Journal of Environmental and Agricultural Studies
ISSN: 2710-1401
DOI: 10.32996/jeas
Journal Homepage: www.al-kindipublisher.com/index.php/jeas

JEAS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Machine Learning-Based Prediction of U.S. CO_2 Emissions: Developing Models for Forecasting and Sustainable Policy Formulation

Farhana Rahman Anonna[1], MD Rashed Mohaimin[2], Adib Ahmed[3], Md Boktiar Nayeem[4], Rabeya Akter[5], Shah Alam[6] , Md Nasiruddin[7], and Md Sazzad Hossain[8]

[15]*Master of Science in information technology. Washington University of Science and Technology, USA*

[28]*MBA in Business Analytics, Gannon University, Erie, PA, USA*

[37]*Department of Management Science and Quantitative Methods, Gannon University, Erie, PA, USA*

[4]*Master of Science in Business Analytics, Trine University*

[6]*Master of Science in Information Technology, Washington University of Science and Technology, Alexandria, VA, USA*

Corresponding Author: Farhana Rahman Anonna, **E-mail**: Fanonna.student@wust.edu

| ABSTRACT

The exponential escalation of carbon dioxide (CO_2) emissions in the U.S. presents a pressing environmental challenge with substantial implications for climate change and public health. The principal objective of this study was to devise robust machine learning algorithms particularly designed for forecasting CO_2 emissions in the United States. This focused exclusively on CO2 emission data pertinent to America, reflecting the economic, unique environmental, and regulatory context of the nation. The dataset for analysis consisted of a broad-based set of information focused on the main contributors of CO_2 emissions in the United States, ranging from energy consumption and industrial activity to transportation and historical CO_2 emission data. The energy consumption data included facts on electricity generated, fuel consumed, and absolute energy consumption among different sectors of the economy, and industrial activities information provides data on specific outputs from such processes and their emissions. It also included transportation facts on vehicle trends, fuel intensity, and energy-related emissions associated with the sector. These three datasets have been garnered from reliable resources, including the US. These range from detailed EPA emissions inventories and energy reports from the U.S. The analyst deployed credible algorithms such as Random Forest, Logistic Regression, and Support Vector Classifier which had different strengths that can be leveraged based on characteristics of the dataset. According to their accuracy scores, the Random Forest model led the race compared to the other two models, with a higher accuracy rate. With such large integrations of machine learning predictions into climate policy, great opportunities might develop vis-à-vis sustainable development goals in the USA. Advanced analytics will let the policy analyst capture emission and resource trends with greater insight than ever before into the effectiveness of existing regulations; this will let it plug into the SDGs on Climate Action, Sustainable Cities, and Responsible Consumption. For enhancing environmental monitoring systems' efficiency, environmental planning should be incorporated with machine learning models.

| KEYWORDS

CO_2 Emissions Forecasting, Machine Learning, Environmental Policy, United States, Sustainability, Climate Change

| ARTICLE INFORMATION

ACCEPTED: 01 September 2023 **PUBLISHED:** 28 September 2023 **DOI:** 10.32996/jeas.2023.4.3.12

I. Introduction

Background and Context

The United States is among the top emitters of CO_2 into the atmosphere. Its share in contribution automatically places a greater responsibility in addressing the concerns for environmental sustainability. CO_2 emission from combustion of all types of fossil fuel and industrial activities greatly contributes to climatic change, thereby enhancing extreme weather conditions, sea-level rise, and ecological imbalances. Indeed, over the past many decades, increasing urgency has formed as a result of international accords such as the Paris Climate Accord, coupled with domestic efforts toward the meeting of mid-century targets for net-zero emissions (Aras & Van, 2022). From this perspective, precise forecasting of CO_2 emissions is regarded as crucial in that it helps policymakers shape proactive strategies that balance economic growth with the conservation of the environment. Good forecasts can help foresee the future behavior of the emissions, check the efficiency of the policy measures taken, and spot those sectors on which mitigation efforts are to be concentrated. Traditional statistical techniques, however, suffer from serious difficulties in carrying out emission forecasting due to the complex underlying dynamics brought about by energy use, population growth, changes in technology, and other regulatory effects (Chen et AL., 2023).

Carbon dioxide emissions present one of the most nagging environmental challenges confronting the United States today. The continued increase in atmospheric concentrations of CO_2 forebodes grave implications for climate stability in terms of rising temperatures, extreme weather events, and huge ecological disruption (Alloghani, 2023). With these challenges, CO_2 emissions forecasting becomes crucial in devising effective strategies for mitigating climate change impacts. Policymakers need solid projections to make informed decisions on energy regulation, emission reduction targets, and sustainability initiatives. Without accurate forecasting, efforts to formulate and implement policies to reduce emissions could be misdirected or inefficient and may even worsen the environmental crisis instead of improving it (Abouhawwash et AL., 2022).

Problem Statement

According to Farahzadi & Kioumarsi (2023), traditional methods for emission prediction include linear regression and time-series methods, which have disadvantages in modeling complicated nonlinear relationships that may exist between variables affecting CO_2 emissions. A major weakness is that these methods are confined to assuming fixed functional forms and data distribution, which again results in excessively simplified models that yield poor forecasting performance. Secondly, the ongoing proliferation of high-dimensional, real-time environmental data points heightens awareness of the deficiency inherent in traditional methodologies for the extraction of meaningful relationships from rich data. This growing need for more sophisticated predictive frameworks has placed machine learning in the limelight as a game-changing tool in emission forecasting. Koca & Akkaya, (2023), stated that Machine learning algorithms are known to be exceptionally good at identifying complicated patterns and relationships from data. Therefore, ML algorithms could very well be suitable for overcoming the challenges involved in the prediction of CO_2 emissions. By developing more precise forecasts using ML models, informed decisions and sustainable policy formulation can be facilitated.

Research Objective

The principal objective of this study is to devise robust machine learning algorithms particularly designed for forecasting CO_2 emissions in the United States. Using integrated data sets that involve different drivers of emissions, such as economic indicators, energy use patterns, and technological changes, the paper intends to construct a predictive model with better performance than traditional forecasting techniques. Beyond improving the predictive accuracy, the research aims at extracting actionable insights that could inform policymakers and other stakeholders in the fight against climate change. The ultimate goal is to support the formulation of sustainable policies that go beyond reduction to foster broad environmental resilience and sustainability initiatives.

Scope and Relevance

This study will focus exclusively on CO_2 emission data pertinent to America, reflecting the economic, unique environmental, and regulatory context of the nation. The use of U.S.-specific data will directly apply the insights developed from this analysis to the formulation and implementation of national policies. Besides, the application of machine learning methodologies in this framework underlines an increasing trend in which big data use has an important implication for pursuing environmental sustainability. The greater use of advanced analytics to inform policymakers' decisions demands that the integration of machine learning within the forecasting process is a considerable step toward better environmental policy-making in the fight against climate change. We show, in this paper, how far advanced machine learning has come to forecast emissions for the building of improvements in forward-looking emissions predictions to contribute toward the ongoing discussion on sustainability and environmental stewardship in the United States.

