ET1556 Introduction to scientific work
September 2024

Predicting Sea Freight Carbon Emissions using Machine Learning Techniques

A Comparative Evaluation of Polynomial Support Vector Machine and a Linear Support Vector Machine

AVSS Vighnesh Mandaleeka

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

This project report is submitted to the Department of Mathematics and Natural Sciences at Blekinge Institute of Technology in partial fulfillment of the requirements for the course ET1556.

Contact Information:
Author:
AVSS Vighnesh Mandaleeka
E-mail: ahma24@student.bth.se

University advisor:
Lecturer Irina Gertsovich
Department of Mathematics and Natural Sciences

Dept. of Mathematics and Natural Sciences
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: Maritime freight plays a significant role in global carbon emissions, necessitating the development of precise prediction models to mitigate these impacts. Leveraging machine learning techniques provides a robust approach to handling extensive datasets and identifying complex patterns in CO2 emissions, making them well-suited for this critical task.
Objectives: This project aims to compare the performance of Linear SVM and Polynomial SVM in predicting CO2 emissions from maritime shipping. The focus is on achieving a balance between prediction accuracy and computational efficiency, with evaluation based on key metrics such as MSE, RMSE, MAE, MedAE, EVS, and R^2.
Methodology: The dataset containing maritime shipping data underwent meticulous preprocessing, including feature scaling and selection. Both Linear SVM and Polynomial SVM models were rigorously trained and tested. Their performance was meticulously evaluated based on the aforementioned metrics to assess accuracy and computational efficiency.
Results: Polynomial SVM demonstrated superior accuracy by capturing non-linear relationships in the data, albeit requiring more memory and longer training times. On the other hand, Linear SVM, while faster and more resource-efficient, exhibited lower prediction accuracy and struggled with complex patterns in the dataset.
Conclusions: The Polynomial SVM model outperformed Linear SVM in accuracy, albeit at the cost of greater computational demand. Future initiatives will focus on exploring alternative kernel functions, fine-tuning hyperparameters, and assessing alternative models such as Random Forests and Neural Networks. Further research will also encompass real-time deployment and the integration of external data sources to enhance predictions and efficiency in maritime CO2 emission forecasting.
Keywords:

1. SVM Support Vector Machine
2. MSE Mean Squared Error
3. RMSE Root Mean Square Error
4. MAE Mean Absolute Error
5. EVS Explained Variance Score
6. MBD Mean Bias Deviation
7. MedAE Median Absolute Error

Contents

List of Figures

Chapter 1

Introduction

1.1 Introduction

According to UN estimates, the world population passed the 8 billion mark on 15 November. Over the past 25 years, the number of people on the planet has increased by one-third, or 2.1 billion. Humanity is expected to grow by another fifth to just under 10 billion around 2050. [10]

In 2022, the shipping industry globally emitted an estimated 858 million tonnes of CO2, compared with 739 million tonnes of CO2 emissions from air transport (domestic and international flights). Additionally, 63 percent of emissions from global shipping came from vessels operated by companies based in OECD countries. [**?**] In order to mitigate CO2 emissions from sea freight, a variety of models utilizing different algorithms such as Linear SVM, Linear Regression, CNN, etc. [7] [13] have been employed. It is proposed that a Polynomial Support Vector Machine (SVM) model could offer enhanced accuracy and performance [8].

This model type is proficient in comprehending complex shipping factors including fuel consumption, route planning, and emissions. The model will undergo training using historical shipping data, encompassing details such as ship type, fuel type, distance traveled, cargo weight, and CO2 emissions.
Leveraging this data, [1]the Polynomial SVM can anticipate emission levels in diverse scenarios and propose optimal shipping routes or fuel-saving practices. The SVM's capability to comprehend intricate patterns will augment its accuracy. Once trained, the model can assist in identifying specific modifications conducive to emission reduction, such as deceleration or the advocacy of more sustainable shipping practices. [8]

1.1.1 Background

The combustion of fossil fuels, deforestation, and industrial processes are significant contributors to CO2 emissions, which greatly impact climate change. Accurately predicting CO2 emissions is essential for developing strategies to reduce emissions and meet environmental targets. Machine learning is increasingly being used for this purpose, as it can analyze historical data to identify patterns and predict future trends. Various machine learning models have been applied to CO2 emissions prediction, each with different strengths. Common algorithms include Linear Regres-

sion, Support Vector Machines (SVM), Random Forest, Gradient Boosting Machines (GBM), Neural Networks, and k-nearest Neighbors (k-NN) [2].

Polynomial Support Vector Machine (SVM) stands out for its ability to handle the complex, non-linear relationships found in large, high-dimensional datasets. Simpler models, like Linear Regression or Linear SVM, struggle with such complexity. While more advanced models, such as Random Forest or Neural Networks, are powerful, they often require significant computational resources, making them less practical for large-scale applications [8]. Polynomial SVM provides a balanced solution, capturing non-linear patterns without becoming too computationally demanding. By mapping data into higher dimensions, it effectively separates complex patterns that linear models miss, making it well-suited for datasets with many interrelated features, like those used for CO2 emissions prediction.

