

Article

The Limits of Emission-Based Learning in 3PL Operations: Evidence from Medical and Pharmaceutical Last-Mile Deliveries

Marzena Kramarz and Mariusz Kmiecik *

Logistics Institute, Department of Organization and Management, Silesian University of Technology, Akademicka 2A, 44-100 Gliwice, Poland; marzena.kramarz@polsl.pl

* Correspondence: mariusz.kmiecik@polsl.pl

Abstract

Medical and pharmaceutical last-mile deliveries are simultaneously expected to be fast, reliable and temperature-safe for patients and to become measurably greener, yet these objectives often pull transport operations in opposite directions. Third-party logistics (3PL) providers are therefore increasingly required not only to report transport CO₂ emissions, but also to learn from them; however, it remains unclear whether the routine operational data they collect are sufficiently informative to enable such emission-based learning in this regulated and service-critical setting. This study examines the predictive limits of machine learning models in estimating CO₂ emissions in medical and pharmaceutical last-mile deliveries performed by a 3PL operator. Using operational data from six customers, we compare global and customer-specific models for the following two dependent variables: total CO₂ emissions per transport operation and CO₂ emissions per pallet. Linear and non-linear models, including linear regression, ElasticNet, Random Forest, HistGradientBoosting and XGBoost, are evaluated using chronological train-test splitting and cross-validation. The results show that global models fail to outperform a naïve benchmark, with negative R² values for both emission measures. Customer-level models reveal substantial heterogeneity as follows: for selected customers, especially those with more regular operational patterns, moderate predictive performance is achieved, while for others, emissions remain largely unpredictable using the available variables. The findings suggest that routine shipment-level data are insufficient for robust emission prediction in 3PL last-mile operations. Emission-based learning requires richer contextual, vehicle, route, traffic and telematics data, as well as customer-sensitive modelling approaches. The study contributes by identifying the data and modelling limits of sustainability intelligence in medical and pharmaceutical last-mile logistics.

Keywords: CO₂ emissions prediction; machine learning; third-party logistics (3PL); last-mile delivery; pharmaceutical and medical logistics

Academic Editor: Wen-Chyuan Chiang

Received: 29 May 2026

Revised: 25 June 2026

Accepted: 29 June 2026

Published: 1 July 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Last-mile logistics has become one of the most important areas of sustainability pressure in contemporary supply chains. It is operationally complex, cost-intensive and environmentally burdensome, particularly in urban and service-critical distribution systems [1]. In the healthcare sector, the importance of last-mile logistics is even greater.

Medical and pharmaceutical products must be delivered efficiently, reliably and safely, in compliance with strict quality and traceability requirements. Delays, temperature deviations or delivery failures may affect the availability of treatment, the integrity of products and, ultimately, patient safety [2]. At the same time, these service-critical requirements make decarbonisation particularly difficult: tight delivery time windows, temperature control and high reliability expectations constrain route consolidation, vehicle choice and load factors, so that reducing CO₂ emissions in healthcare last-mile transport cannot be separated from the clinical and regulatory constraints that shape it. Controlling emissions in this setting is therefore not only a reporting task but a genuine operational trade-off against service quality.

At the same time, third-party logistics providers are increasingly expected to measure, report and reduce the environmental impact of their transport operations. This pressure is driven by sustainability strategies, customer requirements, regulatory expectations and the growing need for transparency in logistics networks. In this context, emission data are no longer only a reporting requirement [3]. They may become a potential source of organisational learning, helping 3PL providers understand when, where and why emissions arise, and how transport operations can be redesigned to reduce their environmental footprint.

However, the ability to learn from emission data should not be taken for granted. In many logistics organisations, available data are limited to basic operational variables, such as distance, shipment size, dates, loading and unloading locations or the number of pallets [4]. Such data may be sufficient for retrospective reporting, but not necessarily for accurate prediction or decision support. Emissions in last-mile transport are shaped by multiple factors that are often missing from routine operational datasets, including vehicle type, load consolidation, route microstructure, traffic conditions, driving behaviour, stop density, weather, delivery time windows and temperature-control requirements [5].

This problem is particularly relevant in medical and pharmaceutical last-mile logistics. These supply chains are characterised by high product heterogeneity, strict regulatory constraints, varying delivery destinations and different service requirements across customers. A single 3PL network may simultaneously serve hospitals, pharmacies, laboratories, distribution points and patient-oriented channels. As a result, emission patterns may differ significantly between customers, depending on routing structures, delivery frequency, product mix, shipment size and consolidation opportunities. Crucially, the same features also shape the carbon footprint of each delivery and, just as importantly, its predictability: heterogeneous products, destinations and service levels translate into heterogeneous routing, consolidation and vehicle-use patterns, so that the conditions under which emissions arise—and can be controlled—differ markedly from one customer to another. This raises the question of whether one global emission prediction model can represent such heterogeneous operations, or whether customer-specific modelling is required.

Although machine learning is increasingly used in sustainable logistics, transport planning and emission modelling [6–8], existing studies often assume that more advanced algorithms will improve prediction quality if sufficient data are available. Less attention has been paid to the following opposite problem: what happens when routine operational data contain too little signal to support reliable prediction [9,10]? This issue is important because failed or weak prediction models may still provide valuable managerial insight. In other words, a model that predicts poorly is not merely a modelling failure but also a diagnostic result: it demonstrates that the variables currently captured do not carry enough signal, and thereby pinpoints which additional data sources (for example vehicle, route or telematics data) would have to be collected before emissions can be predicted and actively managed. They can reveal the limits of existing data systems, identify missing

sources of operational information and show where sustainability intelligence cannot yet be built.

This study addresses this gap by examining the limits of emission-based learning in 3PL medical and pharmaceutical last-mile operations. The aim is to test whether routine operational data collected by a 3PL provider are sufficient to predict CO₂ emissions at the level of individual transport operations and whether such prediction can support the development of sustainability-oriented learning capabilities. The empirical analysis uses operational data for six customers from the medical and pharmaceutical sector. The following five machine learning models are compared: linear regression, ElasticNet, Random Forest, HistGradientBoostingRegressor and XGBoost. The models are estimated both for the full dataset and separately for each customer. The following two dependent variables are analysed: total CO₂ emissions per transport operation and CO₂ emissions per pallet. The interpretation of model performance is anchored against two naïve benchmarks (mean and median predictors), which makes the meaning of negative R² values directly observable.

The study is guided by the following research questions:

RQ1. *How accurately can machine learning models predict CO₂ emissions in medical and pharmaceutical last-mile deliveries performed by a 3PL operator?*

RQ2. *How can emission prediction contribute to the development of sustainability intelligence in 3PL logistics networks?*

This paper makes three connected contributions. Empirically, it shows how far routine operational data can actually go in explaining delivery-level emissions in last-mile healthcare logistics, and where that explanatory power runs out. This limit, however, is not uniform across the network: because customers differ in routing regularity, product mix and consolidation, a single global model tends to hide segment-level structure that becomes visible only when each recipient is modelled on its own. Building on these two results, the study recasts emission prediction as a precondition for emission-based learning rather than a by-product of reporting: measuring carbon footprint is not the same as being able to act on it. Taken together, the findings indicate that machine learning will support sustainability-oriented decisions in 3PL operations only once billing-level data are extended with the contextual, route and telematics information they currently omit.

The remainder of the paper is structured as follows. Section 2 reviews the literature on healthcare last-mile logistics, machine learning in sustainable logistics and emission-based learning and develops an analytical framework. Section 3 describes the data, variables and modelling workflow. Section 4 presents the results for the full dataset and customer-specific models. Section 5 discusses the findings and their implications for 3PL sustainability intelligence. Section 6 concludes the paper and outlines directions for future research.

2. Theoretical Background

2.1. Healthcare Last-Mile Logistics as a Complex and Data-Constrained System

2.1.1. Last-Mile Logistics as a Critical Stage of Healthcare Supply Chains

Last-mile logistics is widely regarded as one of the most critical and operationally demanding stages of supply chain management [1]. While in conventional distribution systems the last mile is often associated with the final movement of goods to the customer, in healthcare supply chains its meaning is broader and more consequential [11]. It refers not only to the physical delivery of products to hospitals or pharmacies, but also to the assurance that medicines, medical devices, diagnostic materials and other health

commodities reach the point of care, laboratory hub, retail pharmacy or patient's home at the required time, in the required condition and with full traceability.

The criticality of last-mile logistics in healthcare stems from its direct connection with access to treatment and continuity of care. In public health systems, last-mile distribution is responsible for ensuring that life-saving medicines and essential commodities are available even in remote or infrastructure-constrained regions [12]. WHO defines last-mile logistics for medicines and health products as the transport of products from a distribution hub or warehouse to the final beneficiary at health facility or community level, stressing the role of information systems in making such flows visible and controllable [13]. Similarly, global health supply chain programmes emphasise that improved last-mile logistics increases commodity availability and reduces stock-outs at service delivery points [14].

In healthcare, the last mile also performs a quality-protection function. Products such as vaccines, biologics, laboratory specimens or temperature-sensitive pharmaceuticals require controlled handling conditions until the moment of use. Any delay, temperature excursion or loss of traceability may reduce product effectiveness, affect diagnostic accuracy or create direct risks for patients. Studies on cold-chain medicines show that the last mile remains particularly vulnerable because temperature control, storage practices and patient-level handling are often less standardised than upstream logistics processes [15].

At the same time, last-mile logistics affects the operational performance of healthcare systems. Delivery reliability influences fill rates, stock availability, expiry reduction, emergency replenishment needs and the ability of healthcare providers to maintain service continuity. Industry analyses also point out that last-mile logistics supports a wide range of clinical, diagnostic and administrative flows, including pharmaceuticals, lab specimens, medical devices, consumables and documentation [16]. Therefore, last-mile logistics should not be treated merely as a transport activity, but as a service-critical interface between supply chain operations and healthcare outcomes.

From a broader logistics perspective, last-mile delivery also connects healthcare supply chains with urban logistics and city freight systems. The growth of home healthcare, direct-to-patient models, decentralised diagnostics and e-pharmacy delivery extends the last mile beyond institutional healthcare facilities. This creates new interactions between healthcare logistics, urban freight regulation, congestion, environmental restrictions and sustainable city logistics. Research on sustainable last-mile delivery highlights the growing importance of stakeholder coordination, alternative delivery modes, environmental performance and urban policy in shaping last-mile solutions [17]. In this sense, healthcare last-mile logistics contributes not only to customer service and patient access, but also to the sustainability of distribution networks and, consequently, entire supply chains.

