National College of Ireland

A Data-Driven Framework for Estimating Hard-to-Measure Value-Chain Emissions

MSc Research Project
Master of Science in AI for Business

Lavanya Budhraja
Student ID: 23300540

School of Computing
National College of Ireland

Supervisor: Charlyn Rosales

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Lavanya Budhraja
Student ID:	23300540
Programme:	MSc in Artificial Intelligence for Business **Year:** 2025
Module:	Practicum
Supervisor:	Charlyn Rosales
Submission Due Date:	28 January 2026
Project Title:	A Data-Driven Framework for Estimating Hard-to-Measure Value-Chain Emissions
Word Count:	5293 **Page Count**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Lavanya Budhraja
Date:	28 January 2026

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Data-Driven Framework for Estimating Hard-to-Measure Value-Chain Emissions

Lavanya Budhraja
23300540

Abstract

The precise measurement of corporate carbon footprints, especially indirect Scope 3 emissions, has become one of the urgent issues facing business enterprises with the intention of addressing their sustainability goals. The existing conventional carbon accounting approaches are associated with the problem of information fragmentation and enormous truncation errors, which restrict the possibilities of the organizations to state the overall effects on the environment. This study fills these gaps by creating a machine-enhanced automated and data-driven structure which can predict emissions using primary operation-related data. The research was carried out through a strict and multistage analytical pipeline made to suit the complexities of the environmental data. To look at the technical feasibility of the proposed solution, an artificial sample of 23,808 corporate profiles was used. This domain-specific imputation was used to determine the quality of data, K-Means clustering (k=4) to realize strategic segmentation, and a set of ensemble predictive models to obtain quantitative emission. The experimental results showed that the XGBoost regressor model performed best in the prediction of aggregate emissions with an R^2 of 0.52, but a Multi-Output Random Forest model performed better in the direct Scope 1 emissions (R^2 of 0.54) than in the stochastic Scope 3 emissions (R^2 of 0.47). The structural analysis revealed existence of specific operational clusters, which signified there was need to have sector-specific baseline so as to have proper reporting. The paper presents a scalable and replicable artefact that demonstrates the viability of algorithmic carbon accounting and provides a route towards reducing the burden of reporting on SMEs and finding the necessary factors which drive corporate carbon intensity.

Keywords: Algorithmic Carbon Accounting, XGBoost Regressor, Multi-Output Random Forest, K-Means Clustering, SME sustainability

1 Introduction

1.1 Background

The growing sense of climate crisis and the increasing number of regulations have fueled the demand to adopt a transparent and accurate corporate carbon accounting both in the European Union in the form of the Corporate Sustainability Reporting Directive (CSRD) and in the U.S. in the form of the climate disclosure rules of the U.S. Securities and Exchange Commission (Adamolekun and Kyiu, 2025) . Although the basic guidelines such as the Greenhouse Gas (GHG) Protocol have unified the assessments of emissions, there are still severe obstacles especially on the limits of the system and quality of data (Matthews, Hendrickson and Weber, 2008). The latter issues are especially experienced by medium-sized enterprises (SMEs), which

frequently do not have specialized resources that would enable them to conduct extensive emissions inventorying (Harangozo and Szigeti, 2017).

The history of carbon accounting is the development of voluntary disclosure to a compulsory one, but the methodologies of calculating the indirect emissions have been abandoned Scope 3 (Csutora and Harangozo, 2017). According to the research by Nuță et al. (2021), carbon accounting schemes on national and corporate levels must change accordingly to be able to meet the effects of urbanization. The largest part of the carbon footprint of the company is often the scope 3 emissions i.e. the overall value chain, which is unfortunately, difficult to measure due to the fragmentation of the data, the lack of visibility of the suppliers, and the excessive reliance on manual, error-prone, methods of estimations (Zvezdov and Hack, 2016).

1.2 Problem statement and motivation

The main issue that will consider this study is the operational bottleneck of traditional carbon accounting: the use of manual and bottom-up data collection on Scope 3 emissions cannot be scaled and is highly prone to large losses of truncation errors (Matthews, Hendrickson and Weber, 2008). The SMEs have a two-fold problem to contend with, they are being pressured to report their emissions to both stakeholders and the regulators, but in many instances, they have to do so with fragmented information and disparate systems.

The reason behind this research is the necessity to close a gap between the availability of data and the necessity to report. The market-valuation risk of the difference between reported amounts and the amount measurable by companies generates a delta, where investors rate companies with opaque carbon profiles more down, as they grow (Adamolekun and Kyiu, 2025). Thus, it is not only an environmental requirement but a business necessity to come up with an automated and data-driven system.

1.3 Research question and objectives

How can structured operational data from medium-sized companies be analyzed to achieve accurate quantification, segmentation, and forecasting of carbon emissions across Scope 1, Scope 2, and Scope 3?

Research Objectives:

- To utilize the K-Means clustering method for the segmentation of operational data profiles to identify distinct corporate emission archetypes.
- To apply regression models such as Random Forest and Gradient Boosting for prediction, thereby providing actionable, scope-specific sustainability insights for both current and future datasets.

1.4 Contribution and structure

The study proposes an automated framework that can be replicated and can be used to predict corporate carbon footprints through the combination of domain-informed data imputation and

ensemble machine learning. The study offers a first-pass screening tool, which reduces the entry barrier to Scope 3 reporting substantially, by testing this method on a sample of 23,808 corporate profiles.

2 Related work

2.1 The data quality crisis in Scope 3 accounting

Switching of corporate carbon accounting towards a regulated compliance rather than voluntary disclosure has highlighted significant methodological issues and this is particularly in terms of Scope 3 reporting. Nguyen et al. (2023) recognize the lack of data and its inconsistency as fundamental impediments, whereas Matthews, Hendrickson and Weber (2008) explain underreporting by the boundary truncation error, when the process of tracing the supply chain is too complicated and expensive. This indicates a problem that is even of a bigger data quality crisis than a regulatory lapse. Harangozo and Szigeti (2017) also state that the use of static and aggregated financial data distorts actual emission profiles, which minimizes validity and accuracy. Product level measurements are not standardized, which is another hindrance to cross-company benchmarking (Russell, 2019). Therefore, it is becoming impossible to conduct traditional manual Life Cycle Assessments (LCA) to large product portfolios, and the gap between the capabilities of ERP systems and the requirement of modern environmental reporting is also acute (Gao, H. et al., 2023).

