MACHINE LEARNING MODELS FOR PREDICTION OF CO2 EMISSION WITH EXPLORATORY DATA ANALYSIS

Dheeravath Janu, Mamatha Arupula, Dr Nazimunnisa
Assistance Professor, Assistance Professor, Associate Professor
Dept of CSE
Sree Dattha Institute of Engineering and Science

ABSTRACT

CO2 emissions play a major role in global warming, leading to serious consequences such as extreme weather events, rising sea levels, and ecological imbalances. To address this pressing issue, it is crucial that we fully understand the factors influencing CO2 emissions in order to develop effective strategies for reduction and sustainability. The growing concern over climate change and its harmful effects on our environment has motivated researchers and policymakers to seek innovative solutions for curbing greenhouse gas emissions, especially CO2 emissions. However, traditional statistical methods have their limitations when it comes to handling large and complex datasets. This is where machine learning steps in as a powerful tool, offering the ability to analyze vast amounts of data and make accurate predictions. This presents a promising avenue for forecasting CO2 emissions and creating sustainable policies. Machine learning allows us to identify hidden patterns and relationships within the data, enabling us to make more precise predictions and reliable forecasts. Therefore, this work focuses on exploring various machine learning models for predicting and forecasting CO2 emissions. Additionally, we plan to incorporate exploratory data analysis (EDA) techniques, which will help us visualize and interpret the data effectively. Through EDA, we can identify crucial features, understand data distributions, and pinpoint outliers that might influence model performance. The significance of our study lies in the valuable insights it can provide to policymakers and environmentalists. By making accurate predictions about CO2 emissions, we can help design effective policies that control and reduce emissions, optimize resource allocation, and promote the shift towards renewable energy sources. Furthermore, precise forecasts can assist in planning adaptation measures to mitigate the impact of climate change.

INTRODUCTION

1.1 Overview

Predicting and forecasting CO2 emissions is of paramount importance in addressing the global climate crisis. This task involves assessing the likely future levels of carbon dioxide (CO2) emissions into the Earth's atmosphere, primarily driven by human activities such as burning fossil fuels, deforestation, and industrial processes. To achieve accurate forecasts, a multi-faceted approach is essential.

Firstly, historical data analysis is crucial. Researchers and climate scientists analyze past emission trends to understand patterns and drivers, including economic growth, energy consumption, and policy changes. This historical context serves as a baseline for forecasting.

Next, various models and methodologies are employed to make predictions. One common approach is using integrated assessment models (IAMs) that combine economic, energy, and environmental data to simulate different scenarios. These models account for factors such as population growth, technological advancements, energy transitions, and policy interventions. They allow for the exploration of "business-as-usual" scenarios and the impact of climate mitigation policies.

Machine learning and artificial intelligence have also played an increasingly significant role in forecasting CO2 emissions. These techniques can analyze complex datasets, identify trends, and make predictions based on real-time information, improving the accuracy of forecasts.

Incorporating geopolitical factors and policy changes is another essential aspect. Government regulations, international agreements like the Paris Agreement, and evolving energy policies significantly influence emissions trajectories. Therefore, forecasting must consider political will and the potential for policy shifts.

Climate events and natural occurrences, such as volcanic eruptions and wildfires, can also have short-term and long-term effects on CO2 emissions. Therefore, including probabilistic elements in forecasting models is vital to account for unforeseen events.

Moreover, public awareness and behavioral changes are crucial factors. As society becomes more environmentally conscious, shifts in consumer preferences, demand for sustainable products, and lifestyle choices can impact emissions. Forecasters must monitor and assess these dynamics.

Finally, uncertainty quantification is an integral part of forecasting. Predicting CO2 emissions involves inherent uncertainty due to the complexity of natural and human systems. Therefore, forecasts typically present a range of scenarios to account for various possible outcomes.

So, predicting and forecasting CO2 emissions involves a comprehensive approach that considers historical data, complex modeling techniques, policy dynamics, natural events, behavioral changes, and uncertainties. Accurate predictions are essential for guiding climate action, influencing policy decisions, and mitigating the impacts of climate change on a global scale.

