Strategie innovative per la logistica: il valore del kitting e assembly nel settore idrotermosanitario

Leogrande, Angelo

LUM UNIVERSITY GIUSEPPE DEGENNARO

2024

Strategie innovative per la logistica: il valore del kitting e assembly nel settore idrotermosanitario

Angelo Leogrande, LUM University Giuseppe Degennaro, leogrande.cultore@lum.it

Abstract

L'articolo esplora l'importanza strategica dell'implementazione dei servizi di kitting e assembly per affrontare problematiche di assegnazione delle risorse in un magazzino operante nel settore idrotermosanitario. Si concentra sulla crescente complessità delle operazioni logistiche in un contesto caratterizzato da una domanda sempre più personalizzata e dalla necessità di garantire tempi di consegna rapidi. Attraverso un'analisi approfondita, il lavoro evidenzia come il kitting e l'assembly siano strumenti fondamentali per ottimizzare i flussi operativi, migliorare l'efficienza e soddisfare le aspettative dei clienti. Il kitting viene descritto come il processo di raggruppamento di componenti per assemblaggi specifici, contribuendo alla riduzione dei tempi operativi e minimizzando gli errori umani. L'assembly, d'altro canto, completa il ciclo producendo kit semi-finiti o finiti, pronti per la distribuzione. L'articolo analizza il valore di questa integrazione, mostrando come essa migliori la gestione degli spazi e la tracciabilità dei materiali, oltre a fornire un vantaggio competitivo. La ricerca adotta un approccio olistico, prendendo in esame sia gli aspetti tecnologici, come l'uso di software di gestione logistico avanzato, sia quelli collaborativi, evidenziando l'importanza del coordinamento tra risorse umane e materiali. Inoltre, include casi studio dettagliati che dimostrano i benefici tangibili delle soluzioni implementate, come la riduzione degli errori, l'aumento dell'efficienza e un impatto positivo sulla sostenibilità. Questo lavoro rappresenta un contributo significativo per le aziende che intendono migliorare la gestione logistica, con un focus su innovazione e ottimizzazione dei processi.

Keywords: Kitting, Assembly, Idrotermosanitario, Machine Learning Regressions, Machine Learning Clustering.

JEL CODES: L91, L92, L94, R40, R41,M1.

1. Introduzione

L'articolo si concentra sull'implementazione dei servizi di kitting e assembly come soluzione strategica per affrontare i problemi logistici legati all'assegnazione delle risorse in un magazzino operante nel settore idrotermosanitario. La crescente complessità delle operazioni logistiche, guidata dall'aumento della domanda di personalizzazione dei prodotti e dalla necessità di garantire consegne rapide e precise, ha spinto molte aziende a ricercare soluzioni innovative per ottimizzare i flussi operativi. Il kitting e l'assembly si sono affermati come strategie chiave per migliorare l'efficienza, ridurre gli errori e soddisfare le aspettative dei clienti. Il kitting, definito come il processo di raggruppamento e confezionamento di componenti necessari per assemblaggi o installazioni specifiche, rappresenta un elemento cruciale per migliorare la gestione delle risorse in magazzino. Esso consente di standardizzare operazioni complesse, riducendo i tempi di lavorazione e minimizzando il rischio di errori umani. L'assembly, d'altra parte, si occupa della combinazione di tali componenti per creare prodotti semi-finito o finiti, pronti per essere distribuiti. L'integrazione di questi due servizi offre alle aziende un vantaggio competitivo, ottimizzando l'uso degli spazi e migliorando la tracciabilità dei materiali. Un aspetto centrale della ricerca è la connessione tra l'adozione di sistemi di kitting e assembly e la capacità di un'organizzazione di rispondere rapidamente ai cambiamenti della domanda del mercato. La pressione esercitata da settori come quello idrotermosanitario, caratterizzato da alti volumi e frequenti variazioni nelle richieste dei

clienti, richiede un approccio altamente dinamico e flessibile. Attraverso l'uso di software di gestione logistico avanzato, l'articolo dimostra come sia possibile aumentare la visibilità e il controllo dei processi, migliorando l'accuratezza e l'efficienza complessiva (Kilic and Durmusoglu, 2012; Ramnath, et al., 2014; Gajjar and Thakkar, 2014).

Un altro elemento distintivo affrontato nell'articolo è la dimensione collaborativa dell'implementazione di questi servizi. Il successo del kitting e dell'assembly non dipende solo dalle tecnologie impiegate, ma anche dalla capacità di coordinare efficacemente le risorse umane e materiali. Esempi pratici tratti dal settore dimostrano come la suddivisione dei ruoli in squadre specializzate, come quelle dedicate al prelievo, all'assemblaggio e alla spedizione, possa ridurre significativamente i tempi operativi e migliorare l'accuratezza degli ordini. L'articolo analizza anche l'impatto economico di tali implementazioni, evidenziando come il kitting e l'assembly possano ridurre i costi operativi attraverso la standardizzazione dei processi e la diminuzione degli errori. In particolare, l'uso di kit preassemblati consente un migliore utilizzo dello spazio di magazzino e una gestione più efficace delle spedizioni, contribuendo a ridurre i costi di trasporto e il consumo di materiali di imballaggio. Questa ottimizzazione operativa non solo migliora la redditività aziendale, ma ha anche un impatto positivo sulla sostenibilità ambientale. Il lavoro si distingue per la sua attenzione all'integrazione tecnologica, esplorando come strumenti come l'Internet of Things (IoT) e i sistemi di tracciamento in tempo reale possano facilitare la gestione di operazioni logistiche complesse. Tuttavia, l'articolo non ignora le sfide associate all'implementazione di queste tecnologie, come i costi iniziali elevati e la necessità di formare il personale per gestire nuovi sistemi digitali. Affrontare tali sfide è fondamentale per garantire che i benefici dell'automazione e della digitalizzazione siano pienamente realizzati. La ricerca sottolinea infine l'importanza della standardizzazione delle procedure operative per risolvere i problemi di assegnazione delle risorse. I modelli matematici utilizzati per analizzare i flussi operativi e i dati empirici forniti attraverso esempi numerici dimostrano come la segmentazione delle attività e l'ottimizzazione delle risorse possano contribuire a un significativo miglioramento dell'efficienza complessiva. Questi strumenti analitici permettono di identificare i colli di bottiglia e di assegnare le priorità alle attività critiche, garantendo una gestione più fluida e coordinata. L'articolo si distingue anche per il suo approccio pratico, presentando esempi concreti di implementazione nel settore idrotermosanitario. Attraverso casi studio dettagliati, vengono mostrati i benefici tangibili di queste strategie, come la riduzione dei tempi operativi, la diminuzione degli errori e un miglioramento complessivo della soddisfazione del cliente. Tali esempi offrono un valido punto di riferimento per altre aziende interessate a implementare soluzioni simili. In conclusione, l'articolo fornisce un contributo significativo alla comprensione dell'importanza dei servizi di kitting e assembly per migliorare la gestione logistica. Esso mette in evidenza non solo i benefici operativi ed economici, ma anche le sfide e le opportunità legate alla loro implementazione. La capacità di integrare tecnologie avanzate con un'efficace gestione delle risorse umane rappresenta il fulcro di una strategia di successo, in grado di garantire un vantaggio competitivo nel contesto odierno, caratterizzato da una crescente complessità e dinamicità dei mercati (Zhao et al., 2021; Bortolini et al., 2020; Maatar et al., 2022).

Dati. Occorre considerare che i dati utilizzati in questo studio sono stati generati attraverso l'intelligenza artificiale. In modo particolare sono stati generati attraverso l'intelligenza artificiale sia i casi di studio relativi alla risoluzione del problema dell'assegnazione, sia il database che è stato utilizzato per addestrare gli algoritmi di machine learning utilizzati sia per la regressione che per la clusterizzazione. L'utilizzo dell'intelligenza artificiale per la data generation è stata realizzata per sopperire alla mancanza di dati operativi relativi alla risoluzione del problema dell'assegnazione nell'ambito di un magazzino idrotermosanitario.

L'articolo continua come di seguito: il secondo paragrafo contiene la literature review, il terzo paragrafo presenta degli esempi di risoluzione del problema dell'assegnazione in un magazzino

idrotermosanitario, il quarto paragrafo contiene l'applicazione di vari algoritmi di machine learning per la regressione, il quinto paragrafo mostra l'applicazione di vari algoritmi di clusterizzazione, il sesto paragrafo indica degli accorgimenti manageriali derivanti dall'analisi svolta, il settimo paragrafo conclude.

2. Literature review

Lo studio di Ahmed, Parvathaneni e Shareef (2023) si concentra sulla riorganizzazione dell'inventario per migliorare l'efficienza e l'utilizzo dello spazio. Gli autori utilizzano un approccio pratico, applicabile direttamente in contesti manifatturieri, dimostrando come una gestione più strategica del layout e del flusso dei materiali possa ridurre i tempi di preparazione dei kit. Sebbene i risultati siano rilevanti, il lavoro appare limitato a scenari specifici e non considera appieno l'impatto di fattori esterni, come le variazioni nella domanda o la crescente digitalizzazione dei processi logistici. Caputo, Pelagagge e Salini (2021) adottano un approccio comparativo, confrontando i sistemi manuali e automatizzati di kitting sotto il profilo economico e organizzativo. L'analisi costi-benefici rappresenta il punto di forza del lavoro, in quanto offre alle aziende uno strumento utile per decidere tra investimenti in automazione e miglioramenti nei processi manuali. Tuttavia, la complessità dell'implementazione dei sistemi automatizzati, soprattutto per le piccole e medie imprese, non viene affrontata in modo esaustivo, lasciando aperte domande su come ridurre le barriere di accesso tecnologiche. Fatima, Mohammed e Shareef (2024) propongono una prospettiva olistica che combina l'ottimizzazione dei carrelli per il kitting e l'analisi dei processi di ricezione dell'acciaio. Questo lavoro si distingue per l'integrazione tra design operativo e analisi dei dati, enfatizzando la necessità di un approccio sistemico per migliorare la qualità complessiva dei processi. Tuttavia, la generalizzabilità delle conclusioni è discutibile, poiché l'analisi si basa su uno studio di caso specifico che potrebbe non essere rappresentativo di altri settori o contesti. El Moussaoui et al. (2021) spostano l'attenzione verso il settore edilizio, esaminando i centri logistici per la costruzione che offrono servizi di kitting. L'articolo presenta un'analisi organizzativa e una mappatura dei costi, sottolineando i vantaggi in termini di riduzione degli sprechi e maggiore coordinamento. Sebbene il lavoro fornisca un contributo utile, non esplora a sufficienza le sfide legate alla scalabilità e all'implementazione pratica di tali centri logistici in progetti di costruzione complessi e frammentati. Dakhli e Lafhaj (2022) introducono una prospettiva innovativa, collegando il kitting alla gestione della supply chain edilizia tramite l'uso del Building Information Modeling (BIM). La combinazione tra approccio lean e tecnologie 4.0 evidenzia le opportunità offerte dalla digitalizzazione, ma il lavoro risulta più teorico che pratico, mancando una validazione empirica robusta che dimostri l'efficacia di queste soluzioni.

Montoya Zapata et al. (2023) si concentrano su una simulazione di sistemi di kitting per il rifornimento di una linea di assemblaggio automobilistico. Il lavoro sottolinea l'importanza della modellazione e della simulazione per anticipare le inefficienze e migliorare le decisioni operative. Tuttavia, l'approccio presentato, pur rigoroso, rimane confinato al livello teorico e manca di una validazione empirica che dimostri i vantaggi concreti del sistema proposto. La sua utilità pratica potrebbe risultare limitata senza ulteriori test su scala reale, un aspetto che spesso caratterizza studi basati su simulazioni. Bueno Viso (2022) presenta una stazione di kitting automatizzata (AGS), ponendo l'accento sull'automazione come strumento per aumentare la precisione e ridurre i tempi di ciclo. L'articolo enfatizza i benefici della robotica nel migliorare la qualità e la velocità dei processi, ma non approfondisce a sufficienza le sfide di implementazione, come i costi elevati, le difficoltà di integrazione con i sistemi esistenti o la resistenza organizzativa al cambiamento. Un confronto con le alternative manuali o ibride avrebbe fornito una prospettiva più equilibrata. Ferrari et al. (2022) offrono una prospettiva educativa, simulando i processi di magazzino automatizzato con studenti di ingegneria. Questo approccio è interessante perché combina teoria e pratica, promuovendo la formazione di competenze utili per il settore industriale. Tuttavia, il focus su un contesto accademico potrebbe ridurre la rilevanza per le applicazioni reali, in cui i vincoli e le complessità operative sono

ben più articolati. Inoltre, non vengono affrontate questioni come la scalabilità o l'impatto economico dell'automazione proposta. Zhang et al. (2022) introducono una soluzione innovativa basata sulla realtà aumentata (AR) per il picking manuale, dimostrando come le tecnologie immersive possano migliorare l'efficienza e ridurre gli errori umani. Sebbene il lavoro sia convincente dal punto di vista tecnico, permangono interrogativi sull'adattabilità di tali soluzioni a diversi settori produttivi, così come sui costi e sull'accessibilità tecnologica per le PMI. Inoltre, l'articolo manca di un'analisi comparativa con altre tecnologie emergenti che potrebbero offrire vantaggi simili. Berroir et al. (2021) affrontano il kitting nel contesto della logistica edilizia, esplorando il potenziale del Consolidation and Coordination Centre (CCC) per ridurre i costi e l'impatto ambientale. Questo contributo è particolarmente rilevante per la crescente attenzione alla sostenibilità, ma le implicazioni pratiche del modello proposto sono limitate a un caso di studio specifico, lasciando aperte domande sulla scalabilità e sulla replicabilità in contesti diversi.

Lo studio di Kule, Patil e Vaity (2020) si concentra sull'introduzione del kitting nelle linee di assemblaggio per migliorare il processo di alimentazione dei materiali. L'approccio proposto mira a ottimizzare l'efficienza riducendo i tempi di inattività e i problemi legati alla gestione tradizionale del magazzino. Tuttavia, l'articolo tende a enfatizzare i benefici senza approfondire le difficoltà pratiche legate all'implementazione, come l'impatto sui costi iniziali o l'adattabilità del sistema a produzioni di diversa complessità. Inoltre, la mancanza di un confronto con approcci alternativi rende il lavoro meno critico di quanto potrebbe essere. Jum'a e Basheer (2023) utilizzano il principio di Pareto per analizzare i servizi a valore aggiunto nei magazzini di un fornitore logistico di terze parti. Questo approccio di qualità si rivela utile per identificare i punti critici e concentrare gli sforzi di miglioramento sui processi più influenti. Sebbene l'analisi fornisca intuizioni interessanti, il caso studio trattato potrebbe risultare limitante: l'assenza di una discussione più ampia su come applicare questa metodologia a contesti diversi riduce la generalizzabilità delle conclusioni. Inoltre, il legame con il kitting rimane marginale, suggerendo la necessità di un approccio più integrato. Costa, Pinto e Gonçalves (2024) propongono un sistema di localizzazione wireless applicato al kitting basato su pick-to-light. Questo contributo evidenzia il potenziale delle tecnologie avanzate per aumentare l'accuratezza e la velocità dei processi. Tuttavia, l'articolo solleva interrogativi sull'effettiva implementabilità del sistema proposto, in particolare per quanto riguarda i costi di adozione e la compatibilità con infrastrutture già esistenti. La tecnologia wireless, pur promettente, potrebbe presentare problemi di affidabilità in ambienti complessi o con numerosi ostacoli fisici. Geraeds e Llamoca (2023) analizzano la previsione del fabbisogno di manodopera nelle linee di assemblaggio a basso volume e alta complessità. Questo studio è particolarmente rilevante per settori dove la personalizzazione del prodotto richiede un equilibrio tra automazione e lavoro umano. Tuttavia, il focus sull'analisi predittiva potrebbe trascurare aspetti legati alla formazione e al coinvolgimento del personale, cruciali per garantire il successo delle operazioni di kitting in contesti dinamici. Infine, Mabeya (2022) esplora come l'uso di sistemi moderni di gestione e tracciabilità possa migliorare la gestione dei materiali SMT (Surface-Mount Technology) nei magazzini. Il lavoro pone enfasi sulla digitalizzazione e sull'automazione come strumenti per ridurre gli errori e migliorare la visibilità nella supply chain. Sebbene l'approccio sia ben articolato, manca una valutazione più critica delle possibili difficoltà, come la resistenza al cambiamento o le competenze richieste per utilizzare le nuove tecnologie.

Lo studio di Dakhli e Lafaj (2022) si inserisce nel più ampio ambito della costruzione lean e della trasformazione digitale nel settore AEC (Architecture, Engineering, and Construction). Gli autori esplorano come l'approccio lean possa essere integrato con tecnologie avanzate per migliorare la gestione dei materiali attraverso il kitting. Sebbene il lavoro presenti una visione interessante e teoricamente solida, risulta poco ancorato alla pratica. Mancano esempi applicativi che dimostrino concretamente i vantaggi del metodo proposto, un limite che riduce l'impatto del contributo rispetto ad altri studi più pragmatici. Bortolini et al. (2020) affrontano il tema della progettazione del layout

dei magazzini per la gestione di kit con attributi fisici variabili. Il loro studio si distingue per l'attenzione ai dettagli operativi, proponendo un procedimento di assegnazione che ottimizza l'efficienza complessiva. Tuttavia, l'approccio proposto appare complesso e altamente specifico, il che potrebbe limitarne l'applicabilità in contesti meno strutturati. Inoltre, l'articolo non analizza a sufficienza le implicazioni economiche delle soluzioni suggerite, lasciando un importante vuoto nella valutazione complessiva. Thai e Norlander (2021) analizzano le potenziali conseguenze dell'introduzione di un nuovo concetto logistico presso Volvo Trucks. Il caso studio offre un'analisi approfondita, evidenziando i benefici ma anche i rischi associati a cambiamenti strutturali nelle operazioni logistiche. Tuttavia, il lavoro si concentra quasi esclusivamente su un contesto aziendale specifico, rendendo difficile generalizzare i risultati. Sarebbe stato utile includere un confronto con altre aziende o settori per ampliare la portata delle conclusioni. Simões, Pinto e Silva (2023) indagano sull'ottimizzazione del consumo energetico nei sistemi di kitting robotizzato per l'industria automobilistica. L'approccio è innovativo, combinando sostenibilità ambientale e automazione. Questo contributo si distingue per la rilevanza contemporanea, considerando l'importanza crescente della riduzione dell'impatto energetico. Tuttavia, il focus sulla robotica può essere visto come un limite in termini di applicabilità per settori con minori risorse tecnologiche. Inoltre, l'articolo non esplora sufficientemente le possibili sinergie tra soluzioni automatizzate e pratiche tradizionali. Tan et al. (2021) presentano una stazione di kitting e confezionamento accessibile, sviluppata nell'ambito del SourceAmerica Design Challenge. Questo contributo si distingue per il suo obiettivo sociale, enfatizzando l'importanza di creare soluzioni logistiche inclusive per lavoratori con disabilità. Sebbene l'approccio sia lodevole, mancano analisi approfondite su come scalare o adattare questa tecnologia ad ambienti industriali più complessi. Inoltre, il lavoro è più descrittivo che analitico, lasciando aperte questioni riguardo all'efficacia pratica del sistema proposto.

Lo studio di Dakhli e Lafaj (2022) si inserisce nel più ampio ambito della costruzione lean e della trasformazione digitale nel settore AEC (Architecture, Engineering, and Construction). Gli autori esplorano come l'approccio lean possa essere integrato con tecnologie avanzate per migliorare la gestione dei materiali attraverso il kitting. Sebbene il lavoro presenti una visione interessante e teoricamente solida, risulta poco ancorato alla pratica. Mancano esempi applicativi che dimostrino concretamente i vantaggi del metodo proposto, un limite che riduce l'impatto del contributo rispetto ad altri studi più pragmatici. Bortolini et al. (2020) affrontano il tema della progettazione del layout dei magazzini per la gestione di kit con attributi fisici variabili. Il loro studio si distingue per l'attenzione ai dettagli operativi, proponendo un procedimento di assegnazione che ottimizza l'efficienza complessiva. Tuttavia, l'approccio proposto appare complesso e altamente specifico, il che potrebbe limitarne l'applicabilità in contesti meno strutturati. Inoltre, l'articolo non analizza a sufficienza le implicazioni economiche delle soluzioni suggerite, lasciando un importante vuoto nella valutazione complessiva. Thai e Norlander (2021) analizzano le potenziali conseguenze dell'introduzione di un nuovo concetto logistico presso Volvo Trucks. Il caso studio offre un'analisi approfondita, evidenziando i benefici ma anche i rischi associati a cambiamenti strutturali nelle operazioni logistiche. Tuttavia, il lavoro si concentra quasi esclusivamente su un contesto aziendale specifico, rendendo difficile generalizzare i risultati. Sarebbe stato utile includere un confronto con altre aziende o settori per ampliare la portata delle conclusioni. Simões, Pinto e Silva (2023) indagano sull'ottimizzazione del consumo energetico nei sistemi di kitting robotizzato per l'industria automobilistica. L'approccio è innovativo, combinando sostenibilità ambientale e automazione. Questo contributo si distingue per la rilevanza contemporanea, considerando l'importanza crescente della riduzione dell'impatto energetico. Tuttavia, il focus sulla robotica può essere visto come un limite in termini di applicabilità per settori con minori risorse tecnologiche. Inoltre, l'articolo non esplora sufficientemente le possibili sinergie tra soluzioni automatizzate e pratiche tradizionali. Tan et al. (2021) presentano una stazione di kitting e confezionamento accessibile, sviluppata nell'ambito del SourceAmerica Design Challenge. Questo contributo si distingue per il suo obiettivo sociale, enfatizzando l'importanza di creare soluzioni logistiche inclusive per lavoratori con disabilità.

Sebbene l'approccio sia lodevole, mancano analisi approfondite su come scalare o adattare questa tecnologia ad ambienti industriali più complessi. Inoltre, il lavoro è più descrittivo che analitico, lasciando aperte questioni riguardo all'efficacia pratica del sistema proposto.

Bottin et al. (2021) esplorano la produzione automatizzata di kit di vendita, proponendo una procedura integrata per ridurre i tempi di setup in contesti caratterizzati da un'elevata varietà di prodotti. La forza dell'articolo risiede nell'applicazione di strumenti pratici per affrontare una problematica comune nelle industrie moderne. Tuttavia, l'approccio sembra limitato a contesti di produzione specifici e non considera adeguatamente l'impatto economico e organizzativo che un sistema altamente automatizzato potrebbe avere su aziende di dimensioni ridotte o meno strutturate. Inoltre, l'attenzione alla riduzione dei tempi di setup, sebbene fondamentale, potrebbe risultare insufficiente in scenari che richiedono anche un'elevata flessibilità operativa. Ferrari et al. (2021) introducono un approccio di simulazione integrata per magazzini "4.0", sottolineando l'importanza di combinare tecnologie avanzate come IoT e big data per migliorare l'efficienza operativa. Lo studio è all'avanguardia e offre un quadro utile per l'implementazione di sistemi intelligenti nei magazzini. Tuttavia, l'articolo non approfondisce le barriere tecniche ed economiche che potrebbero limitare l'applicazione di tali soluzioni, specialmente in contesti caratterizzati da budget limitati o infrastrutture obsolete. La mancanza di un confronto con metodi tradizionali di gestione del magazzino riduce inoltre la comprensibilità del reale valore aggiunto dell'approccio proposto. Buzu (2021) analizza l'effetto della gestione del magazzino sulle prestazioni complessive dello stesso. L'articolo, pur basandosi su un quadro concettuale ben articolato, soffre di un'eccessiva genericità, senza fornire un'analisi sufficientemente approfondita o dati empirici che sostengano le affermazioni proposte. Sebbene venga riconosciuto il ruolo strategico del magazzino all'interno della supply chain, il lavoro non affronta specificamente le complessità del kitting, né propone soluzioni operative innovative. Questo lo rende meno rilevante rispetto agli altri articoli che trattano direttamente tematiche di automazione e ottimizzazione. Tornese et al. (2020) analizzano il modello di magazzino on-demand, descrivendone le caratteristiche principali e il potenziale come soluzione logistica flessibile. Sebbene l'articolo offra spunti interessanti, manca di una valutazione critica sui limiti di questi modelli, come la dipendenza da tecnologie esterne e la vulnerabilità a fluttuazioni della domanda. Inoltre, il lavoro non esplora le implicazioni del modello on-demand sul kitting, lasciando un vuoto importante in termini di applicabilità diretta. Montoya-Zapata et al. (2024) propongono un sistema multi-agente per gestire perturbazioni nel processo di kitting in una linea di assemblaggio automobilistico. L'articolo si distingue per l'uso dell'intelligenza artificiale applicata a scenari complessi e dinamici, dimostrando come i sistemi multi-agente possano migliorare la resilienza e l'efficienza. Tuttavia, il lavoro non esplora sufficientemente l'adattabilità del sistema ad altri settori e non fornisce una valutazione dettagliata dei costi di implementazione, elementi essenziali per valutarne l'impatto industriale su larga scala.

Baglio, Creazza e Dallari (2024) introducono un modello di maturità per valutare l'adozione delle tecnologie della "Logistica 4.0" nell'industria dei 3PL. Il contributo è interessante perché fornisce uno strumento analitico per misurare il grado di implementazione delle tecnologie avanzate, come IoT e blockchain. Tuttavia, l'approccio teorico rischia di essere poco applicabile a realtà aziendali con risorse limitate o con infrastrutture obsolete. Inoltre, la scala di maturità proposta manca di un'analisi dettagliata sull'impatto economico delle tecnologie in questione, un aspetto cruciale per le decisioni strategiche delle imprese. Raja e Venkatachalam (2022) offrono una revisione della letteratura sull'adozione delle tecnologie digitali nei servizi logistici globali. Il lavoro evidenzia come l'innovazione digitale stia trasformando il settore, ma manca di un'analisi critica che discuta le barriere culturali, economiche e operative che molte aziende affrontano nel processo di trasformazione digitale. Inoltre, la revisione si concentra principalmente su casi di successo, trascurando esempi di implementazioni fallite che avrebbero potuto fornire lezioni preziose per una comprensione più completa del fenomeno. Jiang et al. (2021) propongono un approccio di

ottimizzazione spaziale e temporale per i magazzini intelligenti con elevata rotazione delle merci. Questo studio si distingue per l'applicazione di modelli matematici per risolvere problemi pratici, contribuendo alla gestione ottimale dello spazio e dei flussi di materiale. Tuttavia, l'approccio risulta altamente tecnico e potrebbe essere difficile da implementare in contesti aziendali con competenze limitate in analisi quantitativa. Inoltre, l'articolo non affronta il problema della scalabilità del modello in magazzini più grandi o più complessi. Vaka (2020) analizza il sistema Embedded Extended Warehouse Management (EWM) di S/4HANA, evidenziando come questo strumento possa massimizzare l'efficienza della gestione del magazzino. Sebbene il lavoro fornisca un'analisi dettagliata delle funzionalità tecniche del sistema, manca una discussione critica sulle sfide pratiche legate alla sua implementazione, come la formazione del personale o i costi di aggiornamento delle infrastrutture IT. Inoltre, l'articolo appare strettamente legato al software specifico, limitando la sua applicabilità a realtà aziendali che utilizzano altre soluzioni tecnologiche.

L'articolo di Klundt, Towers e Bechkoum (2024) esplora l'applicazione di strategie lean e agili nei centri di distribuzione per fornire servizi a valore aggiunto (VAS). Gli autori sottolineano come la combinazione di approcci lean e agili possa migliorare la flessibilità e ridurre gli sprechi, rispondendo meglio alle esigenze dei clienti. Tuttavia, il lavoro rischia di rimanere troppo teorico, poiché non offre esempi pratici o casi studio che dimostrino concretamente l'applicazione delle strategie proposte. Inoltre, non viene affrontata la possibile tensione tra l'approccio lean, focalizzato sull'eliminazione degli sprechi, e quello agile, orientato a una maggiore flessibilità, che in alcuni casi potrebbe portare a costi aggiuntivi. Bogue (2022) discute l'impatto della pandemia di Covid-19 e delle innovazioni tecnologiche sul mercato dei robot per magazzini, evidenziando come questi abbiano rivoluzionato il settore migliorando l'efficienza e riducendo la dipendenza dal lavoro umano. Sebbene l'articolo offra una panoramica interessante delle tendenze di mercato, non esplora sufficientemente le barriere all'adozione della robotica, come i costi iniziali elevati e la necessità di formazione specifica per il personale. Inoltre, l'impatto sociale della riduzione della manodopera viene toccato solo superficialmente, nonostante rappresenti una questione rilevante sia dal punto di vista etico che economico. Karim et al. (2021) propongono un approccio rivisto per misurare la produttività dei magazzini, introducendo indicatori basati su rapporti standardizzati. Questo studio è particolarmente utile per le aziende che cercano di migliorare il monitoraggio delle prestazioni. Tuttavia, l'articolo manca di un'analisi approfondita sull'applicabilità universale di questi benchmark, che potrebbero risultare meno efficaci in contesti con infrastrutture meno sviluppate o con flussi di lavoro non standardizzati. Inoltre, l'assenza di un confronto con altri metodi di misurazione della produttività riduce la capacità di valutare il valore aggiunto delle proposte degli autori. Yusup (2022) analizza l'impatto delle politiche di magazzino doganale nel porto di Makassar sulla riduzione dei costi logistici. Sebbene l'articolo offra una prospettiva interessante sul legame tra politiche pubbliche e logistica, risulta limitato dal contesto locale. L'assenza di un confronto con altri porti o politiche simili a livello internazionale rende difficile valutare la trasferibilità delle conclusioni. Inoltre, il lavoro non esplora sufficientemente le implicazioni pratiche per gli operatori logistici, limitando il suo valore per i decisori aziendali. Infine, Kujanpää (2024) esamina le pratiche logistiche nella costruzione lean, identificandone sfide e migliori pratiche. L'articolo offre un quadro completo dei vantaggi di un approccio lean, ma non riesce a spiegare come superare le difficoltà operative, come la resistenza al cambiamento culturale o l'adattamento delle tecniche lean a progetti complessi e su larga scala. La mancanza di dati empirici o casi studio approfonditi limita l'impatto del lavoro, lasciando il lettore con più domande che risposte.

Mudyazhezha (2024) esamina il ruolo della logistica e dei magazzini nella riduzione delle perdite post-raccolta nella filiera del mais. L'articolo evidenzia l'importanza di infrastrutture adeguate per preservare la qualità dei raccolti e garantire la sicurezza alimentare. Tuttavia, il lavoro si basa principalmente su una revisione della letteratura, senza proporre soluzioni innovative o studiare casi pratici di successo. Questo limita la trasferibilità delle conclusioni in contesti diversi da quello

specifico dello Zimbabwe. Inoltre, non viene esplorato a fondo il ruolo della tecnologia, che potrebbe offrire soluzioni cruciali per migliorare l'efficienza della filiera agricola. Winkelhaus e Grosse (2022) offrono una prospettiva socio-tecnica sui magazzini intelligenti, sottolineando la necessità di bilanciare innovazioni tecnologiche con le esigenze umane. L'articolo è particolarmente rilevante in quanto affronta il tema della trasformazione digitale non solo dal punto di vista tecnico, ma anche delle interazioni tra sistemi automatizzati e lavoratori. Tuttavia, la ricerca rimane teorica e manca di esempi pratici che mostrino come mitigare i conflitti socio-tecnici, come la resistenza al cambiamento o il rischio di disoccupazione tecnologica. Una maggiore attenzione alla formazione del personale o all'integrazione graduale dei sistemi automatizzati potrebbe rafforzare il contributo degli autori. Lorson, Fügener e Hübner (2023) indagano le interazioni tra esseri umani e sistemi robotizzati nei magazzini. Lo studio si distingue per l'approccio interdisciplinare, che combina elementi di psicologia del lavoro con l'ingegneria industriale, offrendo una comprensione più profonda delle dinamiche di collaborazione. Tuttavia, il lavoro non considera sufficientemente i costi e le difficoltà legate all'implementazione di tali sistemi, come l'adattamento alle infrastrutture esistenti o le limitazioni tecnologiche in contesti meno avanzati. Inoltre, non emerge chiaramente come le aziende possano affrontare i problemi di fiducia e accettazione da parte dei lavoratori, un elemento cruciale per il successo di queste soluzioni. Coe (2021) esplora le strategie del settore 3PL nell'Asia-Pacifico per affrontare la crescente commoditizzazione dei servizi logistici. L'autore offre un'analisi approfondita delle dinamiche competitive in una regione chiave per il commercio globale, enfatizzando la necessità di differenziazione attraverso l'innovazione e il miglioramento del valore aggiunto. Tuttavia, l'articolo si focalizza principalmente sul livello macroeconomico, trascurando le implicazioni pratiche per le singole aziende. Inoltre, sarebbe stato utile un confronto con altre aree geografiche per verificare se le strategie discusse siano applicabili a un contesto globale. Elia, Gnoni e Tornese (2024) analizzano le piattaforme di magazzinaggio on-demand, presentandole come un modello innovativo della sharing economy applicato alla logistica. Il lavoro è interessante perché collega la flessibilità operativa con le opportunità offerte dalla digitalizzazione. Tuttavia, l'articolo non approfondisce le sfide regolatorie e le possibili criticità legate alla sostenibilità di tali modelli, come la dipendenza da infrastrutture digitali avanzate o i potenziali impatti negativi sui lavoratori del settore.

Maderna et al. (2020) presentano un algoritmo di schedulazione online per la collaborazione tra esseri umani e robot nel processo di kitting. Il lavoro si distingue per l'innovazione tecnica, offrendo un contributo interessante nel campo della robotica collaborativa. Tuttavia, il modello proposto risulta altamente specifico e poco adattabile a contesti diversi da quello analizzato. Inoltre, non vengono affrontate le sfide legate all'implementazione pratica, come la resistenza dei lavoratori umani o l'ottimizzazione dei costi. Una maggiore attenzione alle implicazioni sociali ed economiche dell'automazione collaborativa potrebbe migliorare l'impatto del lavoro. Ali et al. (2024) propongono un quadro decisionale per valutare la prontezza di un'organizzazione nell'adozione di magazzini intelligenti. Questo contributo teorico è rilevante, poiché offre un approccio sistematico per identificare i fattori critici di successo nella trasformazione digitale. Tuttavia, il lavoro si basa principalmente su modelli teorici e manca di validazioni empiriche solide. Inoltre, non viene considerato a sufficienza il ruolo delle piccole e medie imprese, che spesso affrontano sfide uniche rispetto alle grandi organizzazioni. L'integrazione di casi studio o esperimenti sul campo avrebbe rafforzato la rilevanza pratica del framework proposto. Bogue (2024) analizza il ruolo dei robot nella logistica, evidenziando come la robotica stia trasformando le operazioni di magazzinaggio e distribuzione. L'articolo offre una panoramica utile delle tendenze emergenti, ma la trattazione rimane generale e manca di una discussione approfondita sulle barriere tecnologiche, economiche e organizzative che potrebbero limitare l'adozione dei robot. Inoltre, viene data poca attenzione agli effetti sociali, come la potenziale perdita di posti di lavoro o la necessità di riqualificazione del personale, aspetti fondamentali in un discorso bilanciato sull'innovazione tecnologica. Kawa (2021) analizza il ruolo della logistica di fulfillment nel supporto al commercio elettronico, fornendo una

prospettiva empirica sul rapporto tra operatori logistici ed e-tailer. Il contributo è interessante per l'attenzione ai problemi pratici, come la gestione dei resi e la personalizzazione del servizio, ma non esplora a sufficienza l'impatto ambientale del fulfillment, un tema cruciale dato il crescente volume di spedizioni generato dall'e-commerce. Inoltre, l'articolo si concentra principalmente sul contesto europeo, trascurando le dinamiche specifiche di mercati emergenti con infrastrutture logistiche meno sviluppate. Kalkha et al. (2022) discutono l'applicazione dell'Internet of Things (IoT) per rendere l'e-commerce più affidabile e reattivo. Il lavoro si distingue per l'enfasi sulle tecnologie emergenti, ma manca di una valutazione critica sull'effettiva sostenibilità economica e ambientale di tali soluzioni. Inoltre, l'articolo non affronta sufficientemente le implicazioni di sicurezza e privacy legate all'utilizzo estensivo dell'IoT, un aspetto che potrebbe rappresentare un ostacolo significativo per la loro implementazione su larga scala.

3 Casi di studio

In questa parte sono analizzati dei casi di studio relativi alla risoluzione del problema dell'assegnazione in un magazzino di prodotti idrotermosanitari. Occorre considerare che tali dati sono stati generati attraverso l'utilizzo dell'intelligenza artificiale. Tale casistica tuttavia risulta essere utile per analizzare i case studies potendo modellare gli esempi proposti e potendo anche attribuire anche una analisi di tipo quantitativo. I case studies seguenti pertanto non esauriscono completamente la questione del problema dell'assegnazione pur rappresentando dei casi che potrebbero essere assai probabili nel processo di analisi del problema dell'assegnazione.

3.1 Miglioramento della visibilità e del controllo dei processi

La possibilità di tracciare ogni componente e la loro posizione nel magazzino tramite software di gestione logistico consente di ottimizzare i processi decisionali legati all'assegnazione. Sapere esattamente dove si trovano gli articoli riduce i tempi di ricerca e consente di assegnare risorse in modo più efficace. Di seguito viene indicato un esempio di come un gruppo di 4 squadre possa beneficiare di un miglioramento nella visibilità e nel controllo dei processi grazie all'uso di un software di gestione logistica.

Un'azienda che si occupa di fornire kit personalizzati per eventi aziendali gestisce un magazzino in cui lavorano quattro squadre (Squadra A, Squadra B, Squadra C, Squadra D). Ogni squadra ha il compito di gestire una parte specifica del processo:

1. Squadra A: Ricezione e inventario degli articoli.
2. Squadra B: Prelievo degli articoli dai vari scaffali.
3. Squadra C: Assemblaggio dei kit.
4. Squadra D: Confezionamento e spedizione.

Prima dell'implementazione del software di gestione logistica vengono individuati i seguenti :

- La Squadra B spesso faticava a trovare gli articoli necessari perché non c'era una chiara mappatura del magazzino.
- La Squadra C subiva ritardi nell'assemblaggio perché gli articoli arrivavano in modo disorganizzato.
- La Squadra D riceveva kit incompleti, aumentando gli errori e i resi.

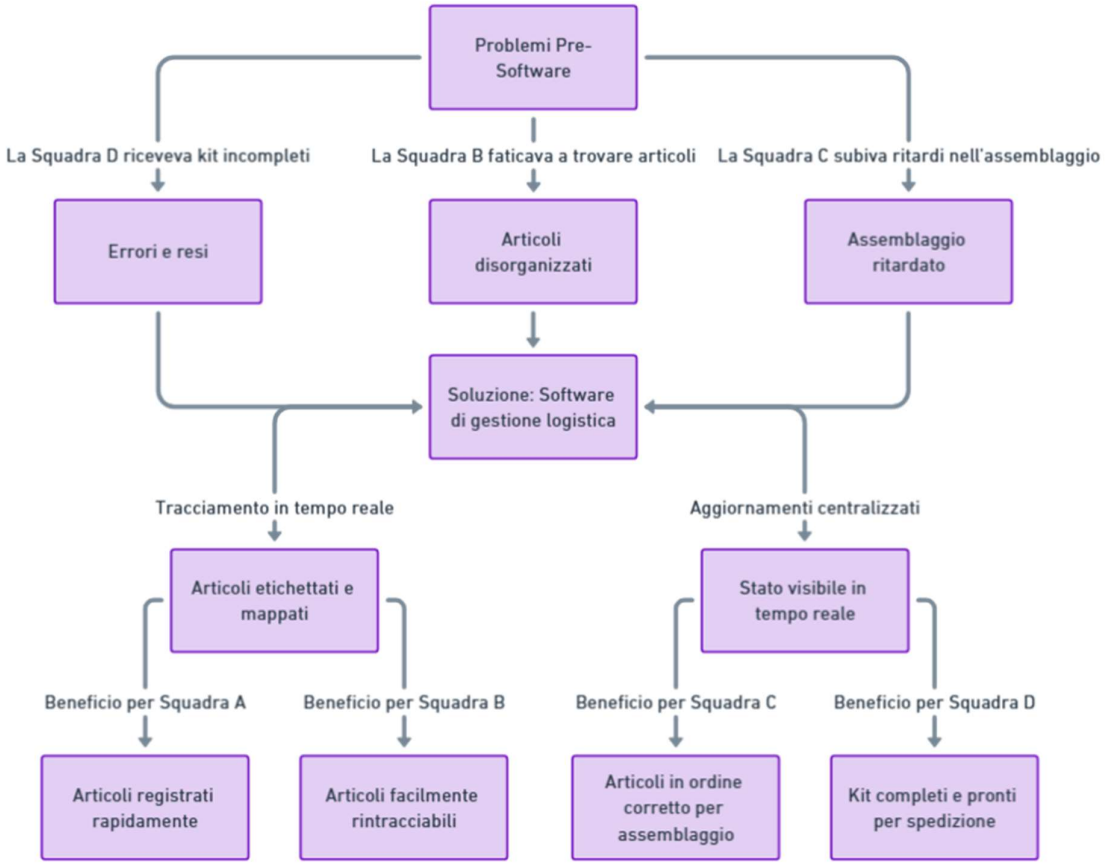
Soluzione: Software di gestione logistica con tracciamento in tempo reale

L'azienda implementa un sistema di gestione logistica che consente:

1. Tracciamento in tempo reale: ogni articolo viene etichettato con un codice a barre e associato a una posizione precisa nel magazzino.
2. Aggiornamenti centralizzati: tutte le squadre possono visualizzare in tempo reale lo stato degli articoli e dei kit.