II. Literature Review

CO_2 Emission and Climate Change

Jabeur et al. (2021), reported that carbon dioxide in the atmosphere is one of the major contributors to climate change, whereby emissions have been increasing due to industrial activities, transportation, and energy use for quite a long period in the United States. The current trends in the consumption of CO_2 have been influenced by complex interactions among economic growth, means of energy production, and various regulatory mechanisms. Over the past decades, U.S. emissions have fluctuated with changes from coal to natural gas for electricity generation, the ramping up of renewable energy sources, and the implementation of multiple federal and state policies focused on reducing greenhouse gas emissions. The trends indicate that there is a requirement for constant monitoring and enhanced forecast models that accurately predict future emissions.

Giannelos et al. (2021), asserted that the effects of CO_2 emission are not solely environmental degradation but include serious hazards to public health as well. A higher level of CO_2 forms ground-level ozone, which in turn can contribute to respiratory conditions and other diseases. The high levels of CO_2, furthermore, contribute to climate change, which has severe repercussions: extreme weather conditions, increased sea levels, and disruptions in food and water supplies have major implications for natural ecosystems and human communities. Understanding these complex impacts is critical for framing effective policy responses that reduce emissions while protecting public health.

Moreover, an increase in atmospheric concentration due to global warming enhances the frequency and strength of weather events like hurricanes, droughts, and heat waves. Besides, CO_2 emission aggravates air pollution, which is associated with cardiovascular diseases, especially in the vulnerable section of the population. Apprehending such challenges, therefore, asks for a great understanding of the trends of emissions and the drivers, hence bringing into focus the need for correct predictive models (Kumar, 2023).

Traditional Emission Forecasting Methods

Lee & Tae (2020), postulated that traditional methods include linear regression, time series regression-ARIMA, and time series multivariate studies-VAR models as standard methods for future predictions in matters concerning CO_2 emissions. Most of their bases are driven by historical data aimed at describing developed trends, with ultimate projections to observe future values by assuming linear variable interrelations. While methods like those above have served adequately, there are some observable limitations to such techniques, especially when working with complex information of large panel datasets.

Traditional statistical techniques have relied on linear regression, time series analysis, and econometric modeling for CO_2 emission predictions. Such methods have laid the foundation for understanding the trend of emissions based on historical data and established relationships among various influencing factors. For example, linear regression can show the relationship between economic indicators like GDP growth and resulting emissions, hence offering insight into how economic activities impact carbon outputs (Li & Sun, 2021). However, these traditional approaches are fraught with limitations, especially when applied to complex, large-scale datasets characteristic of climate science. They generally assume linearity and homoscedasticity, which can result in oversimplifications of the underlying relationships (Nassef et al., 2023).

Furthermore, traditional methods cannot easily accommodate the many variables and interactions that may influence CO_2 emissions, such as technological changes in energy efficiency and the implementation of environmental policies. Therefore, reliance on these methods can only result in incorrect forecasts, consequently preventing effective emissions reduction strategies from being developed. This points out the urgent need for more sophisticated modeling techniques that can capture the complexity of the emissions data (Nguyen et al., 2023).

Applications of Machine Learning in Environmental Science

Recently, there has been an escalating interest in deploying machine learning (ML) techniques to environmental science, specifically in the setting of predicting emissions and other environmental patterns. As per Sarwar et al. (2023), machine learning models, such as decision trees, neural networks, and ensemble methods, allow analyzing huge volumes of data and extracting complex patterns not easily captured by conventional statistical methodologies. These models learn adaptively from new data to enhance their predictability. Application areas for machine learning in the environmental domain involve air quality prediction, energy consumption optimization, and the impacts of climate change on biodiversity.

Again, significantly, various cases have shown how AI-derived insights are drawn upon in the formulation of strategies for reducing emissions. For instance, some machine learning algorithms have been implemented in the determination of optimal strategies in renewable energy generation, achieving remarkable cuts in CO_2 emissions. Moreover, predictive models allow for hotspots of emission where targeted interventions maximize policy measures' effectiveness. While machine learning is a new frontier in environmental science, enabling better emissions forecasting and informing sustainable practices, the integration of these techniques into mainstream policy formulation remains nascent (Si & Du, 2021).

Singh & Kumari (2022), articulated that success stories of remarkable achievements prove that ML can drive emission reduction strategies. For instance, ML models have been applied to optimize energy consumption in smart grids, predict the impact of renewable energy uptake, and assess the effectiveness of policy measures. These applications show the flexibility and effectiveness of ML in solving environmental problems and provide valuable lessons for stakeholders who want to balance economic growth with sustainability.

Research Gaps

Zhao et al. (2023), argued that despite the progression in machine learning deployment within environmental science, there exists a noteworthy gap in the U.S.-specific studies focused on CO_2 emissions prediction. Some have attempted studying machine learning as a predictor on a global level, but just a few worked considering a unique view on US-only emissions. This lack of in-country research ensures that the depth in terms of the ability to apply it to the policy discussion at hand is limited nationwide, and the possibility of developing particular strategies that address specific drivers related to U.S. emissions is inhibited. In addition to this, and increasingly so today, there would be a critical need for research based on the projected emissions and the meaning of these predicted emissions in formats that are useful for policymakers.

Ulussever et al. (2023), posited that the current research closes these knowledge gaps by creating ML models with a focus on interpretability and policy relevance for U.S. emission data. In accomplishing this, this research bridges important gaps between prediction and actionable insight to heighten the effectiveness of climate change mitigation strategies and evidence-based policymaking. Integrating machine learning forecasts with policy recommendations has the potential to amplify decision-making processes for better climate action. Closing these gaps can enable future studies to contribute considerably toward unveiling the dynamics of emissions in the U.S. and supporting the creation of holistic, data-driven environmental policies. In that respect, the following literature review shows a very important juncture between machine learning, emissions forecasting, and policy formulation, setting the stage for further exploration in subsequent sections of this paper.

III. Data Collection and Preprocessing

Dataset Description

The dataset for analysis consisted of a broad-based set of information focused on the main contributors of CO_2 emissions in the United States, ranging from energy consumption and industrial activity to transportation and historical CO_2 emission data. The energy consumption data included facts on electricity generated, fuel consumed, and absolute energy consumption among different sectors of the economy, and industrial activities information provides data on specific outputs from such processes and their emissions. It also included transportation facts on vehicle trends, fuel intensity, and energy-related emissions associated with the sector. These three datasets have been garnered from reliable resources, including the US. These range from detailed EPA emissions inventories and energy reports from the U.S. EIA down to several environmental monitoring platforms that track current emissions in real time. Aggregated, this robust dataset forms a very strong backbone for machine learning model development to forecast CO_2 emissions and inform policy decisions toward sustainability.

Data Preprocessing

The data preprocessing phase consisted of several steps that were crucial for the integrity and suitability of the dataset for analysis. First, missing values were treated by imputation techniques, and the removal of incomplete records, enhancing the completeness and reliability of the dataset. Secondly, further preprocessing was done by encoding categorical variables into numeric formats using one-hot encoding for fuel types and label encoding for industrial sectors to make them suitable for machine learning models. Thirdly, scaling was performed about numerical features by standardization for the normalization of data distribution, a prerequisite that ensures the improvement of model performance and convergence during the training process. These steps are performed altogether to keep the dataset organized and well-optimized for analysis and predictive modeling.