2.2 Methodological stagnation vs. hybrid innovation

The classical estimating methods have been alternating between very thorough evaluations of the process and general Economic Input-Output (EIO) models. Nevertheless, both methods have to cope with practical obstacles in modern reporting. A recent literature review indicates that although the bottom-up models are very accurate, they are also extremely costly, at the same time the EIO methods are suffering from the disadvantage of being too aggregated (Schmidt et al., 2022). To deal with these shortcomings, hybrid 'spend-based' techniques have come up still, even these are unable to correctly represent the industry's non-linear relationship between expenditure and carbon intensity (Csutora and Harangozo, 2017). Current studies are turning more towards automated data-driven systems that allow for predictive insights rather than post hoc auditing. This change is further supported by the results of the comparisons which show that the variations in the emission profiles of different sectors are significant and that the static emission coefficients do not capture this well (Russell, 2019).

2.3 The machine learning paradigm shift

The agreement among the researchers in this field is that the use of Machine Learning (ML) techniques is the best way to go compared to the traditional econometric models as represented in Table 1. In their paper, Gao, H. et al., (2023) prove that the use of non-parametric algorithms such as Random Forest and Gradient Boosting is better than OLS regression in revealing heteroscedasticity in the carbon emission datasets. The advantage is explained by the fact that

the analysis of very complex interdependencies between such variables as fleet size, revenue, and energy consumption, which are largely ignored by linear models due to oversimplification, is possible with ML (Nguyen, Diaz-Rainey and Kuruppuarachchi, 2021; Gao, P. et al., 2023). It can be substantiated by Meng and Noman (2022), who have shown that AI-based models can decrease the error rate in footprint prediction significantly in comparison to the conventional statistical techniques.

In the case of Scope 3 emissions, ensemble methods have managed to produce dependable forecasts even when there is missing data, on the condition that the variety of the feature set is sufficient (Serafeim and Caicedo, 2022). In addition, further studies reveal that ML models are always less affected by the presence of outliers, which are usually a significant problem in the data of corporate emissions (Xu et al., 2025).

Table 1: Summary of machine learning models and reported accuracies in corporate and industrial carbon forecasting literature

Title	Authors	Machine Learning Model(s) Used	Accuracy Reported
Machine-Learning-Based Carbon Footprint Management in the Frozen Vegetable Processing Industry	Magdalena Scherer, Piotr Milczarski	Unsupervised: Canopy, k-means, EM; Supervised: k-NN, MLP, C4.5, RF, SVM	Classification accuracy for clusters up to 100% (RF with k-means/canopy)
Machine Learning Applications for Carbon Emission Estimation (Systematic Review)	Hala Salem Al Nuaimi, Adolf Acquaye, Ahmad Mayyas	Survey: SVM, ANN, LSTM, RF, XGBoost, LightGBM, CatBoost	Sectoral summary: e.g. MAPE <5% in buildings/energy (review, sector-by-sector)
A Review of Building Carbon Emission Accounting and Prediction Models	Huan Gao et al.	ANN, SVM, regression, hybrid/optimization (e.g., GA-BP, FOA-LSSVM)	BP, GA-BP, SVM, LMDI models cited with some MAPE <10% (varies)
Machine Learning Models for Prediction of Scope 3 Carbon Emissions	George Serafeim, Gladys Vélez Caicedo	AdaBoost, Random Forest, k-NN, OLS, Gamma-GLM	AdaBoost MAPE: 6.3–15%, R^2 up to 0.92
Scope 3 emissions: Data quality and machine learning prediction accuracy	Quyen Nguyen et al.	Linear Forest, XGBoost, RF, OLS, ElasticNet	Aggregated Scope 3: MDAPE 72%; best category 60%

Title	Authors	Machine Learning Model(s) Used	Accuracy Reported
An approach for analyzing urban carbon emissions using machine learning models	Peidao Gao et al.	LightGBM, LASSO	R^2: 0.97 (validation), 0.87 (test)
Forecasting carbon dioxide emissions: Two-stage ML procedure	Chunzi Wang, Moye Li, Junpeng Yan	SVR, RF, Ridge, ANN; Two-stage (SVR→ANN)	Hybrid improves RMSE/MAE by 14–43% over single-stage
Predicting-corporate-carbon-footprints	Adam Fishman et al.	XGBoost, Ridge	XGBoost: R^2 up to 0.90
Short-Term Forecasting of CO2 Emission Intensity in Power Grids by Machine Learning	Kenneth Leerbeck et al.	Linear Regression, LASSO, ARIMA/ARIMAX, Softmax Ensemble	RMSE for 6hr avg=37.6, marginal=9.6 $kgCO_2/MWh$
A Machine Learning Approach for Forecasting Environmental Impact of Building Sector	Spyros Giannelos et al.	Linear Regression, ARIMA, Shallow/Deep NN	World: SH-NN MAPE=2.38%; EU: ARIMA/NN MAPE 4.36–4.45%
Predicting-CO2-Emission-Footprint	Zulaikha Akhter et al.	ANN, SVM Regression, RF	ANN: R^2 = 0.95 (train), 0.92 (test)
Who predicts better? ML vs. econometrics for ship emissions	Yoonjin Choi et al.	XGBoost, RF, Elastic Net, OLS	XGBoost: RMSE (test) 0.164, R^2=0.87
A multi-stage method to predict carbon dioxide emissions using dimensionality reduction, clustering, and machine learning techniques	Abbas Mardani, Huchang Liao, Mehrbakhsh Nilashi, Melfi Alrasheedi, Fausto Cavallaro	SVD (dimensionality reduction), SOM (clustering), ANN, ANFIS	SVD-SOM-ANFIS: MAE=0.065, SVD-SOM-ANN: MAE=0.104, single ANFIS: MAE=0.352, MLR (baseline): MAE=0.522
A Data-Driven Machine Learning Framework for Scope 3 Emission Estimation	Lavanya Budhraja	Random Forest (RF) Regressor, XGBoost Regressor	XGBoost: R^2 = 0.5168, RMSE = 848,306 tCO_2e, MAE = 97,864 tCO_2e (test set, total emissions) Random Forest: R^2 = 0.4790, RMSE = 880,906 tCO_2e, MAE = 95,257 tCO_2e (test set, total emissions) Random Forest Scope 1: R^2 = 0.5449 (best for direct emissions) Median error on audited records: −0.2 tCO_2e, IQR: −232 to +259 tCO_2e (audit subset)

