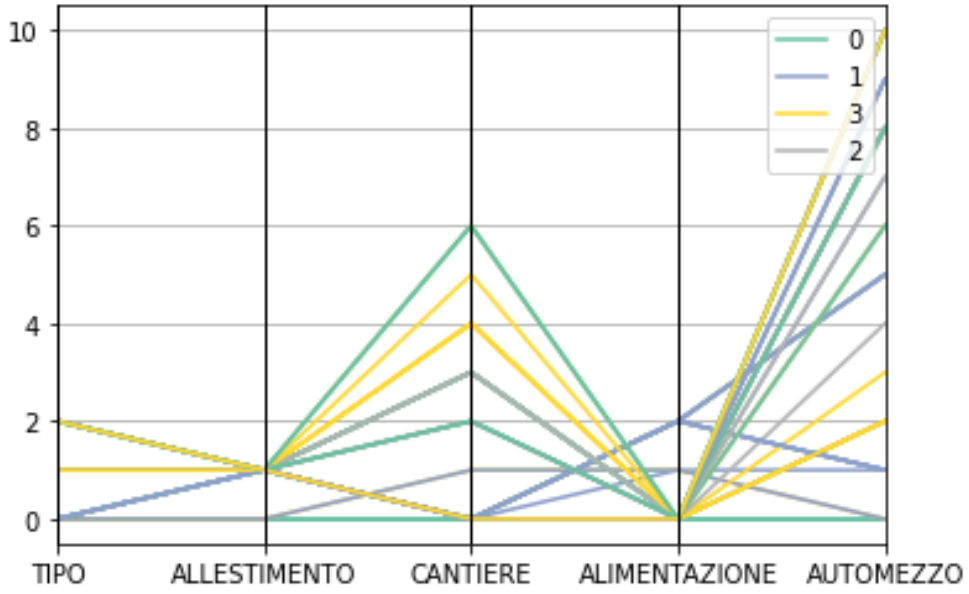

Fig. V.11: Rappresentazione delle correlazioni tra veicoli

Il parallel diagram V.11, è poco rappresentativo anche a causa dell'ordine di grandezza, ovvero del numero di sample che possediamo per ogni cluster, infatti, non risulta nessuna correlazione significativa.

V.4.2 Analisi dei Turni di Servizio

Analisi esplorativa del Calendario Servizi Fiorentino 2021

Il primo dataset analizzato appartiene al cantiere di Firenze.

Dal grafico V.12 risulta che:

- Dati raccolti dal 01-01-2021 al 01-01-2022
- 18 parametri totali monitorati dei turni
- 159 record (non ripetuti)

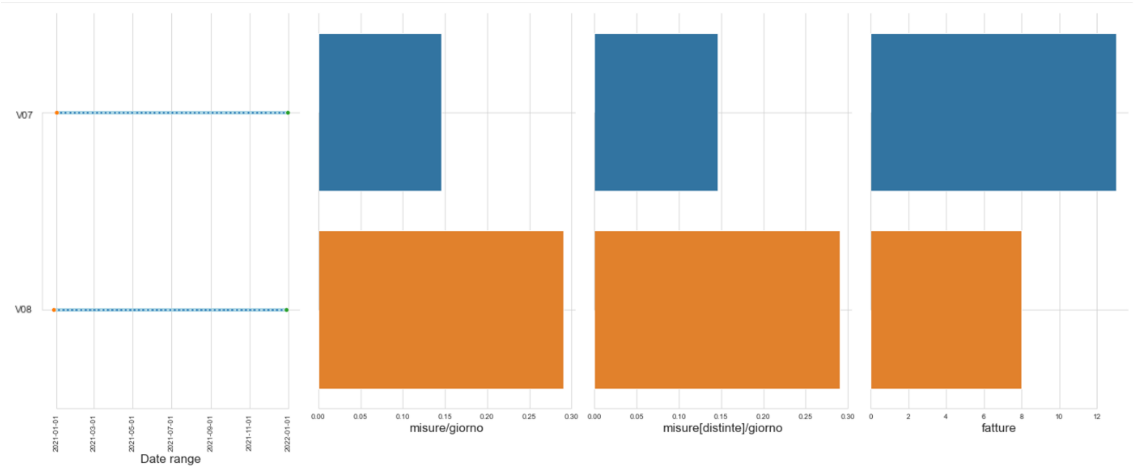

Fig. V.12: Overview del Calendario dei Servizi Fiorentino 2021

- 2 veicoli di interesse monitorati
- In media 0.22 $servizi/giorno/mezzo$ ($<< 1$)
- 33 fatture associate

Per comprendere efficacemente i Turni di Servizio dei veicoli in questione, è stata eseguita un'analisi esplorativa ridotta ad un periodo di tempo, ovvero un mese (Gennnaio 2021). Ciò viene evidenziato nel grafico V.13, che mostra una bassa attività lavorativa dei mezzi:

- 1 $sample/settimana$ per il veicolo V07
- 2 $sample/settimana$ per il veicolo V08

Per questo motivo, un'analisi proposta è la seguente:
La ricerca dei veicoli di interesse del cantiere Fiorentino (V07, V08) in altri cantieri, di altre regioni, per capire se effettivamente vi è un sottoutilizzo dei mezzi oppure quest'ultimi svolgono servizio altrove.

Ricerca dei veicoli di interesse presso altri cantieri

Prima di eseguire la ricerca dei veicoli, è opportuno scendere di livello di dettaglio dell'informazione, pertanto, la ricerca volge l'attenzione, non soltanto ai Cantieri ma

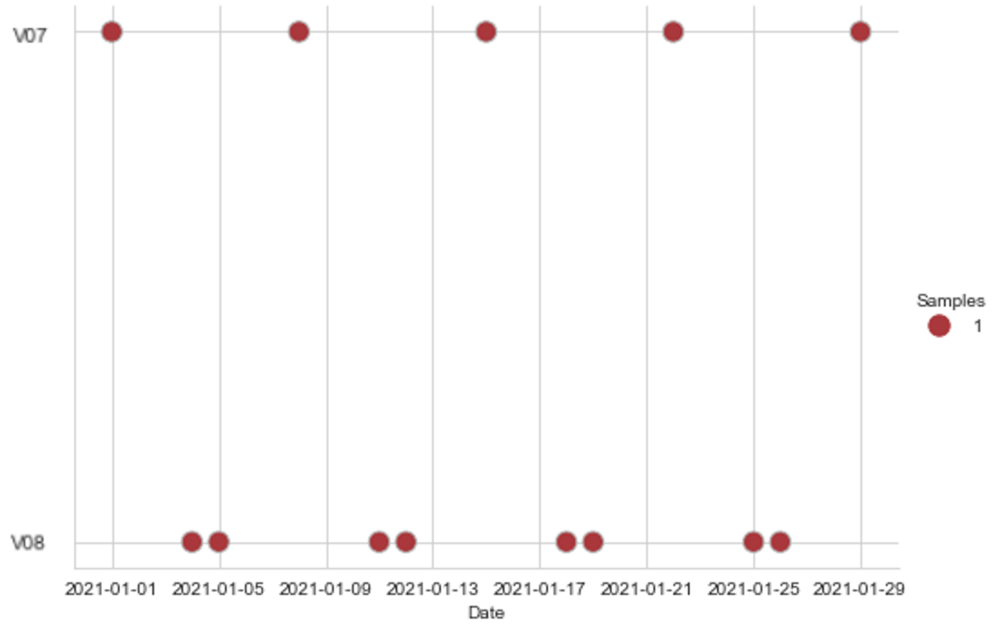

Fig. V.13: Calendario Servizi Fiorentino nel mese di Gennaio 2021

anche ai Comuni presso cui prestano servizio i veicoli (figura V.14).