For the present study, the very characteristics that make healthcare last-mile logistics critical—tight time windows, frequent small-batch deliveries, dispersed destinations and strict handling conditions—are also the operational drivers of its CO₂ emissions. Every additional stop, every detour imposed by a temperature-control or urgency constraint and every sub-optimal load factor translates into additional fuel burn and emissions. The references reviewed above therefore matter for decarbonisation in a very concrete way: understanding and reducing the carbon footprint of healthcare last-mile transport requires data that capture these service-critical constraints, and not only aggregate distances and volumes. This observation directly motivates the empirical question, examined later in the paper, of whether routine operational data are rich enough to explain emissions.

2.1.2. Specificity of Medical and Pharmaceutical Last-Mile Logistics

Medical and pharmaceutical last-mile logistics differs from standard commercial last-mile delivery because it is embedded in highly regulated, risk-sensitive and service-critical supply chains. Healthcare supply chains typically involve manufacturers, distributors, third-party logistics providers, wholesalers, pharmacies, hospitals, laboratories, public agencies and patients. In many public health systems, medical products flow through central medical stores, regional warehouses and service delivery points, whereas in private or mixed systems they may be delivered through wholesalers, 3PL networks, direct-to-pharmacy channels or patient-facing delivery models. The Reach Alliance case study on Tanzania illustrates how public medical supply chains may evolve from fragmented distribution structures towards more direct, data-enabled and GPS-supported delivery to rural health facilities [18].

A central feature of medical and pharmaceutical logistics is product heterogeneity. Products differ in clinical function, sensitivity, regulatory status, value density, shelf life and transport requirements. Some products, such as vaccines, biologics, insulin or selected injectable medicines, require strict temperature control. Others, such as medical devices, consumables, test kits, gloves or documentation, may be less temperature-sensitive but still require delivery reliability, traceability and damage prevention. The public health logistics literature also classifies commodities by disease programme or clinical area, for example HIV/AIDS, malaria, family planning, maternal and child health, neglected tropical diseases or laboratory diagnostics [13].

This heterogeneity has direct implications for last-mile operations. Cold-chain products impose requirements related to temperature monitoring, validated packaging, controlled storage and limited exposure time. Short shelf-life products increase the need for frequent replenishment and careful stock rotation. Laboratory specimens require rapid and reliable transport because delays may affect diagnostic validity. Medical devices and field inventory often require high visibility and accurate allocation across dispersed points of use, as illustrated by logistics service providers' healthcare case studies. The same last-mile network may therefore handle flows with very different service priorities as follows: life-critical urgency, temperature integrity, regulatory compliance, cost efficiency, route consolidation and patient convenience [19]. From the carbon perspective, each of these service priorities maps onto a different emission profile as follows: urgent specimen runs favour dedicated, poorly consolidated trips; cold-chain deliveries add the energy load of active temperature control; and dispersed patient-facing channels increase stop density and detours. Product and service heterogeneity is thus, simultaneously, heterogeneity of emission generation, which is precisely why a single emission pattern is unlikely to hold across customers.

Existing research has addressed several aspects of medical and pharmaceutical last-mile logistics, including public health commodity distribution, cold-chain challenges, inventory visibility, digital logistics information systems, drone and IoT-enabled monitoring, blockchain-based traceability, and smart parcel lockers for pharmaceutical delivery [19,20]. The recent WHO guidance on last-mile logistics information systems confirms the growing policy relevance of medicine and health-product visibility at facility and community level [13]. Studies on digitalisation of last-mile logistics information systems also suggest that better information flows can improve health commodity coverage and may reduce adverse health outcomes [20].

However, the literature remains uneven. A considerable part of the evidence comes from reports, programme evaluations, policy guidance, consultancy publications or case studies rather than from peer-reviewed logistics journals. These sources are valuable because they document real operational practices, especially in low- and middle-income countries, but they often focus on access, availability, stock-outs and inventory control

rather than on environmental performance. Peer-reviewed studies increasingly address cold-chain integrity, routing, digitalisation and technology-enabled traceability, but less attention is paid to whether routine operational data collected by 3PL providers are sufficient to support emissions prediction and sustainability learning.

This creates a relevant research gap. Existing studies usually assume that better data improve visibility and decision-making, yet they rarely test the predictive value of routine 3PL operational data for emission modelling at the level of individual deliveries. Moreover, customer heterogeneity is often underdeveloped. Medical and pharmaceutical clients may differ in delivery frequency, routing patterns, product mix, consolidation levels, service requirements and destination structure. These differences can strongly influence the predictability of emissions, but they are rarely modelled explicitly. Consequently, there is limited evidence on whether a single global emission model can represent heterogeneous healthcare logistics flows, or whether customer-specific models are required. Translated into the carbon dimension, the following two questions remain largely open: whether the operational data routinely held by a 3PL provider carry enough signal to predict delivery-level emissions (RQ1), and whether such prediction could become a basis for emission-based learning rather than mere reporting (RQ2). These two questions motivate the analytical framework developed in Section 2.3.

2.2. Machine Learning for Sustainable Logistics

The literature on sustainable logistics increasingly notes that the development of analytical tools—machine learning methods in particular—is changing how firms assess and limit the environmental impact of their transport operations [6,7]. Unlike classical statistical techniques, ML allows the analysis of data of considerably greater complexity, which opens the way to more precise emission modelling and to the identification of sources of inefficiency [21,22]. For this reason, ML is sometimes regarded as one of the pillars of digital transformation in logistics, although its actual use in practice remains uneven.

2.2.1. The Role of ML in Emission Reduction

Research on transport indicates that ML is applied mainly in two areas. First one, described among others by Yin et al. [23], Zhao and He [24] and Natarajan et al. [25], concerns the modelling of emissions and fuel consumption. Regression and non-linear models—especially those from the boosting family—capture relationships that are difficult to express with simple methods, including the effect of route length, route variability, vehicle load and the parameters of the operating day. Several studies have noted that combining operational data with telematics information allows a markedly more accurate representation of CO₂ emissions than traditional calculation-based approaches. The next area, described among others by Karami and Kashef [26] and Vadlamani and Modashiya [27], concerns operational processes. ML models are used, for example, to detect inefficient routes, identify empty runs or predict situations that lead to increased fuel consumption (such as delays or long stops). In practice, many firms use ML as a decision-support tool rather than an automatic forecasting solution, because prediction quality depends directly on the quality of the input data—and that quality is often limited.

ML also appears in research on demand forecasting [28] and operational planning [29], which indirectly affects emissions. More stable volume forecasts mean fewer unnecessary runs and greater regularity of operations. Although these applications are not designed solely with emissions in mind, the environmental effects are noticeable.

In the work of 3PL operators, functioning under high variability of routes, customers and volumes, ML is applied in several areas. It appears most often in analyses concerning:

- estimation of volumes and scheduling of operations [30,31];
- assessment of delivery punctuality and prediction of disruptions [32];
- route optimisation with respect to time or cost [33];
- emission estimation, especially in urban and last-mile transport [34].

2.2.2. Examples of Machine Learning Methods Used in Logistics

Three groups of methods dominate logistics analyses. Although their applications are similar, they differ in the range of possible interpretations and in their data requirements.

The first group consists of methods based on decision trees. Decision trees and models composed of many trees (Random Forest, Extra Trees) are among the most frequently used tools in operations research [35]. They cope well with mixed-type data and do not require extensive feature preparation. In transport analyses they appear, among others, in models predicting fuel consumption [36] and the risk of delays [37]. Their drawback is limited stability when predictors are very weakly informative, which is often the case in last-mile datasets.

Boosting models—including XGBoost, LightGBM and classical GBM—usually obtain the best predictive results in logistics tasks. They can capture complex non-linear relationships and scale well to large datasets. For this reason, they are popular in emission modelling based on telematics driving measurements, where the data have a spatio-temporal character [38,39].

Neural networks are used mainly in more advanced applications, including the analysis of GPS routes [40] or emission modelling based on detailed telematics data [41]. Sequential architectures (such as LSTM) perform well with time signals [42], and GNN models appear in research on representing the structure of road networks [43]. Because of their high data requirements, these methods are used less often in strictly operational 3PL analyses, where access to telematics is often limited.

Beyond this methodological taxonomy, a growing body of empirical work reports how these algorithms actually perform once they are applied to real logistics and healthcare data, which is decisive for the present study. In a systematic review of machine learning in logistics and supply chain management, Akbari and Do [8] show that tree-based and boosting methods dominate transport and distribution applications, but that the reported performance is highly uneven and depends far more on data availability and quality than on the choice of algorithm. In the pharmaceutical domain specifically, Zhu et al. [44] demonstrate, on two large manufacturer datasets, that advanced ML models outperform classical baselines for drug-demand forecasting only when they are enriched with additional supply-chain features (for example, downstream inventory information); without such enrichment the advanced models do not beat simple references. Comparable evidence in the 3PL context, where ML supports volume estimation and operational scheduling [30,31], and in healthcare supply chains, where prioritised multi-task models have been proposed [45], points in the same direction: ML brings measurable value mainly in segments with regular, well-instrumented operations, and far less where the input data are sparse and noisy. Two conclusions follow for emission modelling. First, the prevailing assumption that more advanced algorithms automatically improve prediction is not supported once data informativeness is taken into account. Second, the studies above examine demand, delivery times or disruptions, but very rarely delivery-level CO₂ emissions estimated from routine 3PL billing data—the gap this paper addresses.

On the basis of the literature review it can be concluded that ML has considerable potential for reducing emissions in logistics, although its effectiveness depends primarily on the availability of high-quality data.

2.3. Analytical Framework: From Emission Reporting to Emission-Based Learning

The literature reviewed above reveals two recurrent assumptions that this study questions. The first is that collecting and reporting emissions data is broadly equivalent to being able to manage them; the second is that algorithmic sophistication is the main lever of predictive performance. Building on the concept of emission-based learning—the organisational capability to convert emission data into actionable knowledge for redesigning transport operations—we propose a three-layer analytical framework that makes these assumptions testable.