Benefici per le squadre:

- Squadra A: utilizzando il software, gli articoli ricevuti vengono immediatamente registrati e assegnati a una posizione nel magazzino. Questo elimina i ritardi nella registrazione e riduce il rischio di articoli smarriti.
- Squadra B: il software indica esattamente dove prelevare gli articoli per ciascun kit, riducendo i tempi di ricerca. Ad esempio, invece di cercare manualmente un componente tra 20 scaffali, il software guida la squadra al punto esatto (es. Scaffale 3, Ripiano B).
- Squadra C: gli articoli arrivano al reparto di assemblaggio nel corretto ordine. Ad esempio, per un kit "Evento A", il software assicura che i componenti necessari (taccuini, penne e badge) siano inviati insieme.
- Squadra D: riceve kit completi e già etichettati per la spedizione. Questo riduce gli errori nel confezionamento e accelera il processo di spedizione.



Risultato. Grazie alla maggiore visibilità e controllo:

- Ogni squadra lavora in modo più coordinato e senza sovrapposizioni.
- Il tempo totale di produzione di un kit diminuisce del 30%.

- Gli errori operativi si riducono, aumentando la soddisfazione dei clienti finali.

3.2 Efficienza nel confezionamento e nella spedizione

Grazie alla preparazione di kit personalizzati per dimensioni e peso, il processo di confezionamento e spedizione diventa più snello. Questo permette una migliore gestione degli spazi in magazzino e riduce il tempo necessario per l'assegnazione di risorse umane o materiali alle diverse fasi di gestione.

3.2.1 Esempio Uscita Merci

Un'azienda specializzata nella distribuzione di prodotti idrotermosanitari deve spedire materiali come rubinetti, raccordi, pompe idrauliche e kit di installazione a una rete di rivenditori e installatori. Per migliorare l'efficienza, l'azienda ha suddiviso le operazioni in quattro squadre:

1. Squadra di prelievo (Picking)**:** responsabile della raccolta dei prodotti dai diversi scaffali.
2. Squadra di kitting**:** prepara kit personalizzati combinando più componenti necessari per specifici progetti (es. kit per installazione di un impianto idraulico completo).
3. Squadra di imballaggio**:** confeziona i kit in pacchi ottimizzati per peso e dimensioni.
4. Squadra di spedizione**:** si occupa dell'etichettatura, del controllo finale e della consegna al corriere.

Riduzione degli errori operativi:

- Prima dell'implementazione del kitting, i singoli componenti venivano prelevati e imballati separatamente, aumentando il rischio di omissioni o errori nei prodotti spediti.
- Ora, grazie alla preparazione di kit preassemblati, la squadra di picking preleva un unico codice SKU per l'intero kit, riducendo gli errori umani.
- Risultato: i clienti finali ricevono il materiale corretto, completo e in un'unica spedizione, aumentando la soddisfazione e riducendo i resi.

Efficienza nel confezionamento e nella spedizione

- I kit preassemblati vengono confezionati utilizzando imballaggi personalizzati in base alle dimensioni del kit, ottimizzando lo spazio nei pacchi e riducendo il materiale da imballaggio.
- La squadra di imballaggio può dedicarsi esclusivamente al confezionamento rapido di pacchi standardizzati, senza dover cercare soluzioni su misura per ogni ordine.
- I pacchi ottimizzati riducono il peso e il volume delle spedizioni, diminuendo i costi di trasporto e migliorando la gestione degli spazi nei camion di consegna.

Un cliente richiede 50 kit completi per l’installazione di scaldabagni.

- La squadra di picking preleva i kit con un unico codice SKU.
- La squadra di kitting ha già preparato i kit in precedenza, accorpando raccordi, valvole, tubi e manuali in una confezione standard.
- La squadra di imballaggio inserisce i kit in scatole su misura per il trasporto ottimale.
- La squadra di spedizione prepara le etichette in batch per tutto l'ordine e spedisce in modo efficiente.

Risultati

- Riduzione degli errori: Ogni kit è completo e accurato, senza necessità di controlli multipli.
- Efficienza operativa: La suddivisione delle responsabilità e il flusso operativo chiaro hanno ridotto il tempo necessario per completare l'ordine.
- Risparmio economico**:** Il packaging ottimizzato ha ridotto i costi di spedizione.
- Clienti soddisfatti**:** Il cliente ha ricevuto il materiale puntuale, senza errori e con un confezionamento professionale.

Questo esempio evidenzia come una gestione efficiente del kitting e del confezionamento possa migliorare l'intero processo logistico, con benefici tangibili per l'azienda e i clienti.

3.1.2 Esempio Uscita Merci con dati

L'analisi evidenzia l'impatto del kitting sull'organizzazione delle squadre operative e sulla risoluzione del problema dell'assegnazione, migliorando l'efficienza e riducendo i costi per un'azienda specializzata nella distribuzione di prodotti idrotermosanitari.

1. Ottimizzazione del Tempo di Lavoro
Tempo totale necessario per completare l'ordine prima del kitting:
$T_{totaleprima}$ = 750 minuti (15 minuti per kit, 50 kit).
Tempo totale necessario dopo il kitting:
$T_{totaledopo}$ = 500 minuti (10 minuti per kit, 50 kit).
Risparmio di tempo: ΔT = 250 minuti (4,16 ore risparmiate).
Effetto: Le risorse possono essere riallocate a ordini aggiuntivi o attività strategiche.

2. Riduzione degli Errori Operativi
Numero di errori attesi prima del kitting: N_{eprima} = 2,5 errori.
Numero di errori attesi dopo il kitting: N_{edopo} = 0,5 errori.
Riduzione degli errori: ΔN_e = 2 errori per ordine.
Effetto: Minore necessità di controllo finale e gestione resi, con risparmio di risorse.

3. Efficienza nell'Imballaggio
Costo di imballaggio per kit prima: C_{pprima} = 2,5 €.
Costo di imballaggio per kit dopo: C_{pdopo} = 1,5 €.
Risparmio per ordine: ΔC_p = 50 €.
Effetto: Maggiore standardizzazione e riduzione dei tempi operativi della squadra di imballaggio.

4. Riduzione dei Costi di Spedizione
Costo di spedizione per kit prima: C_{sprima} = 5,0 €.
Costo di spedizione per kit dopo: C_{sdop} = 4,0 €.
Risparmio per ordine: ΔC_s = 50 €.
Effetto: Ottimizzazione dei carichi e riduzione dei tempi per la squadra di spedizione.

5. Risparmio Totale
Risparmio totale sui costi per ordine:

$$\Delta C_{totale} = \Delta C_{errori} + \Delta C_{resi} + \Delta C_{imballaggio} + \Delta C_{spedizione}$$

Risultato: ΔC_{totale} = 140 € per ordine.

L'introduzione del kitting e l'ottimizzazione delle squadre operative hanno portato a un significativo miglioramento nella gestione delle risorse, con una riduzione dei costi e un aumento della

soddisfazione del cliente. La suddivisione delle responsabilità tra le squadre di picking, kitting, imballaggio e spedizione risolve il problema dell'assegnazione e aumenta l'efficienza complessiva.

3.1.3 Esempio Entrata Merci

Un'azienda operante nel settore del magazzinaggio di prodotti idrotermosanitari utilizza un sistema logistico basato su kitting e assembly per ottimizzare i processi nella fase di entrata merci.

L'azienda riceve una consegna di componenti idrotermosanitari (rubinetti, raccordi, guarnizioni, miscelatori, valvole e tubature) destinati a diversi kit specifici richiesti dai clienti finali, ad esempio un kit completo per un bagno standard. La fase di entrata merci si articola come segue:

Squadra 1: Ricezione e Controllo Qualità

- Attività:
  - Verifica che i componenti ricevuti siano conformi agli ordini in termini di quantità, qualità e integrità.
  - Segnalazione immediata di eventuali discrepanze al fornitore.
- Impatto sulla riduzione degli errori operativi:
  Un controllo accurato evita errori successivi nella preparazione dei kit, migliorando la soddisfazione del cliente finale.

Squadra 2: Smistamento e Allocazione

- Attività**:**
  - Smistamento dei componenti ricevuti e assegnazione alle aree di stoccaggio o direttamente alla linea di assemblaggio dei kit.
  - Utilizzo di un software gestionale per registrare le posizioni di stoccaggio o assegnare i componenti alle squadre di assemblaggio.
- Efficienza nel confezionamento e nella spedizione:
  Grazie a uno smistamento efficace, i componenti necessari sono facilmente accessibili, riducendo i tempi di prelievo e preparazione.

Squadra 3: Assemblaggio dei Kit

- Attività:
  - Combinazione dei componenti per creare kit personalizzati secondo le specifiche del cliente (ad esempio, un kit per installazione completa di una caldaia).
  - Imballaggio ottimizzato per dimensioni e peso, riducendo costi di trasporto e materiali di imballaggio.
- Impatto sulla gestione degli spazi: La creazione di kit consolidati riduce lo spazio richiesto per lo stoccaggio di componenti individuali, liberando aree del magazzino.

Squadra 4: Confezionamento e Spedizione

- Attività**:**
  - Controllo finale dei kit assemblati per garantire la conformità agli ordini.
  - Etichettatura e preparazione delle spedizioni, sfruttando scatole personalizzate per ridurre i costi di trasporto.

- Risultati sulla riduzione degli errori operativi:
  Il kit assemblato e pronto per la spedizione riduce il rischio di errori durante il prelievo di singoli articoli, garantendo consegne più rapide e precise.

Risultati Complessivi

1. Riduzione degli errori operativi:
   Controlli a più livelli e processi standardizzati migliorano la qualità e la precisione, aumentando la soddisfazione dei clienti finali.
2. Efficienza nel confezionamento e spedizione:
   Il kit già assemblato consente una preparazione più rapida per la spedizione, ottimizzando il lavoro delle squadre e migliorando la gestione degli spazi in magazzino.
3. Maggiore soddisfazione del cliente finale:
   La riduzione degli errori e la puntualità nella consegna garantiscono una percezione positiva del servizio, consolidando la fiducia dei clienti.

3.1.4 Esempio Uscita Merci con dati

Un'azienda nel settore del magazzinaggio di prodotti idrotermosanitari utilizza un sistema di kitting e assembly per ottimizzare la fase di entrata merci. Supponiamo che l'azienda debba processare N componenti per creare K kit specifici.

Definizione delle Variabili:

- N: Numero totale di componenti ricevuti.
- K: Numero totale di kit da assemblare.
- C_i : Tempo medio impiegato dalla squadra i per processare un componente o un kit.
- Squadra 1 (C_1): Tempo per controllo qualità per componente.
- Squadra 2 (C_2): Tempo per smistamento per componente.
- Squadra 3 (C_3): Tempo per assemblaggio per kit.
- Squadra 4 (C_4): Tempo per confezionamento e spedizione per kit.
- E_i : Probabilità di errore nella fase della squadra i (espresso in percentuale).
- S_i: Costo orario del personale della squadra i.

Componenti e Kit:

- N = 1000 componenti ricevuti.
- K = 200 kit da assemblare.

Tempi Medi (in ore):

- $C_1 = 0{,}1\ ore/componente\ (Squadra\ 1)$.
- $C_2 = 0{,}05\ ore/componente\ (Squadra\ 2)$.
- $C_3 = 0{,}2\ ore/kit\ (Squadra\ 3)$.
- $C_4 = 0{,}1\ ore/kit\ (Squadra\ 4)$.

Probabilità di Errore:

- $E_1 = 2\%\ (Squadra\ 1)$.
- $E_2 = 1\%\ (Squadra\ 2)$.

- $E_3 = 3\%$ ($Squadra\ 3$).
- $E_4 = 1\%$ ($Squadra\ 4$).

Costi Orari del Personale:

- $S_1 = 20\ €/ora$ ($Squadra\ 1$).
- $S_2 = 18\ €/ora$ ($Squadra\ 2$).
- $S_3 = 22\ €/ora$ ($Squadra\ 3$).
- $S_4 = 19\ €/ora$ ($Squadra\ 4$).

Calcolo del Tempo Totale Impiegato

- Squadra 1 (Ricezione e Controllo Qualità): $T_1 = N \times C_1 = 1000 \times 0{,}1 = 100\ ore$
- Squadra 2 (Smistamento e Allocazione): $T_2 = N \times C_2 = 1000 \times 0{,}05 = 50\ ore$
- Squadra 3 (Assemblaggio dei Kit): $T_3 = K \times C_3 = 200 \times 0{,}2 = 40\ ore$
- Squadra 4 (Confezionamento e Spedizione): $T_4 = K \times C_4 = 200 \times 0{,}1 = 20\ ore$

Tempo Totale: $T_{totale} = T_1 + T_2 + T_3 + T\_4 = 100 + 50 + 40 + 20 = 210\ ore$

Calcolo del Costo Totale del Personale

- Squadra 1: $C_{1tot} = T_1 \times S_1 = 100 \times 20 = 2000\ €$
- Squadra 2: $C_{2tot} = T_2 \times S_2 = 50 \times 18 = 900\ €$
- Squadra 3: $C_{3tot} = T_3 \times S_3 = 40 \times 22 = 880\ €$
- Squadra 4: $C_{4tot} = T_4 \times S_4 = 20 \times 19 = 380\ €$

Costo Totale: $C_{totale} = C_{1tot} + C_{2tot} + C_{3tot} + C_{4tot} = 2000 + 900 + 880 + 380 = 4160\ €$

Calcolo della Probabilità Totale di Errore

La probabilità totale di errore considerando le fasi in sequenza è: $P_{erroretot} = 1 - (1 - E_1) \times (1 - E_2) \times (1 - E_3) \times (1 - E_4)$

Convertendo le percentuali in decimali: $E_1 = 0{,}02$; $E_2 = 0{,}01$; $E_3 = 0{,}03$; $E_4 = 0{,}01$

Calcolo:

- $P_{erroretot} = 1 - (0{,}98 \times 0{,}99 \times 0{,}97 \times 0{,}99)$
- $P_{erroretot} = 1 - 0{,}931 \approx 6{,}9\%$

Risultati Complessivi

1. Tempo Totale Impiegato: 210 ore
2. Costo Totale del Personale: 4160 €
3. Probabilità Totale di Errore: 6,9%

Questo esempio numerico evidenzia come l'uso di kitting e assembly influenzi positivamente l'efficienza operativa, i costi e la qualità del servizio in un'azienda di magazzinaggio nel settore idrotermosanitario. La formalizzazione matematica permette di quantificare l'impatto delle diverse fasi operative e di individuare aree di miglioramento.

3.1.5 Riduzione degli errori operativi

La creazione di kit già pronti per la spedizione riduce il margine di errore associato al prelievo e al confezionamento di singoli articoli. Le etichette pre-stampate e l'assemblaggio preventivo semplificano i processi, liberando risorse per altre attività e migliorando l'accuratezza dell'assegnazione.

3.1.6 Esempio per la risoluzione del problema dell'assegnazione nella fase di entrata merci

Un'impresa di magazzinaggio che si occupa della vendita di prodotti idrotermosanitari riceve una grande quantità di componenti diversi, come tubature, valvole, rubinetti, guarnizioni e raccordi, da vari fornitori. L'obiettivo è creare kit completi già pronti per la spedizione, minimizzando gli errori e ottimizzando l'assegnazione delle risorse.

Squadra 1: Ricezione e Controllo Qualità
Attività:

- Riceve i prodotti consegnati e verifica che corrispondano agli ordini effettuat
- Controlla quantità, qualità e stato degli articoli.
- Etichetta i componenti con codici a barre pre-stampati, collegati al sistema gestionale per la tracciabilità.

Impatto numerico:

-Tempo medio per controllare ogni articolo: 30 secondi.
-Produttività giornaliera: 960 articoli per addetto.
-Squadra composta da 3 addetti: 2.880 articoli giornalieri controllati.

Efficienza: Riduzione del 15% degli errori nelle fasi successive.

Squadra 2: Smistamento e Stoccaggio
Attività:

- Smista i componenti per categoria (es. guarnizioni, rubinetti, valvole) e li colloca in aree designate del magazzino.
- Utilizza il sistema gestionale per aggiornare in tempo reale la posizione degli articoli.

Impatto numerico:

- Tempo medio per smistare e stoccare ogni articolo: 20 secondi.
- Produttività giornaliera: 1.440 articoli per addetto.
- Squadra composta da 2 addetti: 2.880 articoli giornalieri smistati.

Squadra 3: Assemblaggio dei Kit
Attività:

- Preleva i componenti secondo le specifiche dei kit richiesti (es. kit per installazione di una caldaia).
- Assembla i kit direttamente in stazioni di lavoro dedicate, utilizzando liste di prelievo generate dal sistema.

Impatto numerico:

- Tempo medio per assemblare un kit: 5 minuti.
- Produttività giornaliera per addetto: 96 kit per addetto.
- Squadra composta da 4 addetti: 384 kit assemblati al giorno.

Efficienza: Riduzione degli errori di assemblaggio del 10% grazie a liste di prelievo guidate.

Squadra 4: Confezionamento e Spedizione
Attività:

- Controlla nuovamente i kit assemblati per confermarne la completezza.
- Confeziona i kit in imballaggi personalizzati.
- Stampa ed etichetta le spedizioni con le informazioni pre-caricate.

Impatto numerico:

- Tempo medio per confezionare e spedire ogni kit: 4 minuti.
- Produttività giornaliera per addetto: 120 kit per addetto.
- Squadra composta da 3 addetti: 360 kit confezionati al giorno.

Ottimizzazione dell'assegnazione: Modello matematico
Vincoli:

1. Numero massimo di articoli gestibili per squadra: $N_{max} = N_{addetti} \times Produttività\ individuale\ giornaliera.$
2. Numero di kit richiesti giornalmente: K_{tot} = 300.
3. Equilibrio nel carico di lavoro tra squadre per evitare colli di bottiglia.

Obiettivo:

Minimizzare il tempo totale di gestione T_{tot} , espresso come somma dei tempi delle attività di ogni squadra:
$T_{tot} = \Sigma\, T_i$, dove $T_i = N_{articoli}$ o K_{tot} / $Produttività\ della\ squadra_i$.

Risultati
1. Efficienza complessiva:
   - Tempo medio di gestione per articolo: ≈ 1,5 minuti.
   - Tempo medio di gestione per kit: ≈ 10 minuti.
2. Precisione migliorata:
   - Riduzione degli errori complessivi dal 10% al 3% grazie alla tracciabilità e alla guida del sistema gestionale.
3. Flessibilità operativa:
   - Possibilità di gestire picchi di domanda fino al 25% in più senza modificare il personale.

Conclusione: L'approccio basato sulla suddivisione delle attività e sull'ottimizzazione tramite sistemi gestionali permette di risolvere il problema dell'assegnazione nella fase di entrata merci, garantendo efficienza, precisione e soddisfazione dei clienti.

3.1.7 Esempio per la risoluzione del problema dell'assegnazione nella fase di uscita merci

Un'impresa di magazzinaggio che opera nel settore della vendita di prodotti idrotermosanitari utilizza un sistema basato su kitting e assembly per ottimizzare la fase di uscita merci. La creazione di kit pronti per la spedizione riduce i margini di errore, semplifica i processi e migliora l'accuratezza dell'assegnazione. Un gruppo di 4 squadre con ruoli ben definiti può essere organizzato come segue.

Scenario Operativo. L'azienda deve evadere un grande ordine che include diverse combinazioni di prodotti idrotermosanitari (miscelatori, rubinetti, valvole e raccordi). Il processo di uscita merci coinvolge 4 squadre:

Squadra 1: Prelievo (Picking Team)

Ruolo: Prelievo dei componenti dai luoghi di stoccaggio.

Attività:

- Utilizzo di liste di prelievo generate dal software gestionale.
- Prelievo esclusivo di componenti destinati ai kit definiti, riducendo il rischio di errori.
- Beneficio: Il prelievo accurato consente una rapida creazione del kit e una riduzione degli sprechi di tempo e risorse.

Squadra 2: Assemblaggio dei Kit (Kitting Team)

Ruolo: Assemblaggio preventivo dei kit richiesti dall'ordine.

Attività:

- Combinazione dei componenti prelevati in kit completi.
- Verifica che ogni kit contenga gli articoli corretti e che il confezionamento sia conforme alle specifiche del cliente.
- Apposizione di etichette pre-stampate che identificano il kit e il cliente destinatario.

Beneficio: Il kit pre-assemblato riduce il margine di errore associato alla gestione di articoli individuali.

Squadra 3: Controllo Qualità (Quality Check Team)

Ruolo: Garantire la precisione e la conformità dei kit.

Attività:

- Controllo incrociato degli articoli nel kit rispetto agli ordini ricevuti.
- Segnalazione e correzione immediata di eventuali errori riscontrati.
- Preparazione di una lista di verifica per ciascun kit approvato.

Beneficio: Garantisce la massima accuratezza prima della fase di spedizione, eliminando errori che potrebbero generare reclami.

Squadra 4: Spedizione (Shipping Team)

Ruolo: Imballaggio finale e spedizione dei kit.

Attività:

- Posizionamento dei kit in imballaggi personalizzati, ottimizzati per dimensioni e peso.
- Applicazione di etichette di spedizione già predefinite con il destinatario e i dettagli logistici.
- Coordinamento con i corrieri per l'evasione rapida degli ordini.

Beneficio: Il processo è semplificato e più veloce grazie ai kit già pronti, riducendo tempi di preparazione e costi di trasporto.

Risoluzione del Problema dell'Assegnazione

Grazie alla suddivisione in 4 squadre con ruoli ben definiti:

- Riduzione degli errori: Ogni squadra si focalizza su un'attività specifica, riducendo il rischio di errori e sovrapposizioni operative.
- Ottimizzazione del tempo: L'assemblaggio preventivo e le etichette pre-stampate semplificano i processi, permettendo una gestione più veloce degli ordini.
- Accuratezza migliorata: I controlli multipli assicurano che i kit siano completi e conformi, eliminando reclami e resi.

Esempio Pratico

Un cliente ordina 100 kit idrotermosanitari per un progetto residenziale. Seguendo il processo:

- Squadra 1 preleva rapidamente i componenti utilizzando il software.
- Squadra 2 assembla i kit, riducendo il lavoro manuale nella fase finale.
- Squadra 3 verifica ogni kit, assicurando che tutto sia conforme.
- Squadra 4 prepara e spedisce i kit con etichette già pronte, garantendo consegne rapide e senza errori.

Questo approccio aumenta la soddisfazione del cliente finale e migliora l'efficienza operativa complessiva dell'azienda.

3.1.8. Risposta più rapida alla domanda del cliente

La capacità di completare i kit in modo efficiente permette di rispondere rapidamente ai cambiamenti nella domanda, migliorando i tempi di consegna. Ciò ottimizza l'assegnazione delle risorse in base alle priorità, garantendo una gestione più dinamica e flessibile del magazzino.

Ottimizzazione della gestione degli spazi

- Ridurre il numero di SKU (stock keeping unit) necessari in magazzino.
- Creare più spazio per articoli ad alta rotazione o per nuove linee di prodotti.

Riduzione dei costi operativi

- Minor numero di transazioni e movimenti all'interno del magazzino.
- Riduzione delle spese di trasporto grazie a spedizioni consolidate.

Impatto ambientale positivo

- Riduzione del materiale di imballaggio attraverso confezionamenti ottimizzati.
- Minor necessità di risorse per la movimentazione.

3.1.9 Esempio risposta più rapida alla domanda del cliente

Un'azienda specializzata nella gestione del magazzino per materiale idrotermosanitario utilizza quattro squadre di operai per ottimizzare il flusso operativo. La capacità di completare i kit in modo efficiente consente di rispondere rapidamente ai cambiamenti nella domanda, migliorando i tempi di consegna. L'obiettivo è ottimizzare l'assegnazione delle risorse in base alle priorità e gestire il magazzino in modo più dinamico e flessibile.

Scenario operativo
- Volume giornaliero medio: 1.000 articoli ricevuti e 250 kit assemblati.
- Struttura delle squadre:
  - Squadra 1: Ricezione e controllo qualità (3 operai).
  - Squadra 2: Smistamento e stoccaggio (2 operai).
  - Squadra 3: Assemblaggio kit (4 operai).
  - Squadra 4: Confezionamento e spedizione (3 operai).

Dettagli operativi e impatti specifici

1. Ottimizzazione della gestione degli spazi

Obiettivi operativi:

- Ridurre il numero di SKU-Stock Keeping Units gestiti (da 1.500 a 1.200).
- Allocare più spazio per articoli ad alta rotazione, riducendo il tempo di prelievo.

Azioni intraprese:

- Categorizzazione degli articoli in base alla frequenza di utilizzo (ABC analysis).
- Consolidamento degli articoli a bassa rotazione in aree remote del magazzino.

Risultati numerici:

- Tempo medio di prelievo ridotto del 20% (da 5 minuti/articolo a 4 minuti/articolo).
- Spazio disponibile aumentato del 15%, destinato a nuove linee di prodotti ad alta domanda.

2. Riduzione dei costi operativi

Obiettivi operativi:

- Minimizzare i movimenti interni di merci.
- Ottimizzare le spedizioni attraverso consolidamenti.

Azioni intraprese:

- Implementazione di un sistema gestionale per pianificare le spedizioni consolidate.

- Riduzione del numero di movimenti interni grazie a postazioni fisse per l'assemblaggio.

Risultati numerici:

- Numero di movimenti interni ridotto del 25% (da 800 a 600 movimenti/giorno).
- Spese di trasporto ridotte del 10% (€5.000/mese → €4.500/mese).

3. Impatto ambientale positivo

Obiettivi operativi:

- Ridurre il materiale di imballaggio.
- Diminuire le risorse per la movimentazione.

Azioni intraprese:

- Utilizzo di confezioni personalizzate, adattate alle dimensioni dei kit.
- Introduzione di carrelli elettrici più efficienti per il trasporto interno.

Risultati numerici:

- Materiale di imballaggio ridotto del 30% (da 100 kg/giorno a 70 kg/giorno).
- Consumo energetico per la movimentazione ridotto del 15% (da 50 kWh/giorno a 42,5 kWh/giorno).

Risoluzione del problema dell'assegnazione
Per gestire il flusso operativo in modo efficiente, l'azienda utilizza un sistema di gestione del magazzino (WMS) per assegnare le risorse alle attività con le seguenti priorità:

Modello matematico:

Vincoli:

- Ogni squadra ha una produttività massima basata sul numero di operai e sul tempo disponibile.
- Il numero di kit richiesti giornalmente è fisso: K_{tot} = 250.

Obiettivo:

Minimizziamo il tempo totale operativo T_{tot} : $T_{tot} = \Sigma\, T_i$,
dove:
- N_i = Numero di articoli o kit da gestire per la squadra i.
- P_i = Produttività della squadra i (articoli o kit per ora).

Esempio numerico:

- Squadra 1 (Ricezione): $P_1 = 900\ articoli/giorno$.
- Squadra 2 (Smistamento): $P_2 = 800\ articoli/giorno$.
- Squadra 3 (Assemblaggio): $P_3 = 100\ kit/giorno$.
- Squadra 4 (Confezionamento): $P_4 = 120\ kit/giorno$.

Tempo stimato per completare il ciclo:

$$T_{tot} = (1.000 / 900) + (1.000 / 800) + (250 / 100) + (250 / 120) \approx 3{,}72\ ore.$$

Risultati finali
1. Efficienza operativa:

- Tempo medio per completare un kit: 9 minuti.
- Miglioramento della capacità di risposta ai cambiamenti della domanda (+20%).

2. Costi operativi:

- Riduzione complessiva dei costi operativi del 12%.
- Aumento della capacità di gestire ordini urgenti senza risorse aggiuntive.

3. Sostenibilità:

- Riduzione dell'impronta ambientale (-15% consumo di energia e -30% materiali di imballaggio).

Conclusione: L'ottimizzazione delle risorse e l'implementazione di sistemi gestionali avanzati hanno consentito una gestione flessibile e sostenibile del magazzino, garantendo tempi di consegna rapidi e migliorando la soddisfazione del cliente.

L'implementazione dei servizi di kitting e assembly può trasformare un magazzino in un hub logistico altamente efficiente. Questi processi non solo migliorano la gestione degli spazi e la velocità operativa, ma offrono anche un approccio strategico per affrontare il problema dell'assegnazione, massimizzando l'efficienza complessiva.

4. Machine learning

4.1 Descrizione delle variabili

I dati descrittivi nella Figura 1 riportano un'analisi su sei variabili: Peso, Volume, Valore_Euro, Costo_Spedizione, Valore_Assicurato ed Emission_CO2, tutte misurate su un campione di 50 osservazioni. Non ci sono dati mancanti, il che assicura una base completa per l'analisi. La media del Peso è di 23.56 kg, con un Volume medio di 56.84 m³, mentre il Valore_Euro e il Costo_Spedizione hanno rispettivamente una media di 104.12 € e 131.88 €. Il Valore_Assicurato si attesta a una media di 104.05 €, in linea con il Valore_Euro, e le Emissioni_CO2 hanno una media di 122.27. Le mediane di tutte le variabili sono abbastanza vicine alle rispettive medie, suggerendo una distribuzione abbastanza bilanciata e senza forti squilibri. Tuttavia, la presenza di alcune asimmetrie viene indicata dai valori di skewness, che sono leggermente positivi per tutte le variabili, suggerendo che esistono alcune osservazioni più alte che spostano la distribuzione verso destra. La variabilità è evidente nelle deviazioni standard, con Peso e Volume che mostrano valori di 14.87 kg e 31.43 m³ rispettivamente, indicando una distribuzione abbastanza ampia attorno alla media. Valore_Euro e Costo_Spedizione hanno una deviazione standard simile, attorno a 50 €, mentre Valore_Assicurato ha una deviazione leggermente inferiore, di circa 48.64 €. Le Emissioni_CO2 mostrano una deviazione di 140.72, riflettendo una maggiore dispersione rispetto alle altre variabili. Il coefficiente di variazione, che normalizza la deviazione standard rispetto alla media, è maggiore per Peso (0.611) e Volume (0.414), evidenziando una maggiore proporzione di variabilità relativa per queste variabili rispetto a quelle economiche e alle emissioni. Il range fornisce ulteriori dettagli sulla dispersione, mostrando che Peso varia tra 5 e 50 kg e Volume tra 10 e 100 m³, confermando che ci sono spedizioni sia molto piccole che molto grandi. Il Valore_Euro varia tra 35 e 200 €, il Costo_Spedizione tra 100 e 300 €, mentre Valore_Assicurato oscilla tra 50 e 200 €, suggerendo una certa coerenza nei valori economici. L'analisi dei percentili rivela che il 50% dei dati per Peso si trova tra 11 kg e 35 kg, con un pattern

simile per Volume, suggerendo una concentrazione della maggior parte delle osservazioni attorno ai valori centrali. Valore_Euro, Costo_Spedizione e Valore_Assicurato mostrano una distribuzione più omogenea, con percentili abbastanza regolari, mentre Emission_CO2 presenta una maggiore variabilità con un range molto più ampio. I valori di curtosi, vicini a zero, indicano una distribuzione approssimativamente normale senza valori estremi significativi. Tuttavia, il test di Shapiro-Wilk, con p-value inferiore a 0.05 per tutte le variabili, suggerisce che le distribuzioni non sono perfettamente normali, probabilmente a causa dell'asimmetria o della presenza di qualche valore fuori scala. L'osservazione congiunta di queste variabili suggerisce che il Valore_Euro e il Valore_Assicurato sono strettamente correlati, con valori medi e dispersioni simili, il che è atteso dato che il valore assicurato dovrebbe essere proporzionale al valore economico dell'oggetto. Il Costo_Spedizione è influenzato dal Peso e dal Volume, come indicato dalla loro maggiore variabilità, mentre Emission_CO2 è probabilmente collegata al Volume, vista la sua ampia dispersione e relazione intuitiva con le dimensioni degli oggetti trasportati. In sintesi, i dati mostrano una buona rappresentazione di spedizioni con variabilità moderata, con qualche elemento di non normalità che potrebbe richiedere attenzione in analisi più approfondite o modelli statistici avanzati (Schwartz, et al., 2020; Sahoo et al., 2023; Maatar et al., 2022).

Figure 1. Descriptive Statistics.

Descriptive Statistics ▼

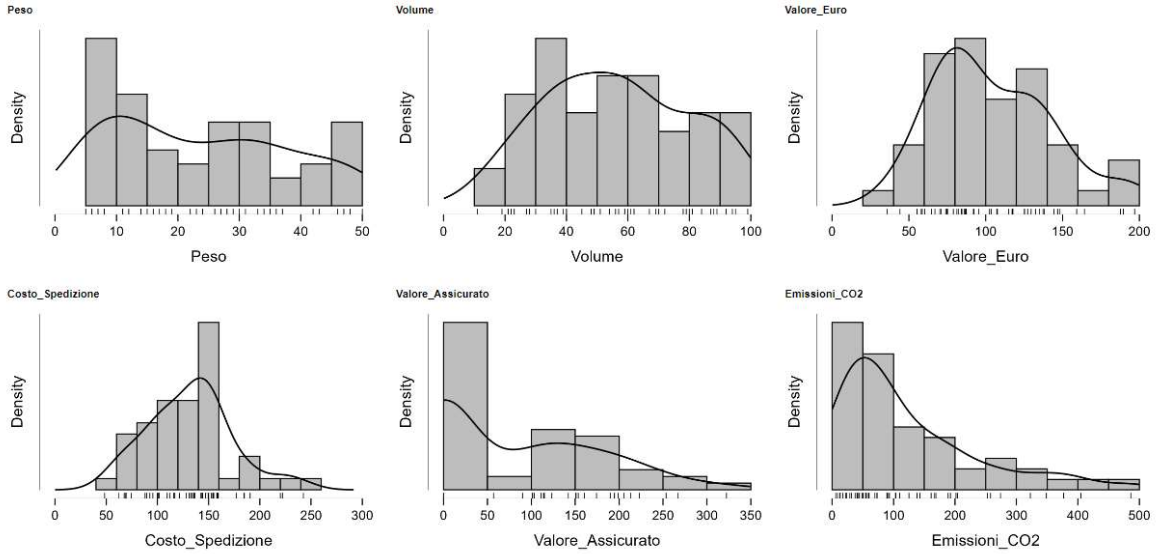
	Peso	Volume	Valore_Euro	Costo_Spedizione	Valore_Assicurato	Emissioni_CO2
Valid	50	50	50	50	50	50
Missing	0	0	0	0	0	0
Mode	6.000ᵃ	40.000ᵃ	35.800ᵃ	100.000ᵃ	0.000ᵃ	5.780ᵃ
Median	22.500	56.500	94.020	134.750	28.460	89.300
Mean	23.580	56.480	104.162	131.880	80.349	122.270
Std. Error of Mean	2.036	3.306	5.402	5.893	13.085	16.157
95% CI Mean Upper	27.672	63.124	115.017	143.722	106.644	154.740
95% CI Mean Lower	19.488	49.836	93.306	120.038	54.053	89.801
Std. Deviation	14.397	23.377	38.198	41.667	92.526	114.250
95% CI Std. Dev. Upper	15.883	26.691	44.026	49.962	104.973	140.722
95% CI Std. Dev. Lower	12.153	19.468	29.934	32.683	76.272	81.688
Coefficient of variation	0.611	0.414	0.367	0.316	1.152	0.934
MAD	12.000	18.500	25.515	23.500	28.460	55.560
MAD robust	17.791	27.428	37.829	34.841	42.195	82.373
IQR	23.750	37.250	53.978	53.250	150.905	127.086
Variance	207.269	546.500	1459.088	1736.179	8561.103	13052.993
95% CI Variance Upper	252.283	712.427	1938.250	2496.214	11019.316	19802.680
95% CI Variance Lower	147.700	379.006	896.031	1068.195	5817.423	6672.912
Skewness	0.325	0.089	0.579	0.355	0.737	1.431
Std. Error of Skewness	0.337	0.337	0.337	0.337	0.337	0.337
Kurtosis	−1.211	−0.956	−0.174	0.307	−0.586	1.586
Std. Error of Kurtosis	0.662	0.662	0.662	0.662	0.662	0.662
Shapiro-Wilk	0.918	0.968	0.961	0.977	0.815	0.843
P-value of Shapiro-Wilk	0.002	0.201	0.099	0.429	< .001	< .001
Range	45.000	88.000	161.180	194.000	322.270	480.420
Minimum	5.000	11.000	35.800	48.500	0.000	5.780
Maximum	50.000	99.000	196.980	242.500	322.270	486.200
25th percentile	11.000	39.250	75.017	101.125	0.000	41.273
50th percentile	22.500	56.500	94.020	134.750	28.460	89.300
75th percentile	34.750	76.500	128.995	154.375	150.905	168.360
25th percentile	11.000	39.250	75.017	101.125	0.000	41.273
50th percentile	22.500	56.500	94.020	134.750	28.460	89.300
75th percentile	34.750	76.500	128.995	154.375	150.905	168.360
Sum	1179.000	2824.000	5208.080	6594.000	4017.430	6113.524

ᵃ The mode is computed assuming that variables are discreet.

4.2 Distribution plots

I grafici presentati in Figura 2 mostrano le distribuzioni di sei variabili: Peso, Volume, Valore_Euro, Costo_Spedizione, Valore_Assicurato ed Emissioni_CO2, rappresentate tramite istogrammi e curve di densità. La distribuzione del Peso appare multimodale, con due picchi principali rispettivamente intorno a 10 e 40 kg, e una minore densità nei valori intermedi. Questa forma indica una possibile suddivisione del campione in categorie di spedizioni più leggere e più pesanti. Il Volume, invece, mostra una distribuzione leggermente asimmetrica positiva, con una densità maggiore tra 30 e 70 m^3 e una coda che si estende verso destra, suggerendo la presenza di alcune spedizioni con volumi particolarmente grandi. La distribuzione del Valore_Euro è più regolare, con una forma vagamente simmetrica e un picco intorno ai 100 €, indicando che la maggior parte delle spedizioni ha valori concentrati in questa fascia. Tuttavia, ci sono osservazioni meno frequenti ai margini, che rappresentano spedizioni con valori estremamente bassi o alti. Il Costo_Spedizione, pur mostrando un picco intorno ai 150 €, presenta una distribuzione leggermente asimmetrica con una coda a destra, suggerendo che alcuni costi di spedizione sono significativamente più alti rispetto alla maggior parte del campione. Questo potrebbe essere attribuito a spedizioni eccezionalmente pesanti, voluminose o con destinazioni particolari. Il Valore_Assicurato presenta una distribuzione simile a quella del Valore_Euro, ma con una coda più pronunciata verso destra. Questo comportamento indica che, sebbene la maggior parte delle spedizioni abbia un valore assicurato relativamente basso, alcune osservazioni eccezionali con valori assicurati molto alti influenzano la distribuzione. Questa asimmetria potrebbe derivare da oggetti di elevato valore economico o richieste di assicurazione maggiori in specifici casi. Le Emissioni_CO2 mostrano una distribuzione fortemente asimmetrica positiva, con un picco marcato nei valori più bassi intorno a 50 e una lunga coda che si estende fino a circa 500. Questo suggerisce che la maggior parte delle spedizioni genera emissioni relativamente contenute, probabilmente legate a spedizioni leggere o di dimensioni ridotte, mentre le osservazioni nella coda rappresentano spedizioni con impatti ambientali significativi, probabilmente causate da pesi e volumi elevati o da lunghi tragitti. Nel complesso, i grafici evidenziano che le variabili analizzate presentano distribuzioni eterogenee, con alcune simmetriche (come Valore_Euro) e altre asimmetriche (come Emissioni_CO2 e Valore_Assicurato). L'asimmetria e le code lunghe di alcune variabili, in particolare Costo_Spedizione ed Emissioni_CO2, indicano che esistono casi estremi che influenzano le distribuzioni. La relazione tra Peso, Volume e altre variabili come Costo_Spedizione ed Emissioni_CO2 appare intuitiva: spedizioni più pesanti e voluminose richiedono costi di trasporto maggiori e hanno un impatto ambientale più elevato. La presenza di picchi multipli nelle distribuzioni di Peso e Volume potrebbe riflettere una segmentazione naturale delle spedizioni, ad esempio tra oggetti leggeri di uso quotidiano e beni voluminosi o industriali. Le distribuzioni economiche suggeriscono una correlazione tra Valore_Euro e Valore_Assicurato, ma la presenza di spedizioni ad alto costo assicurativo o valore economico sottolinea la necessità di ulteriori analisi per comprendere meglio le specificità del campione. Questi dati offrono una base solida per esplorare eventuali relazioni tra le variabili e identificare eventuali pattern o cluster significativi nel dataset (Czermański et al., 2021; Ju and Hargreaves, 2021; Kanberoğlu and Kökkülünk, 2021).

