Quantitative Analysis and Forecasting of Industrial CO_2 Emissions Using Multiple Machine Learning Models

Goenka Neev

Researcher, School Student

Abstract:

In response to escalating climate concerns, precise industrial Carbon Dioxide (CO2) emissions prediction is paramount. Employing advanced Machine Learning (ML) techniques, this study focuses on forecasting industrial CO2 emissions using global data from the Our World In Data Dataset (containing information on annual emissions from cement, coal, flaring, gas, and oil industries). Various regression models including Support Vector Regression (SVR), Linear Regression, and XGBoost were explored, with a primary emphasis on time series forecasting models for yearly CO2 emissions. Leveraging time series forecasting, intricate temporal trends in emissions data are discerned, offering enhanced predictive insights. CO2 prediction literature was reviewed, data collected and preprocessed, and various ML algorithms implemented, followed by hyperparameter tuning. The models, rigorously trained and evaluated, yield accurate emission predictions. Results highlight the superior performances of the Transformer model and the Neural Prophet Library developed by Stanford University in collaboration with Facebook Inc., with RMSE scores of 416.58 and 470.30, impressively low MAPE scores of both 0.01, and relatively lower MAE of 349.07 and 380.40 compared to other tested models. DeepTCN also demonstrates competitive predictive capabilities but falls short of Transformer model and Neural Prophet model accuracy. Traditional models including ARIMA, Naive Forecasting, Auto Regression (AR), Exponential Smoothing, and SARIMA lag in performance compared to both Neural Prophet and Transformer. These findings underscore the promising role of ML in advancing sustainable environmental management and pave the way for subsequent research endeavors.

Keywords: CO2 emissions, Industrial Emissions, Sustainability, Environmental AI, Machine Learning, Time series forecasting.

I. Introduction:

A. Background of the project

Global warming results from rising greenhouse gasses, notably CO_2 (Denchak M., 2023). Unchecked emissions jeopardize capping warming at 1.5 or 2 degrees Celsius (US EPA,2023). While current climate policies are slowing the growth of carbon emissions, more transformative measures are required to halt rising emissions and attain carbon neutrality by 2050 (Yao & Zhang, 2022). To act, we must grasp environmental shifts, aligning progress with conservation (IPCC, 2021).

CO_2, a major greenhouse gas, poses complex challenges like weather extremes and biodiversity loss. Predicting emissions is vital. It equips policymakers with data-driven insights to steer climate action.

At the forefront of global carbon emissions, the energy spent in the industry sector takes the lead when it comes to the amount of emissions released due to energy production, accounting for a substantial 24.2%. To this percentage, if we add the contributions of direct industrial processes, such as cement (3%) and chemicals and petrochemicals (2.2%), we have that the industry sector is responsible for a total of 29.4% of GHG global emissions, thus making it a pressing sector to be transformed (Richie et al, 2020). This is a multifaceted problem, since emissions are not just derived from the burning of fossil fuels for energy production, but are also inherent to the manufacturing of widely used materials, such as steel and cement, where CO_2 is a byproduct (IPCC, Chapter 7: Industry, 2007).

In recent years, the urgency of addressing climate change has intensified, necessitating the adoption of more accurate and data-driven methods, such as Artificial Intelligence (AI). AIs are well known for their reliable results in forecasting, predicting as well as estimating. Infact, Neural Networks (NNs) are the most popular method in forecasting CO_2 emissions and account for 14.29% of journal articles in this topic (Abdullah & Pauzi, 2015)[1].

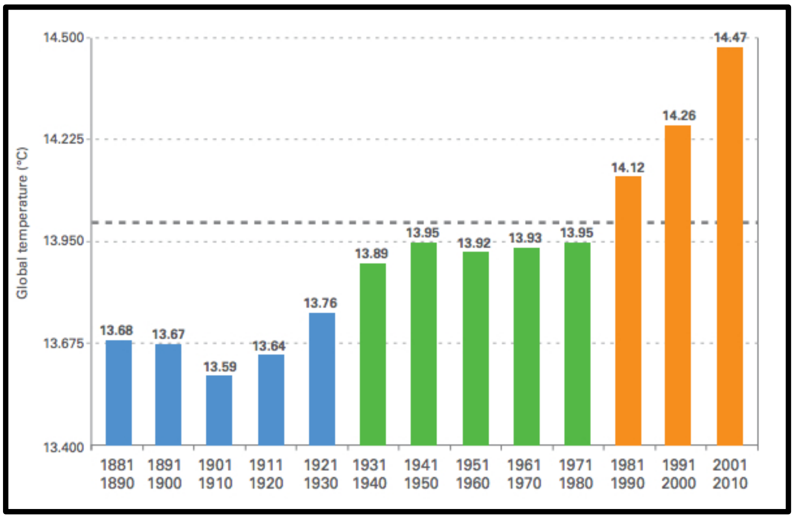


Figure 1. Global Warming.

Notably, prior research has demonstrated the effectiveness of ML algorithms like exponential Gaussian Process Regression (GPR) (Ma et al., 2021) and Autoregressive Integrated Moving Average (ARIMA) models (Gopu et al., 2021) in providing accurate predictions of CO_2 emissions. These models have proven their worth in capturing the temporal characteristics of emissions data.

We've selected specific ML models based on our project's goals and insights from an extensive literature review. These models align with the temporal nature of emissions data. The first one is ARIMA which is ideal for capturing historical patterns in emissions data due to its expertise in handling temporal dependencies, seasonality, and trends. The second, Seasonal Auto-Regressive Integrated Moving Average (SARIMA) extends ARIMA to consider the seasonality aspect of the data. Lastly, Prophet excels at handling data with complex seasonality, a common feature in emissions datasets. It's also adept at managing missing data and outliers, further enhancing our predictions[9].

Our project's core focus is using ML to predict Industrial CO2 Emissions, filling a critical research gap. We rely on the OWID dataset, spanning in global level to understand emissions dynamics[22]. In a similar vein, recent research employed the OWID dataset to investigate CO_2 emissions specifically in the context of the Middle East (Naseef et al, 2023).

Our journey began with in-depth Exploratory Data Analysis(EDA), uncovering hidden trends. These insights shaped our foundational regression models, laying the groundwork for this transformative predictive effort.

In summary, our project addresses a significant research gap, using ML and real-world data to predict CO_2 emissions. This paper outlines our objectives, roadmap, and the value of our approach in advancing emissions forecasting and sustainable decision-making.

B. Objective of the project

The objective of our research is to conduct a comprehensive analysis of CO_2 emissions data from various sources, with a focus on emissions from different industries, to gain insights into historical trends, industry contributions, and the impact of external events. Additionally, the objective includes the development and evaluation of ML models, including time series forecasting techniques, for accurate CO_2 emissions forecasting. The document aims to assess the performance of these models using relevant metrics and address anomalies in the data to ensure robust and reliable emissions forecasting. The ultimate goal is to contribute to a greener future by providing valuable insights and tools for managing and mitigating the impact of rising carbon emissions.