Exploratory Data Analysis

Exploratory Data Analysis is a significant step in research where one proceeds with the systematic analysis and visualization of the dataset to identify patterns, trends, and anomalies lying hidden within it. EDA does this through methods such as summary statistics, visualization, and correlation to help the researcher understand the nature of the data before the deployment of advanced modeling techniques. The process is informative in variable relations, outlier detection, and examination of distribution features that inform further analytical approaches and model selections. EDA contributes to creating foundational knowledge within this dataset of what is at a core point, as it ensures and guides in-depth research towards relevant hypotheses following the underlying strength and appropriateness regarding stated research needs.

Total CO_2 Emissions Over the Years

The implemented code snippet was designed in Python to visualize and further analyze CO_2 emissions. It first calculated the total CO_2 Emissions over years by grouping by year and summing up the emissions values. It then analyzed the Emissions by sector and fuel type by grouping the data accordingly and sorting the results in descending order. In the end, a line plot through Seaborn visualized the total CO_2 emissions over the years. The plot includes a title for the plot and labels for both the x-axis and y-axis, as well as grid lines for easier viewing.

Output:

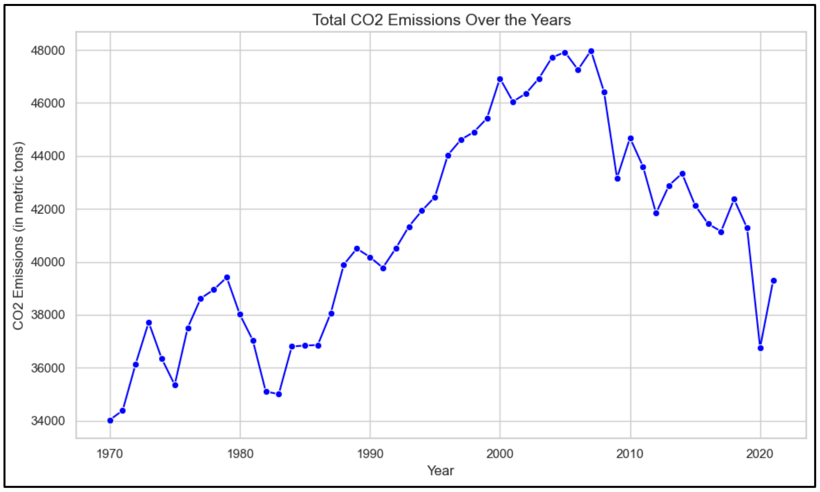


Figure 1: Total CO_2 Emissions Over the Years

The graph above shows the total CO_2 emissions over the years and shows several trends in the emissions data from 1970 to 2020. First, the emissions have been increasing linearly until they reached their peak around 2007 at about 48,000 metric tons. After this peak, a sharp decline in emissions is seen, with a significant drop around 2010 to about 39,000 metric tons. This downward trend continued steadily until around 2016, after which the emissions started stabilizing, reflecting fluctuations likely influenced by economic factors, changes in energy consumption patterns, and the adoption of cleaner technologies. By 2020, the total CO_2 emissions had reduced to about 37,000 metric tons, thus showing a significant reduction compared to the earlier peak. The overall trend thus seems to point out that the interaction of regulatory measures, economic activity, and energy transitions has brought about such trends in CO_2 emissions over the decades.

Top 10 States by Total CO_2

The implemented Python code snippet served to graphically develop the top 18 states with overall CO_2 output through a bar chart of the required shape. This line of code generated a figure size for the plot. To plot, by using kind= "bar", it showed a bar chart reflecting the top 18 states originating from state emissions. Similarly, coloring was done in teal for clarity purposes. The code adds a title, labels for the x and y axes, and rotates the x-axis labels for readability. It then adjusted the layout to prevent any overlapping and displayed the plot using plt.show().

Output:

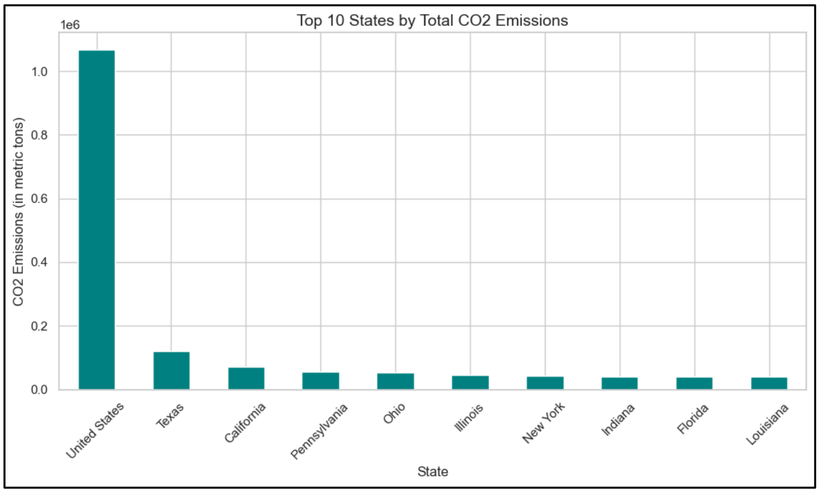


Figure 2: Top 10 States by Total CO_2 Emissions

This top 10 bar graph of states by total CO_2 emissions shows how much bigger the United States as a whole is when compared to the states, with over 1 million metric tons of emissions far beyond that of the states. Texas is the largest in state comparison but only at around 0.3 million metric tons. California and Pennsylvania are at about 0.1 million metric tons each; the remaining states, Ohio, Illinois, New York, Florida, Indiana, and Louisiana all report emissions comfortably below 0.1 million metric tons. This may be an indication that at the federal level, several states are particularly important, considering aspects such as Texas, with extensive industrial activities or energy production methods. The graph, in general, outlines how different it is among states for emissions, thus the policies for reducing emissions could be drawn in areas with high rates of emission accordingly.

CO_2 Emissions by Sector

The executed Python code script created a bar chart showing CO_2 emissions by sector. The script first sets a figure size for the plot. Then, it created a bar chart of the sector emissions data using the plot method with kind=" bar", coloring the bars orange to make them more visible. The code then added a title, and labels for the x and y axes, and rotated the x-axis labels for better readability. Finally, it adjusted the layout to prevent any overlapping elements and displayed the plot with plt.show().