The framework distinguishes (i) a data layer, comprising the variables a 3PL operator routinely records (distance, pallets, locations, dates, derived calendar and utilisation features); (ii) a prediction layer, in which machine learning models attempt to map these variables onto delivery-level emissions; and (iii) a learning/decision layer, in which validated predictions are used to steer consolidation, routing and vehicle allocation toward lower emissions. The central claim of the framework is sequential: the learning layer can only be reached if the prediction layer attains validity, which in turn depends on whether the data layer contains sufficient signal. Reporting, by contrast, requires only the data layer and is therefore necessary but not sufficient for learning.

This logic is moderated by customer heterogeneity. Because medical and pharmaceutical customers differ in routing regularity, product mix, consolidation and service requirements (Section 2.1.2), the signal available in identical operational variables is expected to vary across customers. The framework therefore predicts that predictive validity—and hence the feasibility of emission-based learning—is not a global property of the network but a segment-specific one. The following three propositions summarise the framework and structure the empirical analysis:

P1. *With predictors restricted to routine operational variables, machine learning models do not attain practically useful predictive validity for delivery-level CO₂ emissions at the level of the whole network and do not outperform naïve benchmarks.*

P2. *Predictive validity is moderated by customer operational regularity: it is higher for customers with stable, repeatable supply structures and lower for highly irregular ones, so that global models mask segment-specific signal.*

P3. *Because validity is a precondition for the learning/decision layer, emission-based learning from routine data is feasible only in regular customer segments; elsewhere it requires the data layer to be extended with contextual, vehicle, route and telematics information.*

Sections 4 and 5 evaluate these propositions empirically, using the global analysis to test P1, the per-customer analysis to test P2, and the joint interpretation to test P3.

3. Materials and Methods

3.1. Data Collection

The data analysed concerned the operational data of a 3PL operator relating to transport activity for six customers (recipients of transport processes) from the medical and pharmaceutical industry. The characteristics of the data are described in the later subsection on variables and predictors. Data preprocessing consisted of removing incomplete records (for example, those without an emission level or without the number of pallets in a delivery). For the purposes of the study the date range was also restricted, from 8 December 2021 to:

- Client01—31 July 2024;
- Client02—4 September 2024;

- Client03—23 August 2024;
- Client04—31 July 2024;
- Client05—10 August 2024.

Only direct transport operations were taken into the analysis, that is, without transshipments or additional loading along the route. With respect to data completeness, the raw dataset was not fully complete: a subset of records lacked an emission value or the number of pallets and, because both are indispensable for constructing the target variables, such records were removed list-wise rather than imputed. For the records retained, sporadic missing values among the predictor variables were handled inside the modelling pipeline (median imputation for numerical predictors and most-frequent-value imputation for categorical ones; see Section 3.4), so that no further observations were discarded at the predictor stage. The characteristics of the customers are presented in Table 1. Emission columns (CO₂, CO_{2e}, NO_x, NMHC, and SO_x) are expressed in kg over the full date range.

Table 1. Brief characteristics of the customers (transport recipients).

Customer ID	Share of Transport Days in All Workdays in Date Range	Total Emission [kg] in Full Date Range					Number of Unloading Locations (by ZIP Code)	Number of Unique Unloading Locations
		CO ₂	CO _{2e}	NO _x	NMHC	SO _x		
Client01	94.73%	9.11	9.33	41.62	3.87	8.38	400	845
Client02	90.42%	24.67	25.23	119.58	9.71	18.81	843	3333
Client03	69.90%	12.43	12.71	67.94	4.69	8.56	853	1656
Client04	85.32%	7.93	8.36	32.18	6.47	7.97	342	811
Client05	98.06%	20.66	21.16	89.53	8.56	18.09	574	1298
Client06	69.22%	4.42	4.65	19.10	2.56	4.25	195	277

The data used were divided into target variables and predictors.

3.2. Variables Description

Among the target variables was the CO₂ emission value in grams. This value was assigned to each transport operation carried out and followed directly from the input data.

In addition, a CO₂ per pallet indicator was created as a target variable, obtained by simply dividing the emission value by the number of pallets in a given transport. Neither of the two target variables was subject to logarithmic transformation; preliminary tests did not show a meaningful improvement in the stability of the results, and the original scale allows the errors (RMSE, MAE) to be interpreted directly in grams of CO₂.

The predictors were divided into base predictors derived from the dataset and additional predictors created on the basis of the base data.

The base predictors included:

- Distance;
- Quantity_pallets;
- Loading_ZIP, Unloading_ZIP—categorical, encoded with OneHotEncoder limited to the 100 most frequent categories in the training set (max_categories = 100). Remaining postal codes and values outside the training set are handled through the handle_unknown = “ignore” parameter, which eliminates the leakage risk characteristic of target encoding and at the same time limits the dimensionality of the problem;
- Quantity_pallets_space_on_floor.

The additional predictors were created as follows:

- Transport time = (Unloading_Date – Loading_Date);
- Speed proxy = Distance/Transport time;
- Pallet space utilisation = Quantity_pallets_space_on_floor/Quantity_pallets;
- Seasonal indicator (from the dates: Q1, Q2, Q3, and Q4).

It should be noted that, in medical and pharmaceutical last-mile operations, transport time does not consist solely of driving. It also incorporates time-consuming collection, handover and documentation procedures at hospitals, pharmacies and laboratories, which can vary substantially between stops and between customers. The speed proxy defined above therefore conflates genuine driving dynamics with non-transport dwell time and is likely to be a noisy and only weakly informative predictor of emissions. This is one plausible reason for the limited predictive power of the variable observed in the results below, and it should be read as a caveat when interpreting the role of speed-related features.

The target variables and predictors were used in the procedure specified in the next section. Specifically, the modelling methods used are described first, in Section 3.3, and the procedure in which the target variables and predictors were combined with these methods is then specified in Section 3.4.

3.3. Chosen ML Methods

Five models were used in the analysis, representing both linear and non-linear approaches. The models were chosen so as to allow a comparison of simple reference methods with more complex algorithms that learn non-linear relationships. Table 2 synthesises the main strengths and weaknesses of each of them in the context of the operational data analysed.

Table 2. Strengths and weaknesses of the chosen models.

Model	Strengths	Weaknesses
Linear Regression	Closed form and fast to estimate; each coefficient gives a directly interpretable, signed effect of a predictor, which makes the model a meaningful baseline for judging the gain of the more complex methods.	Captures only additive, linear effects; sensitive to multicollinearity among predictors and to outliers; cannot represent interactions or threshold effects.
ElasticNet	Adds L1 and L2 penalties to linear regression, which stabilises estimation when predictors are numerous or correlated and performs implicit variable selection; less prone to overfitting than ordinary least squares.	Remains a linear model, so it cannot capture non-linear relationships or interactions; requires tuning of the penalty strength and of the L1/L2 ratio.
RandomForest	Models non-linearities and interactions without feature scaling or distributional assumptions; robust to outliers; provides feature-importance measures that help identify the most relevant predictors.	Tends to overfit when the predictors carry little signal; large ensembles are memory-heavy and slower to score; predictions are harder to interpret than linear coefficients.
HistGradientBoostingRegressor or	Histogram-based boosting that is fast and memory-efficient on large datasets; models non-linear effects and interactions; handles missing values natively without prior imputation.	Sensitive to hyperparameters (tree depth, learning rate, number of iterations); can overfit when the signal is weak; lower interpretability than linear models.
XGBoost Regressor	Regularised gradient boosting with strong predictive performance on tabular data; scales to large datasets and handles missing values internally; widely used and well-supported.	Requires careful tuning to perform well; may overfit on low-signal features; more complex to configure and interpret than the linear models.

Source: own elaboration based on [35,41–44].

The use of models of varying complexity made it possible not only to compare their effectiveness, but also to assess whether non-linear structures can be captured using the available operational predictors.

3.4. Research Workflow

The research process comprised the following two stages: (i) analysis of the full dataset for all customers and (ii) analysis carried out separately for each customer. As shown in Figure 1, the workflow was designed as a sequence of ordered steps ensuring consistency and comparability of results across models.

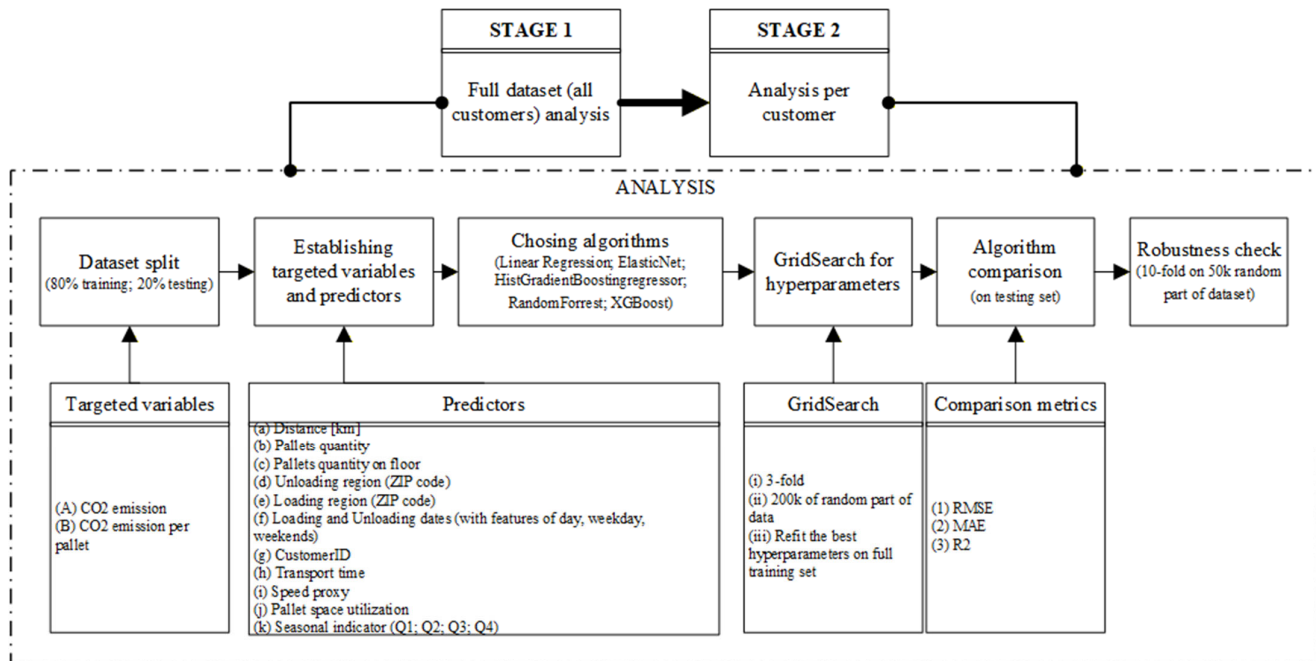


Figure 1. Research workflow. Source: own work.