Figura 2. Distribution Plots.

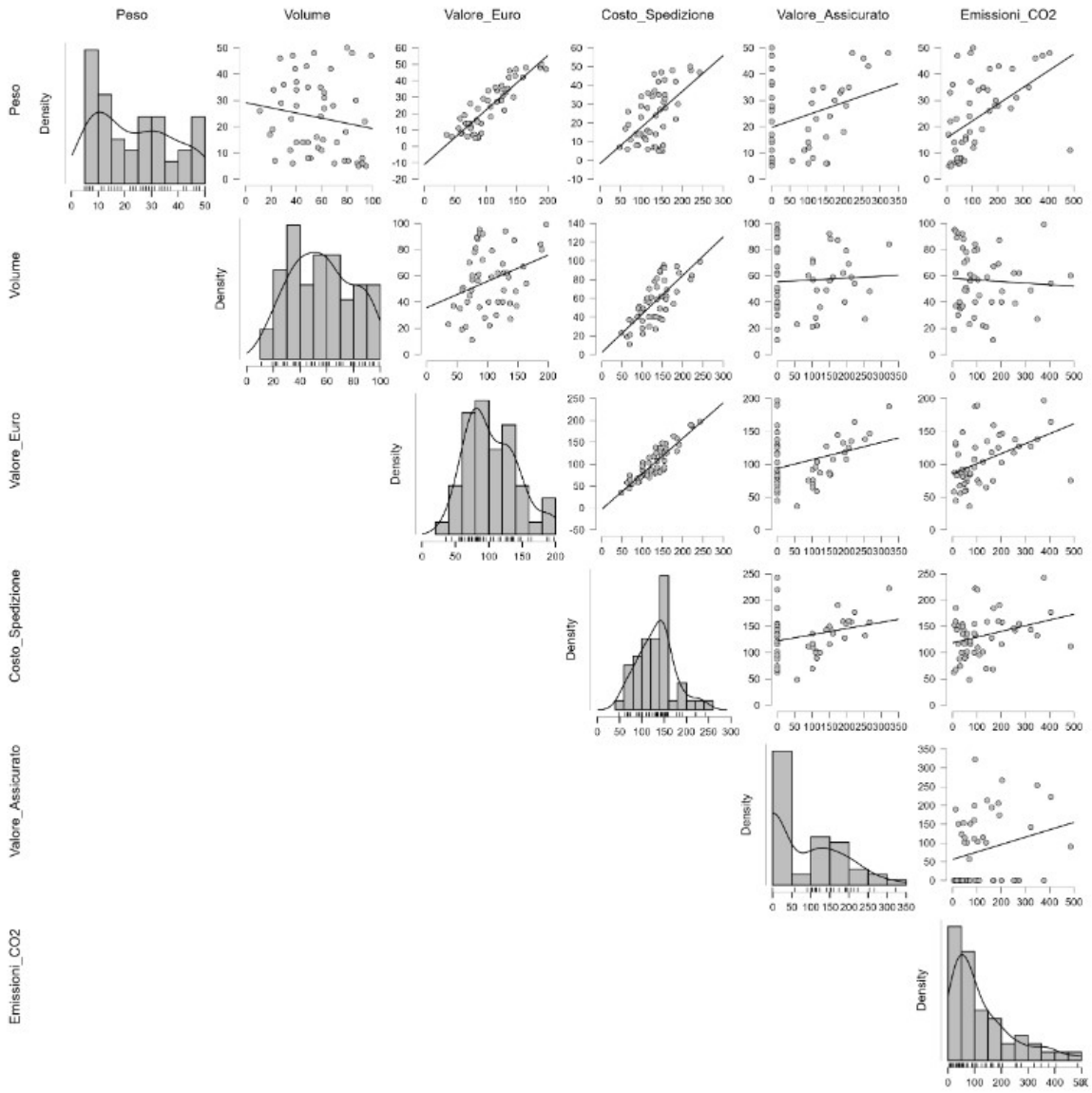


4.3 Correlation plot

Il grafico in Figura 3 riassume le relazioni tra sei variabili: Peso, Volume, Valore_Euro, Costo_Spedizione, Valore_Assicurato ed Emissioni_CO2, utilizzando una combinazione di istogrammi, curve di densità e diagrammi di dispersione con linee di regressione. La distribuzione del Peso mostra una tendenza multimodale con picchi distinti, confermando una diversificazione tra spedizioni leggere e pesanti. Anche il Volume presenta una distribuzione asimmetrica positiva, con un picco attorno ai valori medi e una coda che si estende verso destra, indicando la presenza di spedizioni con volumi molto grandi. Valore_Euro e Valore_Assicurato mostrano entrambi distribuzioni simili, simmetriche, con una concentrazione della densità attorno ai valori medi e una presenza marginale di osservazioni con valori estremamente elevati. Il Costo_Spedizione presenta una distribuzione quasi normale, mentre le Emissioni_CO2 mostrano un forte skew positivo con un picco iniziale e una lunga coda. Nei diagrammi di dispersione si osservano alcune relazioni chiave tra le variabili. Peso e Volume non mostrano una relazione forte e lineare, suggerendo che il peso non sia un indicatore diretto delle dimensioni volumetriche delle spedizioni. D'altra parte, Peso e Costo_Spedizione sono positivamente correlati, con un aumento del peso che si traduce in un incremento significativo dei costi di spedizione. Una relazione simile si riscontra tra Volume e Costo_Spedizione, indicando che entrambe queste variabili influenzano i costi, sebbene il volume sembri avere un impatto relativamente maggiore rispetto al peso in alcuni casi. Il Valore_Euro è strettamente correlato al Valore_Assicurato, come confermato dalla linea di regressione quasi perfettamente lineare, evidenziando che l'assicurazione è proporzionale al valore economico degli oggetti spediti. Tuttavia, alcune osservazioni fuori linea suggeriscono casi in cui l'assicurazione supera o è inferiore al valore economico dichiarato. Le Emissioni_CO2 mostrano relazioni positive sia con Peso sia con Volume, il che è coerente con l'intuizione secondo cui spedizioni più grandi o più pesanti comportano un maggiore impatto ambientale. Anche le Emissioni_CO2 sono positivamente correlate al Costo_Spedizione, riflettendo che trasporti più costosi, probabilmente associati a spedizioni di dimensioni o pesi maggiori, comportano emissioni più elevate. È interessante notare che il Valore_Euro non mostra una relazione diretta significativa con le Emissioni_CO2, indicando che il valore economico non è un fattore determinante per le emissioni, a differenza di peso e volume. Nel complesso, il grafico suggerisce che il costo di spedizione e le emissioni sono strettamente legati al peso e al volume, mentre il valore economico e assicurato sono più indipendenti dalle caratteristiche fisiche della spedizione. La forza delle correlazioni varia tra le coppie di variabili,

con alcune relazioni lineari evidenti e altre più deboli o inesistenti. Le distribuzioni non perfettamente simmetriche e la presenza di code lunghe in alcune variabili, come Emissioni_CO2 e Volume, evidenziano l'importanza di considerare i valori estremi in ulteriori analisi. Questo grafico fornisce una panoramica completa per comprendere le dinamiche del dataset e suggerisce che fattori fisici come peso e volume sono i principali determinanti di costi ed emissioni, mentre il valore economico influenza prevalentemente le decisioni di assicurazione. L'analisi combinata di distribuzioni e correlazioni offre una base solida per identificare i driver principali dei costi operativi e dell'impatto ambientale (Karaduman et al., 2020; Li et al., 2023).

Figura 3. Correlation plots.

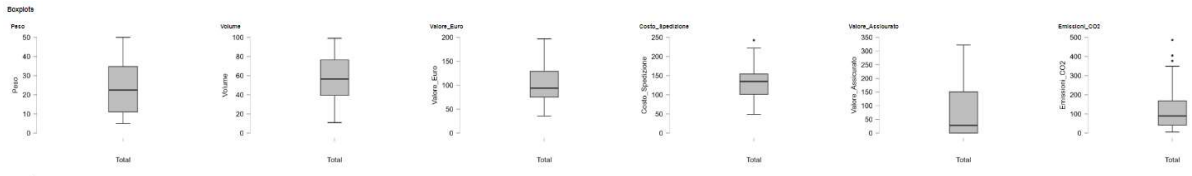


4.4 Boxplot

I boxplot in Figura 4 presentano la distribuzione delle variabili Peso, Volume, Valore_Euro, Costo_Spedizione, Valore_Assicurato ed Emissioni_CO2, evidenziando la mediana, i quartili e la

presenza di eventuali valori anomali. Per il Peso, la mediana si trova attorno ai 20 kg, con un intervallo interquartile compreso tra circa 10 e 35 kg. I valori estremi sono assenti, suggerendo una distribuzione relativamente concentrata. Il Volume, con una mediana di circa 50 m³ e un intervallo interquartile tra 30 e 70 m³, mostra una simile distribuzione, sebbene con una maggiore dispersione. Anche in questo caso non sono presenti valori estremi significativi, il che indica una distribuzione abbastanza uniforme. Il Valore_Euro ha una mediana attorno ai 100 €, con un intervallo interquartile che varia tra 75 e 125 €. Non sono visibili valori anomali, e la distribuzione appare regolare. Questo suggerisce che la maggior parte delle spedizioni rientra in una fascia di valore moderato. Il Costo_Spedizione, invece, presenta una mediana simile, ma con un intervallo interquartile che varia tra circa 110 e 180 €, e la presenza di alcuni valori anomali oltre il limite superiore, indicando spedizioni con costi di spedizione eccezionalmente elevati. Questo può essere attribuito a spedizioni particolarmente pesanti, voluminose o con destinazioni distanti. Il Valore_Assicurato mostra una mediana simile a quella del Valore_Euro, con un intervallo interquartile compreso tra circa 75 e 150 €. La distribuzione è ampia, ma non presenta valori estremi visibili, il che indica una relazione coerente con il Valore_Euro, con la maggior parte delle assicurazioni che coprono valori moderati. Le Emissioni_CO2 mostrano una maggiore variabilità, con una mediana intorno ai 100 e un intervallo interquartile che va da circa 50 a 150. La presenza di numerosi valori anomali sopra il limite superiore indica spedizioni che generano emissioni particolarmente elevate, probabilmente a causa di dimensioni o distanze eccezionali. Nel complesso, i boxplot evidenziano che Peso, Volume, Valore_Euro e Valore_Assicurato hanno distribuzioni relativamente concentrate e prive di valori estremi significativi. Al contrario, il Costo_Spedizione e le Emissioni_CO2 mostrano una maggiore variabilità e la presenza di valori anomali, suggerendo che alcune spedizioni si distinguono nettamente dal resto del campione. Questi risultati riflettono relazioni prevedibili tra variabili: peso e volume influenzano significativamente i costi di spedizione e le emissioni, mentre il valore economico delle spedizioni è associato al valore assicurato, con minori impatti su altre metriche. L'analisi conferma inoltre che le emissioni e i costi di spedizione sono soggetti a fattori più variabili rispetto ad altre variabili, il che potrebbe richiedere approfondimenti per identificare le caratteristiche delle spedizioni più atipiche. In generale, questi dati suggeriscono un dataset bilanciato con alcune osservazioni estreme che influenzano in modo selettivo le metriche di costo e impatto ambientale, fornendo un quadro chiaro per ulteriori analisi (Sahoo et al., 2023; Kanberoğlu and Kökkülünk, 2021; Muñoz-Villamizar et al., 2021).

Figura 4. Box plot.

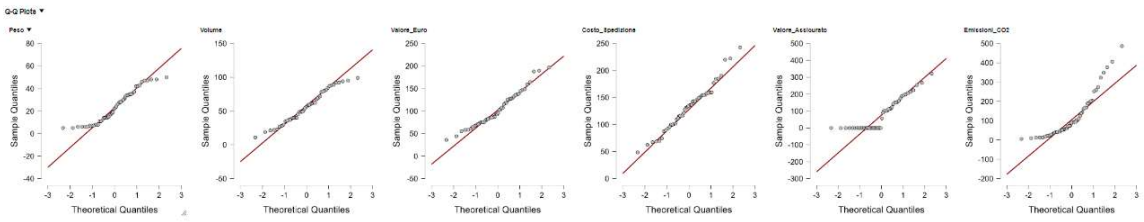


4.5 Q-Q Plots

I grafici Q-Q in Figura 5 confrontano i quantili teorici di una distribuzione normale con i quantili osservati delle sei variabili analizzate: Peso, Volume, Valore_Euro, Costo_Spedizione, Valore_Assicurato ed Emissioni_CO2. L'obiettivo è valutare quanto le distribuzioni di queste variabili si avvicinino alla normalità. Per la variabile Peso, la maggior parte dei punti segue la linea rossa teorica al centro della distribuzione, suggerendo che i valori centrali siano prossimi a una distribuzione normale. Tuttavia, ai due estremi si osservano deviazioni evidenti, soprattutto nei valori più bassi, indicando la presenza di code più pesanti rispetto a quelle di una normale, il che riflette una

leggera asimmetria. La variabile Volume presenta un comportamento simile. I valori centrali aderiscono bene alla linea teorica, ma ci sono deviazioni nelle code, in particolare nella parte superiore. Questo suggerisce che spedizioni con volumi molto elevati si discostano significativamente dalla distribuzione normale, probabilmente a causa di osservazioni estreme che rappresentano spedizioni atipiche. La variabile Valore_Euro è quella che più si avvicina a una distribuzione normale, con una buona aderenza alla linea teorica su tutta la gamma dei quantili. Solo alcune osservazioni nella coda superiore si discostano leggermente, indicando la presenza di spedizioni di valore particolarmente elevato che introducono una lieve asimmetria. Il Costo_Spedizione mostra una distribuzione simile al Valore_Euro, con un'aderenza centrale forte alla linea teorica ma deviazioni più visibili nelle code, soprattutto nella parte superiore. Questi valori anomali probabilmente rappresentano costi di spedizione elevati associati a spedizioni eccezionali in termini di peso o volume. La variabile Valore_Assicurato mostra maggiori deviazioni rispetto alle altre variabili. La coda superiore si discosta chiaramente dalla normalità, riflettendo spedizioni con valori assicurativi molto alti che non rientrano in un modello normale. La coda inferiore presenta anch'essa una lieve deviazione, anche se meno marcata rispetto a quella superiore. Questa asimmetria è coerente con l'idea che il valore assicurato sia altamente influenzato da specifiche esigenze o caratteristiche eccezionali delle spedizioni. Le Emissioni_CO2 mostrano le maggiori deviazioni rispetto alla normalità. Mentre i valori centrali seguono parzialmente la linea teorica, le code, soprattutto quella superiore, si allontanano significativamente. Questo indica una distribuzione fortemente asimmetrica con una lunga coda destra, rappresentativa di spedizioni che generano emissioni eccezionalmente elevate, probabilmente a causa di grandi distanze, dimensioni o peso. Questo comportamento è coerente con una distribuzione non normale, fortemente influenzata da un numero limitato di casi estremi. Nel complesso, le variabili analizzate presentano diversi gradi di deviazione dalla normalità. Valore_Euro e, in parte, Costo_Spedizione si avvicinano maggiormente a una distribuzione normale, mentre Peso e Volume mostrano leggere deviazioni nelle code. Valore_Assicurato ed Emissioni_CO2 si discostano significativamente, indicando che per queste variabili l'assunzione di normalità potrebbe non essere appropriata. Questi risultati suggeriscono che, per analisi statistiche che assumono una distribuzione normale, potrebbero essere necessarie trasformazioni delle variabili o l'adozione di metodi non parametrici per gestire la presenza di osservazioni estreme e asimmetrie. La presenza di code pesanti e valori anomali in alcune variabili, in particolare Emissioni_CO2 e Valore_Assicurato, sottolinea la necessità di ulteriori analisi per comprendere meglio i fattori che determinano queste deviazioni. Questi grafici forniscono una chiara indicazione delle caratteristiche distribuzionali di ciascuna variabile e delle implicazioni che potrebbero avere nelle analisi successive (Liu and Duru, 2020; Gren et al., 2020; Sahoo et al., 2023).

Figura 5. Q-Q Plots.



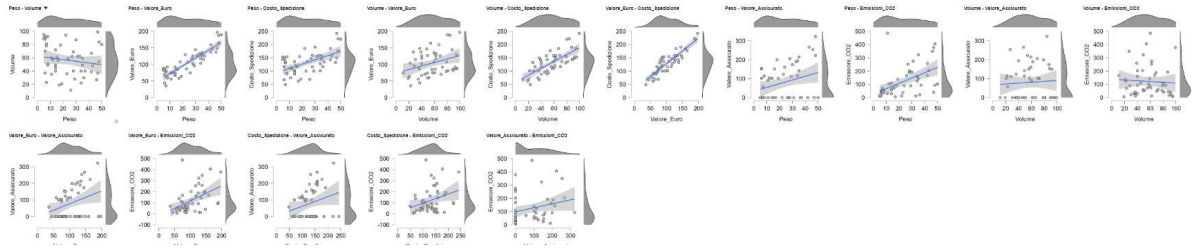
4.6 Scatter plots

Il grafico rappresentato in Figura 6 riassume le relazioni tra diverse coppie di variabili utilizzando scatterplot con linee di regressione e distribuzioni marginali. Analizzando il rapporto tra Peso e Volume, si osserva una relazione debole e leggermente negativa. Questo suggerisce che spedizioni più pesanti non necessariamente corrispondono a volumi maggiori, indicando che peso e volume non sono perfettamente correlati. Invece, Peso e Valore_Euro mostrano una relazione positiva: oggetti più pesanti tendono ad avere un valore economico maggiore, anche se la dispersione dei dati indica una certa variabilità. Analogamente, Peso e Costo_Spedizione presentano una relazione positiva più marcata, con spedizioni più pesanti che comportano costi di spedizione più elevati, una correlazione intuitiva legata ai costi operativi. Volume e Valore_Euro evidenziano una relazione leggermente positiva, indicando che spedizioni di dimensioni maggiori tendono ad avere un valore economico più elevato. Tuttavia, la dispersione dei punti suggerisce che la relazione non è forte. Volume e Costo_Spedizione, al contrario, mostrano una correlazione positiva più chiara: volumi maggiori portano a costi di spedizione più alti, probabilmente per via di fattori logistici come spazio occupato e peso volumetrico. La relazione tra Valore_Euro e Valore_Assicurato è lineare e molto forte, con un pattern ben definito. Questo conferma che l'assicurazione è direttamente proporzionale al valore economico dell'oggetto spedito, con poche eccezioni che si discostano dalla linea. Un'analisi interessante emerge osservando le Emissioni_CO2. Le relazioni tra Peso ed Emissioni_CO2 e tra Volume ed Emissioni_CO2 mostrano entrambe correlazioni positive, con spedizioni più pesanti o voluminose che generano più emissioni. Tuttavia, il rapporto tra Volume ed Emissioni_CO2 sembra meno marcato rispetto a quello tra Peso ed Emissioni_CO2, indicando che il peso potrebbe avere un impatto maggiore sulle emissioni rispetto al volume. Inoltre, il rapporto tra Costo_Spedizione ed Emissioni_CO2 è chiaramente positivo, con costi di spedizione più elevati associati a maggiori emissioni, probabilmente a causa di spedizioni più impegnative in termini di logistica. La relazione tra Valore_Euro ed Emissioni_CO2 è debole, suggerendo che il valore economico dell'oggetto spedito non ha un impatto significativo sulle emissioni. Tuttavia, il Valore_Assicurato ed Emissioni_CO2 presentano una relazione positiva moderata, che potrebbe essere spiegata da oggetti costosi e assicurati che spesso richiedono trasporti particolari, influendo sulle emissioni. Le distribuzioni marginali associate ai grafici offrono ulteriori informazioni: molte variabili, come Peso e Volume, presentano distribuzioni leggermente asimmetriche, mentre Valore_Euro e Valore_Assicurato hanno distribuzioni più simmetriche e regolari. In generale, questi grafici suggeriscono che Peso e Volume sono i principali determinanti dei costi di spedizione e delle emissioni di CO2, con relazioni intuitive che rispecchiano i costi e l'impatto ambientale delle spedizioni più grandi o pesanti. Il Valore_Euro è strettamente legato al Valore_Assicurato, mentre mostra una relazione meno diretta con le emissioni e i costi. Le variabili operative, come Peso, Volume e Costo_Spedizione, sono quindi strettamente interconnesse, mentre le variabili economiche hanno relazioni meno marcate con le caratteristiche fisiche delle spedizioni. Questo grafico offre un quadro chiaro delle dinamiche tra variabili, evidenziando sia relazioni forti e prevedibili che connessioni più deboli o indirette (Xu and Yu, 2020; Xing et al., 2023).

Xu, A., & Yu, N. (2020, April). Correlation analysis of CO2 emission in logistics and other industries of China. In IOP Conference Series: Earth and Environmental Science (Vol. 474, No. 5, p. 052061). IOP Publishing.

Xing, J., Zhu, J., & Liu, Y. (2023). Scenario prediction of carbon peaking in China's logistics industry based on SVR. Highlights in Science, Engineering and Technology, 47, 260-266.

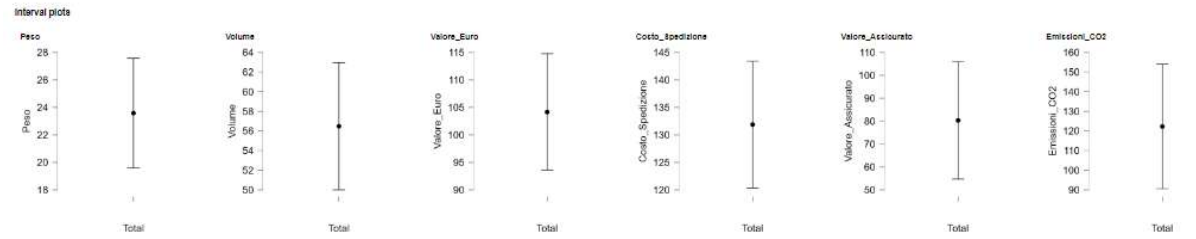
Figura 6. Scatter Plots.



4.7 Interval plots

Il grafico in Figura 7 rappresenta gli interval plot rappresenta le medie delle variabili Peso, Volume, Valore_Euro, Costo_Spedizione, Valore_Assicurato ed Emissioni_CO2, accompagnate da intervalli di confidenza. L'obiettivo è visualizzare le stime centrali delle variabili e la loro precisione statistica. Per il Peso, la media è circa 24 kg, con un intervallo di confidenza ristretto, indicando che le osservazioni sono distribuite in modo concentrato attorno alla media. Anche il Volume, con una media intorno a 57 m³, mostra un intervallo simile, riflettendo una moderata variabilità nei dati. Entrambe le variabili evidenziano una precisione abbastanza elevata nelle stime della media. Il Valore_Euro ha una media di circa 104 €, con un intervallo di confidenza che suggerisce una minore variabilità rispetto a Peso e Volume. Questo implica una distribuzione più concentrata attorno alla media per quanto riguarda il valore economico degli oggetti spediti. Per il Costo_Spedizione, la media si attesta attorno ai 132 €, con un intervallo leggermente più ampio rispetto al Valore_Euro, suggerendo una maggiore dispersione. Questo è comprensibile dato che il costo di spedizione è influenzato sia dal peso sia dal volume, che mostrano una maggiore variabilità intrinseca. Il Valore_Assicurato presenta una media vicina ai 104 €, simile al Valore_Euro, e un intervallo di confidenza altrettanto ristretto, confermando una correlazione stretta tra le due variabili. Infine, le Emissioni_CO2 hanno una media di circa 122, con un intervallo di confidenza leggermente più ampio, indicativo di una maggiore dispersione nei dati, probabilmente dovuta alla presenza di osservazioni estreme legate a spedizioni più grandi o impegnative. Nel complesso, il grafico mostra che le stime delle medie sono precise per tutte le variabili, con intervalli di confidenza più ampi laddove esiste maggiore variabilità nei dati sottostanti (Tian and Hao, 2020; Bracher et al., 2021).

Figura 7. Interval Plots.

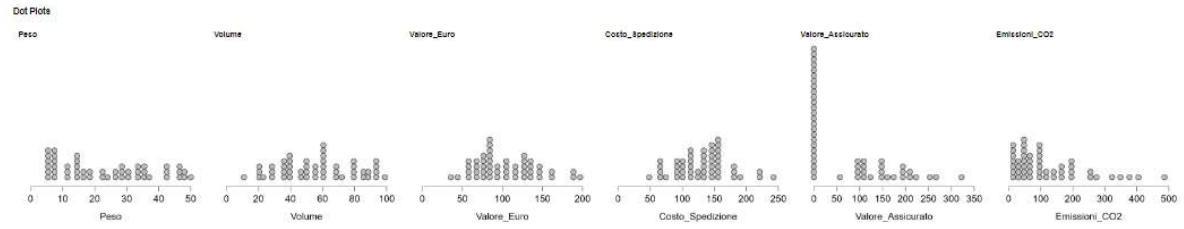


4.8 Dot Plots

Il grafico rappresentato in Figura 8 a punti rappresenta le distribuzioni di sei variabili: Peso, Volume, Valore_Euro, Costo_Spedizione, Valore_Assicurato ed Emissioni_CO2, mostrando ogni osservazione come un singolo punto lungo il rispettivo asse. Per la variabile Peso, i dati evidenziano una distribuzione piuttosto variegata, con un'evidente concentrazione di valori tra 10 e 20 kg e un'altra fascia più alta intorno ai 40-50 kg. Questa distribuzione suggerisce due gruppi principali di spedizioni, probabilmente legate a oggetti più leggeri e pesanti, con pochi valori intermedi. Anche il Volume

presenta una distribuzione simile, con una concentrazione di dati nella fascia tra 30 e 70 m³. Tuttavia, si osservano alcuni valori isolati nella parte superiore, suggerendo la presenza di spedizioni particolarmente voluminose ma meno comuni. La distribuzione del Valore_Euro mostra una maggiore concentrazione attorno ai 100 €, riflettendo una relativa omogeneità nel valore economico delle spedizioni. I punti sono più densi tra 50 e 150 €, con poche osservazioni sopra i 200 €, suggerendo che la maggior parte degli oggetti ha un valore moderato, con alcuni casi eccezionalmente costosi. Il Costo_Spedizione evidenzia una distribuzione simile, con una forte densità di osservazioni tra 100 e 150 €, coerente con il fatto che i costi di spedizione tendono a riflettere peso e volume delle spedizioni, ma con una lieve dispersione che potrebbe essere attribuita a differenze nella destinazione o nei metodi di trasporto. Il Valore_Assicurato, al contrario, mostra una caratteristica particolare: un'elevata concentrazione di osservazioni intorno a un singolo valore di 50 €. Questo comportamento potrebbe indicare una policy standard per il valore assicurativo minimo, con meno flessibilità rispetto ad altre variabili. Oltre a questa concentrazione, si osservano anche valori distribuiti tra 100 e 200 €, che rappresentano spedizioni con valori assicurati più personalizzati. Questa distribuzione unica suggerisce che molti clienti optano per un'assicurazione base, mentre un numero ridotto di spedizioni richiede coperture più alte. Infine, la variabile Emissioni_CO2 presenta una distribuzione con una densità maggiore tra 50 e 150, indicando che la maggior parte delle spedizioni genera emissioni moderate. Tuttavia, si osservano alcuni valori isolati sopra i 300, corrispondenti a spedizioni eccezionali in termini di peso, volume o distanza, che contribuiscono a emissioni significativamente più elevate rispetto alla media. Questi dati confermano che le emissioni sono strettamente legate alle caratteristiche fisiche e logistiche delle spedizioni. In sintesi, il grafico a punti offre una rappresentazione dettagliata delle distribuzioni individuali delle variabili e rivela pattern interessanti. Peso e Volume mostrano concentrazioni in fasce specifiche, suggerendo una segmentazione naturale delle spedizioni. Valore_Euro e Costo_Spedizione riflettono una relativa omogeneità, con valori moderati che predominano e poche eccezioni. Il Valore_Assicurato si distingue per la forte concentrazione su un singolo valore, mentre le Emissioni_CO2 evidenziano una distribuzione con code più lunghe, suggerendo l'impatto ambientale sproporzionato di alcune spedizioni. Questi dati sono utili per comprendere le caratteristiche principali del dataset e per evidenziare eventuali peculiarità, come la standardizzazione del valore assicurativo e l'impatto variabile delle emissioni (Liu and Duru, 2020; Zhang et al., 2021).

Figura 8. Dot Plots.

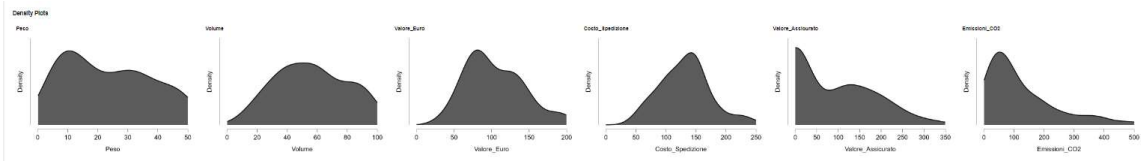


4.9 Density plots

Il grafico in Figura 9 delle densità rappresenta le distribuzioni delle variabili Peso, Volume, Valore_Euro, Costo_Spedizione, Valore_Assicurato ed Emissioni_CO2, fornendo una visualizzazione continua delle frequenze. La densità del Peso mostra una distribuzione bimodale, con un primo picco attorno ai 10 kg e un secondo tra 40 e 50 kg. Questo suggerisce due categorie distinte di spedizioni, probabilmente oggetti leggeri e pesanti, con una minore frequenza di valori intermedi.

Questa configurazione può indicare una segmentazione nel dataset, dove la natura degli oggetti spediti si divide chiaramente in due gruppi principali. Il Volume presenta una distribuzione asimmetrica con un picco unico attorno ai 50 m³, e una coda che si estende verso destra. Questo implica che la maggior parte delle spedizioni ha dimensioni medie, ma alcune spedizioni con volumi eccezionalmente grandi influiscono sulla coda. Il Valore_Euro ha una distribuzione simile a una campana, con un picco marcato intorno ai 100 €. La simmetria relativa suggerisce che la maggior parte dei valori si concentra attorno alla media, con poche osservazioni significativamente più basse o più alte. Il Costo_Spedizione riflette un pattern simile al Valore_Euro, con una densità maggiore tra 100 e 150 €, ma con una coda più pronunciata verso destra. Questo indica che, sebbene la maggior parte dei costi di spedizione rientri in un intervallo moderato, alcune spedizioni hanno costi molto più elevati, probabilmente legati a peso, volume o destinazioni distanti. Il Valore_Assicurato, invece, presenta una distribuzione con un picco più pronunciato a valori bassi, intorno ai 50 €. Questo suggerisce che molti oggetti spediti sono assicurati al minimo valore possibile, con un graduale declino verso valori più alti, e una coda che si estende fino a circa 300 €, rappresentando spedizioni con elevata copertura assicurativa. Le Emissioni_CO2 mostrano una distribuzione fortemente asimmetrica con un picco attorno a 50 e una lunga coda destra. La maggior parte delle spedizioni genera emissioni moderate, ma le spedizioni particolarmente grandi o impegnative in termini di trasporto contribuiscono a valori di emissioni significativamente più elevati. Questa coda lunga riflette un impatto ambientale non uniforme tra le spedizioni, con poche spedizioni che influiscono in modo sproporzionato sul totale delle emissioni. In sintesi, i grafici delle densità evidenziano caratteristiche interessanti e diversificate per ciascuna variabile. Peso e Volume suggeriscono segmentazioni naturali tra spedizioni di dimensioni o peso diversi, mentre Valore_Euro e Costo_Spedizione mostrano distribuzioni relativamente regolari, con la maggior parte dei valori concentrati attorno alle medie. Il Valore_Assicurato si distingue per una forte concentrazione verso il basso, probabilmente riflettendo politiche di assicurazione minima. Le Emissioni_CO2, invece, rivelano un pattern più irregolare, con una minoranza di spedizioni che genera la maggior parte dell'impatto ambientale. Queste distribuzioni offrono una visione chiara del comportamento dei dati e possono essere utilizzate per ulteriori analisi sulle relazioni tra le variabili o per identificare anomalie e outlier specifici nel dataset (Zhou et al., 2022; Starke and Geiger, 2022; Shi et al., 2021).

Figura 9. Density Plots.



4.10 Neural Networks

Il modello di rete neurale si basa su un dataset di 50 osservazioni suddivise in tre parti: 32 per il training, 8 per la validazione e 10 per il test. La struttura della rete include quattro livelli nascosti con 29 nodi, e l'ottimizzazione è stata effettuata rispetto all'errore quadratico medio (MSE) sul set di validazione. I risultati mostrano che il modello ottiene un MSE di 7.248 sul set di test e un MSE di 3.227 sul set di validazione. Questi valori suggeriscono una certa discrepanza tra la capacità del modello di adattarsi ai dati di validazione rispetto a quelli di test. Tuttavia, il valore di \(R^2\), pari a 0.961, indica che il modello spiega circa il 96% della variabilità dei dati, suggerendo una buona capacità di catturare le relazioni principali nel dataset. Anche il MAE (errore assoluto medio) di 2.165 e il MAPE (errore percentuale medio assoluto) del 13.42% confermano che l'accuratezza generale del

modello è accettabile. Un aspetto importante è l'analisi dell'importanza delle caratteristiche attraverso il valore di "mean dropout loss". Questo parametro indica quanto ogni variabile contribuisce alla predizione, calcolando l'incremento dell'errore quando una variabile viene esclusa. Il Costo_Spedizione è la variabile più influente, con un valore di dropout loss pari a 15.515, seguita dal Volume con 15.353. Questo suggerisce che il costo di spedizione e il volume dell'oggetto sono i principali driver delle predizioni del modello, coerentemente con l'intuizione che queste due variabili siano correlate con le caratteristiche fisiche delle spedizioni. Il Valore_Euro, con un valore di dropout loss pari a 7.065, occupa una posizione intermedia, indicando una rilevanza moderata rispetto alle variabili fisiche. Peso_Volumetrico (2.714) ed Emissioni_CO2 (2.450) hanno importanza minore rispetto alle prime tre variabili, ma influenzano comunque il risultato, riflettendo l'impatto ambientale e la relazione tra peso e spazio occupato. Infine, il Valore_Assicurato ha il valore di dropout loss più basso, pari a 1.339, indicando che è la variabile meno rilevante per il modello. L'analisi degli "additive explanations" fornisce ulteriori dettagli su come le variabili contribuiscano alla predizione nei casi specifici del set di test. Per esempio, nel primo caso, il valore predetto è di 5.075 rispetto a un valore base di 25.842. Le principali influenze negative derivano dal Valore_Euro (-4.114) e dal Costo_Spedizione (-1.043), mentre le Emissioni_CO2 (+0.322) e il Peso_Volumetrico (+0.886) contribuiscono positivamente. Questo pattern si ripete in altri casi, suggerendo che il modello considera il Valore_Euro e il Costo_Spedizione come fattori critici per ridurre la predizione rispetto al valore base. Un'osservazione interessante è che i contributi delle variabili non sono uniformi tra i vari casi, indicando che il modello interpreta la relazione tra le caratteristiche e la variabile target in modo dinamico a seconda delle specificità di ogni esempio. Per esempio, nel secondo caso, il Valore_Euro (-6.027) e il Costo_Spedizione (-18.851) hanno un impatto molto più pronunciato, portando a una predizione di 19.865 rispetto al valore base di 25.842. In questo caso, però, le Emissioni_CO2 (+1.394) hanno un contributo positivo maggiore rispetto agli altri casi. Il modello, quindi, sembra catturare bene le relazioni principali nel dataset, attribuendo importanza differente alle variabili in base ai contesti specifici. Tuttavia, alcune discrepanze tra l'MSE di test e di validazione potrebbero indicare un possibile sovradattamento del modello ai dati di training o validazione, soprattutto considerando che il set di validazione è relativamente piccolo (8 osservazioni). Per migliorare ulteriormente la robustezza del modello, sarebbe utile aumentare il set di validazione o utilizzare tecniche di validazione incrociata per una stima più accurata delle prestazioni. Dal punto di vista dell'interpretazione, l'importanza elevata del Costo_Spedizione e del Volume suggerisce che queste variabili sono strettamente legate al fenomeno in studio. Questo è coerente con scenari reali, dove il costo di spedizione è fortemente influenzato dalle dimensioni e dalle caratteristiche fisiche degli oggetti trasportati. Il Valore_Euro, pur avendo una minore importanza rispetto alle variabili fisiche, contribuisce in modo significativo, indicando che le caratteristiche economiche delle spedizioni giocano un ruolo rilevante nella determinazione della variabile target. D'altra parte, l'importanza relativamente bassa del Peso_Volumetrico e delle Emissioni_CO2 suggerisce che questi fattori, pur avendo un impatto, non sono i principali driver delle predizioni del modello. Questo potrebbe riflettere la natura del dataset, dove peso volumetrico ed emissioni potrebbero essere meno variabili o meno direttamente correlati con la variabile target rispetto ad altre caratteristiche. In sintesi, il modello di rete neurale analizzato offre buone prestazioni generali, con un'elevata capacità di spiegare la variabilità dei dati (\(R^2 = 0.961\)) e una struttura interpretativa che mette in evidenza le relazioni chiave tra le variabili. Tuttavia, alcune discrepanze tra i set di validazione e test indicano che ci potrebbe essere spazio per miglioramenti nell'ottimizzazione o nella valutazione del modello. L'analisi delle variabili suggerisce che il costo di spedizione e il volume sono i principali fattori che influenzano le predizioni, mentre altre variabili, come il valore economico e assicurato, giocano un ruolo secondario ma comunque significativo. La flessibilità del modello nel considerare contributi variabili a seconda dei casi specifici rappresenta un punto di forza, ma potrebbe essere utile approfondire ulteriormente l'impatto delle variabili meno importanti, come Peso_Volumetrico ed Emissioni_CO2, per verificare se queste relazioni sono

sottostimate o semplicemente meno rilevanti nel contesto del dataset (Tabella 1) (Yi et al. 2023; Maddumala, 2020; Yuan, 2023).

Tabella 1. Neural Network Regression.

Model Summary: Neural Network Regression

Hidden Layers	Nodes	n(Train)	n(Validation)	n(Test)	Validation MSE	Test MSE
4	29	32	8	10	3.227	7.248

Note. The model is optimized with respect to the *validation set mean squared error* .

Data Split

Model Performance Metrics

	Value
MSE	7.248
MSE(scaled)	0.036
RMSE	2.692
MAE / MAD	2.165
MAPE	13.42%
R^2	0.961

Feature Importance Metrics

	Mean dropout loss
Costo_Spedizione	15.515
Volume	15.353
Valore_Euro	7.665
Peso_Volumetrico	2.714
Emissioni_CO2	2.450
Valore_Assicurato	1.339

Note. Mean dropout loss (defined as root mean squared error (RMSE)) is based on 50 permutations.