1.2 Research Motivation

The motivation for conducting research on predicting and forecasting CO2 emissions is rooted in the profound and pressing challenges posed by climate change, making it one of the most critical issues of our time. Several key motivations drive the need for comprehensive investigations into this field:

- **Global Climate Crisis**: Climate change represents an existential threat to the planet. Rising global temperatures, melting ice caps, extreme weather events, and disruptions to ecosystems are all direct consequences of excessive CO2 emissions. Understanding and forecasting these emissions is fundamental to addressing the root cause of the crisis.
- **Mitigation of Catastrophic Impacts**: Accurate predictions of CO2 emissions are essential for mitigating the most catastrophic impacts of climate change. By quantifying future emissions, researchers can identify the urgency of emission reductions required to avoid the worst-case scenarios, such as sea-level rise, species extinction, and food scarcity.
- **Policy Formulation and Evaluation**: Policymakers at all levels of government require data-driven insights into future emissions to design and assess climate policies effectively. Timely and accurate forecasts guide the development of regulations, incentives, and international agreements aimed at reducing emissions and transitioning to a sustainable, low-carbon economy.
- **Global Cooperation**: International cooperation is critical in the fight against climate change. Accurate CO2 emission forecasts provide a common basis for negotiations and commitments among nations, fostering collaboration on a global scale.
- **Resource Allocation**: In a world with finite resources, understanding emissions trends is essential for optimizing resource allocation. Governments and organizations must prioritize

investments in renewable energy, energy efficiency, and sustainable infrastructure to reduce emissions efficiently.

- **Economic Stability**: Climate change poses significant risks to economic stability. By predicting emissions and their consequences, researchers can help governments and businesses make informed decisions to safeguard economic interests and prevent financial crises related to climate impacts.
- **Innovation and Technological Advancement**: Forecasts of CO2 emissions stimulate innovation by identifying opportunities for research and development in clean energy technologies, carbon capture and storage, and sustainable transportation solutions. This can drive economic growth and job creation while reducing emissions.
- **Human Health and Well-being**: Climate change exacerbates health risks through factors such as heatwaves, air pollution, and the spread of diseases. Accurate predictions enable public health officials to plan for and mitigate these risks, protecting the well-being of communities.
- **Sustainable Development**: Sustainable development is a global priority. Predicting CO2 emissions supports the United Nations Sustainable Development Goals (SDGs) by providing insights into how emissions impact different aspects of development, from poverty reduction to clean water access.
- **Public Awareness and Engagement**: The public plays a crucial role in advocating for climate action. Transparent and accessible emissions data, combined with forecasts that highlight the consequences of inaction, raise public awareness and mobilize individuals and communities to demand and implement sustainable solutions.

1.3 Problem Statement

The problem statement for research on predicting and forecasting CO2 emissions revolves around the critical need to address the global climate crisis through informed, evidence-based decision-making. This encompasses several specific challenges and issues:

- **Climate Change as a Global Emergency**: Climate change represents an existential global emergency, with wide-ranging impacts on ecosystems, economies, and human societies. The problem lies in the fact that the world is currently on a trajectory of rising CO2 emissions, primarily driven by human activities such as burning fossil fuels, deforestation, and industrial processes. The increasing concentration of greenhouse gases, particularly CO2, in the atmosphere is the root cause of rising global temperatures and the associated consequences, including more frequent and severe extreme weather events, rising sea levels, and disruptions to food and water supplies.
- **Inadequate Emission Reductions**: Despite international agreements and growing awareness of the climate crisis, global efforts to reduce CO2 emissions have been inadequate to curb the worst impacts of climate change. One of the primary reasons for this inadequacy is the lack of precise, up-to-date, and region-specific information about future emissions trends.

Policymakers, businesses, and individuals need accurate forecasts to understand the urgency of emission reductions and to develop effective mitigation strategies.

- **Policy and Investment Challenges**: The absence of reliable forecasts hampers the development and implementation of climate policies and sustainable investment decisions. Governments struggle to set emission reduction targets and allocate resources effectively without a clear understanding of future emissions. Likewise, businesses face challenges in aligning their strategies with evolving market dynamics and regulatory changes.
- **Risk and Uncertainty**: Climate change poses significant risks to ecosystems, economies, and public health. These risks are compounded by uncertainty surrounding the magnitude and timing of future emissions. Uncertainty regarding emissions trajectories, coupled with the unpredictable nature of climate-related events, makes it challenging to assess and prepare for the full range of climate impacts.
- **Global Collaboration**: Effective global collaboration in the fight against climate change depends on accurate and transparent emissions data and forecasts. Without a shared understanding of future emissions, it becomes difficult to negotiate and enforce international agreements and commitments.

