



Machine Learning Approaches to Predicting Sea Freight Carbon Emissions

A Comparative Evaluation of decision trees, Support
Vector Machines, and random forests

Jeswanth Naidu Padi
Viswas Setty

This thesis is submitted to the Faculty of Engineering at Blekinge Institute of Technology in partial fulfillment of the requirements for the degree of Bachelor Qualification Plan in Computer Science. The thesis is equivalent to 20 Weeks weeks of full-time studies.

The authors declare that they are the sole authors of this thesis and that they have not used any sources other than those listed in the bibliography and identified as references. They further declare that they have not submitted this thesis at any other institution to obtain a degree.

Contact Information:

Author(s):

Jeswanth Naidu Padi

E-mail: jepd23@student.bth.se

Viswas Setty

E-mail: vise23@student.bth.se

University advisor:

Dr.Lawrence Henesey

Department of Computer Science

Faculty of Engineering
Blekinge Institute of Technology
SE-371 79 Karlskrona, Sweden

Internet : www.bth.se
Phone : +46 455 38 50 00
Fax : +46 455 38 50 57

Abstract

Background: Current concerns about climate change require a thorough understanding of carbon emissions across organizations and their mitigation. The transportation industry's carbon emissions are largely attributed to maritime freight, which resulted in research into predictive models as an approach to predicting and efficiently reducing these emissions. Considering the ability to examine large datasets and identify complex patterns, machine learning techniques present prominent ways for these prediction tasks.

Objectives: The focus of this thesis is to assess how well machine learning algorithms predict the carbon emissions from marine freight. It specifically aims to assess the effectiveness of random forests, decision trees, and support vector machines by using the AIS data. By evaluating these methods' advantages and disadvantages, the study seeks to highlight the best methods for accurate and efficient carbon emission prediction in maritime logistics.

Methods: The chosen methodology to achieve the objectives of the thesis and address its research questions involved conducting experiments. An overall, preprocessed dataset is gathered that includes a variety of factors affecting maritime freight carbon emissions, including cargo categories, route characteristics, vessel specifications, etc. This data is then used to implement and train decision trees, support vector machines, and random forest models. Each model's predictive ability is evaluated using metrics like Mean square error, root mean square error, mean absolute error, and R^2 -score.

Results: The evaluation of decision trees, support vector machines (SVM), and random forests in comparison shows different patterns of performance for many parameters. Considering their ease of understanding and simplicity, decision trees frequently experience overfitting. Support Vector Model works well with high-dimensional data and is resistant to overfitting, although it may not perform as well on large-scale datasets. Using ensemble learning, random forests provide competitive accuracy and generalization performance.

Conclusions: The research highlighted how machine learning techniques can predict carbon emissions from marine freight. random forests, Support vector models, and decision trees all have different pros and cons, which shows how crucial it is to choose the right algorithms depending on the demands of a given application. Future studies could investigate hybrid models or incorporate other features to improve further prediction accuracy and applicability in real-world maritime logistics conditions. Overall, the findings support the advancement of sustainability initiatives in the transportation industry by promoting effective emission reduction strategies and well-informed decision-making.

Keywords: Sea freight, Carbon emissions, Machine learning, Support Vector Machines, decision trees, random forests.

Acknowledgments

We are deeply grateful to Dr.Lawrence Henesey, whose assistance as our university advisor from the Department of Computer Science at the Blekinge Institute of Technology, Sweden, proved crucial at every step during my research. His unfailing support, knowledgeable instruction, and insightful feedback helped us in better understand the topic and guided each phase of my research.

We also wish to express my gratitude to the course responsible, Dr.Prashant Goswami from the Department of Computer Science for his constructive input from thorough assessment throughout the course. Their combined efforts and dedication to this research have been much appreciated and we are immensely thankful for their contributions.

- Jeswanth Naidu Padi
Viswas Setty

List of Acronyms

SVM	Support Vector Machine
MSE	Mean Square Error
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
R^2	R-Squared
AIS	Automatic Identification System

Contents

Abstract	i
Acknowledgments	ii
1 Introduction	1
1.1 On the content	1
1.2 Ethical, societal and sustainability aspects	3
1.2.1 Ethical aspects:	3
1.2.2 Societal aspects:	4
1.2.3 Sustainability aspects:	4
1.3 Aim	4
1.4 Objectives	4
1.5 Research questions	5
1.6 Scope of the thesis	5
1.7 Outline	6
2 Background	7
2.1 Carbon emissions from transportation	7
2.2 Carbon emissions from sea freight	8
2.3 Machine learning	9
2.4 Machine learning algorithms:	10
2.4.1 decision trees	10
2.4.2 random forests:	11
2.4.3 SVMs	11
2.5 Performance metrics	12
2.5.1 Mean Square Error (MSE)	12
2.5.2 Root Mean Square Error (RMSE)	13
2.5.3 Mean Absolute Error (MAE)	13
2.5.4 R-squared score	13
3 Related Work	14
3.1 Research gap established through related work	16
4 Method	17
4.1 Data Collection:	18
4.2 Data Preprocessing:	19
4.3 Feature selection and Dimensionality reduction:	22
4.4 Model Training:	24

4.5	Model Evaluation:	25
4.6	Model Testing	26
5	Results and Analysis	27
5.1	Analysis	29
6	Discussion	31
6.1	Reflections:	32
7	Conclusions and Future Work	34
7.1	Conclusion	34
7.2	Future work	35
	References	36

List of Figures

1.1	Input Output Parameters Flowchart showing the dataset attributes as well as performance metrics	3
2.1	Carbon emissions from transportation sector in the year 2018 [21] . .	8
2.2	Carbon emissions trends in sea freight from 2010–2050 [9]	9
2.3	Pictorial representation of the components in a decision tree [11] . .	10
2.4	Pictorial representation of the random forests [10]	11
2.5	SVM with different categories classified using hyperplane [1]	12
2.6	Formula for Mean Square Error	13
2.7	Formula for Root Mean Square Error	13
2.8	Formula for Mean Absolute Error	13
2.9	Formula for R-Squared Score	13
4.1	Methodology showing the different phases of the experimental procedure followed	18
4.2	Overview of the dataset	19
4.3	Standard scaling	20
4.4	Dataset after handling missing values	20
4.5	K-Means clustering	21
4.6	Categorical Encoding on dataset attributes	21
4.7	Distribution of numerical features	22
4.8	Correlation of Numerical features	23
4.9	Formula for Mean Square Error	25
4.10	Formula for Root Mean Square Error	26
4.11	Formula for Mean Absolute Error	26
4.12	Formula for R-Squared Score	26
5.1	SVM metrics showing highest MSE, RMSE and MAE values	28
5.2	Random Forest regressor metrics showing highest accuracy when compared to SVM and decision tree	28
5.3	decision tree regressor performs decently with lowest training time . .	28
5.4	Visualization of MAE, MSE, R2-score and MSE for each trained model	29

List of Tables

5.1	Evaluation metrics on test set	30
-----	--	----

1.1 On the content

The movement of commodities by cargo is essential to linking markets and maintaining economic growth in the modern world of international trade. However, environmental consequences are associated with this crucial component of the supply chain, as carbon emissions play a major role in air pollution and climate change. It is now essential to address the environmental effects of carbon emissions to meet sustainable development goals. Every year, trucks, airplanes, ships, and trains move billions of tons of cargo throughout the globe. 8 percent of the world's greenhouse gas emissions come from transportation, and that number rises to 11 percent when ports and warehouses are taken into account [2]. The problem of carbon emissions is complex and involves commercial viability, legal compliance, and environmental sustainability. Conventional approaches to this issue frequently fail to deliver thorough and timely solutions. Cargo operations generate enormous amounts of complicated data, which makes a sophisticated analytical approach necessary.

Transport accounts for around one-fifth of global carbon dioxide (CO₂) emissions. The World Resource Institute's Climate Data Explorer provides data from CAIT on the breakdown of emissions by sector. In 2016, global CO₂ emissions (including land use) were 36.7 billion tonnes of CO₂; emissions from transport were 7.9 billion tonnes of CO₂. Transport therefore accounted for $7.9 / 36.7 = 21$ percent of global emissions [21]. According to reports from Sinay.ai, cargo ships emit 16.14 grams of CO₂ per kilometer for every metric ton of goods they transport [3]. Aspects of natural supporting factors in coordination have been examined in prior research in the field, from embracing greener measures to optimizing in practice methods. Although these contributions have advertised smart data, a basic investigation demonstrates a noteworthy vacuum within the writing approximately an all-encompassing, data-driven strategy for determining and optimizing cargo emissions in an assortment of calculated scenarios to see into the outflows of gasses and particles, particularly those of ultrafine particles, from a huge bulk carrier utilizing HFO (3.13 wt percentage) beneath different working circumstances, such as at-berth moving, and ocean-going [13].

To overcome these challenges, researchers have developed predictive models that can predict and eventually lower the carbon emissions related to maritime shipping. One such technology is machine learning. This thesis aims to develop precise predictions of marine freight carbon emissions using machine learning methods, particularly

decision trees, support vector machine (SVM), and random forests. The objective is to assess and compare their performance using key evaluation metrics such as MSE, RMSE, MAE, R^2 -score, etc. A comprehensive evaluation will be carried out by utilizing Automatic Identification System (AIS) data in combination with essential emissions data from maritime freight. Numerous variables will be examined in this research, such as the characteristics of the vessel, the requirements of the route, the kinds of cargo, and operational metrics like fuel consumption and speed.

By processing huge datasets, extracting relevant details, and making accurate predictions, machine learning plays a critical role in fulfilling the goals of predicting carbon emissions in maritime freight. Machine learning models can learn from historical AIS data using supervised learning techniques like decision trees, SVMs, and random forests to find patterns and connections between vessel features, operational parameters, and carbon emissions. To verify accuracy and reliability, these models are evaluated using performance metrics such as MSE, RMSE, MAE, and R^2 -score. In addition to improving the accuracy of emission predictions, machine learning also offers insights into the key causes of these emissions, supporting the development of responsible regulations and the implementation of efficient emission reduction strategies. Additionally, machine learning models are appropriate for real-time applications due to their scalability and adaptability, which ultimately support sustainability activities in the maritime industry [4].

The dataset [1] comprises data from two sources for the Automatic Identification System (AIS): satellite and terrestrial. The first dataset covers port calls made between January 1, 2021, and April 31, 2022, inside the following nations: Sweden, Finland, Estonia, Latvia, Lithuania, and Poland. All cargo and tanker vessels over 65 meters are included, and comprehensive data including port ID, port name, LOCODE, MMSI, IMO, vessel name, destination, type, and arrival and departure times are provided. The second collection provides historical AIS data within a defined zone defined by latitude and longitude coordinates, namely latitudes 54.5 to 55.4 degrees and longitudes 13.0 to 13.5 degrees. Along with cargo and tanker vessels longer than 65 meters, this dataset extends the same time frame as the preceding one, from January 1, 2021, to April 31, 2022. The format of the data sample contains all the necessary characteristics with a 10-minute resolution, such as MMSI, TimePosition, Latitude, Longitude, Speed, Course, Heading, NavStatus, TimeVoyage, IMO, Name, Callsign, VesselType, Draught, TimeETA, and Destination. Figure 1.1 explains our system's input and output parameters, including input, operational, and output parameters.

Input parameters: This is the raw data gathered from the Automatic Identification System, which monitors the motion of ships.