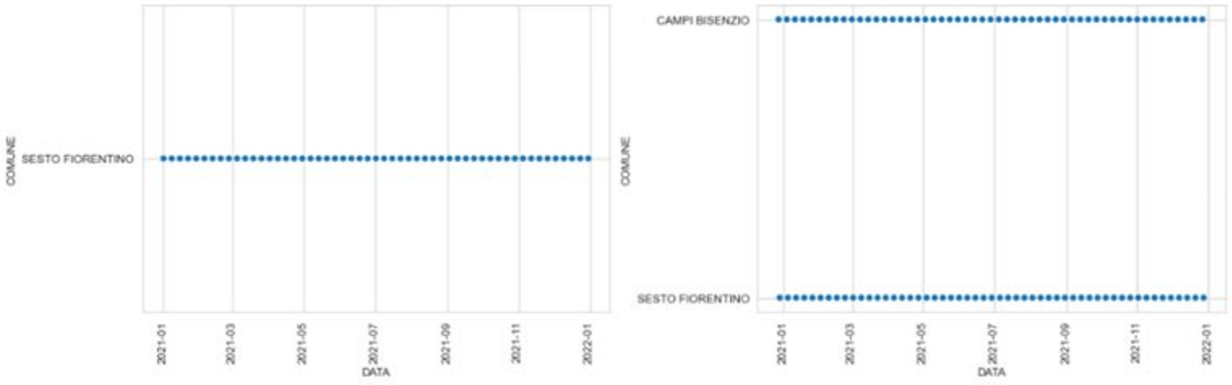

Fig. V.14: Comuni appartenenti al Calendario Servizi Fiorentino 2021 dei veicoli V07 (a sinistra) e V08 (a destra)

I grafici V.15) e V.16, mostrano le attività lavorative dei veicoli presso i vari Comuni. Il periodo di analisi racchiude un anno (Gennaio 2021-Gennaio 2022). Dall'analisi risulta che, sia V07 che V08, prestano un servizio costante in un cantiere (blu), ad eccezione di determinati periodi di tempo (altri colori). Questo comportamento può avere differenti interpretazioni:

- Chi raccoglie i dati non ha segnalato ogni giorno i servizi svolti dai veicoli in altri comuni, quindi, dati mancanti.
- Il veicolo è stato assegnato definitivamente in uno specifico cantiere e in casi eccezionali presta servizio presso altri cantieri.
- Errori di segnalazione dei Turni di Servizio dei veicoli

Per capire il vero motivo di queste anomalie nei dati, è stato eseguito un confronto con i dati raccolti dal sensore P1 che ricordiamo essere il sensore che si occupa del monitoraggio della mobilità di veicoli.

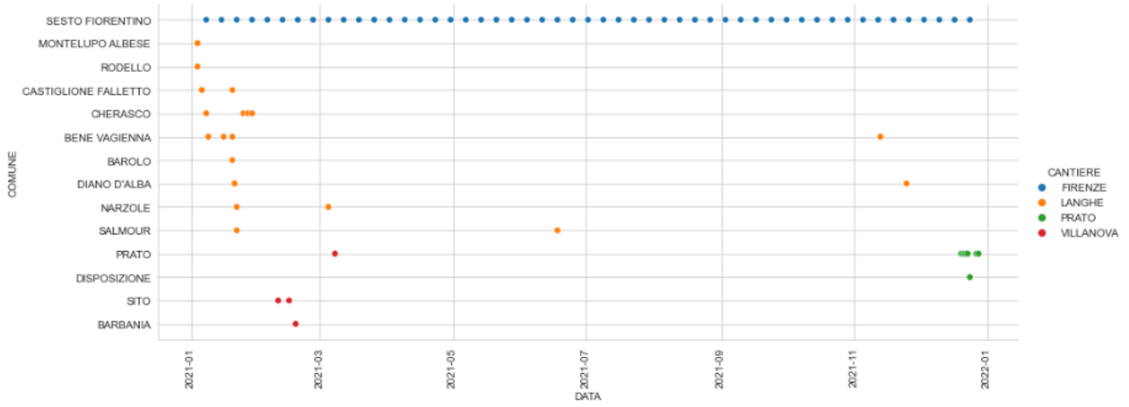

Fig. V.15: Raccolta eseguita dal veicolo V07 nell'anno 2021

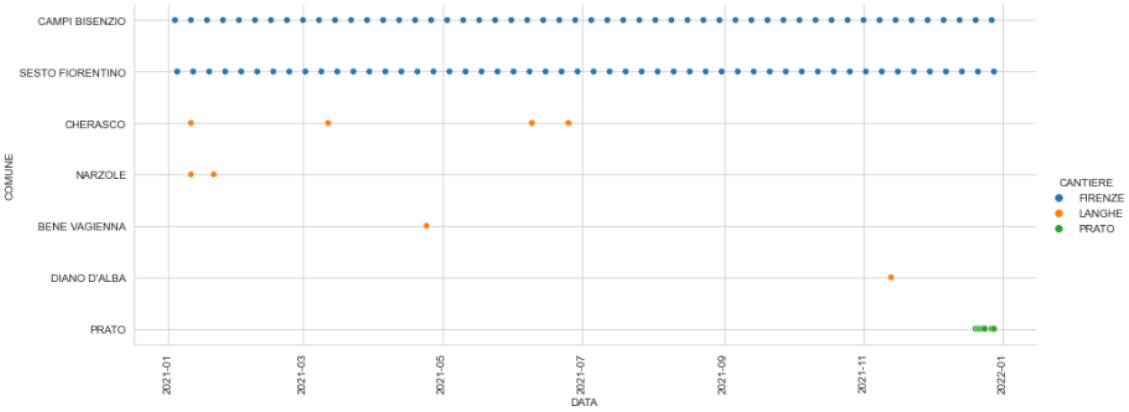

Fig. V.16: Raccolta eseguita dal veicolo V08 nell'anno 2021

Confronto con il fornitore P1

Per eseguire questa analisi, estraiamo un campione di dati che abbia come periodo temporale di monitoraggio (Novembre 2021-Gennaio 2022), questo perché coincide con il periodo in cui i veicoli lavorano su differenti cantieri e sia perché il periodo di raccolta dati del sensore P1 inizia da Maggio 2021