2. Literature Survey

This literature review section is organized as follows. First, the prediction of CO2 emissions is reviewed. Second, studies on the causality among industrial structure, energy consumption, and CO2 emissions are reviewed. Finally, the application of machine learning (ML) to predict CO2 emissions is reviewed. The literature review focuses special attention on research in China.

Sharp increases in carbon dioxide (CO2) emissions strengthen the greenhouse effect, leading to an ongoing increase in the global average temperature. The average annual global emissions of greenhouse gases from 2010 to 2019 were at the highest level in human history. Since then, the growth rate has slowed. Global greenhouse gas (GHG) emissions are expected to peak by 2025 to meet the goal of limiting global warming to 1.5 °C by the end of the century. Specifically, annual CO2 emissions are expected to fall by approximately 48% by 2030 and reach net zero by 2050 [1].

As a developing country, China faces the dual task of developing its economy and protecting the environment. In the past two decades, China’s economy has developed rapidly, and because economic development depends on energy consumption [2,3], China has become a large energy consumer and carbon emitter [4,5]. In 1990, China’s emissions were less than one-quarter of the total of the world’s developed countries. Since 2006, however, China has been the world’s largest carbon emitter [6,7].

China’s CO2 emissions mainly come from electricity generation [8,9], industry [10], construction [11,12], transportation [13,14], and agriculture [15]. Of these, electricity and industry are the two major high-emission sectors, accounting for more than 70% of the total emissions. Thermal power generation currently dominates China’s power structure. The main ways to reduce carbon in the power industry include reducing the proportion of coal power; accelerating the development of non-fossil energy, such as wind and photovoltaic power; and building a clean, low-carbon, safe, and efficient energy system. Second, achieving a low-carbon economy requires adjusting the industrial structure. This includes increasing the proportion of the service industry, which provides economic activity at low consumption and emission levels, and reducing the proportion of the manufacturing industry, which has high consumption and emission levels.

3.PROPOSED SYSTEM

3.1 Overview

The research work should start with a discussion of the findings, including insights gained from EDA, the effectiveness of data preprocessing techniques, the performance of the existing and proposed KNN models, and any recommendations for improving CO2 emission prediction and forecasting using machine learning. Additionally, the research work should discuss the limitations of the study and potential areas for future research. Figure 4.1 shows the proposed system model. The detailed operation illustrated as follows:

Step 1. Exploratory Data Analysis (EDA):

- Data Collection: Gather the dataset containing historical CO2 emission data along with relevant features such as population, GDP, energy consumption, etc.
- Data Inspection: Examine the dataset's structure, including the number of rows and columns, data types, and any missing values.
- Data Visualization: Create various plots and visualizations to gain insights into the data's distribution, trends, and relationships. This may include histograms, scatter plots, correlation matrices, and bar charts.
- Outlier Detection: Identify and handle outliers in the dataset, as extreme values can adversely affect machine learning models.

Step 2. Data Preprocessing:

- Feature Selection: Choose the most relevant features for CO2 emission prediction. This step involves selecting a subset of features that have the most impact on the target variable.
- Handling Missing Data: Address any missing values in the dataset through techniques like imputation or removal of rows/columns with missing data.
- Normalization/Scaling: Scale numerical features to ensure they have similar scales, which can improve the performance of some machine learning algorithms.
- Encoding Categorical Data: If applicable, convert categorical data into numerical format using techniques like one-hot encoding.
- Data Splitting: Divide the dataset into training and testing sets for model development and evaluation.

Step 3. Existing KNN Model:

- Select Existing KNN Model: Choose a standard K-Nearest Neighbors (KNN) regression model as a baseline.
- Hyperparameter Tuning: Use techniques like grid search or cross-validation to find the best hyperparameters (e.g., the number of neighbors) for the KNN model.
- Model Training: Fit the selected KNN model to the training data.

Step 4. Proposed KNN Model:

- Feature Engineering: Create new features or combinations of features that may improve the prediction of CO2 emissions.
- Hyperparameter Tuning: Similar to the existing KNN model, optimize the hyperparameters for the proposed KNN model.
- Model Training: Train the proposed KNN model using the training data.