In the first step the data were split chronologically into a training set (80%) and a test set (20%). Next, the target variables (CO₂ emission and CO₂ emission per pallet) and the set of operational and calendar predictors were defined. Five algorithms were used for modelling—linear regression, ElasticNet, HistGradientBoostingRegressor, RandomForestRegressor and XGBoost—representing both linear and non-linear approaches. Hyperparameter tuning was carried out using a simplified GridSearch: three-fold validation on a random sample of 200 thousand observations, followed by refitting of the best parameter set on the full training set. The models were evaluated on the test set using the following three metrics: RMSE, MAE and R². In addition, a robustness check was performed through 10-fold cross-validation on a random sample of 50 thousand observations. The workflow designed in this way ensured transparency of the procedure, the possibility of comparing algorithms and control over the stability of the results obtained.

The preprocessing pipeline for the sklearn models comprised: SimpleImputer (median for numerical variables, most frequent value for categorical ones), StandardScaler for continuous variables and OneHotEncoder with the parameters max_categories = 100 and handle_unknown = "ignore" for categorical variables. The use of OneHotEncoder instead of target encoding is methodologically important: it eliminates the risk of leakage of test-set statistics into training, and the max_categories = 100 parameter simultaneously guards against the dimensional explosion resulting from the number of unique postal

codes. Rare postal codes (occurring outside the TOP-100 group in the training set) and entirely new values are represented as a vector of zeros in the one-hot space, which corresponds to a “treat as unknown” strategy for external predictors. The XGBoost model in the BigQuery ML environment (BOOSTED_TREE_REGRESSOR) operates on the platform’s native pipeline: it does not use OneHotEncoder, because the handling of categorical variables is internal; it does not apply feature scaling, because the boosting algorithm is insensitive to scale; and observations with missing values are automatically dropped rather than imputed. Despite the differences in preprocessing between the two environments, the resulting test metrics are comparable—which signals that prediction quality is limited mainly by the informativeness of the features, and not by the choice of encoding or scaling technique. The coefficient of determination R^2 was calculated from the standard formula $R^2 = 1 - SS_{\text{res}}/SS_{\text{tot}}$, where SS_{res} is the residual sum of squares of the model on the test set and SS_{tot} is the sum of squared deviations of the test observations from their mean. $R^2 = 0$ corresponds to a model predicting a constant equal to the mean of the test set; $R^2 < 0$ means that the model performs worse than such a trivial prediction. To enable direct interpretation of the negative R^2 values obtained in the analysis, the following two naïve benchmarks were added to each set of results: a constant predictor equal to the mean of y_{train} (mean predictor) and a constant predictor equal to the median of y_{train} (median predictor). The mean predictor serves as a theoretical reference—its R^2 is by definition equal to zero when $\text{mean}(y_{\text{train}}) = \text{mean}(y_{\text{test}})$ and converges to zero in other cases. The median predictor serves as a practical reference for skewed distributions (such as CO₂ emission), where it minimises the mean absolute error and yields $R^2 \approx 0$.

4. Results

4.1. Prediction for the Full Dataset

The prediction analysis was carried out for the full dataset comprising 665,280 observations in the training set and 166,320 observations in the test set (an 80/20 split by loading date). The following two levels of aggregation of the dependent variable were tested: (i) total CO₂ emission per transport (CO₂_Emission) and (ii) CO₂ emission per pallet (CO₂_per_pallet).

For each, four models were estimated in the sklearn environment (Python 3.12)—linear regression, ElasticNet, HistGradientBoostingRegressor and RandomForestRegressor—as well as one model based on the BigQuery ML environment, the Boosted Tree Regressor (XGBoost model). The evaluation results on the test set and in 10-fold cross-validation are presented in Appendix A. In XGBoost, rows containing missing data or categories that are too rare are automatically dropped, whereas in sklearn they are imputed and retained thanks to the full pipeline (imputation, scaling, and limitation of the number of categories). As a result, XGBoost trains and records metrics on a smaller, filtered subset of the data, which leads to the visible differences in the tables that compare these two model environments.

4.1.1. Results for CO₂ per Pallet Emission

For the CO₂_per_pallet variable, the lowest RMSE on the test set was obtained for the ElasticNet model (RMSE = 0.2093; MAE = 0.1277; and $R^2 = -3.62$). Linear regression achieved a worse result (RMSE = 0.2864; MAE = 0.1390; and $R^2 = -7.66$), while the tree ensemble models were decidedly weaker in terms of fit on the test set:

- HistGradientBoostingRegressor—RMSE = 0.4896; MAE = 0.0938; and $R^2 = -24.30$;
- RandomForestRegressor—RMSE = 0.6781; MAE = 0.1190; and $R^2 = -47.54$.

In terms of squared error, both tree models clearly lose to the linear models, and especially to ElasticNet. At the same time, all values of the coefficient of determination R^2 on the test set are strongly negative, which means that none of the models outperforms the simple benchmark of predicting the mean value of CO₂_per_pallet in the training set. The scale of this discrepancy is measurable: the constant mean predictor of y_{train} (Table A1) achieves RMSE = 0.1210 and the median predictor RMSE = 0.0977 (with $R^2 \approx 0$). ElasticNet, with RMSE = 0.2093, is therefore about 2.1 times worse in terms of squared error than the trivial median prediction; HGBR is five times worse, and RandomForest seven times worse. The negative R^2 values are not an artefact of scale here—they reflect a genuine advantage of the constant prediction over the operational models.

The results of 10-fold cross-validation on a sample of 50 thousand observations confirm the weak explanatory ability of the models. The mean RMSE for ElasticNet was about 0.804 (std ≈ 1.05), with $R^2 \approx -0.10$ (std ≈ 0.20), which indicates the absence of a stable improvement over the naïve model and a relatively large variability of prediction quality between folds. For linear regression the mean CV R^2 was even lower (about -0.95), which further underlines the limited usefulness of simple linear models for emission per pallet.

The XGBoost model built on the full feature set (with `data_split_method = CUSTOM` and the `is_train` column) also did not improve quality. The metrics obtained were: RMSE = 0.3916; MAE = 0.3841; and $R^2 = -15.18$. The R^2 value is even more negative here than for the sklearn models—XGBoost achieves an RMSE about four times higher than the naïve median (0.3916 vs. 0.0977) and ranks the weakest of the tested models in terms of fit. For CO₂_per_pallet, the best of the tested models (ElasticNet) minimises the RMSE error but does not achieve a positive R^2 and therefore does not provide statistically satisfactory predictive quality at the level of a single transport relation.

4.1.2. Results for CO₂ Emission

For total CO₂ emission per transport, the results obtained are somewhat better, but still far from the expectations for operational models. The lowest RMSE on the test set was again achieved by the ElasticNet model (RMSE = 0.3395; MAE = 0.2180; $R^2 = -1.14$). The remaining models behaved as follows:

- Linear regression—RMSE = 0.4159; MAE = 0.2323; and $R^2 = -2.21$;
- HistGradientBoostingRegressor—RMSE = 0.7577; MAE = 0.1940; and $R^2 = -9.66$;
- RandomForestRegressor—RMSE = 0.7934; MAE = 0.2313; and $R^2 = -10.69$.

Again, it can be seen that ElasticNet provides the lowest squared error, while the tree models are characterised by markedly higher RMSE with very strongly negative R^2 values. This means that the more complex non-linear models not only fail to improve the fit but actually degrade prediction quality relative to the simple linear models. The naïve benchmarks for CO₂_Emission (Table A2) are RMSE = 0.2634 (mean predictor) and RMSE = 0.2321 (median predictor). ElasticNet, with RMSE = 0.3395, is therefore about 46% worse in terms of RMSE than the median prediction; XGBoost is 73% worse; and HGBR and RandomForest are about 2.3 and 2.4 times worse, respectively. The pattern is consistent with the results for CO₂_per_pallet: the strongly negative R^2 values are a consequence of worse prediction than a constant, not a matter of error scaling. In 10-fold cross-validation the results are slightly more promising:

- For ElasticNet—RMSE ≈ 1.2233 (std ≈ 0.92), MAE ≈ 0.2026 (std ≈ 0.011), and $R^2 \approx 0.0166$ (std ≈ 0.066);
- For HistGradientBoostingRegressor—RMSE ≈ 1.2228 , MAE ≈ 0.1875 , and $R^2 \approx 0.0172$;
- For RandomForestRegressor—RMSE ≈ 1.1680 , MAE ≈ 0.1620 , and $R^2 \approx 0.0213$.

The cross-validation R^2 values for the three non-linear models are slightly positive, but at a level of about 0.02, which means that these models explain only a few percent of

the variance of emissions—and only on average, with no guarantee of transferability to the final test set (where R^2 remains negative). The standard deviations of the metrics are relatively high, which signals a considerable sensitivity of the models to changes in the training sample. The XGBoost model for CO₂_Emission achieved RMSE = 0.4018, MAE = 0.3550, and R^2 = −2.00, which places it between linear regression and ElasticNet in terms of RMSE, but still with a negative R^2 . From the point of view of practical usefulness, this means that deploying this model would not lead to an improvement in forecast quality relative to a simple naïve model.

4.2. Prediction per Customer

In the second stage of the analyses, models were estimated separately for each customer (Client01, Client02, Client03, Client04, Client05, and Client06). In each case the same scheme was applied as follows: an 80/20 split by loading date, the set of operational and calendar features as in the aggregated analysis, and an identical set of models (linear regression, ElasticNet, HistGradientBoostingRegressor, RandomForestRegressor and XGBoost). The aim was to check whether modelling emissions at the level of individual recipients allows higher prediction accuracy than the global model. In general, the results are strongly differentiated between customers and in most cases R^2 on the test set remains negative; however, for some customers (especially Client02, Client04 and Client06) moderately positive R^2 values are observed for the CO₂_Emission variable, which suggests a certain local structure that can be captured by the models.