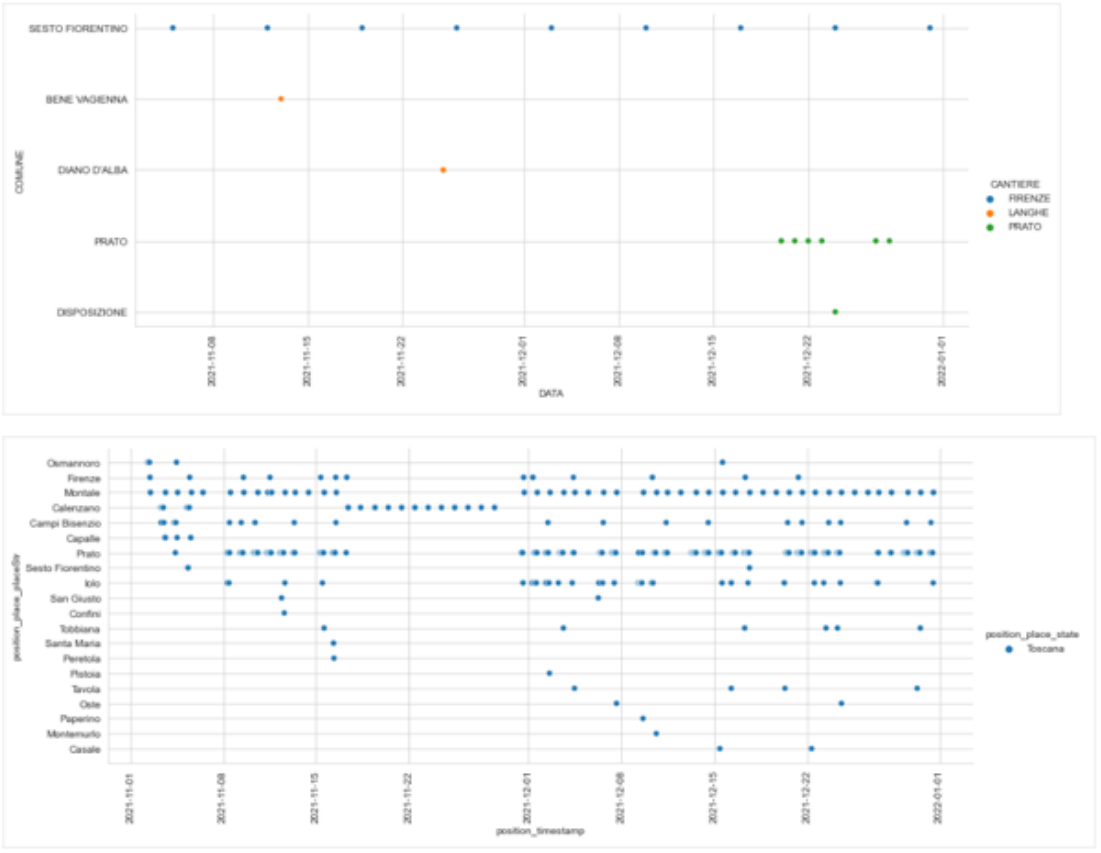

Fig. V.17: Confronto tra i dataset Turni di Servizio e fornitore P1 del veicolo V07

Il grafico V.17 mostra il confronto tra i dati raccolti nel Calendario dei Turni di Servizio 2021 (grafico sopra) e i dati di localizzazione raccolti dal sensore del fornitore P1 (grafico sotto). Confrontando le posizioni geografiche sembrerebbe che il veicolo abbia svolto la raccolta nel Toscano ma che sia stato localizzando in più comuni (non soltanto nei Comuni segnati nel Calendario dei Servizi 2021). Stesso ragionamento vale per il veicolo V08 (figura V.18).

In conclusione, possiamo dire che, per utilizzare i Turni di Servizio dei veicoli e

quindi raggiungere l'obietto di ricerca, si potrebbe pensare o di correggere le anomalie con l'ausilio dei dati di tracciamento del fornitore P1 oppure estrarre ulteriori informazioni da altre feature del dataset dei Turni dei Servizi 2021.

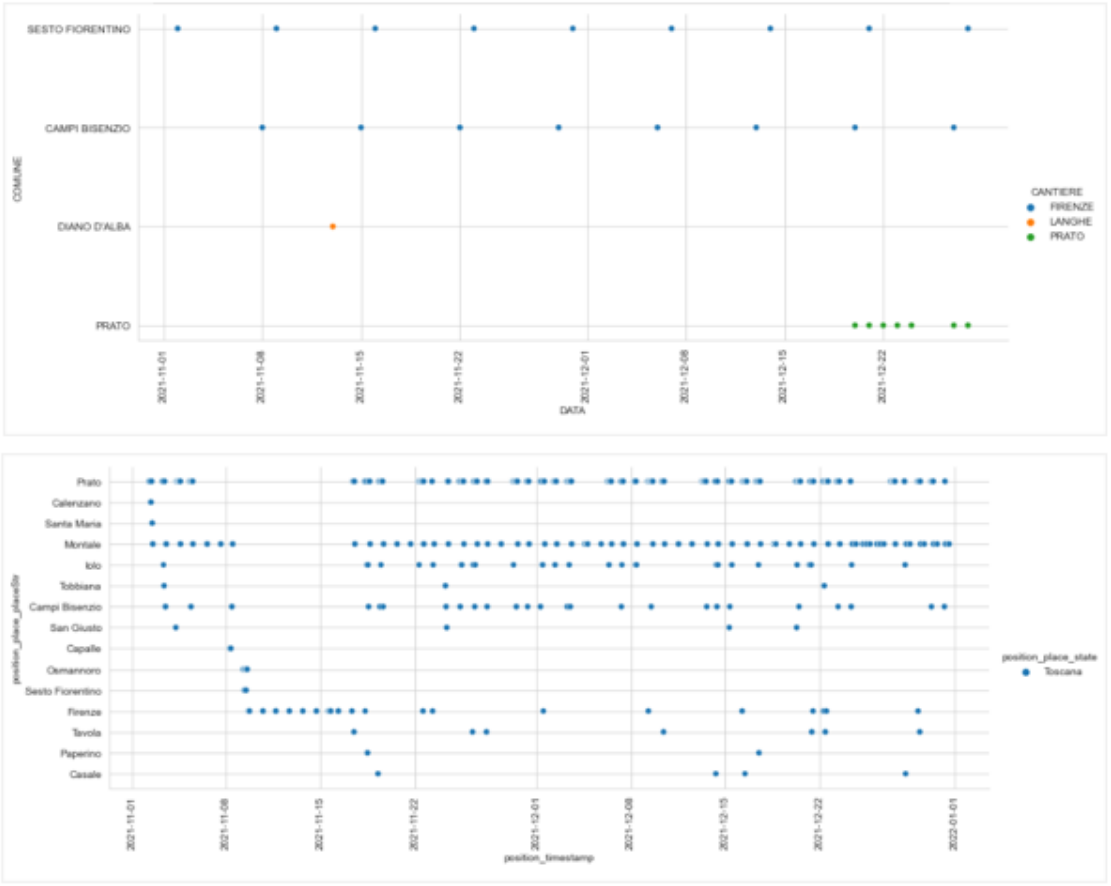

Fig. V.18: Confronto tra i dataset Turni di Servizio e fornitore P1 del veicolo V08

V.4.3 Analisi dei Dipendenti

In questa sezione, inizieremo con un'analisi esplorativa del dataset DipendentList per avere una maggiore conoscenza dei dati. In seguito, sarà eseguito un confronto con i dati dei fornitori che tengono traccia dell'id dell'operatore alla guida di un determinato veicolo, questo per trovare una correlazione tra lo stile di guida degli operatori e la frequenza di guasto dei veicoli.

Analisi esplorativa

Le feature di maggiore interesse sono **IdUtente**, **Data Assunzione** e **Data Cessazione**.

Ogni qualvolta, è settata la Data di Cessazione del contratto di un dipendente e successivamente, la Data di Assunzione è rinnovata, si genera un nuovo IdUtente (numero intero incrementale da 1 a 592).

Ciò vuol dire che per ogni nuovo contratto, è generato un nuovo IdUtente (esempio riportato nella Tabella V.4).

DipendentList					
IdUtente	**Codice Fiscale**	**Cognome**	**Nome**	**Data Assunzione**	**Data Cessazione**
137	AAAAAAAA	Rossi	Mario	02/07/2018	01/10/2018
138	AAAAAAAA	Rossi	Mario	05/11/2018	02/01/2019
139	AAAAAAAA	Rossi	Mario	27/07/2020	27/10/2020

Tab. V.4: Esempio della struttura dei dati di DipendentList

Confronto con i fornitori di dati

I fornitori con cui poter eseguire un confronto sono i seguenti:

1. **Fornitore P2**: è il fornitore di dati che si occupa del monitoraggio in tempo reale del comportamento e posizione dei mezzi.

   La feature di interesse in questo sensore è '**DriverIdentification**' che è un codice alfanumerico, differente dall'IdUtente (intero incrementale). Questo Id è controllato dal fornitore P2 stesso, sarebbe opportuno ed interessante capire la gestione di questo tipo di dato per verificare un eventuale correlazione con l'IdUtente di Dipendentlist. Il grafico V.19 mostra il count di DriverIdentification di P2. La maggior parte dei record sono etichettati come Unknown (191338) mentre i valori unici non nulli sono circa 10k.

2. **Fornitore P6**: è il fornitore di dati che si occupa della gestione del carburante.