This project focuses on Polynomial SVM due to the complexity of the dataset. By comparing it with Linear SVM, we aim to show that Polynomial SVM delivers more accurate predictions while remaining computationally feasible. The evaluation will consider metrics like prediction accuracy, efficiency, and overall performance, highlighting the advantages of Polynomial SVM for this task [8] [13].

1.1.2 Aim and Objectives

Aim: The goal is to compare the prediction accuracy and efficiency of two models for fore-casting CO2 emissions from sea freight: a traditional Support Vector Machine (SVM) and a Polynomial SVM enhanced with dimensionality reduction techniques.

Objectives:

1. **Data Collection and Refinement:-** The task involves gathering and preprocessing a large, complex CO2 emissions dataset. This will ensure data quality through cleaning, normalization, and feature selection, optimizing input for both Polynomial and Linear SVM models.

2. **Evaluate Prediction Accuracy:-** The goal is to compare the prediction accuracy of Polynomial SVM and Linear SVM models on the refined dataset. This comparison will help assess which model more effectively captures non-linear relationships and produces better predictions.

3. **Assess Model Performance:-** The objective is to analyze the computational performance of both models, focusing on training time, memory usage, and scalability. This analysis will determine the efficiency of Polynomial SVM over Linear SVM.

4. **Model Implementation and Parameter Optimization:-** This involves implementing both Polynomial and Linear SVM models, fine-tuning parameters such as polynomial degree and regularization, and evaluating their impact on prediction accuracy and model performance.

Research Question:
How does the use of Polynomial SVM compare to traditional linear SVM in predicting CO2 emissions from maritime shipping in terms of accuracy and performance?
Motivation:-
The large dimensionality of the dataset [1] led to the selection of Polynomial Support Vector machines (SVM) for the CO2 emissions prediction model [8]. High-dimensional data frequently causes overfitting, which hinders the model's capacity for efficient generalization. The model can minimize complexity by concentrating on the most crucial elements through the application of dimensionality reduction techniques [11]. For large-scale or real-time prediction jobs, this method improves prediction accuracy, decreases training time, uses less memory, and boosts the model's overall efficiency.

1.1.3 Scope

The project's scope is limited to evaluating the performance of Polynomial Support Vector Machines (SVM) in comparison to Linear SVM on a large and complex dataset for CO2 emissions prediction. The main goal is to determine whether the Polynomial SVM, with its capability to model non-linear relationships, offers significant improvements in prediction accuracy and overall performance compared to the simpler Linear SVM. The evaluation will focus on metrics such as accuracy, training time, and computational efficiency, providing a clear comparison between the two approaches within the context of the dataset used.

1.1.4 Outline

Chapter 1: Introduction, background, aims, objectives, and ethical aspects. Chapter 2: Related work in machine learning and CO2 emissions modeling. Chapter 3: Methodology, data collection, SVM implementation, and model performance evaluation. Chapter 4: Results and analysis, comparing Polynomial and Linear SVM performance. Chapter 5: Discussion of results and interpretation, addressing model advantages and limitations. Chapter 6: Conclusion, summarizing findings and providing recommendations for future work.

1.1.5 Ethical, societal and sustainability aspects

1.1.5.1 Ethical Aspects

We use AIS and other publicly available data in our research. This data does not include any private information about individuals and is not confidential. It only shows vessel routes and is not linked to any specific nation or person, so there are no GDPR concerns.

1.1.5.2 Societal Aspects

Using machine learning to predict CO2 emissions from sea freight can help us influence regulations for transporting goods by cargo. This could have a big impact on climate change and global warming. It could also help people who are at risk from

sea or ocean-related natural disasters. Additionally, this information could encourage the use of more environmentally friendly transportation methods instead of shipping by sea. This would help create a more sustainable future.

1.1.5.3 Sustainability Aspects

The findings support the adoption of initiatives that promote sustainable freight transportation by helping to identify relevant environmental regulations and recommendations. Using data mining techniques and machine learning algorithms to analyze cargo emissions will drive lasting changes in the industry for better reduction and control of CO2 emissions.

Chapter 2

Related Work

2.1 Related Work

Jeswanth Naidu's [7] thesis examines the use of machine learning algorithms to predict carbon emissions from sea freight. The study evaluates decision trees, support vector machines, and random forests, using data from the Automatic Identification System (AIS). It involves data collection, preprocessing, feature selection, and model training. The performance of these models is assessed using metrics such as Mean Square Error, Root Mean Square Error, Mean Absolute Error, and R-squared. The research highlights the advantages and disadvantages of each model, ultimately demonstrating the effectiveness of machine learning in reducing maritime carbon emissions.