To enable a synthetic comparison of results across customers and targets, Table 3 presents, for each customer × variable combination, the best ML model (selected by the lowest test RMSE) together with two naïve benchmarks. The column $\Delta\text{RMSE} = \text{RMSE_ML} - \text{RMSE_naive_mean}$ indicates for which combinations operational modelling brings a real improvement over the trivial prediction

Table 3. Per-customer summary—best ML model vs naïve benchmarks.

Client	Target	Best ML Model	RMSE ML	MAE ML	R^2 ML	RMSE Naive (Mean)	R^2 Naive (Mean)	RMSE Naive (Median)	R^2 Naive (Median)	$\Delta\text{RMSE} = \text{ML} - \text{Naive_Mean}$
Client01	CO2_Emission	BQML XGBoost	0.2601	0.2543	−8.2528	0.1194	−0.9511	0.0855	−0.0003	+0.1406
Client01	CO2_per_pallet	RandomForest	0.0969	0.0389	−9.6901	0.0457	−1.3699	0.0297	−0.0041	+0.0513
Client02	CO2_Emission	BQML XGBoost	0.3553	0.3255	−2.4491	0.2260	−0.3960	0.2114	−0.2216	+0.1293
Client02	CO2_per_pallet	BQML XGBoost	0.3867	0.3840	−33.6443	0.0804	−0.4997	0.0706	−0.1575	+0.3063
Client03	CO2_Emission	BQML XGBoost	0.3777	0.3703	−9.8067	0.4144	−11.9943	0.1484	−0.6663	−0.0367
Client03	CO2_per_pallet	HistGradientBoosting	0.2418	0.1575	−5.1912	0.2454	−5.3811	0.1227	−0.5945	−0.0037
Client04	CO2_Emission	ElasticNet	0.1380	0.0876	0.1409	0.1529	−0.0569	0.1684	−0.2833	−0.0149
Client04	CO2_per_pallet	HistGradientBoosting	0.0223	0.0082	−2.0562	0.0136	−0.1311	0.0135	−0.1232	+0.0087
Client05	CO2_Emission	BQML XGBoost	0.3733	0.3465	−3.0050	0.2170	−0.3531	0.1867	−0.0014	+0.1563

Client05	CO ₂ _per_pallet	BQML XGBoost	0.3922	0.3916	−177.5166	0.0776	−6.0042	0.0296	−0.0186	+0.3146
Client06	CO ₂ _Emission	BQML XGBoost	1.0190	0.5155	0.1426	1.1005	−0.0000	1.1167	−0.0298	−0.0815
Client06	CO ₂ _per_pallet	HistGradBoost	0.2422	0.1593	−0.0368	0.2437	−0.0497	0.2411	−0.0277	−0.0015

Table 3 shows that in five of the twelve combinations (Client03 for both targets, Client04 for CO₂_Emission, and Client06 for both targets) the best ML model actually beats the mean predictor. In the remaining seven combinations, ML modelling stays worse than naive_mean, sometimes substantially (for example, Client05 for CO₂_per_pallet: Δ RMSE = +0.3146). The median predictor remains a harder boundary to beat than the mean predictor—in none of the twelve combinations does the best ML model fall simultaneously below both baselines.

For the CO₂_per_pallet variable, prediction quality remains generally low, and R² values on the test set are usually strongly negative. The main results of the study are presented below. To show this variability simultaneously for all models, Table 4 presents the full test results for each customer and each model, together with two naïve benchmarks in the final rows.

Table 4. Per-customer detailed test results for CO₂_Emission (sklearn + BQML XGBoost) with naïve benchmarks.

Client	Model	RMSE (Test)	MAE (Test)	R ² (Test)
Client01	LinearRegression	39.4920	0.2745	−13986.93
Client01	ElasticNet	0.4060	0.0966	−0.4789
Client01	HistGradBoost	0.2726	0.1258	−9.1551
Client01	RandomForest	0.3043	0.0905	−9.7960
Client01	BQML XGBoost	0.2601	0.2543	−8.2528
Client01	Naive—mean	0.1194	0.1033	−0.9511
Client01	Naive—median	0.0855	0.0510	−0.0003
Client02	LinearRegression	4.4276	0.2756	−2.7905
Client02	ElasticNet	3.0181	0.2554	−0.7585
Client02	HistGradBoost	2.9970	0.2510	−0.7345
Client02	RandomForest	1.7898	0.3867	0.3867
Client02	BQML XGBoost	0.3553	0.3255	−2.4491
Client02	Naive—mean	0.2260	0.1694	−0.3960
Client02	Naive—median	0.2114	0.1020	−0.2216
Client03	LinearRegression	0.5471	0.3186	−12.0000
Client03	ElasticNet	0.4720	0.2929	−9.7000
Client03	HistGradBoost	0.4651	0.2756	−9.4123
Client03	RandomForest	0.5947	0.3210	−13.8300
Client03	BQML XGBoost	0.3777	0.3703	−9.8067
Client03	Naive—mean	0.4144	0.4055	−11.9943
Client03	Naive—median	0.1484	0.1112	−0.6663
Client04	LinearRegression	0.3446	0.1568	−4.36
Client04	ElasticNet	0.1380	0.0876	0.1409
Client04	HistGradBoost	0.1499	0.0930	0.1310
Client04	RandomForest	0.1496	0.0902	0.0900
Client04	BQML XGBoost	0.3316	0.3122	−3.96
Client04	Naive—mean	0.1529	0.0941	−0.0569
Client04	Naive—median	0.1684	0.0971	−0.2833
Client05	LinearRegression	1.3395	0.4112	−50.5700

Client05	ElasticNet	1.6681	0.4350	−39.2500
Client05	HistGradBoost	1.5128	0.4011	−18.6300
Client05	RandomForest	1.5261	0.4198	−15.4500
Client05	BQML XGBoost	0.3733	0.3465	−3.0050
Client05	Naive—mean	0.2170	0.1475	−0.3531
Client05	Naive—median	0.1867	0.0738	−0.0014
Client06	LinearRegression	1.0820	0.5465	0.0333
Client06	ElasticNet	1.2481	1.0743	0.5418
Client06	HistGradBoost	1.1176	1.0669	0.4589
Client06	RandomForest	1.1385	1.0481	0.4393
Client06	BQML XGBoost	1.0190	0.5155	0.1426
Client06	Naive—mean	1.1005	0.4611	−0.0000
Client06	Naive—median	1.1167	0.3376	−0.0298

For Client01, all models obtain strongly negative R^2 values on the test set, despite relatively small absolute errors. The lowest RMSE was achieved by the XGBoost model (RMSE = 0.2601; MAE = 0.2543; and R^2 = −8.25), and among the sklearn models by HistGradientBoostingRegressor (RMSE = 0.2726; MAE = 0.1258; and R^2 = −9.16). The extremely negative R^2 of linear regression (−13,987) also points here to a very small variance of emissions combined with non-zero prediction errors. The naïve median for this customer achieves RMSE = 0.0855—all tested ML models are at least three times worse than it.

For Client02 a model with a positive R^2 on the test set already appears. RandomForestRegressor achieved RMSE = 1.7898, MAE = 0.3867, and R^2 = 0.39, which means it explains about 39% of the variance of emissions. The ElasticNet and HistGradientBoosting models achieve lower RMSE (about 3.0) than linear regression (RMSE = 4.4276), but still with negative R^2 . XGBoost (RMSE = 0.3553; MAE = 0.3255; and R^2 = −2.45) does not match RandomForest in quality. The CV results (R^2 $\approx$ 0.13 for RF) confirm the relative stability of this model. The naïve median (RMSE = 0.2114), however, remains unattainable for all ML models—this qualifies as a result that is partly useful in the sense of explained variance, but not in the sense of absolute error.

For Client03 the results are mixed. On the test set all sklearn models have negative R^2 (for example, RandomForest: RMSE = 0.5947; R^2 = −13.83), while XGBoost (RMSE = 0.3777; R^2 = −9.81) does not improve the situation either. On the other hand, in 10-fold CV both HistGradientBoosting and RandomForest achieve high positive mean R^2 values (about 0.59 and 0.68 respectively), which suggests that the models can capture part of the emission structure in the training samples but have difficulty generalising to the final test set (probably because of the specifics of the test period or a change in the operations profile). This is one of the two customers for which the best ML model beats naïve_mean (Δ RMSE = −0.037 for XGBoost vs. naïve_mean RMSE = 0.4144).

For Client04, for CO₂_Emission, the following interesting contrast is observed: linear regression achieves the lowest RMSE (0.3446) and MAE (0.1568), but with a negative R^2 (−4.36), while the other models (ElasticNet, HGBR, and RandomForest) have a much higher RMSE (about 1.50) with slightly positive R^2 on the test set (0.09–0.13). In practice this means that these models better represent the variability of emissions (in the sense of variance), but at the cost of a higher absolute error. XGBoost (RMSE = 0.3316; MAE = 0.3122; and R^2 = −3.96) is intermediate—in terms of RMSE it approaches linear regression, but without a positive R^2 . ElasticNet achieves R^2 = 0.14—the only R^2 both above zero and below naïve_mean (RMSE 0.138 vs. 0.153)—which makes Client04 one of the two cases in which ML has a real advantage over the benchmark.

For Client05 all models stay at the level of negative R^2 on the test set (for example, linear regression: RMSE = 1.3395; R^2 = -50.57; ElasticNet: RMSE = 1.6681; R^2 = -39.25; RandomForest: RMSE = 1.5261; R^2 = -15.45; XGBoost: RMSE = 0.3733; and R^2 = -3.01). The CV R^2 values are only slightly above zero (up to about 0.21), which suggests that emissions for this customer are particularly difficult to predict from the available features (probably owing to high heterogeneity of order types and routes). The naïve median (RMSE = 0.1867) is unattainable here—all models are more than twice as bad.