C. Scope of the project

Our research focuses on predicting total CO_2 emissions on a global scale, as well as the CO_2 emissions from specific industrial processes, including industries such as coal, cement, flaring, gas, oil, and others with significant carbon outputs. We limit the scope to historical data analysis and prediction, leaving the implementation of emission reduction strategies outside the immediate purview of this project.

The study's scope includes the collection and preprocessing of relevant data from diverse industrial sources, feature engineering to identify critical predictors and the application of various ML algorithms for prediction. We aim to evaluate and compare the performance of different models, allowing us to identify the most effective methods for accurate emission predictions.

It is essential to recognize that while our research contributes valuable insights, the prediction of CO_2 emissions is inherently complex, influenced by various dynamic factors beyond the industrial domain, such as policy changes, economic shifts, and technological advancements. By leveraging our models, we aim to bridge the gap between innovation and practical emissions management, empowering data-driven decision-making for emissions reduction and sustainability. As such, our project acknowledges and addresses these challenges within the defined scope while setting the groundwork for future research and collaborative efforts to refine and expand our methodology.

In summation, the convergence of ML and environmental sustainability presents a promising opportunity to address the critical issue of industrial CO_2 emissions. By outlining the background, objectives, and scope of our project, we aim to make a meaningful contribution to the ongoing efforts toward a greener and more sustainable future.

II. Data Collection and Preprocessing

A. Data Collection

This research focuses on analyzing global CO_2 emissions attributed to the industrial sector. The study involved an extensive search for a dataset that offers consistent and continuous measurements spanning multiple years. While various CO_2 emissions datasets were considered, most lacked a detailed breakdown of sector-specific contributions, especially within the industrial sector. After thorough investigation, four datasets were initially selected for evaluation. The Emission Database for Global Atmospheric Research

(EDGAR) dataset, covering 1971 to 2021, accounted only for emissions from the oil and gas industry[7]. The Intergovernmental Panel on Climate Change (IPCC) dataset concentrated on overall energy emissions, including those from oil and gas[18]. The Carbon Monitor dataset, with daily CO_2 emissions data, was available only from 2019, limiting its utility for identifying long-term trends[5]. The dataset that came closest to meeting the research objectives was the Our World In Data (OWID) Dataset[22]. This dataset provides yearly CO_2 measurements for various industrial sectors like cement, coal, flaring, gas, and oil, from 1750 to 2021 on a global scale.

We chose not to include cumulative industrial CO_2 emissions or consumption-based metrics in our model. This decision was based on the need to gain insights into yearly fluctuations in CO_2 emissions, whether they were increasing or decreasing. Cumulative metrics tend to obscure specific annual changes because they aggregate past data up to the present. Additionally, we did not consider consumption-based metrics for this study. Our primary focus remained on the industrial sector, enabling a more precise assessment of emissions originating from industrial activities.

B. Data Cleaning

The OWID dataset contains extensive information on total CO_2 emissions and Coal CO_2 emissions dating back to 1750. However, data for other industries like cement is available only from 1880, with 47.79% missing values. Oil-related CO_2 emissions data starts from 1855 with 38.6% missing values, gas-related data from 1882 with 48.53% missing values, and flaring-related data from 1950 with 73.53% missing values.

Column Name	Missing Values Percentage
Year	0.00%
Total CO_2	0.00%
Cement CO_2	47.79%
Coal CO_2	0.00%
Flaring CO_2	73.53%
Gas CO_2	48.53%
Oil CO_2	38.60%

Table 1. Percentage of Missing Values in OWID by Industry Since 1750

Given that original dataset values are available from 1880 for all columns except gas and flaring, our analysis has concentrated on data from this year onwards. This approach is driven by the fact that emissions from most industries were insubstantial before this point in time due to their limited prevalence and integration within mainstream practices.

Column Name	Missing Values Percentage
Year	0.00%

Total CO_2	0.00%
Cement CO_2	0.00%
Coal CO_2	0.00%
Flaring CO_2	49.30%
Gas CO_2	1.41%
Oil CO_2	0.00%

Table 2. Percentage of Missing Values in OWID by Industry Since 1880

To address missing gas and flaring values since 1880, we chose to fill them with zeros. This decision aligns with historical context, as emissions from these industries were negligible during that period. The consistent proximity to zero (in million tonnes) for emissions values since data collection began for these industries further supports this choice. We intentionally avoided using alternative imputation techniques like mean, random, or mode to maintain the accuracy of the dataset and reflect historical realities.

Exploratory Data Analysis

EDA is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods[12]. This process aids in comprehending the data, revealing patterns, trends, relationships, and anomalies within it.

We performed a comprehensive EDA of the OWID dataset, focusing on historical CO_2 emissions. We also examined CO_2 emissions within specific industries like coal, cement, gas, oil, and flaring to gain insights into yearly fluctuations. This approach allowed us to visualize how CO_2 emissions evolved in each industry over time. Throughout our analysis, we used various graphical tools such as line plots, box plots, and heatmaps to effectively communicate our findings.

Descriptive Statistical Analysis

Descriptive statistics is the process of using and analyzing the summary statistic that quantitatively describes or summarizes features[16].

From the Summary Table and Boxplot, we can deduce that the contribution by different industries on CO_2 emissions is in this order - Coal, Oil, Gas, Cement and Flaring with coal being the major contributor to flaring least one.

	Year	Total CO_2	Cement CO_2	Coal CO_2	Flaring CO_2	Gas CO_2	Oil CO_2
Data Count	142.000	142.000	142.000	142.000	142.000	142.000	142.000
Mean	1950.500	12106.835	317.373	5538.816	129.960	1789.140	4263.348
Standard Deviation	41.136	11314.988	453.690	3977.845	151.931	2320.239	4513.949

Minimum	1880.000	853.705	0.000	838.340	0.000	0.000	15.364
Quartile 1 (25%)	1915.250	3173.079	0.148	2838.551	0.000	32.648	201.486
Medium	1950.500	6191.534	71.409	3967.049	75.618	384.671	1695.332
Quartile 3 (75%)	1985.750	20551.891	435.955	8306.316	250.716	3102.437	8967.448
Maximum	2021.000	37123.852	1672.592	15051.513	439.254	7921.830	12345.653

Table 3. Descriptive Analysis of CO^2 Emissions by Industry

Following image shows a box plot analysis of the CO2 Emission.