   Le feature di interesse in questo sensore sono, oltre ad il '**driver_id**' (numero intero incrementale), '**last_name**' e '**first_name**'. Pertanto, è possibile fare

Fig. V.19: Count di DriverIdentification

un confronto anche con i dati personali degli operatori. Per trovare una correlazione tra driver_id e IdUtente e i dati personali dei dipendenti, eseguiamo il merge dei due dataset come mostrato in Tabella V.5.

driver_id	**last_name**	**first_name**	**IdUtente**	**Cognome**	**Nome**
330	Verdi	Giuseppe	0	NaN	NaN
334	Rossi	Giuseppe	224	Rossi	Mario
335	Bianchi	Gianluca	433	Bianchi	Gianluca

Tab. V.5: Merge tra i dataset P6 e DipendentList

Come si evince dall tabella V.5, purtroppo, il merge non ha riscontrato successo perché si verificano differenti anomalie:

- Alcuni valori risultano mancanti per gli attributi IdUtente, Cognome e Nome (riga 1).
- Alcuni valori di last_name risultano correlati al Cognome ma driver_id e first_name hanno valore differete rispettivamente per IdUtente e Nome (riga 2).
- Alcuni valori di i first_name e last_name risultano correlati ai dati personali di un operatore ma hanno differente id (riga 3).

3. **Fornitore P12**: è il fornitore di dati che si occupa della gestione dei rifornimenti.

   La feature di interesse in questo sensore è '**DriverCode**' che è un numero intero incrementale confrontabile con IdUtente ma dal grafico V.20, si nota che

abbia dei valori > di 592 mentre l'IdUtente è un numero incrementale (da 1 a 592) .

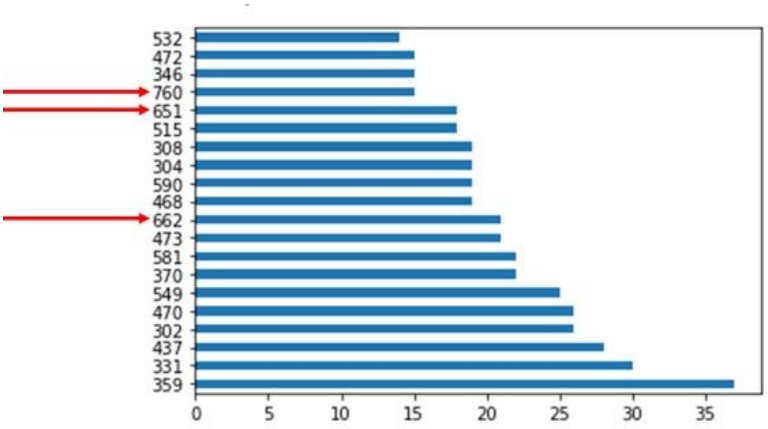

Fig. V.20: Count di DriverCode

Per trovare una correlazione tra gli DriverCode e IdUtente eseguiamo il merge dei due dataset come mostrato in Tabella V.6.

Dalla Tabella si evince che i dati personali dei dipendenti, risultano essere

DriverCode	Targa	IdUtente	Codice Fiscale	Cognome	Nome
581	V01	0	NaN	NaN	NaN
371	V02	371	AAAAAAAA	Rossi	Mario
909	V03	0	NaN	NaN	NaN

Tab. V.6: Merge tra i dataset P12 e DipendentList

etichettati come NaN; poiché, non esiste un IdUtente corrispondente.

V.5 Analisi Predittiva

V.5.1 LSTM

Il modello rincorrente è allenato su una time series unica con la coppia targa-categoria (sezione V.3), ciò vuol dire che è costruito su un singolo truck.
La variabile da predire, su una finestra temporale di 30 giorni, è il numero di Km che un veicolo può percorrere prima del prossimo guasto. Il modello dovrà predire,

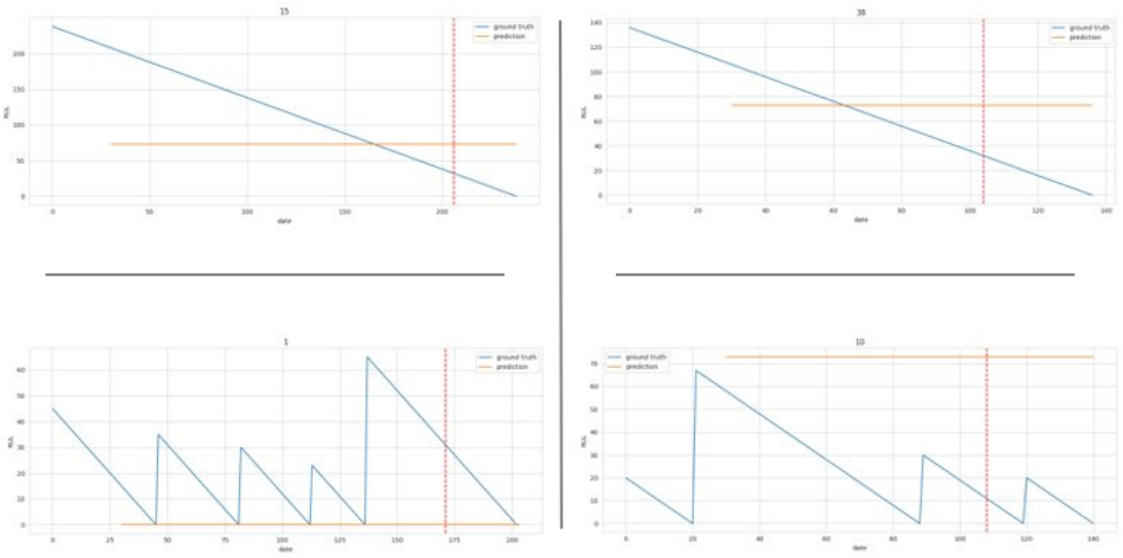

Fig. V.21: Risultati del modello LSTM

non il giorno esatto in cui avviene il guasto ma un periodo di 7 giorni antecedenti, in modo da poter prepararsi alla riparazione. Nella Tabella V.7 sono illustrati i parametri settati per allenare l'algoritmo.

	Experimented value
seg_len	30
hot_period	7
dt	10
Test set size	33 trucks
Train batch size	4

Tab. V.7: Parametri sperimentali settati per LSTM

Risultati sperimentali LSTM

Abbiamo analizzato il problema come un task di classificazione (etichetta '1' nei 7 giorni precedenti) o di regressione (RUL) in termini di giorni. Alcuni risultati sperimentali li possiamo analizzare nella figura V.21 che mostra da una parte la classificazione pura e la regressione dall'altra.

V.5.2 XGBoost

Dopo aver terminato i processi di feature engineering e feature selection (visti nella sezione IV.6), otteniamo un dataset in forma tabulare in cui ciascun record rappresenta un truck in un determinato giorno e il riassunto del suo comportamento sulla base delle variabili in input. Ciascun record lo etichettiamo a:

- **0** se è un comportamento senza failure
- **1** se temporalmente questo record precede l'informazione temporale presente nella fattura relativa a quel determinato truck

Come visto nell'approccio precedente (sezione V.5.1), il task da eseguire è ma la variabile da predire è il numero di Km rimanenti che un truck può percorrere prima del prossimo guasto. Mentre la fattura da predire è quella associata all'impianto di lubrificazione motore.
Il periodo temporale utilizzato (ovvero hotperiod) precedente alla fattura che ci permette di verificare, non il guasto puntuale in quel determinato giorno ma ad N giorni prima, è settato a 10 giorni.

Risultati sperimentali XGBoost

Dopo aver etichettato il dataset, gli esperimenti riportati sono 3:

1. Il modello è allenato e testato, considerando tutti i dati a disposizione
2. Il modello è allenato e testato su un singolo truck, come mostarto nella Fig. V.22
3. Il modello è allenato su n-1 truck e testato su quello rimanente, come mostarto nella Fig. V.23

Abbiamo riportato la distribuzione dell'etichette che, come si deduce dal dataset, è molto sbilanciata poiché i failure sono minori del periodo in cui il truck funziona correttamente. Nella tabella V.8 sono riportati i risultati relativi ai 3 esperimenti.

