CHAPTER 5

PREDICTIVE ANALYSIS

Our predictive analysis aims to predict CO_2 emissions from maritime operations by applying various machine learning models. We focus on utilizing these models to accurately predict emissions based on different operational parameters, such as total time spent at sea, fuel consumption, and emissions during berthing and travel. By doing so, we seek to identify patterns and factors that significantly influence CO_2 emissions in the maritime sector.

Additionally, we aim to evaluate the most effective and accurate machine learning techniques for this prediction task. By comparing the performance of different models, such as Support Vector Machines, Decision Trees, Artificial Neural Networks, and Random Forests, we determine which algorithm provides the most reliable predictions.

➢ **Sub Objectives**

- To predict CO_2 emissions of maritime operations by applying various machine learning models.
- To evaluate the most effective and accurate machine learning technique for predicting CO_2 emissions in the maritime freight industry.

5.1 Exploratory data Analysis

Figure 5:1 depicts the flowchart of the predictive analysis. In the initial phase, we provide a succinct overview of the dataset, summarizing key characteristics essential for contextualizing the problem. Firstly, we received the maritime data from EMSA (2021) for the year 2021. The information in the dataset includes the Annual Time spent at sea [hours] vs Total CO_2 emissions [m tonnes], CO_2 emissions from all voyages between different ports [m tonnes] vs Total CO_2 emissions [m tonnes], CO_2 emissions produced when ships depart from various ports to start their journeys [m tonnes] vs Total CO_2 emissions [m tonnes], CO_2 emissions from all voyages to destination ports [m tonnes] vs Total CO_2 emissions [m tonnes], CO_2 emissions while ships