2.4 Structural segmentation and feature engineering

One of the main difficulties that comes up during carbon prediction is the differing nature of companies, which is very important yet often neglected. Different related studies indicate that the application of the same forecasting model on different industry data leads to poor generalization performance. Hence, a multi-stage framework involving unsupervised clustering methods like K-Means is proposed to first divide entities into similar clusters and then regression models are applied (Schmidt et al., 2022; Csutora and Harangozo, 2017). As an example, by using comparable clustering procedures, Scherer and Milczarski (2021) have applied the food processing sector to handle the variability in feedstock, which proves the multi-stage method. This method permits the identification of emission drivers per sector, thus guaranteeing service-based companies are not subjected to the same assessments as the heavy industry manufacturers (Russell, 2019). In addition to this, using dynamic features has been proved to have a great impact on increasing the prediction power of the model, especially when it comes to using the data of temporal energy consumption (Gao, H. et al., 2023).

2.5 Strategic value and financial materiality

The new carbon accounting literature highlights that the practice of carbon accounting is shifting from a mere technical necessity to a source of business value with the help of the carbon accounting strategy. Adamolekun and Kyiu (2025) offer arguments in favor of this, claiming that the market value of the company goes up with the amount of carbon disclosed, because investors cut off companies that don't report environmentally. This increases the demand for carbon measurement models to be accurate in their predictions. Besides, Bocken and Allwood (2012) assert that carbon accounting's end goal is to change the stakeholders' decisions hence, the models must be of high accuracy and interpretability at the same time. Among the techniques is the Random Forest that is mainly suited due to the accuracy and understanding of the features.

3 Research methodology

3.1 Research framework and process model

The methodology used in the research was designed to be scientifically reproducible and rigorous and made use of a bespoke iterative lifecycle adapted to the demands of environmental data. This systematic approach ensured that the progression from raw data acquisition to model evaluation remained aligned with the core business objective of automating Scope 3 carbon accounting. The methodology is capable of being expanded to address the hybrid demands of enterprise-level carbon estimation since it combines domain-conscious imputation and multi-output machine learning prediction.

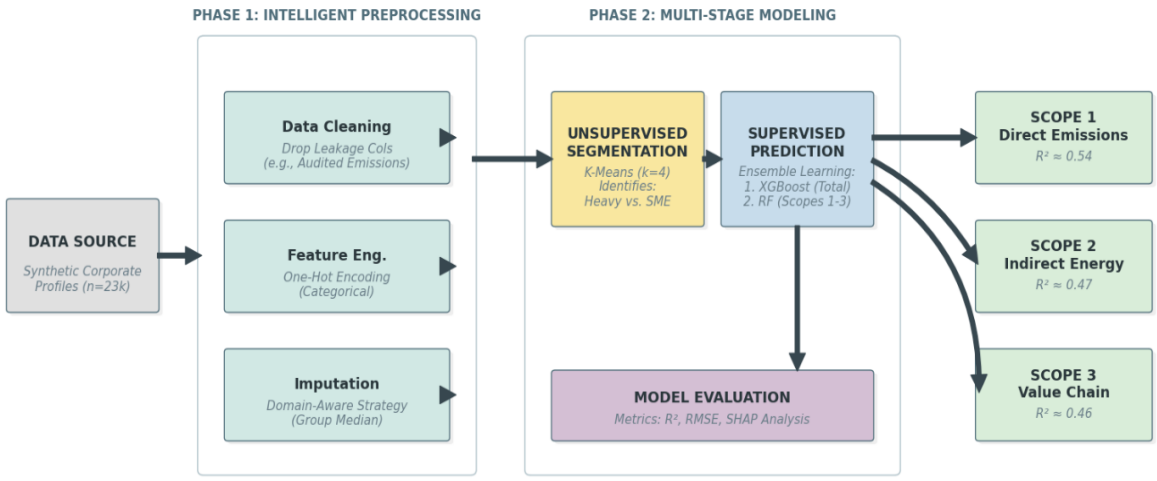

Figure 1: Integrated data processing and multi-stage modeling pipeline

The procedure commenced with a thorough contextual analysis of the data fragmentation inherent in indirect emissions reporting, followed by distinct phases of data acquisition, cleaning, strategic segmentation, and predictive modelling shown in Figure 1. This structured workflow ensured that the resulting artefact not only addressed the technical challenges of missing data but also mitigated the truncation errors common in traditional boundary estimation methods.

3.2 Data acquisition and characteristics

The research comprised of a structured dataset that contained 23,808 records, which had medium-sized companies in different industrial sectors. This dataset was used to resemble the statistical distribution and variety of real-world firms and had 59 different features from high-level firmographics such as revenue and office area to detailed operational metrics like fleet composition, electricity consumption, and procurement spend.