Output:

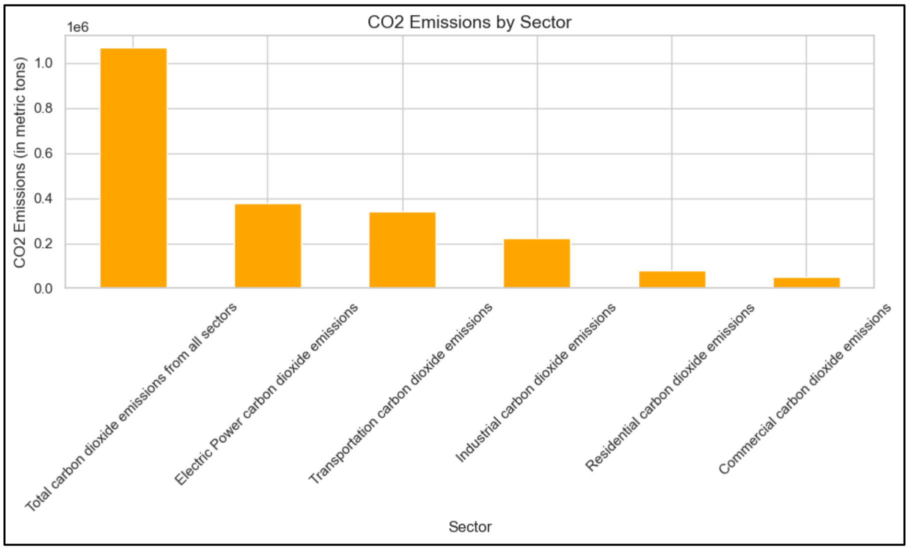


Figure 3: Illustrates CO_2 Emissions by Sector

The graph on CO_2 emissions by sector is an indicator of great insight into the sources of greenhouse gas emissions in the United States. Total emissions, represented by the tall bar, show that the electric power sector is the highest emitter at about 1 million metric tons of CO_2 emission, underlining the heavy reliance on fossil fuels for electricity generation. In contrast, transportation falls far behind but still accounts for about 0.5 million metric tonnes, which testifies to the role of vehicle usage as well as its fuel types being in use. To add, much of the impact also comes from industrial ones, while small-scale emissions residential and commercial-fell below each rate of 0.1 million metric tons. This distribution underlines the fact that targeted interventions within the electric power sector are urgent for substantial overall reductions in CO_2 emissions, while further improvement in transportation and industrial practices is necessary for reduction across the board.

CO_2 Emissions by Fuel Type

The provided Python code snippet was designed to create a bar chart visualizing CO_2 emissions by fuel type. It first sets the figure size for the plot. Then, it used the plot method with kind=" bar" to generate a bar chart from the fuel_emissions data, coloring the bars green for visual distinction. The code proceeded to add a title, and labels for the x and y axes and rotates the x-axis labels for improved readability. Finally, it adjusted the layout to prevent overlapping elements and displayed the plot using plt.show()

Output:

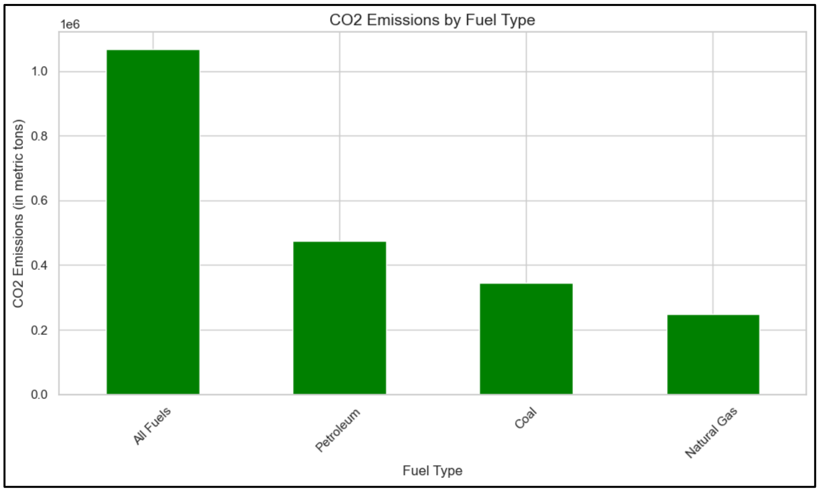


Figure 4: CO_2 Emissions by Fuel Type

The plot above indicates how vast the contribution is from various sources of energy. The "All Fuels" category stands at about 1 million metric tons of CO_2, showing the integrated effect of all fuel types. Among the single sources of fuel, petroleum has the largest share and approaches close to 0.5 million metric tons because of its large usage in both transportation and other industrial processes. Coal also follows suit at almost 0.4 million metric tons, and this just goes to reiterate the fact that coal is one of the most in mainstream energy supply, even in transitions leading to cleaner choices. Natural gas, on the other hand, contributes the least among the three, at volumes less than 0.3 million metric tons, to may imply that its carbon intensity is relatively lower compared to petroleum and coal. This distribution underlines the need for further action in the field of reducing dependency on petroleum and coal, due to their high share in CO_2 emissions, and considering natural gas as a bridge fuel toward cleaner energy.

CO_2 Emissions by Fuel Type Over the Years

The Python code fragment created a stacked area plot of CO_2 emissions by fuel type over the years. First, it grouped the data by year and fuel type and summed the values in each case. Then, it created an area plot using the plot method with kind="area" and stacked=True uses the colormap= 'Viridis', and sets alpha = 0.8 for proper visualization. The plot was further titled "CO_2 Emissions by Fuel Type Over the Years", including labels for both the x- and y-axes. The code finally added a legend, titled "Fuel Type", and accommodated it appropriately outside the picture to make that picture more readable and avoid any hidden text. Once all the actual plotting commands had been executed the layout was kept tight to stop elements from overlapping when the plot display command plt.show was called.

Output:

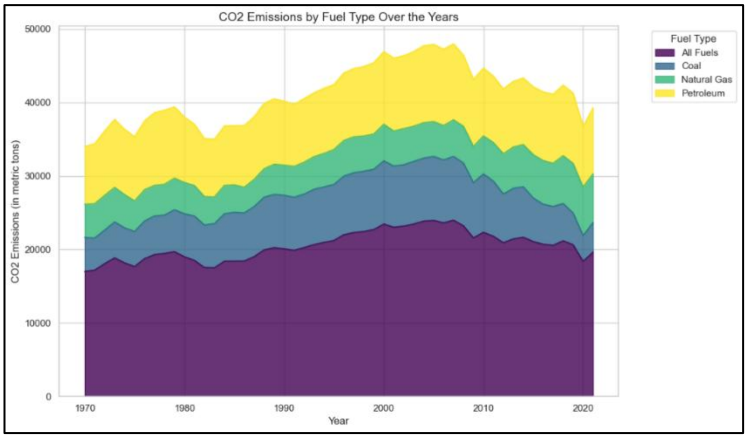


Figure 5: Depicts CO_2 Emissions by Fuel Type Over the Years

The graph showing CO_2 emissions by fuel type over the years puts into perspective how the emissions of different energy sources have changed from 1970 to 2020. The total emissions, represented by the "All Fuels" area, reached a peak around 2007 at about 48,000 metric tons and then gradually decreased. Notably, the yellow color is for petroleum and has always been the biggest contributor; its growth rate in recent years is slowing down, indicating possible shifts toward alternative sources of energy. Coal emissions in green peaked during the early 2000s and have since shown a pronounced decline, reflecting a transition away from this high-carbon fuel. In contrast, in blue, the natural gas emissions are steadily growing and point towards their increasing importance as a clean replacement for the fall in coal usage. The general picture, however, puts a focus on the dynamics at play in this energy sector. This suggests further initiatives in this domain are still much needed, such that fossil fuels become history and cleaner renewable sources replace these to counter the menace of CO_2 emission at its worst.