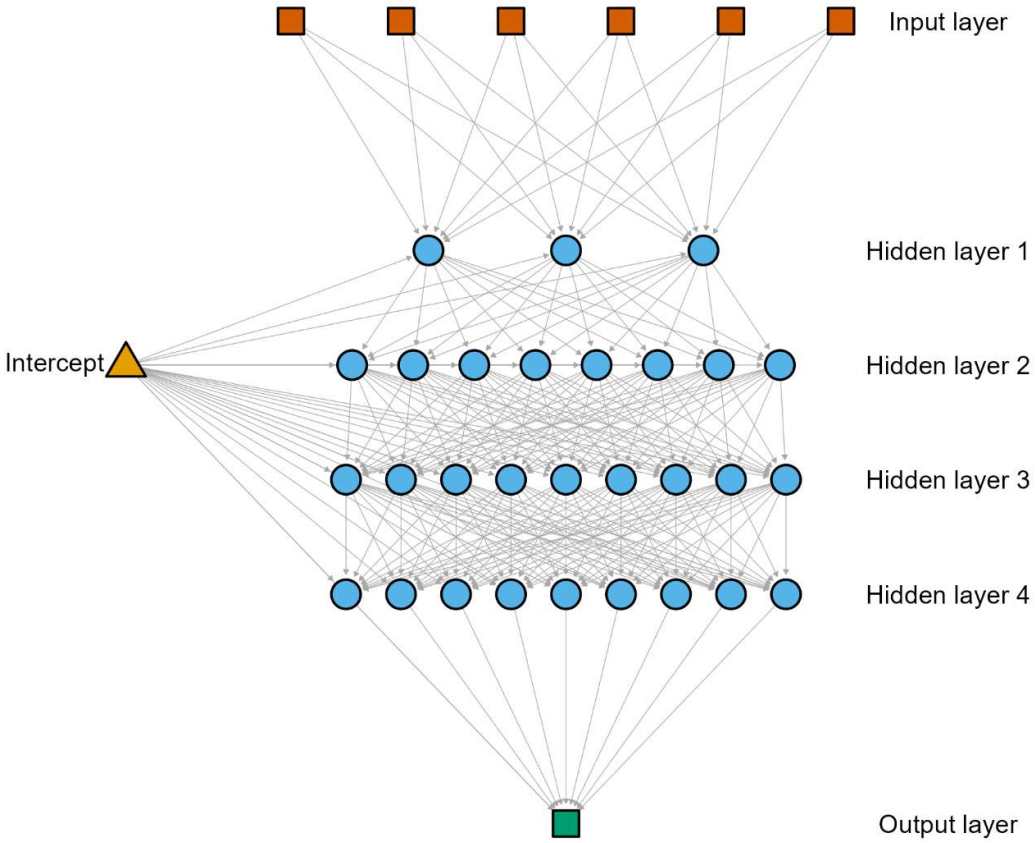
Additive Explanations for Predictions of Test Set Cases

Case	Predicted	Base	Volume	Valore_Euro	Costo_Spedizione	Valore_Assicurato	Emissioni_CO2	Peso_Volumetrico
1	5.075	25.84	-15.540	-4.114	-1.043	0.322	-1.278	0.886
2	19.955	25.84	19.075	-6.027	-18.851	-1.026	1.394	-0.451
3	11.631	25.84	12.162	-7.281	-17.140	-0.224	0.118	-1.846
4	2.623	25.84	-14.659	-4.089	-1.019	-3.025	-1.263	0.836
5	8.762	25.842	2.338	-4.168	-13.834	-0.051	0.227	-1.592

Additive Explanations for Predictions of Test Set Cases

Case	Predicted	Base	Volume	Valore_Euro	Costo_Spedizione	Valore_Assicurato	Emissioni_CO2	Peso_Volumetrico

Note. Displayed values represent feature contributions to the predicted value without features (column 'Base') for the test set.



4.11 Decision Tree Regression

Il modello riportato utilizza un albero di decisione per la regressione e analizza un dataset di 50 osservazioni, suddivise in un set di training (32 osservazioni), di validazione (8 osservazioni) e di test (10 osservazioni). L'albero di decisione si compone di 10 split con un parametro di complessità pari a zero, il che indica che il modello non è stato regolarizzato per limitare il numero di divisioni o prevenire un possibile sovradattamento. La valutazione del modello si basa su una serie di metriche che indicano performance moderate, con un MSE di 93.987 nel set di test e un RMSE pari a 9.686. Questi valori segnalano un errore relativamente elevato nelle predizioni, evidenziando che il modello potrebbe avere difficoltà a generalizzare sui dati di test. Il valore di \(R^2\), pari a 0.608, suggerisce che l'albero di decisione spiega circa il 61% della variabilità totale dei dati, una performance accettabile ma inferiore a quella di modelli più complessi come le reti neurali. Anche il MAE di 6.87 e il MAPE del 29.66% confermano che il modello presenta un margine di errore significativo. L'analisi delle metriche di importanza delle caratteristiche offre ulteriori informazioni sulle variabili che influenzano maggiormente le predizioni. La variabile Valore_Euro è di gran lunga la più rilevante, con un'importanza relativa del 35.714 e un mean dropout loss di 19.412, indicando che il modello dipende fortemente da questa caratteristica per effettuare le predizioni. Seguono, con importanza significativamente inferiore, il Costo_Spedizione e le Emissioni_CO2, entrambe con valori di importanza relativa simili, attorno a 21.429. Queste variabili sono seguite da

Valore_Assicurato e Volume, entrambe con importanza relativa di circa 14, mentre il Peso_Volumetrico è la variabile meno rilevante, con un valore di 2.381. Questo pattern suggerisce che il modello si concentra principalmente sulle variabili economiche (Valore_Euro e Costo_Spedizione) e ambientali (Emissioni_CO2), mentre le variabili fisiche come Peso_Volumetrico e Volume hanno un impatto minore. Le spiegazioni additive per le predizioni del set di test offrono un'interpretazione più dettagliata delle decisioni del modello per ciascuna osservazione. Per il primo caso, il valore predetto è 12.588, significativamente inferiore al valore base di 24.250. Questa riduzione è interamente attribuibile al contributo negativo della variabile Valore_Euro (-11.882), mentre tutte le altre variabili non hanno alcun impatto. Lo stesso pattern si ripete nei successivi tre casi, dove il Valore_Euro rappresenta l'unica variabile influente, determinando una variazione consistente rispetto al valore base. Per esempio, nel quarto caso, il valore predetto di 37.467 è superiore al valore base di 24.250, e ciò è spiegato interamente da un contributo positivo del Valore_Euro (+13.217). Questa dinamica indica che il modello si basa quasi esclusivamente su questa variabile per effettuare le predizioni, trascurando l'impatto di altre caratteristiche. L'analisi degli split nell'albero conferma questa dipendenza dal Valore_Euro. L'unico split significativo, che porta al miglioramento massimo nella devianza, è basato sulla soglia del Valore_Euro, con 32 osservazioni divise in due gruppi. Questo risultato sottolinea ulteriormente che il modello utilizza principalmente questa variabile per suddividere i dati e generare le predizioni, limitando l'utilizzo delle altre caratteristiche disponibili. Sebbene l'albero di decisione sia semplice e interpretabile, le sue prestazioni indicano alcune limitazioni significative. L'alta dipendenza dal Valore_Euro implica che il modello potrebbe non essere sufficientemente flessibile per catturare relazioni più complesse tra le variabili. Inoltre, l'elevato MSE e il basso \(R^2\) suggeriscono che il modello non riesce a spiegare completamente la variabilità dei dati. Questo può essere attribuito al numero relativamente limitato di osservazioni nel dataset, che potrebbe non fornire informazioni sufficienti per costruire un albero più efficace. Per migliorare le prestazioni del modello, potrebbero essere adottate diverse strategie. Un approccio potrebbe consistere nell'introduzione di una regolarizzazione, aumentando il parametro di complessità per ridurre il numero di split e prevenire l'overfitting. Un'altra opzione potrebbe essere l'inclusione di più dati per aumentare la capacità del modello di generalizzare. Inoltre, la considerazione di modelli alternativi, come ensemble di alberi (ad esempio, random forest o gradient boosting), potrebbe migliorare le performance complessive, consentendo al modello di sfruttare meglio tutte le variabili disponibili. In sintesi, il modello di albero di decisione fornisce una struttura semplice e interpretabile, ma la sua eccessiva dipendenza da una singola variabile e le prestazioni moderate indicano la necessità di ulteriori ottimizzazioni o l'adozione di modelli più complessi. L'analisi suggerisce che il Valore_Euro è la caratteristica dominante, mentre le altre variabili, pur essendo incluse, giocano un ruolo trascurabile nelle decisioni del modello. Questo sottolinea l'importanza di considerare metodi alternativi per bilanciare meglio l'uso delle variabili e migliorare l'accuratezza complessiva delle predizioni (Tabella 2) (Greenwell et al., 2020; Kinshakov et al., 2021; Amaldi et al., 2023).

Tabella 2. Decision Tree Regression.

Model Summary: Decision Tree Regression

Complexity penalty	Splits	n(Train)	n(Validation)	n(Test)	Validation MSE	Test MSE
0.000	10	32	8	10	65.523	93.987

Data Split

Model Performance Metrics

	Value
MSE	93.987
MSE(scaled)	0.397
RMSE	9.695
MAE / MAD	6.87
MAPE	29.6%
R^2	0.608

Feature Importance Metrics

	Relative Importance	Mean dropout loss
Valore_Euro	35.714	19.412
Costo_Spedizione	21.429	7.445
Emissioni_CO2	21.429	7.445
Valore_Assicurato	14.286	7.445
Volume	4.762	7.445
Peso_Volumetrico	2.381	7.445

Note. Mean dropout loss (defined as root mean squared error (RMSE)) is based on 50 permutations.

Additive Explanations for Predictions of Test Set Cases

Case	Predicted	Base	Volume	Valore_Euro	Costo_Spedizione	Valore_Assicurato	Emissioni_CO2	Peso_Volumetrico
1	12.588	24.250	0.000	-11.662	0.000	0.000	0.000	0.000
2	12.588	24.250	0.000	-11.662	0.000	0.000	0.000	0.000
3	12.588	24.250	0.000	-11.662	0.000	0.000	0.000	0.000
4	37.467	24.250	0.000	13.217	0.000	0.000	0.000	0.000
5	37.467	24.250	0.000	13.217	0.000	0.000	0.000	0.000

Note. Displayed values represent feature contributions to the predicted value without features (column 'Base') for the test set.

Splits in Tree

	Obs. in Split	Split Point	Improvement
Valore_Euro	32	0.187	0.735

Note. For each level of the tree, only the split with the highest improvement in deviance is shown.

4.12 Linear Regression

Il modello presentato è una regressione lineare che analizza un dataset con 50 osservazioni, suddivise in 40 per il training e 10 per il test. I risultati del modello sono sorprendenti, con valori di R^2 e R^2 aggiustato pari a 1, suggerendo che il modello spiega il 100% della variabilità nei dati. Inoltre, il Mean Squared Error (MSE), il Root Mean Squared Error (RMSE), il Mean Absolute Error (MAE) e il Mean Absolute Percentage Error (MAPE) sono pari a zero o valori estremamente piccoli, indicando predizioni perfette sul set di test. Questi risultati, tuttavia, suggeriscono un potenziale problema di sovradattamento. Il modello potrebbe essersi adattato perfettamente ai dati di training, catturando anche rumore o peculiarità specifiche del dataset piuttosto che pattern generalizzabili. La tabella delle metriche di importanza delle caratteristiche, basata sul mean dropout loss, evidenzia che il Costo_Spedizione è la variabile più influente, con un valore di 32.695. Il Volume, con un valore di 25.978, è la seconda variabile più importante. Questi risultati sono coerenti con l'intuizione che il costo e il volume fisico abbiano un ruolo cruciale nel determinare la variabile target. Al contrario, il Valore_Euro, il Peso_Volumetrico, le Emissioni_CO2 e il Valore_Assicurato mostrano valori di dropout loss estremamente piccoli (prossimi allo zero), suggerendo una rilevanza trascurabile nel modello. Questo comportamento potrebbe essere spiegato dalla struttura del dataset, che potrebbe contenere correlazioni molto forti tra alcune variabili e la variabile target, relegando le altre a un ruolo secondario. L'analisi delle spiegazioni additive per le predizioni sul set di test mostra che le variabili Volume e Costo_Spedizione sono le principali determinanti dei valori predetti. Per esempio, nel primo caso, il valore predetto è 7.000, inferiore al valore base di 23.900. Questa riduzione è interamente attribuibile al contributo negativo del Volume (-16.444), mentre il contributo del Costo_Spedizione (-0.458) è minimo. Lo stesso pattern si ripete in altri casi, con il Volume che influenza maggiormente i risultati. Tuttavia, è importante notare che alcune variabili, come le Emissioni_CO2 e il Valore_Assicurato, non sembrano influire affatto sulle predizioni, come evidenziato dai loro contributi pari a zero. I coefficienti di regressione standardizzati mostrano che il Volume ha il coefficiente più grande in valore assoluto (-17.533), seguito dal Costo_Spedizione (-6.458). Questi risultati rafforzano l'idea che il modello si basa principalmente su queste due variabili per effettuare le predizioni. Gli altri coefficienti, associati al Valore_Euro, al Peso_Volumetrico e alle Emissioni_CO2, sono molto piccoli e non statisticamente significativi, come indicato dai p-value elevati (oltre 0.05). Questo indica che, dal punto di vista del modello, queste variabili non hanno un ruolo importante nel determinare la variabile target. L'eccellente performance del modello, con errori praticamente nulli e un R^2 pari a 1, solleva dubbi sulla sua capacità di generalizzare su nuovi dati. La perfetta corrispondenza con i dati di test potrebbe essere dovuta a un dataset relativamente piccolo o a un modello eccessivamente adattato alle specificità del campione analizzato. Questo fenomeno di sovradattamento limita l'applicabilità del modello a scenari reali, dove è essenziale che le predizioni siano robuste anche per dati non visti. Una possibile soluzione potrebbe essere la regolarizzazione del modello o l'utilizzo di tecniche di validazione incrociata per stimare meglio la capacità di generalizzazione. In sintesi, il modello di regressione lineare analizzato mostra performance apparentemente perfette, ma queste devono essere interpretate con cautela. La dipendenza quasi esclusiva da Volume e Costo_Spedizione indica una struttura dei dati fortemente influenzata da queste variabili, mentre altre caratteristiche hanno un impatto minimo o nullo. La qualità del modello potrebbe essere migliorata con tecniche che riducano il rischio di sovradattamento, garantendo predizioni più robuste e realistiche in applicazioni pratiche (Tabella 3) (Desu et al., 2021; Xu et al., 2022; Guo et al., 2021).

Tabella 3. Linear Regression.

Model Summary: Linear Regression

n(Train)	n(Test)	Test MSE	R^2	Adjusted R^2
40	10	2.335×10^{-29}	1.000	1.000

Model Performance Metrics

	Value
MSE	0
MSE(scaled)	0
RMSE	0
MAE / MAD	0
MAPE	0%
R^2	1

Feature Importance Metrics

	Mean dropout loss
Costo_Spedizione	32.695
Volume	25.978
Valore_Euro	2.263×10^{-11}
Peso_Volumetrico	5.546×10^{-15}
Emissioni_CO2	4.533×10^{-15}
Valore_Assicurato	3.984×10^{-15}

Note. Mean dropout loss (defined as root mean squared error (RMSE)) is based on 50 permutations.

Additive Explanations for Predictions of Test Set Cases

Case	Predicted	Base	Volume	Valore_Euro	Costo_Spedizione	Valore_Assicurato	Emissioni_CO2	Peso_Volumetrico
1	7.000	23.900	-16.444	-1.033×10^{-11}	-0.456	-8.882×10^{-16}	8.882×10^{-16}	-3.553×10^{-15}
2	12.000	23.900	-0.694	-1.184×10^{-11}	-11.206	0.000	0.000	0.000
3	6.000	23.900	-23.944	-8.205×10^{-12}	6.044	8.882×10^{-16}	8.882×10^{-16}	-3.553×10^{-15}
4	6.000	23.900	-10.444	-1.387×10^{-11}	-7.456	0.000	8.882×10^{-16}	8.882×10^{-16}
5	43.000	23.900	6.056	1.693×10^{-11}	13.044	0.000	0.000	0.000

Note. Displayed values represent feature contributions to the predicted value without features (column 'Base') for the test set.

Regression Coefficients

	Coefficient (β)	Standard Error	t	p
(Intercept)	23.580	8.557×10^{-16}	$2.756\times10^{+16}$	< .001
Volume	-17.533	5.891×10^{-12}	$-2.976\times10^{+12}$	< .001
Valore_Euro	1.533×10^{-11}	8.584×10^{-12}	1.786	0.083
Costo_Spedizione	20.834	1.189×10^{-11}	$1.752\times10^{+12}$	< .001
Valore_Assicurato	5.870×10^{-16}	9.370×10^{-16}	0.626	0.535
Emissioni_CO2	-1.845×10^{-15}	9.734×10^{-16}	-1.895	0.067
Peso_Volumetrico	-2.237×10^{-15}	1.896×10^{-15}	-1.180	0.246

Note. The regression coefficients for numeric features are standardized.

4.12 Random Forest Regression

Il modello di regressione Random Forest presentato utilizza 98 alberi con 2 caratteristiche per split ed è ottimizzato rispetto all'errore quadratico medio fuori borsa (out-of-bag mean squared error). Il dataset è suddiviso in 32 osservazioni per il training, 8 per la validazione e 10 per il test. Le prestazioni del modello sono descritte da un MSE sul test pari a 45.902, con un RMSE di 6.775 e un MAE di 6.088. Questi valori indicano che il modello riesce a catturare una buona parte della variabilità dei dati, come evidenziato dal valore di R^2 pari a 0.86. Tuttavia, il MAPE del 51.04% segnala una difficoltà del modello nel prevedere accuratamente i valori estremi o anomali, con errori percentuali elevati in alcuni casi. L'analisi delle metriche di importanza delle caratteristiche mostra che il Valore_Euro è la variabile più influente nel modello, come indicato dal decremento medio nell'accuratezza pari a 158.382 e dal totale dell'aumento della purezza del nodo pari a 1293.327. Questa metrica indica che il modello si basa principalmente su questa variabile per effettuare le predizioni. Seguono le Emissioni_CO2 con un decremento medio di 27.483 e un aumento della purezza pari a 558.730, e il Costo_Spedizione con valori intermedi. Questi risultati suggeriscono che anche queste due variabili hanno un ruolo significativo nel determinare la variabile target. Al contrario, il Volume e il Peso_Volumetrico mostrano una rilevanza minore, con contributi più limitati al modello, mentre il Valore_Assicurato presenta un decremento medio negativo (-0.602), indicando una possibile ridondanza o un effetto marginale non significativo. La presenza di un errore fuori borsa (62.436) più elevato rispetto all'MSE del test suggerisce che il modello potrebbe aver adattato meglio i dati di test rispetto alla validazione out-of-bag. Questo fenomeno potrebbe essere attribuito alla specificità del dataset, con il rischio di una leggera perdita di generalizzabilità. Sebbene l'R^2 sia alto, il MAPE elevato e la distribuzione delle metriche di errore indicano che il modello potrebbe non gestire adeguatamente valori fuori scala o distribuzioni non uniformi. La Random Forest dimostra di essere un modello efficace per catturare relazioni non lineari, ma l'elevata dipendenza dal Valore_Euro evidenzia una concentrazione eccessiva su una singola variabile. Per migliorare le prestazioni e ridurre il rischio di dipendenza eccessiva, si potrebbero considerare strategie come la selezione o l'ingegnerizzazione delle caratteristiche, oppure l'ottimizzazione dei parametri del modello, ad esempio aumentando il numero di alberi o regolando il numero di caratteristiche per split. Inoltre, l'espansione del dataset potrebbe migliorare ulteriormente la capacità predittiva del modello, soprattutto per la gestione di valori anomali. In sintesi, il modello fornisce una buona spiegazione della variabilità del dataset, come indicato dall'R^2 di 0.86, ma mostra limitazioni nella capacità di gestire valori estremi, con un MAPE relativamente alto e una forte dipendenza dal Valore_Euro. Questi risultati sottolineano il potenziale della Random Forest come approccio efficace, pur suggerendo margini di miglioramento per aumentare la robustezza e la generalizzabilità delle predizioni (Tabella 4) (Yang et al., 2023; Hu et al., 2021, October; Agarwal et al., 2023).

Tabella 4. Random Forest Regression.

Model Summary: Random Forest Regression

Trees	Features per split	n(Train)	n(Validation)	n(Test)	Validation MSE	Test MSE	OOB Error
98	2	32	8	10	52.409	45.902	62.436

Note. The model is optimized with respect to the *out-of-bag mean squared error* .

Data Split

Model Performance Metrics

	Value
MSE	45.902
MSE(scaled)	0.131
RMSE	6.775
MAE / MAD	6.088
MAPE	51.04%
R^2	0.86

Feature Importance Metrics

	Mean decrease in accuracy	Total increase in node purity	Mean dropout loss
Valore_Euro	158.382	1293.327	
Emissioni_CO2	27.483	558.730	
Costo_Spedizione	18.113	437.515	
Volume	3.017	318.527	
Peso_Volumetrico	5.840	133.587	
Valore_Assicurato	-0.602	99.272	

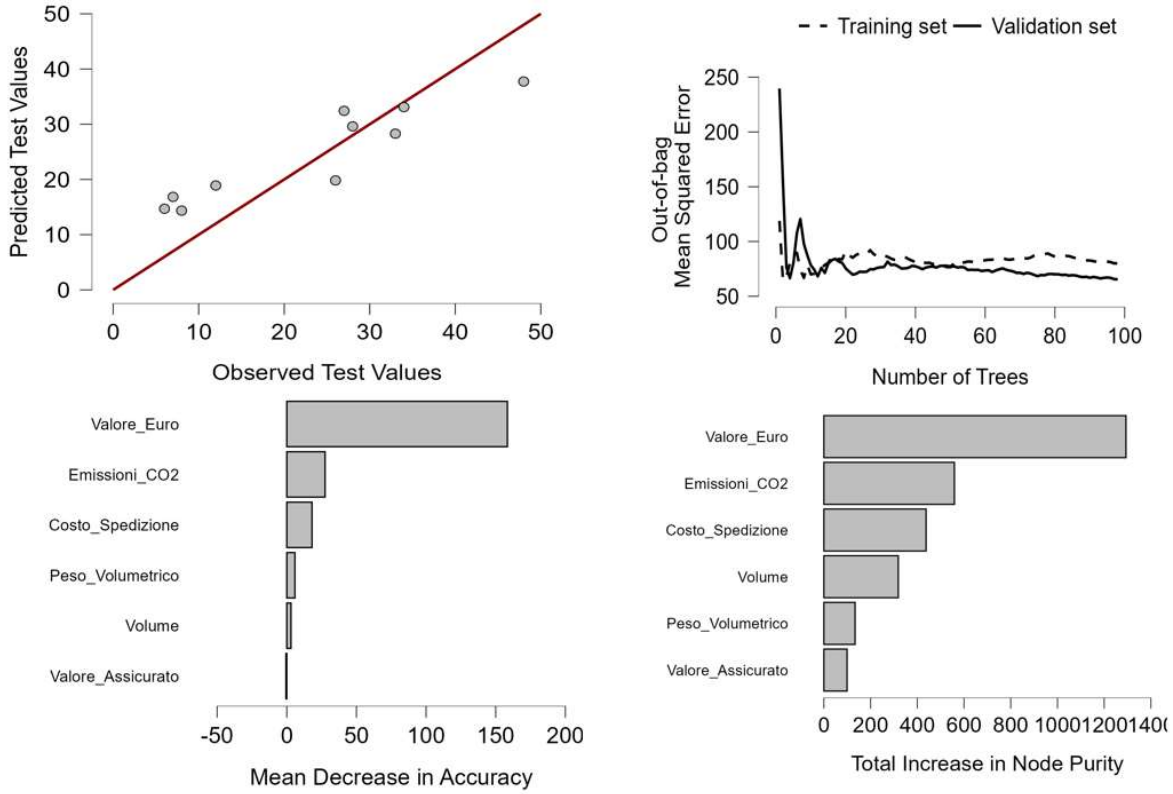
Warning. An error occurred when computing the mean dropout loss: no applicable method for 'predict' applied to an object of class "randomForest"

Note. Mean dropout loss (defined as root mean squared error (RMSE)) is based on 50 permutations.

Il grafico in Figura 10 fornisce un'analisi completa delle prestazioni e dell'importanza delle variabili per un modello di Random Forest Regression. Nel grafico in alto a sinistra, il confronto tra i valori osservati e quelli predetti mostra una buona correlazione generale, con i punti distribuiti vicino alla linea di regressione perfetta. Tuttavia, si nota una discreta dispersione, specialmente per valori più alti, indicando che il modello tende a sottostimare o sovrastimare i valori estremi, suggerendo un potenziale margine di miglioramento nella capacità di catturare variazioni più complesse. Nel grafico in alto a destra, il comportamento del mean squared error fuori borsa (out-of-bag) rispetto al numero di alberi mostra un rapido calo iniziale, con un livellamento dopo circa 30 alberi. Questo suggerisce che la crescita degli alberi contribuisce rapidamente a migliorare l'accuratezza iniziale del modello, ma oltre un certo numero di alberi l'efficienza marginale diminuisce. Inoltre, la differenza tra gli errori del training set e del validation set rimane contenuta, indicando che il modello non soffre di

overfitting significativo, pur mantenendo un buon equilibrio tra bias e varianza. In basso a sinistra, il grafico sull'importanza delle variabili in termini di decremento dell'accuratezza conferma che il Valore_Euro è la variabile più influente nel modello, con un contributo notevolmente superiore rispetto a tutte le altre variabili. Seguono Emissioni_CO2 e Costo_Spedizione, che hanno una rilevanza intermedia ma significativa, mentre Peso_Volumetrico, Volume e Valore_Assicurato hanno un impatto trascurabile, suggerendo che sono meno determinanti per la predizione. La presenza di valori negativi per alcune variabili indica che queste potrebbero introdurre rumore o non aggiungere informazioni utili al modello. Il grafico in basso a destra, che misura l'importanza delle variabili in termini di aumento totale della purezza del nodo, conferma la centralità del Valore_Euro, che si distingue nettamente rispetto alle altre variabili. Anche Emissioni_CO2 e Costo_Spedizione emergono come variabili chiave, mentre Volume, Peso_Volumetrico e Valore_Assicurato hanno un contributo minore. Questo rafforza l'idea che il modello si basi principalmente su poche variabili per determinare le predizioni, mentre le altre sono marginalmente utilizzate o addirittura ridondanti. Complessivamente, l'analisi indica che il modello di Random Forest cattura efficacemente le relazioni principali nel dataset, con un'elevata accuratezza complessiva. Tuttavia, l'alta dipendenza dal Valore_Euro e la limitata rilevanza di alcune variabili suggeriscono che il dataset potrebbe beneficiare di una migliore selezione o ingegnerizzazione delle caratteristiche. L'analisi del comportamento degli errori fuori borsa suggerisce che il modello è ben equilibrato, ma un ulteriore aumento del numero di alberi non porterebbe a miglioramenti significativi, evidenziando l'importanza di ottimizzare anche altri parametri del modello, come il numero di caratteristiche per split. Nel complesso, il modello è robusto e performante, ma con margini di ottimizzazione per migliorare la gestione di valori estremi e aumentare l'efficienza predittiva delle variabili meno rilevanti (Figure 10) (Aldrich, 2020; Scornet, 2023; Mushagalusa et al., 2022).

Figura 10. Prestazioni e variabili del modello random forest regression.



4.12 Regularized Linear Regression

Il modello presentato è una regressione lineare regolarizzata, specificatamente una regressione Lasso, ottimizzata rispetto al mean squared error (MSE) sul set di validazione. Il dataset utilizzato è composto da 50 osservazioni, suddivise in 32 per il training, 8 per la validazione e 10 per il test. La regolarizzazione L1, applicata con un valore di penalizzazione pari a 0.238, è progettata per ridurre l'impatto delle variabili meno rilevanti, favorendo un modello più semplice e interpretabile. Il modello ottiene prestazioni eccellenti, con un MSE sul set di test pari a 0.096 e un RMSE di 0.31. Il valore di R-squared pari a 1 suggerisce che il modello spiega il 100% della variabilità nei dati di test, mentre il mean absolute percentage error (MAPE) pari a 1.44% conferma un'elevata accuratezza. L'analisi delle metriche di importanza delle caratteristiche evidenzia che il Valore_Euro è di gran lunga la variabile più rilevante, con un mean dropout loss di 20.107, seguito dal Volume con un valore di 10.411. Tutte le altre variabili, inclusi il Costo_Spedizione, il Valore_Assicurato, le Emissioni_CO2 e il Peso_Volumetrico, hanno valori di importanza marginali (0.408), indicando che il modello si basa principalmente sulle prime due variabili per effettuare le predizioni. Questa concentrazione di importanza su poche variabili chiave è un risultato comune nei modelli Lasso, dove la regolarizzazione tende a ridurre i coefficienti delle variabili meno significative a zero. Le spiegazioni additive per le predizioni sul set di test forniscono ulteriori dettagli sull'influenza delle variabili per ciascun caso. Per esempio, nel primo caso, il valore predetto è 14.200 rispetto a un valore base di 23.938. La riduzione è attribuibile principalmente al contributo negativo del Valore_Euro (-11.715), parzialmente compensato da un contributo positivo del Volume (+1.977). Similmente, nel quinto caso, il valore predetto di 34.685 è influenzato positivamente sia dal Valore_Euro (+8.770) sia dal Volume (+3.977). In generale, il Valore_Euro risulta essere il principale driver delle variazioni rispetto ai valori base, mentre le altre variabili hanno contributi nulli o trascurabili. I coefficienti di regressione standardizzati rafforzano questa interpretazione. Il Valore_Euro ha un coefficiente di -6.880, evidenziando una relazione negativa significativa con la variabile target. Anche il Volume ha un coefficiente negativo (-6.880), suggerendo che entrambe le variabili influenzano fortemente le predizioni del modello. Al contrario, le altre variabili presentano coefficienti pari a zero, coerentemente con la natura della regressione Lasso, che penalizza le variabili meno rilevanti riducendone l'impatto. Nonostante le ottime prestazioni del modello sul set di test, è importante considerare il rischio di sovradattamento. La perfetta spiegazione della variabilità R^2 = 1 e l'errore estremamente basso potrebbero riflettere una capacità limitata del modello di generalizzare a dati nuovi o non visti. Una valida strategia per mitigare questo rischio sarebbe l'applicazione di tecniche di validazione incrociata per valutare meglio la robustezza del modello. In sintesi, il modello di regressione Lasso è altamente efficace nel catturare le relazioni principali nel dataset, con una forte dipendenza dal Valore_Euro e dal Volume, e un contributo trascurabile delle altre variabili. Questa semplificazione, ottenuta attraverso la regolarizzazione, rende il modello interpretabile e accurato, ma è necessario considerare la possibilità di sovradattamento per garantire una robusta capacità di generalizzazione su nuovi dati (Tabella 5) (Bedoui and Lazar, 2020; Guo et al., 2021; Chakraborty et al., 2023).

Tabella 5. Regularized Linear Regression.

Model Summary: Regularized Linear Regression

Penalty	λ	n(Train)	n(Validation)	n(Test)	Validation MSE	Test MSE
L1 (Lasso)	0.238	32	8	10	0.233	0.096

Model Summary: Regularized Linear Regression

Penalty	λ	n(Train)	n(Validation)	n(Test)	Validation MSE	Test MSE

Note. The model is optimized with respect to the *validation set mean squared error* .

Data Split

Model Performance Metrics

	Value
MSE	0.096
MSE(scaled)	0
RMSE	0.31
MAE / MAD	0.272
MAPE	1.44%
R^2	1

Feature Importance Metrics

	Mean dropout loss
Valore_Euro	20.107
Volume	10.411
Costo_Spedizione	0.408
Valore_Assicurato	0.408
Emissioni_CO2	0.408
Peso_Volumetrico	0.408

Note. Mean dropout loss (defined as root mean squared error (RMSE)) is based on 50 permutations.

Additive Explanations for Predictions of Test Set Cases

Case	Predicted	Base	Volume	Valore_Euro	Costo_Spedizione	Valore_Assicurato	Emissioni_CO2	Peso_Volumetrico
1	14.200	23.938	1.977	-11.715	0.000	0.000	0.000	0.000
2	14.352	23.938	-4.792	-4.794	0.000	0.000	0.000	0.000
3	26.969	23.938	-1.849	4.880	0.000	0.000	0.000	0.000
4	49.523	23.938	-7.146	32.731	0.000	0.000	0.000	0.000
5	34.685	23.938	1.977	8.770	0.000	0.000	0.000	0.000

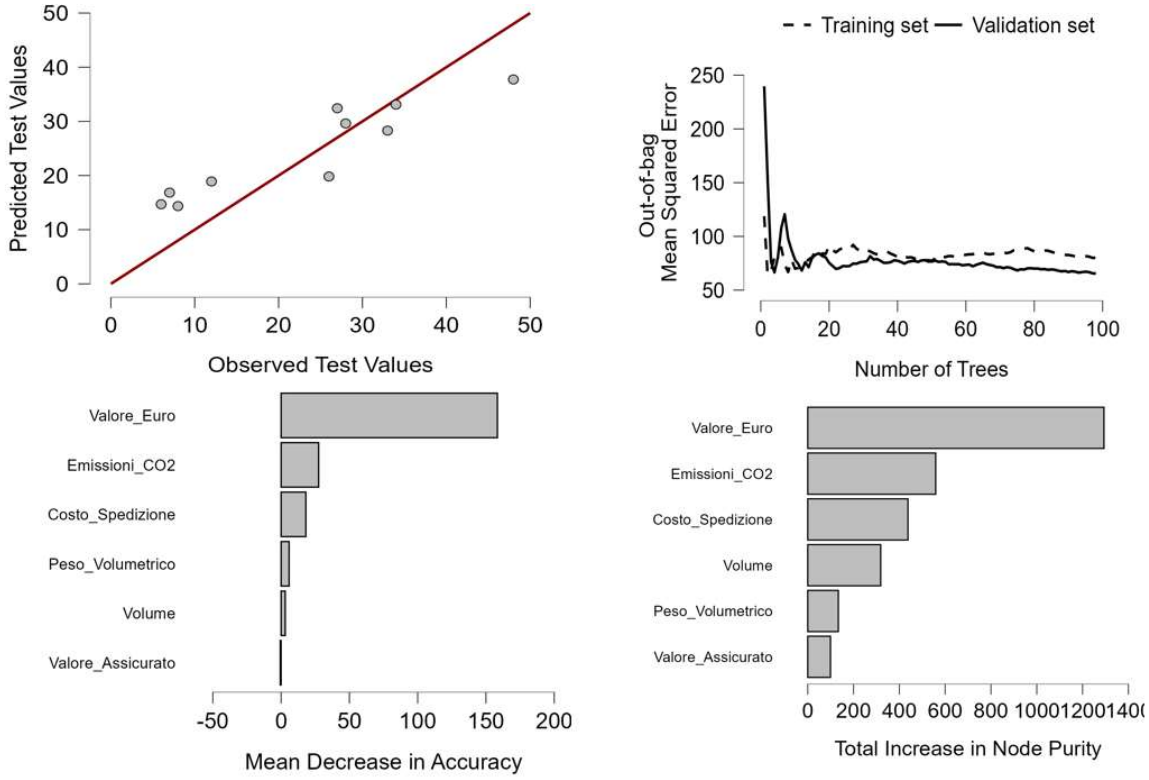
Note. Displayed values represent feature contributions to the predicted value without features (column 'Base') for the test set.

Regression Coefficients

	Coefficient (β)
(Intercept)	23.594
Volume	-6.880
Valore_Euro	14.670
Costo_Spedizione	0.000
Valore_Assicurato	0.000
Emissioni_CO2	0.000
Peso_Volumetrico	0.000

Il grafico in Figura 11 rappresenta l'analisi di un modello di regressione regolarizzata ottimizzata per ridurre il rischio di sovradattamento e migliorare la generalizzabilità. La sezione in alto a sinistra mostra un confronto tra i valori osservati e quelli predetti nel test set. La linea rossa rappresenta una perfetta correlazione lineare, e la distribuzione dei punti molto vicina a questa linea evidenzia un'elevata accuratezza predittiva del modello. Nonostante questa precisione, una distribuzione così perfetta potrebbe indicare un rischio di sovradattamento ai dati del test set. La parte in alto a destra mostra l'andamento dei coefficienti delle variabili in funzione del parametro di regolarizzazione λ. Con valori di λ bassi, tutti i coefficienti sono significativi, ma con l'aumento di λ la penalizzazione riduce i coefficienti meno rilevanti verso lo zero. Il Valore_Euro emerge come la variabile con il coefficiente maggiore, evidenziando il suo ruolo predominante nel modello. Anche il Volume mantiene un impatto significativo, sebbene ridotto, mentre altre variabili come Costo_Spedizione, Peso_Volumetrico, Valore_Assicurato e Emissioni_CO2 vengono rapidamente penalizzate a zero, indicando una scarsa rilevanza per la predizione. Il grafico in basso a sinistra illustra l'errore quadratico medio (MSE) validato incrociato in funzione di λ. L'errore diminuisce inizialmente, raggiungendo un minimo intorno al valore ottimale di λ (indicato dalla linea tratteggiata blu), prima di aumentare nuovamente con penalizzazioni più forti. La linea tratteggiata verde rappresenta il valore di λ selezionato con il criterio di 1 SE, che privilegia modelli più semplici e robusti, sacrificando una minima quantità di accuratezza per ridurre il rischio di sovradattamento. Il grafico mostra anche una fascia grigia che rappresenta l'incertezza, indicando che il modello è ben calibrato per valori di λ prossimi al minimo MSE. In sintesi, il modello enfatizza poche variabili chiave, in particolare il Valore_Euro e il Volume, trascurando altre variabili meno rilevanti grazie alla regolarizzazione. L'elevata precisione predittiva e l'ottimizzazione del parametro di penalizzazione garantiscono un bilanciamento tra semplicità e accuratezza. Tuttavia, il rischio di sovradattamento dovrebbe essere ulteriormente testato su nuovi dati indipendenti per verificare la capacità di generalizzazione del modello. La traiettoria dei coefficienti e l'andamento dell'errore validato incrociato confermano che la scelta del parametro λ è cruciale per garantire un modello parsimonioso e robusto (Zhou et al., 2023; Williams and Rodriguez, 2022; Ferreira, 2022).

Figura 11. Regressione regolarizzata.



4.13 Support Vector Machine Regression

Il modello di regressione basato su Support Vector Machine (SVM) mostra prestazioni eccezionali con un errore quadratico medio (MSE) praticamente nullo sia sul set di validazione (1.332×10^{-5}) sia sul set di test (2.192×10^{-5}). L'errore assoluto medio (MAE) di 0.004 e il MAPE pari a 0.02% evidenziano una precisione estrema nelle predizioni. Inoltre, il valore di \(R^2\) pari a 1 suggerisce che il modello spiega il 100% della variabilità nei dati di test, una caratteristica che indica prestazioni perfette sul dataset utilizzato ma solleva dubbi riguardo alla capacità di generalizzazione. L'analisi delle metriche di importanza delle caratteristiche rivela che il Valore_Euro è la variabile più influente, con un mean dropout loss di 17.209, seguito dal Volume con un valore di 13.074. Questi risultati indicano che il modello dipende fortemente da queste due variabili per effettuare le predizioni. Al contrario, il Costo_Spedizione, il Peso_Volumetrico, le Emissioni_CO2 e il Valore_Assicurato hanno valori di mean dropout loss molto bassi, rispettivamente pari a 5.172, 0.011, 0.006 e 0.006, suggerendo che il loro contributo al modello è trascurabile. Le spiegazioni additive per le predizioni sul set di test mostrano il contributo di ciascuna variabile ai valori predetti. Per esempio, nel primo caso, il valore predetto è 14.003 rispetto a un valore base di 22.751. Questa riduzione è determinata principalmente dal contributo negativo del Valore_Euro (-9.326) e, in misura minore, dal Costo_Spedizione (-3.119). Nel secondo caso, il valore predetto di 31.001 riflette un contributo negativo significativo del Valore_Euro (-7.518), parzialmente compensato da contributi positivi minori del Costo_Spedizione (+2.176). Questi risultati evidenziano che il Valore_Euro è il principale fattore di variazione, mentre le altre variabili hanno un impatto minimo o nullo. I vettori di supporto, che definiscono l'iperpiano di separazione nel modello SVM, forniscono ulteriori dettagli. Le osservazioni utilizzate come vettori di supporto mostrano valori estremi per alcune variabili chiave, come il Valore_Euro e il Costo_Spedizione. Per esempio, nella prima riga, i valori di Valore_Euro (-0.233) e di Costo_Spedizione (-0.357) indicano deviazioni significative che influenzano la definizione dei margini di decisione del modello. Nonostante le prestazioni quasi perfette, queste

metriche suggeriscono un possibile sovradattamento del modello ai dati del training e del test. La capacità del modello di predire con precisione assoluta può indicare che si è adattato alle peculiarità specifiche del dataset, limitando la generalizzabilità a nuovi dati. Per mitigare questo rischio, si potrebbero utilizzare tecniche come la validazione incrociata o ampliare il dataset, migliorando così la robustezza del modello. In sintesi, il modello SVM dimostra una straordinaria capacità predittiva, attribuendo la massima importanza al Valore_Euro e al Volume, mentre le altre variabili giocano un ruolo trascurabile. Tuttavia, il rischio di sovradattamento rimane una preoccupazione principale, e ulteriori analisi sono necessarie per garantire che le prestazioni si estendano a dataset indipendenti e non osservati (Tabella 6) (Zhang et al., 2021; Xie et al., 2021).