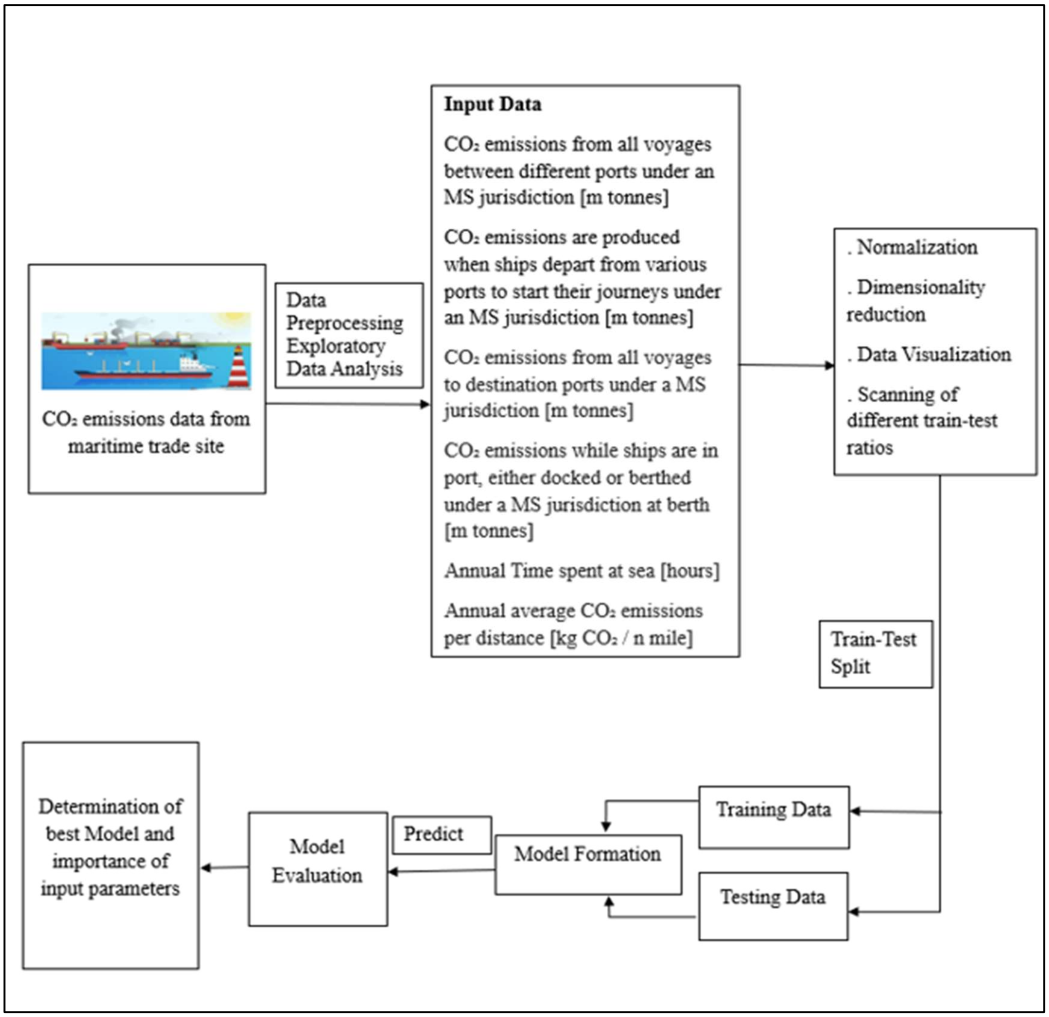

Figure 5:1 Flow chart of ML model construction in the study

are in port, either docked or berthed [m tonnes] vs Total CO_2 emissions [m tonnes]. This analysis plays a pivotal role in data analysis, offering a comprehensive examination, summarization, and elucidation of data to discern key patterns and trends. By employing exploratory analysis, we can detect outliers and discern similarities among variables, enabling us to extract the most significant insights from our dataset. The analysis comprises several essential steps: summarizing statistics, visualizing data, conducting correlation analysis, addressing data cleaning, and missing values, and identifying and eliminating outliers.

Table 5:1 Definitions of parameters used in ML models

Parameter	**Name in ML models**
Technical efficiency	TE
Total fuel consumption [m tonnes]	TFC
Total CO_2 emissions [m tonnes]	TCE
CO_2 emissions from all voyages between different ports under an MS jurisdiction [m tonnes]	CED
CO_2 emissions are produced when ships depart from various ports to start their journeys. under a MS jurisdiction [m tonnes]	CEV
CO_2 emissions from all voyages to destination ports under an MS jurisdiction [m tonnes]	CEVD
CO_2 emissions while ships are in port, either docked or berthed under an MS jurisdiction at berth [m tonnes]	CEDB
Annual average Fuel consumption per distance [kg / n mile]	AAF
Annual average CO_2 emissions per distance [kg CO_2 / n mile]	AAC
Annual Time spent at sea [hours]	ATS

Table 5:2 Description of the data

	TE	TFC	TCE	CED	CEV	CEVD	CEDB	ATS	AAF	AAC
0	31.75	3473.74	11071.53	8143.43	1098.95	946.93	882.22	2346.25	104.75	333.86
1	57.84	722.50	2289.71	89.00	997.00	1108.00	96.00	1324.00	58.50	185.39
5	44.93	2928.66	9389.28	4862.00	1216.52	1617.94	1692.82	2053.08	135.88	435.62
12	11.49	349.50	1111.89	129.62	495.64	449.23	37.39	351.15	94.14	299.51
14	10.78	2899.46	9227.98	5165.07	1539.56	1557.48	965.88	2197.95	109.23	347.64

Table 5:3 Descriptive statistics of the numeric data used in ML models

	TFC	TCE	CED	CEV	CEVD	CEDB	ATS	AAF	AAC
count	7968.000000	7968.000000	7968.000000	7968.000000	7968.000000	7968.000000	7968.000000	7968.000000	7968.000000
mean	3705.168184	11545.302430	2789.846234	3896.097897	4099.790104	756.638312	2476.198341	118.769837	371.308005
std	4207.335919	13009.031293	6255.634982	5073.052558	4966.848455	1118.437391	1417.824005	72.644190	224.890357
min	71.270000	79.830000	0.030000	0.640000	0.960000	3.890000	45.000000	19.230000	1.180000
25%	1348.667500	4242.912500	320.682500	1160.725000	1331.442500	209.825000	1331.250000	75.427500	237.017500
50%	2263.805000	7128.725000	855.015000	2262.110000	2506.220000	425.755000	2171.470000	95.935000	300.995000
75%	4129.007500	12925.617500	2594.315000	4298.325000	4664.992500	926.782500	3428.755000	133.102500	415.460000
max	42759.000000	133812.360000	112005.000000	48664.600000	45226.930000	20224.000000	8036.000000	680.960000	2183.160000

5.2 Data pre-processing

Data pre-processing encompasses the cleaning, transformation, and organization of data to render it suitable for modelling or analysis. This involves a series of steps and techniques applied to raw data before analysis, including the identification and removal of any missing, illogical, ambiguous, or irrelevant data. Once data transformation is complete, data reduction is performed, involving the selection or extraction of features that contribute to predicting the target variable. Heatmaps are often utilized for this purpose.