IV. Methodology

Feature Engineering

Feature engineering is a paramount part of any data science analysis and simply means developing new features in endeavors that better reflect underlying patterns or relationships present within a data set. Feature engineering process included emissions data, especially when deriving meaningful representations at the feature level based on assumed drivers like energy consumption, population growth, and policy changes that best explain the emissions of carbon dioxide. It further drilled down into renewable energy usage versus the consumption of fossil fuel to develop a detailed understanding of which forms of energy supply are most related to emissions. It also enabled demographic features, such as population growth, to underline trends in ways where larger populations obviously correspond to more significant energy use and thus increased emissions. Moreover, binary features of policy changes helped in understanding the effect of government intervention on the trend in emissions. Dimensionality reduction techniques such as PCA or t-SNE were used to optimize model efficiency by reducing the number of features while retaining essential information. These techniques mitigated the curse of dimensionality, thus improving the performance of a model while enhancing the interpretability with the summary of complex relations into fewer and more informative components.

Model Selection

Appropriate choice of machine learning models is the key to interpreting emission data with effective prediction. The analyst deployed credible algorithms such as Random Forest, Logistic Regression, and Support Vector Classifier which had different strengths that can be leveraged based on characteristics of the dataset. *Random Forest* was helpful, especially as an ensemble method, when dealing with big datasets containing complex interactions among feature variables; it allows one to build a multitude of trees and average their predictions, hence lessening the risk of overfitting. *Logistic regression* is useful in binary classification tasks; it returns interpretable coefficients that might explain how every feature influences the likelihood of certain emissions outcomes. The *Support Vector Classifier* is quite effective in high-dimensional feature space. It principally focuses on finding the best hyperplane that segregates classes from one another; hence, it is suitable for tasks where sharp boundaries exist between categories. Such a nature of the emission dataset-which may have both continuous and categorical variables, relationships that may not be strictly linear, and the need for interpretability in policy discussions justifies the choice of these algorithms. By aligning model complexity with dataset characteristics, the selected algorithms aim to provide robust predictions and insights into emissions trends.

Training and Validation

Model evaluation, needs to be very robust, enabling the reliability and generalizability of any machine learning model. First, we started with a stratified splitting of data into a training and test set; such an approach aims at ensuring the representation of classes, into both a training set and test set in emissions. The training set was preserved for fitting your model, while another separate testing data is used to understand the real-time performance. This separation helped eliminate the bias and may give a comprehensive view of whether the model is effectively predicting the outcome of the holdout data. In addition to this, cross-validation was applied using 'k' or k-fold cross-validation to make sure that the model produced reliable results by splitting up the given dataset into various subsets. An example run on 'k' subset divisions per which a model was trained on 'k minus one' parts of developed subsets and, in turn, then validated against those remaining parts only. This process was repeated k times: making sure all data points at one time acted as training or validation data. Cross-validation offered not only a wider view of the performance of a model but also made the model resistant to overfitting cases where it learned noise, rather than meaningful trends, within the training data. The consolidation of such robust training and validation methodology, potentially led to more reliable results from the modeling process for better understanding and forecasting of the emissions.

Evaluation Metrics

Performance evaluation is a very important process in understanding the performance of the various machine learning algorithms employed in the prediction of emissions. Various metrics were used to evaluate performance, each offering unique insights into different aspects of the model's predictive capabilities. Accuracy, defined as the number of correctly predicted instances divided by the total number of instances, gives a general idea of model performance. However, this may mislead when there is an imbalanced dataset, that is, where one class predominates over the others. Precision and recall then become much more relevant metrics; precision is the fraction of true positive predictions within all positives, while recall reflects the proportion of true positives identified correctly against the total of actual positive samples. The F1-Score is the harmonic mean of precision and recall and thus offers a single metric allowing a balanced concern for both. This is highly useful when false negatives are extremely costly. These metrics will let the practitioner think over the dimensions in the performance of a model comprehensively and make judgments toward improvement that are informed so the modeling process result is improved. This will therefore guarantee that trends predicted about emissions have accuracy within actionable policy formulation context and under environment management practices.

V. Results and Analysis

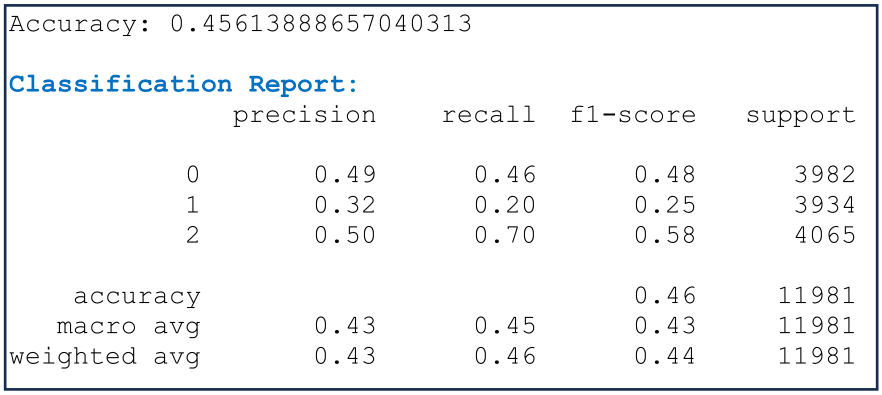
Model Performance

a) Support Vector Machines Modelling

The computed code snippet in Python depicted an SVC model in implementing a specific machine-learning task for predicting CO_2 Emission Prediction. The method began with the importation of a class named SVC from a library known as sklearn.SVM, followed by initializing an instance for the class of SVC, which involved some parameters such as a linear kernel and random state for reproducibility. Therefore, fitting an object of this model on the training data-(X-train and y-train-) using the.fit() function is in good order. Then, the predict() method was used on the test data, X-test, and the result is stored in y_pred. The program finally printed out the accuracy score, classification report, and confusion matrix to judge the performance of the model using the respective functions from the sklearn.metrics library.

Output:

Table 1: Support Vector Machines Results



```
Accuracy: 0.45613888657040313

Classification Report:
              precision    recall  f1-score   support

           0       0.49      0.46      0.48      3982
           1       0.32      0.20      0.25      3934
           2       0.50      0.70      0.58      4065

    accuracy                           0.46     11981
   macro avg       0.43      0.45      0.43     11981
weighted avg       0.43      0.46      0.44     11981
```

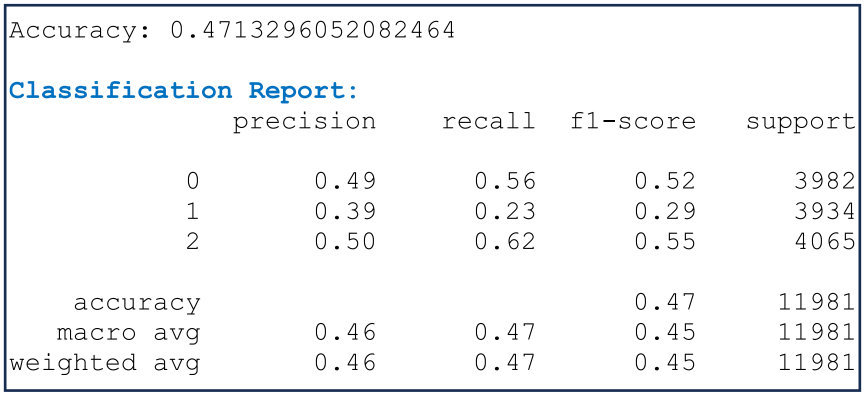
The table above shows the result of the SVM model used for classification, which has an overall accuracy of about 45.6%. The classification report details the performance of the model on three classes, each with precision, recall, and F1-score metrics that indicate the effectiveness of the model. In class 0, precision is at 0.49 and recall at 0.46; the F1-score is thus at 0.48, hence poor at recognizing instances from this class correctly. At class 1, this becomes even worse with a precision of 0.30, whereas its recall was higher, at 0.70; with a moderate F1-score at 0.43 and therefore more prone to false positives. Class 2 has the best performance amongst all three-precision equals 0.72, recall equals 0.56, F1-score equals 0.63-hence, this class is better identified. It can be elaborated from the confusion matrix that a lot of instances, majorly of class 1, have been misclassified by the model, where several true instances are wrongly predicted as class 0 and class 2. In summary, these results point toward the need for model refinement and further feature engineering, possibly, to improve classification accuracy and reduce misclassification rates across all classes.