- **Port Calls Data:** Details regarding the names of the ports, ship identifiers (IMO and MMSI), ship types, and arrival and departure times of ships are all included in this data.
- **Historical AIS Data:** Detailed records of ships' movements within a designated geographic area are updated every ten minutes. These logs include

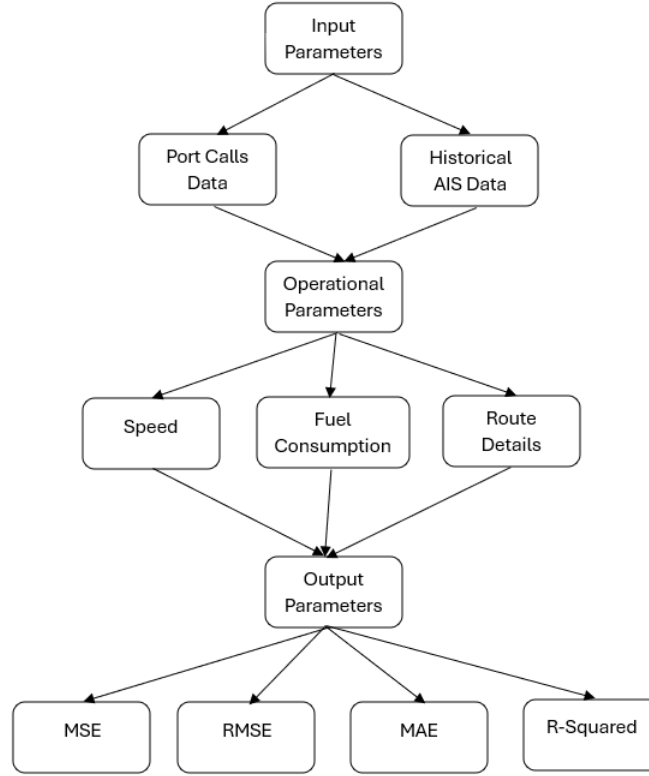


Figure 1.1: Input Output Parameters Flowchart showing the dataset attributes as well as performance metrics

locations (latitude and longitude), speeds, courses, directions, and other navigational statuses.

Operational parameters: Additional significant information such as the vessel's speed, consumption of fuel, and route details. These operational parameters are the main inputs to the models for predicting the carbon emissions in maritime freight.

Output Parameters: The operational parameters are used to get the carbon emissions. These values are then fed to the machine learning models like decision trees, SVMs, and random forests to get the predictions. Further, each model's performance is evaluated using metrics like MSE, RMSE, MAE, and R-squared, which are the output parameters of our system.

1.2 Ethical, societal and sustainability aspects

1.2.1 Ethical aspects:

The AIS and other data we use in our research are freely accessible to everyone and are in the public domain or any open-source data. The data doesn't contain any private information about specific individuals, it is not closed. Since the data we're using only provides the route of the vessels and has nothing to do with any specific nation or person, there are no GDPR concerns.

1.2.2 Societal aspects:

Using advanced machine learning methods to predict sea freight emissions precisely, this work can help influence policy decisions that seek to mitigate the environmental impact of marine transportation. Reducing emissions from maritime freight helps fight climate change and enhances air quality, which benefits vulnerable people and coastal communities globally in terms of health and well-being. Additionally, by encouraging environmentally friendly practices in the marine sector, the research supports a larger social movement toward environmentally friendly transportation networks, supporting international efforts to build a more resilient and sustainable future.

1.2.3 Sustainability aspects:

The results support the development of initiatives that promote sustainable freight transportation by helping to identify environmental regulations and recommendations. The integration of data mining methodologies to address cargo emissions aligns to attain industrial modifications over the long term through environmental concern adaptation and mitigation.

1.3 Aim

This research aims to evaluate and compare the effectiveness of machine learning algorithms, namely decision trees, SVMs, and random forests, in predicting sea freight carbon emissions. By conducting a comprehensive analysis of performance metrics (e.g., MSE, RMSE, MAE, R^2 -score) of the AIS data, the study seeks to identify the most accurate and efficient algorithm for predicting carbon emissions in maritime logistics.

1.4 Objectives

The objectives of this research are:

- To obtain and preprocess the AIS data and relevant other data on emissions from marine freight, considering various factors including vessel characteristics, route details, cargo, operational parameters (e.g., speed, fuel consumption), etc.
- To implement machine learning models based on decision trees, SVMs, and random forests for predicting sea freight emissions.
- To conduct a comparative evaluation of the performance metrics (e.g., MSE, RMSE, MAE, R^2 -score) and to analyze the interpretability and scalability of the implemented machine learning algorithms.

1.5 Research questions

RQ1: When training each learning model on sea freight carbon emissions dataset, which algorithm among decision trees, SVMs, and random forests results in the highest effectiveness in predicting carbon emissions in maritime shipping when considering AIS data?

Motivation: The focus of this research question is to address the need for accurate carbon emissions prediction in maritime freight, which is essential for reducing negative environmental impact. By introducing machine learning as a method, we assess the effectiveness of various machine learning algorithms, such as decision trees, SVMs, and random forests, to determine which one is the most accurate in detecting and analysing useful patterns in the raw AIS data.

RQ2: When assessing the trained machine learning algorithms, how is each algorithm performing on metrics like MSE, RMSE, MAE, and R^2 -score?

Motivation: This question aims to compare the performance of different machine learning algorithms (e.g., decision trees, SVMs, and random forests) in predicting sea freight emissions. By examining metrics such as MSE, RMSE, MAE, and R^2 -score, we seek to determine the most effective algorithm for predicting carbon emissions in sea freight logistics.

1.6 Scope of the thesis

The scope of this thesis extends to include a careful examination of how well machine learning algorithms like decision trees, SVMs, and random forests can predict marine freight carbon emissions. The objective of this research is to collect and preprocess AIS data in addition to relevant emissions data from maritime freight, taking into account a wide range of characteristics such as vessel features, route specifications, cargo types, and operational parameters including speed and fuel consumption. To predict marine freight emissions, machine learning models will be implemented and also performance metrics including MSE, RMSE, MAE, and R^2 -score will be compared.

- Research Question 1 (RQ1) looks into the factors that affect an algorithm's performance to determine which one is the most effective in predicting carbon emissions in maritime shipping using AIS data.
- Research Question 2 (RQ2) is concerned with identifying differences in the machine learning algorithms' performance and explaining the causes of these deviations. Metrics including MSE, RMSE, MAE, and R^2 -score will be particularly examined.

This research aims to use machine learning methods to reduce the environmental impact of maritime freight by accurately predicting carbon emissions. This thesis tries to provide useful insights for improving carbon emissions prediction in sea shipping by carefully comparing and evaluating different methods.

1.7 Outline

The outline of the thesis includes-

- **Chapter 1** which introduces the importance of predicting sea freight carbon emissions and outlines the thesis's objectives and research questions, providing context for the research focus on maritime logistics emissions.
- Foundational concepts regarding maritime emissions and machine learning techniques are discussed in **chapter 2**.
- **Chapter 3** is the literature review of existing research on predicting maritime carbon emissions and identifying gaps and areas for further exploration.
- Details of the research approach, including data collection, preprocessing, and implementation of machine learning algorithms (decision trees, SVMs, and random forests) are mentioned in **chapter 4**.
- This **chapter 5** presents experimentation findings, focusing on machine learning algorithms' performance in predicting carbon emissions.
- In **chapter 6**, we interpret results, comparing algorithm performance and discussing implications for future research and applications.
- In the end, **chapter 7** summarizes key findings, identifies the most effective algorithm, and suggests directions for future research.

2.1 Carbon emissions from transportation

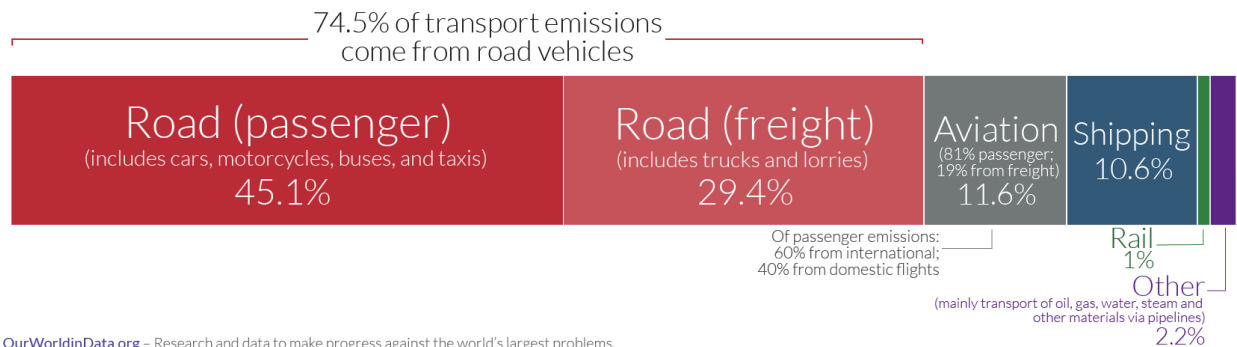
Carbon emission is the release of carbon compounds into the atmosphere. To talk about carbon emissions is simply to talk of greenhouse gas emissions, the main contributors to climate change. Since greenhouse gas emissions are often calculated as carbon dioxide equivalents, they are often referred to as “carbon emissions” when discussing global warming or the greenhouse effect. Since the industrial revolution, the burning of fossil fuels has increased, which directly correlates to the increase of carbon dioxide levels in our atmosphere and, thus, the rapid increase of global warming [25]. Carbon dioxide emissions are the primary driver of global climate change. It’s widely recognized that to avoid the worst impacts of climate change, the world needs to urgently reduce emissions [22]. Transport accounts for around one-fifth of global carbon dioxide (CO₂) emissions [24% if we only consider CO₂ emissions from energy].

- Road travel accounts for three-quarters of transport emissions. Most of this comes from passenger vehicles – cars and buses – which contribute 45.1 percent. The other 29.4 percent comes from trucks carrying freight.
- Since the entire transport sector accounts for 21 percent of total emissions, and road transport accounts for three-quarters of transport emissions, road transport accounts for 15 percent of total CO₂ emissions.
- Aviation – while it often gets the most attention in discussions on action against climate change – accounts for only 11.6 percent of transport emissions. It emits just under one billion tonnes of CO₂ each year – around 2.5 percent of total global emissions [we look at the role that air travel plays in climate change in more detail in another article]. International shipping contributes a similar amount, at 10.6 percent.
- Rail travel and freight emits very little – only 1 percent of transport emissions. Other transport – which is mainly the movement of materials such as water, oil, and gas via pipelines – is responsible for 2.2 percent [21].

Figure 2.1 shows the global transport emissions in 2018 by the International Energy Agency (IEA).

Global CO₂ emissions from transport

This is based on global transport emissions in 2018, which totalled 8 billion tonnes CO₂.
Transport accounts for 24% of CO₂ emissions from energy.



OurWorldinData.org – Research and data to make progress against the world's largest problems.

Data Source: Our World in Data based on International Energy Agency (IEA) and the International Council on Clean Transportation (ICCT).

Licensed under CC-BY by the author Hannah Ritchie.

Figure 2.1: Carbon emissions from transportation sector in the year 2018 [21]

2.2 Carbon emissions from sea freight

Cargo ships are responsible for a significant amount of carbon dioxide emissions, which contribute to climate change and air pollution. According to the International Maritime Organization (IMO), shipping accounts for approximately 2.5 percent of global greenhouse gas emissions, with cargo ships being the largest contributor. In comparison to other modes of transportation, cargo ships emit significantly more carbon dioxide per unit of cargo transported. For example, a single large container ship can emit as much pollution as 50 million cars. This is because cargo ships run on heavy fuel oil, which is a low-quality fuel that emits high levels of sulfur dioxide and nitrogen oxides. The impact of cargo ship emissions on the environment is significant. The emissions from cargo ships contribute to the warming of the planet, leading to melting ice caps, rising sea levels, and extreme weather events. In addition, the pollutants released by cargo ships can have negative effects on human health, including respiratory problems, heart disease, and cancer [7].

International shipping emissions had a year-on-year growth of 4.9 percent in 2021, rising to approximately 700 million metric tons of carbon dioxide (Mt CO). This was higher than 2019 and accounted for roughly 11 percent of total global transportation CO emissions that year. Emissions from international shipping have risen almost 90 percent since 1990, owing to increasing seaborne trade and the growing number of ships crossing the world's oceans. While GHG emissions within the shipping industry have been on an upward trajectory, releases of sulfur dioxide (SO) have been falling since 2009 due to imposed limits on sulfur content in marine fuels. Stricter regulations were introduced by the International Maritime Organization (IMO) in 2020, lowering the upper limit of sulfur content to 0.5 percent (previously 3.5 percent). It is estimated that these new regulations could have reduced shipping SO emissions by 77 percent that year, relative to 2019 levels [14]. Figure 2.2 shows the increasing trends of the number of carbon emissions by cargo ships over the period 2010–2050 [9].