Client06 is the most promising case. For the CO₂_Emission variable the models obtain clearly positive R^2 values on the test set:

- linear regression—RMSE = 1.0820; MAE = 0.5465; and R^2 = 0.0333;
- ElasticNet—RMSE = 1.2481; MAE = 1.0743; and R^2 = 0.5418;
- HistGradientBoosting—RMSE = 1.1176; MAE = 1.0669; and R^2 = 0.4589;
- RandomForest—RMSE = 1.1385; MAE = 1.0481; and R^2 = 0.4393;
- XGBoost—RMSE = 1.0190; MAE = 0.5155; and R^2 = 0.1426.

ElasticNet stands out in particular in terms of R^2 (over 0.54), while XGBoost provides the lowest RMSE. The positive R^2 values also hold in cross-validation (up to about 0.46 for RandomForest), which indicates a relatively regular emission structure for this customer. These results suggest that, for certain configurations of the delivery network (and order profile), the operational data of a 3PL operator may contain enough signal to enable practically useful emission modelling at the level of a single recipient. Importantly, this is the only customer for which the best ML model beats naïve_mean (Δ RMSE = -0.08 for XGBoost vs. 1.1005).

The overall picture from Table 3 is that only two of the six customers (Client04 and Client06) have the best ML model that is simultaneously lower in RMSE than naïve_mean and has $R^2 > 0$. These are the two segments in which investment in ML modelling brings measurable value.

Results for the CO₂_Emission variable are generally better than for emission per pallet, but still clearly differentiated between customers.

5. Discussion

Before discussing the findings in detail, Table 5 summarises the main empirical results of the study, the evidence supporting them and their implications, and maps them onto the two research questions and the three propositions formulated in Section 2.3.

Table 5. Summary of main findings.

Main Finding	Supporting Evidence	Implication (RQ/Proposition)
Global models do not beat naïve benchmarks for either target; R^2 is negative throughout	Tables A1–A3; ElasticNet best yet $R^2 < 0$ (-1.14/-3.62)	Routine data insufficient at network level (RQ1; supports P1)
More complex non-linear models do not improve, and often worsen, fit vs. ElasticNet	Tables A1 and A2 (HGBR, RF, XGBoost more negative R^2)	Algorithm choice secondary to predictor quality (RQ1; P1)
Strong heterogeneity of emission predictability across customers	Tables 3 and 4 (R^2 from -177.5 to +0.54)	Global models mask local signal (RQ1; supports P2)
Only Client04 and Client06 have a best model with	Table 3 (Δ RMSE < 0 and $R^2 > 0$)	ML pays off only in regular segments (RQ1/RQ2; P2–P3)

RMSE < naive_mean and R^2 >
0

Billing-derived data cannot represent driving dynamics, real route or dwell time	Negative global R^2 ; speed-proxy caveat (Section 3.2)	Emission-based learning needs richer data (RQ2; P3)
Emission prediction is a precondition for, not equivalent to, emission-based learning	Joint reading of global and per-customer results	Reporting $\neq$ learning; segment-specific models advised (RQ2; P3)

5.1. Findings and Implications for Further Research—Full Dataset

The analysis of the full dataset showed that, when predictors are restricted to the routine operational variables of the 3PL operator—mainly distance, the number of pallets, simple seasonality indicators and postal codes—none of the tested machine learning models obtained a positive coefficient of determination R^2 on the test set, either for CO₂_Emission or for CO₂_per_pallet. This means that in this configuration no algorithm outperformed the naïve benchmarks—the mean predictor and the median predictor. To reinforce this point, Tables A1–A3 compare the ML models with the naïve benchmarks as follows: the RMSE of ElasticNet is 2.1 times higher than the naïve median for CO₂_per_pallet and 1.5 times higher for CO₂_Emission; the tree models and XGBoost lose to the naïve median by factors of 1.7× to 7×. The negative R^2 values are therefore not a feature of error scale, but a real indicator of worse prediction than a constant. Put differently, the failure of the global models is direct evidence that billing-derived operational data, however complete and accurate for accounting and ESG-reporting purposes, are insufficient for accurate delivery-level emission prediction: the information that would actually explain emissions—instantaneous driving dynamics, the real executed route, the achieved load consolidation and on-site dwell time—is simply not present in the billing record. This is a key empirical confirmation of proposition P1. Such a result, although it may be disappointing from the perspective of predictive effectiveness, is in itself an important empirical conclusion.

In the comparison between algorithms, several clear patterns were obtained. First, the lowest RMSE and MAE values on the test set were consistently achieved by the ElasticNet model, that is, a moderately regularised linear regression. This suggests that, with the current set of predictors, the relationships between features and emissions are to a large extent (quasi-)linear, and that L1/L2 regularisation helps to reduce the influence of noise and feature collinearity. The use of L1/L2 regularisation for this purpose is consistent with observations in the literature [45]. In addition, the more complex non-linear models (HistGradientBoostingRegressor, RandomForestRegressor) and XGBoost did not bring the expected improvement in quality—in many cases they were characterised by higher RMSE values and a more negative R^2 than the linear models. This is consistent with the results of works showing that, under conditions of limited feature informativeness, increasing model complexity leads to overfitting on the training data rather than to an improvement in prediction quality [46,47].

From a methodological perspective, these results can be interpreted as a strong but valuable “negative result” (negative results associated with the use of ML are shown, among others, by Shmueli [9], who emphasises the high importance of low predictability in inference, and by Yarkoni and Westfall [10], who reveal that prediction often shows that variables have too little signal strength). The study shows that an attempt to predict CO₂ emissions at the level of individual last-mile operations, based only on basic operational variables, is insufficient to build a useful, generalising ML model. Emissions in medical and pharmaceutical distribution networks are shaped by a number of

additional factors—such as route microstructure, vehicle characteristics, traffic and weather conditions, time windows, the degree of load consolidation or temperature regimes—which remain outside the current set of predictors. The literature on emission estimation in road transport emphasises that accurate carbon footprint modelling requires data of high spatio-temporal resolution as well as technical information on vehicles and driving style (for example, models based on telematics, CAN-bus data and sensor measurements [48,49]); the results of this study empirically confirm this thesis in the context of 3PL last-mile operations.

From the point of view of the concept of emission-based learning in 3PL, the results obtained suggest that the passive collection of data on emissions and on the basic characteristics of orders alone is not enough to build a cognitive capability for predicting and controlling the carbon footprint. The full potential of machine learning can be exploited only when emission data are combined with richer sources of information (for example, telematics data, TMS/WMS systems, time-window records, weather data, and traffic-intensity information). The implication for further research is therefore the need to: (i) extend the range of predictors with contextual and technical information, (ii) consider multilevel models (such as hierarchical mixed models) combining the order, route and vehicle levels, and (iii) test hybrid architectures combining global models for the whole network with specialised local models (for example, per customer, per route type). In this sense, the negative results obtained do not testify to the ineffectiveness of machine learning as such but rather mark the limits of prediction based on data that are too sparse, and indicate the direction for developing emission analytics in last-mile logistics.

5.2. Findings and Implications for Further Research—Per-Customer Analysis

The analysis carried out separately for each customer revealed phenomena that were not visible in the global analysis, while at the same time showing the diversity of operational patterns in contract logistics. The most important conclusion is the high emission heterogeneity between customers—both in terms of the level of emissions and in terms of the predictability of those emissions. On the basis of the analysis carried out, three main insights can therefore be drawn.

The first insight is that (i) machine learning models obtain significantly different results depending on the customer, which indicates that operational patterns, order profiles and route configurations are strongly dependent on their specifics (even among customers from a single industry). For some recipients, such as Client01 and Client05, emissions remain largely unpredictable on the basis of the available features—all tested algorithms generate negative R^2 values, close to those observed in the global analysis. This means that the structure of their operations is characterised by high variability, irregular routes, a variable degree of consolidation or other contextual variables that are not reflected in the dataset. A different picture appears for customers Client02, Client04, and in particular Client06. For them, both linear models (ElasticNet) and non-linear tree models achieve moderately positive R^2 values, which testifies to the existence of more regular, repeatable emission patterns. It is notable that even simpler models—such as ElasticNet—can capture part of the structure of the variability, which suggests more stable relationships between features (such as distance, the number of pallets, and seasonality) and CO₂ emissions in these operational segments. This phenomenon is consistent with differences in order structures, distribution-network configurations and route repeatability between 3PL customers, where it has been shown that predictive models perform better in segments with greater operational regularity and a lower level of exogenous variability. Similar conclusions have already been drawn, among others, in the case of demand-volume prediction [31], as well as for various network configurations and mechanisms [50], and the results of this study confirm this thesis in the context of

transport emissions. This pattern provides direct empirical support for proposition P2: predictive validity is moderated by customer operational regularity, and the global model masks the signal that exists in the more regular segments.

The next issue is (ii) the relative advantage of the per-customer approach over the global model. Compared with the analysis of the full dataset, the per-customer approach in several cases leads to an improvement in prediction quality. The most pronounced benefits are observed in the CO₂_Emission models for customers with more stable logistics patterns—for example, Client02, Client04 and Client06—where higher (sometimes positive) R² values and lower RMSE were recorded than in the global models. This means that customer segmentation is not only operationally justified but also constitutes a promising analytical direction. A concrete managerial recommendation follows directly from this result: rather than pursuing a single universal emission model for the whole network, 3PL operators should build dedicated predictive models for their key customers with regular, repeatable supply structures (such as Client06, and to a lesser extent Client02 and Client04), and treat highly irregular customers (such as Client01 and Client05) as cases where with the data currently available, emissions can realistically only be reported rather than predicted and optimised. In practice this means routing analytical investment toward the segments where it is demonstrably effective, instead of diluting it across a one-size-fits-all system.

The results across customers confirm the observation from the global analysis, namely that (iii) the choice of algorithm is of secondary importance relative to the quality of the predictors. Both the tree models (RandomForest, HistGradientBoosting) and XGBoost proved competitive only in selected cases—primarily in segments characterised by greater regularity (for example, Client06). In the remaining cases the non-linear models did not outperform the linear models, and sometimes achieved worse R² values, which suggests overfitting on a set of predictors that is too poorly informative. Importantly, XGBoost in many cases achieved a relatively low RMSE but still a negative R², which is a signal that the model minimises absolute errors well but cannot represent the variability of emissions—similar conclusions follow, among others, from the work of Singh and Dubey [51] and Yin et al. [23].