Per ciascuno di essi è riportato:

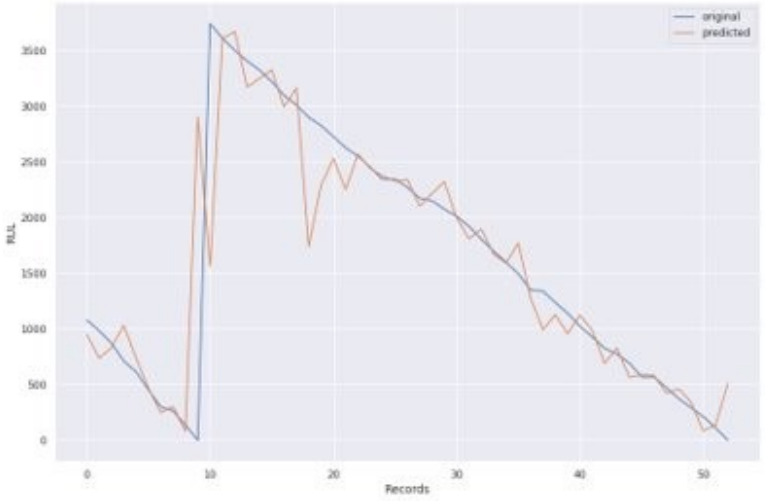

Fig. V.22: Modello generato su un singolo truck (V02)

Esperimenti	**F1-score class 0 (failure)**	**F1-score class 1 (failure)**
Tutti i dati a disposizione sono utilizzati per allenare/testare il modello	0.98	0.68
Il modello è allenato e testato su un singolo truck	0.97	0.74
Il modello è allenato su n-1 truck e testato su quello rimanente	0.95	**0.0**

Tab. V.8: Risultati sperimentali dell'algoritmo XGBoost

- f1-score relativo alla **classe 0** maggioritaria
- f1-score relativo alla **classe 1** minoritaria

Si volge l'attenzione alla seconda colonna (classe 1) perché ci dice quanto effettivamente il modello sia in grado di predire un guasto ma i valori risultano essere molto bassi in tutti e tre gli esperimenti. Un po' più alto è quello associato al secondo esperimento perché il modello è allenato e testato su un singolo truck, tuttavia, questo valore potrebbe essere frutto di regole fittizie che il modello ha imparato: Questo perché se, ad esempio, non avessimo aggregato l'odometro, ma avessimo la-

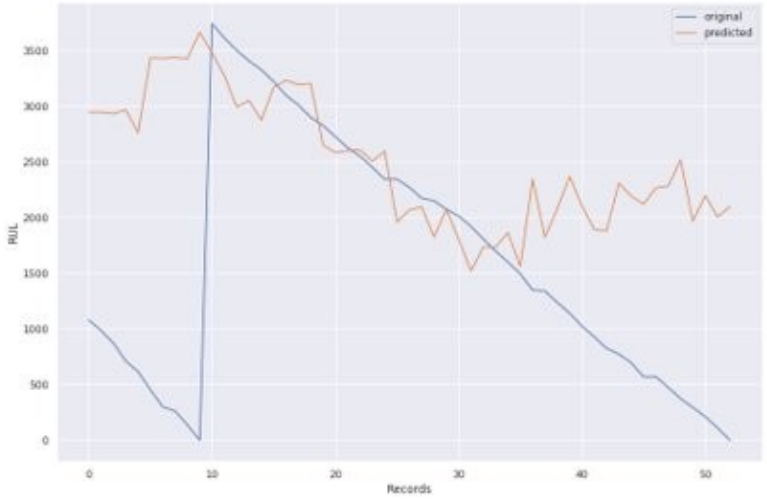

Fig. V.23: Modello allenato su n-1 trucks (V07, V05, V14) e testato su un truck (V02)

sciato il suo valore puntuale, il modello riconoscendo che il guasto avviene ad un determinato valore di odometer, etichetta ad 1 valori superiori al valore puntuale di odometer mentre a 0 valori più bassi di quest'ultimo. Questa sarebbe una custom rule per un singolo truck e non generalizzabile.

Conclusioni e Sviluppi Futuri

Questo lavoro di tesi ha permesso lo studio approfondito di diverse tecniche di analisi dei Big Data per la costruzione di un modello di manutenzione predittiva in ambito veicolare.
Il progetto è stato proposto da un'azienda italiana: SEA (Soluzioni Eco Ambientali) che si occupa della raccolta di rifiuti, per mezzo di veicoli, monitorati da differenti sensori installati sul telaio e sull'allestimento.
L'obiettivo dell'azienda è quello di attivare un piano di manutenzione preventiva dei guasti sui veicoli, al fine di contenere i rischi, riducendo al minimo i costi e aumentando i profitti, ma anche di ridurre l'impatto ambientale. La flotta di veicoli analizzata si compone di 58 veicoli di differenti tipologie, monitorati da diversi fornitori di dati.
Uno degli aspetti più rilevanti di questa ricerca, è stata l'analisi esplorativa e caratterizzazione preliminare dei dati della flotta. Un'accurata esplorazione dei dati è utile per migliorare la qualità di previsione dei guasti del modello.
Sono stati analizzati e confrontati i dati raccolti da 3 provider, su 40 veicoli, nel corso di 9 mesi, evidenziando problemi e sfide; poiché, informazioni anomale dell'attività del veicolo, come quelle riportate in questo lavoro, possono portare a caratteristiche di input inaffidabili e di conseguenza a una previsione incerta.
La presenza di informazioni multi-provider per lo stesso veicolo (anche se più costoso), ha permesso di rimediare al frequente problema dei dati mancanti.
Un altro concetto da non sottovalutare, al fine di etichettare in maniera ottimale i dati, è l'inferenza delle caratteristiche di contesto sul guasto di un veicolo. Di fatto, cercare una funzione che vada a mappare il degrado nel tempo del veicolo, può si-

curamente portare vantaggi di precisione nella costruzione di un modello predittivo. Pertanto, parte del lavoro è stato dedicato all'analisi dei Turni di Servizio, che danno informazioni sull'attività lavorativa del veicolo, ma anche all'analisi dei Dipendenti, poiché lo stile di guida di ciascuno di essi influisce sul deterioramento delle componenti meccaniche. Dalle analisi sono emerse problematiche legate, anche in questo caso, alla raccolta dei dati. Principalmente, dovute alla frequenza di dati mancanti. Dopo un'attenta ricerca su tutte le possibili tematiche da affrontare e analisi da eseguire, per raggiungimento dell'obiettivo, sono stati estesi due modelli di manutenzione predittiva. I due algoritmi non hanno prodotto risultati significativi in termini di accuratezza e precisione, a causa dell'insufficienza delle informazioni raccolte.

Rimane comunque l'impostazione del lavoro che potrà essere utilizzato per uno sviluppo futuro, che includerà ulteriori analisi promettenti, quali:

- **Analisi dei rifornimenti**, utile per stimare i consumi di carburante in rapporto ai Km percorsi da ciascun veicolo.
- **Ipotesi di correlazione tra Km percorsi e coordinate geografiche dell'officina** per comprovare la data del giorno dell'intervento di manutenzione, indicata nelle fatture.

Grazie ai risultati sperimentali, ottenuti dalle analisi eseguite in questi mesi, l'azienda sta svolgendo differenti cambiamenti:

1. Miglioramento dell'estrazione dei dati raccolti dai sensori.
2. Aumento del valore e della qualità del prodotto acquistato, esponendo ai fornitori di dati diverse problematiche legate alla rilevazione di dati dai sensori.
3. Maggiore accuratezza semantica nell'etichettatura del tipo di guasto di manutenzione.
4. Aumento degli investimenti in tecnologia. Ad esempio, l'impiego di piattaforme professionali che possano tenere traccia di informazioni importanti come i Turni di Servizio e dei Dipendenti.