Fongyiu Wong researched [12] predicting carbon emissions allowances trade using machine learning. The study utilized both long-term and short-term prediction models to forecast carbon emissions trading volume. Various machine learning algorithms such as linear regression, decision trees, SVMs, extreme gradient boosting, and random forests were employed for long-term prediction. To enhance accuracy, a time series forward multi-step hybrid intelligent prediction model was used for short-term prediction. The short-term prediction model training process involved a combination of reinforcement learning and the hidden Markov model, followed by the integration of neural network and hidden Markov model with reinforcement learning. The research concluded that there exists a nonlinear relationship between the lag of transaction volume/frequency and price in the future. Among the five long-term prediction models, Random Forest performed best, with the smallest mean absolute error of 24.42, while the mean absolute error of short-term prediction was 6.79.

In a study by Babasaheb S. Satpute et al. [9] machine learning methods were used to predict CO2 emissions using a dataset that included characteristics of over 7500 automobiles. The initial model resulted in a mean absolute error (MAE) of 3.24 and a mean squared error (MSE) of 30.00. However, this model may have limited efficacy due to its assumption of a linear correlation between characteristics and CO2 emissions. The decision tree model outperformed the initial model with an MAE of 1.86 and an MSE of 14.27. The Random Forest model further improved the predicted accuracy with an MSE of 2.32 and an MAE of 1.83. In contrast, the SVM model performed comparatively worse with an MSE of 406.86 and an MAE of 9.54.

In 2018, Pooja Kadam and Suhasini Vijayumar [5] presented a CO2 emission prediction model using machine learning, offering valuable insights. Traditionally, researchers have utilized statistical techniques such as regression, t-test, derivation, and ANOVA Test for prediction. Machine learning introduces diverse techniques to train the machine based on experience. The paper's regression model is employed to predict CO2 emissions based on historical data. To enhance the prediction model, a trial and error approach was implemented to minimize error. The Root Mean Square Error (RMSE) value is 0.2557. The primary objective of the experiment is to achieve a low RMSE value.

Chapter 3

Method

3.1 Method

The research methodology employed in this study encompasses the application of machine learning techniques to accurately estimate carbon emissions from marine freight. Python was selected as the programming language for its robustness and versatility. The data collection process entails a comprehensive gathering of operational parameters, environmental conditions, vessel features, and time-related aspects, ensuring a comprehensive understanding of the factors influencing emissions.

Through meticulous variable selection and rigorous data preprocessing, we ascertain the accuracy and suitability of the dataset for analysis. The training of machine learning models, including Linear Support Vector Machine, and Polynomial Support Vector Machine, involves meticulous Hyperparameter tuning to optimize performance. Performance metrics such as MSE, RMSE, MAE, EVS, MedAE, MBD and R^2 [4] [6] [?]score are utilized during testing and validation to rigorously assess the predictive capabilities of the models. To address research question 1.1.2

we apply a blend of quantitative and qualitative research techniques, comparing the effectiveness of 2 types of SVMs [13] [8], in predicting carbon emissions from shipping. The efficiency of each algorithm is quantitatively evaluated, providing a rigorous assessment of their predictive power. Additionally, an extensive literature review was conducted to contextualize the study within the existing body of knowledge, providing valuable insights into the problem domain and identifying areas requiring further investigation.

This qualitative research component forms a robust foundation for identifying the research gap and guiding future studies. Finally, the models are objectively evaluated based on their ability to accurately predict new instances on unseen data after being trained on a subset of the original data, ensuring the robustness and reliability of our findings4.

3.1.1 Data Collection

During the data collection phase, we are thoroughly examining the specifics of the data utilized in our research. The dataset employed in this study is a fusion of Automatic Identification System (AIS) data [1]sourced from both satellite and terrestrial

channels, offering a comprehensive insight into maritime activities. The initial segment of the dataset encompasses 19 port calls occurring in specific countries, such as Sweden, Finland, Estonia, Latvia, Lithuania, and Poland, within the time frame of January 1, 2021, to April 31, 2022. This dataset, which is unlabeled, encompasses cargo and tanker ships exceeding 65 meters in length and provides extensive details, including port ID, name, LOCODE, MMSI, IMO, vessel name, destination, type, as well as arrival and departure timestamps.

3.1.2 Data Preprocessing

Data preprocessing is a crucial step in machine learning where we clean the data and make sure it is ready to use for training a model so handling the missing values and scaling them as required to use for training SVM models.