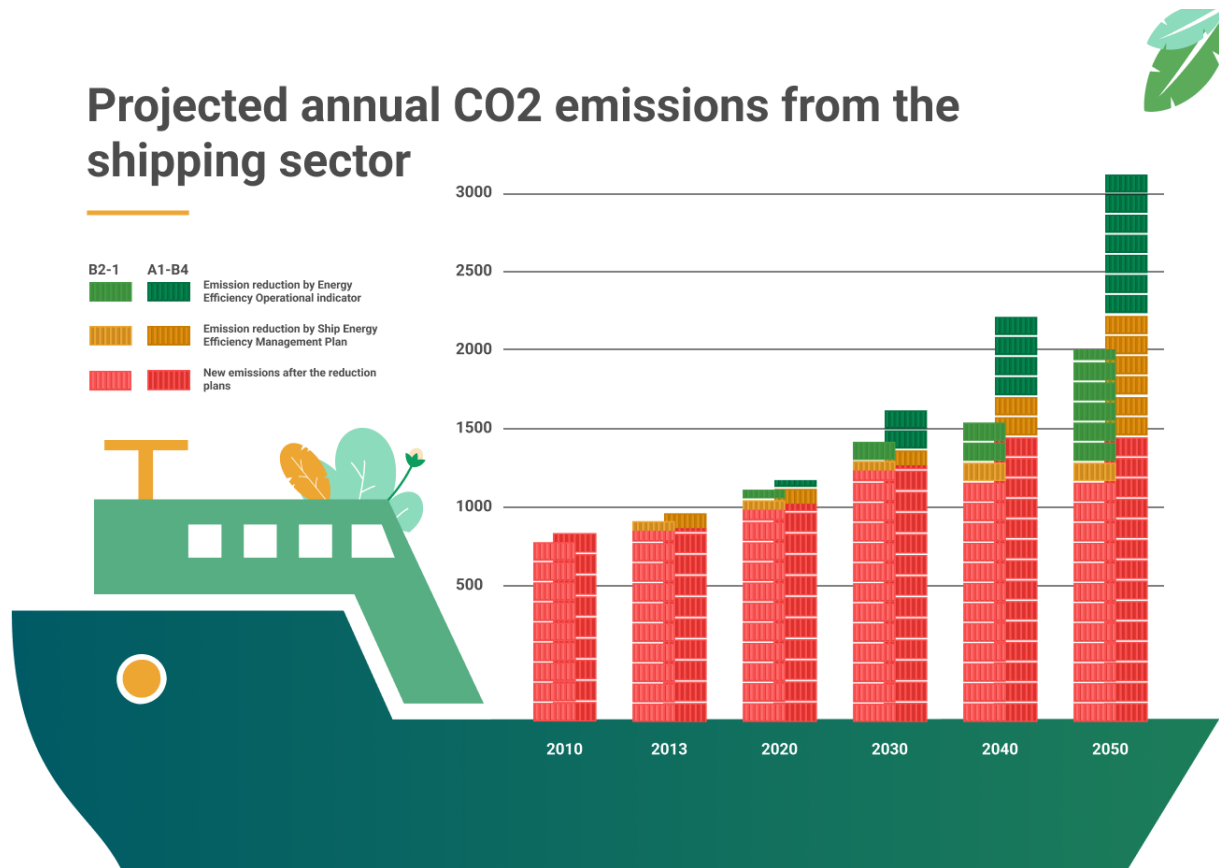


Figure 2.2: Carbon emissions trends in sea freight from 2010–2050 [9]

2.3 Machine learning

Machine learning is a subfield of artificial intelligence (AI) that uses algorithms trained on data sets to create self-learning models that are capable of predicting outcomes and classifying information without human intervention. Machine learning is used today for a wide range of commercial purposes, including suggesting products to consumers based on their past purchases, predicting stock market fluctuations, and translating text from one language to another. AI refers to the general attempt to create machines capable of human-like cognitive abilities, and machine learning specifically refers to the use of algorithms and data sets to do so. Several different types of machine learning power the many different digital goods and services we use every day. To help you get a better idea of how these types differ from one another, here's an overview of the four different types of machine learning primarily in use today.

1. Supervised learning: In supervised machine learning, algorithms are trained on labeled data sets that include tags describing each piece of data. In other words, the algorithms are fed data that includes an “answer key” describing how the data should be interpreted. Supervised machine learning is often used to create machine learning models used for prediction and regression purposes.

2. Unsupervised learning: Unsupervised machine learning uses unlabeled data sets to train algorithms. In this process, the algorithm is fed data that doesn't include tags, which requires it to uncover patterns on its own without any outside guidance. Unsupervised machine learning is often used by researchers and data scientists to identify patterns within large, unlabeled data sets quickly and efficiently.

3. Semi-supervised learning: Semi-supervised machine learning uses both unlabeled and labeled data sets to train algorithms. Generally, during semi-supervised machine learning, algorithms are first fed a small amount of labeled data to help direct their development and then fed much larger quantities of unlabeled data to complete the model. Semi-supervised machine learning is often employed to train algorithms for regression and prediction purposes if large volumes of labeled data are unavailable [12].

2.4 Machine learning algorithms:

2.4.1 decision trees

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both regression and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes, and leaf nodes. A decision tree starts with a root node, which does not have any incoming branches. The outgoing branches from the root node then feed into the internal nodes, also known as decision nodes. Based on the available features, both node types conduct evaluations to form homogenous subsets, which are denoted by leaf nodes, or terminal nodes. The leaf nodes represent all the possible outcomes within the dataset. Figure 2.3 is the pictorial representation of the decision tree [11].

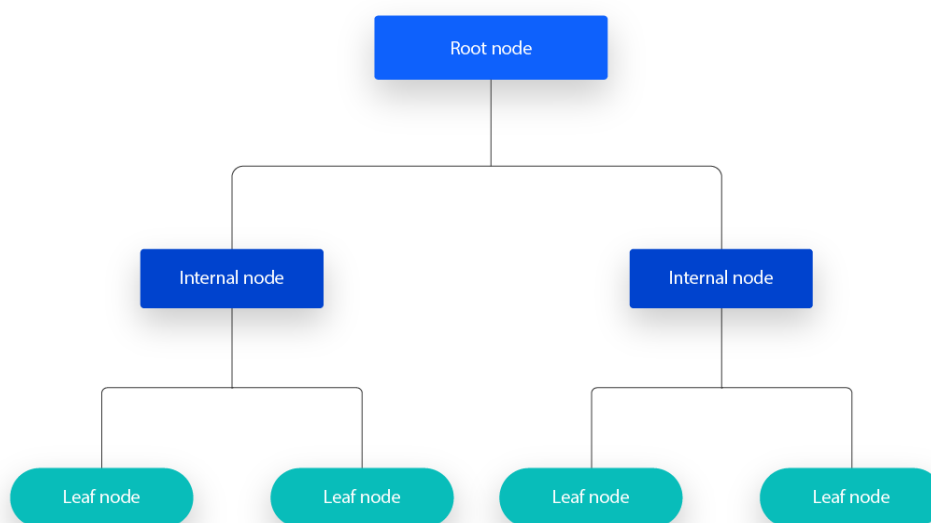


Figure 2.3: Pictorial representation of the components in a decision tree [11]

2.4.2 random forests:

The Random Forest algorithm is a powerful tree-learning technique in Machine Learning. It works by creating some decision trees during the training phase. Each tree is constructed using a random subset of the data set to measure a random subset of features in each partition. This randomness introduces variability among individual trees, reducing the risk of overfitting and improving overall prediction performance. In prediction, the algorithm aggregates the results of all trees, either by voting. This collaborative decision-making process, supported by multiple trees with their insights, provides an example of stable and precise results. random forests are widely used for regression and regression functions, which are known for their ability to handle complex data, reduce overfitting, and provide reliable predicts in different environments [10]. Figure 2.4 is the pictorial representation of the random forests.

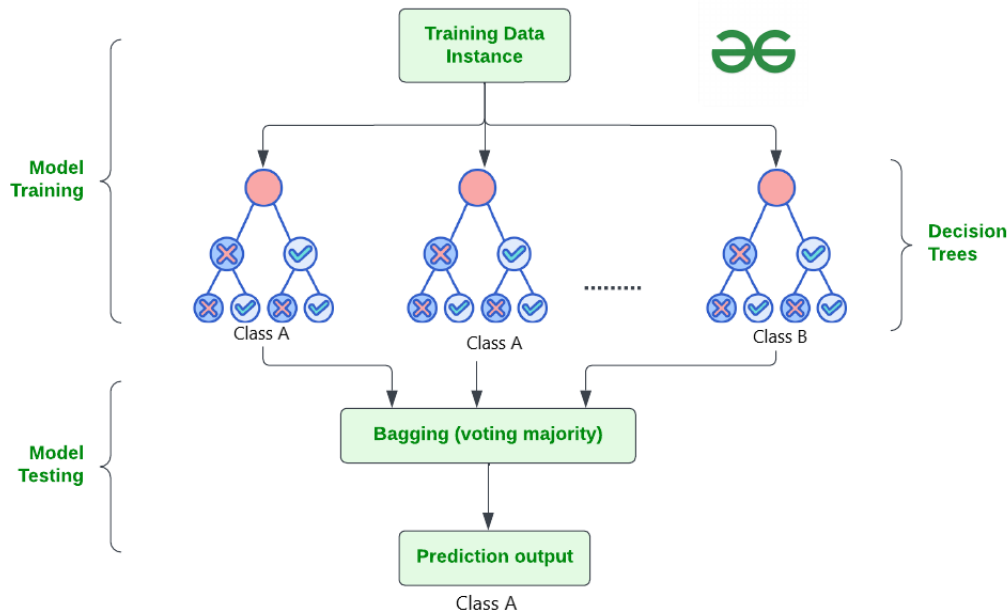


Figure 2.4: Pictorial representation of the random forests [10]

2.4.3 SVMs

SVM or SVM is one of the most popular supervised learning algorithms, which is used for regression as well as regression problems. However, primarily, it is used for regression problems in machine learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category. This best decision boundary is called a hyperplane. This algorithm chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence algorithm is termed a Support Vector Machine. It can be used for Face detection, image regression, text categorization, etc. Figure 2.5 is an SVM diagram in which two different categories are classified using a decision boundary or hyperplane.

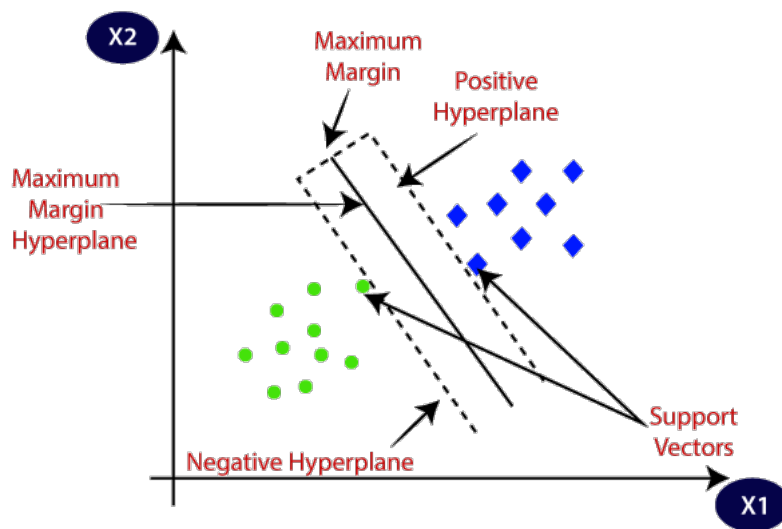


Figure 2.5: SVM with different categories classified using hyperplane [1]

Types of SVMs: SVM can be of two types:

Linear SVM: Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and regressor is used called as Linear SVM regressor.

Non-linear SVM: Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data, and the regressor used is called Non-linear SVM regressor [5].