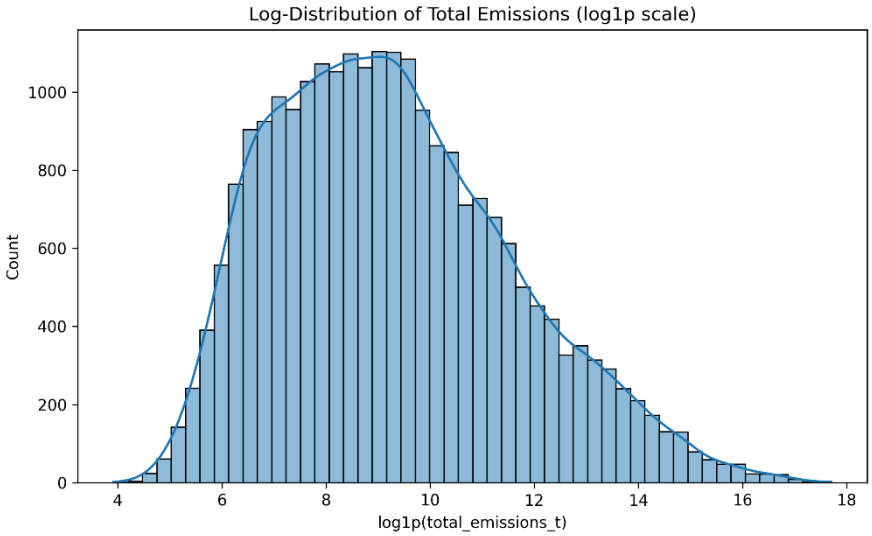

Figure 2: Log-transformed distribution of total emissions (tCO2e)

By this, it was made possible to conduct a stringent validation of the machine learning framework's skill in dealing with the data sparsity and fragmentation which are characteristic of Scope 3 accounting without the risk of disclosing confidential information. A Python-based analytical environment was used to ingest the dataset for its initial inspection and statistical profiling. As illustrated in Figure 2, the log-transformation of total emissions reveals a normalized distribution, addressing the skewness inherent in corporate environmental data.

3.3 Data preprocessing and quality assurance

Data preprocessing was given top priority to make sure that the input features for the machine learning algorithms were reliable. The first exploratory analysis pointed out that there were critical numeric fields with huge amounts of missing data, namely total_procurement _spend_usd, fleet_cars and IT equipment inventories. To avoid compromising the statistical integrity of the data, a domain-aware imputation strategy was created and applied. A tailored function was used to calculate the missing values by taking the median of certain groups that were defined by industry and size_strata. This way the values that were filled in were not only relevant to the operational profile of the company but also avoided the pitfalls of a global average that could bias the values.

The leakage removal process was carried out in a very strict manner and consequently eliminated features that might hinder the model’s performance. Model input features were removed that reflected the target variable, such as audited_total_emissions_t, capex_emissions_t, and natural_gas_emissions_t, as they would not be accessible at the time of prediction in a real-world environment. After this cleaning process, categorical variables (industry, country) were transformed using One-Hot Encoding so that they could be processed by machines, whereas numerical features were standardized using a Standard Scaler to make them comparable in range for the distance-based algorithms used in the next step.

3.4 Unsupervised learning: Strategic segmentation

To manage the differences in company emissions, an unsupervised learning step was taken to group companies into different operating k-Means clustering algorithms was used for this purpose because it is the most effective one in dividing large datasets. The best number of clusters (k) was found out by performing an analysis of Silhouette scores and Inertia values iteratively. The evaluation gave very similar Silhouette scores for both k=3(0.2001) and k=4 (0.1966), indicating that the groups had the same strength of connection. By chance, k=3 got a bit higher numerical score, the research nevertheless went for a four-cluster solution (k=4) based on interpretability from the domain viewpoint.

The four-cluster structure increased the distinction between energy-intensive industries, logistics-heavy firms and service-based SMEs and thus it was more suitable for the targeted sustainability interventions than the larger three-cluster model.

3.5 Supervised learning: Emission quantification

After segmentation, a supervised predictive modeling approach was adopted to measure carbon emissions in the three categories of Scopes 1, 2, and 3. The ColumnTransformer was used to directly integrate feature engineering into the modeling pipeline, which made the scaling and encoding steps automatic to avoid data leakage between the training and testing sets. The two methods of ensemble learning, Random Forest Regressor and XGBoost, were employed to understand the complex and non-linear interactions which existed between the different operational activities and the resulting carbon production.

We set random forest model to have 200 estimators and a minimum sample split of 4, but the XGBoost model was enabled at a learning rate of 0.05 and a max depth of 6. These hyperparameters were determined by heuristic performance evaluations during the experimental setup as the best ones to offer a good balance between computational efficiency and predictive accuracy. A Multi-Output Random Forest Regressor was then used for the granular prediction of the individual emission scopes, giving the ability to predict Scopes 1, 2, and 3 at the same time and considering the inherent correlations between the direct and indirect emission sources.

3.6 Evaluation strategy and statistical techniques

The validation of models was performed through a hold-out validation strategy. A fixed random seed (42) was used to split the dataset into 80% for training and 20% for testing, which helped in establishing reproducibility. The accuracy of the models was determined by using Root Mean Squared Error (RMSE) which penalizes larger prediction errors and the Coefficient of Determination (R^2) which gives the proportion of variance in emissions that is explained by the models. To illuminate the "black box" nature of ensemble models and to meet the business requirement for explainability, SHAP (SHapley Additive exPlanations) values were calculated. This statistical method made it possible to identify major contributors to emissions, such as revenue and procurement spending, by quantifying the marginal contribution of each feature to the final prediction.

4 Design specifications

4.1 High-level system architecture

The suggested approach was designed in a way that consists of several parts, a batch-processing machine learning pipeline capable of processing large amounts of data and predicting every single emission in detail. The system took advantage of the component-based design pattern supported by the Scikit-Learn Pipeline framework instead of going for the usual monolithic software approach. The structure was logically divided into four layers: Data Ingestion and Validation, Preprocessing and Transformation, Unsupervised Segmentation, and Supervised Inference. By adopting this layered strategy, the concerns were separated which made it possible to optimize both the data cleaning logic and the predictive modeling

algorithms independently. The data flow started with the raw CSV data, then it went through strict schema validation, feature transformation, and finally, it was processed in parallel using ensemble model architectures to get the final Scope 1, 2, and 3 estimates.