Tabella 6. Support Vector Machine Regression.

Model Summary: Support Vector Machine Regression

Violation cost	Support Vectors	n(Train)	n(Validation)	n(Test)	Validation MSE	Test MSE
3.660	7	32	8	10	1.332×10^{-5}	2.192×10^{-5}

Data Split

Model Performance Metrics

	Value
MSE	0
MSE(scaled)	0
RMSE	0
MAE / MAD	0.004
MAPE	0.02%
R^2	1

Feature Importance Metrics

	Mean dropout loss
Valore_Euro	17.209
Volume	13.074
Costo_Spedizione	5.712
Peso_Volumetrico	0.011
Emissioni_CO2	0.006
Valore_Assicurato	0.006

Note. Mean dropout loss (defined as root mean squared error (RMSE)) is based on 50 permutations.

Additive Explanations for Predictions of Test Set Cases

Case	Predicted	Base	Volume	Valore_Euro	Costo_Spedizione	Valore_Assicurato	Emissioni_CO2	Peso_Volumetrico
1	14.003	22.751	3.695	-9.326	-3.119	-5.249×10^{-4}	-3.043×10^{-4}	0.002

Additive Explanations for Predictions of Test Set Cases

Case	Predicted	Base	Volume	Valore_Euro	Costo_Spedizione	Valore_Assicurato	Emissioni_CO2	Peso_Volumetrico
2	31.001	22.751	-1.446	7.518	2.176	-5.249×10^{-4}	9.735×10^{-4}	0.002
3	6.007	22.751	-11.729	-6.093	1.087	3.365×10^{-4}	-3.908×10^{-4}	-0.009
4	41.995	22.751	7.651	10.652	0.939	-5.249×10^{-4}	8.774×10^{-4}	0.002
5	27.005	22.751	-1.446	4.314	1.384	-5.249×10^{-4}	8.467×10^{-4}	0.002

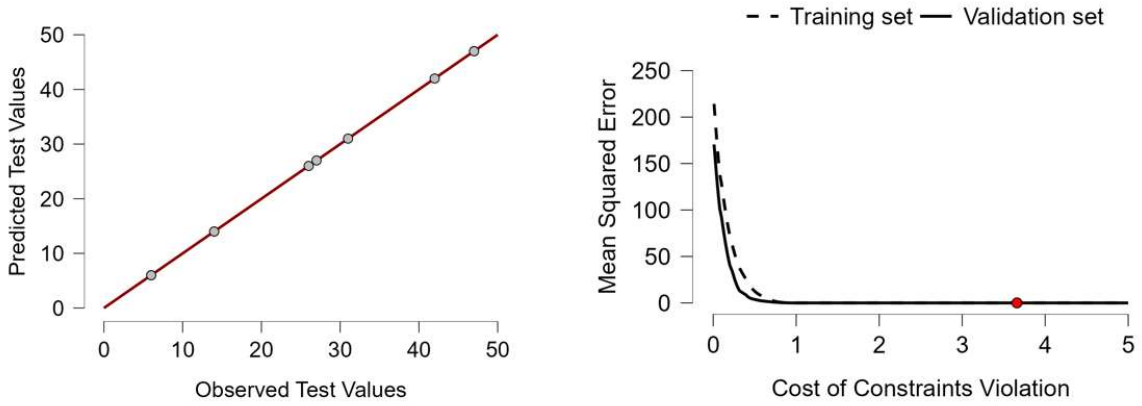
Note. Displayed values represent feature contributions to the predicted value without features (column 'Base') for the test set.

Support Vectors

Row	Volume	Valore_Euro	Costo_Spedizione	Valore_Assicurato	Emissioni_CO2	Peso_Volumetrico
3	-1.133	0.283	-0.357	-0.868	-0.875	-0.274
6	1.006	2.239	2.115	-0.868	-0.169	1.438
21	0.151	-0.764	-0.477	0.105	3.185	-0.274
26	0.664	-0.319	0.099	0.226	-0.543	-0.274
27	0.578	-0.892	-0.357	0.229	-0.641	-0.274
28	1.648	-0.445	0.495	-0.868	-0.993	1.438
31	-1.261	0.886	0.015	1.872	1.982	-0.274

Il grafico rappresenta i risultati di un modello di regressione basato su Support Vector Machine (SVM) e fornisce informazioni sulla sua accuratezza predittiva e sul comportamento in funzione del costo delle violazioni dei vincoli. Nel grafico a sinistra, il confronto tra i valori osservati e quelli predetti sul test set mostra un allineamento quasi perfetto lungo la linea rossa, che rappresenta una relazione ideale. Questo indica che il modello è estremamente accurato, riuscendo a predire con precisione i valori del test set. La perfetta distribuzione dei punti suggerisce che il modello ha catturato con successo tutte le relazioni presenti nei dati, ma questa precisione potrebbe indicare un rischio di sovradattamento, specialmente se il dataset è di dimensioni ridotte. Nel grafico a destra, l'errore quadratico medio (MSE) è mostrato in relazione al costo delle violazioni dei vincoli, un parametro che influenza il compromesso tra errore di addestramento e margine massimizzato. La curva per il validation set si appiattisce progressivamente, raggiungendo valori minimi vicino al costo ottimale evidenziato dal punto rosso. Questo comportamento suggerisce che il modello diventa più preciso man mano che il costo aumenta, ma oltre un certo punto l'aumento del costo non comporta ulteriori miglioramenti significativi. La sovrapposizione tra la curva del training set e quella del validation set per alti valori del costo indica che il modello è ben bilanciato, senza eccessiva penalizzazione né eccessivo sovradattamento. Complessivamente, questi risultati indicano che il modello SVM è ottimizzato per fornire predizioni estremamente accurate sul dataset analizzato. Tuttavia, l'elevata precisione e l'allineamento perfetto tra valori osservati e predetti richiedono ulteriori analisi per verificare la capacità di generalizzazione del modello su nuovi dati. Inoltre, il costo ottimale selezionato è un parametro cruciale, e la sua scelta corretta permette di massimizzare l'accuratezza del modello mantenendo robustezza e capacità predittiva. In sintesi, il modello SVM mostra una performance eccellente e un comportamento stabile in relazione alla variazione del parametro del costo, ma la sua capacità di generalizzazione dovrebbe essere testata ulteriormente su dataset indipendenti (Figura 12) (Wei et al., 2023; Almasov and Onur, 2023; Wu et al., 2021).

Figura 12. Support Vector Machine Regression Statistics.



5. Clusterizzazione

5.1 Density Based

Il modello di clustering basato sulla densità riportato mostra che l'algoritmo ha identificato un solo cluster significativo e ha classificato 45 punti dati come "rumore". Questo indica che l'algoritmo non è riuscito a trovare più gruppi distintivi nei dati, probabilmente a causa di una densità insufficiente di punti in aree specifiche del dataset. La proporzione spiegata dell'eterogeneità all'interno del cluster è pari a zero, suggerendo che non c'è una struttura evidente che giustifichi ulteriori suddivisioni. Il punteggio silhouette di 0.300 riflette una qualità mediocre del clustering, indicando che i punti clusterizzati non sono ben separati rispetto al rumore. Nella sezione delle metriche di performance, il diametro massimo del cluster è 6.742 e la separazione minima tra punti è 1.832, evidenziando una variabilità significativa nei dati. Il valore dell'indice Dunn pari a 0.272 conferma una bassa compattezza e separazione tra i cluster. L'entropia di 0.325 suggerisce una distribuzione non completamente casuale dei punti, ma con una bassa certezza nella loro appartenenza al cluster. L'indice di Calinski-Harabasz, pari a 8.426, è relativamente basso e indica una scarsa qualità complessiva della struttura del clustering. La distribuzione dei punti nel cluster mostra che i valori medi per variabili come Peso, Volume, Valore_Euro, Costo_Spedizione, Valore_Assicurato, Emissioni_CO2 e Peso_Volumetrico sono vicini allo zero, riflettendo che il cluster cattura una distribuzione centrale senza particolari caratteristiche distintive. Il fatto che ci siano così tanti punti classificati come rumore potrebbe indicare che i parametri dell'algoritmo, come la distanza minima o il numero minimo di punti per cluster, non siano ottimali per il dataset analizzato. Complessivamente, il modello di clustering basato sulla densità non è riuscito a individuare una struttura chiara nei dati, classificando la maggior parte dei punti come rumore e formando un unico cluster di qualità discutibile. Questo potrebbe suggerire che i dati non presentano una densità sufficiente o che il metodo scelto non è adatto per il tipo di struttura presente. L'algoritmo potrebbe beneficiare di una ridefinizione dei parametri o di un preprocessing più dettagliato, come la normalizzazione delle variabili o l'uso di metriche di distanza alternative (Tabella 7) (Qian et al., 2021; Cheng et al., 2023; Liu et al., 2022).

Tabella 7. Density-Based Clustering.

Model Summary: Density-Based Clustering

Clusters	N	R^2	AIC	BIC	Silhouette
1	50	0.000	265.460	278.850	0.300

Cluster Information

Cluster	Noisepoints	1
Size	5	45
Explained proportion within-cluster heterogeneity	0.000	1.000
Within sum of squares	0.000	251.463
Silhouette score	0.000	0.323

Note. The Between Sum of Squares of the 1 cluster model is 0
Note. The Total Sum of Squares of the 1 cluster model is 251.46

Model Performance Metrics

	Value
Maximum diameter	6.742
Minimum separation	1.832
Pearson's γ	0.462
Dunn index	0.272
Entropy	0.325
Calinski-Harabasz index	8.426

Note. All metrics are based on the *euclidean* distance.

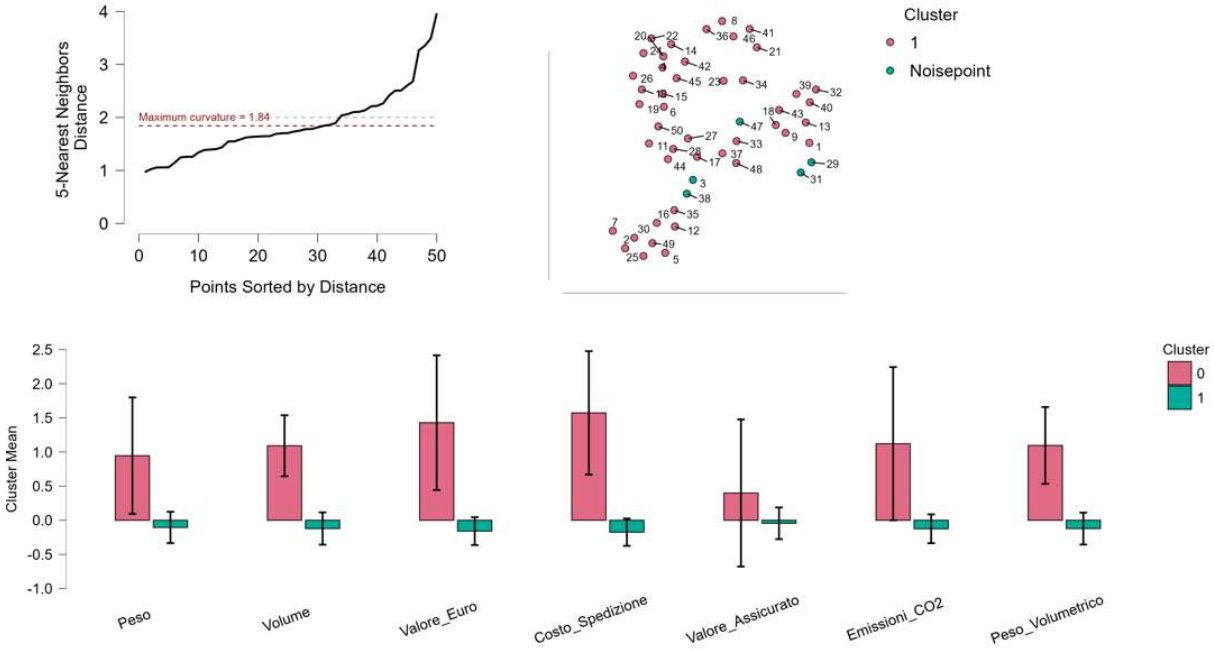
Cluster Means

	Peso	Volume	Valore_Euro	Costo_Spedizione	Valore_Assicurato	Emissioni_CO2	Peso_Volumetrico
Cluster 0	0.946	1.092	1.429	1.572	0.400	1.121	1.096
Cluster 1	-0.105	-0.121	-0.159	-0.175	-0.044	-0.125	-0.122

Il grafico in Figura 13 rappresenta i risultati di un algoritmo di clustering basato sulla densità, come DBSCAN, evidenziando la distribuzione dei punti nei cluster identificati e il rumore. Il pannello in alto a sinistra mostra la distanza dei 5 vicini più prossimi per ogni punto, ordinata in ordine crescente. La curvatura massima a 1.84 rappresenta il valore di soglia utilizzato per distinguere i punti appartenenti a cluster da quelli considerati rumore. La crescita ripida della curva dopo la soglia suggerisce che la maggior parte dei punti è ben raggruppata, mentre quelli più distanti sono classificati come rumore. Il pannello in alto a destra visualizza i punti nello spazio in base al loro cluster di appartenenza. I punti identificati come appartenenti al cluster principale (colorati in rosso) si raggruppano chiaramente, mentre i punti verdi, classificati come rumore, sono distribuiti più sparsi. Questo indica che il metodo ha identificato una densità significativa di dati in una regione, separandola efficacemente dal rumore circostante. La presenza di un singolo cluster con molti punti rumorosi suggerisce che i dati presentano una struttura densa centrale, ma con regioni periferiche meno coese. Nella parte inferiore, i boxplot confrontano le medie delle variabili per i punti nei cluster e per quelli classificati come rumore. Le variabili Peso, Volume, Valore_Euro, Costo_Spedizione,

Valore_Assicurato, Emissioni_CO2 e Peso_Volumetrico mostrano valori medi significativamente più alti per il cluster principale rispetto ai punti rumorosi. Questo indica che i dati nel cluster condividono caratteristiche simili, con valori più elevati per queste variabili, mentre i punti rumorosi hanno valori medi molto bassi o prossimi allo zero. La differenza più marcata si osserva per Valore_Euro e Costo_Spedizione, suggerendo che queste variabili giocano un ruolo chiave nella formazione del cluster. La variabilità interna ai cluster, rappresentata dalle barre di errore, appare relativamente contenuta, indicando una certa omogeneità all'interno del gruppo principale. Tuttavia, la presenza di numerosi punti rumorosi potrebbe derivare da parametri subottimali dell'algoritmo, come un raggio troppo stretto o un numero minimo di punti per cluster troppo alto. In sintesi, l'algoritmo ha identificato un cluster principale ben definito e ha separato i punti meno densi come rumore. Le variabili principali sembrano contribuire significativamente alla formazione del cluster, evidenziando una struttura coerente nei dati centrali. Tuttavia, il numero elevato di punti rumorosi suggerisce che i parametri del metodo potrebbero essere ulteriormente ottimizzati per migliorare la qualità del clustering e ridurre il numero di punti esclusi (Deng, 2020; Bushra and Yi, 2021; Wang et al., 2022).

Figura 13. Statistiche per Density Based Clustering.



5.2 Fuzzy C-Means

Il modello di clustering fuzzy c-means ha identificato tre cluster nei dati, ciascuno con dimensioni diverse: 27, 12 e 11 punti per i cluster 1, 2 e 3 rispettivamente. La proporzione spiegata dell'eterogeneità intra-cluster mostra che il cluster 1 spiega il 48.3% della variabilità interna, mentre i cluster 2 e 3 spiegano il 28.7% e il 25.0%, rispettivamente. Questo suggerisce che il cluster 1 è il gruppo più denso e significativo, mentre gli altri due cluster hanno una struttura meno definita. I punteggi silhouette per ciascun cluster sono bassi, con il punteggio più alto pari a 0.208 per il cluster 1. Questi valori indicano che i cluster non sono ben separati e che potrebbe esserci sovrapposizione tra i gruppi. I centroidi dei cluster, rappresentati dai valori medi per ciascuna variabile, evidenziano

differenze nei profili dei cluster. Il cluster 1 ha valori negativi o vicini allo zero per quasi tutte le variabili, suggerendo che include punti dati con valori più bassi rispetto agli altri gruppi. Il cluster 2 presenta valori più elevati per Volume e Valore_Euro, indicando che rappresenta punti con una maggiore concentrazione su queste variabili. Il cluster 3 mostra valori ancora più alti per Valore_Assicurato e Valore_Euro, suggerendo che rappresenta un sottogruppo distinto di punti con caratteristiche specifiche legate a queste variabili. Le metriche di performance del modello indicano una qualità di clustering moderata. Il diametro massimo di un cluster è 5.628, mentre la separazione minima tra cluster è 0.385, suggerendo una sovrapposizione significativa tra i cluster. L'indice di Dunn pari a 0.085 conferma una scarsa compattezza e separazione tra i cluster. L'entropia del modello è 1.008, il che indica che i punti dati non appartengono esclusivamente a un singolo cluster, riflettendo la natura fuzzy del modello. L'indice di Calinski-Harabasz è pari a 12.346, un valore moderato che suggerisce una discreta qualità del clustering. Complessivamente, il modello ha identificato tre cluster con differenze evidenti nelle medie delle variabili, ma la qualità complessiva del clustering è limitata. I punteggi silhouette bassi e l'elevata sovrapposizione tra i cluster indicano che i gruppi non sono nettamente distinti. È possibile che i dati richiedano un preprocessing più approfondito o che il numero di cluster non sia ottimale per rappresentare la struttura intrinseca dei dati. Inoltre, i centroidi dei cluster suggeriscono che Valore_Euro, Volume e Valore_Assicurato giocano un ruolo chiave nella definizione dei gruppi, mentre altre variabili come Peso ed Emissioni_CO2 sembrano meno influenti. Ulteriori analisi potrebbero includere la variazione del numero di cluster o l'uso di metodi di clustering alternativi per migliorare la separazione e la significatività dei gruppi identificati (Tabella 8) (Zhao et al., 2023; Khan et al., 2021; Liu et al., 2022).

Tabella 8. Fuzzy C-Means Clustering.

Model Summary: Fuzzy C-Means Clustering

Clusters	N	R^2	AIC	BIC	Silhouette
3	50	0.250	228.740	263.160	0.170

Note. The model is optimized with respect to the *BIC* value.

Cluster Information

Cluster	1	2	3
Size	27	12	11
Explained proportion within-cluster heterogeneity	0.483	0.267	0.250
Within sum of squares	93.011	51.486	48.246
Silhouette score	0.208	0.100	0.167
Center Volume	-0.597	0.250	0.447
Center Valore_Euro	0.211	0.554	0.463
Center Peso	0.519	0.453	0.260
Center Costo_Spedizione	-0.144	0.523	0.556
Center Valore_Assicurato	-0.712	0.929	1.071
Center Emissioni_CO2	-0.195	-0.754	0.507

Note. The Between Sum of Squares of the 3 cluster model is 64.26
Note. The Total Sum of Squares of the 3 cluster model is 257.01

Model Performance Metrics

	Value
Maximum diameter	5.628

Model Performance Metrics

	Value
Minimum separation	0.365
Pearson's γ	0.385
Dunn index	0.065
Entropy	1.008
Calinski-Harabasz index	12.346

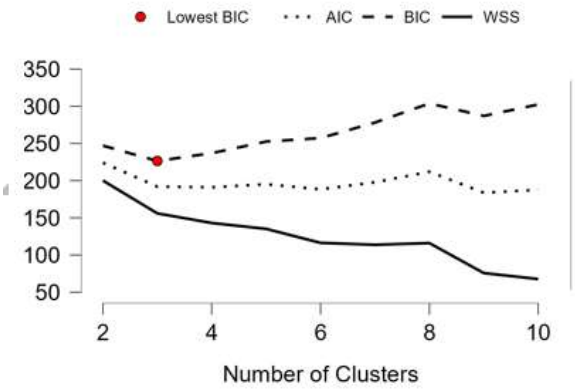
Note. All metrics are based on the *euclidean* distance.

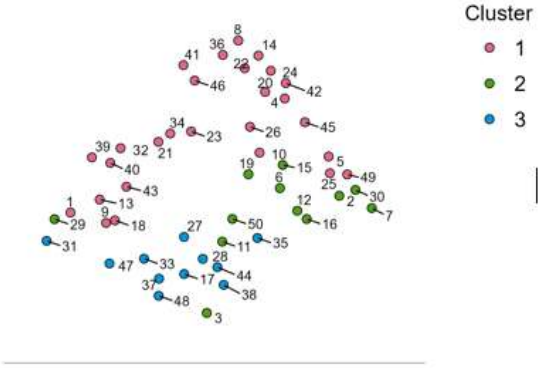
Cluster Means

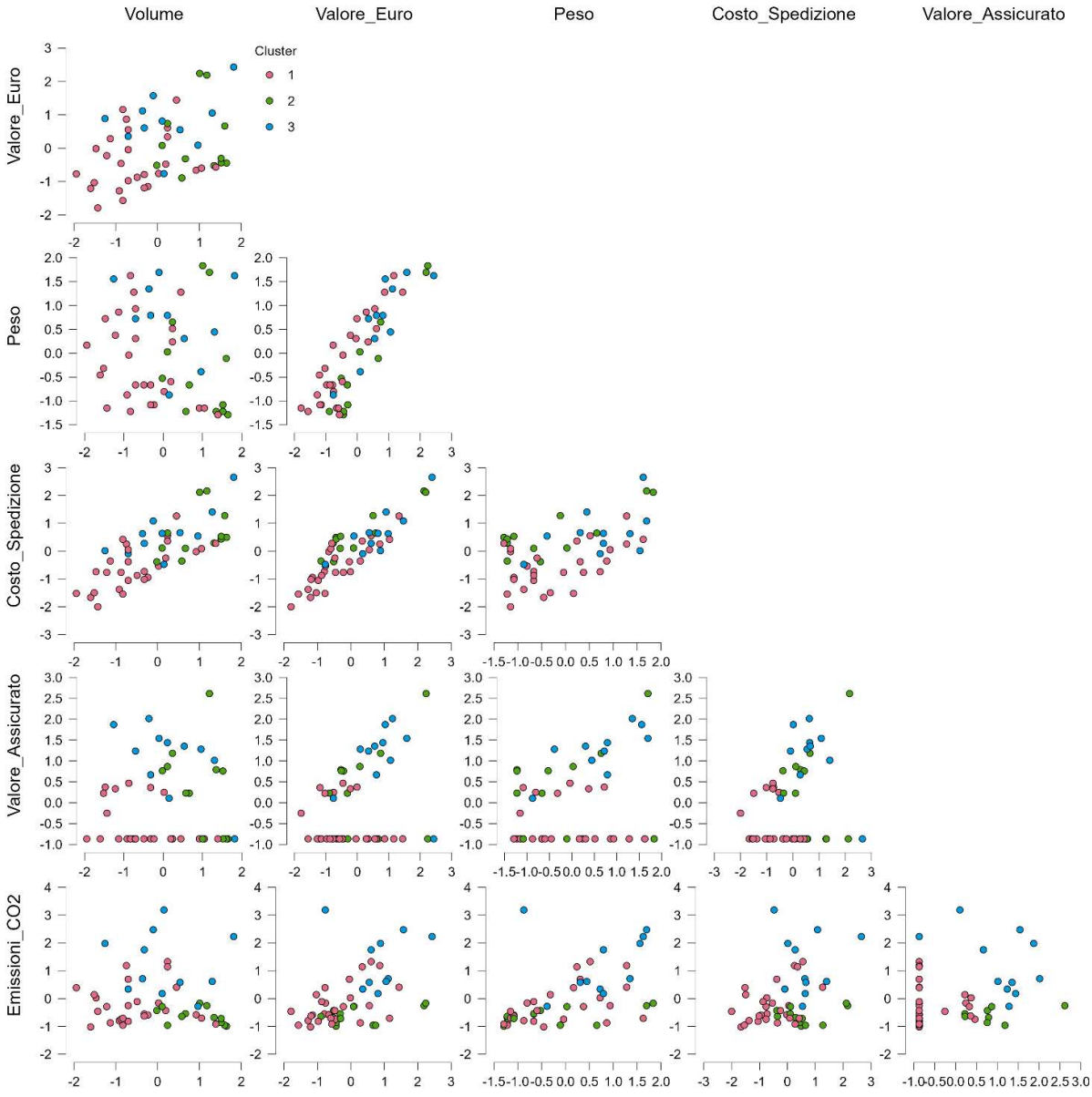
	Volume	Valore_Euro	Peso	Costo_Spedizione	Valore_Assicurato	Emissioni_CO2
Cluster 1	-0.501	-0.414	-0.182	-0.547	-0.578	-0.233
Cluster 2	0.949	0.206	-0.260	0.619	0.330	-0.625
Cluster 3	0.193	0.792	0.730	0.667	1.060	1.253

Il grafico in Figura 14 rappresenta i risultati di un modello di clustering fuzzy c-means con l'identificazione di tre cluster principali. Nel pannello in alto a sinistra, vengono valutati diversi numeri di cluster sulla base di criteri statistici come il Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC) e Within-Cluster Sum of Squares (WSS). Il valore ottimale, corrispondente al punto rosso, indica che la scelta di tre cluster rappresenta il miglior compromesso tra complessità del modello e capacità di rappresentare i dati. La decrescita del WSS suggerisce che l'aggiunta di cluster riduce l'eterogeneità interna, ma i valori crescenti di BIC e AIC oltre il terzo cluster indicano che ulteriori suddivisioni non migliorano significativamente il modello. Il grafico in alto a destra visualizza la distribuzione spaziale dei punti nei tre cluster (rosso, verde e blu). I cluster appaiono relativamente distinti nello spazio bidimensionale, anche se alcune sovrapposizioni sono visibili. La distribuzione irregolare dei punti suggerisce che i cluster catturano bene le differenze principali nei dati, ma alcuni punti sembrano collocarsi al margine tra i gruppi, evidenziando la natura fuzzy dell'algoritmo. La parte inferiore del grafico mostra una matrice di scatter plot, con la distribuzione dei punti per ciascuna combinazione di variabili. I punti sono colorati in base al cluster di appartenenza, offrendo un'analisi visiva delle relazioni tra variabili all'interno dei cluster. Le variabili Volume e Valore_Euro mostrano una chiara separazione tra i cluster, suggerendo che queste dimensioni sono le più influenti nella formazione dei gruppi. In particolare, il cluster verde sembra avere valori più elevati per Volume e Valore_Euro rispetto agli altri gruppi, mentre il cluster rosso è caratterizzato da valori medi più bassi su quasi tutte le variabili. Per Peso e Costo_Spedizione, la separazione è meno netta, indicando una minore rilevanza di queste variabili nella differenziazione dei cluster. Infine, variabili come Emissioni_CO2 e Valore_Assicurato mostrano una separazione visibile, ma con maggiore sovrapposizione rispetto alle prime due variabili principali. Complessivamente, il modello di clustering fuzzy c-means è riuscito a identificare tre gruppi con differenze significative nelle caratteristiche principali, in particolare per Volume e Valore_Euro. Tuttavia, la presenza di sovrapposizioni tra cluster e di punti vicini ai confini indica che le transizioni tra i gruppi non sono nette, come ci si aspetta da un algoritmo fuzzy. Questo approccio offre una rappresentazione utile per i dati con strutture complesse, ma potrebbe beneficiare di un'ulteriore analisi per ottimizzare la separazione e verificare la robustezza dei cluster identificati. L'aggiunta di variabili o un preprocessing più approfondito potrebbe migliorare ulteriormente la qualità della segmentazione (Irfiyanda et al., 2022; Xiong et al., 2020).

Figura 14. Fuzzy c-Means clustering.

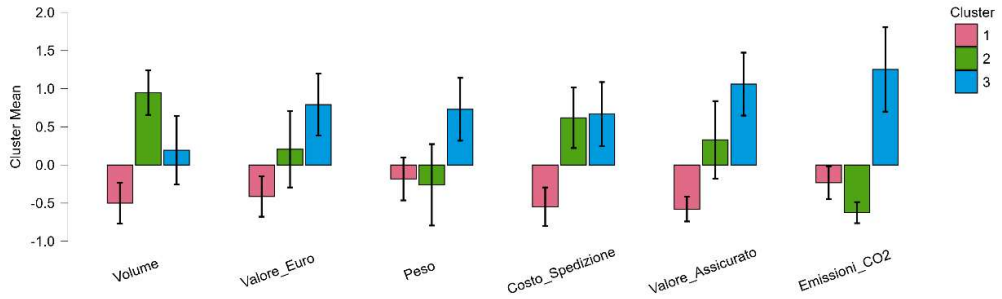

Lowest BIC
AIC
BIC
WSS
350
300
250
200
150
100
50
2
4
6
8
10
Number of Clusters

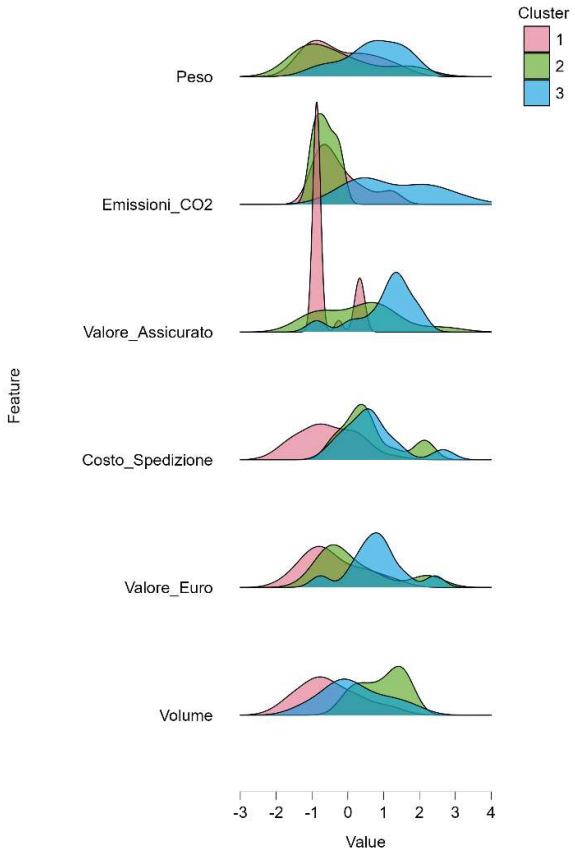

Cluster
1
2
3


Volume
Valore_Euro
Peso
Costo_Spedizione
Valore_Assicurato
Cluster
1
2
3
Valore_Euro
Peso
Costo_Spedizione
Valore_Assicurato
Emissioni_CO2

Il grafico in Figura 15 rappresenta un'analisi dei risultati di clustering ottenuti con il metodo Fuzzy C-Means. La parte superiore mostra un confronto delle medie dei cluster per ciascuna variabile tramite un grafico a barre con intervalli di errore. I cluster (indicati dai colori rosa, verde e blu per cluster 1, 2 e 3, rispettivamente) evidenziano differenze significative nelle variabili, con il cluster 2 che tende ad avere valori più elevati per molte delle variabili, mentre il cluster 1 presenta generalmente valori inferiori. Per esempio, il cluster 2 mostra i valori più alti per Volume, Valore_Euro e Costo_Spedizione, mentre il cluster 1 ha i valori più bassi su queste stesse variabili. Il cluster 3, invece, si distingue per valori intermedi, con una maggiore variabilità rispetto agli altri due gruppi. La variabile Peso mostra una differenza meno marcata tra i cluster, indicando che ha un'influenza minore nella formazione dei gruppi rispetto a variabili come Valore_Euro e Volume. La sezione inferiore del grafico mostra le distribuzioni di densità delle variabili per ciascun cluster. Le curve sovrapposte evidenziano come i valori delle variabili siano distribuiti tra i cluster. Per esempio, la variabile Valore_Euro presenta una chiara separazione, con il cluster 2 che domina i valori più elevati, mentre il cluster 1 si concentra su valori bassi. La variabile Emissioni_CO2, invece, mostra una maggiore sovrapposizione tra i cluster, suggerendo che non contribuisce in modo significativo alla separazione dei gruppi. Allo stesso modo, Costo_Spedizione e Valore_Assicurato presentano differenze visibili tra i cluster, ma con una sovrapposizione che ne riduce l'utilità come dimensioni discriminanti. Queste osservazioni riflettono la natura fuzzy del clustering, in cui i punti non appartengono esclusivamente a un singolo cluster ma hanno probabilità di appartenenza distribuite su più gruppi. Questo approccio è utile per analizzare dati con confini sfumati tra i gruppi, consentendo una comprensione più dettagliata delle relazioni sottostanti. In sintesi, l'analisi suggerisce che Valore_Euro, Volume e Costo_Spedizione sono le variabili più rilevanti nella definizione dei cluster, con una chiara separazione tra i gruppi. Tuttavia, la sovrapposizione in variabili come Peso ed Emissioni_CO2 indica che non tutte le dimensioni contribuiscono ugualmente alla segmentazione. Questo risultato potrebbe essere ulteriormente migliorato ottimizzando il numero di cluster o applicando tecniche di selezione delle caratteristiche per ridurre la complessità del modello (Dey et al., 2021; Pantula et al., 2020; Chang et al., 2023).

Figura 15. Fuzzy c-Means Clustering.





5.3 Hierarchical

L'analisi riportata si riferisce a un modello di clustering gerarchico che ha identificato sei cluster nei dati. La distribuzione dei punti all'interno dei cluster varia notevolmente, con il cluster 4 che contiene il maggior numero di punti (20) e il cluster 1 il numero più basso (3). L'eterogeneità intra-cluster varia significativamente, con i cluster più piccoli che spiegano una proporzione minima della variabilità, mentre il cluster 4, essendo il più grande, spiega una porzione maggiore, anche se non particolarmente elevata in termini assoluti. I punteggi silhouette sono anch'essi variabili tra i cluster: il cluster 1 ha il valore più alto (0.415), indicando una buona coesione interna e separazione dagli altri gruppi, mentre il cluster 3 ha un valore pari a zero, suggerendo che i suoi membri potrebbero sovrapporsi con altri cluster. Le metriche di performance del modello indicano una qualità di clustering moderata. Il diametro massimo del cluster è pari a 3.396, mentre la separazione minima tra cluster è 0.947, valori che suggeriscono una certa distanza tra i gruppi, ma non una separazione netta. L'indice di Dunn, pari a 0.279, riflette una scarsa compattezza e separazione complessiva. L'entropia del modello, pari a 1.519, suggerisce una distribuzione moderatamente dispersa dei punti tra i cluster. Tuttavia, l'indice di Calinski-Harabasz è relativamente alto (24.729), indicando che il modello è stato in grado di catturare una struttura significativa nei dati. I valori medi per ciascun cluster evidenziano le differenze nelle caratteristiche principali. Il cluster 6 si distingue per i valori medi più elevati di Valore_Euro e Valore_Assicurato, indicando che i punti in questo cluster sono associati a queste variabili. Al contrario, i cluster 1 e 5 presentano valori medi negativi su quasi tutte le variabili, suggerendo che rappresentano gruppi con caratteristiche opposte. Il cluster 4, il più numeroso, mostra

valori medi prossimi allo zero, rappresentando un gruppo più equilibrato e meno estremo rispetto agli altri. I valori medi negativi per Peso, Volume e altre variabili in alcuni cluster indicano che questi gruppi sono composti da dati con caratteristiche specifiche e distinguibili. Complessivamente, il modello di clustering gerarchico è riuscito a identificare sei gruppi distinti nei dati, ma con una qualità complessiva moderata, come indicato dai punteggi silhouette e dalle altre metriche. La sovrapposizione tra alcuni cluster, evidenziata dai punteggi silhouette bassi per alcuni gruppi, suggerisce che i confini tra i cluster non sono ben definiti. Questo risultato potrebbe essere migliorato ottimizzando il numero di cluster o applicando tecniche di preprocessing per migliorare la separazione delle caratteristiche. La variazione delle dimensioni dei cluster e i loro valori medi indicano che il modello è stato in grado di catturare una diversità significativa nei dati, anche se con alcune limitazioni nella coesione interna (Tabella 9) (Ran et al., 2023; Eibeck et al., 2024; Abdalzaher et al., 2023).

Tabella 9. Hierarchical Clustering.

Model Summary: Hierarchical Clustering

Clusters	N	R^2	AIC	BIC	Silhouette
6	50	0.738	124.300	181.660	0.330

Note. The model is optimized with respect to the *BIC* value.

Cluster Information

Cluster	1	2	3	4	5	6
Size	3	9	1	20	7	10
Explained proportion within-cluster heterogeneity	0.041	0.178	0.000	0.512	0.079	0.190
Within sum of squares	2.641	11.427	0.000	32.903	5.090	12.243
Silhouette score	0.415	0.395	0.000	0.252	0.478	0.349

Note. The Between Sum of Squares of the 6 cluster model is 180.7
Note. The Total Sum of Squares of the 6 cluster model is 245

Model Performance Metrics

	Value
Maximum diameter	3.396
Minimum separation	0.947
Pearson's γ	0.555
Dunn index	0.279
Entropy	1.519
Calinski-Harabasz index	24.729

Note. All metrics are based on the *euclidean* distance.

Cluster Means

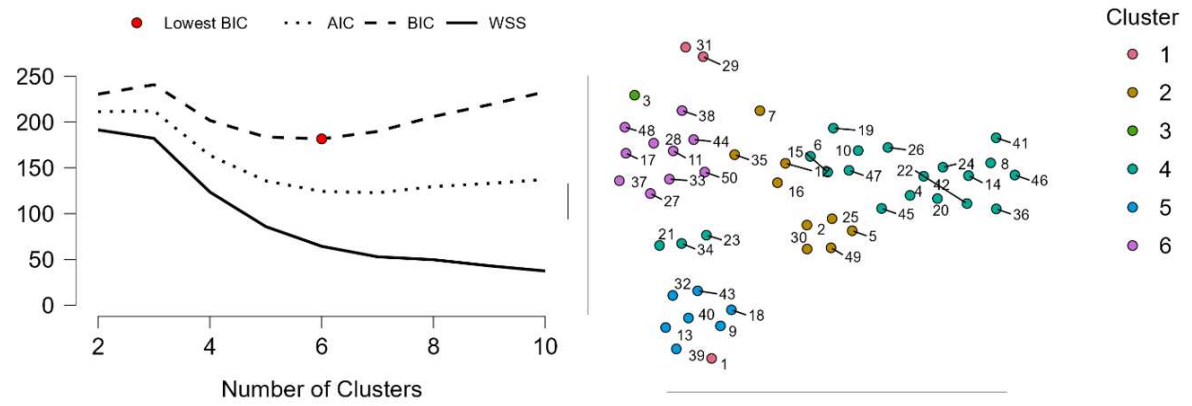
	Peso	Volume	Valore_Euro	Costo_Spedizione	Valore_Assicurato
Cluster 1	1.580	1.092	2.037	2.011	-0.868
Cluster 2	-0.990	1.329	-0.310	0.435	-0.264
Cluster 3	1.696	1.177	2.189	2.163	2.615
Cluster 4	-0.582	-0.615	-0.852	-0.920	-0.237
Cluster 5	0.823	-0.522	0.538	0.130	-0.868

Cluster Means

	Peso	Volume	Valore_Euro	Costo_Spedizione	Valore_Assicurato
Cluster 6	0.835	-0.046	0.777	0.538	1.319

Il grafico in Figura 16 presenta i risultati di un'analisi di clustering gerarchico, evidenziando il numero ottimale di cluster e la distribuzione dei punti nei gruppi. Nel pannello a sinistra, sono visualizzate tre metriche utilizzate per determinare il numero ideale di cluster: il Bayesian Information Criterion (BIC), l'Akaike Information Criterion (AIC) e la somma dei quadrati intra-cluster (WSS). Il punto rosso indica il numero ottimale di cluster (sei), in corrispondenza del valore minimo del BIC. Questo suggerisce che la scelta di sei cluster rappresenta il miglior compromesso tra complessità del modello e capacità di rappresentare la struttura dei dati. Il pannello a destra mostra la distribuzione spaziale dei punti assegnati ai sei cluster. I punti sono colorati in base al cluster di appartenenza, evidenziando una chiara separazione visiva tra alcuni gruppi, mentre altri cluster presentano una certa sovrapposizione. In particolare, i cluster 1, 2 e 4 sembrano avere una buona coesione interna e una separazione chiara dagli altri gruppi, mentre i cluster 3 e 6 mostrano una maggiore dispersione. La distribuzione irregolare dei punti nei gruppi riflette le diverse densità e caratteristiche delle regioni nei dati. La combinazione di una WSS decrescente con valori crescenti di BIC e AIC per numeri di cluster superiori a sei suggerisce che ulteriori suddivisioni non migliorano significativamente la qualità del clustering, ma aumentano la complessità del modello. Questo equilibrio tra semplificazione e rappresentazione accurata è un elemento chiave nella scelta del numero di cluster in un'analisi gerarchica. In sintesi, il modello di clustering gerarchico con sei cluster offre una rappresentazione utile dei dati, catturando variazioni significative con una buona separazione visiva tra alcuni gruppi. Tuttavia, la sovrapposizione tra alcuni cluster e la distribuzione non uniforme dei punti suggeriscono che i dati potrebbero beneficiare di ulteriori analisi o di un preprocessamento mirato per migliorare la separazione e la compattezza dei cluster (Zhou et al., 2021; Xu et al., 2020; Nguyen et al., 2021).