Taken together, these results allow the contributions of the study to be stated more sharply than in the introduction. First, the paper provides rare empirical evidence that routine, billing-level 3PL data are insufficient for delivery-level emission prediction in medical and pharmaceutical last-mile logistics—a negative but actionable result that delimits where sustainability intelligence can currently be built. Second, by contrasting a global model with six customer-level models, it shows that emission predictability is a segment-specific property and that aggregate analyses can hide usable local signal; this reframes the design choice from “which algorithm” to “which customer, with which data”. Third, through the framework of Section 2.3, it reframes emission prediction as a precondition for, rather than a synonym of, emission-based learning, and offers a concrete, testable decision rule (predict-and-optimise for regular segments, report-only for irregular ones) for 3PL sustainability practice.

5.3. Research Limitations

The study has several limitations that must be taken into account when interpreting the results. One of the most important limitations was the available computing power, and for this reason the model-fitting procedure was simplified as follows: preliminary hyperparameter tuning was carried out on a random sample of 200 thousand observations, cross-validation on a sample of 50 thousand observations, and grid search was limited to three folds. Although such an approach is consistent with practices used in research based on large datasets, it may lead to an underestimation of the optimal

hyperparameters of the non-linear models and to an understatement of their potential predictive capabilities.

The analysis covered one 3PL operator and six transport recipients, which limits the possibility of generalisation. The results are representative of the given distribution network but need not carry over to other sectors of medical logistics or to 3PL operations with different route and fleet structures. All predictive models were based on a single-step point-prediction architecture. The study did not include more advanced sequential or geospatial methods (such as LSTM, GNN, and Transformer-based spatio-temporal models), which in the literature frequently show higher effectiveness in tasks with a spatio-temporal structure. The emission data were declarative in nature and came from the operator's billing systems, which means they may contain measurement errors and simplifications related to the calculation model. The lack of independent verification (for example, from onboard telematics) limits the ability to assess the level of systematic error.

6. Conclusions

The aim of the study was to verify empirically to what extent machine learning models can predict CO₂ emissions in last-mile transport in medical logistics (RQ1) and whether emission prediction can constitute a basis for developing cognitive capabilities in the area of sustainability for a 3PL operator (RQ2). The results of the analyses lead to several key conclusions.

With regard to RQ1, the study clearly shows that, with the currently available operational predictors—comprising distance, the number of pallets, basic calendar features and encoded postal codes—ML models are not able to effectively predict CO₂ emissions at the level of a single transport operation. None of the tested models obtained a positive coefficient of determination R^2 on the test set in the global analysis, and the best results (ElasticNet) only minimised the absolute error without, however improving the explanatory ability relative to the naïve model. This indicates that last-mile emissions are determined by factors that remain outside the current set of features—such as road conditions, vehicle characteristics, driving style, delivery time windows, the degree of load consolidation or route microstructure. At the same time, the per-customer analysis revealed that prediction effectiveness is strongly differentiated and depends on the specifics of the recipient. In selected cases, especially for customers with more regular order and route schemes (for example, Client06, and to a lesser extent Client02 and Client04), the models obtained positive R^2 on the test set. This confirms the existence of local structures that can be captured by ML, which suggests that global models may mask the signal present in separated operational segments. Consequently, emission prediction in 3PL should be designed taking into account the segmentation of customers and routes, rather than as one universal solution for the whole network.

With regard to RQ2, the results indicate that emission prediction can be an element of building cognitive capabilities in the area of sustainability only when it is based on richer and more contextual data than those present in the analysed system. The study proves that the passive collection of operational data is not enough to create models enabling organisational learning or decision-making that reduces the carbon footprint. At the same time, the results provide the following important guidance: where logistics profiles are stable and repeatable, the possibility of genuine learning from emissions appears, that is, of using prediction to steer operations. For this to be possible on a wider scale, however, a 3PL operator must integrate emission data with telematics information, traffic conditions, vehicle parameters, weather data and information from IT systems.

Author Contributions: Conceptualization, M.K. (Marzena Kramarz) and M.K. (Mariusz Kmiecik); methodology, M.K. (Marzena Kramarz) and M.K. (Mariusz Kmiecik); software, M.K. (Marzena

Kramarz) and M.K. (Mariusz Kmiecik); validation, M.K. (Marzena Kramarz) and M.K. (Mariusz Kmiecik); formal analysis, M.K. (Marzena Kramarz) and M.K. (Mariusz Kmiecik); investigation, M.K. (Marzena Kramarz) and M.K. (Mariusz Kmiecik); resources, M.K. (Marzena Kramarz) and M.K. (Mariusz Kmiecik); data curation, M.K. (Marzena Kramarz) and M.K. (Mariusz Kmiecik); writing—original draft preparation, M.K. (Marzena Kramarz) and M.K. (Mariusz Kmiecik); writing—review and editing, M.K. (Marzena Kramarz) and M.K. (Mariusz Kmiecik); visualization, M.K. (Marzena Kramarz) and M.K. (Mariusz Kmiecik); supervision, M.K. (Marzena Kramarz) and M.K. (Mariusz Kmiecik); project administration, M.K. (Marzena Kramarz) and M.K. (Mariusz Kmiecik); funding acquisition, M.K. (Marzena Kramarz) and M.K. (Mariusz Kmiecik). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by European Funding for Silesia Region 2021-2027 grant number FESL.10.25-IZ.01-07G5/23-004.

Data Availability Statement: Authors are not permitted to share the data.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Dependent variable: CO₂ per pallet (models from sklearn)—with naïve benchmark.

Model	RMSE (Test)	MAE (Test)	R ² (Test)	RMSE CV (Mean)	RMSE CV (Std)	MAE CV (Mean)	MAE CV (Std)	R ² CV (Mean)	R ² CV (Std)	Best Params (GS)
LinearRegression	0.2864	0.1390	−7.6596	0.8554	1.0208	0.1133	0.0122	−0.9542	1.5597	-
ElasticNet	0.2093	0.1277	−3.6234	0.8036	1.0511	0.1085	0.0136	−0.1006	0.1994	$\alpha = 0.001$, l1_ratio = 0.1
HistGradientBoostingRegressor	0.4896	0.0938	−24.3028	0.8055	1.0555	0.1003	0.0121	−0.0683	0.1206	depth = 10, iter = 200, lr = 0.05, l2 = 0.1
RandomForestRegressor	0.6781	0.1190	−47.5407	0.8453	1.0303	0.0829	0.0150	−0.6351	1.1524	n_est = 400, depth = None, sqrt, leaf = 1
Naive—mean(y_train)	0.1210	0.0828	−0.5458	—	—	—	—	—	—	$\hat{y} = 0.1019$
Naive—median(y_train)	0.0977	0.0319	−0.0073	—	—	—	—	—	—	$\hat{y} = 0.0383$

Source: own work.

Table A2. Dependent variable: CO₂ emission (models from sklearn)—with naïve benchmark.

Model	RMSE (Test)	MAE (Test)	R ² (Test)	RMSE CV (Mean)	RMSE CV (Std)	MAE CV (Mean)	MAE CV (Std)	R ² CV (Mean)	R ² CV (Std)	Best Params (GS)
LinearRegression	0.4159	0.2323	−2.2124	1.2475	0.8931	0.2086	0.0137	−0.4753	1.2425	-
ElasticNet	0.3395	0.2180	−1.1404	1.2233	0.9249	0.2026	0.0112	0.0166	0.0662	$\alpha = 0.001$, l1_ratio = 0.1
HistGradientBoostingRegressor	0.7577	0.1940	−9.6598	1.2228	0.9307	0.1875	0.0109	0.0172	0.1074	depth = 10, iter = 200, lr = 0.05, l2 = 0.0

RandomForest Regressor	0.7934	0.2313	−10.6889	1.1680	0.9311	0.1620	0.0136	0.0212	0.4251	n_est = 400, depth = None, sqrt, leaf = 1
Naive— mean(y_train)	0.2634	0.1581	−0.2883	—	—	—	—	—	—	$\hat{y} = 0.2060$
Naive— median(y_train)	0.2321	0.0736	−0.0009	—	—	—	—	—	—	$\hat{y} = 0.0883$

Source: own work.

Table A3. Dependent variable: CO₂ per pallet and CO₂ emission (XGBoost from BigQuery ML)—with naïve benchmark.

Variable	Model	RMSE (Test)	MAE (Test)	R ² (Test)	Note
CO ₂ _per_pallet	BQML XGBoost	0.3916	0.3841	−15.1848	XGBoost did not improve over sklearn; strongly negative R ² indicates worse than naïve baseline.
CO ₂ _Emission	BQML XGBoost	0.4018	0.3550	−1.9982	RMSE comparable to LinearRegression; R ² still below 0, i.e., worse than mean predictor.
CO ₂ _per_pallet	Naive— mean(y_train)	0.1210	0.0828	−0.5458	$\hat{y} = 0.1019$
CO ₂ _per_pallet	Naive— median(y_train)	0.0977	0.0319	−0.0073	$\hat{y} = 0.0383$
CO ₂ _Emission	Naive— mean(y_train)	0.2634	0.1581	−0.2883	$\hat{y} = 0.2060$
CO ₂ _Emission	Naive— median(y_train)	0.2321	0.0736	−0.0009	$\hat{y} = 0.0883$

Source: own work.

References

- Ha, N.T.; Akbari, M.; Au, B. Last mile delivery in logistics and supply chain management: A bibliometric analysis and future directions. *Benchmarking Int. J.* **2023**, *30*, 1137–1170.
- Aleidi, F.; Alomair, S.; Alharbi, H.; AlSalamah, N. Stability of refrigerated medications at room temperature: Implications for transport, delivery, and patient safety. *Cureus* **2025**, *17*, e93213. <https://doi.org/10.7759/cureus.93213>.
- Adusei, S.; Nuertey, D.; Poku, E. Last-mile distribution and commodity availability, security and access: The moderating role of supply chain integration. *Benchmarking Int. J.* **2025**, *32*, 369–397.
- Olkiewicz, M.; Dyczkowska, J. Green logistics at selected logistics operators in Poland. *Sustainability* **2025**, *17*, 4587. <https://doi.org/10.3390/su17104587>.
- Mogire, E.; Kilbourn, P.; Luke, R. Electric vehicles in last-mile delivery: A bibliometric review. *World Electr. Veh. J.* **2025**, *16*, 52. <https://doi.org/10.3390/wevj16010052>.
- de la Torre, R.; Corlu, C.G.; Faulin, J.; Onggo, B.S.; Juan, A.A. Simulation, optimization, and machine learning in sustainable transportation systems: Models and applications. *Sustainability* **2021**, *13*, 1551.
- Makanadar, A.; Shahane, S. Urban mobility: Leveraging AI, machine learning, and data analytics for smart transportation planning—A case study on New York City. In *Proceedings of the 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*; IEEE: New York, NY, USA, 2023; pp. 1–6.
- Akbari, M.; Do, T.N.A. A systematic review of machine learning in logistics and supply chain management: Current trends and future directions. *Benchmarking Int. J.* **2021**, *28*, 2977–3005. <https://doi.org/10.1108/BIJ-10-2020-0514>.
- Shmueli, G. To explain or to predict? *Stat. Sci.* **2010**, *25*, 289–310.
- Yarkoni, T.; Westfall, J. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspect. Psychol. Sci.* **2017**, *12*, 1100–1122.