Step 5. Prediction:

- Predict CO2 Emissions: Use both the existing and proposed KNN models to predict CO2 emissions for the testing dataset.

Step 6. Performance Estimation:

- Mean Absolute Error (MAE): Calculate the MAE to quantify the average absolute difference between predicted and actual CO2 emissions.
- Mean Squared Error (MSE): Compute the MSE to measure the average squared difference between predicted and actual emissions.
- Root Mean Squared Error (RMSE): Calculate the RMSE by taking the square root of MSE, providing a measure in the original unit (e.g., Mt).
- R-squared (R2) Score: Determine the R2 score to evaluate how well the model explains the variance in CO2 emissions.
- Comparison: Compare the performance metrics between the existing and proposed KNN models to assess whether the proposed model provides better predictions.

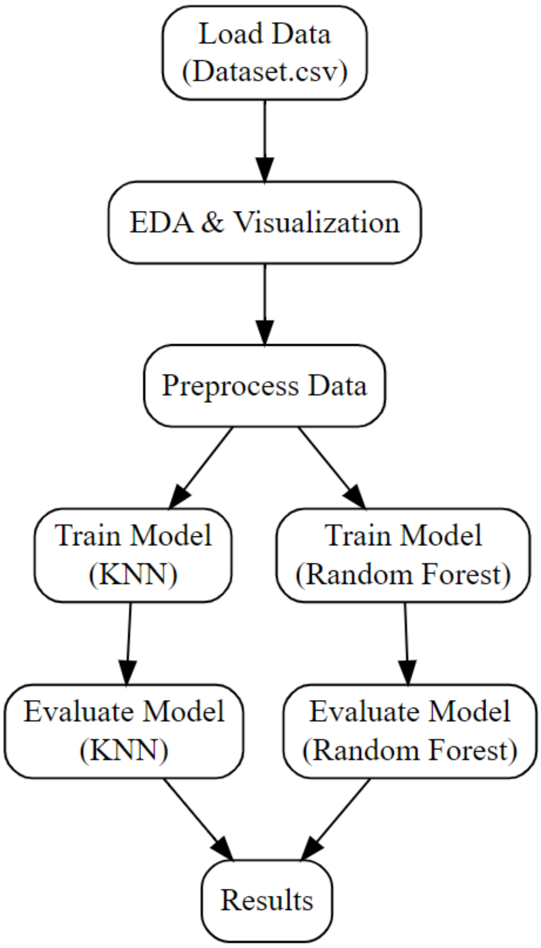


Fig. 3.1: Block diagram of proposed system.

3.2 RFC Model

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

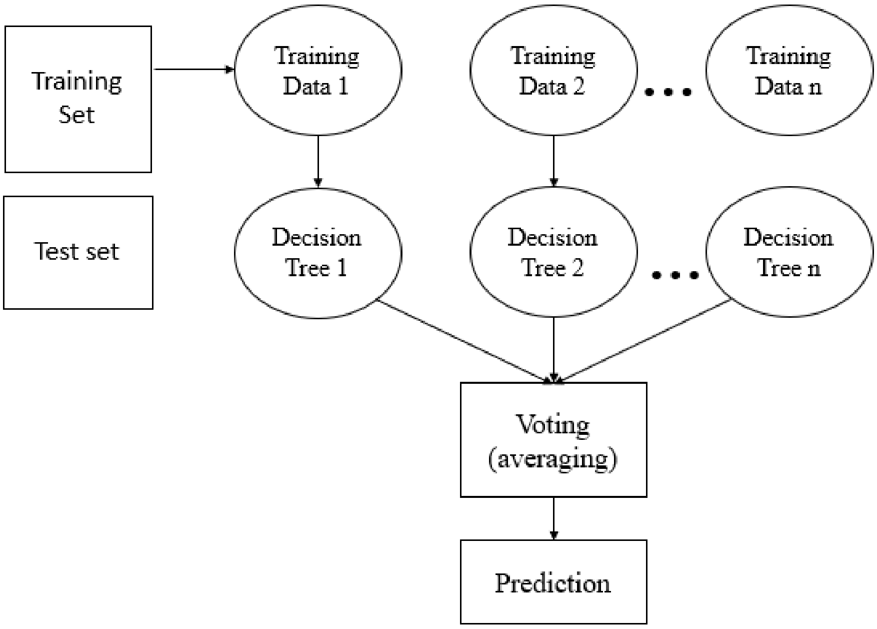


Fig. 3.1: Random Forest algorithm.