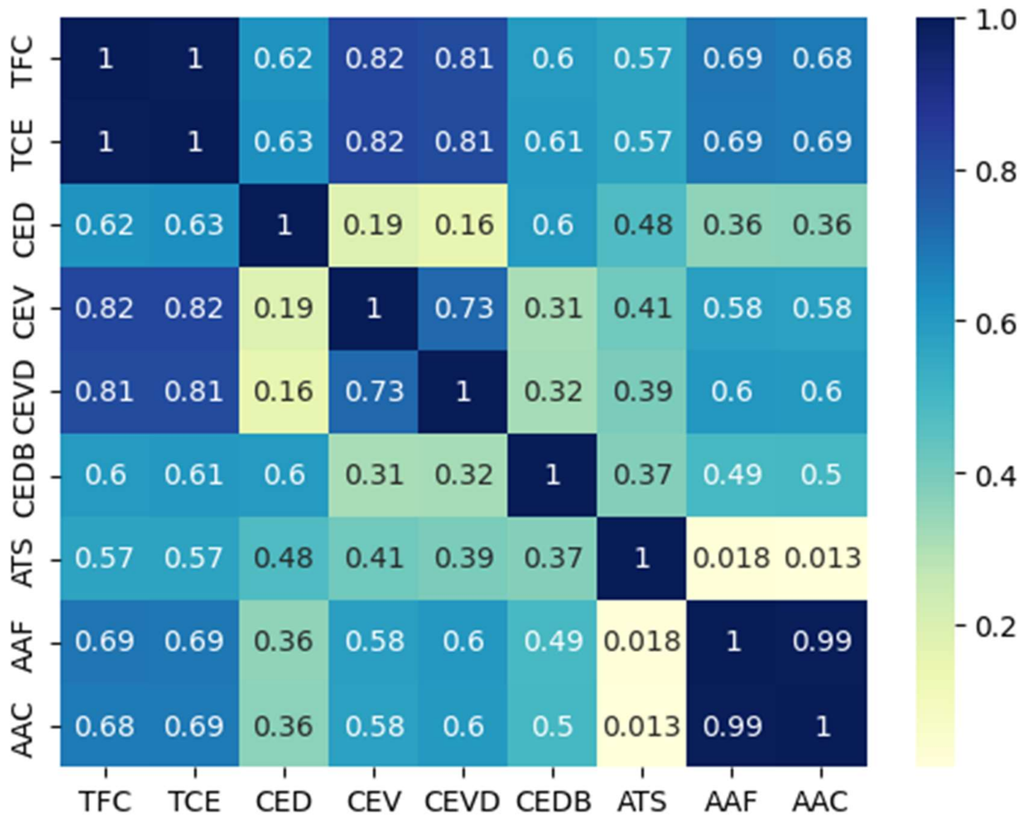

Figure 5:2 Pearson's correlation matrix

5.3 Machine Learning Algorithms

The model building step is a crucial step, as it involves selecting a suitable predictive models based on the nature of problem, training and testing the model, evaluation and validation performed. CO_2 emissions from all voyages between different ports [m tonnes], CO_2 emissions produced when ships depart from various ports to start their journeys [m tonnes], CO_2 emissions from all voyages to destination ports [m tonnes], CO_2 emissions while ships are in port, either docked or berthed [m tonnes], Annual Time spent at sea [hours] were treated as independent feature, while Total CO_2 emissions [m tonnes] serve as the output or target feature. In this proposed paper, four models are used. Decision tree, Random Forest, Support Vector Machine and Feed forward neural network are these models. We use the supervised learning algorithm because we already know the dataset's inputs and outputs. The dataset split into training and test data and then predict the CO_2 emissions. In the final step, wherein the performance metrics of our models were compared to determine the most effective one. Google Colab served as the platform, with data manipulation and analysis carried out using the Pandas library. Visualization was facilitated by Seaborn and Matplotlib, while prediction tasks utilized the Scikit-learn (sklearn) library. These libraries are integral components of the Python programming language; thus, Python was utilized exclusively for all code in this work.

5.4 Results

5.4.1 Data visualization

Figure 5:3 to Figure 5:10 depict scatter plots demonstrating the correlation between the independent variable and the target variable. The data visualization offers insights into the relationship between various factors and total CO_2 emissions within the maritime industry. Figure 5:3 illustrates the correlation between Annual Time spent at sea [hours] and Total CO_2 emissions [m tonnes]. Total CO_2 emissions [m tonnes] plotted over Annual Time spent at sea indicating a gradual increase in CO_2 emissions as the time spent at sea increases. From Figure 5:3, it is evident that there exists a positive correlation between the annual time spent at sea and total CO_2 emissions. As the duration of time spent at sea increases, there is a corresponding gradual rise in CO_2 emissions. This insight suggests that vessel activity duration plays a significant role in determining the amount of CO_2 emitted, implying potential opportunities for optimizing fuel efficiency and implementing emission reduction strategies. Furthermore, the observed correlation underscores the importance of considering voyage duration in assessing and mitigating the environmental impact of maritime operations. By identifying this relationship, stakeholders can focus efforts on developing measures to minimize emissions during extended periods at sea, thereby contributing to sustainable practices in the maritime industry.