11. Brown, J.; Guiffrida, A. Carbon emission comparison of last-mile delivery versus customer pickup. *Int. J. Logist. Res. Appl.* **2014**, *17*, 503–521. <https://doi.org/10.1080/13675567.2014.907397>.
12. Miko, N.U.; Abbas, U. Determinants of efficient last-mile delivery: Evidence from health facilities and Kaduna Health Supplies Management Agency. *J. Humanit. Logist. Supply Chain Manag.* **2024**, *14*, 4–16.
13. World Health Organization. *Routine Health Information System and Health Facility and Community Data for Neglected Tropical Diseases: Last-Mile Logistics Information System for Medicines and Health Products*; WHO: Geneva, Switzerland, 2025.
14. Reproductive Health Supplies Coalition. *Summary of Last Mile Distribution Models for Health Commodities Discussion*; International Association of Public Health Logisticians (IAPHL): Arlington, VA, USA, 2014.
15. Bhatnagar, A.; Gupta, V.; Tandon, P.; Saksena, T.; Ranjan, A.; Gandh, P.; Garcha, S.; Kapoor, A. Last mile delivery of cold chain medicines—Challenges and recommendations. *Indian J. Pharm. Biol. Res.* **2018**, *6*, 34–41. <https://doi.org/10.30750/ijpbr.6.1.6>.
16. L.E.K. Consulting. *The True Value of Last-Mile Logistics in Healthcare*; L.E.K. Consulting: London, UK, 2025.
17. Kiba-Janiak, M.; Marcinkowski, J.; Jagoda, A.; Skowrońska, A. Sustainable last mile delivery on e-commerce market in cities from the perspective of various stakeholders. Literature review. *Sustain. Cities Soc.* **2021**, *71*, 102984. <https://doi.org/10.1016/j.scs.2021.102984>.
18. Westley, E.; Moro Martin, L. *Driving Last-Mile Solutions to Ensure Access to Public Health Commodities*; USAID Global Health Supply Chain Program—Procurement and Supply Management (GHSC-PSM): Arlington, VA, USA, 2022.
19. Daultani, Y.; Chan, F.T.; Pratap, S.; Goswami, M. Modeling drone-enabled last-mile blood delivery systems for emergency healthcare. *Int. J. Logist. Res. Appl.* **2025**, 1–22.
20. Fritz, J.; Herrick, T.; Gilbert, S.S. Estimation of health impact from digitalizing last-mile logistics management information systems (LMIS) in Ethiopia, Tanzania, and Mozambique: A Lives Saved Tool (LiST) model analysis. *PLoS ONE* **2021**, *16*, e0258354.
21. Kumar, V.; Pallathadka, H.; Sharma, S.K.; Thakar, C.M.; Singh, M.; Pallathadka, L.K. Role of machine learning in green supply chain management and operations management. *Mater. Today Proc.* **2022**, *51*, 2485–2489.
22. Zhu, J.; Wu, Y.; Liu, Z.; Costa, C. Sustainable optimization in supply chain management using machine learning. *Int. J. Manag. Sci. Res.* **2025**, *8*, 1–8.
23. Yin, C.; Wu, J.; Sun, X.; Meng, Z.; Lee, C. Road transportation emission prediction and policy formulation: Machine learning model analysis. *Transp. Res. Part D Transp. Environ.* **2024**, *135*, 104390.
24. Zhao, L.; He, Z. Using machine learning to analyze the factors influencing city-level transportation carbon emissions. *Energy* **2025**, *333*, 137355.
25. Natarajan, Y.; Wadhwa, G.; Sri Preethaa, K.R.; Paul, A. Forecasting carbon dioxide emissions of light-duty vehicles with different machine learning algorithms. *Electronics* **2023**, *12*, 2288.
26. Karami, Z.; Kashef, R. Smart transportation planning: Data, models, and algorithms. *Transp. Eng.* **2020**, *2*, 100013.
27. Vadlamani, S.; Modashiya, M. Improving transportation planning using machine learning. In *Encyclopedia of Data Science and Machine Learning*; IGI Global: Hershey, PA, USA, 2023; pp. 3076–3088.
28. Plakandaras, V.; Papadimitriou, T.; Gogas, P. Forecasting transportation demand for the US market. *Transp. Res. Part A Policy Pract.* **2019**, *126*, 195–214.
29. Lechtenberg, S.; Hellingrath, B. Guiding practitioners of road freight transport to implement machine learning for operational planning tasks. *Transp. Res. Procedia* **2025**, *82*, 1839–1857.
30. Gabellini, M.; Calabrese, F.; Galizia, F.G.; Ronchi, M.; Regattieri, A. An information-sharing and cost-aware custom loss machine learning framework for 3PL supply chain forecasting. *Comput. Ind. Eng.* **2025**, 111573.
31. Kmiecik, M. Supply and demand prediction by 3PL for assortment planning. *Manag. Sci. Lett.* **2025**, *15*, 97–112.
32. Toivonen, K. Forecasting Demand in 3PL Warehouse. Master's Thesis, Aalto University, Espoo, Finland, 2024.
33. Liu, X.; Bo, G. Q-learning algorithm for fourth party logistics route optimization considering tardiness risk. In *Proceedings of the 2022 International Conference on Cyber-Physical Social Intelligence (ICCSI)*; IEEE: New York, NY, USA, 2022; pp. 268–273.
34. Oršič, J.; Jereb, B.; Obrecht, M. Sustainable operations of last mile logistics based on machine learning processes. *Processes* **2022**, *10*, 2524.
35. Benjatanont, S.; Tantuico, D. *Application of Linear Models, Random Forest, and Gradient Boosting Methods to Identify Key Factors and Predict Truck Dwell Time for a Global 3PL Company*; Massachusetts Institute of Technology: Cambridge, MA, USA, 2020.
36. Ateş, O. Determining the appropriate transportation option with the decision tree. *Int. J. New Find. Eng. Sci. Technol.* **2023**, *1*, 76–82.

37. Sahoo, R.; Pasayat, A.K.; Bhowmick, B.; Fernandes, K.; Tiwari, M.K. A hybrid ensemble learning-based prediction model to minimise delay in air cargo transport using bagging and stacking. *Int. J. Prod. Res.* **2022**, *60*, 644–660.
38. Nsir, N.; Alzubi, A.B.; Adegboye, O.R. Enhancing sustainable supply chain performance prediction using an augmented algorithm-optimized XGBoost in Industry 4.0 contexts. *Sustainability* **2025**, *17*, 10344.
39. Tong, C. Optimizing multi-tier supply chain ordering with LNN+XGBoost: Mitigating the bullwhip effect. *arXiv* **2025**, arXiv:2507.21383.
40. Dabiri, S.; Heaslip, K. Inferring transportation modes from GPS trajectories using a convolutional neural network. *Transp. Res. Part C Emerg. Technol.* **2018**, *86*, 360–371.
41. Erramaline, A.; Badard, T.; Cote, M.P.; Duchesne, T.; Mercier, O. Identification of road network intersection types from vehicle telemetry data using a convolutional neural network. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 475.
42. Nguyen, H.D.; Tran, K.P.; Thomassey, S.; Hamad, M. Forecasting and anomaly detection approaches using LSTM and LSTM autoencoder techniques with the applications in supply chain management. *Int. J. Inf. Manag.* **2021**, *57*, 102282.
43. Wu, X.; Huang, Y.; Bao, W. Using GNN to identify supply chain network structure. *Int. J. Big Data Intell. Technol.* **2025**, *6*, 46–55.
44. Zhu, X.; Ninh, A.; Zhao, H.; Liu, Z. Demand forecasting with supply-chain information and machine learning: Evidence in the pharmaceutical industry. *Prod. Oper. Manag.* **2021**, *30*, 3231–3252. <https://doi.org/10.1111/poms.13426>.
45. Kar, S.; Mohanty, M.K.; Thakurta, P.K.G. Efficient healthcare supply chain: A prioritized multi-task learning approach with task-specific regularization. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108249.
46. Maulud, D.; Abdulazeez, A.M. A review on linear regression comprehensive in machine learning. *J. Appl. Sci. Technol. Trends* **2020**, *1*, 140–147.
47. Giudici, P.; Raffinetti, E.; Riani, M. Robust machine learning models: Linear and nonlinear. *Int. J. Data Sci. Anal.* **2025**, *20*, 1043–1050.
48. Faris, W.F.; Rakha, H.A.; Kafafy, R.I.; Idres, M.; Elmoselhy, S. Vehicle fuel consumption and emission modelling: An in-depth literature review. *Int. J. Veh. Syst. Model. Test.* **2011**, *6*, 318–395.
49. Mądziel, M. Instantaneous CO₂ emission modelling for a Euro 6 start-stop vehicle based on portable emission measurement system data and artificial intelligence methods. *Environ. Sci. Pollut. Res.* **2024**, *31*, 6944–6959.
50. Kramarz, M.; Kmiecik, M. Configuration of sustainable distribution networks as a determinant of logistics coordination mechanism selection. *Sustainability* **2025**, *17*, 7994.
51. Singh, M.; Dubey, R.K. Deep learning model based CO₂ emissions prediction using vehicle telematics sensors data. *IEEE Trans. Intell. Veh.* **2021**, *8*, 768–777.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.