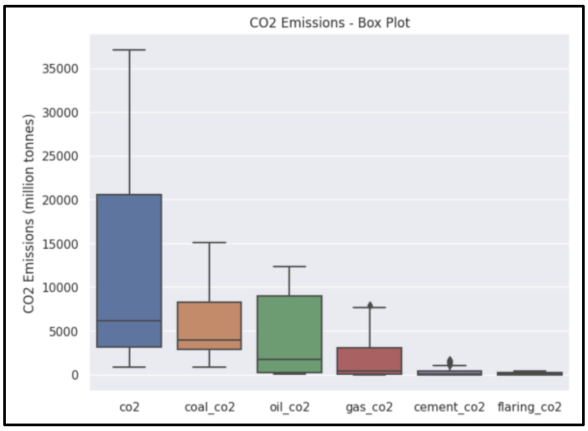


Figure 1. CO$_2$ Emission Box Plot

Univariate Analysis

1. Total CO2 Emissions

The data vividly portrays a substantial increase in global CO$_2$ emissions since 1880. Over the course of 140 years, emissions surged from 853.71 million tonnes in 1880 to a staggering 37,123.85 million tonnes in 2021, marking a more than 43-fold rise. Despite temporary declines in emissions after events like World War II, the 2008-2009 recession, and the COVID-19 pandemic, these reductions were short-lived. The overall trend has been consistent growth in CO$_2$ emissions, highlighting the formidable challenge of effectively curbing and controlling them.

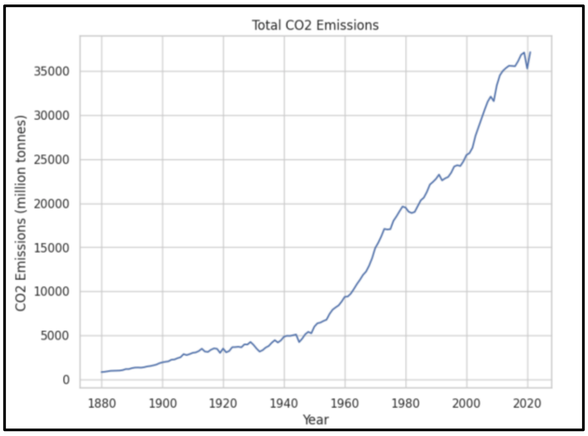


Figure 2. Total CO2 Emissions

Bivariate Analysis

1. Coal vs. Total CO2 Emissions

Coal has historically been a major contributor to CO_2 emissions. Between 1880 and 1920, coal significantly dominated CO_2 emissions, surpassing contributions from other industries. During this period, coal emissions nearly matched total CO_2 emissions, while other industries played a minor role. After 1920, a clear divergence emerged between total CO_2 emissions and those from coal, indicating the increasing involvement of other industries and sources in CO_2 emissions.

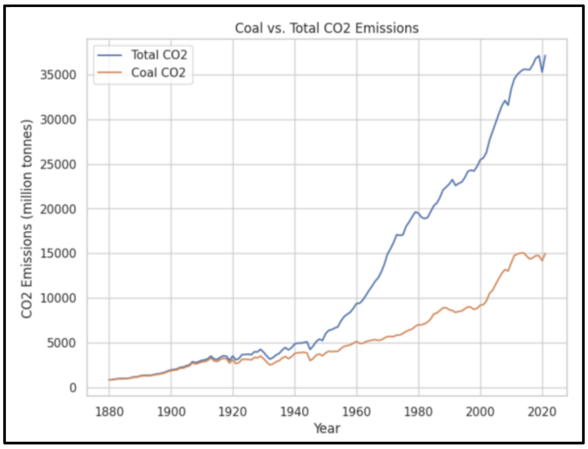


Figure 3. Coal vs. Total CO_2 Emissions

2. Cement vs. Total CO2 Emissions

Between 1880 and 2021, the cement industry's maximum contribution to CO_2 emissions reached 1672.59 million tonnes, constituting a relatively small portion of the total CO_2 emissions. While there was a noticeable increase in the cement industry's contribution starting around 1960, it still represents a minor percentage of the total emissions.

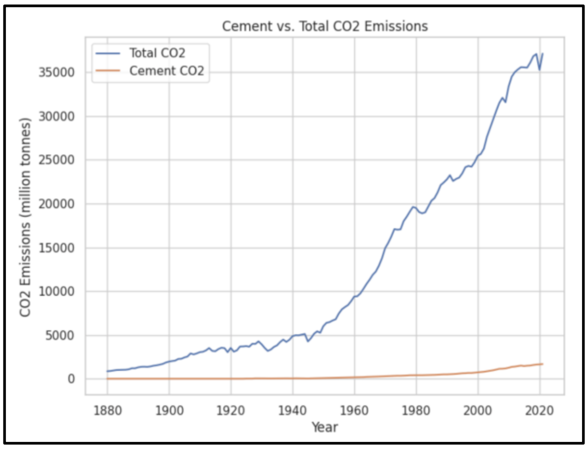


Figure 4. Cement vs. Total CO2 Emissions

3. Oil vs. Total CO2 Emissions

Between 1920 and 2021, CO_2 emissions from the oil industry surged dramatically, increasing by 33-fold. They rose from 350.95 million tonnes in 1920 to a staggering 11,837.16 million tonnes in 2021. Notably, there was a distinct drop in oil-related CO_2 emissions during the 1980s, likely linked to historical events like the Iran Revolution and the Iran-Iraq war, which led to soaring oil prices and a global economic slowdown[26]. This caused a temporary reduction in oil-related emissions, but this decline was short-lived, as emissions from oil resumed their upward trend afterward.

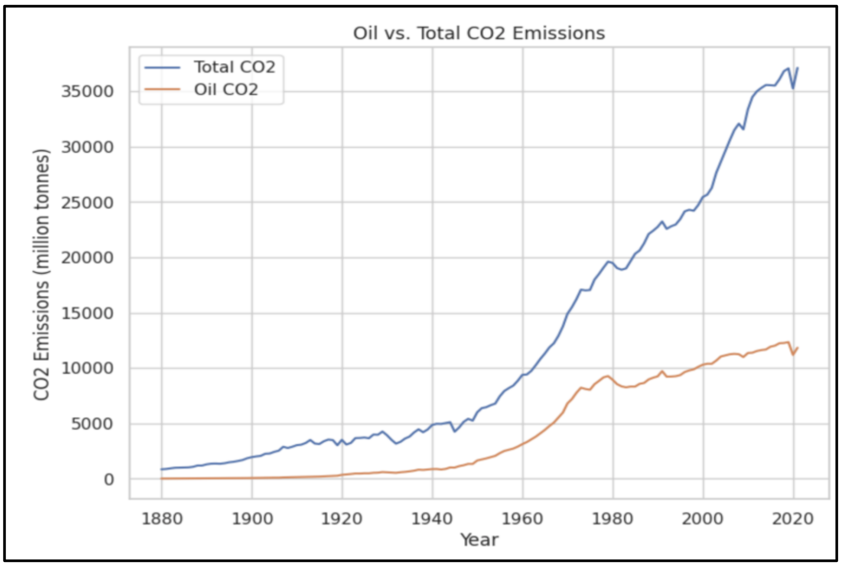


Figure 5. Oil vs. Total CO_2 Emissions

4. Gas vs. Total CO2 Emissions

The graph shows a significant increase in CO_2 emissions from the gas industry after 1940. Between 1940 and 2021, gas-related CO_2 emissions increased by a remarkable factor of 51, rising from 153.51 million tonnes in 1940 to a substantial 7921.83 million tonnes in 2021. Throughout this period, these emissions consistently followed an upward trend, with occasional minor fluctuations.