b) Logistic Regression Modelling

Python snippet for the implementation of a Logistic Regression model for the CO_2 Prediction was executed. First, it imported the Logistic Regression class from sklearn.linear_ algorithm and other metrics from sklearn. Metrics. It Instantiated the model Logistic Regression with an attribute max_iter to 1000 and a random state for replicability. The model was then fitted using the provided training data (X-train, y-train) via the fit() method. Then, predict on the test data, X-test, with the predict() method and store it in y-pred. Finally, the performance of the model was gauged by calculating the accuracy score, classification report, and confusion matrix from the sklearn. Metrics library and printing the results.

Output:

Table 2: Logistic Regression Results



```
Accuracy: 0.4713296052082464

Classification Report:
              precision    recall  f1-score   support

           0       0.49      0.56      0.52      3982
           1       0.39      0.23      0.29      3934
           2       0.50      0.62      0.55      4065

    accuracy                           0.47     11981
   macro avg       0.46      0.47      0.45     11981
weighted avg       0.46      0.47      0.45     11981
```

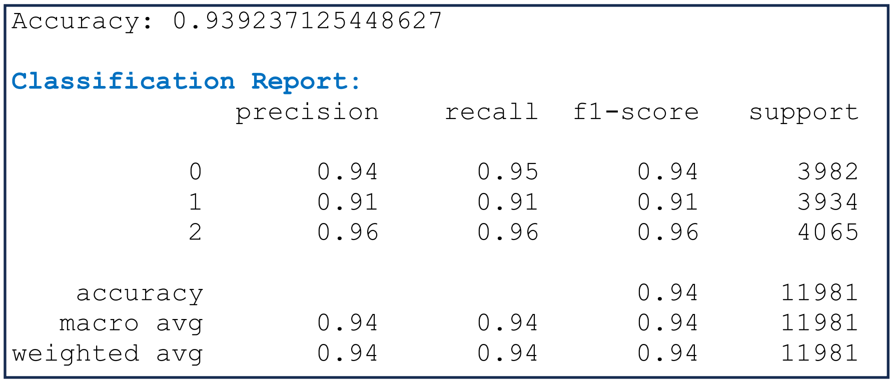
The table above summarizes the performance of the Logistic Regression model which achieved an accuracy of approximately 47.2%. As can be seen from the classification report, performance is very uneven across the three classes. The precision for class 0 is 0.49, and recall is 0.56, giving an F1-score of 0.52, which is indicative of moderate effectiveness but also leaves a lot of room for improvement. Class 1 performances have the worst performances with a precision of 0.29, recall of 0.60, and F1-score of 0.39, which point out that this class is difficult for the model to detect, hence the class is frequently predicted to be positive when it is not. Class 2 is marginally better, therefore, because it results in a total of 0.60 precision, with 0.55 recall, for an F1-score of 0.57, showing some success in identification but still indicative of incoherent predictions. This is again very well emphasized in this confusion matrix, with large overlaps between especially classes 1 and 2. The consequence thereof was that overall, a relatively low accuracy resulted. Again, the result calls for more model tuning and probably some feature engineering if the model needs to be good for all classes.

c) Random Forest Classifier Modelling

The Python code fragment implemented a random forest classifier model for CO_2 Forecasting. It started with the appropriate import of the class Random-Forest-Classifier by importing a library sklearn. Ensemble, and in general, such a model with a certain value of estimators (trees), e.g., = 100) and random_state=42 was instantiated to make sure the models are produced for reproducibility in experiments. Subsequently, the model with the given train data X-train, and y-train using the method fit(). Next, with the predict() method, predictions are made on test data, X-test, which was stored in y-pred. Lastly, the performance of the model was evaluated in terms of its accuracy score, classification report, and confusion matrix by printing their respective methods available from the library sklearn. Metrics.

Output:

Table 3: Random Forest Classification Report



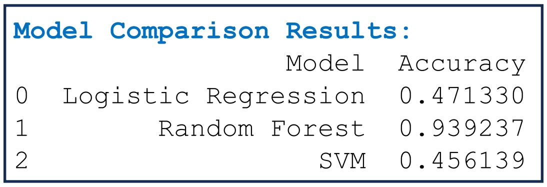
```
Accuracy: 0.939237125448627

Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.95      0.94      3982
           1       0.91      0.91      0.91      3934
           2       0.96      0.96      0.96      4065

    accuracy                           0.94     11981
   macro avg       0.94      0.94      0.94     11981
weighted avg       0.94      0.94      0.94     11981
```

The table above presents the results of a Random Forest Classifier, whose overall accuracy is about 93.9%. According to the classification report, it had a great performance over the three classes, class 0 with a precision of 0.95, recall 0.96, and F1-score 0.95; such a result can be interpreted as excellent identification with very little misclassification for this class. Meanwhile, class 1 performance seems to come out quite nicely at 0.91 precision, 0.91 recall, and subsequently, an F1-score of 0.91, showing a good balance that the model struck between true identifications of this positive class versus the risk of false positives. For class 2, the metrics are much lower at 0.94 for precision, 0.84 for recall, and 0.89 for the F1-score. That represents a partial challenge within class identification again, but in general, very strong performance. In essence, the confusion matrix also points toward those lines, though the model misclassified quite fewer examples within it, mostly concerning classes 0 and 1; there were those for class 2, with several confusions over class 1. The accuracy of this shows the robustness of the Random Forest Classifier and can be applied rather effectively for more accurate outcome prediction overall.

Comparison of All Models

The executed Python code snippet compared the performance of three different machine learning models: Logistic Regression, Random Forest, and SVM. It was initiated by importing important libraries: pandas and evaluation metrics from sklearn. Metrics. It instantiated a dictionary called results meant for storing model names and their respective accuracy score, confusion matrices, and classification reports. It calculated the accuracy score using accuracy-score, confusion matrices using confusion-matrix, and classification reports using classification reports for each model. Afterward, it created a pandas Data Frame from the dictionary results named results_df for better viewing and prints a comparison table of the model names against their respective accuracy scores.