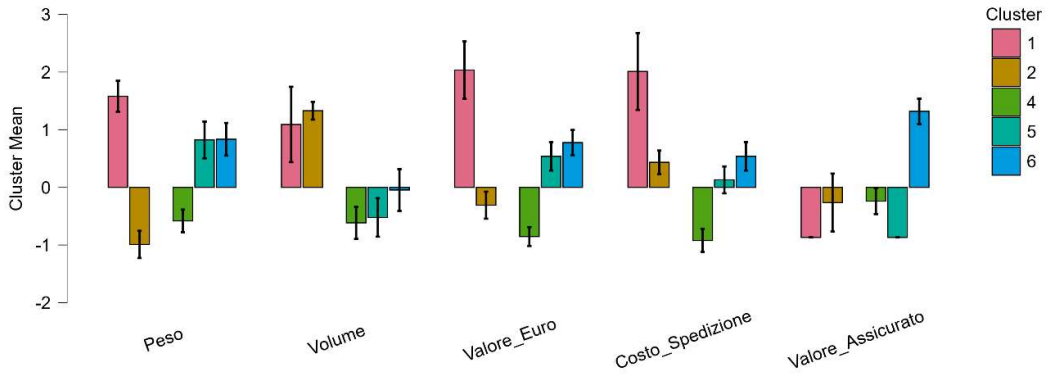
Figura 16. Statistiche relative allo hierarchical clustering.



Il grafico in Figura 17 mostra i valori medi delle variabili considerate per ciascun cluster individuato tramite il clustering gerarchico. Ogni cluster è rappresentato da una barra colorata, con intervalli di errore che indicano la variabilità all'interno del gruppo. L'analisi evidenzia differenze significative nelle variabili tra i cluster, riflettendo la segmentazione ottenuta. Il cluster 1 si distingue per i valori medi più alti su quasi tutte le variabili, in particolare su Peso, Volume, e Valore_Euro, suggerendo che rappresenti un gruppo di osservazioni con caratteristiche estreme. In contrapposizione, il cluster 4 presenta valori medi nettamente negativi per la maggior parte delle variabili, posizionandosi come l'opposto del cluster 1. Questi due cluster sembrano rappresentare i gruppi più distinti nel dataset. Il cluster 6 mostra valori moderatamente positivi per alcune variabili, come Valore_Assicurato, mentre mantiene un profilo più bilanciato rispetto a variabili come Peso e Volume, indicando che potrebbe

rappresentare un segmento intermedio o più omogeneo. Al contrario, il cluster 5 mostra valori medi prevalentemente negativi, con una variabilità relativamente bassa, suggerendo una maggiore coerenza interna. Il cluster 2 è caratterizzato da valori medi vicini allo zero per la maggior parte delle variabili, indicando che rappresenta un gruppo equilibrato e meno estremo rispetto agli altri. Tuttavia, mostra una leggera positività per Costo_Spedizione, distinguendosi leggermente da altri cluster con valori più negativi. Il cluster 3 presenta valori moderatamente positivi per Volume e Valore_Euro, ma negativi per altre variabili, come Costo_Spedizione e Valore_Assicurato, suggerendo una distribuzione più specifica che lo distingue dagli altri gruppi. Le differenze tra i cluster suggeriscono che le variabili Peso, Volume, e Valore_Euro siano i principali driver della segmentazione, mentre Valore_Assicurato e Costo_Spedizione mostrano una variabilità meno marcata. L'uso di intervalli di errore fornisce un'indicazione della coesione interna di ciascun cluster: i cluster con barre più lunghe, come il 6 e il 5, potrebbero avere una maggiore dispersione interna rispetto a quelli con barre più corte, come il cluster 1. Complessivamente, il clustering gerarchico ha identificato gruppi con profili distinti che riflettono variazioni significative nelle variabili. Tuttavia, la presenza di sovrapposizioni nei valori medi suggerisce che alcune variabili potrebbero avere un impatto minore nella distinzione tra gruppi. Questo risultato potrebbe essere ulteriormente raffinato aggiungendo più variabili o applicando tecniche complementari di analisi (Yang et al., 2020; Côme et al., 2021; Mandal et al., 2022).

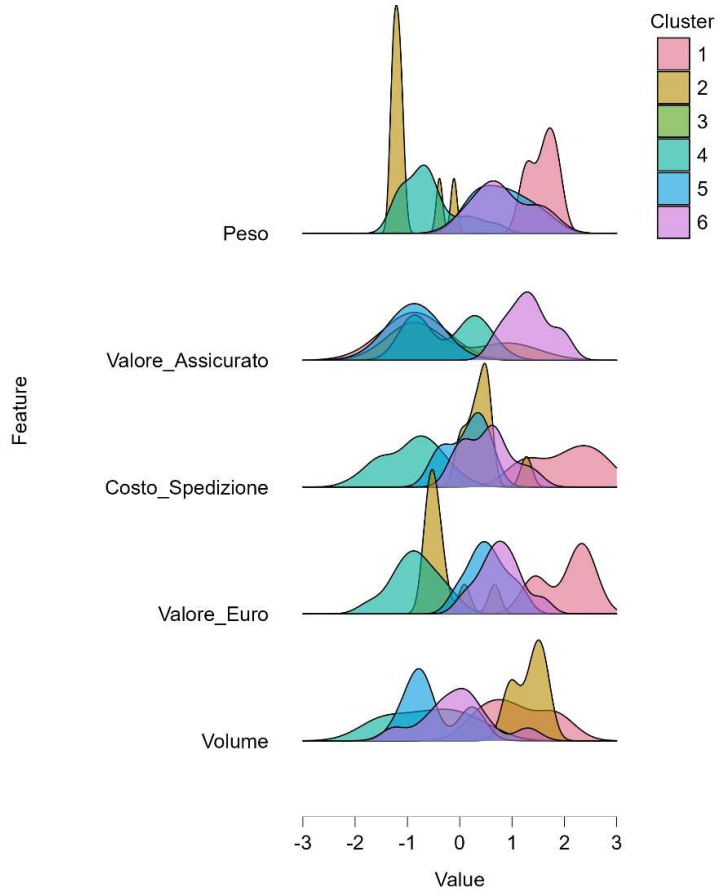
Figura 17. Statistiche Hierarchical Clustering.



Il grafico in Figura 18 mostra le distribuzioni delle variabili all'interno dei cluster identificati tramite hierarchical clustering. Ogni curva rappresenta la densità di probabilità di una variabile per un determinato cluster, evidenziando come le osservazioni siano distribuite nei vari gruppi. I cluster sono rappresentati da colori distinti per facilitare il confronto. Per la variabile Peso, i cluster 1 e 4 mostrano distribuzioni molto distinte: il cluster 1 presenta una densità concentrata su valori positivi elevati, mentre il cluster 4 è caratterizzato da valori prevalentemente negativi. Gli altri cluster, come il 2 e il 6, mostrano distribuzioni più vicine allo zero, indicando gruppi meno estremi rispetto alle osservazioni. La variabile Valore_Assicurato evidenzia una sovrapposizione maggiore tra i cluster, specialmente tra i cluster 5 e 6, suggerendo che questa variabile non contribuisce significativamente alla separazione dei gruppi. Tuttavia, il cluster 4 si distingue con una densità concentrata verso valori negativi, mentre il cluster 1 mantiene valori leggermente positivi. Nel caso del Costo_Spedizione, il cluster 1 domina i valori positivi elevati, mentre il cluster 4 presenta una distribuzione completamente negativa. I cluster rimanenti, come il 5 e il 6, si sovrappongono in una fascia centrale, indicando che questa variabile contribuisce in modo moderato alla definizione dei gruppi. La variabile Valore_Euro mostra una buona separazione tra i cluster. In particolare, il cluster 1 si distingue con valori positivi elevati, mentre il cluster 4 mantiene valori negativi. Anche il cluster 6 mostra una distribuzione ben definita verso valori leggermente negativi, mentre i cluster 2 e 5 si sovrappongono maggiormente in

una fascia intermedia. Per la variabile Volume, si osserva una separazione interessante tra il cluster 1, che domina i valori positivi elevati, e il cluster 4, che mantiene una distribuzione negativa. I cluster 2 e 6 presentano valori più bilanciati, mentre il cluster 5 ha una distribuzione più diffusa, indicando una maggiore variabilità interna. Complessivamente, le variabili Peso, Valore_Euro e Costo_Spedizione sembrano avere un ruolo chiave nella distinzione dei cluster, offrendo una separazione più chiara tra i gruppi. Tuttavia, altre variabili, come Valore_Assicurato, mostrano sovrapposizioni significative tra alcuni cluster, riducendo la loro capacità discriminante. Questo risultato suggerisce che i cluster identificati riflettono differenze strutturali nei dati, ma ulteriori analisi potrebbero migliorare la comprensione dei fattori che guidano la segmentazione (Thamrin and Wijayanto, 2021; Yang et al., 2020; Dahl et al., 2023).

Figura 18. Distribuzioni Hierarchical Clustering.



5.4 Model Based Clusters

I dati presentati riguardano un'analisi di clustering basata su modelli, che ha segmentato il dataset in quattro cluster, ognuno con caratteristiche specifiche. Il modello è stato ottimizzato utilizzando il criterio BIC, raggiungendo un valore di 172.610. Il coefficiente di determinazione (R^2) di 0.614 indica che il modello riesce a spiegare una porzione moderata della varianza totale dei dati, ma lascia spazio a una possibile ottimizzazione. Il punteggio medio di silhouette è 0.270, suggerendo una coesione interna e una separazione tra cluster limitate, con margini di miglioramento. Il Cluster 1 è composto da sei osservazioni ed è caratterizzato da valori medi positivi per tutte le variabili, con valori elevati per peso, valore economico e costo di spedizione. Il punteggio silhouette di 0.170 suggerisce una bassa coesione interna e una possibile sovrapposizione con gli altri cluster. La proporzione di eterogeneità intra-cluster spiegata è dello 0.195, segnalando che il cluster è moderatamente

omogeneo. Il Cluster 2 comprende otto osservazioni ed è caratterizzato da valori medi negativi per tutte le variabili, in particolare volume e valore economico. Questo cluster rappresenta osservazioni che si distinguono per i loro bassi valori. Tuttavia, il punteggio silhouette di 0.549 indica una migliore coesione interna rispetto al Cluster 1, evidenziando una segmentazione più definita. Il Cluster 3 è il più numeroso, con 19 osservazioni, ed è caratterizzato da valori medi negativi su tutte le variabili, sebbene meno estremi rispetto al Cluster 2. Questo cluster rappresenta un gruppo più ampio di osservazioni con profili relativamente omogenei. Tuttavia, il punteggio silhouette di 0.147 è basso, suggerendo sovrapposizioni con gli altri cluster. Il Cluster 4 comprende 17 osservazioni, con valori medi negativi su tutte le variabili. Il punteggio silhouette di 0.314 indica una coesione interna moderata. Questo cluster si distingue per una varianza intra-cluster relativamente elevata, riflettendo una certa dispersione tra le osservazioni. Le metriche di performance del modello, come l'indice di Calinski-Harabasz pari a 24.475 e l'entropia di 1.282, suggeriscono una qualità di clustering moderata. La distanza massima tra punti all'interno dei cluster è di 4.488, indicando la presenza di alcune osservazioni distanti che potrebbero influire sulla compattezza dei cluster. In sintesi, il modello di clustering basato su modelli ha identificato quattro gruppi distinti, ognuno con caratteristiche specifiche. Tuttavia, i punteggi silhouette relativamente bassi e la varianza intra-cluster elevata per alcuni gruppi suggeriscono che la segmentazione potrebbe essere migliorata. Una possibile soluzione potrebbe essere quella di esplorare un diverso numero di cluster o di affinare i parametri del modello per ottenere una segmentazione più precisa e ben definita (Tabella 10) (Altieri et al., 2021; García-Ordás et al., 2023; Gergely and Vargha, 2021).

Tabella 10. Model Based Clustering.

Model Summary: Model-Based Clustering

Clusters	N	R^2	AIC	BIC	Silhouette
4	50	0.614	134.370	172.610	0.270

Note. The model is diagonal with varying volume and equal shape.
Note. The model is optimized with respect to the *BIC* value.

Cluster Information

Cluster	1	2	3	4
Size	6	8	19	17
Explained proportion within-cluster heterogeneity	0.195	0.061	0.496	0.248
Within sum of squares	18.438	5.719	46.847	23.365
Silhouette score	0.170	0.549	0.147	0.314

Note. The Between Sum of Squares of the 4 cluster model is 150.35
Note. The Total Sum of Squares of the 4 cluster model is 244.72

Model Performance Metrics

	Value
Maximum diameter	4.488
Minimum separation	0.603
Pearson's γ	0.512
Dunn index	0.134
Entropy	1.282
Calinski-Harabasz index	24.475

Model Performance Metrics

	Value

Note. All metrics are based on the *euclidean* distance.

Cluster Means

	Peso	Volume	Valore_Euro	Costo_Spedizione	Valore_Assicurato
Cluster 1	1.430	0.942	1.821	1.781	0.428
Cluster 2	-1.134	1.257	-0.483	0.274	-0.321
Cluster 3	0.662	-0.286	0.496	0.216	0.299
Cluster 4	-0.710	-0.604	-0.970	-0.999	-0.334

I dati presentati si riferiscono al clustering basato su modelli, con quattro componenti che rappresentano gruppi distinti all'interno del dataset. Le probabilità di miscelazione indicano la proporzione stimata di osservazioni appartenenti a ciascun componente. Il Component 1 ha una probabilità di miscelazione dello 0.118, il che suggerisce che rappresenta un piccolo sottogruppo di osservazioni con valori medi positivi per tutte le variabili, in particolare per il valore economico e il costo di spedizione. La matrice di covarianza per il Component 1 mostra una varianza relativamente alta per il peso e il valore assicurato, indicando una maggiore variabilità interna per queste variabili. Il Component 2 ha una probabilità di miscelazione dello 0.162 e rappresenta un gruppo con valori medi negativi per quasi tutte le variabili, eccetto per il volume. Questo cluster si distingue per una maggiore variabilità interna nel volume e nel costo di spedizione, come evidenziato dalla matrice di covarianza. Questo suggerisce che, all'interno di questo cluster, esiste una certa dispersione nei valori relativi a queste variabili. Il Component 3 è il più grande, con una probabilità di miscelazione dello 0.380, e rappresenta un gruppo con valori medi leggermente negativi per quasi tutte le variabili. La matrice di covarianza mostra una varianza relativamente alta per il peso e il valore assicurato, indicando una moderata dispersione dei dati. Questo componente può rappresentare un cluster generalizzato con una maggiore omogeneità rispetto agli altri gruppi. Il Component 4 ha una probabilità di miscelazione dello 0.340 e si caratterizza per valori medi negativi per tutte le variabili. Tuttavia, il valore economico e il costo di spedizione mostrano le medie più basse, suggerendo che questo cluster rappresenta un gruppo distinto con osservazioni che hanno caratteristiche significativamente inferiori rispetto agli altri gruppi. La matrice di covarianza indica una minore variabilità interna rispetto agli altri cluster, suggerendo una maggiore compattezza interna. Complessivamente, le probabilità di miscelazione riflettono una distribuzione non uniforme delle osservazioni tra i cluster, con il Component 3 che domina rispetto agli altri. Le medie e le matrici di covarianza forniscono indicazioni dettagliate sulla natura di ciascun cluster, evidenziando differenze significative nei valori medi e nella variabilità interna. Questo suggerisce che il clustering basato su modelli è stato in grado di identificare gruppi distinti, ma i risultati potrebbero beneficiare di ulteriori ottimizzazioni o verifiche rispetto ad altri metodi di clustering per garantire la robustezza delle conclusioni (Tabella 11) (Zhuang et al., 2023; Choi et al., 2023; Yang et al., 2020).

Tabella 11. Statistics for Model Based Clustering.

Mixing Probabilities

	Mixing probability
Component 1	0.118
Component 2	0.162
Component 3	0.380

Mixing Probabilities

	Mixing probability
Component 4	0.340

Means

	Peso	Volume	Valore_Euro	Costo_Spedizione	Valore_Assicurato
Component 1	1.449	0.917	1.827	1.773	0.398
Component 2	-1.136	1.249	-0.489	0.266	-0.315
Component 3	0.660	-0.264	0.505	0.234	0.309
Component 4	-0.698	-0.617	-0.964	-1.002	-0.332

Covariance Matrix for Component 1

	Peso	Volume	Valore_Euro	Costo_Spedizione	Valore_Assicurato
Peso	0.281	0.000	0.000	0.000	0.000
Volume	0.000	0.727	0.000	0.000	0.000
Valore_Euro	0.000	0.000	0.192	0.000	0.000
Costo_Spedizione	0.000	0.000	0.000	0.354	0.000
Valore_Assicurato	0.000	0.000	0.000	0.000	1.466

Covariance Matrix for Component 2

	Peso	Volume	Valore_Euro	Costo_Spedizione	Valore_Assicurato
Peso	0.047	0.000	0.000	0.000	0.000
Volume	0.000	0.121	0.000	0.000	0.000
Valore_Euro	0.000	0.000	0.032	0.000	0.000
Costo_Spedizione	0.000	0.000	0.000	0.059	0.000
Valore_Assicurato	0.000	0.000	0.000	0.000	0.245

Covariance Matrix for Component 3

	Peso	Volume	Valore_Euro	Costo_Spedizione	Valore_Assicurato
Peso	0.236	0.000	0.000	0.000	0.000
Volume	0.000	0.610	0.000	0.000	0.000
Valore_Euro	0.000	0.000	0.161	0.000	0.000
Costo_Spedizione	0.000	0.000	0.000	0.297	0.000
Valore_Assicurato	0.000	0.000	0.000	0.000	1.231

Covariance Matrix for Component 4

	Peso	Volume	Valore_Euro	Costo_Spedizione	Valore_Assicurato
Peso	0.162	0.000	0.000	0.000	0.000
Volume	0.000	0.418	0.000	0.000	0.000
Valore_Euro	0.000	0.000	0.110	0.000	0.000
Costo_Spedizione	0.000	0.000	0.000	0.203	0.000
Valore_Assicurato	0.000	0.000	0.000	0.000	0.843

Scale of the Covariance

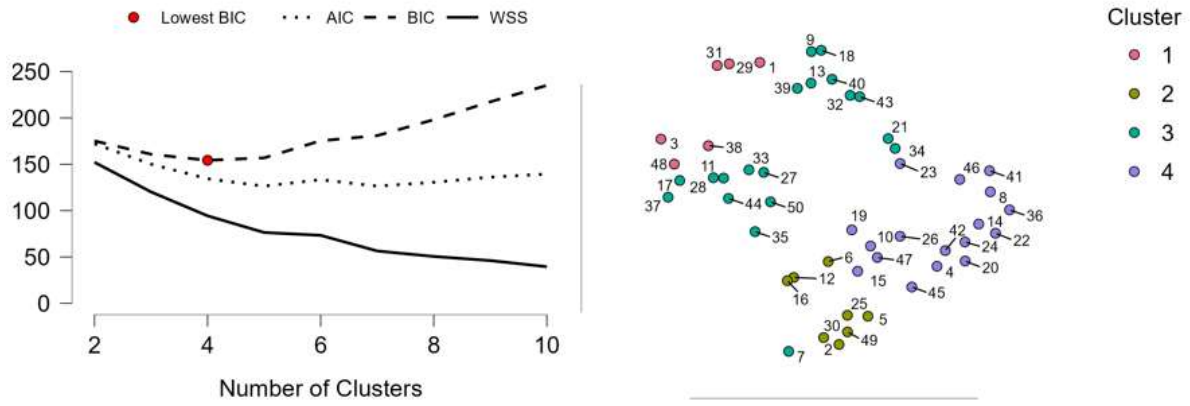
	Scale
Component 1	0.459
Component 2	0.077
Component 3	0.385
Component 4	0.264

Shape of the Covariance Matrix

	Peso	Volume	Valore_Euro	Costo_Spedizione	Valore_Assicurato
Component 1	0.613	1.584	0.418	0.771	3.196
Component 2	0.613	1.584	0.418	0.771	3.196
Component 3	0.613	1.584	0.418	0.771	3.196
Component 4	0.613	1.584	0.418	0.771	3.196

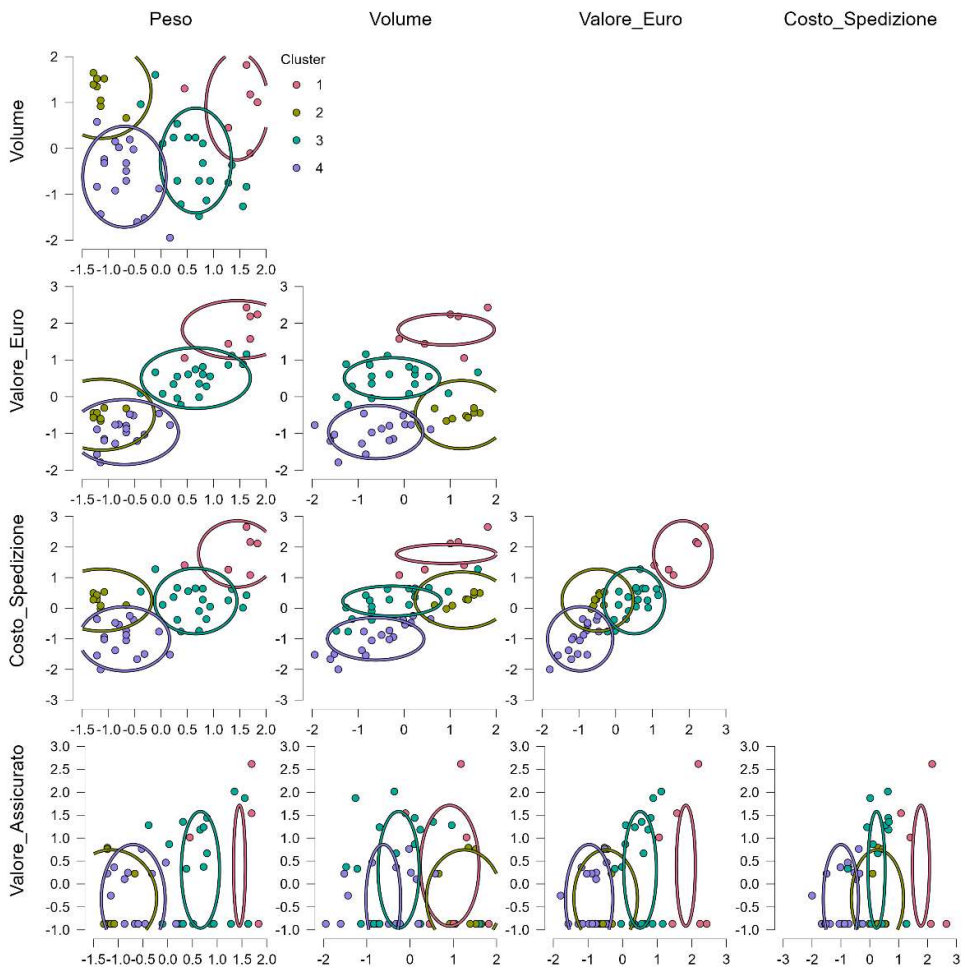
I dati rappresentati in figura 19 relativi al clustering basato su modelli evidenziano la selezione di quattro cluster come configurazione ottimale, in base ai criteri di valutazione presentati. Nel grafico a sinistra, il numero di cluster è determinato considerando il valore minimo del Bayesian Information Criterion (BIC), rappresentato dal punto rosso. Il calo del valore del BIC fino a quattro cluster suggerisce che questa configurazione rappresenta un buon compromesso tra la complessità del modello e l'aderenza ai dati. Dopo quattro cluster, il BIC inizia ad aumentare, indicando un'eventuale sovrassegmentazione. Il grafico a destra visualizza la distribuzione delle osservazioni nei quattro cluster. Le osservazioni sono distribuite in modo relativamente bilanciato, ma alcuni cluster, come il Cluster 1 (rosso), sembrano contenere meno punti rispetto agli altri. Questo suggerisce che il Cluster 1 potrebbe rappresentare un sottogruppo di dati con caratteristiche uniche o meno rappresentate nella popolazione complessiva. I Cluster 3 (verde) e 4 (viola) contengono un numero maggiore di osservazioni, indicando una maggiore densità o eterogeneità interna. La rappresentazione spaziale dei cluster mostra che ciascun gruppo è ben separato, anche se ci sono alcune osservazioni vicine ai confini tra i cluster, che potrebbero essere influenzate da sovrapposizioni nei dati. Questo è comune nei metodi di clustering basati su modelli, dove la separazione è influenzata dalla stima delle distribuzioni sottostanti. Complessivamente, i dati suggeriscono che il modello basato su cluster ha identificato gruppi distinti e significativi, supportato dal miglior valore di BIC con quattro cluster. Tuttavia, un'analisi più approfondita delle caratteristiche interne dei cluster potrebbe essere utile per comprendere meglio le dinamiche sottostanti e verificare se i cluster riflettono strutture significative nei dati o se ci sono sovrapposizioni che potrebbero richiedere ulteriori ottimizzazioni (Yang and Wu, 2023; Gergely and Vargha, 2021; Ngatchou-Wandji and Bulla, 2011).

Figura 19. Dati relativi al model based clustering.



Il grafico mostra la distribuzione delle osservazioni in uno spazio bidimensionale, con i punti suddivisi in quattro cluster, evidenziati da ellissi che rappresentano la densità stimata per ciascun cluster. L'approccio utilizzato è il clustering basato su modelli, che assume che i dati siano generati da distribuzioni probabilistiche sottostanti. Le ellissi rappresentano quindi le aree in cui i punti appartenenti a ciascun cluster hanno la maggiore probabilità di trovarsi. Ogni dimensione dello spazio bidimensionale riflette una combinazione di variabili significative, come il peso, il volume, il valore economico o altri attributi. Le ellissi sovrapposte in alcuni grafici suggeriscono che ci possono essere sovrapposizioni tra cluster, il che implica che le osservazioni di questi gruppi potrebbero condividere alcune caratteristiche comuni. In particolare, il Cluster 1 sembra avere una distribuzione più isolata in alcune dimensioni, suggerendo che le osservazioni in questo gruppo siano significativamente diverse dagli altri cluster. D'altra parte, i Cluster 3 e 4 mostrano una maggiore sovrapposizione in alcune proiezioni, indicando che i dati in queste regioni potrebbero essere meno distinguibili. La distribuzione non uniforme delle dimensioni e delle posizioni delle ellissi suggerisce che i cluster presentano livelli diversi di variabilità interna. Alcuni gruppi, come il Cluster 2, mostrano ellissi più compatte, il che può indicare che i dati sono più omogenei all'interno del cluster. Al contrario, cluster come il Cluster 4 presentano ellissi più ampie, segnalando una maggiore variabilità interna o una possibile eterogeneità nelle caratteristiche del gruppo. L'analisi complessiva dei dati mostra che il modello basato su cluster è riuscito a identificare gruppi significativi, ma alcune sovrapposizioni suggeriscono che potrebbe essere necessaria una valutazione aggiuntiva per determinare se i cluster catturano davvero strutture distinte nei dati o se vi è il rischio di errata classificazione (Figura 20) (Wu and Li, 2022; Rapp et al., 2020; Asamov and Ben-Israel, 2022).

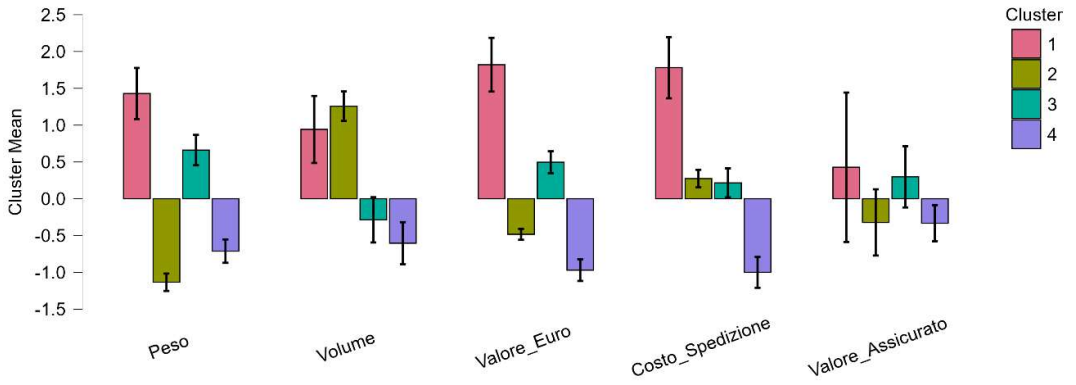
Figura 20. Dati relativi al model based clustering



Il grafico mostra le medie normalizzate delle caratteristiche per quattro cluster identificati attraverso un metodo di clustering. Ogni barra rappresenta la media di una variabile all'interno di un cluster specifico, con barre di errore che indicano la variabilità all'interno del cluster. Il Cluster 1 è caratterizzato da valori positivi predominanti in quasi tutte le variabili, suggerendo che le osservazioni in questo gruppo hanno valori significativamente superiori rispetto alla media normalizzata per tutte le caratteristiche considerate. In particolare, il peso, il volume e il valore economico sembrano avere un impatto notevole nel differenziare questo cluster dagli altri. Il Cluster 2, al contrario, mostra valori fortemente negativi per la maggior parte delle variabili, indicando che le osservazioni in questo gruppo sono generalmente al di sotto della media normalizzata, in particolare per il peso e il costo di spedizione. Il Cluster 3 presenta una combinazione di valori positivi e negativi, con una tendenza verso valori leggermente superiori alla media per il volume e il valore economico, ma al di sotto della media per il peso e il valore assicurato. Questo indica che le osservazioni in questo cluster potrebbero rappresentare un gruppo con caratteristiche miste o intermedie. Il Cluster 4 è caratterizzato da valori negativi in quasi tutte le variabili, con un impatto particolarmente forte per il costo di spedizione e il valore assicurato, suggerendo che le osservazioni in questo cluster rappresentano una categoria più omogenea e meno rilevante in termini di variabili chiave. Le barre di errore relativamente ampie per alcune variabili, in particolare nel Cluster 4, indicano una variabilità interna maggiore, il che potrebbe

riflettere un'eterogeneità significativa tra le osservazioni all'interno del cluster. In generale, il grafico suggerisce che le variabili considerate contribuiscono in modo differente alla definizione di ciascun cluster, con alcune variabili che sembrano essere più determinanti di altre nel separare i gruppi. Questo tipo di analisi è utile per comprendere quali caratteristiche influenzano maggiormente la formazione dei cluster e può guidare ulteriori indagini o interventi basati sui dati (Figura 21) (Singh and Singh, 2020; Starovoitov and Golub, 2021; de Amorim and Makarenkov, 2021).

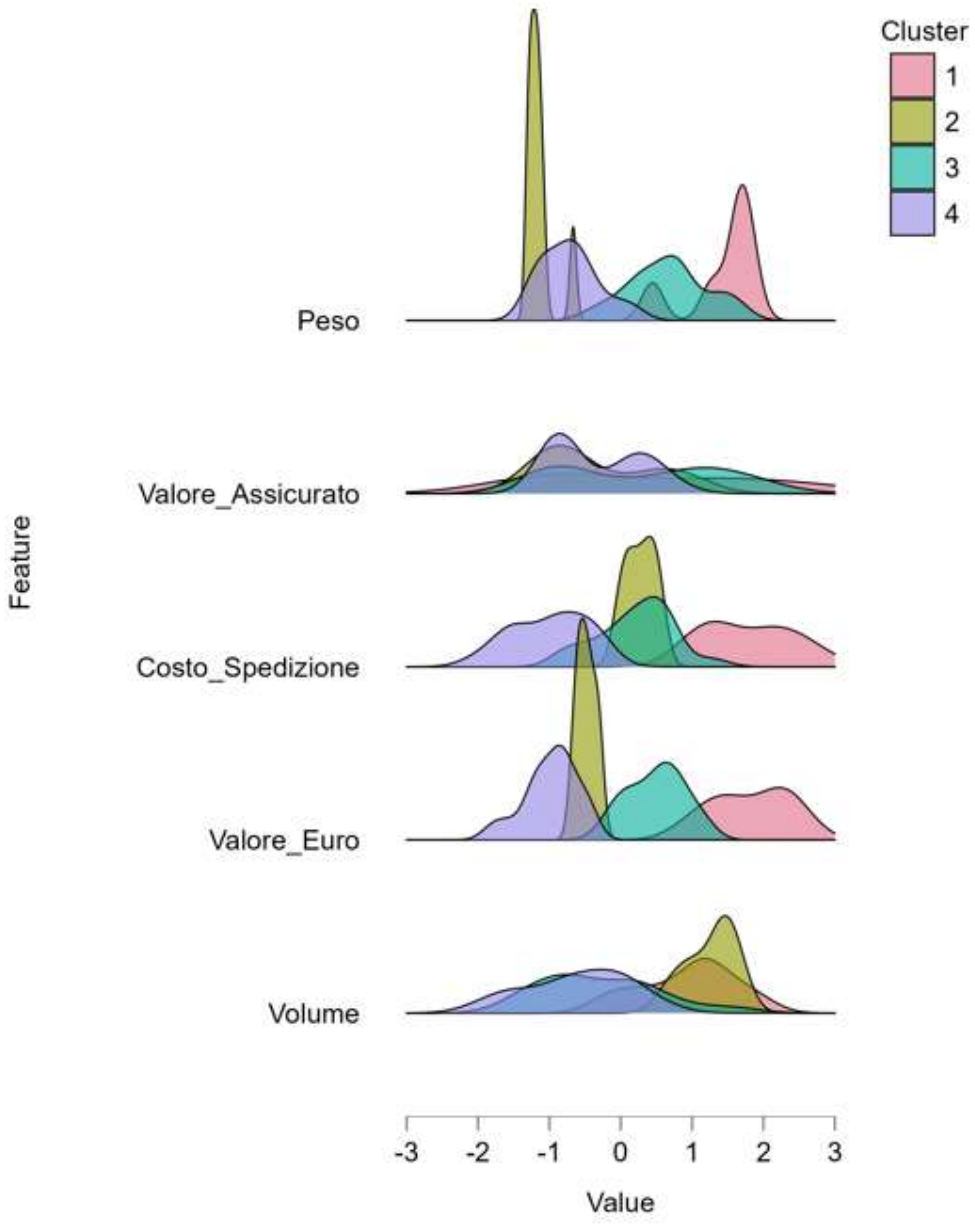
Figura 21. Dati relativi al model based clustering.



Il grafico mostra la distribuzione delle variabili analizzate all'interno dei quattro cluster identificati attraverso il model-based clustering. Ogni curva rappresenta la distribuzione di una specifica variabile per ciascun cluster. Le distribuzioni evidenziano le differenze tra i cluster in termini di peso, valore assicurato, costo di spedizione, valore economico e volume. Per la variabile peso, i cluster presentano distribuzioni significativamente diverse. Il Cluster 2 mostra un picco molto pronunciato vicino al valore zero, suggerendo una forte concentrazione di osservazioni attorno a quel valore. Al contrario, i Cluster 1 e 4 hanno distribuzioni più sparse, con il Cluster 1 che tende verso valori positivi e il Cluster 4 che si concentra su valori negativi. Il valore assicurato appare meno differenziato tra i cluster, con sovrapposizioni significative nelle distribuzioni. Tuttavia, il Cluster 1 tende verso valori leggermente positivi, mentre il Cluster 3 ha una distribuzione che si sposta verso valori negativi. Per quanto riguarda il costo di spedizione, il Cluster 2 mostra nuovamente una distribuzione molto stretta attorno al valore zero, mentre i Cluster 1 e 4 si distinguono con distribuzioni più ampie. In particolare, il Cluster 1 ha valori positivi predominanti, mentre il Cluster 4 è caratterizzato da valori più negativi. Il valore economico evidenzia una separazione più marcata tra i cluster. Il Cluster 1 mostra una distribuzione nettamente positiva, indicando che le osservazioni in questo gruppo hanno valori economici più alti rispetto alla media. Il Cluster 3 si concentra su valori negativi, mentre il Cluster 4 mostra una sovrapposizione con il Cluster 2, ma con una tendenza verso valori leggermente negativi. Infine, il volume mostra un pattern simile al valore economico, con il Cluster 1 che si posiziona su valori significativamente positivi. Il Cluster 3 ha una distribuzione che si concentra su valori più negativi, mentre i Cluster 2 e 4 hanno sovrapposizioni, ma con tendenze differenti. In sintesi, il model-based clustering evidenzia una buona separazione tra i cluster per alcune variabili, come il peso, il valore economico e il volume, mentre altre, come il valore assicurato, mostrano distribuzioni meno differenziate. Questo suggerisce che alcune caratteristiche sono più discriminanti di altre nella definizione dei cluster, il che potrebbe riflettere differenze strutturali tra i gruppi analizzati. La concentrazione stretta attorno a valori specifici, osservata soprattutto nel Cluster 2, indica

un'omogeneità interna elevata, mentre la variabilità maggiore in altri cluster suggerisce una maggiore eterogeneità (Figura 22) (Fraley and Raftery, 2002; Zhang et al., 2021; Bucci et al., 2022).

Figura 22. Dati relativi al model based clustering



5.5 Neighborhood-Based

Il report riguarda un'analisi di clustering condotta utilizzando il metodo K-Means con quattro cluster identificati. I dati forniscono informazioni su come i punti sono stati distribuiti nei cluster, insieme a metriche che valutano le performance del modello e le caratteristiche di ciascun gruppo. La dimensione dei cluster varia notevolmente, con il Cluster 1 che contiene solo quattro elementi e il Cluster 4 che ne include 19. Questa distribuzione disomogenea potrebbe indicare la presenza di gruppi

con caratteristiche molto specifiche (Cluster 1) rispetto ad altri più eterogenei (Cluster 4). La proporzione di eterogeneità spiegata all'interno dei cluster è bassa per i Cluster 1 e 2, ma più alta per i Cluster 3 e 4, suggerendo che i cluster più grandi catturano una maggiore variabilità dei dati. I punteggi di silhouette variano da 0.206 per il Cluster 1 a valori negativi (-0.329) per il Cluster 4. I valori positivi indicano che i punti sono ben raggruppati e distinti dagli altri cluster, mentre i valori negativi indicano una sovrapposizione tra cluster. Il punteggio silhouette negativo del Cluster 4 suggerisce che questo gruppo potrebbe essere meno ben definito rispetto agli altri. I centri dei cluster evidenziano differenze significative nelle variabili. Ad esempio, il Cluster 1 si caratterizza per valori elevati di peso (1.401) e valore economico (1.978), mentre il Cluster 2 si distingue per valori bassi di peso (-0.957) e valore economico (-0.311). Il Cluster 3 ha valori medi più uniformi, mentre il Cluster 4 si caratterizza per valori bassi in quasi tutte le variabili analizzate. Le metriche di performance del modello, come il Dunn Index (0.171) e il Calinski-Harabasz Index (22.708), indicano un discreto livello di separazione tra i cluster. Tuttavia, il massimo diametro (5.277) e la minima separazione (0.902) suggeriscono che alcuni cluster potrebbero essere relativamente vicini tra loro, contribuendo alla difficoltà nella distinzione osservata in alcuni casi. Analizzando le medie dei cluster, il Cluster 1 si distingue chiaramente come il gruppo con valori sopra la media in quasi tutte le variabili, il che lo rende un gruppo ben definito con caratteristiche elevate. Al contrario, il Cluster 4 si caratterizza per valori negativi in tutte le variabili, rappresentando un gruppo opposto al Cluster 1. I Cluster 2 e 3 mostrano profili intermedi, con il Cluster 2 che presenta valori negativi per molte variabili tranne il peso volumetrico, mentre il Cluster 3 ha valori prossimi alla media ma leggermente negativi. In sintesi, l'analisi K-Means ha identificato quattro cluster con caratteristiche ben distinte. Tuttavia, la dimensione ridotta di alcuni cluster e i punteggi silhouette negativi indicano la possibilità di sovrapposizioni e suggeriscono che una revisione del numero di cluster o una diversa configurazione del modello potrebbe migliorare i risultati. L'eterogeneità nei punteggi delle variabili tra i cluster offre informazioni utili per interpretare le differenze strutturali nei dati (Tabella 12) (Gupta and Chandra, 2020; Celebi et al., 2013).