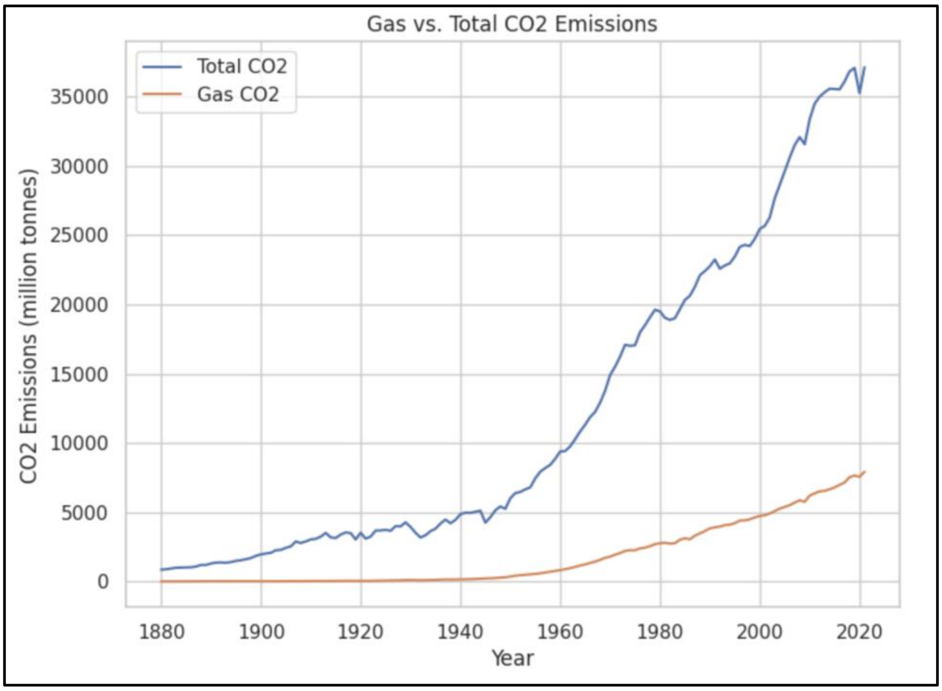


Figure 6. Gas vs. Total CO_2 Emissions

5. Flaring vs. Total CO2 Emissions

Flaring has had a relatively minor impact on CO_2 emissions compared to other industries. Nevertheless, it's worth noting that flaring emissions have increased significantly, rising from 73.63 million tonnes in 1950 to 416.53 million tonnes in 2021, marking a fivefold increase.

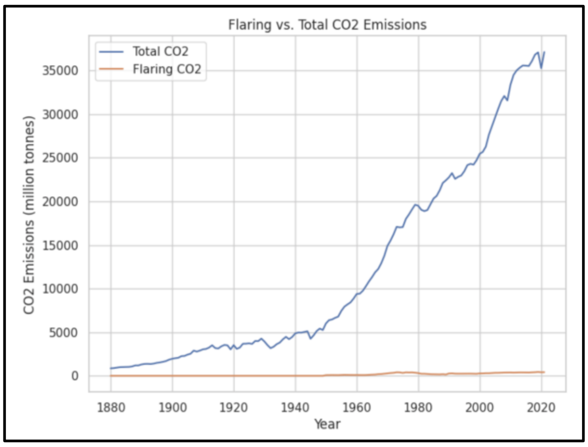


Figure 7. Flaring vs. Total CO_2 Emissions

Multivariate Analysis

1. Comparison of different industries CO2 Emissions

Between 1880 and 2021, CO_2 emissions from various industries have consistently risen with occasional minor declines. Coal historically played a dominant role in these emissions. However, a notable shift occurred in 1968 when the oil industry became the leading contributor to CO_2 emissions until 2004. After that, coal resumed its position as the primary contributor among these industries.

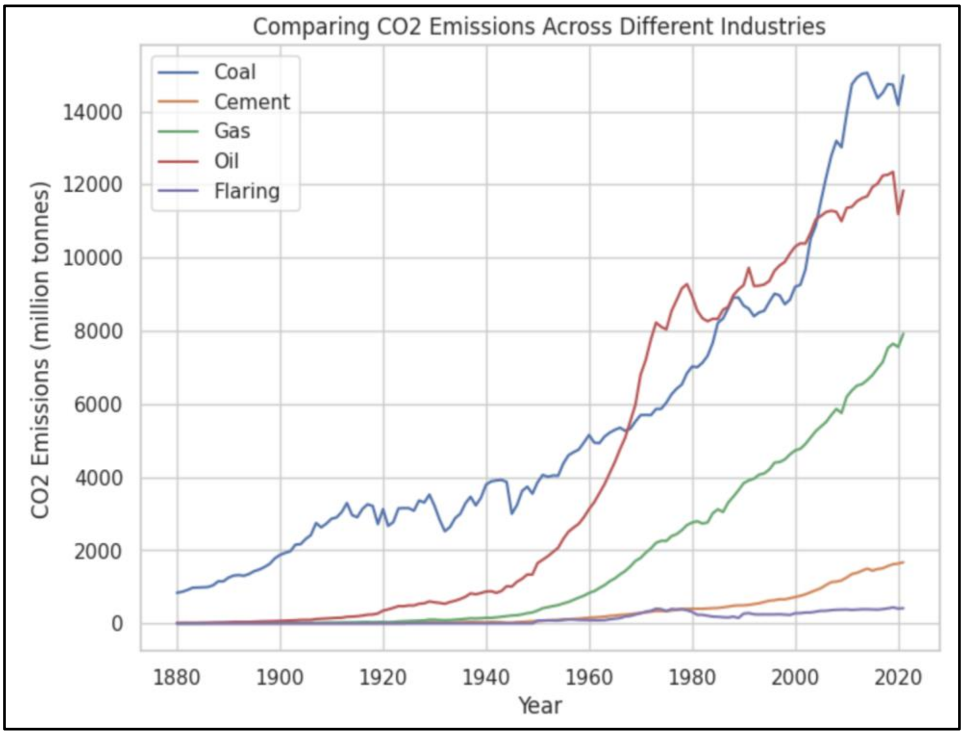
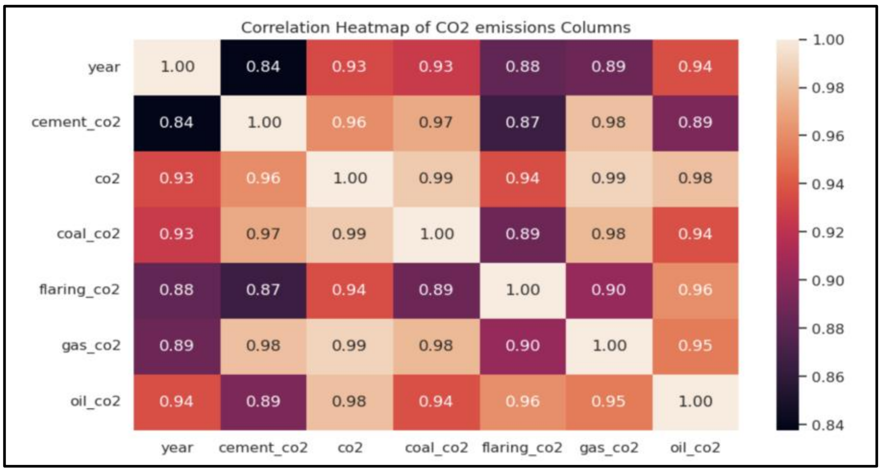