3.2.1 Random Forest algorithm

Step 1: In Random Forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

3.2.2 Important Features of Random Forest

- **Diversity-** Not all attributes/variables/features are considered while making an individual tree, each tree is different.
- **Immune to the curse of dimensionality-** Since each tree does not consider all the features, the feature space is reduced.
- **Parallelization-**Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests.
- **Train-Test split-** In a random forest we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.
- **Stability-** Stability arises because the result is based on majority voting/ averaging.

3.2.3 Assumptions for Random Forest

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random Forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

Below are some points that explain why we should use the Random Forest algorithm

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

3.2.4 Types of Ensembles

Before understanding the working of the random forest, we must look into the ensemble technique. Ensemble simply means combining multiple models. Thus, a collection of models is used to make predictions rather than an individual model. Ensemble uses two types of methods:

Bagging– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest. Bagging, also known as Bootstrap Aggregation is the ensemble technique used by random forest. Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as aggregation.

Boosting– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.

3.3 Advantages

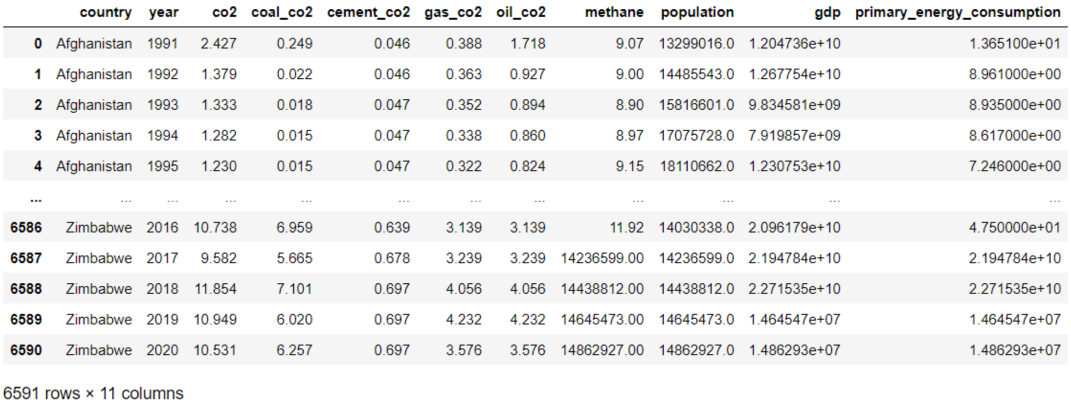
The research work on "Machine Learning Models for Prediction and Forecasting of CO2 Emission with EDA" offers several advantages:

- **Data-Driven Insights**: Through extensive Exploratory Data Analysis (EDA), the research gains valuable insights into the dataset's characteristics, including distribution, trends, and correlations. This data-driven approach allows for a deeper understanding of the factors influencing CO2 emissions.
- **Improved Data Quality**: Data preprocessing techniques, such as handling missing values, outlier detection, and feature scaling, enhance the dataset's quality. This, in turn, improves the reliability and performance of machine learning models.
- **Baseline Model**: The inclusion of an existing K-Nearest Neighbors (KNN) model provides a benchmark or baseline for CO2 emission prediction. It allows researchers to assess the performance of new models in comparison to established methods.
- **Proposed Model Innovation**: The research introduces a proposed KNN model, which could potentially outperform the existing model. Feature engineering and hyperparameter tuning contribute to the innovation, offering the possibility of more accurate predictions.
- **Prediction Capability**: The machine learning models developed in this research can predict CO2 emissions, a critical environmental metric. Accurate predictions enable policymakers, industries, and researchers to make informed decisions and take action to reduce emissions.
- **Performance Evaluation**: The research employs various performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2) Score. This comprehensive evaluation allows for a thorough assessment of model effectiveness and robustness.