4.2 Data transformation pipeline design

The fundamental engineering design was established based on a ColumnTransformer framework that was responsible for the processing of different types of corporate sustainability data. This design choice dealt with the simultaneous processing of continuous financial metrics and nominal categorical attributes without any data leakage occurring. Continuous variables, to be more specific, revenue and office area, were sent through a standardization pipeline. This pipeline comprised a SimpleImputer set up with a median strategy to immunize missing value treatment against the influence of outliers and then, a StandardScaler to bring the feature's variance down to unit scale z-scores. K-Means clustering had been imposed to adhere to the centroids based on the Euclidean distance, hence this normalization design was a strict necessity for the clustering component. Nominal variables like industry and country were routed simultaneously through the OneHotEncoder. The encoder was made to process unknown categories during inference to keep the system strong in the production environments where new industry categories could appear in the future data streams. The design strictly did not allow target-dependent variables to be part of the input matrix to avoid data leakage, thus ensuring that columns such as audited emissions were architecturally segregated into the target vector.

4.3 Predictive modelling framework

The construction of the modeling layer was done in a way that it consisted of individual estimators, and the Multi-Output Regression framework was utilized to cope with the complexity of the different emission scopes. At the core of the predictive system was the Random Forest Regressor. In essence, the algorithm generates many decision trees that are not correlated with each other, after which they are trained, and the average of the individual trees' predictions is given as the outcome. To satisfy the necessity of emission partitioning by source, the framework made use of a Multi-Output Regressor wrapper. The meta-estimator altered the primary setup of the Random Forest in a manner that every input sample gave rise to a multi-dimensional target vector, thus preserving the statistical association of the different domains.

For validation and benchmarking, a secondary XGBoost architecture was incorporated. This model with a gradient boosting framework works sequentially and not in parallel as Random Forest does. In its essence, it generates a combined strong prediction model from several weak models, usually decision trees, where each new model tries to fix the mistakes made by the previous model by reducing the irregularized loss function. The addition of this secondary architecture made it possible to carry out performance comparisons which in turn served the research aim to trace the best algorithmic approach to carbon quantification.

5 Implementation

The entire setup was performed in a cloud-based Python 3.10 environment, specifically on Google Colab, making use of Scikit-Learn (v1.4+), XGBoost, and a modular Jupyter setup for complete reproducibility. A Pandas-based data ingestion pipeline was used to load 23,808 corporate records which had 59 features, and a constant global random seed was used for getting non-random results.To deal with data quality, domain-aware imputation (industry and size strata) was first applied, followed by a ColumnTransformer for One-Hot Encoding and Z-score normalization, resulting in a leakage-free matrix which was split into 19,046 training samples and 4,762 test samples.K-Means clustering (k=4) divided firms into four profiles and cluster IDs were added to improve the predictive power. Among the models, XGBoost was the one with the highest performance (R^2 = 0.5168), whereas the Multi-Output Random Forest correctly predicted all emission scopes, reaching 0.5449 for Scope 1. SHAP analysis pointed out revenue and procurement spend as the main drivers, and automated CSV exports of imputed data, segmentation and predictions confirmed the system's practicality for large-scale carbon footprint estimation.

6 Evaluation

The evaluation strategy was created to test the main quality of the developed models and to check if they were suitable for the process of carbon quantification automation. A detailed discussion of the experimental results is presented in this section, where not only the predictive performance of different algorithms is compared, but also the structural insights obtained from segmentation are analyzed.

6.1 Model performance benchmarking

When assessing the most efficient predictive architecture, evaluation of three ensemble learning methods Random Forest, XGBoost, and Gradient Boosting was performed by training them on the same 80% training partition and testing on a 20% hold-out set.

Table 2: Model performance benchmarking (Total Emissions)

Model Architecture	R^2 Score	RMSE (tonnes CO2e)	MAE (tonnes CO2e)
XGBoost Regressor	**0.5168**	**848,306**	97,864
Random Forest Regressor	0.4790	880,906	**95,257**
Gradient Boosting Regressor	0.4681	890,089	111,225

Random Forest, XGBoost, and Gradient Boosting were the methods that performed accordingly, with XGBoost scoring R^2 of 0.5168, which corresponds to RMSE of about 848,306 tonnes in terms of error. The R^2 value for Random Forest was 0.4790 while for Gradient Boosting it was 0.4681, thus proving that boosted trees are able to comprehend non-linear associations in complicated emission datasets to a greater extent than the other methods.

The comparative performance metrics for the three ensemble architectures are summarized in Table 2.

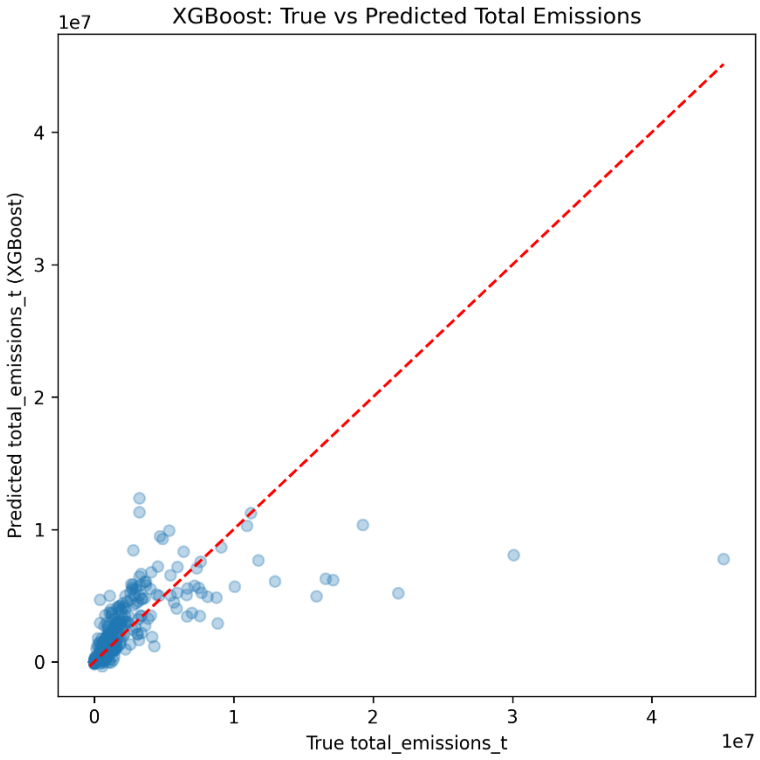

Figure 3: Predicted vs. Actual total emissions (XGBoost, R^2=0.52)