Tabella 12. K-Means Clustering.

Model Summary: K-Means Clustering

Clusters	N	R^2	AIC	BIC	Silhouette
4	50	0.597	194.250	247.790	0.310

Note. The model is optimized with respect to the *BIC* value.

Cluster Information

Cluster	1	2	3	4
Size	4	10	17	19
Explained proportion within-cluster heterogeneity	0.116	0.111	0.441	0.332
Within sum of squares	16.026	15.312	61.015	45.900
Silhouette score	0.206	0.492	0.217	0.329
Center Peso	1.401	-0.957	0.777	-0.486
Center Volume	1.327	1.263	-0.217	-0.750
Center Valore_Euro	1.978	-0.311	0.640	-0.825
Center Costo_Spedizione	2.085	0.401	0.354	-0.967
Center Valore_Assicurato	0.473	-0.215	0.365	-0.313
Center Emissioni_CO2	0.605	-0.724	0.809	-0.470

Cluster Information

Cluster	1	2	3	4
Center Peso_Volumetrico	1.438	1.267	-0.274	-0.724

Note. The Between Sum of Squares of the 4 cluster model is 204.75
Note. The Total Sum of Squares of the 4 cluster model is 343

Model Performance Metrics

	Value
Maximum diameter	5.277
Minimum separation	0.902
Pearson's γ	0.557
Dunn index	0.171
Entropy	1.258
Calinski-Harabasz index	22.708

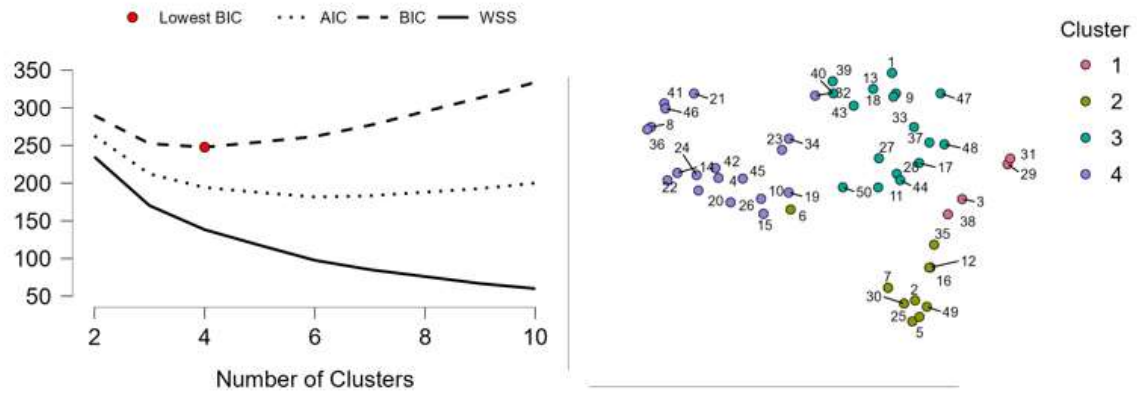
Note. All metrics are based on the *euclidean* distance.

Cluster Means

	Peso	Volume	Valore_Euro	Costo_Spedizione	Valore_Assicurato	Emissioni_CO2	Peso_Volumetrico
Cluster 1	1.401	1.327	1.978	2.085	0.473	0.605	1.438
Cluster 2	-0.957	1.263	-0.311	0.401	-0.215	-0.724	1.267
Cluster 3	0.777	-0.217	0.640	0.354	0.365	0.809	-0.274
Cluster 4	-0.486	-0.750	-0.825	-0.967	-0.313	-0.470	-0.724

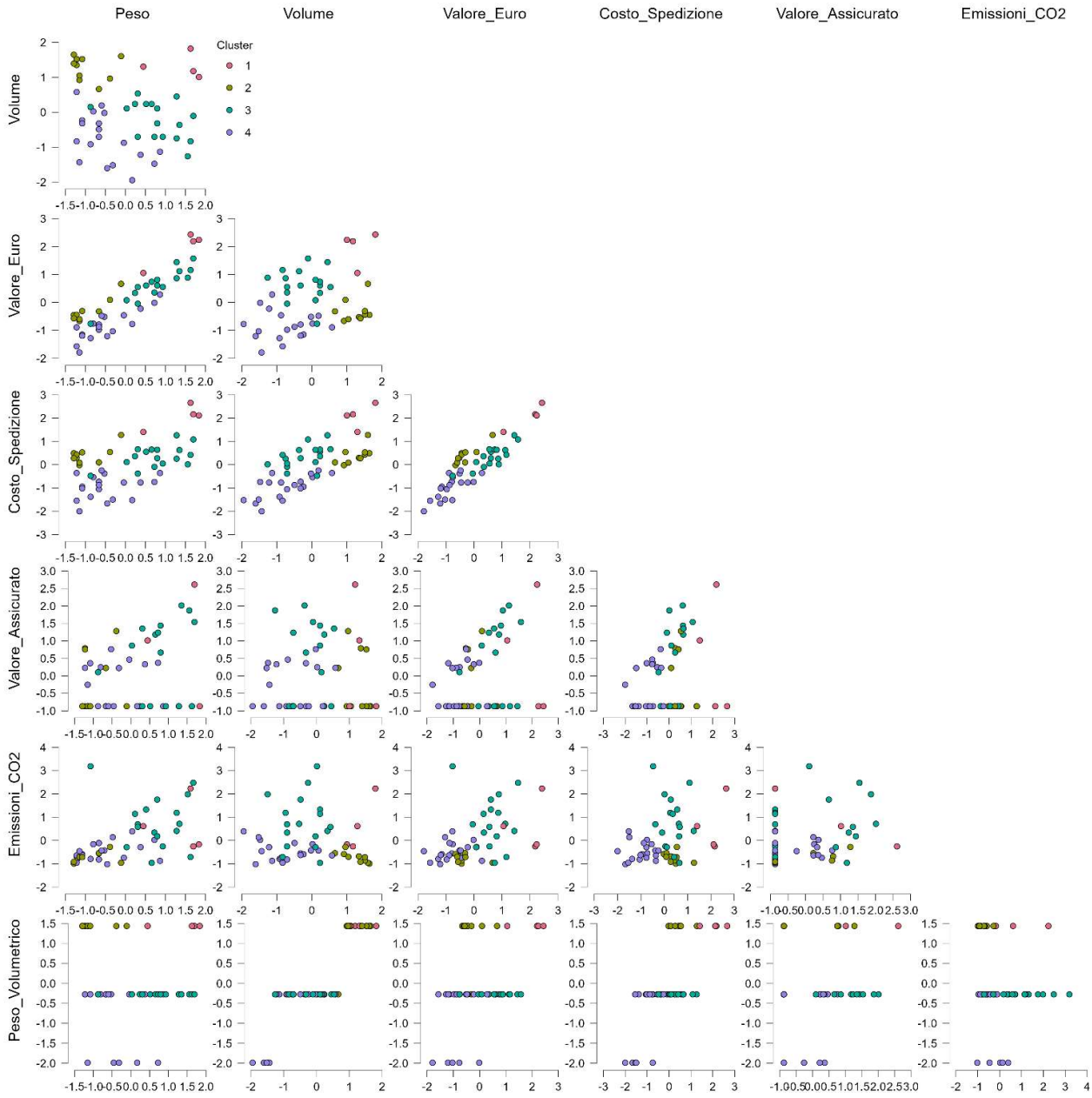
Il grafico a sinistra in Figura 22 mostra la scelta del numero ottimale di cluster per il clustering basato sul metodo Neighborhood-Based, utilizzando criteri statistici come AIC, BIC e la somma dei quadrati intra-cluster (WSS). Il punto rosso evidenziato rappresenta il valore minimo del criterio BIC, suggerendo che il modello con quattro cluster è il più appropriato. L'andamento del WSS decresce man mano che aumenta il numero di cluster, indicando una riduzione della variabilità interna ai cluster, ma con un beneficio marginale ridotto oltre i quattro cluster. Questa analisi indica che quattro cluster offrono un buon compromesso tra semplicità del modello e accuratezza. Il grafico a destra rappresenta la distribuzione dei punti dati suddivisi nei quattro cluster identificati. I punti colorati riflettono l'assegnazione a ciascun cluster, e il metodo Neighborhood-Based sembra aver creato gruppi relativamente ben definiti nello spazio multidimensionale. Tuttavia, si osserva una certa sovrapposizione tra i cluster, specialmente nelle regioni di confine. Questo potrebbe essere il risultato della vicinanza tra i dati di cluster diversi o di caratteristiche non abbastanza distintive per separare chiaramente i gruppi. In generale, il metodo Neighborhood-Based mostra un'adeguata capacità di segmentare i dati, ma l'analisi visiva evidenzia aree in cui il modello potrebbe essere migliorato attraverso un'ulteriore ottimizzazione o l'inclusione di caratteristiche aggiuntive (Khan et al., 2022; Neeraj and Maurya, 2020; Wang, 2006).

Figura 22. Statistiche per il Neighborhood-Based.



Il grafico in Figura 23 a matrice mostra le relazioni bivariate tra le variabili Peso, Volume, Valore_Euro, Costo_Spedizione, Valore_Assicurato, Emissioni_CO2 e Peso_Volumetrico, evidenziando i cluster identificati dal metodo Neighborhood-Based tramite colori distinti. Ogni scatterplot illustra come i dati si distribuiscono e come i cluster si separano lungo due dimensioni alla volta. La distribuzione dei punti suggerisce che alcune variabili, come Peso e Volume o Valore_Euro e Costo_Spedizione, presentano correlazioni evidenti, con tendenze lineari ben definite. Questo è indicativo di relazioni dirette tra tali variabili che possono aver facilitato la separazione dei cluster. I cluster appaiono generalmente distinti, ma in alcune aree dello spazio multidimensionale si osservano sovrapposizioni, in particolare tra il cluster 3 e il cluster 4. Questo potrebbe indicare una difficoltà del modello nel distinguere gruppi vicini quando i dati presentano caratteristiche simili. Alcune variabili come Peso_Volumetrico mostrano una distribuzione fortemente concentrata lungo un singolo valore, suggerendo che non contribuiscono significativamente alla separazione dei cluster. Al contrario, variabili come Peso e Volume sembrano giocare un ruolo chiave nella suddivisione, dato che i cluster appaiono più chiaramente separati in questi scatterplot. Nel complesso, il metodo Neighborhood-Based ha individuato cluster relativamente ben definiti per molte combinazioni di variabili, sebbene alcuni miglioramenti potrebbero essere raggiunti attraverso una migliore selezione delle caratteristiche o una diversa parametrizzazione del modello. L'analisi suggerisce una buona capacità di identificare gruppi coerenti, ma i confini tra alcuni cluster richiedono ulteriori verifiche per confermarne la robustezza (Ren et al., 2024; Liu et al., 2016).

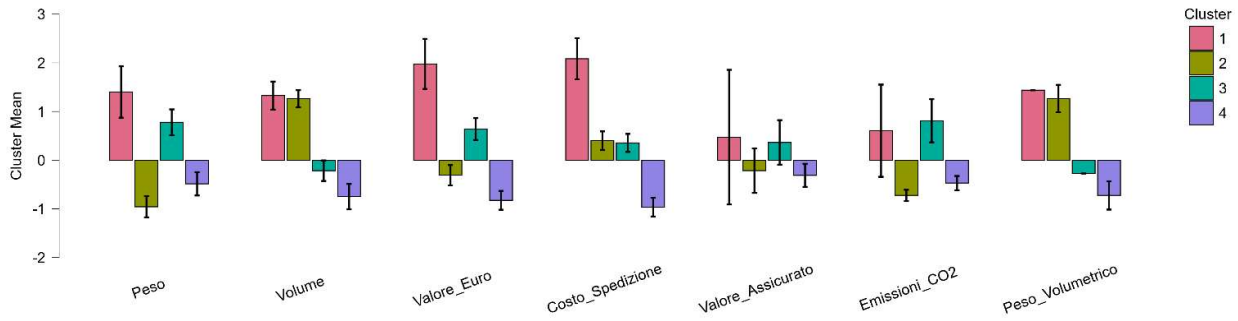
Figura 23. Statistiche per la clusterizzazione con algoritmo Neighborhood-Based.

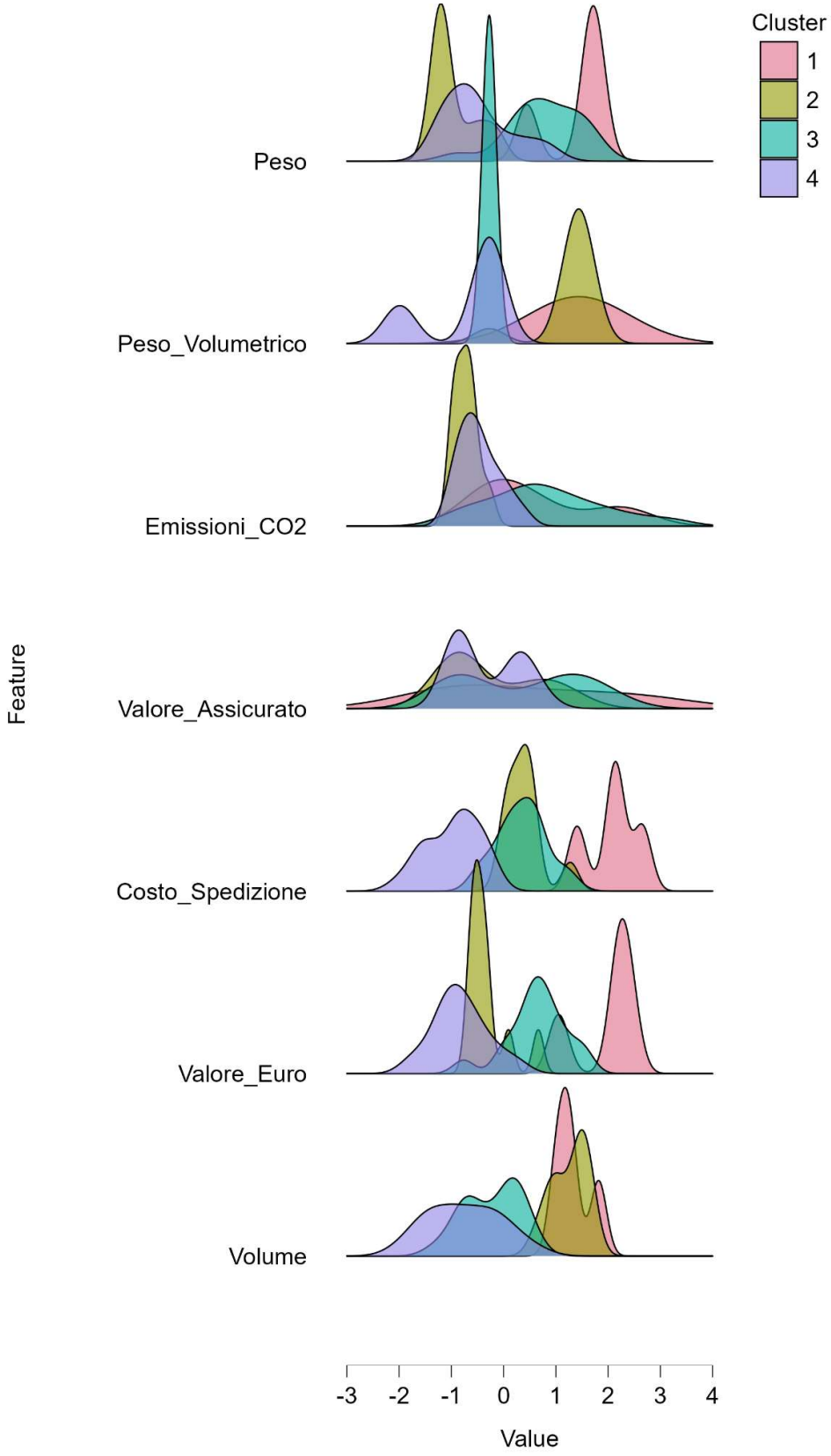


Il grafico in Figura 24 mostra i cluster ottenuti tramite un algoritmo Neighborhood-Based applicato a un insieme di dati. Le medie dei cluster sono rappresentate per ogni variabile considerata (Peso, Volume, Valore_Euro, Costo_Spedizione, Valore_Assicurato, Emissioni_CO2 e Peso_Volumetrico), con barre di errore che indicano la variazione all'interno di ciascun cluster. Ogni cluster è identificato da un colore specifico: rosa per il cluster 1, verde per il cluster 2, azzurro per il cluster 3 e viola per il cluster 4. La variabile Peso presenta una netta differenza tra i cluster, con valori medi positivi marcati per il cluster 1, negativi per il cluster 2, e più vicini a zero per i cluster 3 e 4. Questa distribuzione suggerisce che il Peso sia una variabile discriminante significativa nella separazione dei cluster. Una simile tendenza si osserva per il Volume, con il cluster 1 caratterizzato da una media significativamente maggiore rispetto agli altri, mentre il cluster 2 mostra valori medi negativi. I cluster 3 e 4 si posizionano più vicini allo zero, indicando una distribuzione equilibrata per questa

variabile. La variabile Valore_Euro mostra differenze marcate, con il cluster 1 che domina con un valore medio elevato, mentre gli altri cluster presentano valori inferiori, con il cluster 2 che si distingue per una media particolarmente negativa. Il Costo_Spedizione segue una tendenza simile, con valori più alti nel cluster 1 e valori significativamente negativi nel cluster 4. Questo pattern sottolinea l'importanza del Costo_Spedizione nella caratterizzazione dei cluster, suggerendo un'associazione tra i costi più elevati e i valori elevati delle altre variabili. Il Valore_Assicurato mostra una distribuzione meno pronunciata rispetto alle altre variabili, con differenze più contenute tra i cluster. Questo potrebbe indicare che il Valore_Assicurato abbia un ruolo meno rilevante nella separazione dei gruppi. Anche le Emissioni_CO2 mostrano una variazione limitata tra i cluster, sebbene il cluster 1 mantenga una media positiva leggermente più alta rispetto agli altri. Infine, la variabile Peso_Volumetrico si distingue per una netta differenza tra cluster 1 e cluster 4, con valori medi rispettivamente positivi e negativi, mentre i cluster 2 e 3 si trovano in una posizione intermedia. Nel complesso, il grafico evidenzia come alcune variabili, in particolare Peso, Volume e Valore_Euro, abbiano un'influenza predominante nella definizione dei cluster. I cluster sono ben distinti in termini di medie delle variabili principali, ma alcune variabili come Valore_Assicurato e Emissioni_CO2 sembrano avere un impatto meno significativo. Questo suggerisce che l'algoritmo Neighborhood-Based ha fatto affidamento su alcune variabili chiave per separare i dati in gruppi distinti (Li et al., 2014; Koumetio Tekouabou et al., 2023).

Figura 24. Statistiche per clusterizzazione con algoritmo Neighborhood-Based.




Cluster
1
2
3
4
Peso
Peso_Volumetrico
Emissioni_CO2
Valore_Assicurato
Costo_Spedizione
Valore_Euro
Volume
Feature
-3
-2
-1
0
1
2
3
4
Value

5.6 Random Forest Clustering

Il modello presentato in Figura 25 è una segmentazione basata su clustering utilizzando l'algoritmo di Random Forest. I risultati indicano che sono stati identificati quattro cluster con un totale di 50 osservazioni. L'indice R quadro pari a 0,587 evidenzia una capacità discreta del modello nel spiegare la variabilità complessiva dei dati. Tra i criteri di valutazione globale, i valori di AIC (141,290), BIC (179,540) e il punteggio medio di silhouette (0,240) suggeriscono che il modello ottimizza l'allocazione dei dati rispetto al BIC, ma il basso valore della silhouette media indica una qualità modesta della separazione tra i cluster. Analizzando la struttura dei cluster, il Cluster 1 è il più numeroso con 21 osservazioni, seguito dal Cluster 2 con 13 osservazioni, il Cluster 3 con 11 e il Cluster 4 con 5. La proporzione spiegata dell'eterogeneità interna è fortemente sbilanciata: il Cluster 1 copre il 73,2% dell'eterogeneità, mentre il Cluster 4 contribuisce solo con il 3,2%. Questo indica che il Cluster 1 ha una maggiore concentrazione di variabilità interna rispetto agli altri. I valori della somma dei quadrati interni riflettono una distribuzione simile, con 74,098 per il Cluster 1 e valori significativamente inferiori per gli altri. Per quanto riguarda il punteggio silhouette, il Cluster 4 ha il punteggio più alto (0,429), indicando una buona separazione rispetto agli altri cluster, mentre il Cluster 1 ha il punteggio più basso (0,027), suggerendo una sovrapposizione sostanziale con i cluster vicini. Le metriche di performance mostrano valori moderati in termini di separazione e compattezza. La distanza massima tra le osservazioni è 5,133, mentre la separazione minima tra cluster è 0,654. Pearson's gamma (0,420) indica una relazione moderata tra le distanze effettive e le distanze in classifica tra punti. Tuttavia, il valore di Dunn Index (0,127) evidenzia che il modello non garantisce una netta distinzione tra cluster, mentre l'entropia di 1,278 suggerisce una certa dispersione nella distribuzione delle osservazioni. L'indice di Calinski-Harabasz pari a 21,753 è relativamente alto, suggerendo che i cluster sono ragionevolmente compatti e separati. L'analisi delle medie dei cluster su cinque variabili mostra una chiara differenziazione. Il Cluster 1 presenta valori medi positivi per tutte le variabili, con un peso e un volume inferiori rispetto agli altri cluster. Questo cluster potrebbe rappresentare oggetti leggeri e compatti con valori economici e costi di spedizione medi. Il Cluster 2 mostra valori negativi per tutte le variabili, con un volume medio alto; ciò potrebbe indicare prodotti di grandi dimensioni ma a basso valore economico. Il Cluster 3 ha valori più bassi di peso e volume ma i costi e valori assicurati negativi e bassi, suggerendo una natura economica degli oggetti in questo cluster. Infine, il Cluster 4 si distingue per un peso medio moderato, volumi negativi elevati e valori negativi per costo e assicurazione, indicando potenzialmente oggetti grandi e pesanti ma con bassi costi e valori. In sintesi, il modello segmenta i dati in gruppi distinti ma con alcune sovrapposizioni tra cluster. Le caratteristiche dei cluster rivelano pattern specifici per ciascun gruppo, fornendo una base utile per approfondire l'interpretazione e l'utilizzo pratico dei risultati. Tuttavia, i punteggi di silhouette e Dunn indicano che potrebbero esserci margini di miglioramento nella definizione della struttura dei cluster (Yi et al., 2022; Marquart and Koca Marquart, 2021; Bicego and Escolano, 2021).

Figura 25. Random Forest Clustering.

Model Summary: Random Forest Clustering ▼

Clusters	N	R^2	AIC	BIC	Silhouette
4	50	0.587	141.290	179.540	0.240

Note. The model is optimized with respect to the *BIC* value.

Cluster Information

Cluster	1	2	3	4
Size	21	13	11	5
Explained proportion within-cluster heterogeneity	0.732	0.136	0.100	0.032
Within sum of squares	74.098	13.766	10.167	3.263
Silhouette score	0.027	0.413	0.340	0.429

Note. The Between Sum of Squares of the 4 cluster model is 143.71
Note. The Total Sum of Squares of the 4 cluster model is 245

Model Performance Metrics

	Value
Maximum diameter	5.133
Minimum separation	0.654
Pearson's γ	0.420
Dunn index	0.127
Entropy	1.278
Calinski-Harabasz index	21.753

Note. All metrics are based on the *euclidean* distance.

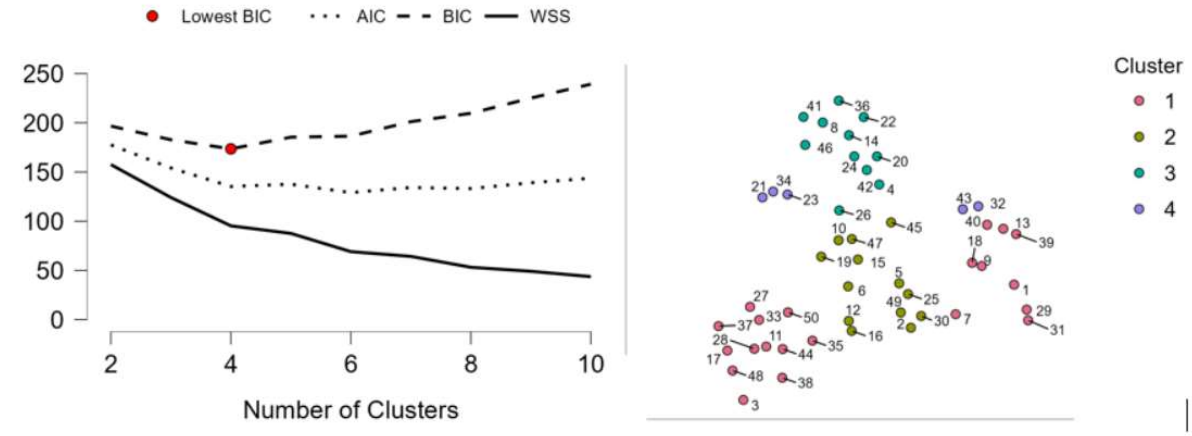
Cluster Means

	Peso	Volume	Valore_Euro	Costo_Spedizione	Valore_Assicurato
Cluster 1	0.899	0.226	0.969	0.811	0.441
Cluster 2	−1.007	0.845	−0.559	0.015	−0.161
Cluster 3	−0.729	−0.938	−1.148	−1.293	−0.601
Cluster 4	0.446	−1.081	−0.092	−0.602	−0.113

L'immagine in Figura 26 mostra un'analisi dei risultati di una Random Forest Clustering che combina l'identificazione del numero ottimale di cluster e la loro rappresentazione spaziale. A sinistra, il grafico riporta i valori di BIC, AIC e somma dei quadrati interni (WSS) in funzione del numero di cluster. A destra, una rappresentazione grafica dei punti suddivisi nei quattro cluster. Nel grafico a sinistra, si osserva che il BIC (linea tratteggiata più spessa) ha il valore minimo con quattro cluster, indicato dal punto rosso. Questo suggerisce che il modello raggiunge un equilibrio ottimale tra complessità (numero di cluster) e capacità di spiegare la variabilità dei dati. Anche l'AIC (linea tratteggiata sottile) segue una tendenza simile, sebbene decresca più lentamente. La somma dei quadrati interni (WSS, linea continua) diminuisce con l'aumentare del numero di cluster, riflettendo una riduzione della varianza interna, ma con guadagni decrescenti dopo quattro cluster. Questa analisi

conferma la scelta di quattro cluster come il compromesso ottimale. Nel grafico a destra, i punti sono distribuiti in uno spazio bidimensionale, colorati in base all'appartenenza ai cluster. I quattro cluster sono distintamente separati, ma si nota una certa sovrapposizione nei confini tra i cluster 2 e 3, oltre a un'aggregazione più compatta del cluster 4. Il cluster 1 appare sparso e meno denso, coerentemente con il basso punteggio silhouette riscontrato in precedenza. I cluster 2 e 3 mostrano una buona separazione interna, con un maggior livello di omogeneità rispetto al cluster 1. Il cluster 4, invece, è molto ben definito, come indicato dal punteggio silhouette più alto. In sintesi, la scelta di quattro cluster è supportata dai valori ottimali di BIC e AIC, nonché dalla distribuzione visiva delle osservazioni. Tuttavia, la qualità della separazione è moderata per alcuni cluster, suggerendo che potrebbe essere utile migliorare il modello o esplorare ulteriori variabili per affinare la segmentazione. La rappresentazione spaziale fornisce un'interpretazione intuitiva dei gruppi, mostrando la distribuzione e la densità delle osservazioni all'interno di ciascun cluster (Huang and Chen, 2021; Dhanaraj et al., 2021; Fan and Lu, 2022).

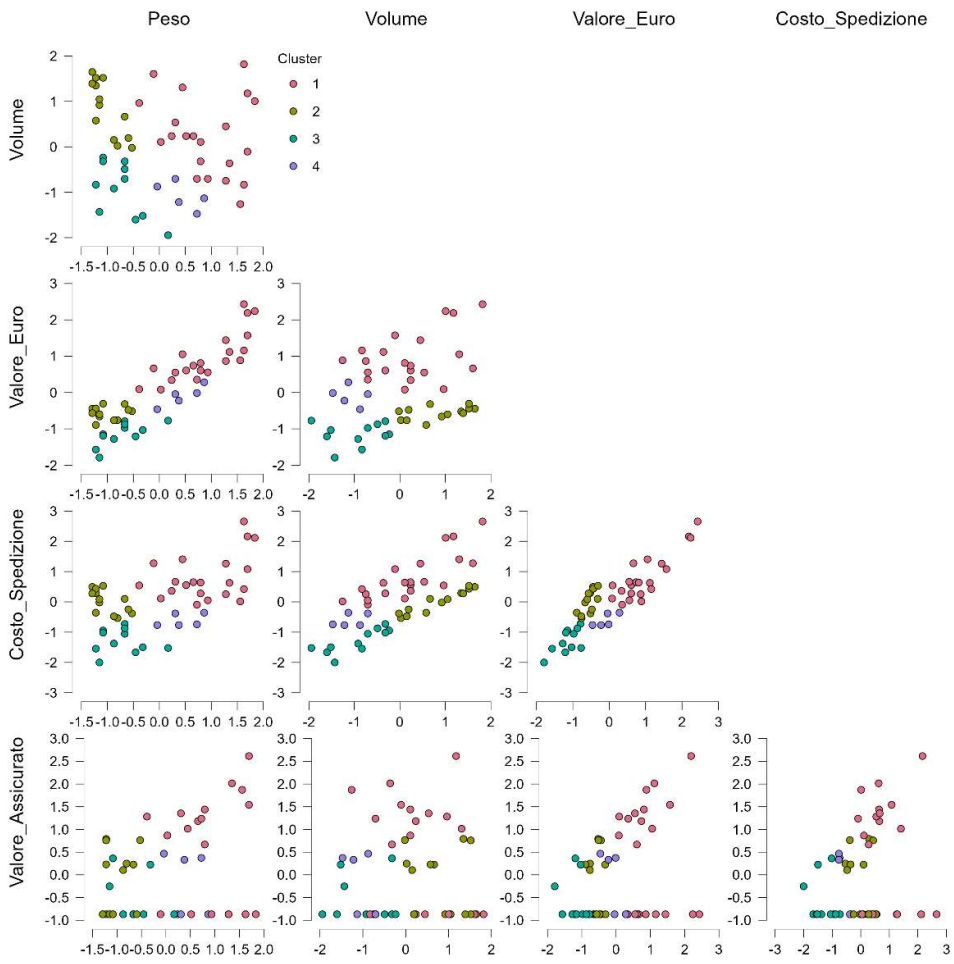
Figura 26. Random Forest Clustering.



L'immagine in Figura 27 mostra una rappresentazione grafica delle relazioni tra le variabili utilizzate per il clustering generato da un modello di Random Forest. Ogni punto rappresenta un'osservazione, e i colori indicano l'appartenenza ai quattro cluster. Le variabili in analisi includono Peso, Volume, Valore_Euro, Costo_Spedizione e Valore_Assicurato, e la matrice di grafici evidenzia le relazioni bivariate tra queste. Dalle distribuzioni si nota una chiara separazione dei cluster lungo alcune dimensioni, con schemi distintivi visibili in relazione a specifiche variabili. Per esempio, il Cluster 1, rappresentato dal colore rosso, tende ad avere valori medi o elevati per Peso, Volume e Valore_Euro, riflettendo probabilmente oggetti di valore con un peso consistente. Inoltre, per il Cluster 1, si osserva una correlazione positiva tra Peso e Valore_Euro, suggerendo che oggetti più pesanti siano generalmente associati a un valore maggiore. Questo cluster si distingue anche per Costo_Spedizione, con valori tendenzialmente più alti rispetto agli altri gruppi. Il Cluster 2, indicato in giallo, si caratterizza per valori bassi in quasi tutte le dimensioni. Peso e Volume mostrano una correlazione negativa con Valore_Euro, indicando che in questo cluster gli oggetti più leggeri e compatti tendono ad avere un valore più basso. Anche il Costo_Spedizione è generalmente più contenuto per questo gruppo, riflettendo probabilmente oggetti economici e poco voluminosi. Il Cluster 3, in verde, presenta una maggiore variabilità interna. Anche se Peso e Volume mostrano una certa correlazione positiva con Valore_Euro, i valori medi sono inferiori rispetto al Cluster 1 ma superiori al Cluster 2.

Ciò suggerisce che questo gruppo rappresenta una categoria intermedia, con oggetti che non eccellono in alcuna dimensione ma mostrano una distribuzione più ampia tra le variabili. Il Cluster 4, indicato in blu, è il più piccolo e mostra una concentrazione distinta di osservazioni caratterizzate da valori estremamente bassi di Valore_Euro e Costo_Spedizione. Questo cluster sembra rappresentare oggetti di basso valore economico e peso ridotto, con poca variabilità interna. La correlazione tra Peso, Volume e Valore_Euro è minima, suggerendo che questi oggetti sono abbastanza omogenei. In generale, la matrice di grafici mostra anche alcune correlazioni trasversali che caratterizzano l'intero dataset. Ad esempio, Peso e Volume sono strettamente correlati in tutti i cluster, indicando che questi due parametri tendono a variare insieme. Allo stesso modo, Valore_Euro e Costo_Spedizione mostrano una correlazione positiva in più cluster, il che è intuitivo dato che oggetti di valore più alto spesso richiedono costi di spedizione maggiori. L'eterogeneità tra i cluster è ben rappresentata in questo grafico, ma alcune sovrapposizioni tra gruppi sono evidenti, soprattutto tra i Cluster 2 e 3. Questo è coerente con i punteggi silhouette relativamente bassi osservati nel modello, suggerendo che i confini tra questi due cluster non sono nettamente definiti. Tuttavia, i Cluster 1 e 4 sono distintamente separati rispetto alle altre categorie, mostrando una maggiore coesione interna. Questa visualizzazione evidenzia che il clustering basato su Random Forest è riuscito a identificare gruppi con caratteristiche specifiche, ma la qualità della separazione varia a seconda delle variabili e dei cluster considerati. Ulteriori analisi potrebbero essere utili per migliorare la definizione dei confini tra cluster, soprattutto considerando nuove variabili o metodi di pre-elaborazione dei dati. La comprensione di queste differenze è essenziale per applicazioni pratiche come la segmentazione di prodotti o la classificazione di oggetti basata su caratteristiche fisiche ed economiche (Rothacher and Strobl, 2024; Marquart and Koca Marquart, 2021; Liu et al., 2021).

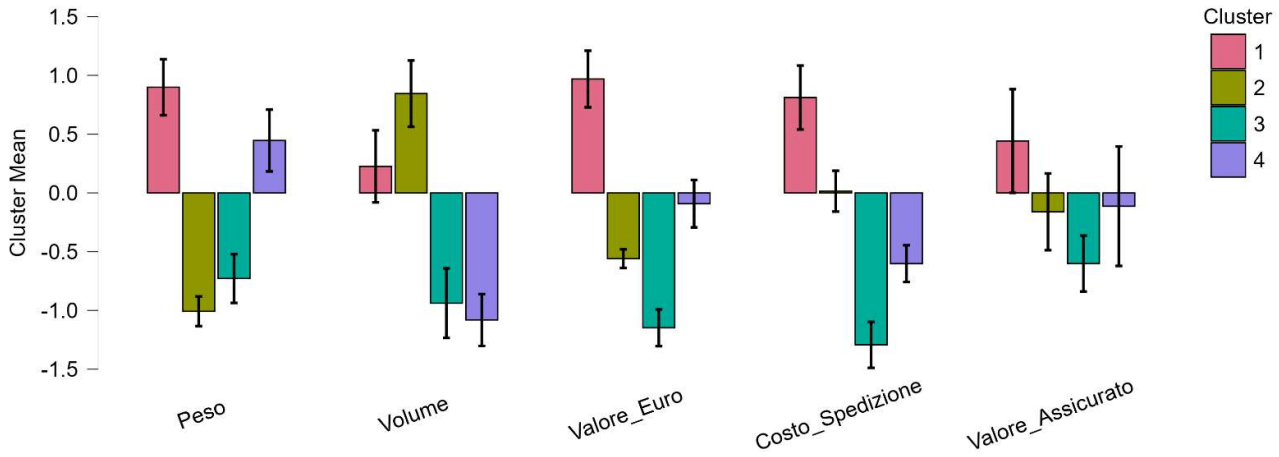
Figura 27. Random Forest Clustering.



Il grafico in Figura 28 mostra le medie delle variabili analizzate nei quattro cluster identificati dal modello di Random Forest Clustering, accompagnate dagli intervalli di errore. Le variabili considerate includono Peso, Volume, Valore_Euro, Costo_Spedizione e Valore_Assicurato. Questa rappresentazione offre una sintesi delle caratteristiche principali di ciascun cluster, evidenziandone le differenze. Il Cluster 1 (rosso) si distingue per medie elevate in quasi tutte le variabili, in particolare per Peso, Volume e Valore_Euro, dove i valori sono significativamente superiori rispetto agli altri cluster. Questo indica che il Cluster 1 rappresenta oggetti pesanti e voluminosi con un alto valore economico. Anche il Costo_Spedizione e il Valore_Assicurato mostrano valori positivi, coerenti con l'idea che oggetti di valore più elevato comportino costi di spedizione e assicurazione maggiori. Gli intervalli di errore per questo cluster sono relativamente stretti, suggerendo una buona coerenza interna. Il Cluster 2 (giallo) è caratterizzato da medie negative per tutte le variabili, anche se il Volume mostra un valore vicino allo zero, indicando oggetti più compatti ma comunque omogenei in termini di peso e valore economico basso. Questo cluster probabilmente rappresenta oggetti meno significativi in termini di valore e peso. Gli intervalli di errore sono più ampi rispetto al Cluster 1, suggerendo una maggiore eterogeneità all'interno di questo gruppo. Il Cluster 3 (verde) presenta medie moderate ma negative per Peso, Volume e Valore_Euro, con un notevole calo nel Costo_Spedizione e nel Valore_Assicurato. Questo cluster potrebbe rappresentare oggetti di dimensioni e valori mediamente più bassi rispetto agli altri cluster. Gli intervalli di errore sono relativamente stretti, indicando una certa consistenza nei dati. Il Cluster 4 (viola) si distingue per medie molto basse in Valore_Euro, Costo_Spedizione e Valore_Assicurato, pur mantenendo un Peso e un Volume negativi ma più vicini allo zero rispetto al Cluster 2. Questo suggerisce che il Cluster 4 è costituito da oggetti di basso valore economico, con spese di spedizione e assicurazione ridotte al

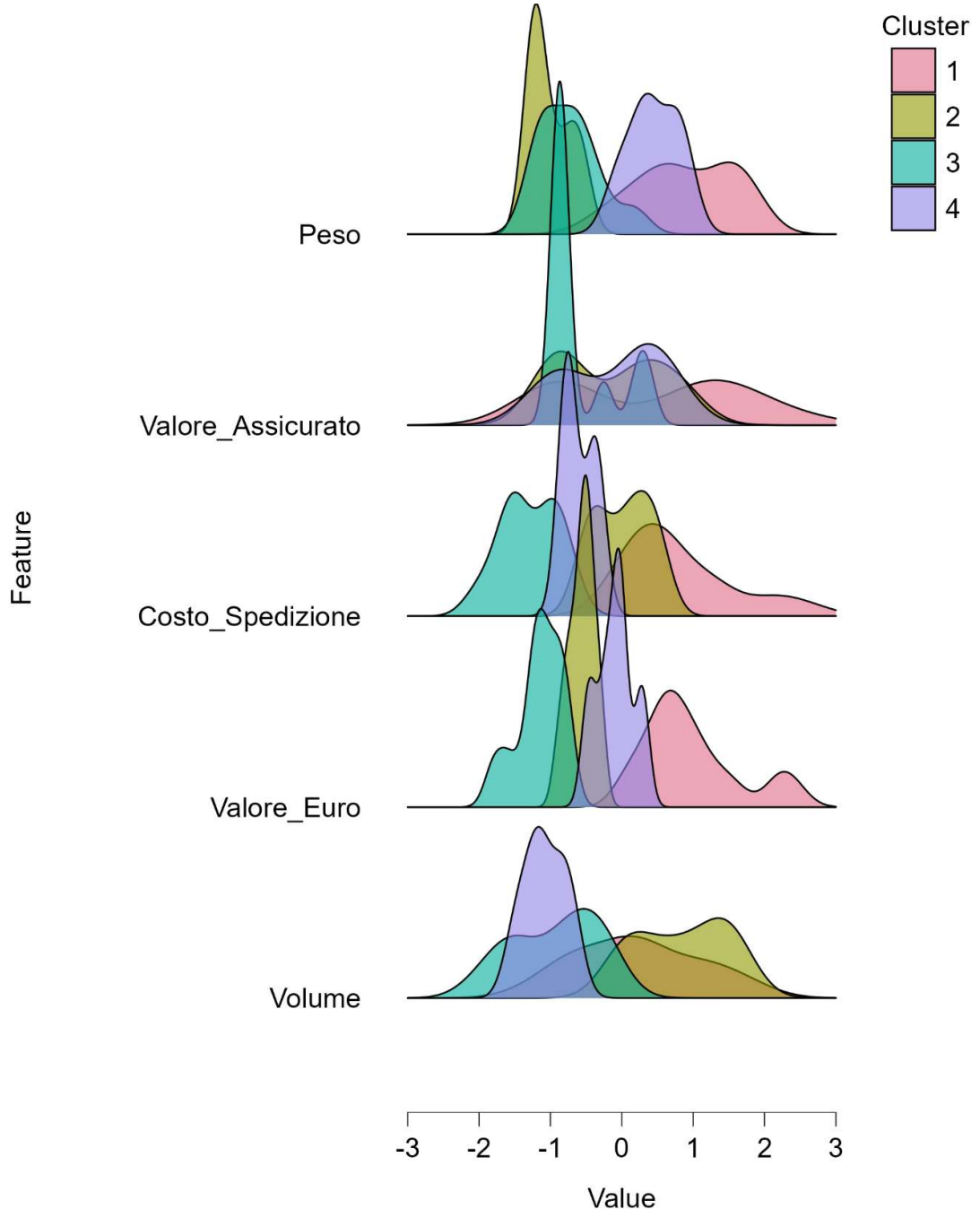
minimo. Gli intervalli di errore sono stretti, suggerendo una coerenza interna marcata. Dall'analisi emerge una chiara distinzione tra i cluster, con il Cluster 1 che rappresenta gli oggetti di maggior valore e dimensioni, mentre il Cluster 4 si posiziona agli estremi opposti come gruppo di oggetti economici e compatti. Il Cluster 2 rappresenta oggetti di basso valore e dimensioni minime, mentre il Cluster 3 mostra una natura più intermedia. L'eterogeneità negli intervalli di errore evidenzia che alcuni cluster (come il 2) sono meno coesi di altri, suggerendo la possibilità di ulteriori approfondimenti per raffinare la segmentazione. Questi risultati forniscono un quadro utile per applicazioni pratiche come la gestione di inventari, la definizione di costi di spedizione o la valutazione del rischio assicurativo (Marquart and Koca Marquart, 2021; Xu et al., 2022; Georgakis, et al., 2023).