2.5 Performance metrics

In Machine Learning it is key to be able to correctly evaluate the model being produced to guarantee that the predictions are accurately describing the intended phenomenon. However, there are so many different performance metrics that Data Scientists can use (MSE, RMSE, MAE, R^2 -score, etc.) that it is often overwhelming to choose which one to use. Selecting the right metric for a specific model, however, is key to being able to measure the performance of the model objectively and in the right setting [8].

2.5.1 Mean Square Error (MSE)

Mean Square Error (MSE) is a measure of the average squared difference in a regression task between the values that were predicted and the actual values. It measures the typical size of the errors the model produces. From the given formula n is the number of observations, $\text{cap}(y_i)$ is the model's predicted value, and y_i is the actual value which is the y -test. Figure 2.6 shows the formula for mean square error.

$$(1/n) * \sum (y_i - \hat{y}_i)^2$$

Figure 2.6: Formula for Mean Square Error

2.5.2 Root Mean Square Error (RMSE)

The average magnitude of errors in the same units as the target variable is measured by RMSE, which is the square root of the MSE. In contrast to MSE, it provides a more logical interpretation. Figure 2.7 shows the formula for root mean square error.

$$\sqrt{(MSE)}$$

Figure 2.7: Formula for Root Mean Square Error

2.5.3 Mean Absolute Error (MAE)

The mean absolute difference (MAE) between the expected and actual values is a measurement. Compared to MSE, it offers a more reliable measure of error since it is less susceptible to outliers. Figure 2.8 shows the formula for mean absolute error.

$$(1/n) * \sum |y_i - \hat{y}_i|$$

Figure 2.8: Formula for Mean Absolute Error

2.5.4 R-squared score

The statistical metric known as R-squared quantifies the percentage of the target variable's volatility that can be accounted for by the regression model. Higher values indicate better model fit; the range is 0 to 1 [6]. From the formula shown below, SS tot is the total sum of squares and SS res is the sum of squared residuals. Figure 2.9 shows the formula for mean square error.

$$R^2 = 1 - (SS_{res} / SS_{tot})$$

Figure 2.9: Formula for R-Squared Score

This chapter mainly consists of some literature reviews of previous research papers that are related to carbon emissions.

Huang et al. [15] focused on Guangdong province, China, due to its high GDP and carbon emissions. They utilized economic, demographic, and openness data to predict carbon emissions. Among the six models (K-Nearest Neighbor, Back Propagation neural network, Random Forest, multiple linear regression model, XGBoost, and LightGBM) compared, RF, XGB, and LGB outperformed others, with LGB demonstrating superior efficiency. The LGB model achieved an R^2 value >0.97 , MAPE of 10.734 percent, consuming only 175KB memory, with an uptime of 0.06s. LGB's success is attributed to its fast, distributed, high-performance gradient-boosting framework based on decision trees, enabling higher accuracy and efficiency.

In 2018, Pooja Kadam and Suhasini Vijayumar [17] provided insight of CO₂ emission prediction model using machine learning. Traditionally researchers have used statistical techniques such as regression, t-test, derivation, and ANOVA Test for prediction. Machine learning provides different techniques to train the machine based on experience. The regression model in the paper is used to predict CO₂ emission based on historical data. For better prediction model trial and error approach was applied to obtain lower error. The Root Mean Square Error (RMSE) value is 0.2557. The main objective of the experiment is to achieve a low RMSE value.

The maritime industry is a significant source of global carbon emissions, and the accurate prediction of emissions in this domain is of paramount importance. Vasilis Michalakopoulos et al. [19] conducted a comparative analysis of machine learning algorithms for predicting CO₂ emissions in the maritime domain. Using a unique dataset, comprising historical maritime data for different vessel voyages, including vessel type, voyage information, environmental factors and employed decision trees, many regression, and artificial neural network algorithms. Performance evaluation is conducted using established metrics such as R^2 , Mean Absolute Error, Normalized Root Mean Squared Error, and Symmetric Mean Biased Error. The findings reveal that the Extra Trees Regressor and Multi-layer Perceptron regressor algorithms outperform the other methods in terms of prediction accuracy demonstrating the least amount of error.

Carbon Emission Prediction is important because China is the world's largest emitter of carbon, and an analysis of the factors affecting China's carbon emissions will help the country to save energy and reduce emissions in the future. This problem has been studied in the literature with 9 methods and 9 kinds of data. Most of

the data were collected in the National Bureau of Statistics of China. Hongxuan Tan [24] predicts carbon emission with 44 Macroeconomic Variables and 7 machine learning models, with data from 31 provinces in China from 2011 to 2019. The study uses machine learning algorithms to discuss the impact of macroeconomic indicators in the dataset on carbon emissions in China. From the analysis, it is found that the AdaBoost model is the most precise, and Coal consumption variables are more important in the prediction.

This research done by Lida Anna Joshy et al. [16] presents an observational and predictive study aimed at offering a comparative analysis of various brands and vehicle types concerning their fuel consumption and CO₂ emissions. The paper successfully demonstrates the power of regression analysis and machine learning in predicting CO₂ emissions from vehicles. By following a comprehensive pipeline encompassing data preprocessing, exploratory data analysis, model training, and evaluation, this research has successfully crafted precise predictive models capable of estimating emissions based on vehicle attributes. The comparison of various regression algorithms highlighted the superiority of the XGBoost Regressor in achieving the highest accuracy, emphasizing the significance of employing advanced techniques for intricate prediction tasks. The paper's findings underscore the importance of vehicle attributes like engine size, fuel consumption, and class in influencing emission levels.

Fongyiu Wong [26] did research on carbon emissions allowances trade prediction based on machine learning. Long-term prediction and short-term prediction models were adopted to predict carbon emissions trading volume. Linear regression, decision trees, SVMs, extreme gradient boosting and random forests were adopted to perform long-term prediction. Additionally, to improve the accuracy and effect of the prediction, they adopted a time series forward multi-step hybrid intelligent prediction model to perform short-term prediction. In the short-term prediction model training process, reinforcement learning and the hidden Markov model were combined first, followed by the combination of neural network and hidden Markov model with reinforcement learning as a bridge. The conclusion of the machine learning-based carbon emissions allowances trade prediction can be summarized as a nonlinear relationship between the lag of volume/frequency of transactions and price in the future. Among the five long-term prediction models, Random Forest performs best with the smallest mean absolute error (24.42) and the mean absolute error of short-term prediction is 6.79.

Babasaheb S. Satpute et al. [23] examined the use of machine learning methods to predict CO₂ emissions using a dataset that included characteristics of more than 7500 automobiles. The model yielded a mean absolute error of 3.24 and a mean squared error of 30.00. Although the model offers an easy-to-understand method, its efficacy may be restricted due to its presumption of a linear correlation between characteristics and CO₂ emissions. The decision tree model performed better with an MAE of 1.86 and an MSE of 14.27. With an MSE of 2.32 and an MAE of 1.83, the Random Forest model improved the predicted accuracy even more. With an MSE of 406.86 and an MAE of 9.54, the SVM model performed comparatively worse than the earlier models.

3.1 Research gap established through related work

Although prior research has explored various methods for emissions prediction in container ships, there exists a research gap in the comprehensive comparison and assessment of machine learning algorithms specifically developed for predicting carbon emissions in marine freight. To close this gap and support developments in emissions management in the maritime sector, the thesis compares decision trees, SVMs, and random forests to determine which algorithm is best for estimating carbon emissions in maritime freight. These algorithms are selected because -

- **random forests** are a good fit for *simulating the complicated patterns* of marine freight carbon emissions because they are good at capturing complex interactions within datasets.
- **decision trees** provide *feature importance transparency*, which helps identify key factors affecting the carbon emissions from marine freight.
- Due to their ability to deal with *high-dimensional data*, **SVMs** are a valuable tool for examining the various characteristics linked to carbon emissions from marine freight.

These algorithms are also used in wide use across multiple domains, robustness, flexibility, high prediction accuracy, and capacity to manage non-linearity [20] [18]. These algorithms provide a strong foundation for comparative research aimed at predicting carbon emissions in maritime freight. This thesis aims to train machine learning models that can accurately predict sea freight carbon emissions using raw AIS data for cargo and tanker vessels in Sweden, Finland, Estonia, Latvia, Lithuania and Poland. By comparing the performance of these models, this thesis will provide useful information, insights and patterns into trends in sea freight carbon emissions in these countries. The findings of this study can have a significant impact on maritime freight by improving the accuracy of carbon emissions prediction. This in turn might serve as a basis for better practices to be followed in the maritime industry in these regions.

The methodology of this research involves the experimentation approach which includes machine learning approaches to estimate marine freight carbon emissions in a structured manner. We used Python as the programming language for this research. First, a comprehensive collection of data from several sources is conducted, covering operational parameters, ambient conditions, vessel features, and time-related aspects. After feature engineering to select relevant variables impacting emissions, data preprocessing guarantees dataset accuracy and suitability for analysis. The process of training machine learning models, such as decision trees, SVMs, and random forests, with parameters changed for better performance, involves employing the engineered data. During testing and validation, performance metrics including MSE, RMSE, MAE, and R^2 score are used to evaluate how predictive the models are.

We will compare three machine learning algorithms — decision trees, SVMs, and random forests. To answer research question 1 [RQ1] using the quantitative and qualitative research techniques as the scientific method. Every algorithm will be put into practice and assessed based on how well it predicts carbon emissions from shipping. Each algorithm's efficiency will be evaluated quantitatively using performance metrics like MSE, RMSE, MAE, and R^2 score. To place the study problem within the framework of the body of current understanding, an in-depth review of the literature was done. Understanding the nuances of the problem domain, identifying prior research efforts, and evaluating other scholars' techniques were the goals of this qualitative research component. Understanding the methods, tactics, and conclusions of earlier research was possible through the literature review, which offered insightful background information that helped to frame the research gap and point out areas in need of more study. The models will be evaluated objectively by testing their ability to predict new instances accurately on unseen data after being trained on a subset of the original data.

To predict maritime freight emissions, research question 2 [RQ2] compares the performance of three machine learning algorithms: random forests, SVMs, and decision trees. This statistical methodology is used within the scientific method. MSE, RMSE, MAE, and R^2 score are the main metrics for performance that are evaluated to give an in-depth evaluation of expected outcomes and model efficiency. To maintain fairness, the assessment employs an identical test dataset, exposing the advantages and disadvantages of every algorithm in terms of emissions prediction using measures including MSE, RMSE, MAE, and R^2 .

Figure 4.1 explains the experimentation procedure of the research in a flowchart.