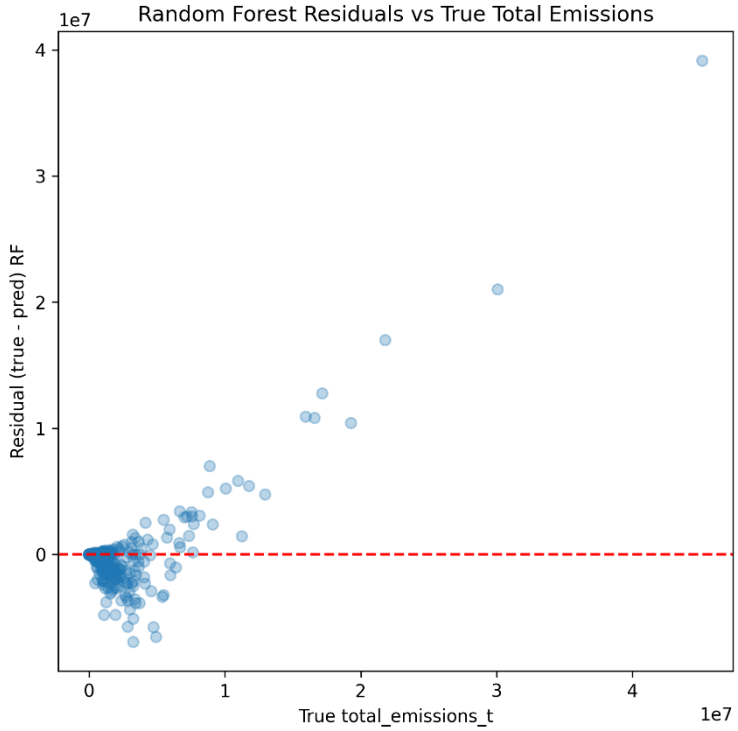

Figure 4: Residual analysis plot for Total Emissions

A moderate level of R^2 (0.52) indicates the model's limited ability to predict, which, however, is relatively acceptable considering the presence of the high-dimensional noise and unobserved factors that are characteristic of environmental datasets. Prediction versus actual values exhibiting the expected trend is illustrated in Figure 3, and Figure 4 showing no correlation of residuals supports the case that the model remains unbiased due to the underlying variability.

6.2 Scope-specific granularity analysis

The objective of this study was to employ the Multi-Output Random Forest Regressor to produce simultaneously both aggregate and scope-specific emission estimates through the joint prediction of Scopes 1, 2, and 3. Scope 1 achieved the highest accuracy (R^2 = 0.5449), indicating that direct emissions, such as those from the fleet and fuel used, can be reliably traced back to operational characteristics like fleet_cars and natural_gas_m3_annual which are observable.

Table 3: Performance of the multi-output random forest model across the three distinct emission scopes

Target Variable	R^2 Score	RMSE (tonnes CO2e)	MAE (tonnes CO2e)
Scope 1 Emissions	**0.5449**	31,154	4,109
Scope 2 Emissions	0.4706	18,568	1,927
Scope 3 Emissions	0.4693	8,44,165	90,688

However, Scope 2 (R^2 = 0.4706) and Scope 3 (R^2 = 0.4693) were not as successfully predicted. Table 3 contains the R^2 and RMSE scores for each scope, highlighting the predictive gap between direct and indirect emissions.

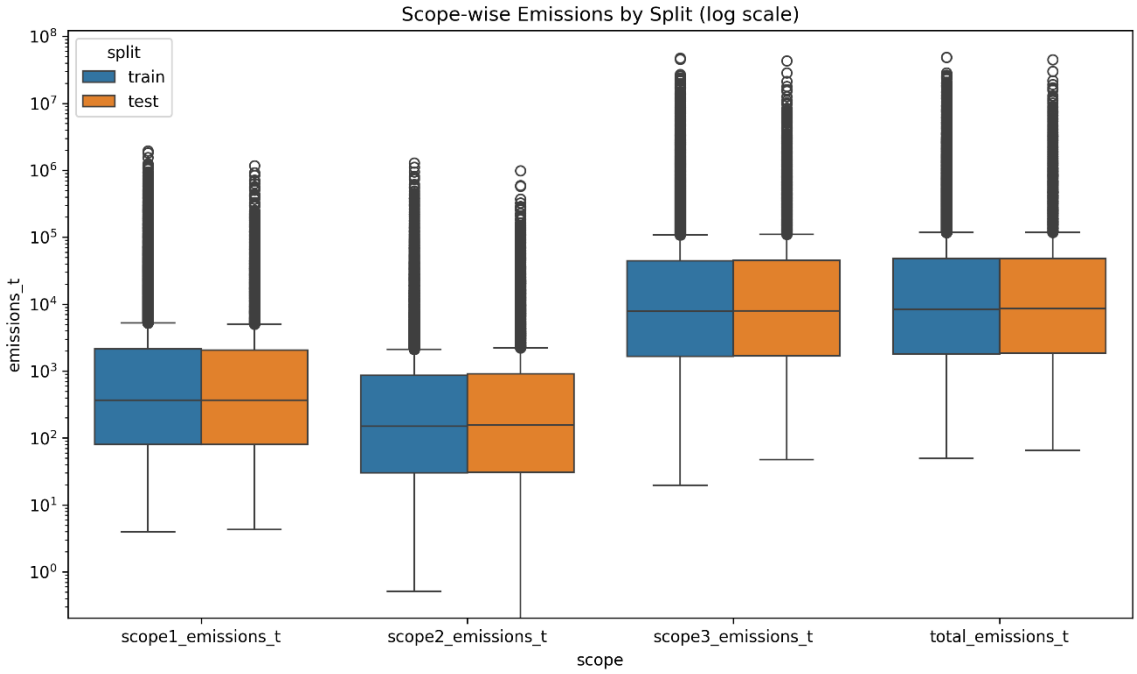

Figure 5: Boxplot analysis of emissions by Scope (Log scale)

This observation is in line with the existing data, it is moreover more difficult to track scope 3 emissions since they are the result of various complex upstream and downstream activities within the supply chain that are not usually reported in corporate disclosures. The boxplot analysis in Figure 5 illustrates this situation, as it points out the higher variance and the presence of extreme outliers in Scope 3 while, conversely, Scope 1 shows a more compact distribution.