Figure 5:4 shows the correlation between CO_2 emissions from all voyages between different ports [m tonnes] and total CO_2 emissions. This depiction reveals that a significant portion of voyages result in high CO_2 emissions. Moving forward, Figure 5:5 portrays the interrelation between CO_2 emissions produced when ships depart from various ports to commence their journeys [m tonnes] and total CO_2 emissions. Notably, this visualization highlights the substantial CO_2 emissions generated when ships embark on their journeys. Similarly, Figure 5:6 underscores the significant CO_2 emissions from all voyages to destination ports. This observation underscores the environmental impact of maritime activities, particularly during departure phases. Hereafter, Figure 5:7 provides insight into the relationship between CO_2 emissions while ships are in port, either docked or berthed at berth [m tonnes], and Total CO_2 emissions [m tonnes]. This depiction sheds light on the emissions of CO_2 during the period when ships are stationed in port, offering valuable insights into port-related emissions. These visualizations, derived from machine learning techniques, contribute to a deeper understanding of the dynamics influencing CO_2 emissions within the maritime sector.

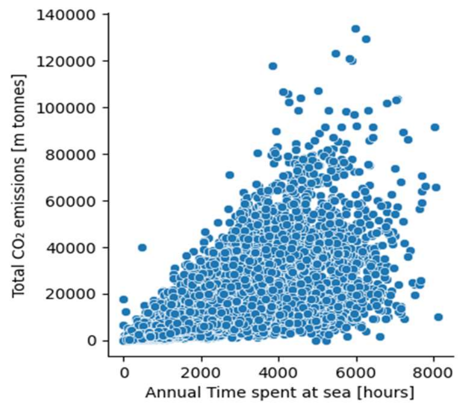

Figure 5:3Total CO_2 emissions [m tonnes] plotted over Annual Time spent at sea [hours]

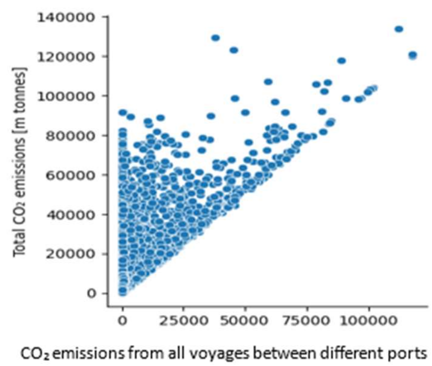

Figure 5:4Total CO_2 emissions plotted over CO_2 emissions from all voyages between different ports

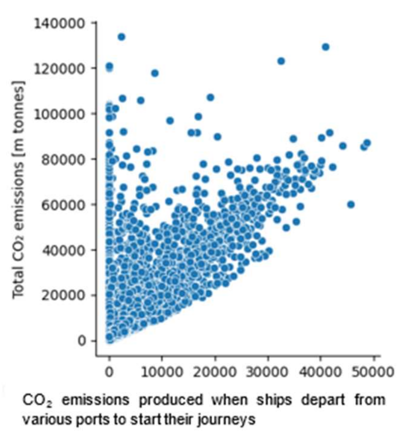

Figure 5:5Total CO_2 emissions plotted over CO_2 emissions produced when ships depart from various ports to start their journeys

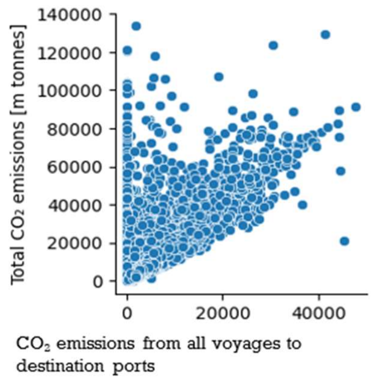

Figure 5:6Total CO_2 emissions plotted over CO_2 emissions from all voyages to destination ports

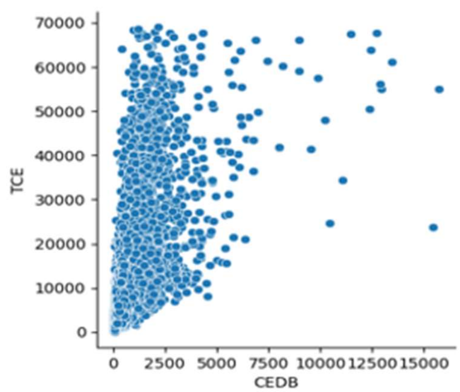

Figure 5:7Total CO_2 emissions plotted over CO_2 emissions while ships are in port, either docked or berthed at berth