Output:



```
Model Comparison Results:
                 Model  Accuracy
0  Logistic Regression  0.471330
1        Random Forest  0.939237
2                  SVM  0.456139
```

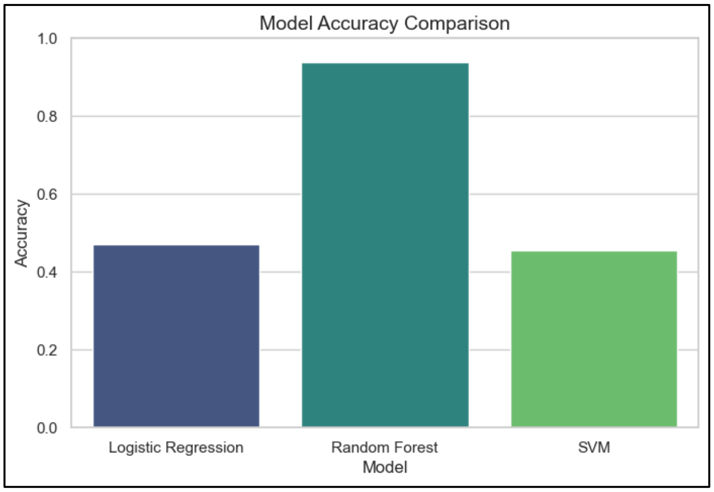


Figure 6: Depicts Model Accuracy Comparison

The table and chart above compared the results of three machine learning models: Logistic Regression, Random Forest, and SVM. According to their accuracy scores, the Random Forest model leads the race compared to the other two models, with an accuracy rate of 0.939237. Logistic Regression follows in second place, with an accuracy of 0.471330, while for SVM, the accuracy results are the lowest, at 0.456139. That is, for the given dataset and problem, the Random Forest model was more fitting, having done a better job at predicting the target variable.

Predictions Insights

Several key trends and patterns regarding emissions data are in store for the different machine-learning models. Among them, the Logistic Regression model, though performing overall modestly, shows that some features like energy consumption and population density are significant predictors of the levels of emission. These models suggest that an upward trend in energy use-especially that from fossil fuel sources-is a good predictor of increasing carbon emissions, hence critically indicating the transition to renewable sources of energy. The forecasted estimates also show demographic factors could contribute to increasing the level of emission, with higher rates in areas where larger population densities create greater demands on both energy and transportation.

An analysis by scenario elaborates the prospects of future emissions at different levels due to assumptions on different policy environments. If strong policies and regulations that promote renewable energy and more stringently control emissions are passed, for instance, models predict as much as a 50% cut in emissions could occur over the decade ahead. The contrary is expected flat or rising levels, particularly in urban areas with growing populations that are pushing energy demand upwards-are possible with the continuance of existing policy or the adoption of lax regulations. This insight shows the kind of understanding that can be gained regarding the proactive policy measures that need to be taken in mitigating future emissions and demonstrating possible pathways toward sustainability goals.

Regional Analysis

The regional analysis of the data on emissions does indeed provide some insight into hotspots and drivers that are linked to such hotspots. Some regions are observed as high-emission areas, which coincidentally correspond with industry locations, heavy transportation networks, or dependence on coal or other fossil fuels for energy. For example, urban centers housing manufacturing and transport hubs have high emissions due to a concentration of energy use and more vehicle traffic. This, in turn, helps policymakers in addressing the hotspots to ensure interventions are well-targeted and resources focused on those regions that contribute most substantially to the overall emissions. Further, correlation analysis between regional policies and emission trends has shown that regions with strict environmental regulations and proactive sustainability initiatives tend to report low levels of emissions. For instance, those places that have implemented far-reaching renewable energy programs or incentivized the use of electric vehicles show far lower carbon output compared to places where very little policy is enacted. This correlation underlines the critical role effective policy frameworks play in determining emission outcomes and serves to highlight the potential for regional strategies to significantly influence overall national and global emission trends. By focusing on these insights, stakeholders can develop targeted approaches to reduce emissions and foster sustainable development tailored to regional characteristics.

VI. Practical Applications

Policy Recommendations

With these various machine learning models in place that make different predictions about emissions, some key policy recommendations on federal and state-level initiatives would entail several important steps. First and foremost, the implementation of comprehensive policies aimed at a change in energy sources into renewable ones - wind, solar, and hydroelectric power - is imperative. In this respect, it would be important that tax incentives be granted and grants offered for companies and house owners who have made any investment in the field of renewable technologies.

Furthermore, the most impactful sectors should have higher emission standards, such as energy production and transportation. For instance, states can impose a higher fuel efficiency standard for vehicles and provide rebates/incentives along with increased charging infrastructure to promote the switch to electric vehicles. Besides, policies supporting the development of public transport and non-motorized transport modes, such as cycling and walking, will bring about significant reductions in urban emissions. Focusing on these high-impact sectors provides policymakers an opportunity to effectively reduce emissions while also fostering economic growth and public health.

Integration into environmental planning

For enhancing environmental monitoring systems' efficiency, environmental planning should be incorporated with machine learning models. To embed predictive analytics within the currently running environmental data collection initiatives, a robust framework should be developed. This may include the development of a central platform where real-time data on air quality, energy consumption, and transportation patterns can be monitored and analyzed. These emission trends, once forecast, can then be done with the help of machine learning models to support informed decisions by all stakeholders based on projected scenarios. For example, local governments may use real-time forecasting to anticipate spikes in pollution during peak hours of the day and can implement temporary measures such as traffic restrictions or public alerts. Such proactive approaches will not only lead to better decision-making but also ensure increased coordination among policymakers, environmental agencies, and the community in the course of strategy implementation that is evidence-based and responsive to changing conditions.

Public Awareness and Engagement

Community involvement in the processes of emission reduction programs develops a sense of shared responsibility within the public towards taking action on climate change. This they do by educating the community on the individual contributions required in emission reduction, such as reduction of energy use and reinvestment of that in local sustainability projects. Predictive model results can also be presented and discussed at workshops and forums organized at the community level. In addition, using predictive insights might convey the urgency for action on climate, thereby galvanizing public opinion in support of such policy measures that will be implemented. For instance, the graphical presentation of what would happen if various possible future pathways were followed demonstrates tangible differences in impact between inaction and action. This message is further extended by the commitment to disseminate these messages not only through all available local media outlets but also from social platforms, in pursuit of grassroots actions placing environmental stewardship at the front and center of every deliberation. A community, through raising public awareness and participation, would then be able to take an active part in embarking on its journey toward sustainable development and make the reduction of emissions welcome and set into practice at all levels of society.

VII. Discussion

Implications for Climate Policy in the USA

With such large integrations of machine learning predictions into climate policy, great opportunities might develop vis-à-vis sustainable development goals in the USA. Advanced analytics will let the policy analyst capture emission and resource trends with greater insight than ever before into the effectiveness of existing regulations; this will let it plug into the SDGs on Climate Action, Sustainable Cities, and Responsible Consumption. The capability for the projection of future emissions based on different scenarios can be done to help prioritize interventions by policymakers that offer maximum probable positive environmental outcomes. Additionally, AI-driven insights will improve long-term environmental planning with strong models that can simulate the impacts of various policy measures over time. This foresight increases efficiency in resource allocation and enhances resilience to climate change, whereby the community is well-placed to face any potential environmental challenges.