Figure 8. Comparison of CO_2 Emissions Across Industries

2. Correlation using heatmap

The heatmap clearly illustrates a direct correlation between CO_2 emissions and emissions from various industries such as coal, oil, cement, gas, and flaring. Each of these industries exhibits a correlation of over 94% with CO_2 emissions. Notably, the correlation is exceptionally high, reaching up to 99%, for both coal and gas industries. Furthermore, even the CO_2 emissions originating from different industries display correlations, with a minimum value of 87%.



	year	cement_co2	co2	coal_co2	flaring_co2	gas_co2	oil_co2
year	1.00	0.84	0.93	0.93	0.88	0.89	0.94
cement_co2	0.84	1.00	0.96	0.97	0.87	0.98	0.89
co2	0.93	0.96	1.00	0.99	0.94	0.99	0.98
coal_co2	0.93	0.97	0.99	1.00	0.89	0.98	0.94
flaring_co2	0.88	0.87	0.94	0.89	1.00	0.90	0.96
gas_co2	0.89	0.98	0.99	0.98	0.90	1.00	0.95
oil_co2	0.94	0.89	0.98	0.94	0.96	0.95	1.00

Figure 9. Correlation Heatmap of CO_2 Emission Sources

EDA Findings

Our in-depth analysis of OWID data spanning from 1880 to 2021 yields several significant insights:

1. CO_2 emissions have increased dramatically by a factor of approximately 43 over this period, with occasional minor declines.
2. Coal is the leading contributor to CO_2 emissions in 2021 (40.4%), followed by CO_2 (31.9%), Gas (21.3%), Cement (4.5%), and Flaring (1.1%).
3. Historically, coal was the main source of CO_2 emissions, but in 1968, the oil industry became the primary contributor, maintaining this position until 2004, illustrating changing contributors over time.
4. A strong correlation exists between overall CO_2 emissions and emissions from specific industries, emphasizing the connection between industrial activities and carbon footprints.
5. Reduced CO_2 emissions coincide with significant events like geopolitical shifts and pandemics, highlighting the impact of external factors on emission trends.
6. Achieving substantial CO_2 emission reductions under standard conditions is challenging, emphasizing the need for proactive measures to address this issue.

In conclusion, our comprehensive analysis of OWID data over more than a century provides valuable insights into the trajectory of CO_2 emissions, the varying roles of different industries, and the imperative to take proactive steps in managing and mitigating the impact of rising carbon emissions.

Methodology

This study initially employed various regression models, including Linear Regression, Polynomial Regression, SVR/SVM, Random Forest, and XGBoost. While these models provided valuable insights into the contributing factors, they proved to be inadequate for precise emission forecasting. Consequently, our research direction shifted towards the utilization of time series forecasting methods.

Subsequently, our focus shifted towards leveraging time series forecasting methods. Our exploration extended across a multitude of models like Naive Forecasting, AutoRegression (AR), DeepTCN, ARIMA,Exponential Smoothing, SARIMA and Neural Prophet.This transition to time series forecasting techniques marked an exploratory journey that uncovered the intricate temporal dynamics of CO_2 emissions.

Each of these models brought a unique perspective to our analysis. Naive Forecasting provided a baseline reference, AR revealed lagged dependencies, DeepTCN demonstrated the power of deep learning architectures, ARIMA captured trend and seasonality, Exponential Smoothing addressed data smoothing, SARIMA refined seasonal forecasting, and Neural Prophet delivered adaptability and performance, all contributing uniquely to our comprehensive understanding and accurate forecasting of CO_2 emissions.

Model Training and Evaluation

Model Training

In the pursuit of accurate CO_2 emission forecasting, we prioritized robust validation methods to ensure the credibility of our models. To mitigate the risks of overfitting and data-related challenges, we methodically partitioned our dataset into three distinct segments: a training set spanning from 1880 to 2000, dedicated to model training; a validation set covering the years 2001 to 2011, utilized for fine-tuning and evaluating model performance; and lastly, a test set encompassing the period from 2012 to 2021, employed for the ultimate assessment of our model's capabilities.

The training set allowed our models to grasp data patterns, while the validation set helped us adjust model parameters and identify optimal settings. To maintain data integrity, we strictly separated the validation and test sets. The validation set was solely used for model tuning, while the test set remained untouched until the final model evaluation.

By employing these techniques, we fortified our forecasting models against data problems, ensuring reliable CO_2 emission forecasts. This approach instilled confidence in the accuracy of our predictions, aligning with our goal of contributing to a greener future.

Evaluation Metrics

In our pursuit of accurate CO_2 emissions forecasting, we carefully selected a set of evaluation metrics that aligned with our project's objectives. These metrics included Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE).

We chose these metrics to provide a well-rounded assessment of our forecasting models. MAE helped us understand the average forecasting error, RMSE captured the magnitude of errors while penalizing larger discrepancies, MAPE provided a percentage-based perspective on relative errors.

By using this suite of metrics, we gained valuable insights into different aspects of forecasting accuracy, allowing us to make informed decisions about model selection and refinement.

Results and Discussion

Among the regression models, SVR stands out as the top-performing model with the lowest RMSE (30.72) and MAPE (0.004) values. In contrast, Polynomial Regression, Random Forest, Linear Regression, and XGBoost all show considerably higher RMSE and MAPE values, highlighting their limited effectiveness in modeling complex time-dependent data. Particularly, Linear Regression and XGBoost perform poorly, emphasizing that standard regression methods are ill-suited for time series forecasting.