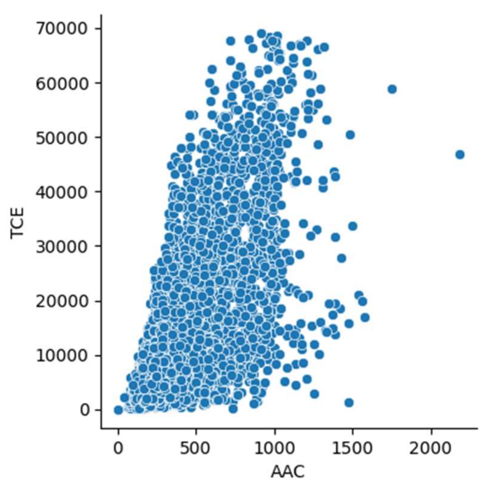

Figure 5:8Total CO_2 emissions plotted over Annual average CO_2 emissions per distance [kg CO_2 / n mile]

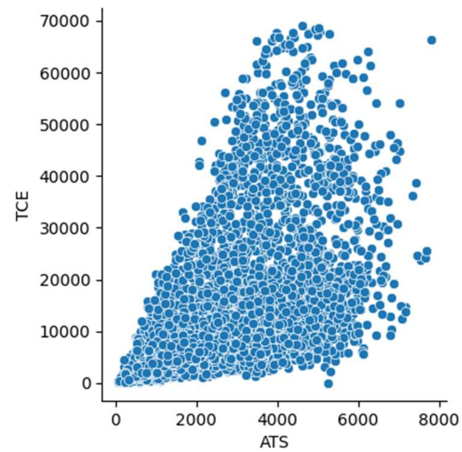

Figure 5:9Total CO_2 emissions plotted over Annual Time spent at sea [hours]

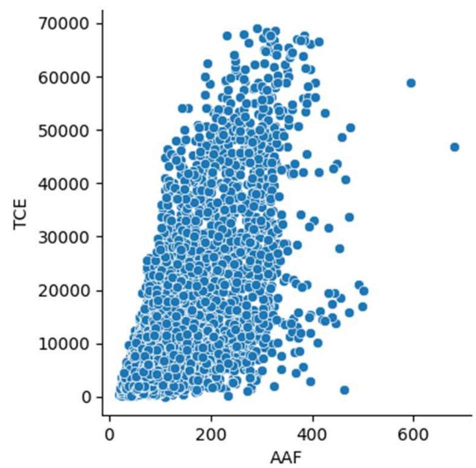

Figure 5:10Total CO_2 emissions plotted over Annual average Fuel consumption per distance [kg / n mile]

Figure 5:8 illustrates a plot of Total CO_2 Emissions versus Annual Average CO_2 emissions per distance in kilograms of CO_2 per nautical mile. There is also a positive correlation exists, indicating that ships with higher CO_2 emissions per nautical mile contribute significantly to total emissions. This relationship emphasizes the importance of improving fuel efficiency and emission reduction strategies. In such a way, the data supports the need for regulatory standards and technological upgrades to minimize CO_2 emissions in maritime operations. And, Figure 5:9 depicts Total CO_2 Emissions versus Annual Time Spent at Sea in hours. There is a clear positive correlation between Annual Time Spent and Total CO_2 Emissions, suggesting that as ships spend more time at sea, their CO_2 emissions tend to increase. As these ships contribute significantly to total emissions, targeted strategies for emission reduction could have a substantial environmental benefit. Whereas, Figure 5:10 presents a plot of Total CO_2 Emissions versus Annual Average Fuel consumption per distance in kilograms per nautical mile. The plot demonstrates a positive correlation between these two, indicating that higher fuel consumption per nautical mile leads to increased CO_2 emissions. This correlation underscores the direct link between fuel efficiency and environmental impact. The plot highlights the potential benefits of adopting advanced propulsion systems and fuel-saving technologies. Upgrading to more efficient engines or utilizing alternative fuels can significantly reduce fuel consumption and emissions.

5.4.2 Performance evaluation of machine learning models

Predictive analysis entails the utilization of statistical algorithms, machine learning techniques, and data mining to forecast future events. In this study, the primary aim is to identify the most effective model for predicting CO_2 emissions. To achieve this goal, four machine learning algorithms namely; Support Vector machine, Random Forest, Decision Tree, and Artificial Neural Network are employed on the dataset. The dataset is partitioned into training and testing data with an 80:20 ratio. An Artificial Neural Network model is implemented, with variations in epochs and batch sizes to optimize neural network performance. The results of these analyses are presented below.