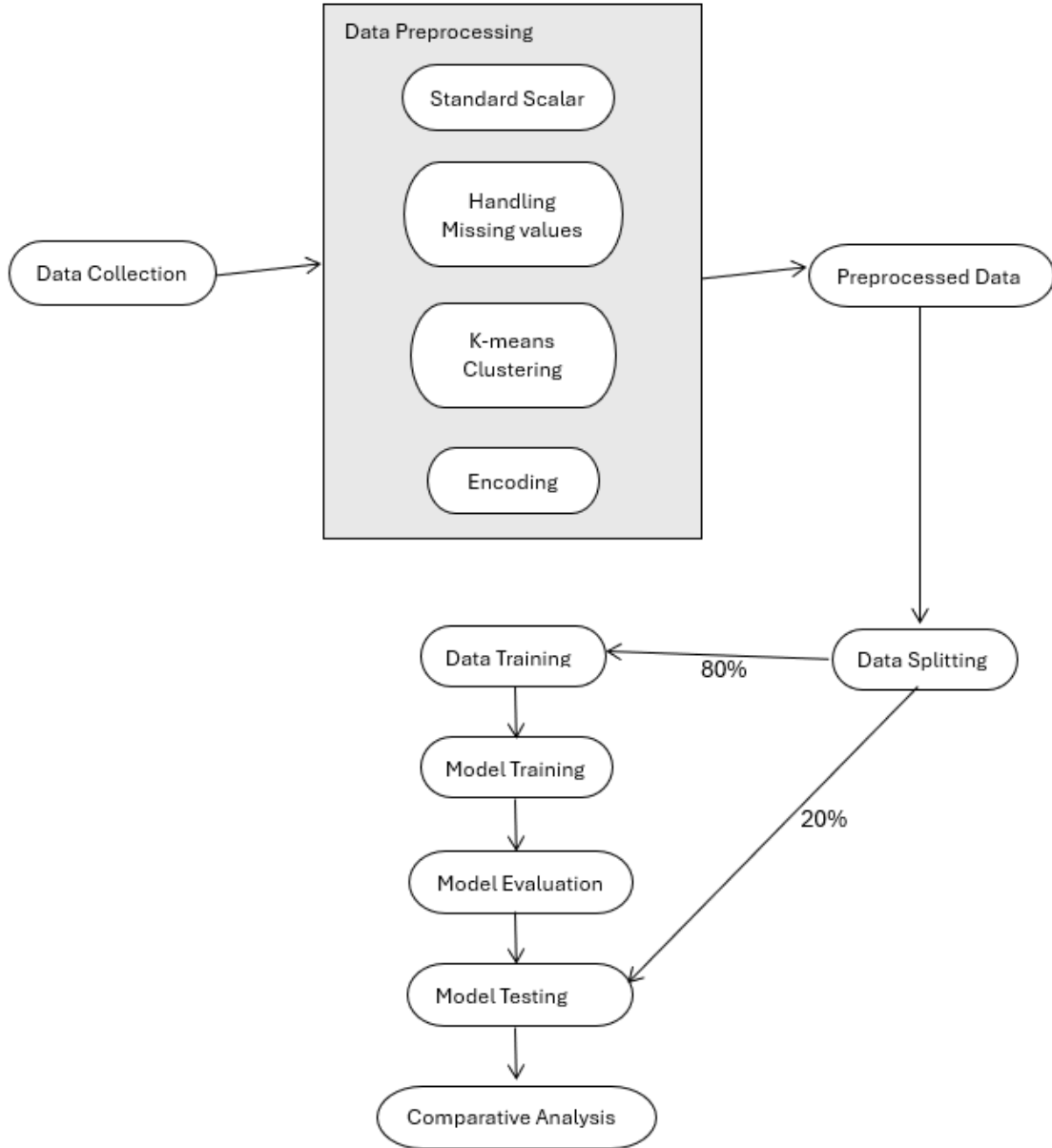
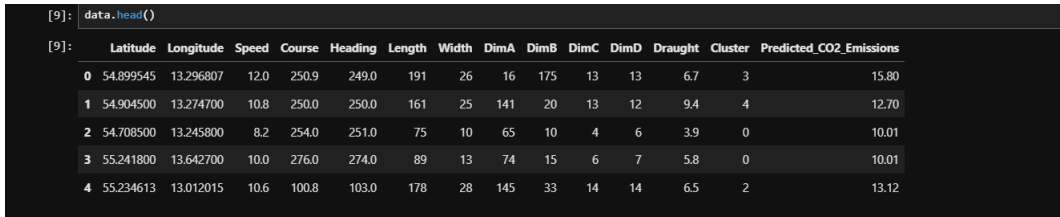


Figure 4.1: Methodology showing the different phases of the experimental procedure followed

4.1 Data Collection:

In data collection, we discuss the details of the data that we used in the research. The dataset used for this research is a combination of Automatic Identification System (AIS) data [1] obtained from satellite and terrestrial sources, offering a thorough understanding of maritime activities. The first section of the dataset covers port

calls that take place in certain countries, such as Sweden, Finland, Estonia, Latvia, Lithuania, and Poland, between January 1, 2021, and April 31, 2022. It is an unlabeled dataset that includes cargo and tanker ships longer than 65 meters and provides comprehensive information including port ID, name, LOCODE, MMSI, IMO, vessel name, destination, type, and arrival and departure timestamps attributes. We handled missing values to guarantee accuracy and preserve the dataset's integrity, this was done by replacing NaN, blank or missing values with techniques such as replacing with the median value, deleting duplicate rows or columns with missing values or advanced methods such as KMeans clustering. This clustering function was used on the data to find underlying patterns and to sort related data points. Then we created a streamline process for predicting emissions. Using the pipeline standard scalar we standardize the features and then fit in a linear regression to ensure that our new dataset is consistent and efficient. The dataset's second section displays historical AIS data within predefined geographic coordinates, including latitudes between 54.5 and 55.4 degrees and longitudes between 13.0 and 13.5 degrees. It covers the same period and comprises cargo and tanker vessels longer than 65 meters, much like the original dataset. Figure 4.2 is the overview of the considered dataset.



	Latitude	Longitude	Speed	Course	Heading	Length	Width	DimA	DimB	DimC	DimD	Draught	Cluster	Predicted_CO2_Emissions
0	54.899545	13.296807	12.0	250.9	249.0	191	26	16	175	13	13	6.7	3	15.80
1	54.904500	13.274700	10.8	250.0	250.0	161	25	141	20	13	12	9.4	4	12.70
2	54.708500	13.245800	8.2	254.0	251.0	75	10	65	10	4	6	3.9	0	10.01
3	55.241800	13.642700	10.0	276.0	274.0	89	13	74	15	6	7	5.8	0	10.01
4	55.234613	13.012015	10.6	100.8	103.0	178	28	145	33	14	14	6.5	2	13.12

Figure 4.2: Overview of the dataset

Important attributes like latitude, longitude, speed, course, heading, length, width, dimA, dimB, dimC, and draught are included in the 10-minute interval data samples within both segments. With a huge number of rows, the dataset offers a wealth of information for study. With the use of these characteristics, comprehensive examination and prediction of maritime freight emissions are made possible by providing a detailed perspective of the movements, characteristics, and navigational details of the vessel.

4.2 Data Preprocessing:

The dataset was preprocessed before analysis to make sure it was suitable for machine learning modeling and of high quality. As part of this, redundant data was cleaned up to get rid of any unnecessary or duplicate entries. The dataset was kept intact by handling missing values by substituting them with the corresponding feature mean.

Standard scalar: To enhance model performance and lower computational complexity, redundant or unnecessary features were eliminated from the dataset. A Standard Scaler was used in data cleaning to get a standardized distribution, with a zero mean and standard deviation of one. It standardizes features by subtracting the

mean value from the feature and then dividing the result by the feature standard deviation. Figure 4.3 explains the standard scalar employed in the data preprocessing process.

```
[10]: scaler = StandardScaler()
      X_scaled = scaler.fit_transform(dt)
```

Figure 4.3: Standard scaling

Handling missing values: To guarantee accuracy and preserve the dataset's integrity, missing values were substituted with the mean of the corresponding feature. This step involves addressing missing values in the dataset, which is crucial for ensuring the quality and integrity of the data. The missing values can be in different forms such as Nan, Blank values, null, or Continuous zeroes. This can be handled by using techniques such as imputation (e.g., replacing missing values with the mean, median, or mode of the feature), deletion of rows or columns with missing values, or advanced imputation methods such as K-nearest neighbors (KNN) imputation. We replaced any missing values with the mean of the feature. From fig 4.4 we handled all the missing values by replacing them with the mean of the feature.

```
[5]: Latitude           False
      Longitude          False
      Speed             False
      Course            False
      Heading           False
      Length            False
      Width             False
      DimA              False
      DimB              False
      DimC              False
      DimD              False
      Draught           False
      Cluster           False
      Predicted_CO2_Emissions False
      dtype: bool
```

Figure 4.4: Dataset after handling missing values

Clustering: On the given dataset, one data analysis method called K-means clustering was performed to find underlying patterns and place related data points in one group. The given fig 4.5 is a code snippet that explains the clustering performed as K-means clustering was used to divide up vessels in the context of maritime freight data according to their movement patterns, traits, and other factors that were noted in the dataset. K-means clustering divided the dataset into clusters, each of which represents a unique collection of vessels with comparable behaviors or characteristics. These clusters were created by taking into account factors such as vessel type, speed,

course, destination, and arrival time. This clustering technique helps in discovering factors impacting carbon emissions in maritime transportation and offers helpful data on vessel operations and preferred routes.

```
[11]: kmeans = KMeans(n_clusters=5, random_state=42) # Adjust the number of clusters as needed
      cluster_labels = kmeans.fit_predict(X_scaled)
```

Figure 4.5: K-Means clustering

Encoding: A preprocessing procedure called "categorical encoding" was used to translate categorical data into a numerical representation that machine learning algorithms can understand. The encoding code snippet was mentioned in the figure 4.6. This categorical encoding involves converting categorical variables into a numerical format that can be used by machine learning algorithms. Common encoding techniques include one-hot encoding, label encoding, and target encoding. This step is essential because most machine learning algorithms require numerical input data. Categorical variables must be encoded for our machine learning models to properly use numerical input, which is how these models are trained. In this research, the widely known method of one-hot encoding is employed to encode categorical values. It consists of setting up binary dummy variables inside the categorical variable for every category. A binary column with a value of 1 denoting the presence of a category and a value of 0 denoting its absence is used to represent each category.

The categorical variable "Cluster" will be in a format that was used as input for decision trees, support vector models, and random forests after one-hot encoding. Subsequently, these algorithms can acquire knowledge from the encoded categorical data as well as numerical characteristics such as dimension ratios and other input factors. This allows for equitable comparisons between various machine learning models and facilitates precise predicts of cargo emissions.

```
if 'Cluster' in data.columns:
    data = pd.get_dummies(data, columns=['Cluster'], prefix='Cluster')
```



```
data.head()
```

	Latitude	Longitude	Speed	Course	Heading	Length	Width	DimA	DimB	DimC	...	Length to Width Ratio	Length to Draught Ratio	Width to Draught Ratio	Speed times Draught
0	54.899545	13.296807	12.0	250.9	249.0	191	26	16	175	13	...	7.346154	28.507463	3.880597	80.40
1	54.904500	13.274700	10.8	250.0	250.0	161	25	141	20	13	...	6.440000	17.127660	2.659574	101.52
2	54.708500	13.245800	8.2	254.0	251.0	75	10	65	10	4	...	7.500000	19.230769	2.564103	31.98
3	55.241800	13.642700	10.0	276.0	274.0	89	13	74	15	6	...	6.846154	15.344828	2.241379	58.00
4	55.234613	13.012015	10.6	100.8	103.0	178	28	145	33	14	...	6.357143	27.384615	4.307692	68.90

Figure 4.6: Categorical Encoding on dataset attributes

4.3 Feature selection and Dimensionality reduction:

Visualizing Numerical Feature Distribution using Pairplot: One of the popular visualization methods for showing the pairwise correlations between numerical features in a dataset is pair plot. Every variable is associated with every other variable in a grid of scatterplots produced by it. It frequently displays histograms or estimations of the kernel densities for each variable along the diagonal. We can visually examine the correlations between every pair of variables in your dataset, such as the linear relationship between emissions and speed, using the pair plot. We can identify any possible patterns, trends, or correlations between the input features and the estimated CO2 emissions by looking at the scatterplots. As shown below in fig 4.7 this graphic visualization can provide insight into the factors that influence CO2 emissions and how they interact with one another. Additionally, it can help in locating any anomalies or outliers in the data that would require additional research. All things considered, the pairplot is a useful tool for learning more about the connections in your dataset, which can help you when comparing various machine-learning models.