Models Name	RSME	MAPE	MAE
Linear Regression	647.40	0.0096	342.49
SVR	30.72	0.01	22.90
Polynomial Regression	107.18	0.01	67.39
Random Forest	177.26	0.02	128.79
XGBoost	1656.81	0.0404	1465.88

Table 4. Error Rates From Regression Models

Among the various time series forecasting models evaluated, the Transformer models and the Neural Prophet model emerged as the top performers, delivering the most accurate predictions with RSME's of 416.58 and 444.30 respectively, an impressively low MAPE of both 0.01, and relatively low MAE's of 349.07 and 367.04 respectively. These results suggest these models, with their ability to capture changepoints and incorporate lags effectively, excels in capturing the underlying patterns and trends in the time series data. In contrast, DeepTCN exhibits an RSME of 1002.43, a MAPE of 0.03, and a MAE of 936.85, indicating that while it offers predictive capabilities, it lacks the precision achieved by the Neural Prophet and the Transformer model in this specific CO_2 emissions forecasting task. Notably, models such as ARIMA, Naive Forecasting, AR, Exponential Smoothing, and SARIMA underperformed compared to Neural Prophet and DeepTCN, further highlighting the significance of model selection in achieving robust time series forecasting results.

Models Name	Parameters	RSME	MAPE	MAE
Transformer	Lags = 5	416.58	0.01	349.07
Neural Prophet	Lags=1	444.30	0.01	367.04
DeepTCN	Lags = 5	1002.43	0.03	936.85
ARIMA	(p,d,q) = 4,3,4	1598.56	0.04	1598.56
Naive Forecasting	Best K parameter = 1	1815.88	4.48	1634.85
Auto Regression (AR)	lags = 7	2263.04	0.06	2046.95
Exponential Smoothing	seasonality parameter = 17	2892.56	6.74	2462.70
SARIMA	order (p,d,q) = (1,1,1), seasonal (p,d,q,s) = (1,1,1,12)	2900.00	9.09	2427.60

Table 5. Error Rates from Time Series Forecasting Models

Deployment & CI/CD Integration

An essential aspect of the deployment process was the establishment of a robust Continuous Integration and Continuous Deployment (CI/CD) framework. The core objective of this research was to leverage ML techniques to enhance the accuracy of CO_2 emissions predictions, thus facilitating informed decision-making for sustainability goals.

The deployment strategy was meticulously devised through a thorough and systematic analysis of diverse CD platforms, encompassing Streamlit, Hugging Face + Gradio, Dash, Panel, Flask, and FastAPI. This comprehensive evaluation, expounded upon in the referenced source[13], encompassed an assessment of key performance indicators such as operational metrics, scalability potential, processing velocity, and user-friendliness. Post-evaluation, the shortlisted platforms, namely Tf Serving, Hugging Face + Gradio, and Streamlit, were distinguished for their exceptional proficiencies. Ultimately, Streamlit emerged as the preferred choice for deployment, a decision substantiated by its seamless integration with GitHub, uncomplicated deployment mechanisms, and customizable attributes. This selection, albeit not attaining zenith with regard to real-time capabilities and scalability, was weighed against its pragmatic merits.

The deployment process encompassed several key steps, beginning with the utilization of an Neural Prophet model for initial testing, implemented through the Statsmodels library[23]. Serialisation of the model was facilitated using the Pickle module[20], allowing for efficient loading and reusability. The resultant dashboard application was constructed using Dash and Plotly libraries[8], thereby providing an interactive and visual interface for users to engage with the predictions and recommendations.

To ensure seamless accessibility and utilization, a CI/CD pipeline was deemed indispensable. Leveraging the Render platform, the deployment was orchestrated with auto-deployment from a GitHub repository, bolstered by free TLS certificates, a global Content Delivery Network (CDN), and private networks[21]. The CI/CD workflow was established through steps that encompassed the creation of 'requirements.txt' to specify dependencies, the integration of Gunicorn in the requirements list, and the formulation of 'render.yaml' to define deployment settings. Notably, the 'render.yaml' creation process was partly automated using Dash-Tools, but conflicts with Statsmodels necessitated its exclusion from 'requirements.txt'. The deployment process culminated with the initiation of the application using the Gunicorn command.

The significance of CI/CD in this research context was underpinned by its role in ensuring a streamlined and automated deployment pipeline, thus bridging the gap between model development and real-world accessibility. This approach not only enhanced the accessibility of the CO_2 emissions prediction and recommendation system but also empowered decision-makers to make informed choices in their pursuit of emissions forecasting and sustainable industrial practices.

Discussion

Practical Applications and Model Implications

The attainment of the stated research objectives, centered around Predicting Industrial CO_2 Emissions with ML and the creation of a Recommendation System for CO_2 emissions forecasting, holds paramount significance in addressing the multifaceted challenges posed by global warming and its implications. The successful realization of these objectives will usher in a range of practical applications and far-reaching implications across various sectors, offering substantial advancements in environmental management, industrial sustainability, business decision-making, climate accord compliance, scientific progress, and accountability measures.

Improved Environmental Management and Policy Formulation

The proposed ML model's ability to accurately predict industrial CO_2 emissions bears the potential to revolutionize environmental management and policy formulation. By providing real-time insights into emission trends and trajectories, policymakers and regulatory bodies can devise targeted strategies to mitigate carbon footprints and curb greenhouse gas emissions[11]. This enhanced understanding will facilitate the development of evidence-based policies that foster sustainable practices and drive the transition towards a low-carbon economy, aligning with global efforts to combat climate change[15].

Enhanced Industrial Sustainability

The implementation of the research outcomes into industrial operations can yield tangible benefits in terms of enhanced sustainability. Industries can leverage the predictive model to optimize production processes, identify emission hotspots, and fine-tune operational practices to minimize environmental impact[25]. This aligns with the growing corporate emphasis on sustainable production, enabling industries to reduce resource consumption, improve energy efficiency, and diminish emissions. Moreover, informed decision-making driven by the model can foster innovation in clean technologies and eco-friendly manufacturing practices[6].

Informed Business Decision-Making

The fusion of ML driven emissions forecasting and recommendation systems empowers businesses with valuable insights for strategic decision-making. Corporations can integrate emissions projections into their operational planning, enabling proactive adjustments to meet regulatory requirements and market demands[2]. Furthermore, the model's capacity to anticipate emission fluctuations can optimize supply chains, reduce costs associated with emissions penalties, and improve risk management[3]. By embedding sustainability considerations into business strategies, organizations can bolster their competitive edge and contribute to broader environmental stewardship.

Global Climate Accord Compliance

The model's capability to forecast CO_2 emissions aligns harmoniously with international climate accords, such as the Paris Agreement. Accurate emission predictions can assist countries in setting and achieving emissions reduction targets, while also fostering transparency and accountability in reporting[17]. This facilitates collective global efforts to limit temperature rise and mitigate the adverse effects of climate change. The research outcomes can serve as a robust tool for nations to monitor their progress, strengthen commitments, and collaborate towards achieving shared climate objectives.