Figura 28. Random Forest Clustering.



L'immagine in Figura 29 mostra le distribuzioni delle variabili chiave utilizzate per il clustering basato su Random Forest. Ogni grafico rappresenta una variabile (Peso, Volume, Valore_Euro, Costo_Spedizione e Valore_Assicurato), con la suddivisione delle densità stimata per ciascun cluster. I cluster sono distinti da colori diversi e riflettono le differenze nelle caratteristiche delle osservazioni.La variabile Peso mostra una chiara distinzione tra cluster. Il Cluster 1 (rosso) ha valori mediamente più elevati, con una distribuzione ampia.

Figura 29. Random Forest Clustering.



6. Soluzioni manageriali

L'implementazione dei servizi di kitting e assembly rappresenta una strategia cruciale per ottimizzare le operazioni logistiche, migliorare l'efficienza e rispondere alle crescenti esigenze del mercato. Questi processi, che includono il raggruppamento di componenti e la loro combinazione in prodotti finiti o semi-finiti, consentono di ridurre i tempi operativi, migliorare la gestione degli spazi e minimizzare gli errori umani. Attraverso l'adozione del kitting e dell'assembly, i magazzini possono beneficiare di una gestione più fluida degli ordini, una tracciabilità avanzata e una maggiore puntualità nelle consegne. Per massimizzare i benefici di queste strategie, è essenziale integrare tecnologie avanzate come i sistemi di gestione logistico (WMS). Questi strumenti permettono di tracciare in tempo reale la posizione dei materiali, ottimizzando il flusso delle attività e garantendo una maggiore precisione operativa. L'utilizzo di tecnologie complementari come codici a barre, RFID e IoT può migliorare ulteriormente la visibilità e l'efficienza complessiva delle operazioni di magazzino. La tecnologia, però, deve essere accompagnata da una formazione adeguata del personale. È necessario investire in programmi di aggiornamento continuo per garantire che gli operatori siano in grado di utilizzare efficacemente le nuove tecnologie e di seguire le migliori pratiche operative. La standardizzazione dei processi operativi è un'altra chiave per ridurre le variabili e aumentare l'efficienza, supportata dall'introduzione di automazioni come stazioni di kitting robotizzate, che migliorano la precisione e riducono i tempi di esecuzione. Un approccio collaborativo è fondamentale per il successo delle operazioni. La suddivisione del personale in squadre specializzate, ad esempio team per il prelievo, l'assemblaggio e la spedizione, garantisce che ogni fase del processo venga eseguita con la massima efficienza. L'adozione di sistemi centralizzati di gestione migliora il coordinamento tra i team, riducendo i tempi morti e gli errori. L'analisi e il monitoraggio dei dati operativi sono strumenti essenziali per identificare colli di bottiglia e aree di miglioramento. Metriche come il tempo medio per ordine, la percentuale di errori e i costi operativi devono essere regolarmente monitorate per supportare decisioni informate e ottimizzare le operazioni. Inoltre, un magazzino deve essere progettato per essere flessibile e scalabile, in modo da adattarsi rapidamente ai cambiamenti nella domanda o nei volumi operativi. Un layout modulare e soluzioni software personalizzabili possono fornire l'agilità necessaria per affrontare queste sfide. Un altro aspetto rilevante è l'adozione di strategie sostenibili. Ridurre gli sprechi di materiali, utilizzare imballaggi ottimizzati e consolidare le spedizioni sono pratiche che non solo diminuiscono i costi ma migliorano anche l'impatto ambientale del magazzino, diventando un importante vantaggio competitivo. La collaborazione con fornitori e clienti è altrettanto importante: prevedere le necessità e ottimizzare le consegne contribuisce a ridurre i tempi di installazione o utilizzo per i clienti, migliorando la loro esperienza e fidelizzazione. Infine, l'innovazione deve essere al centro della gestione del magazzino. Integrare tecnologie emergenti come l'intelligenza artificiale e il machine learning può aiutare a prevedere la domanda, ottimizzare i flussi operativi e mantenere il magazzino competitivo e all'avanguardia. Investire nella ricerca e sviluppo è essenziale per sostenere un vantaggio competitivo a lungo termine, garantendo che il magazzino rimanga un asset strategico per l'azienda. Questi approcci, combinati, permettono di creare un sistema logistico più efficiente, flessibile e sostenibile, capace di rispondere rapidamente alle esigenze di un mercato in continua evoluzione (Zhou and He, 2020; Zhao et al., 2021; El Moussaoui et al., 2021).

7. Conclusioni

Le conclusioni del documento evidenziano l'importanza strategica dell'implementazione dei servizi di kitting e assembly per affrontare le sfide logistiche e operative nei magazzini, in particolare nel settore idrotermosanitario. Attraverso l'analisi condotta, è emerso come questi strumenti possano migliorare significativamente l'efficienza, ridurre gli errori operativi e ottimizzare l'uso delle risorse. Il kitting consente di raggruppare componenti specifici in modo strutturato, semplificando i processi successivi e riducendo il rischio di errori umani. L'assembly completa il ciclo operativo, creando kit semi-finiti o finiti pronti per la spedizione o l'installazione, aumentando così la reattività del magazzino alle richieste dei clienti. Un aspetto cruciale emerso è l'importanza dell'integrazione

tecnologica, come l'adozione di software avanzati di gestione logistica e strumenti di tracciamento in tempo reale. Questi sistemi non solo migliorano la visibilità delle operazioni, ma facilitano anche il coordinamento tra diverse squadre operative, riducendo i tempi di inattività e ottimizzando l'allocazione delle risorse. La standardizzazione dei processi e l'automazione rappresentano ulteriori pilastri per aumentare la produttività e garantire una gestione ottimale degli spazi. Le conclusioni del documento sottolineano anche l'importanza della collaborazione tra risorse umane e tecnologie, evidenziando come la formazione del personale sia essenziale per massimizzare i benefici delle innovazioni implementate. Inoltre, l'approccio analitico e basato sui dati, combinato con l'adozione di pratiche sostenibili, contribuisce non solo alla riduzione dei costi operativi ma anche a un impatto ambientale positivo. In sintesi, l'implementazione dei servizi di kitting e assembly, supportata da un uso intelligente della tecnologia e da una gestione strategica delle risorse, rappresenta un'opportunità concreta per migliorare la competitività e soddisfare le crescenti esigenze dei clienti. Questi strumenti permettono di trasformare il magazzino da un semplice centro operativo a un elemento strategico nella catena del valore aziendale, garantendo maggiore flessibilità, efficienza e sostenibilità.

8. References

Abdalzaher, M. S., Krichen, M., Yiltas-Kaplan, D., Ben Dhaou, I., & Adoni, W. Y. H. (2023). Early detection of earthquakes using iot and cloud infrastructure: A survey. Sustainability, 15(15), 11713.

Agarwal, A., Kenney, A. M., Tan, Y. S., Tang, T. M., & Yu, B. (2023). MDI+: A flexible random forest-based feature importance framework. arXiv preprint arXiv:2307.01932.

Ahmed, S., Parvathaneni, D., & Shareef, I. (2023). Reorganization of inventory to improve kitting efficiency and maximize space utilization. Manufacturing Letters, 35, 1366-1377.

Aldrich, C. (2020). Process variable importance analysis by use of random forests in a shapley regression framework. Minerals, 10(5), 420.

Ali, S. S., Kaur, R., Gupta, H., Ahmad, Z., & Jebahi, K. (2024). A decision-making framework for determinants of an organisation's readiness for smart warehouse. Production Planning & Control, 35(14), 1887-1908.

Almasov, A., & Onur, M. (2023, June). Life-Cycle Production Optimization with Nonlinear Constraints Using a Least-Squares Support-Vector Regression Proxy. In SPE EuropEC-Europe Energy Conference featured at the 84th EAGE Annual Conference & Exhibition. OnePetro.

Altieri, F., Pietracaprina, A., Pucci, G., & Vandin, F. (2021). Scalable distributed approximation of internal measures for clustering evaluation. In Proceedings of the 2021 SIAM International Conference on Data Mining (SDM) (pp. 648-656). Society for Industrial and Applied Mathematics.

Amaldi, E., Consolo, A., & Manno, A. (2023). On multivariate randomized classification trees: l0-based sparsity, vc dimension and decomposition methods. Computers & Operations Research, 151, 106058.

Argente, J. N., Fortus, C. M., & Lacsamana, P. M. (2023). FAL Services: Improving Productivity and Efficiency through Maximizing Facilities using Third Party Logistics. Ani: Letran Calamba Research Report, 19(1), 1-1.

Asamov, T., & Ben-Israel, A. (2022). A probabilistic ℓ1 method for clustering high-dimensional data. Probability in the Engineering and Informational Sciences, 36(2), 433-448.

Baglio, M., Creazza, A., & Dallari, F. (2024). 'Logistics 4.0'technologies in the 3PL industry: a maturity model. Production Planning & Control, 1-17.

Bedoui, A., & Lazar, N. A. (2020). Bayesian empirical likelihood for ridge and lasso regressions. Computational Statistics & Data Analysis, 145, 106917.

Berroir, F., Guernaccini, P., Boje, C., & Maatar, O. (2021, July). Reducing construction logistics costs and embodied carbon with CCC and kitting: a case study. In Proc. 29th Annual Conference of the International Group for Lean Construction (IGLC) (pp. 935-944).

Bicego, M., & Escolano, F. (2021, January). On learning random forests for random forest-clustering. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 3451-3458). IEEE.

Bogue, R. (2022). Warehouse robot market boosted by Covid pandemic and technological innovations. Industrial Robot: the international journal of robotics research and application, 49(2), 181-186.7

Bogue, R. (2024). The role of robots in logistics. Industrial Robot: the international journal of robotics research and application, 51(3), 381-386.

Bortolini, M., Faccio, M., Gamberi, M., & Pilati, F. (2020). Assembly kits with variable part physical attributes: warehouse layout design and assignment procedure. Assembly automation, 40(6), 857-868.

Bottin, M., Faccio, M., Minto, R., & Rosati, G. (2021). Sales kit automated production: An integrated procedure for setup reduction in case of high products variety. Applied Sciences, 11(21), 10110.

Bracher, J., Ray, E. L., Gneiting, T., & Reich, N. G. (2021). Evaluating epidemic forecasts in an interval format. PLoS computational biology, 17(2), e1008618.

Bucci, A., Palomba, G., & Rossi, E. (2022). starvars: An R Package for Analysing Nonlinearities in Multivariate Time Series. THE R JOURNAL, 14, 208-226.

Bueno Viso, M. (2022). Automated AGS Kitting Station.

Bushra, A. A., & Yi, G. (2021). Comparative analysis review of pioneering DBSCAN and successive density-based clustering algorithms. IEEE Access, 9, 87918-87935.

Buzu, A. (2021). The effect of Warehousing management on Warehouse performance. Available at SSRN 3951785.

Caputo, A. C., Pelagagge, P. M., & Salini, P. (2021). A model for planning and economic comparison of manual and automated kitting systems. International Journal of Production Research, 59(3), 885-908.

Celebi, M. E., Kingravi, H. A., & Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert systems with applications, 40(1), 200-210.

Chakraborty, M., Shakir Mahmud, M., Gates, T. J., & Sinha, S. (2023). Analysis and prediction of human mobility in the United States during the early stages of the COVID-19 pandemic using regularized linear models. Transportation research record, 2677(4), 380-395.

Chang, S. C., Chuang, W. C., & Jeng, J. T. (2023). New Interval Improved Fuzzy Partitions Fuzzy C-Means Clustering Algorithms under Different Distance Measures for Symbolic Interval Data Analysis. Applied Sciences, 13(22), 12531.

Cheng, D., Xu, R., Zhang, B., & Jin, R. (2023). Fast density estimation for density-based clustering methods. Neurocomputing, 532, 170-182.

Choi, Y. G., Ahn, S., & Kim, J. (2023). Model-based clustering of mixed data with sparse dependence. IEEE Access.

Coe, N. M. (2021). Coping with commoditization: The third-party logistics industry in the Asia-Pacific. Competition & Change, 25(3-4), 281-307.

Côme, E., Jouvin, N., Latouche, P., & Bouveyron, C. (2021). Hierarchical clustering with discrete latent variable models and the integrated classification likelihood. Advances in Data Analysis and Classification, 15(4), 957-986.

Costa, M., Pinto, V. H., & Gonçalves, G. (2024, June). Wireless Localization System Applied to a Kitting Pick-to-light System. In International Conference Innovation in Engineering (pp. 252-263). Cham: Springer Nature Switzerland.

Czermański, E., Cirella, G. T., Oniszczuk-Jastrząbek, A., Pawłowska, B., & Notteboom, T. (2021). An energy consumption approach to estimate air emission reductions in container shipping. Energies, 14(2), 278.

Dahl, D. B., Andros, J., & Carter, J. B. (2023). Cluster analysis via random partition distributions. Statistical Analysis and Data Mining: The ASA Data Science Journal, 16(2), 135-148.

Dakhli, Z., & Lafaj, Z. (2022). A Lean-Based View. Lean Construction 4.0: Driving a Digital Revolution of Production Management in the AEC Industry, 82.

Dakhli, Z., & Lafhaj, Z. (2022). The Interplay between Construction Supply Chain and BIM through Kitting: A Lean-Based View. In Lean Construction 4.0 (pp. 82-97). Routledge.

de Amorim, R. C., & Makarenkov, V. (2021). Improving cluster recovery with feature rescaling factors. Applied Intelligence, 51, 5759-5774.

Deng, D. (2020, September). DBSCAN clustering algorithm based on density. In 2020 7th international forum on electrical engineering and automation (IFEEA) (pp. 949-953). IEEE.

Desu, S. S. T., Srijith, P. K., Rao, M. P., & Sivadasan, N. (2021, June). Adiabatic quantum feature selection for sparse linear regression. In International Conference on Computational Science (pp. 98-112). Cham: Springer International Publishing.

Dey, S., De, S., & Paul, S. (2021, January). A new approach of data clustering using quantum inspired particle swarm optimization based fuzzy c-means. In 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 59-64). IEEE.

Dhanaraj, R. K., Ramakrishnan, V., Poongodi, M., Krishnasamy, L., Hamdi, M., Kotecha, K., & Vijayakumar, V. (2021). Random forest bagging and x-means clustered antipattern detection from SQL query log for accessing secure mobile data. Wireless communications and mobile computing, 2021(1), 2730246.

Eibeck, A., Shaocong, Z., Mei Qi, L., & Kraft, M. (2024). Research data supporting" A Simple and Efficient Approach to Unsupervised Instance Matching and its Application to Linked Data of Power Plants".

El Moussaoui, S., Lafhaj, Z., Leite, F., Fléchard, J., & Linéatte, B. (2021). Construction logistics centres proposing kitting service: Organization analysis and cost mapping. Buildings, 11(3), 105.

Elia, V., Gnoni, M. G., & Tornese, F. (2024). On-Demand Warehousing Platforms: Evolution and Trend Analysis of an Industrial Sharing Economy Model. Logistics, 8(4), 93.

Fan, Z., & Lu, L. (2022, April). Tourism Data Clustering Analysis Based on Random Forest Algorithm. In 2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC) (pp. 1211-1214). IEEE.

Fatima, U., Mohammed, D., & Shareef, I. (2024). A holistic approach to kitting cart optimization and steel receiving analysis for process improvement. Manufacturing Letters, 41, 1716-1727.

Ferrari, A., Mangano, G., Zenezini, G., & Carlin, A. (2022). A real simulation of automated warehouses processes: an academic experience with engineering students. ... SUMMER SCHOOL FRANCESCO TURCO. PROCEEDINGS.

Ferrari, A., Zenezini, G., Carlin, A., & Rafele, C. (2021). An integrated simulation modelling approach for a warehouse 4.0.

Ferreira, A. R. (2022). Estimation of knots location and number in the splines regression models using an optimization approach (Doctoral dissertation, Universidade de São Paulo).

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. Journal of the American statistical Association, 97(458), 611-631.

Gajjar, J. M., & Thakkar, H. R. (2014). Improvement in material feeding system through introducing kitting concept in lean environment of MSME: a review study. INTERNATIONAL JOURNAL OF ENGINEERING DEVELOPMENT AND RESEARCH, 2(1), 891-900.

García-Ordás, M. T., Alaiz-Moretón, H., Casteleiro-Roca, J. L., Jove, E., Benítez-Andrades, J. A., García-Rodríguez, I., ... & Calvo-Rolle, J. L. (2023). Clustering techniques selection for a hybrid regression model: a case study based on a solar thermal system. Cybernetics and Systems, 54(3), 286-305.

Georgakis, A., Gatziolis, D., & Stamatellos, G. (2023). A primer on clustering of forest management units for reliable design-based direct estimates and model-based small area estimation. Forests, 14(10), 1994.

Geraeds, E. P., & Llamoca, E. L. C. (2023). Manpower prediction for kitting in high complex, low volume assembly lines.

Gergely, B., & Vargha, A. (2021). How to use model-based cluster analysis efficiently in person-oriented research. Journal for Person-Oriented Research, 7(1), 22.

Greenwell, B. M., Boehmke, B. C., & Gray, B. (2020). Variable Importance Plots-An Introduction to the vip Package. R J., 12(1), 343.

Gren, I. M., Brutemark, A., Jägerbrand, A. K., & Svedén, J. B. (2020). Costs of air pollutants from shipping: a meta-regression analysis. Transport reviews, 40(4), 411-428.

Guo, Y., Wang, W., & Wang, X. (2021). A robust linear regression feature selection method for data sets with unknown noise. IEEE Transactions on Knowledge and Data Engineering, 35(1), 31-44.

Gupta, M. K., & Chandra, P. (2020). An empirical evaluation of K-means clustering algorithm using different distance/similarity metrics. In Proceedings of ICETIT 2019: Emerging Trends in Information Technology (pp. 884-892). Springer International Publishing.

Hu, P., Zhang, J., & Li, N. (2021, October). Research on Flight Delay Prediction Based on Random Forest. In 2021 IEEE 3rd International Conference on Civil Aviation Safety and Information Technology (ICCASIT) (pp. 506-509). IEEE.

Huang, Z., & Chen, D. (2021). A breast cancer diagnosis method based on VIM feature selection and hierarchical clustering random forest algorithm. IEEE Access, 10, 3284-3293.

Irfiyanda, C., Andreswari, R., & Hamami, F. (2022, November). Customer Segmentation Using Fuzzy C-Means Algorithm in Telco Industry. In 2022 International Conference of Science and Information Technology in Smart Administration (ICSINTESA) (pp. 1-4). IEEE.

Jiang, Z. Z., Wan, M., Pei, Z., & Qin, X. (2021). Spatial and temporal optimization for smart warehouses with fast turnover. Computers & Operations Research, 125, 105091.

Ju, Y., & Hargreaves, C. A. (2021). The impact of shipping CO2 emissions from marine traffic in Western Singapore Straits during COVID-19. Science of The Total Environment, 789, 148063.

Jum'a, L., & Basheer, M. E. (2023). Analysis of warehouse value-added services using pareto as a quality tool: A case study of third-party logistics service provider. Administrative Sciences, 13(2), 51.

Kalkha, H., Khiat, A., Bahnasse, A., & Ouajji, H. (2022). Toward a reliable and responsive E-commerce with IoT. Procedia Computer Science, 198, 614-619.

Kanberoğlu, B., & Kökkülünk, G. (2021). Assessment of CO2 emissions for a bulk carrier fleet. Journal of Cleaner Production, 283, 124590.

Karaduman, H. A., Karaman-Akgül, A., Çağlar, M., & Akbaş, H. E. (2020). The relationship between logistics performance and carbon emissions: an empirical investigation on Balkan countries. International Journal of Climate Change Strategies and Management, 12(4), 449-461.

Karim, N. H., Abdul Rahman, N. S. F., Md Hanafiah, R., Abdul Hamid, S., Ismail, A., Abd Kader, A. S., & Muda, M. S. (2021). Revising the warehouse productivity measurement indicators: ratio-based benchmark. Maritime Business Review, 6(1), 49-71.

Kawa, A. (2021). Fulfilment as logistics support for E-tailers: An empirical studies. Sustainability, 13(11), 5988.

Khan, A. A., Shameem, M., Nadeem, M., & Akbar, M. A. (2021). Agile trends in Chinese global software development industry: Fuzzy AHP based conceptual mapping. Applied Soft Computing, 102, 107090.

Khan, G. A., Hu, J., Li, T., Diallo, B., & Zhao, Y. (2022). Multi-view low rank sparse representation method for three-way clustering. International Journal of Machine Learning and Cybernetics, 13, 233-253.

Kilic, H. S., & Durmusoglu, M. B. (2012). Design of kitting system in lean-based assembly lines. Assembly automation, 32(3), 226-234.

Kinshakov, E., Parfenenko, Y., & Shendryk, V. (2021). Comparative analysis of methods for prediction continuous numerical features on big datasets. Technology audit and production reserves, 6(2), 62.

Klundt, E., Towers, N., & Bechkoum, K. (2024). Lean and Agile Supply Strategies in Distribution Centres to Deliver Value-Added Services (VAS). Logistics, 8(3), 67.

Koumetio Tekouabou, S. C., Diop, E. B., Azmi, R., & Chenal, J. (2023). Artificial intelligence based methods for smart and sustainable urban planning: a systematic survey. Archives of Computational Methods in Engineering, 30(2), 1421-1438.

Kujanpää, V. (2024). Logistics in lean construction: Exploration of processes, challenges and best practices.

Kule, C., Patil, S. B., & Vaity, S. (2020). Improvement in Material Feeding by Introducing Kitting in the Assembly Line. In Proceedings of International Conference on Intelligent Manufacturing and Automation: ICIMA 2020 (pp. 407-417). Springer Singapore.

Li, J., Anser, M. K., Tabash, M. I., Nassani, A. A., Haffar, M., & Zaman, K. (2023). Technology-and logistics-induced carbon emissions obstructing the Green supply chain management agenda: Evidence from 101 countries. International Journal of Logistics Research and Applications, 26(7), 788-812.

Li, T., Sun, S., Sattar, T. P., & Corchado, J. M. (2014). Fight sample degeneracy and impoverishment in particle filters: A review of intelligent approaches. Expert Systems with applications, 41(8), 3944-3954.

Liu, H., Deng, A., Wang, X., & Yue, G. (2022, December). Density Peaks Clustering Algorithm Based on Density Stratification and Subcluster Fusion. In Proceedings of the 2022 5th International Conference on Algorithms, Computing and Artificial Intelligence (pp. 1-7).

Liu, J., & Duru, O. (2020). Bayesian probabilistic forecasting for ship emissions. Atmospheric environment, 231, 117540.

Liu, J., Anavatti, S., Garratt, M., & Abbass, H. A. (2022). Modified continuous ant colony optimisation for multiple unmanned ground vehicle path planning. Expert Systems with Applications, 196, 116605.

Liu, S., Maljovec, D., Wang, B., Bremer, P. T., & Pascucci, V. (2016). Visualizing high-dimensional data: Advances in the past decade. IEEE transactions on visualization and computer graphics, 23(3), 1249-1268.

Liu, Z., Wen, T., Sun, W., & Zhang, Q. (2021). Feature-weighting and clustering random forest. International Journal of Computational Intelligence Systems, 14(1), 257-265.

Lorson, F., Fügener, A., & Hübner, A. (2023). New team mates in the warehouse: Human interactions with automated and robotized systems. Iise Transactions, 55(5), 536-553.

Maatar, O., Trost, R., De Bruyne, I., Van Dromme, H., & Berroir, F. (2022, September). Smart logistics for urban construction sites (CCC). In IOP Conference Series: Earth and Environmental Science (Vol. 1078, No. 1, p. 012046). IOP Publishing.

Mabeya, F. (2022). Improved Warehouse for SMT Material Management using Modern Technology Retrieval System and Better Traceability (Master's thesis, Minnesota State University, Mankato).

Maddumala, V. R. (2020). A Weight Based Feature Extraction Model on Multifaceted Multimedia Bigdata Using Convolutional Neural Network. Ingénierie des Systèmes d'Information, 25(6).

Maderna, R., Poggiali, M., Zanchettin, A. M., & Rocco, P. (2020, May). An online scheduling algorithm for human-robot collaborative kitting. In 2020 IEEE international conference on robotics and automation (ICRA) (pp. 11430-11435). IEEE.

Mandal, A., Perrot, M., & Ghoshdastidar, D. (2022). A Revenue Function for Comparison-Based Hierarchical Clustering. arXiv preprint arXiv:2211.16459.

Marquart, I., & Koca Marquart, E. (2021). RFCC: Random Forest Consensus Clustering for Regression and Classification. Available at SSRN 3807828.

Montoya Zapata, S., Klement, N., Silva, C., Gibaru, O., & Lafou, M. (2023, September). Simulation of a Kitting System for the Replenishment of an Automotive Assembly Line. In International Workshop on Service Orientation in Holonic and Multi-Agent Manufacturing (pp. 177-188). Cham: Springer Nature Switzerland.

Montoya-Zapata, S., Klement, N., Silva, C., Gibaru, O., & Lafou, M. (2024). Multi-agent system for perturbations in the kitting process of an automotive assembly line. Engineering Applications of Artificial Intelligence, 135, 108679.

Mudyazhezha, E. E. (2024). Reduction of Post-Harvest Losses in the Maize Value Chain: A Review of Warehousing Literature. The Zimbabwe Journal of Business, Economics and Management, 3(1).

Muñoz-Villamizar, A., Velázquez-Martínez, J. C., Haro, P., Ferrer, A., & Mariño, R. (2021). The environmental impact of fast shipping ecommerce in inbound logistics operations: A case study in Mexico. Journal of Cleaner Production, 283, 125400.

Mushagalusa, C. A., Fandohan, A. B., & Glèlè Kakaï, R. (2022). Random Forests in Count Data Modelling: An Analysis of the Influence of Data Features and Overdispersion on Regression Performance. Journal of Probability and Statistics, 2022(1), 2833537.

Neeraj, K. N., & Maurya, V. (2020). A review on machine learning (feature selection, classification and clustering) approaches of big data mining in different area of research. Journal of critical reviews, 7(19), 2610-2626.

Ngatchou-Wandji, J., & Bulla, J. (2011). On choosing a mixture model for clustering.

Nguyen, S. D., Nguyen, V. S. T., & Pham, N. T. (2021). Determination of the optimal number of clusters: A fuzzy-set based method. IEEE Transactions on Fuzzy Systems, 30(9), 3514-3526.

Nilsson, E., & Jayaraman, V. (2021). Selection of Warehouse Automation System (s) for component storage-in a multiple assembly line manufacturing context.

Pantula, P. D., Miriyala, S. S., & Mitra, K. (2020). An evolutionary neuro-fuzzy C-means clustering technique. Engineering Applications of Artificial Intelligence, 89, 103435.

Qian, L., Plant, C., & Böhm, C. (2021, December). Density-based clustering for adaptive density variation. In 2021 IEEE International Conference on Data Mining (ICDM) (pp. 1282-1287). IEEE.

Raghuram, P., & Arjunan, M. K. (2022). Design framework for a lean warehouse–a case study-based approach. International Journal of Productivity and Performance Management, 71(6), 2410-2431.

Raja, R., & Venkatachalam, S. (2022). Adoption of digital technology in global third-party logistics services providers: A review of literature. FOCUS: Journal of International Business, 9(1), 105-129.

Ramnath, B. V., Kumar, C. S., Mohamed, G. R., Venkataraman, K., Elanchezhian, C., & Sathish, S. (2014). Analysis of occupational safety and health of workers by implementing ergonomic based kitting assembly system. Procedia Engineering, 97, 1788-1797.

Ran, X., Xi, Y., Lu, Y., Wang, X., & Lu, Z. (2023). Comprehensive survey on hierarchical clustering algorithms and the recent developments. Artificial Intelligence Review, 56(8), 8219-8264.

Rapp, T., Peters, C., & Dachsbacher, C. (2020). Visual analysis of large multivariate scattered data using clustering and probabilistic summaries. IEEE Transactions on Visualization and Computer Graphics, 27(2), 1580-1590.

Ren, Y., Pu, J., Yang, Z., Xu, J., Li, G., Pu, X., ... & He, L. (2024). Deep clustering: A comprehensive survey. IEEE Transactions on Neural Networks and Learning Systems.

Rocamora, F. R., Camaclang, C., & Cervantes, J. (2023). A Feasibility of Outsourcing Warehouse and Kitting and Services in Automotive Companies at District 2, Laguna. Ani: Letran Calamba Research Report, 19(1), 1-1.

Rothacher, Y., & Strobl, C. (2024). Identifying informative predictor variables with random forests. Journal of Educational and Behavioral Statistics, 49(4), 595-629.

Sahoo, R., Bhowmick, B., & Tiwari, M. K. (2023). Developing a model to optimise the cost of consolidated air freight considering the varying scenarios. International Journal of Logistics Research and Applications, 26(8), 1035-1059.

Sarimi, A. F., Nasir, N., Rashid, H., & Azis, N. A. (2023). Development and simulation via flexsim for kitting trolley design of rear car seat assembly process. Journal of Applied Engineering Design & Simulation (JAEDS), 3(2), 27-37.

Schwartz, H., Gustafsson, M., & Spohr, J. (2020). Emission abatement in shipping–is it possible to reduce carbon dioxide emissions profitably?. Journal of Cleaner Production, 254, 120069.

Scornet, E. (2023, February). Trees, forests, and impurity-based variable importance in regression. In Annales de l'Institut Henri Poincare (B) Probabilites et statistiques (Vol. 59, No. 1, pp. 21-52). Institut Henri Poincaré.

Shi, W., Tong, C., Zhang, A., Wang, B., Shi, Z., Yao, Y., & Jia, P. (2021). An extended Weight Kernel Density Estimation model forecasts COVID-19 onset risk and identifies spatiotemporal variations of lockdown effects in China. Communications biology, 4(1), 126.

Simões, M. J., Pinto, T., & Silva, C. (2023, November). Optimization of the Energy Consumption for Robotic Kitting in the Automotive Industry. In Iberian Robotics conference (pp. 483-494). Cham: Springer Nature Switzerland.

Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. Applied Soft Computing, 97, 105524.

Starke, M., & Geiger, C. (2022). Field setup and assessment of a cloud-data based crane scale (ccs) considering weight-and local green wood density-related volume references. Croatian Journal of Forest Engineering: Journal for Theory and Application of Forestry Engineering, 43(1), 29-45.

Starovoitov, V. V., & Golub, Y. I. (2021, September). Data normalization in machine learning. In Informatics (Vol. 18, No. 3, pp. 83-96).

Tan, C. W., Robin, K. R., Chuang, K. H., & Humpal, A. (2021). SourceAmerica Design Challenge Accessible Kitting and Packaging Station.

Thai, V. T., & Norlander, H. (2021). Analysing the consequences of a future implementation of a new logistics concept-A case study at Volvo Trucks.

Thamrin, N., & Wijayanto, A. W. (2021). Comparison of Soft and Hard Clustering: A Case Study on Welfare Level in Cities on Java Island: Analisis cluster dengan menggunakan hard clustering dan soft clustering untuk pengelompokkan tingkat kesejahteraan kabupaten/kota di pulau Jawa. Indonesian Journal of Statistics and Its Applications, 5(1), 141-160.

Tian, C., & Hao, Y. (2020). Point and interval forecasting for carbon price based on an improved analysis-forecast system. Applied Mathematical Modelling, 79, 126-144.

Tornese, F., Unnu, K., Gnoni, M. G., & Pazour, J. A. (2020). On-demand warehousing: main features and business models. XXV Summer School.

Vaka, D. K. (2020). Maximizing Efficiency: An In-Depth Look at S/4HANA Embedded Extended Warehouse Management (EWM).

Wang, H. (2006). Nearest neighbors by neighborhood counting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(6), 942-953.

Wang, Z., Ye, Z., Du, Y., Mao, Y., Liu, Y., Wu, Z., & Wang, J. (2022, October). AMD-DBSCAN: An Adaptive Multi-density DBSCAN for datasets of extremely variable density. In 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA) (pp. 1-10). IEEE.

Wei, F. F., Chen, W. N., Mao, W., Hu, X. M., & Zhang, J. (2023). An Efficient Two-Stage Surrogate-Assisted Differential Evolution for Expensive Inequality Constrained Optimization. IEEE Transactions on Systems, Man, and Cybernetics: Systems.

Williams, D. R., & Rodriguez, J. E. (2022). Why overfitting is not (usually) a problem in partial correlation networks. Psychological Methods, 27(5), 822.

Winkelhaus, S., & Grosse, E. H. (2022). Smart warehouses—a sociotechnical perspective. In The Digital Supply Chain (pp. 47-60). Elsevier.

Wu, H., & Li, Y. F. (2022). Clustering spatially correlated functional data with multiple scalar covariates. IEEE Transactions on Neural Networks and Learning Systems, 34(10), 7074-7088.

Wu, J., Wang, Y. G., Tian, Y. C., Burrage, K., & Cao, T. (2021). Support vector regression with asymmetric loss for optimal electric load forecasting. Energy, 223, 119969.

Xie, M., Xie, L., & Zhu, P. (2021). An Efficient Feature Weighting Method for Support Vector Regression. Mathematical Problems in Engineering, 2021(1), 6675218.

Xiong, J., Liu, X., Zhu, X., Zhu, H., Li, H., & Zhang, Q. (2020). Semi-supervised fuzzy c-means clustering optimized by simulated annealing and genetic algorithm for fault diagnosis of bearings. Ieee Access, 8, 181976-181987.

Xu, J., Teng, J., & Yao, A. C. C. (2022). Relaxing the feature covariance assumption: Time-variant bounds for benign overfitting in linear regression. arXiv preprint arXiv:2202.06054.

Xu, Q., Zhang, Q., Liu, J., & Luo, B. (2020). Efficient synthetical clustering validity indexes for hierarchical clustering. Expert Systems with Applications, 151, 113367.

Xu, T., Zhu, R., & Shao, X. (2022). On variance estimation of random forests. stat, 1050, 26.

Yang, J., Han, S., & Chen, Y. (2023). Prediction of traffic accident severity based on random forest. Journal of Advanced Transportation, 2023(1), 7641472.

Yang, L., & Wu, T. T. (2023). Model-based clustering of high-dimensional longitudinal data via regularization. Biometrics, 79(2), 761-774.

Yang, W., Wang, X., Lu, J., Dou, W., & Liu, S. (2020). Interactive steering of hierarchical clustering. IEEE Transactions on Visualization and Computer Graphics, 27(10), 3953-3967.

Yang, Y. C., Lin, T. I., Castro, L. M., & Wang, W. L. (2020). Extending finite mixtures of t linear mixed-effects models with concomitant covariates. Computational Statistics & Data Analysis, 148, 106961.

Yi, J., Yan, H., Wang, H., Yuan, J., & Li, Y. (2023, October). DeepSTA: A Spatial-Temporal Attention Network for Logistics Delivery Timely Rate Prediction in Anomaly Conditions. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (pp. 4916-4922).

Yi, Y., Sun, D., Li, P., Kim, T. K., Xu, T., & Pei, Y. (2022). Unsupervised random forest for affinity estimation. Computational visual media, 8, 257-272.

Yuan, C. (2023, July). Emergency Transportation and Structural Optimization Issues of E-commerce Logistics Network Parcels. In 2023 International Conference on Data Science and Network Security (ICDSNS) (pp. 1-7). IEEE.

Yusup, M. F. B. (2022). Impact Analysis Of Implementation Of Bonded Warehouse Policy In Makassar Port New Port On Logistic Costs. Maritime Park Journal of Maritime Technology and Society, 17-25.

Zapata, S. M., Klement, N., Silva, C., Gibaru, O., & Lafou, M. (2022, June). Collective intelligence application in a kitting picking zone of the automotive industry. In International Joint Conference on Mechanics, Design Engineering & Advanced Manufacturing (pp. 410-420). Cham: Springer International Publishing.

Zhang, C., Zhang, W., Luo, W., Gao, X., & Zhang, B. (2021). Analysis of influencing factors of carbon emissions in China's logistics industry: a GDIM-based indicator decomposition. Energies, 14(18), 5742.

Zhang, J., Wang, S., He, W., Li, J., Cao, Z., Wei, B., & Wang, M. (2022). Material kitting in selective assembly: A manual order picking system based on augmented reality. The International Journal of Advanced Manufacturing Technology, 123(1), 675-686.

Zhang, Y., Schnell, P., Song, C., Huang, B., & Lu, B. (2021). Subgroup causal effect identification and estimation via matching tree. Computational Statistics & Data Analysis, 159, 107188.

Zhang, Y., Wang, G., Chung, F. L., & Wang, S. (2021). Support vector machines with the known feature-evolution priors. Knowledge-Based Systems, 223, 107048.

Zhao, J., Zheng, Y., Seppänen, O., Tetik, M., & Peltokorpi, A. (2021). Using real-time tracking of materials and labor for kit-based logistics management in construction. Frontiers in Built Environment, 7, 713976.

Zhao, L., Zhao, F., & Che, W. W. (2023). Distributed adaptive fuzzy fault-tolerant control for multi-agent systems with node faults and denial-of-service attacks. Information Sciences, 631, 385-395.

Zhou, B., & He, Z. (2020). A material handling scheduling method for mixed-model automotive assembly lines based on an improved static kitting strategy. Computers & Industrial Engineering, 140, 106268.

Zhou, L., Simon, J. B., Vardi, G., & Srebro, N. (2023). An agnostic view on the cost of overfitting in (kernel) ridge regression. arXiv preprint arXiv:2306.13185.

Zhou, S., Liu, F., & Song, W. (2021). Estimating the optimal number of clusters via internal validity index. Neural Processing Letters, 53, 1013-1034.

Zhou, Y., Ma, N., Wang, Q., Wang, Z., Chen, C., Tao, J., ... & Cheng, Y. (2022). Bimodal distribution of size-resolved particle effective density: Results from a short campaign in a rural environment over the North China Plain. Atmospheric Chemistry and Physics, 22(3), 2029-2047.

Zhuang, Y., Chen, X., & Yang, Y. (2023, July). Likelihood adjusted semidefinite programs for clustering heterogeneous data. In International Conference on Machine Learning (pp. 43326-43346). PMLR.