1. **Handling Missing Values** In the real-world data we can have some missing values that can cause an error in the model training due to its null values rather than an empirical value which can mitigate the accuracy and performance of the model we are training so removing them is a crucial step and mandatory step in data preprocessing In this project the dataset have null numeric values in speed and draught columns so we initially converted the non-numeric values to NaN and then followed by removing the NaN values as shown in 3.1.

```
df['Speed'] = pd.to_numeric(df['Speed'], errors='coerce')
df['Draught'] = pd.to_numeric(df['Draught'], errors='coerce')
df = df.dropna(subset=['Speed', 'Draught'])
```

Figure 3.1: Removing Null Values

2. **Removing Outliers** Outliers can also impact the performance of models by skewing the results which ultimately leads to overfitting of the model in this project we removed the outliers for 99th percentile for the speed and draught columns using a process called Quantile Filtering 3.2this is used to only retain the rows where the values of the columns speed and draught are 99th percentile.

```
df = df[(df['Speed'] <= df['Speed'].quantile(0.99)) &
        (df['Draught'] <= df['Draught'].quantile(0.99))]
```

Figure 3.2: Outlier removal

3. **Data Standardization** Standardization makes sure that every feature contributes equally to model training by converting them into a common scale. This is particularly important for models like SVM, which are sensitive to the

feature scales done by implementing the code in 3.3.

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Figure 3.3: Standardization

4. **Label Encoding** The SVR models cannot handle categorical values directly, so we need to convert them into numeric values, which makes the training process easier. In our dataset, we have a column called Vessel Type, which is categorical, so we need to make it numeric 3.4.

```
le = LabelEncoder()
df['VesselType'] = le.fit_transform(df['VesselType'])
```

Figure 3.4: Encoding Categorical Values

3.1.3 Feature Engineering And Dimensionality Reduction

Feature Engineering: is a crucial part of data preprocessing where we create new features by transforming the existing ones to improve the model's performance in this project we have used several feature engineering techniques like [3] [?]

1. **Interaction Terms:** We have created a new feature by multiplying two more features to capture interactions between them this helps to catch the non-linear relationships between variables 3.5.

```
df['Speed_Draught'] = df['Speed'] * df['Draught']
df['Length_Width'] = df['Length'] * df['Width']
```

Figure 3.5: Feature Generation

2. **Trigonometric Transformations:** We used the sine and cosine transformations to introduce the period patterns between the length and width columns which can be useful to the model to capture the oscillatory behavior in the data 3.6.

```
df['sin_Length'] = np.sin(df['Length'])
df['cos_Width'] = np.cos(df['Width'])
```

Figure 3.6: Trigonometric Transformation

3. **Exponential Transformation:** Exponential Transformation is applied to feature Speed to capture the exponential relationships which can be useful in modeling the non-linear data 3.7.

```
df['exp_Speed'] = np.exp(df['Speed'] / 100)
```

Figure 3.7: Exponential Tranformation

4. **Clustering KMeans:** We have applied the K-means clustering to create new features using the unsupervised learning this step groups the vessels into clusters for potentially capturing the hidden patterns within the data3.8.

```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=5, random_state=42)
df['Cluster'] = kmeans.fit_predict(df.drop(columns='CO2_Emissions'))
```

Figure 3.8: Clustering

5. **Dimensionality Reduction:** Dimensionality Reduction is a process where we simplify the data without loosing significant information so we are applying the Principal Component Ananlysis (PCA) one of the dimensionality reduction techniques that helps in reducing the complexity of the data we are working on while retaining the key variance as shown in 3.9 [11].

```
from sklearn.decomposition import PCA
pca = PCA(n_components=5)
X_pca = pca.fit_transform(df.drop(columns='CO2_Emissions'))
df['PCA1'] = X_pca[:, 0]
df['PCA2'] = X_pca[:, 1]
```

Figure 3.9: Dimensionality Reduction PCA

3.1.4 Model Training

The primary objective of regression analysis is to make predictions regarding continuous output variables. In the context of this specific project, the goal is to forecast a variable such as CO2 emissions, which is directly influenced by various parameters such as the ship's speed, draught, length, and width. Regression emerges as the most suitable technique for the following reasons:

1. **Continuous Target Variable:** In contrast to classification, which is concerned with predicting discrete labels, regression is better equipped to predict continuous outcomes such as CO2 emissions.

2. **Complex Relationships:**Real-world data often entails intricate relationships. In the case of CO2 emissions, the relationship with input features may not be

purely linear, potentially involving non-linear dependencies, including higher-order terms. Regression models possess the capability to effectively capture and represent these complex relationships.

SVR : Support Vector Machines (SVM) are widely used for classification tasks, and they can also be adapted for regression problems, which is known as Support Vector Regression (SVR). The primary objective of SVR is to determine a function that accurately approximates the mapping between input features and the target variable while adhering to specific error tolerances.

SVM Regressor: It effectively identifies a hyperplane that best fits the data within a margin of tolerance, known as epsilon (ϵ-insensitive loss function). This approach allows SVM to identify the points within this margin and disregard data points beyond a certain distance.