The analysis highlights a consistent upward trend in CO_2 emissions originating from the maritime industry up to 2021. This surge in emissions can be attributed to various factors, including the expansion of industrial activities and the escalation of international trade. To gain deeper insights into emission sources, the study categorizes emissions into distinct activities, such as annual time spent at sea, ship departures from various ports, voyages between different ports under a maritime jurisdiction, voyages to destination ports under a maritime jurisdiction, and emissions generated while ships are in port. By dissecting emissions based on these activities, governing bodies can more effectively monitor and address areas requiring improvement. This granular breakdown facilitates targeted interventions aimed at reducing emissions and enhancing environmental sustainability within the maritime sector.

The principal aim of this study is to predict CO_2 emissions through Support Vector Machine, Random Forest (RF), Decision Tree (DT), and Artificial Neural Network algorithms, subsequently comparing the efficacy of these models. By harnessing predictive modeling techniques, the research endeavors to furnish valuable insights into emission patterns, thereby aiding policymakers and stakeholders in crafting informed strategies for emission reduction and fostering sustainable maritime practices. Actual and predicted results of SVM, RF, DT, and ANN are depicted in Figure 5:11, Figure 5:12, Figure 5:13 and Figure 5:14 respectively.

In the initial phase, SVM and Random Forest techniques were employed to forecast CO_2 emissions, based on training and test data. Additionally, Decision Tree and ANN models were utilized to discern their efficacy in predicting CO_2 emissions from ships.

To assess the accuracy of various methods, Table 5:4 presents actual values and predicted results for total CO_2 emissions from ships for each method, respectively. Meanwhile, SVM, a machine learning model, seeks to minimize error by identifying a function that envelops more original points within the tube created by the hyperplane and its border while reducing slack. The comprehensive performance of ANN and DT models notably surpassed that of RF and SVM in forecasting total CO_2 emissions from ships. However, ANN exhibited relatively higher accuracy levels compared to any other model used in the study, including SVM. Renowned for its predictive capabilities, ANN has been extensively utilized in forecasting algorithms. Overall, total CO_2 emissions from ships were forecasted using all four aforementioned models. Nonetheless, findings suggest that ANN should be prioritized for predicting CO_2 emissions from ships to yield more accurate results.

5.4.3 The forecasting of total CO_2 emissions using SVM

The Support Vector Machine (SVM) model was applied to the dataset using a polynomial kernel with a degree of two. The model yielded an R-squared (R^2) value of 0.9553, indicating a moderate result for the prediction of CO_2 emissions. While this R-squared value suggests a reasonable degree of correlation between the predicted and true values, it falls short compared to other models. Analyzing specific instances from the dataset provides further insights into the SVM model's predictions. For instance, in the third scenario, where the true CO_2 emission value was 3582.93 metric tonnes, the SVM model predicted 3773.12 metric tonnes. Similarly, in the sixth scenario, with a true value of 11160.55 metric tonnes, the model predicted 10446.03 metric tonnes. These results highlight the SVM model's ability to make predictions but also underscore the need for further refinement or exploration of alternative modeling techniques to improve accuracy.