Scientific Advancement and Data-Driven Research

The integration of ML techniques for emissions prediction contributes to the advancement of scientific knowledge. The model's analysis of intricate emission patterns, coupled with its ability to process vast datasets, can lead to groundbreaking discoveries in climate science and atmospheric dynamics[14]. Researchers can leverage the model's outputs to refine climate models, validate hypotheses, and deepen the understanding of emission drivers. This synergy between data-driven research and predictive modeling catalyzes the evolution of climate science, fostering a more comprehensive comprehension of global warming dynamics.

Accountability Measures

The deployment of the model introduces an avenue for accountability, both within industries and among nations. Accurate emissions forecasting enables proactive monitoring and enforcement of emissions regulations, ensuring compliance with emission reduction targets and standards[9]. This, in turn, fortifies the credibility of environmental policies and encourages a culture of responsible emissions management.

The availability of data-backed insights into emission trends holds potential for legal and regulatory enforcement, thereby establishing a framework for holding entities accountable for their carbon contributions.

Future work

While the OWID dataset provided comprehensive CO_2 emissions data for various industries, it lacked detailed industrial factors for CO_2 emissions. Future work could involve collaborating with industrial and government sources to incorporate specific industrial emission data, integrating external variables specific to industry such as industrial economic indicators, industrial policy changes and industrial technological advancements. This will enhance the model's ability to capture the complexity of emissions sources and provide additional context leading to accuracy of CO_2 emission predictions.

Addressing missing values is an ongoing challenge in dataset analysis. While our current approach of filling missing values with zeros is reasonable for industries with historical insignificance, we can delve deeper by leveraging data from similar regions or industries which can lead to a more accurate representation of emissions trends.

Furthermore, we have tried to make multivariate predictions including all the features using the Transformer models and LSTMs. Nevertheless, this approach yielded notable outcomes for certain features, while demonstrating relatively subdued performance for other distinct features. Remarkably, each individual model exhibited strengths and weaknesses in predicting specific features. Strikingly, these underperforming predictions often resulted in the formation of a linear forecasting trend or a consistently flat trajectory, lending a nuanced dimension to the overall forecasting landscape. Being powerful and advanced ML algorithms specifically designed for time series data and to capture temporal dependencies and patterns more effectively, we suggest as future work exploring ways to overcome these limitations, as these algorithms rightfully used have the potential to yield even better results to the ones obtained. ## The efficacy of advanced machine learning algorithms tailored for time series data, including gated recurrent units (GRU) and transformer-based architectures, has been established in effectively capturing temporal dependencies and intricate patterns, consequently resulting in successful forecasting outcomes. However, to unlock their full potential, a more concentrated dataset, such as daily or weekly data, is imperative to further optimize these models and extract their temporal dependencies even better in order to achieve optimal performance.##

Future relevance is crucial for any research and machine model. While our current approach maintains relevance for predicting future CO_2 levels, in a few years, it might become less reliable. Therefore, there is a need for future work to establish a real-time monitoring system that continuously updates the model with the latest emission data. This feedback loop can help the model adapt to changing conditions, providing more up-to-date and accurate predictions for policy and decision-makers. It is also important for the model to adapt and investigate online learning techniques from new data as it becomes available, to learn and stay relevant in a dynamically changing environment.

In the future, integrating Geographical Information Systems (GIS) data can play a very vital role to account for regional differences in emissions patterns. This spatial analysis enriches the accuracy of predictions by capturing localized factors that influence emissions trends.

Collaboration with experts from various fields, such as climate science, economics, and policy, can enrich our understanding. Their insights can help refine the model's inputs and assumptions, leading to more comprehensive and accurate predictions.

In closing, while our current study has illuminated key aspects of CO_2 emissions prediction, the future beckons with opportunities for refinement and expansion. Embracing these directions can contribute to a deeper understanding of emissions dynamics, better-informed decision-making, and a more sustainable future for our planet. As researchers, our commitment to progress remains steadfast, driven by the potential to make a positive impact on our global environment.

Parameter Recommendations

Based on the results obtained from our ML models for predicting industrial CO_2 emissions, we recommend the following parameters to enhance model accuracy and generalization capabilities -

- **Temporal Granularity:** Choose an appropriate temporal granularity (e.g., hourly, daily, monthly) for training the model. Finer granularity provides more detail but can introduce noise, while coarser granularity might miss short-term variations.
- **Handling Seasonality:** Address seasonality in the data. Models like ARIMA and Prophet handle this inherently, but for others, consider decomposing the time series data to explicitly account for seasonality. Tune seasonality parameters based on observed periodic fluctuations.
- **Hyperparameter Tuning:** Regularly tune model hyperparameters. For ARIMA and SARIMAX, optimize parameters like p, d, q, and their seasonal counterparts using techniques like grid search or random search. For the Prophet, fine-tune parameters related to seasonality, holidays, and change points.
- **Training-Validation Split:** Because of the time-series nature, avoid traditional cross-validation and opt for a rolling-window or expanding-window validation approach. This ensures validation on data that follows the chronological order of the training data.
- **Regular Model Updates:** As new emissions data becomes available, retrain models to reflect recent patterns and trends. Establish a schedule for updates, such as monthly, quarterly, or annually, based on data acquisition rates and prediction requirements.
- **Regular Model Evaluations:** Choose an appropriate evaluation metric (e.g., RMSE or MAE) based on emissions data characteristics. Consistently use this metric to assess model performance over time and to compare different models or versions.

Adhering to recommended parameters can enhance predictive model accuracy and robustness, enabling more precise forecasting and impactful emission predictions. It's important to note that parameter tuning is an ongoing, iterative process requiring regular reviews and adjustments to accommodate evolving data and changing business or research needs.

Conclusion

In conclusion, we delved into the intricate task of predicting CO_2 emissions using sophisticated ML techniques, deriving insights from diverse datasets and sourcing data from the prominent OWID database. This exploration spanned over a century of global carbon emission dynamics. Our initial efforts centered on EDA, revealing the historical intricacies and key relationships within the data. Strategic handling of missing values ensured a sound analytical foundation. Through statistical analyses, the overwhelming contribution of coal and oil to emissions became evident, and we discerned industry correlations as well as their evolving roles over time.

Our exploration extended into the realm of predictive modeling, spanning from conventional regression like SVM, Polynomial Regression to the various time series forecasting methods like Transformers, Neural Prophet, DeepTCN, ARIMA, SARIMA, among others. The selection of models was a meticulous process guided by robust evaluation criteria. Performance metrics such as MAE, RMSE, and MAPE played a pivotal role in assessing each model's efficacy. Between the regression models, the SVR exhibited exceptional performance, showcasing the lowest RMSE of 30.72, an impressively low MAPE of 0.004 and a R2 score of 0.999 which we can not forget to mention. This is showing us its ability to capture the correlations between the different features and data points, but underscoring its proficiency in handling complex time-dependent data and performing well on unseen data, as we could expect from a regression model. Among these methodologies explored and experimented with, as stated earlier, the Transformers and Neural Prophet models emerged as formidable frontrunners with impressively low RSME's, low MAPE's and low MAE's as well, showcasing their ability to predict and forecast values closely located to the actual values. As well as their ability to adeptly capture changepoints and incorporate lags demonstrated its capacity to extract nuanced patterns from the CO_2 emissions data.