In this project, we are working with two types of regressors: Linear SVM Regressor (Support Vector Regression, SVR) and Polynomial SVR. Both are grounded in the Support Vector Machine (SVM) concept, yet they diverge in their approach to handling the complexity of the data. Here is a detailed breakdown of each:

1. **Linear SVM:** In this scenario, the SVM regressor seeks to establish a linear correlation between the features (such as speed and draught) and the target (CO2 emissions). If the data exhibits a simple correlation, this approach can be effective. However, if the relationship is more intricate, it may not fully capture the data structure, as indicated by the lower R^2 scores in the evaluation results 3.10.

$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n} \max(0, |y_i - (\mathbf{w}^T\mathbf{x}_i + b)| - \epsilon)$$

Figure 3.10: Mathematical Representation of Linear SVM

Where:
w is the weight vector.
b is the bias term.
C is the regularization parameter.
ϵ is the margin of error allowed.

2. **Polynomial SVM:** This method increases complexity by introducing polynomial terms, enabling the model to accommodate more intricate, non-linear correlations. This helps in better capturing the non-linear relationship between features and emissions, resulting in slightly improved performance compared to Linear SVM in terms of R^2 score and errors 3.11.

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j + r)^d$$

Figure 3.11: Mathematical Representation of Polynomial SVM

Where:
d is the degree of the polynomial.
r is a constant that can shift the kernel function.

3.1.5 Model Evaluation

Upon completion of model training, a comprehensive evaluation was conducted to assess the performance of the Linear SVM and Polynomial SVM regression models in predicting CO2 emissions. Various essential metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Median Absolute Error (MedAE), Explained Variance Score (EVS), and R-squared score (R^2), were meticulously computed to gauge model performance [3].

1. **Mean Squared Error (MSE):** was utilized as a prominent metric to measure the average squared differences between predicted and actual values, with lower MSE values indicating superior model performance and closer alignment between predicted and actual values 3.12.

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Mean Squared Error.

Figure 3.12: MSE

2. **The Root Mean Squared Error (RMSE):** served as a crucial error metric, providing insights into the average disparity between predictions and actual values, expressed in the same units as the target variable 3.13.

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Root Mean Squared Error.

Figure 3.13: RMSE

3. **the Mean Absolute Error (MAE) and Median Absolute Error (MedAE):** facilitated the assessment of the average magnitude of errors, regardless of their direction3.14, with MedAE offering increased robustness against outliers3.15.

$$\frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i |$$

Mean Absolute Error.

Figure 3.14: MAE

$$\text{MedAE}(y, \hat{y}) = \text{median}(| y_1 - \hat{y}_1 |, \ldots, | y_n - \hat{y}_n |)$$

Figure 3.15: MedAE

4. **R-squared Score (R^2):** played a pivotal role in quantifying the proportion of variance in the dependent variable predictable from the independent variables, with higher R^2 values signifying superior model performance and enhanced variance explanation.

5. **Explained Variance Score (EVS):** provided valuable insights into the model's explanatory power, with higher scores indicating a more comprehensive capture of data variance and, consequently, improved performance3.16.

$$(y, \hat{y}) = 1 - \frac{Var(y - \hat{y})}{Var(y)}$$

Figure 3.16: EVS

3.1.6 Model Testing

In this project, we are conducting a thorough assessment of the performance of Linear SVM and Polynomial SVM regression models in predicting CO2 emissions using unseen data. Here is a detailed breakdown of the steps involved:

1. **Train-Test Split:**
   (a) The dataset is meticulously partitioned into training and testing sets to ensure a robust analysis.
   (b) The training set is utilized to comprehensively instruct the model and discern complex patterns within the data.
   (c) The testing set is rigorously employed to evaluate the models' performance on new, unseen data, closely simulating real-world scenarios where precise predictions are essential.
   (d) Typically, we opt for an 80/20 or 70/30 split to allocate a substantial portion of the data for training while reserving an ample portion for comprehensive testing.

2. **Model Testing:**

(a) We meticulously test them using the dedicated testing set. The models adeptly employ the acquired knowledge from the training phase to predict the CO2 emissions for the test data.

(b) The predicted values are meticulously compared with the actual target values in the test set using a comprehensive suite of evaluation metrics such as RMSE, MSE, etc.

(c) Following thorough testing of both the Linear SVM and Polynomial SVM on the identical test set, their performances are meticulously compared based on the aforementioned metrics.

Chapter 4

Results and Analysis

4.1 Results

In this section, we present the results obtained from training and evaluating the Linear SVM and Polynomial SVM models on various dataset sizes: 5,000 4.1, 10,000 4.2, 25,0004.3, and 50,000 samples 4.4. The models were evaluated based on key performance metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) score Explained Variance Score(EVS), Mean Bias Deviation(MBD),Median Absolute Error(MedAE) [3].