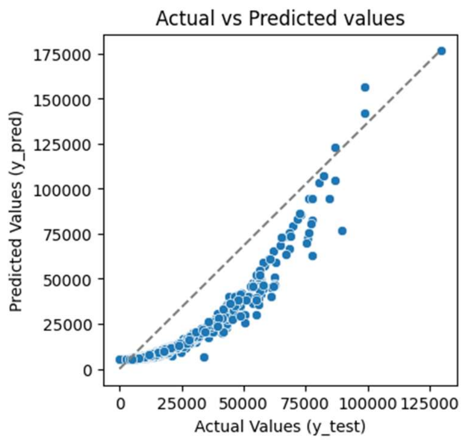

Figure 5:11SVM

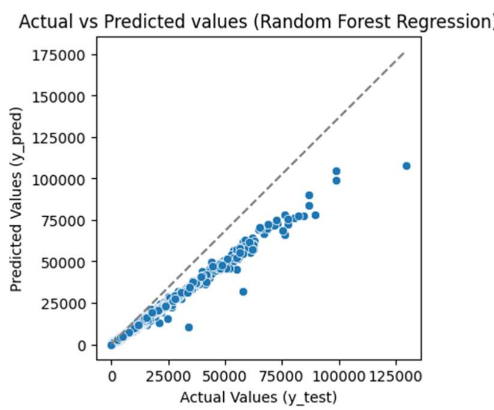

Figure 5:12RF

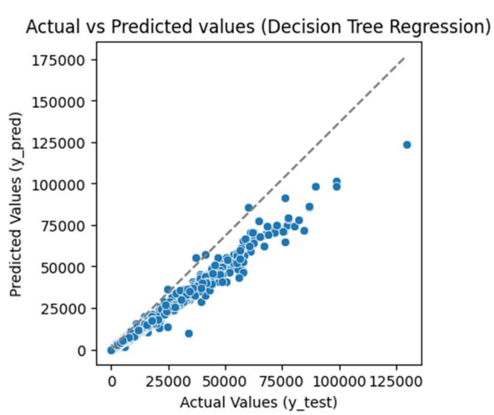

Figure 5:13DT

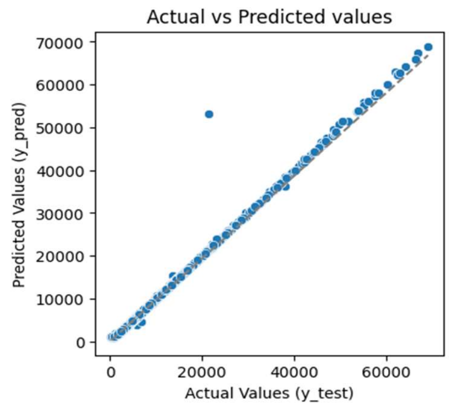

Figure 5:14ANN

The plot shows the actual values of the target variable on the x-axis and the predicted values on the y-axis. The points are scattered around the diagonal line, which indicates a good fit between the actual and predicted values.

Table 5:4 Forecast the total CO_2 emission form ships

Sr. No.	True Value	SVM	RF	DT	ANN
1	3582.93	3773.12	3606.4695	3589.09	3568.20
2	25903.29	23152.11	25095.6911	25307.23	25802.87
3	7335.31	7705.60	7327.6850	7324.01	7357.48
4	5798.84	5928.97	5836.9702	5844.07	5794.68
5	5084.22	5371.54	5098.3385	5087.39	5083.75
6	11160.55	10446.03	11201.6759	11337.82	11111.33
7	8709.08	8831.19	8741.7130	8685.52	8697.00
8	3456.15	3461.61	3456.3436	3466.51	3444.14
9	5235.92	5493.54	5243.6875	5220.34	5077.83
10	6843.12	6634.40	6794.2919	6808.66	6808.86

5.4.4 The forecasting of total CO_2 emissions using RF

The Random Forest (RF) Regression model was employed to predict CO_2 emissions in the dataset, yielding an impressive R-squared (R^2) value of 0.9276. This high R-squared value indicates a remarkably accurate prediction for CO_2 emissions, with a strong correlation between the predicted values and the actual CO_2 emissions in the dataset. Specifically, for the instances presented in the results, the predicted values closely matched the true values, demonstrating the efficacy of the RF algorithm in capturing the underlying patterns and trends in CO_2 emissions data. For example, in the third scenario, where the true value of CO_2 emissions was 7335.31 metric tonnes, the RF model predicted a value of 7327.68 metric tonnes, showcasing a close approximation. Similarly, in the fifth scenario, where the true value was 5084.22 metric tonnes, the RF model predicted a value of 5098.33 metric tonnes, indicating a relatively accurate estimation. These results underscore the robustness of the RF Regression model in accurately predicting CO_2 emissions, making it a valuable tool for analysing and mitigating environmental impact in maritime activities.

5.4.5 The forecasting of total CO_2 emissions using DT

The Decision Tree Regression model was applied to the dataset, yielding a Mean Squared Error (MSE) value of 183,89 and an impressive R-squared (R^2) value of 0.9553. This R-squared value signifies the model's exceptional performance in predicting CO_2 emissions, with 95.53% of the variance in the data being explained by the model. In, the scatter plot illustrates the relationship between the predicted and true values for the Decision Tree model. The majority of points cluster closely around the diagonal line, indicating strong alignment between the actual and predicted values. However, there is one outlier point, suggesting a slight deviation from the overall trend. Examining specific instances from the dataset further reinforces the model's efficacy. For instance, in the second scenario, where the true CO_2 emission value was 25903.2 metric tonnes, the model predicted 25307.23 metric tonnes. Similarly, in the sixth scenario, with a true value of 11160.55 metric tonnes, the model predicted 11337.82 metric tonnes. These results underscore the accuracy and reliability of the Decision Tree Regression model in predicting CO_2 emissions, making it a valuable tool for assessing and mitigating environmental impact in maritime operations.

5.4.6 The forecasting of total CO_2 emissions using ANN

The feed-forward neural network was chosen for application to the dataset due to its suitability for the dataset's nature. The model was executed with various parameter values to optimize results. The neural network employed the Relu activation function with the Adam optimizer and underwent training with epochs and batch sizes set at 50 and 32, respectively. The model yielded a Mean Absolute Error (MAE) value of 105.86

and an impressive R-squared (R^2) value of 0.9636, indicating its effectiveness in predicting CO_2 emissions.