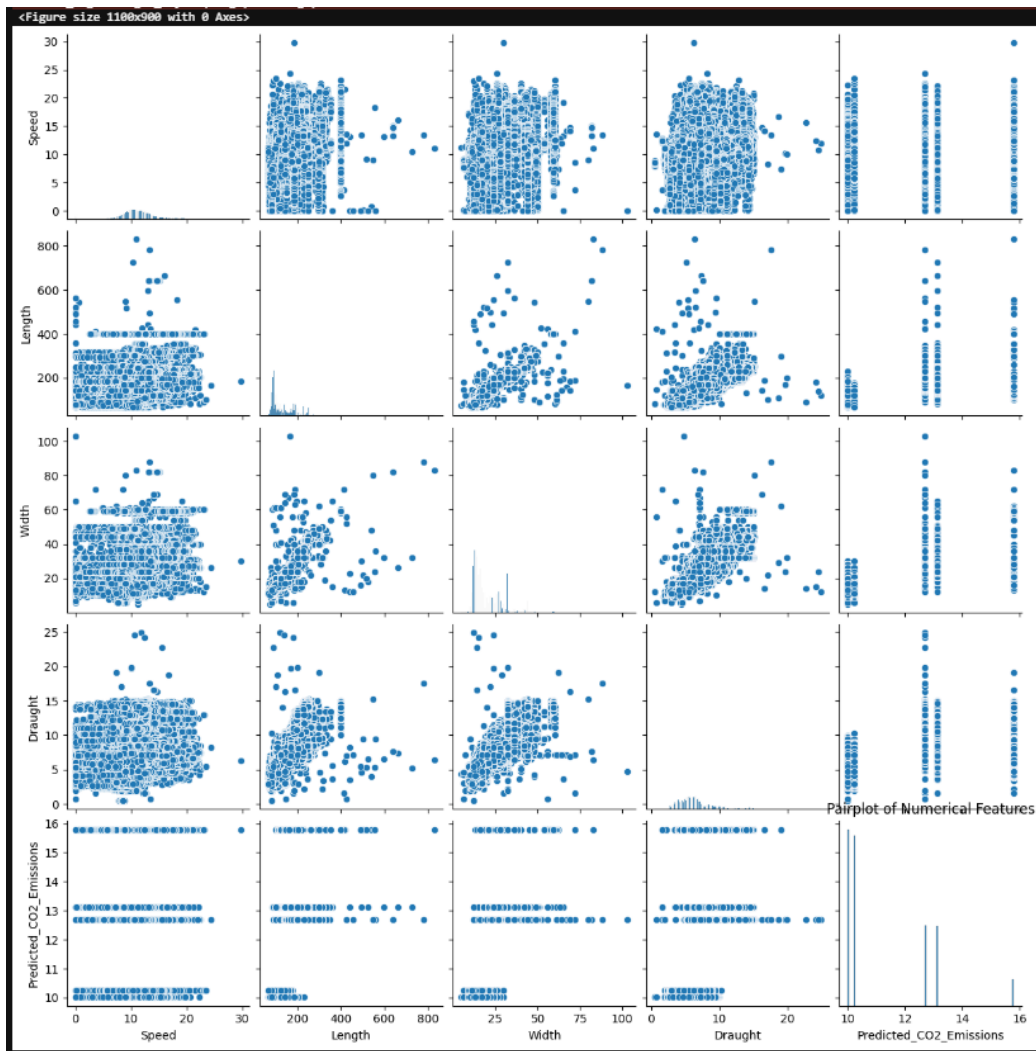


Figure 4.7: Distribution of numerical features

Correlation between Numerical Features using Heatmap: A heatmap is a data visualization where colors are used to represent the values in a matrix. When patterns in any type of data need to be observed. Heatmaps are frequently used to depict correlation matrices, cluster analysis, and other related data. The heatmap is a useful tool for investigating visually the correlations between the predicted CO2 emissions and the input features namely speed, and size. We can determine which input features has the best link with expected emissions and comprehend how changes to these features might impact emissions levels by examining the heatmap. Insights from this research can be very helpful in maximizing the movement of freight while reducing its environmental impact. The matrix of correlations between numerical features is displayed using a heatmap. The linear link between two variables is measured using correlation to determine its strength and direction. Finding strongly correlated features can aid in feature engineering or feature selection by reducing redundancy and enhancing model performance. The attributes we use in this heatmap are Speed, Length, Width, Draught, and predicted CO2 emissions on a scale of 0.4 to 1 as shown below in fig 4.8.

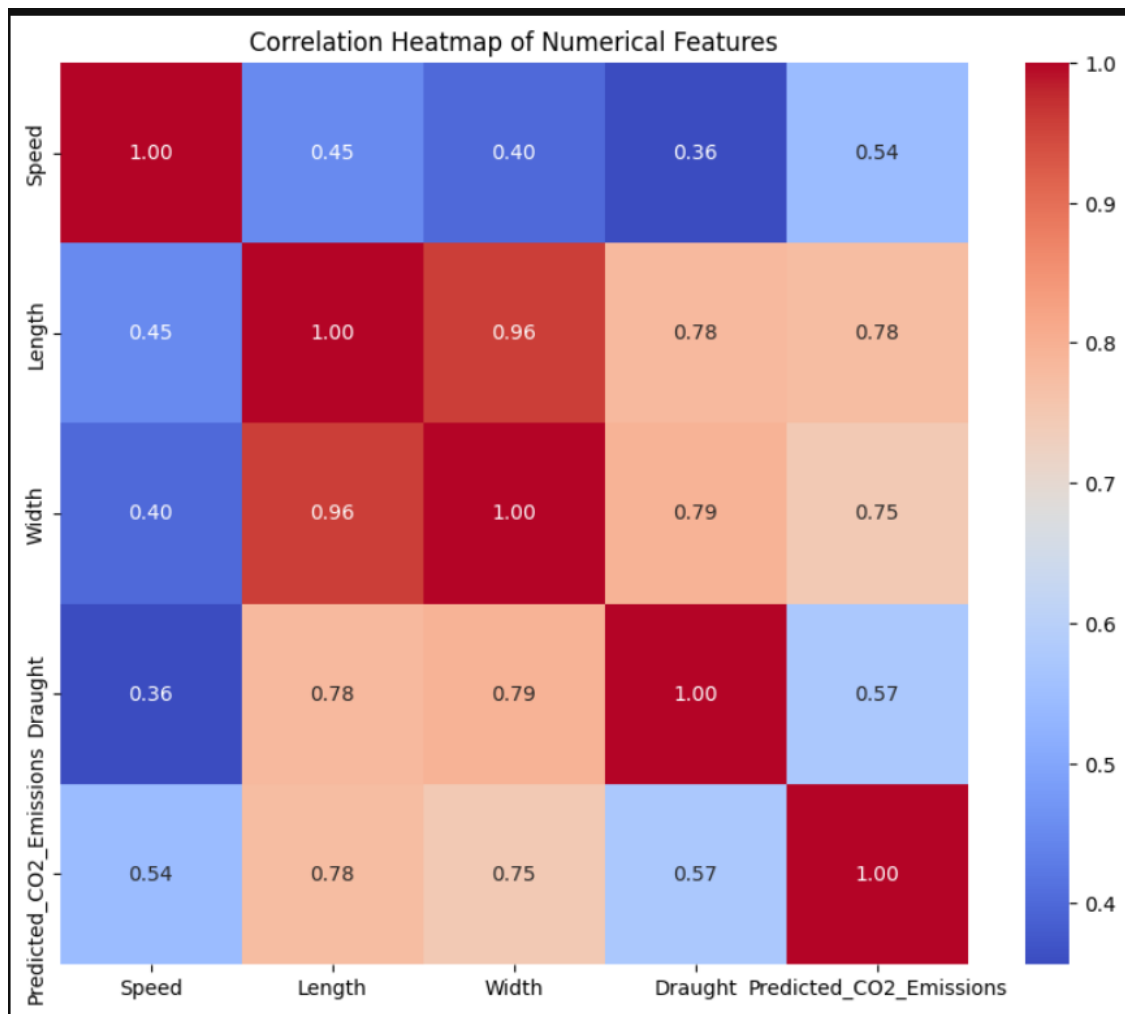


Figure 4.8: Correlation of Numerical features

Calculating Dimension Ratios: Dimension ratios refer to the ratios between different dimensions or aspects of features. For example, in geometric shapes, dimension ratios can represent ratios between lengths, widths, heights, etc. Deriving new features from the physical dimensions of cargo vessels that already exist is a step in the dimension ratio calculation process. Feature engineering, which creates new features from preexisting ones to provide extra data that could be helpful for predictive modeling, is involved in calculating these dimension ratios. These attributes may have an impact on their emissions and performance. By adding these ratios to the initial physical dimensions, you provide the models with additional detailed knowledge about the vessels, which may enhance their ability to anticipate.

4.4 Model Training:

We proceeded with regression because our objective is to predict carbon emissions in maritime shipping, which involves predicting a continuous numeric value (i.e., the amount of carbon emissions). Regression analysis is specifically designed for predicting continuous outcomes based on input variables, making it suitable for our research. Carbon emissions in maritime shipping are measured as continuous quantities, typically in metric tons or kilograms. Regression analysis allows you to model and predict these continuous values. Regression models are well-suited for predicting continuous outcomes with reasonable accuracy. By training regression models on historical data, you can make predictions about future carbon emissions based on current or anticipated ship characteristics and operational conditions. Train Support Vector Regressor, decision tree Regressor, and Random Forest Regressor models on the training data using default hyperparameters.

Support Vector Regressor (SVM): The SVM model was initialized with a radial basis function (RBF) kernel, which is commonly used for non-linear regression tasks. The training data were used to fit the model, optimizing the hyperparameters for improved performance. Hyperparameters such as the regularization parameter (C) and kernel coefficient (gamma) were tuned using techniques like grid search or randomized search to optimize model performance. For regression tasks, the SVM method is modified and called SVR, to predict CO₂ emissions based on certain features. The model is initialized with a radial basis function (kernel='rbf') and trained on the training data (X-train, y-train). After training, predictions are made on the test data (X-test). It works especially well with datasets that have non-linear patterns and complex interactions.

decision trees: decision trees were initialized without any specific hyperparameters to allow for exploration of the model's default behavior. The training data were used to train the decision tree model, which recursively partitions the feature space to make predictions. A decision tree regressor is instantiated and trained on the training data (X-train, y-train). The trained model then predicts the CO₂ emissions for the test data (X-test). The accuracy of the decision tree model is computed by comparing its predictions to the true values (y-test). However, there seems to be a typo in the code where the regression report is called for y-predict, which is undefined. It

should be replaced with dt-predictions for the regression report. decision trees were selected for their simplicity and interpretability, allowing for easy understanding of the decision-making process.

Random Forest: A Random Forest regressor was employed, consisting of an ensemble of decision trees. The number of trees in the forest was set to 100, and other hyperparameters such as maximum depth and minimum samples per leaf were optimized during model training. A Random Forest regressor is created with 100 decision trees (n-estimators=100) and trained on the training data. Similar to the decision tree model, predictions are made on the test data, and the accuracy of the Random Forest model is evaluated. random forests were chosen for their ability to handle high-dimensional data and mitigate overfitting.

4.5 Model Evaluation:

After training each model, performance evaluation was conducted using appropriate regressor evaluation metrics. At first, we wanted to perform Accuracy Metrcision such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared score (R^2) computed to assess the performance of SVM, decision trees, and random forests in predicting CO2 emissions. MSE is a popular metric for assessing how well regression models perform. The mean squared errors between the expected and actual values are computed. Greater MSE values signify greater discrepancies between the observed data and the predicts. As it implies lesser variances between expected and actual values, a lower MSE denotes higher model performance.

$$(1/n) * \sum (y_i - \hat{y}_i)^2$$

Figure 4.9: Formula for Mean Square Error

Where n is the number of observations, cap(y_i) is the model's predicted value, and y_i is the actual value which is the y-test.

By taking the square root of the mean squared errors between the expected and actual values, RMSE is a variation of MSE. It is expressed in the same units as the target variable and shows the standard deviation of the prediction errors. Better model performance, with fewer differences between predicted and actual values, is indicated by a lower RMSE, similar to the MSE. Because RMSE is expressed in the same units as the target variable, it offers a more understandable way to quantify prediction error. As with MSE and RMSE, a lower MAE indicates better model performance, with smaller deviations between predicted and actual values. MAE provides a straightforward measure of the average prediction error.