Results for size 5000

```
Linear SVM Results:
Mean Squared Error: 566518274329.58
Mean Absolute Error: 455626.83
R2 Score: -0.1316
Training Time: 0.43              seconds
Memory Usage: 196.78      MB

Additional Linear SVM Results:
Root Mean Squared Error (RMSE): 752674.082
Explained Variance Score (EVS): 0.069
Mean Bias Deviation (MBD): 317591.168
Median Absolute Error (MedAE): 186966.0198

Polynomial SVM Results:
Mean Squared Error: 503312531169.99
Mean Absolute Error: 453023.58
R2 Score: -0.005
Training Time: 38.3111           seconds
Memory Usage: 234.148     MB

Additional Polynomial SVM Results:
Root Mean Squared Error (RMSE): 709445.227
Explained Variance Score (EVS): 0.169
Mean Bias Deviation (MBD): 295968.0073
Median Absolute Error (MedAE): 208256.078
```

Figure 4.1: Dataset Size=5000

Results for size 10000

```
Linear SVM Results:
Mean Squared Error: 579438758048.50
Mean Absolute Error: 453759.207
R2 Score: -0.0437
Training Time: 1.5870          seconds
Memory Usage: 208.84       MB

Additional Linear SVM Results:
Root Mean Squared Error (RMSE): 761208.748
Explained Variance Score (EVS): 0.126
Mean Bias Deviation (MBD): 307329.986
Median Absolute Error (MedAE): 183393.318

Polynomial SVM Results:
Mean Squared Error: 433278151694.29
Mean Absolute Error: 442037.588
R2 Score: 0.219
Training Time: 158.1333          seconds
Memory Usage: 211.027      MB

Additional Polynomial SVM Results:
Root Mean Squared Error (RMSE): 658238.673
Explained Variance Score (EVS): 0.3500
Mean Bias Deviation (MBD): 269192.887
Median Absolute Error (MedAE): 215535.5341
```

Figure 4.2: Dataset Size=10000

Results for size 25000

```
Linear SVM Results:
Mean Squared Error: 972930383823.76
Mean Absolute Error: 452142.587
R2 Score: 0.081
Training Time: 10.307            seconds
Memory Usage: 295.988      MB

Additional Linear SVM Results:
Root Mean Squared Error (RMSE): 986372.335
Explained Variance Score (EVS): 0.1945
Mean Bias Deviation (MBD): 345621.42
Median Absolute Error (MedAE): 147793.711

Polynomial SVM Results:
Mean Squared Error: 456500755583.11
Mean Absolute Error: 441976.03
R2 Score: 0.569
Training Time: 979.9319          seconds
Memory Usage: 241.773      MB

Additional Polynomial SVM Results:
Root Mean Squared Error (RMSE): 675648.396
Explained Variance Score (EVS): 0.6472
Mean Bias Deviation (MBD): 287549.707
Median Absolute Error (MedAE): 207739.076
```

Figure 4.3: Dataset Size=25000

Results for size 50000

```
Linear SVM Results:
Mean Squared Error: 514993712130.60
Mean Absolute Error: 325102.378
R2 Score: 0.300
Training Time: 52.8042          seconds
Memory Usage: 287.332      MB

Additional Linear SVM Results:
Root Mean Squared Error (RMSE): 717630.623
Explained Variance Score (EVS): 0.384
Mean Bias Deviation (MBD): 247861.7530
Median Absolute Error (MedAE): 103951.491

Polynomial SVM Results:
Mean Squared Error: 358630821927.139
Mean Absolute Error: 396613.867
R2 Score: 0.51300
Training Time: 3884.725          seconds
Memory Usage: 92.246      MB

Additional Polynomial SVM Results:
Root Mean Squared Error (RMSE): 598857.9313
Explained Variance Score (EVS): 0.5935
Mean Bias Deviation (MBD): 243546.3466
Median Absolute Error (MedAE): 191257.431
```

Figure 4.4: Dataset Size=50000

The training and testing of model was done in multiple iterations and observed for any drastic change in the accuracy scores of the model but no major difference was found3.The accuracy of the model is keep getting good with increase of size of the dataset we are using a detailed analysis of the result is discussed further.

4.2 Analysis of Results:

1. **Trend with Increasing Dataset Size:** As the dataset size increases from 5,000 to 50,000 samples, both Linear SVM and Polynomial SVM consistently demonstrate significantly improved performance metrics, particularly in terms of lower MSE and RMSE values. This clear trend indicates that larger datasets enable the models to learn more effectively from the data, resulting in markedly improved accuracy in predicting CO2 emissions.

2. **Polynomial SVM Performance:** Across all dataset sizes, Polynomial SVM consistently outperforms Linear SVM. The MSE and RMSE values for Polynomial SVM consistently remain lower, clearly indicating its superior ability to capture the underlying relationships in the data, especially the non-linear patterns present in CO2 emissions.

3. **Linear SVM Performance:** While Linear SVM exhibits decent performance, its MSE and RMSE values consistently remain higher compared to Polynomial SVM. This compellingly suggests that Linear SVM may not sufficiently capture the complexity of the dataset, especially as the data size grows.