Tutti questi cambiamenti positivi adottati dall'azienda saranno utili al conseguimento del risultato e questo progetto di tesi ne fornisce una solida base.

Bibliografia

[1] Ilenia.gemito@gmail.com, "Industria 4.0: Cos'é, tecnologie e incentivi a vantaggio del business," Oct 2021.

[2] K. Schwab, *The Fourth Industrial Revolution.* Portfolio Penguin, 2017.

[3] "Industria 4.0: La prima rivoluzione industriale del iii millennio - workgroup," Mar 2019.

[4] K. Schwab, "The fourth industrial revolution: What it means and how to respond," Jan 2016.

[5] Simonapolitini, "Sempre più industria 4.0 in italia: Gli ultimi dati dell'osservatorio del polimi," Oct 2021.

[6] "2020 deloitte and mapi smart manufacturing ecosystems study - press release. deloitte united states," Oct 2020. Retrieved March 17, 2022.

[7] "Preparing for our future: Developing a common strategy for key enabling technologies in the eu - current situation of key enabling technologies in europe. (n.d.).," Sep 2009. Retrieved March 17, 2022.

[8] Mecalux, "Tecnologie abilitanti: Cosa sono e come migliorano la logistica."

[9] "Tecnologie abilitanti industria 4.0: Definizione, benefici e rischi. make..," April 2021. Retrieved March 17, 2022.

[10] "Realtá aumentata, cos'é? - strumenti, esempi e sviluppi futuri. intelligenza artificiale. (n.d.).." Retrieved March 17, 2022.

[11] Moore, "Software reviews, businessmaster ii+, ledger for cp/m systems," *InfoWorld, InfoWorld Media Group*, vol. 20, p. 31, Dec. 1982.

[12] C. Ridolfi, "Cos'é l'internet of things? tutto quello che c'é da sapere sull'internet delle cose," Apr 2021.

[13] "https://www.axelero.it/storia-internet-of-things/." Retrieved March 17, 2022.

[14] "https://www.robotiko.it/linternet-delle-cos-e-definizione-esempi/."

[15] "il piano nazionale impresa 4.0 e gli altri strumenti a supporto dell'innovazione digitale delle imprese," Mar 2018.

[16] P. Mell, T. Grance, *et al.*, "The nist definition of cloud computing," 2011.

[17] ISACA, "Csx cybersecurity fundamentals study guide," vol. 190, p. 5, 2015.

[18] A. De Mauro, M. Greco, and M. Grimaldi, "A formal definition of big data based on its essential features," *Library Review*, 2016.

[19] "Big data, cosa sono? - utilizzo, gestione e privacy.." Retrieved March 17, 2022.

[20] A. Sardi, E. Sorano, V. Cantino, and P. Garengo, "Big data and performance measurement research: trends, evolution and future opportunities," *Measuring Business Excellence*, 2020.

[21] "Big data, cosa sono? - utilizzo, gestione e privacy." Retrieved March 17, 2022.

[22] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM J. Res. Dev.*, vol. 3, pp. 210–229, 1959.

[23] T. M. Mitchell and M. Learning, "Mcgraw-hill," *New York*, pp. 154–200, 1997.

[24] E. Alpaydin, "Introduction to machine learning, 3rd editio. ed," 2014.

[25] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4. Springer, 2006.

[26] B. Apolloni, A. Ghosh, F. Alpaslan, and S. Patnaik, *Machine learning and robot perception*, vol. 7. Springer Science & Business Media, 2005.

[27] S.-I. Ao, B. B. Rieger, and M. Amouzegar, *Machine learning and systems engineering*, vol. 68. Springer Science & Business Media, 2010.

[28] G. László, O. Györgi, and H. Walk, *Machine Learning for Financial Engineering*. Imperial College Press, 2012.

[29] X. W. Gong Y, "Machine learning for multimedia content analysis."

[30] J. Yu and D. Tao, *Modern machine learning techniques and their applications in cartoon animation research*, vol. 4. John Wiley & Sons, 2013.

[31] A. Fielding, *Machine learning methods for ecological applications*. Springer Science & Business Media, 1999.

[32] S. Mitra, S. Datta, T. Perkins, and G. Michailidis, *Introduction to machine learning and bioinformatics*. CRC Press, 2008.

[33] Z. R. Yang, *Machine learning approaches to bioinformatics*, vol. 4. World scientific, 2010.

[34] T. J. Cleophas and A. H. Zwinderman, "Machine learning in medicine," in *Springer Netherlands*, 2013.

[35] J. D. Malley, K. G. Malley, and S. Pajevic, *Statistical learning for biomedical data*. Cambridge University Press, 2011.

[36] G. Ifrah, *The Universal History of Computing: From the abacus to the Quantum Computer*. Wiley, 2001.

[37] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain.," *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[38] P. Werbos and P. John, "Beyond regression : new tools for prediction and analysis in the behavioral sciences /," 01 1974.

[39] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[40] "Support-vector networks," 1995.

[41] G. P. Shapiro and W. Frawley, "Knowledge discovery in databases," *AAAI/MIT Press, Cambridge, MA*, p. 39, 1991.

[42] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[43] G. E. Hinton, "Learning multiple layers of representation," *Trends in cognitive sciences*, vol. 11, no. 10, pp. 428–434, 2007.

[44] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[45] V. Cherkassky and F. M. Mulier, *Learning from data: concepts, theory, and methods*. John Wiley & Sons, 2007.

[46] "Kdd process in data mining - javatpoint."

[47] R. O. Duda, P. E. Hart, *et al.*, *Pattern classification and scene analysis*, vol. 3. Wiley New York, 1973.

[48] D. Heckerman, "Bayesian networks for data mining," *Data mining and knowledge discovery*, vol. 1, no. 1, pp. 79–119, 1997.

[49] A. Shamim, H. Hussain, and M. U. Shaikh, "A framework for generation of rules from decision tree and decision table," in *2010 International Conference on Information and Emerging Technologies*, pp. 1–6, IEEE, 2010.

[50] A. Suresh, R. Udendhran, and M. Balamurgan, "Hybridized neural network and decision tree based classifier for prognostic decision making in breast cancers," *Soft Computing*, vol. 24, no. 11, pp. 7947–7953, 2020.

[51] Priyanka and D. Kumar, "Decision tree classifier: A detailed survey," *International Journal of Information and Decision Sciences*, vol. 12, no. 3, pp. 246–269, 2020.

[52] A. S. Eesa, A. M. Abdulazeez, and Z. Orman, "A dids based on the combination of cuttlefish algorithm and decision tree," *Science Journal of University of Zakho*, vol. 5, no. 4, pp. 313–318, 2017.

[53] F.-J. Yang, "An extended idea about decision trees," in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 349–354, IEEE, 2019.

[54] J. Liang, Z. Qin, S. Xiao, L. Ou, and X. Lin, "Efficient and secure decision tree classification for cloud-assisted online diagnosis services," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 4, pp. 1632–1644, 2019.

[55] A. Brifcani and A. Issa, "Intrusion detection and attack classifier based on three techniques: a comparative study," *Eng. & Tech. Journal*, vol. 29, no. 2, pp. 368–412, 2011.

[56] A. S. Eesa, Z. Orman, and A. M. A. Brifcani, "A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems," *Expert systems with applications*, vol. 42, no. 5, pp. 2670–2679, 2015.

[57] P. H. Swain and H. Hauska, "The decision tree classifier: Design and potential," *IEEE Transactions on Geoscience Electronics*, vol. 15, no. 3, pp. 142–147, 1977.

[58] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR).[Internet]*, vol. 9, pp. 381–386, 2020.