An additional measure for assessing regression models is the MAE. The mean of the absolute discrepancies between the expected and actual values is computed. Com-

$$\sqrt{(\text{MSE})}$$

Figure 4.10: Formula for Root Mean Square Error

pared to MSE and RMSE, MAE is less susceptible to outliers. A lower MAE denotes better model performance, with smaller discrepancies between predicted and actual values, similar to MSE and RMSE. The Mean average prediction error (MAPE) is measured in an easy-to-understand manner.

$$(1/n) * \sum |y_i - \hat{y}_i|$$

Figure 4.11: Formula for Mean Absolute Error

A statistical metric known as R-squared (R^2) quantifies the percentage of the target variable's variance that the regression model accounts for. Higher values indicate better model fit; the range is 0 to 1. A regression model is said to be well-fitted if its R^2 value is closer to 1, which denotes that it accounts for a significant amount of the variance in the target variable. On the other hand, a model with an R^2 value that is closer to 0 indicates poor fit and little variance explanation.

$$R^2 = 1 - (\text{SS}_{\text{res}} / \text{SS}_{\text{tot}})$$

Figure 4.12: Formula for R-Squared Score

, where SS tot is the total sum of squares and SS res is the sum of squared residuals.

The Support Vector Regressor, decision tree Regressor, and Random Forest Regressor models' performances are quantified by means of these assessment metrics. We can determine which model performs best in terms of prediction accuracy and model fit by comparing these metrics across various models.

4.6 Model Testing

We are using cross-validation where we split the dataset into minor parts and when just a small dataset is available for training, a specific technique called cross-validation is employed for model testing, especially for assessing the performance of a model on new, or unknown data. Using various combinations of training and validation data, the model is trained and assessed numerous times via cross-validation, which divides the available dataset into multiple subsets (folds). This makes it possible to estimate the model's performance more accurately and lessens the impact of randomness and dataset variability. Each regression model's performance was determined by analyzing the outcomes of model testing and cross-validation. The assessment metrics of several models were compared, and their variability was examined, to obtain an understanding of the advantages and disadvantages of each strategy.

Chapter 5

Results and Analysis

The results and analysis of our comparative study on cargo emissions utilizing the Support Vector Regression, decision tree Regressor, and Random Forest Regressor regression models are presented in this section. The study included preprocessing the dataset, and feature engineering, which included calculating dimension ratios, choosing a model, and evaluating it using performance measures like R-squared Score (R^2 Score), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Squared Error (MSE).

We can confirm the results in our thesis using a combination of model testing and cross-validation techniques.

First, we use model testing, wherein the regression models are trained on a subset of the dataset and their performance is assessed on a different, held-out subset. By going through this procedure, we can evaluate how effectively the models estimate carbon emissions in maritime shipping and generalize to new data. To measure each model's performance, we use a variety of assessment metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) score.

Furthermore, we use cross-validation as a reliable method of validating results, especially in cases where dataset sizes are constrained. We acquire more accurate estimations of the model's performance metrics by using k-fold cross-validation, where the dataset is divided into k equal-sized folds, each of which is utilized as a test set while the model is trained on the remaining data. This iterative procedure offers a thorough evaluation of the models' capacity to generalize and precisely predict carbon emissions across various data subsets.

Therefore, we have developed a rigorous and methodical validation process that makes use of cross-validation and model-testing strategies to guarantee the validity and dependability of our study findings when it comes to estimating carbon emissions in maritime shipping.

The regression metrics of the test set for the SVM regressor are shown in Figure 5.1. While the validation test's MSE score is 53578.43 and the R^2 -score is 0.65, the test's MSE score is 238.14.

```
SVM Model Evaluation :  
Mean Square Error (MSE): 53578.43  
Root Mean Square Error (RMSE): 238.14  
Mean Absolute Error (MAE): 119.7  
R-squared score: 0.65
```

Figure 5.1: SVM metrics showing highest MSE, RMSE and MAE values

The regression metrics of the test set for the random forest regressor are shown in Figure 5.2. While the validation test's MSE score is 35785.41 and the R^2 -score is 0.78, and the test's MSE score is 48.5 which suggests that it is an accurate model.

```
Random Forest Model Evaluation :  
Mean Square Error (MSE): 35785.41  
Root Mean Square Error (RMSE): 159.65  
Mean Absolute Error (MAE): 48.5  
R-squared score: 0.78
```

Figure 5.2: Random Forest regressor metrics showing highest accuracy when compared to SVM and decision tree

The regression metrics of the test set for the decision tree regressor are shown in Figure 5.3. While the validation test's MSE score is 40744.35 and the R^2 -score is 0.72, the test's MSE score is 81.4 which is approximately accurate concerning each other.

```
Decision Tree Model Evaluation :  
Mean Square Error (MSE): 40744.35  
Root Mean Square Error (RMSE): 162.97  
Mean Absolute Error (MAE): 81.4  
R-squared score: 0.72
```

Figure 5.3: decision tree regressor performs decently with lowest training time

5.1 Analysis

It is clear from the results that the Random Forest Regressor performed better on all criteria than the SVR and decision tree Regressor. Its MSE, RMSE, and MAE values were the lowest, indicating better prediction accuracy and reduced error rates. Furthermore, when compared to the other models, the Random Forest Regressor had the greatest R²-Score, indicating a better match to the data.

There are various reasons for the observed variations in performance between the regression models. The improved predictive performance of the Random Forest Regressor can probably be attributed to its ensemble method, which makes use of numerous decision trees and feature randomness. The high dimensionality and non-linearities of the dataset may have presented challenges for SVR, despite its ability to capture intricate correlations in the data. Because the decision tree Regressor tended to overfit the training set, it performed worse than Random Forest in terms of error rates, even if it was easier to understand and use.

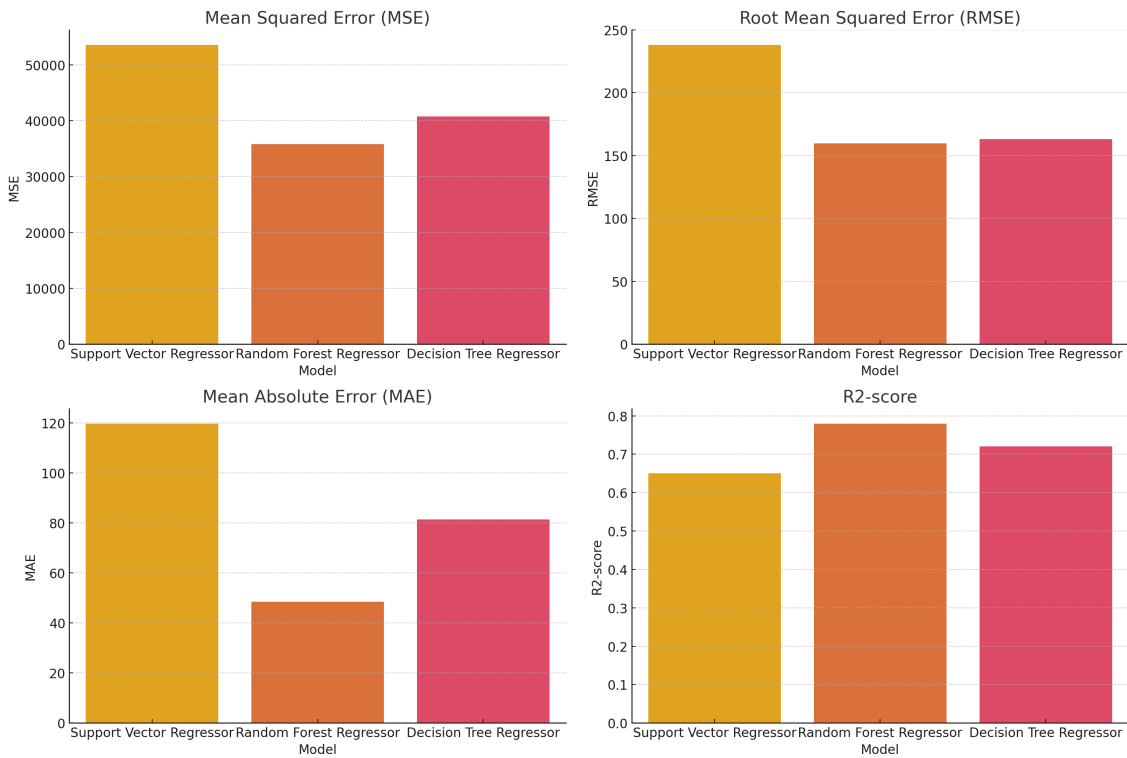


Figure 5.4: Visualization of MAE, MSE, R2-score and MSE for each trained model

The research results as shown in figure 5.4 showcase the summary of evaluation metrics on the test set and highlight how crucial it is to choose the right regression model to predict cargo emissions accurately. It is the graphical representation of results acquired after verifying with model testing and cross-validation.

Models	MSE	RMSE	MAE	R ²
Support vector Regressor	53578.43	238.14	119.7	0.65
Random Forest Regressor	35785.41	159.65	48.5	0.78
decision tree Regressor	40744.35	162.97	81.4	0.72

Table 5.1: Evaluation metrics on test set

A thorough comparison of the three machine learning models—Support Vector Regressor (SVR), Random Forest Regressor, and decision tree Regressor—based on four important regression metrics is given in the tabular 5.1 as given above.

These metrics indicate that although SVR can predict quite well, its errors are larger than those of the other models. It is clear from the lower error metrics and higher R² that the Random Forest Regressor performs better at accurately estimating CO₂ emissions. The Random Forest Regressor outperforms decision tree Regressors in terms of accuracy, despite its superior performance over SVR. Based on the AIS data, the Random Forest Regressor proved to be the most effective model for estimating CO₂ emissions in marine shipping, as it exhibited superior performance in all criteria. The Random Forest Regressor’s capacity to handle big datasets and capture intricate feature interactions is responsible for its consistently good performance.

We trained the model using supervised machine learning techniques such as support vector regressor, decision tree regressor, and random forest regressor. Before doing the training, we extracted characteristics to determine which aspects are essential for predicting carbon emissions. The Standard scalar has been utilized for feature scaling, and vessel clusters have been identified using K-means clustering. We have trained the model on three distinct supervised algorithms based on the features that we have chosen, allowing us to predict the model's accuracy. Finally model testing and cross-validation to verify the results.

RQ1: When training each learning model on sea freight carbon emissions dataset, which algorithm among decision trees, SVMs, and random forests results in the highest effectiveness in predicting carbon emissions in maritime shipping when considering AIS data?

The highly effective model in predicting carbon emissions was the Random Forest Regressor, which strikes a 0.78 in R^2 Score on a scale of 0 to 1 as shown in figure 5.2, compromising between model complexity and predict accuracy. Based on the evaluation measures, the random forest, SVM, and decision tree models' performances are contrasted. When it comes to projecting carbon emissions, the model with the lowest MSE, RMSE, and MAE values as well as the greatest R^2 Score is said to be the most accurate. Finding the best algorithm and important contributing elements provides insightful information for improving sustainability initiatives in the maritime shipping sector. Precise estimation of carbon emissions aids in well-informed decision-making and permits focused actions to mitigate ecological consequences. To further understand emissions predicting for cargo vessels and investigate new features and modeling approaches that may improve prediction performance, more research is necessary.

An essential aspect of the shipping industry's sustainability initiatives is the ability of machine learning algorithms to predict carbon emissions in maritime shipping. In this work, using the dataset from the Automatic Identification System (AIS), we examine the performance of three carefully chosen machine learning algorithms in estimating carbon emissions: decision trees, SVMs, and random forests. Finding the emission prediction algorithm with the maximum efficiency and studying the elements influencing better performance are the main goals.

RQ2: When assessing the trained machine learning algorithms, how is each algorithm performing on metrics like MSE, RMSE, MAE, and R^2 -score?

Our goal in this study is to compare the created machine learning algorithms in terms of performance measures like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared Score (R^2). Gaining insight into the advantages and disadvantages of each method in estimating carbon emissions in maritime shipping based on Automatic Identification System (AIS) data requires an understanding of these variations.

Standard regression performance indicators are used to assess models for estimating carbon emissions, such as decision trees, SVMs, and random forests. To evaluate how well the models work on untested data, the dataset is divided into training and testing sets. For every model, we compute its MSE, RMSE, MAE, and R^2 using its predictions on the testing set. Variations in the algorithms' performance measures are examined and interpreted to determine the underlying causes.