4. **R^2 Scores:** The R^2 scores for both models consistently improve with larger datasets, with Polynomial SVM achieving significantly higher R^2 scores. This robustly reinforces its exceptional capacity to effectively explain the variance in CO2 emissions. The increasing R^2 scores indicate that the models become increasingly predictive as they are trained on larger datasets.

5. **Practical Implications:** The results unequivocally underscore the critical importance of selecting an appropriate model based on the dataset size and the complexity of the relationships present in the data. For applications necessitating high accuracy in predicting CO2 emissions, Polynomial SVM unequivocally stands out as the preferred choice, while Linear SVM may still find utility in resource-constrained environments or where computational efficiency is a paramount concern.

Chapter 5

Discussion

5.1 Discussion

Research Question:
How does the use of Polynomial SVM compare to traditional Linear SVM in predicting CO2 emissions from maritime shipping in terms of accuracy and performance?

Introduction and Restating the Research Question: This study aimed to compare the performance of Polynomial SVM and Linear SVM regression models for predicting CO2 emissions in maritime shipping. The research sought to determine which model provides the best balance between prediction accuracy and computational efficiency, particularly for handling complex, non-linear relationships in the dataset.

Answer to Research Question: The findings demonstrate that Polynomial SVM consistently outperformed Linear SVM in terms of predictive accuracy. The Polynomial SVM model had lower error metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), and achieved higher values for Explained Variance Score (EVS) and R-squared (R^2). This indicates its stronger ability to capture non-linear relationships within the maritime shipping CO2 emissions dataset. However, this performance came at a cost—Polynomial SVM required significantly more computational resources, including longer training time and higher memory usage. On the other hand, Linear SVM, while less accurate, offered faster training and more efficient resource use, making it more practical for real-time or resource-constrained scenarios.

Relationship of Results to Literature: The results align with existing literature that underscores the strengths of Polynomial SVM in modeling non-linear data. Previous research has highlighted how Polynomial SVM can effectively capture complex relationships, particularly in environmental modeling. However, as noted in the literature, this accuracy comes at a higher computational cost, especially in terms of training time and memory consumption, which is consistent with our findings. Linear SVM, on the other hand, is known to perform adequately for simpler, more linear datasets and requires fewer resources, which mirrors our study's results.

Discussion of Findings: The key finding of this research highlights the accuracy vs. efficiency trade-off between the two models. On the one hand, Polynomial SVM

offers superior accuracy by better capturing the non-linearities in the data, making it more suitable for complex datasets like maritime CO2 emissions. On the other hand, Linear SVM is far more efficient in terms of training speed and memory usage, making it better suited for simpler models or resource-constrained applications where prediction time and computational cost are critical factors.

Potential Conflicting Interpretations: However, it could also be argued that in some real-world scenarios, such as maritime shipping logistics, the emphasis might be on achieving faster predictions with fewer computational resources. Another possible interpretation is that, despite the superior accuracy of Polynomial SVM, the incremental improvement in prediction accuracy may not justify the additional computational burden in certain applications. The specific application context—whether it prioritizes accuracy or efficiency—will largely determine which model is more appropriate.

Unexpected Findings and Limitations: An unexpected finding in this study was the substantial increase in computational costs associated with the Polynomial SVM. Memory usage and training times exceeded initial expectations, highlighting the challenge of using this model in real-time or large-scale applications. Moreover, while both models benefited from preprocessing steps such as scaling and feature engineering, the performance gap remained substantial, reinforcing the necessity of using Polynomial SVM for non-linear datasets.
The study's scope was limited to SVM models, meaning that other advanced machine learning models (such as Random Forests or Neural Networks) were not explored. These models could potentially provide better accuracy while balancing computational efficiency. Future work should consider these alternative approaches to optimize performance further.

Chapter 6

Conclusions and Future Work

6.1 Conculsion

Upon completion of the model training, we conducted a comprehensive assessment to evaluate the performance of the Linear Support Vector Machine (SVM) and Polynomial SVM regression models in predicting CO2 emissions. We assessed key metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Median Absolute Error (MedAE), Explained Variance Score (EVS), and R-squared score (R^2) to gauge model performance3.

The Mean Squared Error (MSE) served as a pivotal metric for measuring the average squared variances between predicted and actual values. Lower MSE values indicate better model performance and closer alignment between predicted and actual values. The Root Mean Squared Error (RMSE) offered insights into the average differences between predictions and actual values, expressed in the same units as the target variable.

Moreover, the Mean Absolute Error (MAE) and Median Absolute Error (MedAE) were instrumental in evaluating the average magnitude of errors, irrespective of their direction, with MedAE providing increased robustness against outliers. The R-squared Score (R^2) quantified the proportion of variance in the dependent variable that is predictable from the independent variables. Higher R^2 values signify better model performance and improved variance explanation. Additionally, the Explained Variance Score (EVS) provided valuable insights into the model's explanatory power, with higher scores indicating a more comprehensive capture of data variance and improved performance.