6.3 Strategic segmentation results

The unsupervised learning phase successfully partitioned the corporate dataset into four distinct clusters (k=4), revealing significant structural heterogeneity in the population.

Table 4: Summarizes the structural segmentation of the corporate population

Cluster ID	Company Count (n)	Population Share	Avg. Total Emissions (tCO2e)	Primary Characteristics
Cluster 2	**12,555**	**52.70%**	39,08,422	**Dominant Group:** Standardized medium-large enterprises.
Cluster 0	8,079	33.90%	71,788	**Low Emitters:** Likely service-based or digital SMEs.
Cluster 3	1,693	7.10%	**1,61,09,804**	**Heavy Emitters:** Energy-intensive industrial/logistics giants.
Cluster 1	1,481	6.20%	1,01,968	**Niche SMEs:** Specialized low-emission firms.

The statistical profiles and operational characteristics of these four archetypes are detailed in Table 4. The "multi-stage" methodology is confirmed by this segmentation: the framework, by separating the high variance "heavy emitters" in Cluster 3 from the SME population in Clusters 0 and 1, circumvents the risk of using a general regression model that fits all diverse corporate populations.

6.4 Feature importance & explainability

To meet the business requirement of transparency, SHAP (SHapley Additive exPlanations) analysis was performed. The analysis revealed that company size metrics, both financial and operational, are the most potent indicators of carbon emission. More precisely, revenue_usd used to be the main character in the plot (Feature Importance > 0.80) and then came total_procurement_spend. The reliance on financial metrics to such a great extent supports the "spend-based" estimation method usually adopted in hybrid Life Cycle Assessment (LCA) approaches, thereby asserting that monetary outflows are an accurate surrogate for the upstream supply chain action (Scope 3).

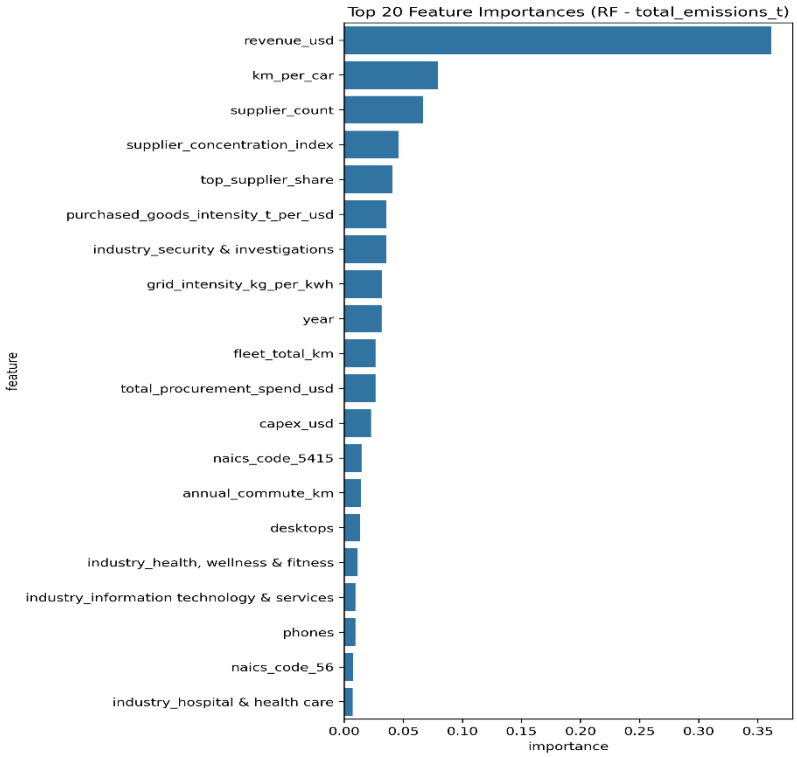

Figure 6: SHAP Summary plot ranking the top features by their marginal contribution to the model's output

Also, the importance of fleet_cars as an indicator proves that the model has properly understood the physical connection among asset ownership, and direct combustion through learning. Revenue_usd and procurement_spend are the main predictors, which are in accordance with the strong correlation that exists between financial scale and environmental footprint, as ranked in Figure 6.

7 Conclusion and future work

7.1 Research synthesis and achievement of objectives

The research aimed at developing an automated machine learning framework which can calculate corporate carbon footprints, primarily addressing the issues of data fragmentation and truncation error in Scope 3 accounting (Matthews et al., 2008). A multi-stage pipeline including domain-aware imputation, K-Means clustering, and ensemble predictive modeling was conducted on 23,808 corporate records to present its feasibility by merely relying on primary operational data. XGBoost performed the best ($R^2 = 0.5168$), validating the use of algorithmic alternatives to manual auditing (Gao, H. et al., 2023). Clustering discovered a predominant moderate-emission group (Cluster 2, n = 12,555) and a smaller, heavy-emitter group (Cluster 3, n = 1,693), revealing that industry-specific baselining is vital for accurate carbon accounting (Jiang et al., 2019).