[59] J. Mrva, Š. Neupauer, L. Hudec, J. Ševcech, and P. Kapec, "Decision support in medical data using 3d decision tree visualisation," in *2019 E-Health and Bioengineering Conference (EHB)*, pp. 1–4, IEEE, 2019.

[60] C. Z. Janikow, "Fuzzy decision trees: issues and methods," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 28, no. 1, pp. 1–14, 1998.

[61] G. Stein, B. Chen, A. S. Wu, and K. A. Hua, "Decision tree classifier for network intrusion detection with ga-based feature selection," in *Proceedings of the 43rd annual Southeast regional conference-Volume 2*, pp. 136–141, 2005.

[62] I. S. Damanik, A. P. Windarto, A. Wanto, S. R. Andani, W. Saputra, *et al.*, "Decision tree optimization in c4. 5 algorithm using genetic algorithm," in *Journal of Physics: Conference Series*, vol. 1255, p. 012012, IOP Publishing, 2019.

[63] R. C. Barros, M. P. Basgalupp, A. C. De Carvalho, and A. A. Freitas, "A survey of evolutionary algorithms for decision-tree induction," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 3, pp. 291–312, 2011.

[64] G. Gupta *et al.*, "A self explanatory review of decision tree classifiers," in *International conference on recent advances and innovations in engineering (ICRAIE-2014)*, pp. 1–7, IEEE, 2014.

[65] J. Chen, K. Li, Z. Tang, K. Bilal, S. Yu, C. Weng, and K. Li, "A parallel random forest algorithm for big data in a spark cloud computing environment," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 4, pp. 919–933, 2016.

[66] J. Chen, K. Li, Z. Tang, K. Bilal, S. Yu, C. Weng, and K. Li, "A parallel random forest algorithm for big data in a spark cloud computing environment," *IEEE*

Transactions on Parallel and Distributed Systems, vol. 28, no. 4, pp. 919–933, 2016.

[67] Y. Singh and A. S. Chauhan, "Neural networks in data mining.," *Journal of Theoretical & Applied Information Technology*, vol. 5, no. 1, 2009.

[68] A. Mosallam, K. Medjaher, and N. Zerhouni, "Data-driven prognostic method based on bayesian approaches for direct remaining useful life prediction," *Journal of Intelligent Manufacturing*, vol. 27, no. 5, pp. 1037–1048, 2016.

[69] R. Yam, P. Tse, L. Li, and P. Tu, "Intelligent predictive decision support system for condition-based maintenance," *The International Journal of Advanced Manufacturing Technology*, vol. 17, no. 5, pp. 383–391, 2001.

[70] M. Mesgarpour, D. Landa-Silva, and I. Dickinson, "Overview of telematics-based prognostics and health management systems for commercial vehicles," in *International conference on transport systems telematics*, pp. 123–130, Springer, 2013.

[71] P. Zschech, "A taxonomy of recurring data analysis problems in maintenance analytics.," in *ECIS*, p. 197, 2018.

[72] J. Furch, T. Turo, Z. Krobot, and J. Stastny, "Using telemetry for maintenance of special military vehicles," in *International Workshop on Modelling and Simulation for Autonomous Systems*, pp. 392–401, Springer, 2017.

[73] J. Furch, T. Turo, Z. Krobot, and J. Stastny, "Using telemetry for maintenance of special military vehicles," in *International Workshop on Modelling and Simulation for Autonomous Systems*, pp. 392–401, Springer, 2017.

[74] Y. Peng, M. Dong, and M. J. Zuo, "Current status of machine prognostics in condition-based maintenance: a review," *The International Journal of Advanced Manufacturing Technology*, vol. 50, no. 1, pp. 297–313, 2010.

[75] S. Kumar, E. Dolev, and M. Pecht, "Parameter selection for health monitoring of electronic products," *Microelectronics Reliability*, vol. 50, no. 2, pp. 161–168, 2010.

[76] D. Kwon, M. R. Hodkiewicz, J. Fan, T. Shibutani, and M. G. Pecht, "Iot-based prognostics and systems health management for industrial applications," *IEEE Access*, vol. 4, pp. 3659–3670, 2016.

[77] K. Javed, R. Gouriveau, N. Zerhouni, and P. Nectoux, "A feature extraction procedure based on trigonometric functions and cumulative descriptors to enhance prognostics modeling," in *2013 ieee conference on prognostics and health management (phm)*, pp. 1–7, IEEE, 2013.

[78] B. Saha and K. Goebel, "Figure 3 from uncertainty management for diagnostics and prognostics of batteries using bayesian techniques: Semantic scholar," Jan 1970.

[79] S. Pal, P. S. Heyns, B. H. Freyer, N. J. Theron, and S. K. Pal, "Tool wear monitoring and selection of optimum cutting conditions with progressive tool wear effect and input uncertainties," *Journal of Intelligent Manufacturing*, vol. 22, no. 4, pp. 491–504, 2011.

[80] A. Heng, S. Zhang, and A. Tan, "et mathew, j.(2009)," *Rotating machinery prognostics: State of the art, challenges and opportunities. Mechanical Systems and Signal Processing*, vol. 23, pp. 724–739.

[81] R. Isermann, *Fault-diagnosis systems: an introduction from fault detection to fault tolerance*. Springer Science & Business Media, 2005.

[82] G. J. Vachtsevanos and G. J. Vachtsevanos, *Intelligent fault diagnosis and prognosis for engineering systems*, vol. 456. Wiley Online Library, 2006.

[83] "Model-based prognostic techniques," 2003.

[84] F. Chaari, T. Fakhfakh, and M. Haddar, "Analytical modelling of spur gear tooth crack and influence on gearmesh stiffness," *European Journal of Mechanics-A/Solids*, vol. 28, no. 3, pp. 461–468, 2009.

[85] Z. Zhang, Y. Wang, and K. Wang, "Fault diagnosis and prognosis using wavelet packet decomposition, fourier transform and artificial neural network," *Journal of Intelligent Manufacturing*, vol. 24, no. 6, pp. 1213–1227, 2012.

[86] D. He, R. Li, and E. Bechhoefer, "Stochastic modeling of damage physics for mechanical component prognostics using condition indicators," in *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, vol. 48999, pp. 1259–1264, 2009.

[87] M. Schwabacher, "(pdf) a survey of data driven prognostics - researchgate," 2005.

[88] G. P. Zhang, "Time series forecasting using a hybrid arima and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, 2003.

[89] N. Gorjian, L. Ma, M. Mittinty, P. Yarlagadda, and Y. Sun, "A review on degradation models in reliability analysis," in *Engineering asset lifecycle management*, pp. 369–384, Springer, 2010.

[90] T. Benkedjouh, K. Medjaher, N. Zerhouni, and S. Rechak, "Health assessment and life prediction of cutting tools based on support vector regression," *Journal of intelligent manufacturing*, vol. 26, no. 2, pp. 213–223, 2015.

[91] G. E. Box and G. M. Jenkins, "Time series analysis: Forecasting and control san francisco," *Calif: Holden-Day*, 1976.

[92] F. L. Lewis, *Applied optimal control and estimation*. Prentice Hall PTR, 1992.

[93] R. S. Tsay, "Time series and forecasting: Brief history and future research," *Journal of the American Statistical Association*, vol. 95, no. 450, pp. 638–643, 2000.

[94] W. Wu, J. Hu, and J. Zhang, "Prognostics of machine health condition using an improved arima-based prediction method," in *2007 2nd IEEE Conference on Industrial Electronics and Applications*, pp. 1062–1067, Ieee, 2007.

[95] J. Yan, M. Koc, and J. Lee, "A prognostic algorithm for machine performance assessment and its application," *Production Planning & Control*, vol. 15, no. 8, pp. 796–801, 2004.

[96] J. Lee, J. Ni, D. Djurdjanovic, H. Qiu, and H. Liao, "Intelligent prognostics tools and e-maintenance," *Computers in industry*, vol. 57, no. 6, pp. 476–489, 2006.