Our analysis revealed that while the Polynomial SVM demonstrated superior performance in capturing non-linear relationships, with higher R^2 and lower MSE, it also necessitated greater memory consumption and longer training times, suggesting a trade-off between performance and resource utilization. In contrast, the Linear SVM, while showing faster training and more efficient memory usage, displayed slightly inferior accuracy, particularly in capturing complex data patterns.

6.2 Future Work

While the results from this project provide valuable insights into the performance of SVM models for CO2 emissions prediction, there are several areas where further exploration could be beneficial. Below are some suggested directions for future work:

1. **Explore Other Kernel Functions:** In this project, we compared Linear and Polynomial SVMs, but there are other kernels, such as Radial Basis Function (RBF) or Sigmoid kernels, that may strike a better balance between performance and computational efficiency. Testing different kernel functions could yield improved results with less computational overhead.

2. **Hyperparameter Tuning:** While basic hyperparameters were used in this project, further fine-tuning through Grid Search or Random Search could potentially improve both the accuracy and efficiency of the models. Specifically, adjusting parameters such as C (regularization parameter), the degree of the polynomial kernel, and gamma (for RBF kernels) could enhance model performance.

3. **Feature Engineering:** In future iterations, more sophisticated feature engineering techniques could be applied to improve model performance. This may include feature scaling, feature selection, and dimensionality reduction techniques like PCA (Principal Component Analysis) or LDA (Linear Discriminant Analysis), which could help both models make more informed predictions by reducing noise in the dataset.

4. **Handling Larger Datasets:** Although the current dataset was sufficient for model comparison, scaling these models to handle much larger datasets could be beneficial. Applying SVM models to larger datasets could also involve parallelization techniques or even moving towards more scalable machine learning methods like Deep Learning for CO2 emission prediction.

5. **Comparison with Other Models:** While this project focused on SVM, it would be insightful to compare these results with other regression models such as Decision Trees, Random Forests, and Neural Networks. These models, particularly ensemble techniques like Random Forests or Gradient Boosting, may offer a better trade-off between accuracy and computational efficiency.

References

[1] "AIV_bth.SE_flytnow - Google Drive." [Online]. Available: https://drive.google.com/drive/folders/1x45VUmqEOERXg72d1N-VojuvE5K5Zx0b

[2] M. Balat, H. Balat, and N. Acici, "Environmental Issues Relating to Greenhouse Carbon Dioxide Emissions in the World," *Energy Exploration & Exploitation*, vol. 21, no. 5, pp. 457–473, Oct. 2003, publisher: SAGE Publications Ltd STM. [Online]. Available: https://doi.org/10.1260/014459803322986286

[3] G. Dong and H. Liu, *Feature Engineering for Machine Learning and Data Analytics*. CRC Press, Mar. 2018, google-Books-ID: 661SDwAAQBAJ. [Online]. Available: https://books.google.se/books?id=661SDwAAQBAJ

[4] A. Gunawardana and G. Shani, "A Survey of Accuracy Evaluation Metrics of Recommendation Tasks." [Online]. Available: https://www.jmlr.org/papers/volume10/gunawardana09a/gunawardana09a.pdf

[5] 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

[6] H. M and S. M.N, "A Review on Evaluation Metrics for Data Classification Evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, pp. 01–11, Mar. 2015. [Online]. Available: http://www.aircconline.com/ijdkp/V5N2/5215ijdkp01.pdf

[7] J. N. Padi and V. Setty, *Machine Learning Approaches to Predicting Sea Freight Carbon Emissions : A Comparative Evaluation of Decision Trees, Support Vector Machines, and Random Forests*, 2024. [Online]. Available: https://urn.kb.se/resolve?urn=urn:nbn:se:bth-26710

[8] A. Patle and D. S. Chouhan, "SVM kernel functions for classification," in *2013 International Conference on Advances in Technology and Engineering (ICATE)*, Jan. 2013, pp. 1–9. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6524743?casa_token=4S_i-YAeRcgAAAAA:5AyAgPEclZgpJH_RMpk4yCj6D3Z3Eo4rvRQLJfOYY4oBysOnxz5OlkcnBrXCEKWXIXip253dqA

[9] 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

[10] A. Shi, "Population Growth and Global Carbon Dioxide Emissions." [Online]. Available: https://iussp.org/sites/default/files/Brazil2001/s00/S09_04_Shi.pdf

[11] T. Tr, "Dimensionality Reduction: A Comparative Review." [Online]. Available: https://members.loria.fr/moberger/Enseignement/AVR/Exposes/TR_Dimensiereductie.pdf

[12] 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

[13] Y. Zhang, "Support Vector Machine Classification Algorithm and Its Application," in *Information Computing and Applications*, C. Liu, L. Wang, and A. Yang, Eds. Berlin, Heidelberg: Springer, 2012, pp. 179–186. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-34041-3_27

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