In Figure 5:14, the predicted versus true values for the Artificial Neural Network (ANN) model are depicted. This visualization illustrates the dataset's training under the ANN model, demonstrating the alignment between predicted and true values. While deviations are observed, the majority of data points closely follow the diagonal line, indicating strong predictive capability. Specifically, examining instances from the dataset reinforces the model's performance. For example, in the fifth scenario where the true CO_2 emission value was 5084.22 metric tonnes, the ANN model predicted 5083.75 metric tonnes. Similarly, in the seventh scenario, with a true value of 8709.08 metric tonnes, the model predicted 8697.00 metric tonnes. These results underscore the ANN model's accuracy and reliability in predicting CO_2 emissions, showcasing its potential as a valuable tool for analysing and addressing environmental impact in maritime activities.

	Algorithm	R2	MAPE
0	SVM	0.89014	0.10511
1	RF	0.92760	0.01137
2	DT	0.95530	0.01439
3	ANN	0.96360	0.01415

Table 5:5 R-squared and MAPE values for each algorithm

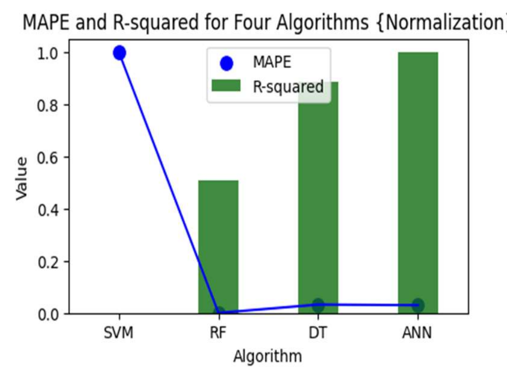

Figure 5:15 Comparison of the normalized value of R-squared and MAPE (Mean Absolute Percentage Error)

5.5 Discussion

The comparative analysis of the four algorithms – Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), and Artificial Neural Network (ANN) – offers valuable insights into their effectiveness in predicting CO_2 emissions in maritime activities. Each algorithm's performance is assessed based on its ability to accurately capture the variance in the data and provide reliable predictions. Figure 5:15 shows the comparison of the normalized value of R-squared and MAPE (Mean Absolute Percentage Error) values. And furthermore, Table 5:5 shows the R-square values obtained for SVM, RF, DT, and ANN – 0.8471, 0.9276, 0.9553, and 0.9636, respectively – indicate varying levels of predictive accuracy. The highest R-square value achieved by ANN suggests its superior performance in accurately capturing the underlying patterns in CO_2 emissions data. This indicates that ANN is particularly adept at modeling the complex relationships present in the dataset, resulting in highly accurate predictions. Following closely behind, DT demonstrates robust predictive capabilities with an R-square value of 0.9553. This suggests that DT effectively captures the variability in the data and provides reliable predictions for CO_2 emissions in maritime activities. The performance of RF is also noteworthy, with an R-square value of 0.9276, indicating commendable predictive accuracy. RF's ability to construct a decision tree that accurately predicts CO_2 emissions showcases its effectiveness as a modeling technique. Although SVM exhibits a slightly lower R-square value of 0.8471 compared to the other algorithms, it still demonstrates notable predictive power. SVM's performance highlights its ability to capture the general trend in CO_2 emissions data and provide valuable insights into the factors influencing emissions in maritime activities.

The varying performances of the four algorithms highlight the importance of selecting the appropriate modeling technique based on the specific characteristics of the dataset and the desired level of prediction accuracy. While RF and DT offer robust performances and are relatively easy to interpret, ANN's superior predictive accuracy makes it a promising choice for modeling complex relationships in CO_2 emissions data. SVM, while presenting a moderate performance, still provides valuable insights into CO_2 emissions patterns and can complement other modeling techniques in comprehensive analysis. Overall, the comparative analysis of the four algorithms underscores the significance of leveraging advanced machine learning techniques for predicting CO_2 emissions in maritime activities, facilitating informed decision-making and effective environmental management strategies.

5.6 Conclusion

This study emphasizes the significance of employing machine learning methodologies to predict CO_2 emissions from ships during various port activities. Through comparative analysis, it was evident that the ANN and DT models outperformed SVM and RF, showcasing superior accuracy in CO_2 emission prediction. Particularly, the ANN model demonstrated exceptional performance, warranting its recommendation for practical application in estimating total CO_2 emissions from maritime ships. By addressing data gaps and offering insights into emission patterns, the ANN model contributes to advancing our understanding of maritime emissions and promotes sustainable practices within the industry. Moreover, the study highlights the potential extension of this predictive approach to model other greenhouse gases globally, leveraging evidence from regions such as Asia. Overall, the comparative analysis of CO_2 emissions from ships provides valuable insights into the relative impact of port activities on emission trends, facilitating informed decision-making for emission reduction strategies and fostering environmental sustainability in the maritime sector.