7.2 Critical evaluation and limitations

The pipeline went through a process of truthfulness, however, the overall R^2 of 0.52 pointed out the main drawbacks. Almost half of the fluctuations in the data were not accounted for, which pointed out that firms' static characteristics like income and number of vehicles could not fully reflect the changing nature of Scope 3 Supply Chains. As mentioned by Harangozo and Szigeti (2017), company-level data lacks the detail of product-level lifecycles which makes it difficult to model indirect emissions. The synthetic data set, which was crucial for privacy compliance, likely acted as a barrier in revealing causal connections. Moreover, the SHAP analysis indicated that financial metrics had been given too much importance whilst the impacts of technological decarbonization had been neglected. Thus, before the model can be accepted as compliant with the regulatory disclosure requirements, it must undergo additional feature enrichment.

7.3 Future research directions

The commercial viability of the research will depend on accomplishing the steps of its development. The integration of dynamic external data sources aimed at reducing residual error will be the focus of the upcoming work.

1. **Real-Time Grid Integration:** the precision of Scope 2 carbon footprint estimation (R^2 = 0.47) can be improved with the help of carbon intensity APIs for real-time grid integration. Leerbeck et al. (2020) assert that time-of-use accounting not just considers the emissions but also records the peaks that static annual averages might ignore, especially in the case of energy-intensive companies.
2. **Temporal Modeling (LSTM):** By applying LSTM-based architecture instead of static ensemble models, the non-linear temporal emission patterns could be better captured. According to Wang et al. (2023), LSTMs are significantly better than Random Forests in predicting dynamic environmental trajectories.
3. **Sector-Specific Emission Factors:** It is possible that the incorporation of emission factors at the level of industry as extra traits would lead to a better performance of the clustering. The association of carbon emissions with economic functions (Jiang et al., 2019) could solve the low cohesion issue detected in the SME clusters (Clusters 0 and 1).

References

Adamolekun, G. and Kyiu, A. (2025) ‘Corporate carbon footprint and market valuation of restructuring announcements’, Review of Quantitative Finance and Accounting, 64, pp. 595–620.

Al Nuaimi, H.S., Acquaye, A. and Mayyas, A. (2025) ‘Machine learning applications for carbon emission estimation’, Resources, Conservation & Recycling Advances, 27, 200263.

Bocken, N.M.P. and Allwood, J.M. (2012) ‘Strategies to reduce the carbon footprint of consumer goods by influencing stakeholders’, Journal of Cleaner Production, 35, pp. 118–129.

Csutora, M. and Harangozo, G. (2017) ‘Twenty years of carbon accounting and auditing – A review and outlook’, Society and Economy, 39(4), pp. 459–480.

Gao, H. et al. (2023) ‘A Review of Building Carbon Emission Accounting and Prediction Models’, Buildings, 13(7), 1617.

Gao, P. et al. (2023) ‘An approach for analyzing urban carbon emissions using machine learning models’, Indoor and Built Environment, 32(8), pp. 1657–1667.

Giannelos, S. et al. (2023) ‘A Machine Learning Approach for Generating and Evaluating Forecasts on the Environmental Impact of the Buildings Sector’, Energies, 16, 2915.

Harangozo, G. and Szigeti, C. (2017) ‘Corporate carbon footprint analysis in practice – with a special focus on validity and reliability issues’, Journal of Cleaner Production, 167, pp. 1177–1183.

Jiang, M. et al. (2019) ‘The structural roles of sectors and their contributions to global carbon emissions: A complex network perspective’, Journal of Cleaner Production, 208, pp. 426–435.

Leerbeck, K. et al. (2020) ‘Short-Term Forecasting of CO2 Emission Intensity in Power Grids by Machine Learning’, arXiv preprint arXiv:2003.05740.

Mardani, A. et al. (2020) ‘A multi-stage method to predict carbon dioxide emissions using dimensionality reduction, clustering, and machine learning techniques’, Journal of Cleaner Production, 275, 122942.

Matthews, H.S., Hendrickson, C.T. and Weber, C.L. (2008) ‘The Importance of Carbon Footprint Estimation Boundaries’, Environmental Science & Technology, 42(16), pp. 5839–5842.

Meng, Y. and Noman, H. (2022) ‘Predicting CO2 Emission Footprint Using AI through Machine Learning’, Atmosphere, 13, 1871.

Nguyen, Q., Diaz-Rainey, I. and Kuruppuarachchi, D. (2021) ‘Predicting corporate carbon footprints for climate finance risk analyses: A machine learning approach’, Energy Economics, 95, 105129.

Nguyen, Q. et al. (2023) ‘Scope 3 emissions: Data quality and machine learning prediction accuracy’, PLOS Climate, 2(11), e0000208.

Nuță, F.M. et al. (2021) ‘National Carbon Accounting—Analyzing the Impact of Urbanization and Energy-Related Factors upon CO2 Emissions in Central-Eastern European Countries’, Energies, 14, 2775.

Russell, S. (2019) Estimating and Reporting the Comparative Emissions Impacts of Products. Working Paper. Washington, DC: World Resources Institute.

Scherer, M. and Milczarski, P. (2021) ‘Machine-Learning-Based Carbon Footprint Management in the Frozen Vegetable Processing Industry’, Energies, 14, 7778.

Schmidt, M., Nill, M. and Scholz, J. (2022) ‘Determining the Scope 3 Emissions of Companies’, Chemical Engineering & Technology, 45(7), pp. 1–14.

Serafeim, G. and Caicedo, G.V. (2022) Machine Learning Models for Prediction of Scope 3 Carbon Emissions. Working Paper 22-080. Harvard Business School.

Wang, C., Li, M. and Yan, J. (2023) ‘Forecasting carbon dioxide emissions: application of a novel two-stage procedure based on machine learning models’, Journal of Water and Climate Change, 14(2), pp. 477–493.

Xu, L. et al. (2025) ‘Who predicts better? A comparison of machine learning and econometrics in forecasting CO2 emissions from global shipping’, Energy, 338, 138967.

Zvezdov, D. and Hack, S. (2016) ‘Carbon footprinting of large product portfolios. Extending the use of Enterprise Resource Planning systems to carbon information management’, Journal of Cleaner Production, 135, pp. 1267–1275.