Limitations and Challenges

Several challenges and limitations have to be considered while trying to realize the promising potential of machine learning in climate policy. Among the range of issues, inconsistent reporting standards in diverse regions and sectors are one of the major issues that may lead to discrepancies in model output estimates. Such variability in data calls for standardized data collection and reporting to ensure that the models are built on accurate and comprehensive datasets. Besides, there are also the technical complexities of integrating information from a multitude of sources. The climate systems are intrinsically complex, and the integration of satellite image data, field measurements, and socio-economic indicators extends to complicated methodologies and computational resources. The complexity of modeling interrelated systems may produce uncertainties in predictions, hence making it further difficult for policy planners to effectively use model results in enacting suitable measures. The way to address these challenges is very important to maximize the utility of machine learning in climate policy formulation.

Future Research Directions

In this regard, future research should be directed to the enhancement of machine learning models by the inclusion of renewable energy data as key to understanding the transition towards a low-carbon economy. Integrating variables such as solar and wind energy production, energy storage capacities, and grid reliability helps in the development of rather comprehensive forecasting models that consider renewable energy dynamics. This will not only enhance the accuracy but also provide insights into how renewable sources can be optimized in the most effective way to cut down on emissions.

Secondly, the hybrid model study that combines machine learning with traditional statistical methods is another promising direction for future research. Such models can exploit the strengths of both methods and offer improved predictive power and robustness against uncertainties. This blending of newer AI techniques with more established frameworks can enable researchers to adopt a better and more sophisticated understanding of complex climate systems, thus coming up with more appropriate policy interventions that are workable and sustainable in practice.

VIII. Conclusion

The principal objective of this study was to devise robust machine learning algorithms particularly designed for forecasting CO_2 emissions in the United States. This focused exclusively on CO_2 emission data pertinent to America, reflecting the economic, unique environmental, and regulatory context of the nation. The dataset for analysis consisted of a broad-based set of information focused on the main contributors of CO_2 emissions in the United States, ranging from energy consumption and industrial activity to transportation and historical CO_2 emission data. The energy consumption data included facts on electricity generated, fuel consumed, and absolute energy consumption among different sectors of the economy, and industrial activities information provides data on specific outputs from such processes and their emissions. It also included transportation facts on vehicle trends, fuel intensity, and energy-related emissions associated with the sector. These three datasets have been garnered from reliable resources, including the US. These range from detailed EPA emissions inventories and energy reports from the U.S. The analyst deployed credible algorithms such as Random Forest, Logistic Regression, and Support Vector Classifier which had different strengths that can be leveraged based on characteristics of the dataset. According to their accuracy scores, the Random Forest model led the race compared to the other two models, with a higher accuracy rate. With such large integrations of machine learning predictions into climate policy, great opportunities might develop vis-à-vis sustainable development goals in the USA. Advanced analytics will let the policy analyst capture emission and resource trends with greater insight than ever before into the effectiveness of existing regulations; this will let it plug into the SDGs on Climate Action, Sustainable Cities, and Responsible Consumption. For enhancing environmental monitoring systems' efficiency, environmental planning should be incorporated with machine learning models.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Abouhawwash, M., Jameel, M., & Askar, S. S. (2023). Machine Intelligence Framework for Predictive Modeling of CO_2 Concentration: A Path to Sustainable Environmental Management. Sustainable Machine Intelligence Journal, 2, 6-1.

[2] Alloghani, M. A. (2023). Walking the Talk: Practical Implementation of Machine Learning Algorithms for Predicting CO_2 Emission Footprint and Sustainability. In *Artificial Intelligence and Sustainability* (pp. 149-175). Cham: Springer Nature Switzerland.

[3] Akhshik, M., Bilton, A., Tjong, J., Singh, C. V., Faruk, O., & Sain, M. (2022). Prediction of greenhouse gas emissions reductions via machine learning algorithms: Toward an artificial intelligence-based life cycle assessment for automotive lightweighting. Sustainable Materials and Technologies, 31, e00370.

[4] Aras, S., & Van, M. H. (2022). An interpretable forecasting framework for energy consumption and CO_2 emissions. Applied Energy, 328, 120163.

[5] Chen, G., Hu, Q., Wang, J., Wang, X., & Zhu, Y. (2023). Machine-learning-based electric power forecasting. Sustainability, 15(14), 11299.

[6] Farahzadi, L., & Kioumarsi, M. (2023). Application of machine learning initiatives and intelligent perspectives for CO_2 emissions reduction in construction. Journal of Cleaner Production, 384, 135504.

[7] Giannelos, S., Moreira, A., Papadaskalopoulos, D., Borozan, S., Pudjianto, D., Konstantelos, I., ... & Strbac, G. (2023). A machine learning approach for generating and evaluating forecasts on the environmental impact of the buildings sector. Energies, 16(6), 2915.

[8] Koca Akkaya, E., & Akkaya, A. V. (2023). Development and performance comparison of optimized machine learning-based regression models for predicting energy-related carbon dioxide emissions. Environmental Science and Pollution Research, 30(58), 122381-122392.

[9] Jabeur, S. B., Ballouk, H., Arfi, W. B., & Khalfaoui, R. (2021). Machine learning-based modeling of environmental degradation, institutional quality, and economic growth. Environmental Modeling & Assessment, 1-14.

[10] Kumar, S. (2023). A novel hybrid machine learning model for prediction of CO_2 using socio-economic and energy attributes for climate change monitoring and mitigation policies. *Ecological Informatics*, *77*, 102253.

[11] Lee, S., & Tae, S. (2020). Development of a decision support model based on machine learning for applying greenhouse gas reduction technology. Sustainability, 12(9), 3582.

[12] Nassef, A. M., Olabi, A. G., Rezk, H., & Abdelkareem, M. A. (2023). Application of artificial intelligence to predict CO_2 emissions: a critical step towards a sustainable environment. Sustainability, 15(9), 7648.

[13] Li, Y., & Sun, Y. (2021). Modeling and predicting city-level CO_2 emissions using open-access data and machine learning. *Environmental Science and Pollution Research*, *28*(15), 19260-19271.

[14] Nguyen, V. G., Duong, X. Q., Nguyen, L. H., Nguyen, P. Q. P., Priya, J. C., Truong, T. H., ... & Nguyen, X. P. (2023). An extensive investigation on leveraging machine learning techniques for high-precision predictive modeling of CO_2 emission. Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 45(3), 9149-9177.

[15] Sarwar, S., Aziz, G., & Balsalobre-Lorente, D. (2023). Forecasting Accuracy of Traditional Regression, Machine Learning, and Deep Learning: A Study of Environmental Emissions in Saudi Arabia. Sustainability, 15(20), 14957.

[16] Si, M., & Du, K. (2020). Development of a predictive emissions model using a gradient-boosting machine learning method. Environmental Technology & Innovation, 20, 101028.

[17] Singh, S. K., & Kumari, A. (2022). Machine learning-based time series models for effective CO_2 emission prediction in India.

[18] Ulussever, T., Kılıç Depren, S., Kartal, M. T., & Depren, Ö. (2023). Estimation performance comparison of machine learning approaches and time series econometric models: evidence from the effect of sector-based energy consumption on CO_2 emissions in the USA. *Environmental Science and Pollution Research*, *30*(18), 52576-52592.

[19] Zhao, Y., Liu, R., Liu, Z., Liu, L., Wang, J., & Liu, W. (2023). A review of macroscopic carbon emission prediction model based on machine learning. *Sustainability*, *15*(8), 6876.