The performance metrics (MSE, RMSE, MAE, and R^2) for each machine learning method are compiled in Table 5.1. Performance indicators vary throughout decision trees, SVMs, and random forests. When compared to decision trees and SVM, random forest regularly exhibits lower MSE, RMSE, and MAE, demonstrating its higher prediction accuracy. Additionally, Random Forest shows higher R^2 values, indicating a better fit between the model and the data.

Achieving precise predictions of carbon emissions in maritime transportation requires careful consideration of algorithm selection, as evidenced by the differences in performance metrics across decision trees, SVMs, and random forests. By helping industry stakeholders implement efficient predictive models, the study advances knowledge of machine learning algorithm performance in marine sustainability applications. In light of the results, we advise using Random Forest to predict carbon emissions in maritime shipping with high accuracy and dependability using AIS data.

6.1 Reflections:

The research gap regarding a comprehensive comparison and evaluation of machine learning techniques for predicting carbon emissions in marine freight has been significantly filled by this thesis. While several techniques for predicting emissions in container ships have been studied in the past, few studies specifically assess the performance of machine learning models in this field. The most efficient technique for calculating carbon emissions using AIS data is determined by methodically contrasting decision trees, SVMs, and random forests.

The results show that random forests estimate carbon emissions better than decision trees and SVMs, with the lowest values for MSE, RMSE, and MAE and the greatest R^2 score of 0.78. This highlights the Random Forest's capacity to identify

complex patterns and relationships within the data and shows a superior balance between model complexity and predictability. This understanding is significant because it suggests that random forests can be applied successfully in real-world scenarios to produce accurate emissions predictions, facilitating better decision-making and focused efforts to reduce the environmental impact of maritime freight.

The thesis also investigates the characteristics that lead to better performance, including the algorithm's robustness against overfitting and ability to handle high-dimensional data. This analysis clarifies the reasons why random forests perform better in this situation and supports its use over alternative techniques. To guarantee that the models are trained and tested on well-prepared data, the study implements a comprehensive evaluation approach that includes feature scaling with the Standard Scalar and vessel clustering using K-means. This careful methodology serves as further evidence of the accuracy of the results and the value of the findings.

Furthermore, comparing performance metrics between several models provides valuable insight into the advantages and disadvantages of every approach. Industry participants hoping to use machine learning methods for sustainability projects and emissions prediction will find this information useful. It closes the research gap with useful suggestions and practical conclusions by highlighting the significance of choosing the appropriate algorithm based on the particular requirements of the application. In conclusion, by proving the accuracy of machine learning algorithms in estimating carbon emissions, this thesis not only fills the indicated research gap but also advances the subject of marine sustainability. The study's findings offer possibilities for research into hybrid models and other features that could boost prediction accuracy even more and improve maritime logistics' environmental management over time.

7.1 Conclusion

Finally, for the objective of predicting carbon emissions from maritime freight, this research provides a comprehensive comparison of machine learning techniques, particularly decision trees, SVMs, and random forests. The goal of the study was to address the pressing need for trustworthy emission prediction models in the transportation sector, especially for maritime logistics, to support carbon mitigation measures and allow well-informed decision-making. An extensive analysis of Automatic Identification System (AIS) data comprising a range of factors impacting freight emissions in maritime logistics has been carried out in this research project. With this careful evaluation, the research has successfully clarified the unique advantages and disadvantages of each machine learning method used.

First of all, decision trees' basic simplicity and interpretability have been highlighted. decision trees attract attention because of these features, which provide a clear understanding of the model's decision-making process. However, the research has highlighted a significant disadvantage of decision trees, which is their vulnerability to overfitting. Regularization methods must be put in place to mitigate this behavior, which causes the model to identify noise in the data instead of true patterns, hence reducing the influence on prediction accuracy. Second, the flexibility of SVMs in managing high-dimensional data has been demonstrated. SVMs are particularly useful for collecting complicated patterns in maritime freight data because they perform well in situations where the feature space is large and complex. But the study has also highlighted a potential problem, especially when dealing with big datasets. Due to the algorithm's increasing computational cost as dataset sizes increase, SVM scalability may have practical drawbacks.

Finally, using ensemble learning approaches to achieve competitive accuracy and generalization performance, random forests have emerged as a potential method. random forests capture challenging correlations in the data while reducing the danger of overfitting associated with individual decision trees by combining the predictions of several trees. The utilization of an ensemble strategy results in enhanced performance in carbon emission prediction tasks by increasing robustness against noise and unpredictability in the dataset, in addition to improving predictive accuracy. Overall, the research's comprehensive evaluation provides significant clarity on the advantages and disadvantages of using decision trees, SVMs, and random forests

to predict carbon emissions from maritime freight. However, for estimating carbon emissions from maritime freight, the Random Forest algorithm is thought to be the most appropriate. The technique of ensemble learning combines the advantages of many decision trees, resulting in competitive accuracy, resilience against overfitting, and versatility with intricate datasets. random forests provide the optimum blend of accuracy and durability in this situation, even though decision trees and SVM have advantages of their own.

7.2 Future work

The research can be further developed in following ways:

- Later research may pursue deeper feature engineering tactics, investigating innovative approaches to derive significant insights from marine data. Model performance may be enhanced and hidden patterns may be found using methods including feature interaction engineering, spatial clustering, and temporal aggregation.
- Algorithms could be made more predictive by refining them further with advanced optimization techniques and algorithmic improvements. This involves studying improved regularization strategies for decision trees, scalable SVM implementations for huge datasets, and ensemble approaches that go beyond random forests.
- A deeper knowledge of emission dynamics in marine logistics could be obtained by incorporating additional external data sources, such as meteorological data, port traffic statistics, and operational characteristics of the vessels, into the prediction models.
- It may be possible to increase the practical utility of adaptive prediction models by creating ones that can dynamically adapt in real time to shifting operational and environmental situations. Investigating adaptive model architectures and online learning methods to handle changing data streams is part of this.
- Creating real-time prediction models with the ability to monitor and predict carbon emissions from maritime freight could have a significant impact on how the transportation sector makes decisions and approaches emissions reduction.

References

- [1] “AIV_bth.SE_flytnow - Google Drive.” [Online]. Available: <https://drive.google.com/drive/folders/1x45VUmqEOERXg72d1N-VojuvE5K5Zx0b>
- [2] “Freight Transportation | MIT Climate Portal.” [Online]. Available: <https://climate.mit.edu/explainers/freight-transportation>
- [3] “How Much Does the Shipping Industry Contribute to Global CO2 Emissions?” section: GHG Emissions. [Online]. Available: <https://sinay.ai/en/how-much-does-the-shipping-industry-contribute-to-global-co2-emissions/>
- [4] “Maritime Artificial Intelligence & Machine Learning: Ultimate Guide.” [Online]. Available: <https://spire.com/maritime/maritime-artificial-intelligence-and-machine-learning/>
- [5] “Support Vector Machine (SVM) Algorithm - Javatpoint.” [Online]. Available: <https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm>
- [6] “Top Performance Metrics in Machine Learning: A Comprehensive Guide.” [Online]. Available: <https://www.v7labs.com/blog/performance-metrics-in-machine-learning>
- [7] “Cargo Ship CO2 Emissions: Compared With Cars, Planes, Trains,” Jun. 2023, section: Nautical Science. [Online]. Available: <https://maritimepage.com/cargo-ship-co2-emissions/>
- [8] “Regression Metrics,” Oct. 2023, section: Machine Learning. [Online]. Available: <https://www.geeksforgeeks.org/regression-metrics/>
- [9] “Shipping emissions: best tip to reduce emissions+ save money,” Aug. 2023, section: Blog. [Online]. Available: <https://www.container-xchange.com/blog/shipping-emissions/>
- [10] “Random Forest Algorithm in Machine Learning,” Feb. 2024, section: AI-ML-DS. [Online]. Available: <https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/>
- [11] “What is a Decision Tree? | IBM,” Mar. 2024. [Online]. Available: <https://www.ibm.com/topics/decision-trees>
- [12] “What Is Machine Learning? Definition, Types, and Examples,” Mar. 2024. [Online]. Available: <https://www.coursera.org/articles/what-is-machine-learning>
- [13] T. Chu-Van, Z. Ristovski, A. M. Pourkhesalian, T. Rainey, V. Garaniya, R. Abbassi, S. Jahangiri, H. Enshaei, U.-S. Kam, R. Kimball, L. Yang,

- A. Zare, H. Bartlett, and R. J. Brown, "On-board measurements of particle and gaseous emissions from a large cargo vessel at different operating conditions," *Environmental Pollution*, vol. 237, pp. 832–841, Jun. 2018. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0269749117329056>
- [14] T. t. p. g. i. S. a. n. l. f. t. i. g. b. c. o. c. D. t. v. u. cycles and S. C. D. M. u.-t.-D. D. T. R. i. t. Text, "Topic: Shipping emissions worldwide." [Online]. Available: <https://www.statista.com/topics/11288/shipping-emissions-worldwide/>
- [15] Z. Huang, C. Huang, and Z. Wen, "Comparison of Carbon Emission Forecasting in Guangdong Province Based on Multiple Machine Learning Models," in *2022 IEEE 5th International Conference on Knowledge Innovation and Invention (ICKII)*, Jul. 2022, pp. 90–93, iSSN: 2770-4785. [Online]. Available: <https://ieeexplore.ieee.org/document/9983576>
- [16] L. A. Joshy, R. K. Sambandam, D. Vetriveeran, and J. Jenefa, "Regression Analysis using Machine Learning Algorithms to Predict CO2 Emissions," in *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)*, Feb. 2024, pp. 444–448. [Online]. Available: <https://ieeexplore.ieee.org/document/10499094>
- [17] P. Kadam and S. Vijayumar, "Prediction Model: CO2 Emission Using Machine Learning," in *2018 3rd International Conference for Convergence in Technology (I2CT)*, Apr. 2018, pp. 1–3. [Online]. Available: <https://ieeexplore.ieee.org/document/8529498>
- [18] A. Kumar, "Random Forest for prediction," Jun. 2020. [Online]. Available: <https://towardsdatascience.com/random-forest-ca80e56224c1>
- [19] V. Michalakopoulos, L. Ilias, P. Kapsalis, S. Mouzakis, and D. Askounis, "Comparison of Machine Learning Algorithms For Predicting CO2Emissions in the maritime domain," in *2023 14th International Conference on Information, Intelligence, Systems & Applications (IISA)*, Jul. 2023, pp. 1–4. [Online]. Available: <https://ieeexplore.ieee.org/document/10345936>
- [20] H. H. Nguyen, "A Complete View of Decision Trees and SVM in Machine Learning," Jan. 2019. [Online]. Available: <https://towardsdatascience.com/a-complete-view-of-decision-trees-and-svm-in-machine-learning-f9f3d19a337b>
- [21] H. Ritchie and M. Roser, "Cars, planes, trains: where do CO2 emissions from transport come from?" *Our World in Data*, Mar. 2024. [Online]. Available: <https://ourworldindata.org/co2-emissions-from-transport>
- [22] —, "CO emissions," *Our World in Data*, Jan. 2024. [Online]. Available: <https://ourworldindata.org/co2-emissions>
- [23] B. S. Satpute, R. Bharati, and W. P. Rahane, "Predictive Modeling of Vehicle CO2 Emissions Using Machine Learning Techniques: A Comprehensive Analysis of Automotive Attributes," in *2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, Nov. 2023, pp. 511–516. [Online]. Available: <https://ieeexplore.ieee.org/document/10390183>

- [24] H. Tan, “Carbon Emission Prediction with Macroeconomic Variables and Machine Learning,” in *2022 3rd International Conference on Clean and Green Energy Engineering (CGEE)*, Aug. 2022, pp. 52–56. [Online]. Available: <https://ieeexplore.ieee.org/document/9976625>
- [25] E. E. Team, “Carbon Emission Defined & Explained,” Dec. 2022. [Online]. Available: <https://ecolife.com/dictionary/carbon-emission/>
- [26] F. Wong, “Carbon emissions allowances trade amount dynamic prediction based on machine learning,” in *2022 International Conference on Machine Learning and Knowledge Engineering (MLKE)*, Feb. 2022, pp. 115–120. [Online]. Available: <https://ieeexplore.ieee.org/document/9763576>