By highlighting historical emission trends and industry contributions, our research not only underscores the urgent need for intervention in carbon management but also charts a course for future investigations. Ultimately, our findings emphasize the power of data-driven insights in steering our world towards a sustainable trajectory.

References

1. Abdullah, Lazim, and Herrini Mohd Pauzi. “Methods in forecasting carbon dioxide emissions: A decade review.” Jurnal Teknologi, vol. 75, no. 1, 23 June 2015, https://doi.org/10.11113/jt.v75.2603.
2. Adebowale, Kayode, and Robin Uzel. “Forecasting CO2 Emissions in Sweden with a Bayesian Neural Network.” Diva-Portal, KTH Royal Institute of Technology, 2023, www.diva-portal.org/smash/get/diva2:1773402/FULLTEXT02.pdf.
3. Alpan, Kezban, et al. “Design and simulation of a global model for carbon emission reduction using IOT and Artificial Intelligence.” Procedia Computer Science, vol. 204, 2022, pp. 627–634, https://doi.org/10.1016/j.procs.2022.08.076.
4. Bernstein, Lenny, and Joyashree Roy. “Industry - IPCC.” IPCC, www.ipcc.ch/site/assets/uploads/2018/02/ar4-wg3-chapter7-1.pdf. Accessed 2 Feb. 2024.
5. “Carbon Monitor.” Carbon Monitor, carbonmonitor.org/. Accessed 2 Feb. 2024.
6. “Climate Change 2021: The Physical Science Basis.” IPCC, 2021, www.ipcc.ch/report/ar6/wg1/.
7. Crippa M. “CO2 Emissions of All World Countries.” EDGAR, EDGAR, 2022, edgar.jrc.ec.europa.eu/report_2022?vis=pop#data_download.
8. “Dash Python User Guide.” Plotly, dash.plotly.com/. Accessed 2 Feb. 2024.
9. Degot, Charlotte, et al. “Reduce Carbon and Costs with the Power of AI.” BCG Global, BCG Global, 6 Nov. 2023, bcg.com/publications/2021/ai-to-reduce-carbon-emissions.
10. Delhotal, Casey, and Jochen Harnisch. “Mitigation of Climate Change.” IPCC, 2007, archive.ipcc.ch/publications_and_data/ar4/wg3/en/ch7-en.html.
11. Denchak, Melissa. “Greenhouse Effect 101.” NRDC, 5 June 2023, www.nrdc.org/stories/greenhouse-effect-101.
12. “Exploratory Data Analysis.” EPA, United States Environmental Protection Agency, 10 Aug. 2023, www.epa.gov/caddis-vol4/exploratory-data-analysis#:~:text=Exploratory%20Data%20Analysis%20(EDA)%20is,step%20in%20any%20data%20analysis.

13. Goenka, Neev. “Deployment Framework Comparison.” Google Slides, Google, 2023, docs.google.com/presentation/d/1zJoG8ghqw0mP1gtZeC8VOj7F3C-62TvSi7FPNFqD7yk/edit?usp=sharing.
14. Gopu, Pooja, et al. “Time series analysis using Arima model for air pollution prediction in Hyderabad City of India.” Advances in Intelligent Systems and Computing, 2021, pp. 47–56, https://doi.org/10.1007/978-981-33-6912-2_5.
15. “Greenhouse Gases.” EPA, United States Environmental Protection Agency, 14 July 2023, www.epa.gov/report-environment/greenhouse-gases.
16. Hayes, Adam. “Descriptive Statistics: Definition, Overview, Types, Example.” Investopedia, Investopedia, www.investopedia.com/terms/d/descriptive_statistics.asp. Accessed 2 Feb. 2024.
17. Ma, Ning, et al. ``Can machine learning be applied to carbon emissions analysis: An application to the CO2 emissions analysis using gaussian process regression.” Frontiers in Energy Research, vol. 9, 24 Sep. 2021, https://doi.org/10.3389/fenrg.2021.756311.
18. Minx, Jan C., and William F. Lamb. “A Comprehensive and Synthetic Dataset for Global, Regional and National Greenhouse Gas Emissions by Sector 1970-2018 with an Extension to 2019.” IPCC, 20 Dec. 2022, Accessed 2 Feb. 2024.
19. Nassef, Ahmed M., et al. “Application of artificial intelligence to predict CO2 emissions: Critical Step Towards Sustainable Environment.” Sustainability, vol. 15, no. 9, 6 May 2023, p. 7648, https://doi.org/10.3390/su15097648.
20. “Pickle - Python Object Serialization.” Python, 1 Feb. 2024, docs.python.org/3/library/pickle.html.
21. Render, render.com/. Accessed 2 Feb. 2024.
22. Ritchie, Hannah, et al. “CO_2 and Greenhouse Gas Emissions.” Our World in Data, 28 Dec. 2023, ourworldindata.org/co2-and-greenhouse-gas-emissions.
23. “Statsmodels.Tsa.Arima.Model.ARIMA.” Statsmodels, 14 Dec. 2023, www.statsmodels.org/stable/generated/statsmodels.tsa.arima.model.ARIMA.html.
24. “Unprecedented Impacts of Climate Change Disproportionately Burdening Developing Countries, Delegate Stresses, as Second Committee Concludes General Debate | UN Press.” United Nations, United Nations, 8 Oct. 2019, press.un.org/en/2019/gaef3516.doc.htm.
25. Yao, Jiaxiong, and Yunhui Zhao. “Structural breaks in carbon emissions: A machine learning analysis.” IMF Working Papers, vol. 2022, no. 009, Jan. 2022, p. 1, https://doi.org/10.5089/9798400200267.001.
26. Zeidel, Ronen. “Implications of the Iran-Iraq War.” E-International Relations, 13 Oct. 2013, www.e-ir.info/2013/10/07/implications-of-the-iran-iraq-war/.

ANNEXURE

Abbreviations:

ABB.	Explanation
AI	Artificial Intelligence
AR	Auto Regression
ARIMA	Autoregressive Integrated Moving Average
CDN	Content Delivery Network

CI/CD	Continuous Integration and Continuous Deployment
CO2	Carbon Dioxide
EDA	Exploratory Data Analysis
EDGAR	Emission Database for Global Atmospheric Research
GIS	Geographical Information Systems
GPR	Gaussian Process Regression
GRU	Gated Recurrent Units
IPCC	Intergovernmental Panel on Climate Change
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
NN	Neural Networks
OWID	Our World In Dataset
RMSE	Root Mean Squared Error
SARIMA	Seasonal Auto-Regressive Integrated Moving Average
SVM	Support Vector Machine
SVR	Support Vector Regression