[97] T. Wang, J. Yu, D. Siegel, and J. Lee, "A similarity-based prognostics approach for remaining useful life estimation of engineered systems," in *2008 international conference on prognostics and health management*, pp. 1–6, IEEE, 2008.

[98] N. Gebraeel, M. Lawley, R. Liu, and V. Parmeshwaran, "Residual life predictions from vibration-based degradation signals: a neural network approach," *IEEE Transactions on industrial electronics*, vol. 51, no. 3, pp. 694–700, 2004.

[99] B. Satish and N. Sarma, "A fuzzy bp approach for diagnosis and prognosis of bearing faults in induction motors," in *IEEE Power Engineering Society General Meeting, 2005*, pp. 2291–2294, IEEE, 2005.

[100] R. Huang, L. Xi, X. Li, C. R. Liu, H. Qiu, and J. Lee, "Residual life predictions for ball bearings based on self-organizing map and back propagation neural network methods," *Mechanical systems and signal processing*, vol. 21, no. 1, pp. 193–207, 2007.

[101] Z. Lei, L. Xingshan, Y. Jinsong, and G. ZhanBao, "A genetic training algorithm of wavelet neural networks for fault prognostics in condition based maintenance," in *2007 8th International Conference on Electronic Measurement and Instruments*, pp. 2–584, IEEE, 2007.

[102] A. P. Vassilopoulos, E. F. Georgopoulos, and V. Dionysopoulos, "Artificial neural networks in spectrum fatigue life prediction of composite materials," *International Journal of Fatigue*, vol. 29, no. 1, pp. 20–29, 2007.

[103] Z. Tian, "An artificial neural network method for remaining useful life prediction of equipment subject to condition monitoring," *Journal of intelligent Manufacturing*, vol. 23, no. 2, pp. 227–237, 2012.

[104] D. Brezak, D. Majetic, T. Udiljak, and J. Kasac, "Tool wear estimation using an analytic fuzzy classifier and support vector machines," *Journal of Intelligent Manufacturing*, vol. 23, no. 3, pp. 797–809, 2012.

[105] A. Gajate, R. Haber, R. Del Toro, P. Vega, and A. Bustillo, "Tool wear monitoring using neuro-fuzzy techniques: a comparative study in a turning process," *Journal of Intelligent Manufacturing*, vol. 23, no. 3, pp. 869–882, 2012.

[106] S. Purushothaman, "Tool wear monitoring using artificial neural network based on extended kalman filter weight updation with transformed input patterns," *Journal of Intelligent Manufacturing*, vol. 21, no. 6, pp. 717–730, 2010.

[107] S. H. Yeo, L. P. Khoo, and S. Neo, "Tool condition monitoring using reflectance of chip surface and neural network," *Journal of Intelligent Manufacturing*, vol. 11, no. 6, pp. 507–514, 2000.

[108] E. Ramasso, M. Rombaut, and N. Zerhouni, "Joint prediction of continuous and discrete states in time-series based on belief functions," *IEEE transactions on cybernetics*, vol. 43, no. 1, pp. 37–50, 2012.

[109] T. Wang, J. Yu, D. Siegel, and J. Lee, "A similarity-based prognostics approach for remaining useful life estimation of engineered systems," in *2008 international conference on prognostics and health management*, pp. 1–6, IEEE, 2008.

[110] "Villanova canavese, torino: Sea soluzioni eco ambientali."

[111] D. A. F. P. B. C. C. Simone Monaco, Paolo Bethaz and T. Cerquitelli1, "Exploring waste-collection fleet data: challenges in a real-world use case from multiple data providers," 2022.

[112] S. Verma, "Why switch to jupyterlab from jupyter-notebook?," Jan 2022.

[113] S. S, "What is pandas in python? everything you need to know," Nov 2021.

[114] I. Mikulić, I. Bošković, and G. Zovak, "Effects of driving style and vehicle maintenance on vehicle roadworthiness," *Promet-Traffic&Transportation*, vol. 32, no. 5, pp. 667–677, 2020.

Elenco delle figure

I.1 Tappe evolutive della storia industriale [3] 10
I.2 Confronto del numero di applicazioni impiegate nel 2019 e nel 2020 . 13
I.3 Le Tecnologie Abilitanti dell'Industria 4.0 [8] 14

II.1 Categorie di algoritmi di apprendimento automatico in base alla natura dei dati di training 27
II.2 Processo KDD in data mining [46] 30
II.3 Decision Tree [60] 36
II.4 Processo di costruzione dell'algoritmo RF [66] 37
II.5 Modello Perceptron 38
II.6 Rete Neurale Feedforward 39

III.1 Logo di SEA Soluzioni Eco Ambientali [110] 46
III.2 Cantieri attivi di SEA nelle regioni d'Italia [110] 47

IV.1 Schema del progetto 52
IV.2 Panoramica delle fatture e dell'attività giornaliera dei sensori sui veicoli associati [111] 55
IV.3 Conteggio pulito dei fallimenti, separati per categoria [111]. 57
IV.4 Matrice di correlazione (feature di P1) 62

V.1 Diagramma di Venn dell'aggregazione semantica di alto livello delle caratteristiche di 3 fornitori di dati di tracciamento [111] 66

V.2 Confronto tra il contachilometri e il tempo di lavoro del motore per 3 veicoli monitorati da P1 e P3 [111] 68
V.3 Distribuzione giornaliera della velocità per ogni provider. Due punti evidenziano la velocità media misurata per v17 [111] 69
V.4 Analisi nel tempo per P1 [111] 71
V.5 Analisi nel tempo per P2 [111] 72
V.6 Analisi nel tempo per P3 [111] 73
V.7 Numero massimo di eventi per ogni coppia targa-categoria 74
V.8 Frequenza dei guasti dell'impianto di sensoristica del V01 75
V.9 Frequenza di guasto di altri tipi di impianti dei veicoli V02, V03, V04, V05 75
V.10 Gruppi di cluster per K=4 77
V.11 Rappresentazione delle correlazioni tra veicoli 78
V.12 Overview del Calendario dei Servizi Fiorentino 2021 79
V.13 Calendario Servizi Fiorentino nel mese di Gennaio 2021 80
V.14 Comuni appartenenti al Calendario Servizi Fiorentino 2021 dei veicoli V07 (a sinistra) e V08 (a destra) 80
V.15 Raccolta eseguita dal veicolo V07 nell'anno 2021 81
V.16 Raccolta eseguita dal veicolo V08 nell'anno 2021 81
V.17 Confronto tra i dataset Turni di Servizio e fornitore P1 del veicolo V07 82
V.18 Confronto tra i dataset Turni di Servizio e fornitore P1 del veicolo V08 83
V.19 Count di DriverIdentification 85
V.20 Count di DriverCode 86
V.21 Risultati del modello LSTM 87
V.22 Modello generato su un singolo truck (V02) 89
V.23 Modello allenato su n-1 trucks (V07, V05, V14) e testato su un truck (V02) 90

Elenco delle tabelle

III.1 Fornitori di dati 48

IV.1 Ipotesi di correlazioni tra Guasti e Caratteristiche del veicolo 57
IV.2 Ipotesi di correlazioni tra Guasti e il Contesto in cui il veicolo vive . . 58
IV.3 Cantiere cui appartengono i veicoli di interesse nel 2021 e nel 2022 . . 59
IV.4 Lista dei Dipendenti di SEA 61

V.1 Calcolo di delta_data del veicolo V01 76
V.2 Calcolo delle statistiche di alcuni veicoli di esempio 76
V.3 Numero di veicoli appartenenti a ciascun cluster 78
V.4 Esempio della struttura dei dati di DipendentList 84
V.5 Merge tra i dataset P6 e DipendentList 85
V.6 Merge tra i dataset P12 e DipendentList 86
V.7 Parametri sperimentali settati per LSTM 87
V.8 Risultati sperimentali dell'algoritmo XGBoost 89