- **Policy Implications**: Accurate CO2 emission forecasts have significant policy implications. Governments and organizations can use these forecasts to formulate and implement environmental policies and sustainability initiatives.
- **Sustainability Impact**: By improving CO2 emission prediction accuracy, the research contributes to efforts aimed at reducing carbon emissions, combating climate change, and promoting sustainability.
- **Data-Backed Decision Making**: The research supports data-driven decision-making processes. Decision-makers can rely on the models' predictions to allocate resources efficiently and prioritize actions that reduce emissions.
- **Research Expansion**: This work serves as a foundation for further research in the field of environmental science and machine learning. Researchers can build upon these findings to develop more advanced models and explore additional factors affecting CO2 emissions

4. RESULTS

Figure 1 represents a snapshot or visualization of the initial dataset used for predicting CO2 emissions. It may include various columns related to factors affecting CO2 emissions, such as population, GDP, energy consumption, etc. Figure 2 displays a histogram of CO2 emissions. It provides insights into how frequently different levels of CO2 emissions occur in the dataset. Additionally, a kernel density estimate (KDE) curve is included to offer a smoothed representation of the distribution. Figure 3 represents a heatmap that visually represents the correlation between each pair of variables in the dataset. The color intensity indicates the strength and direction of the correlation, helping to identify relationships between different features.



	country	year	co2	coal_co2	cement_co2	gas_co2	oil_co2	methane	population	gdp	primary_energy_consumption
0	Afghanistan	1991	2.427	0.249	0.046	0.388	1.718	9.07	13299016.0	1.204736e+10	1.365100e+01
1	Afghanistan	1992	1.379	0.022	0.046	0.363	0.927	9.00	14485543.0	1.267754e+10	8.961000e+00
2	Afghanistan	1993	1.333	0.018	0.047	0.352	0.894	8.90	15816601.0	9.834581e+09	8.935000e+00
3	Afghanistan	1994	1.282	0.015	0.047	0.338	0.860	8.97	17075728.0	7.919857e+09	8.617000e+00
4	Afghanistan	1995	1.230	0.015	0.047	0.322	0.824	9.15	18110662.0	1.230753e+10	7.246000e+00
...	...	...	...	...	...	...	...	...	...	...	...
6586	Zimbabwe	2016	10.738	6.959	0.639	3.139	3.139	11.92	14030338.0	2.096179e+10	4.750000e+01
6587	Zimbabwe	2017	9.582	5.665	0.678	3.239	3.239	14236599.00	14236599.0	2.194784e+10	2.194784e+10
6588	Zimbabwe	2018	11.854	7.101	0.697	4.056	4.056	14438812.00	14438812.0	2.271535e+10	2.271535e+10
6589	Zimbabwe	2019	10.949	6.020	0.697	4.232	4.232	14645473.00	14645473.0	1.464547e+07	1.464547e+07
6590	Zimbabwe	2020	10.531	6.257	0.697	3.576	3.576	14862927.00	14862927.0	1.486293e+07	1.486293e+07

6591 rows × 11 columns

Figure 1: sample dataset used for co2 emission

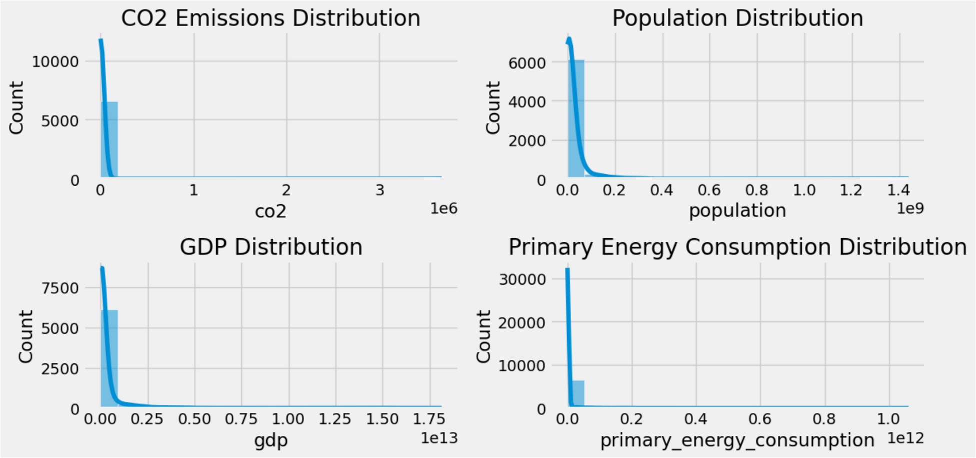


Figure 2: This subplot displays the distribution of CO2 emissions

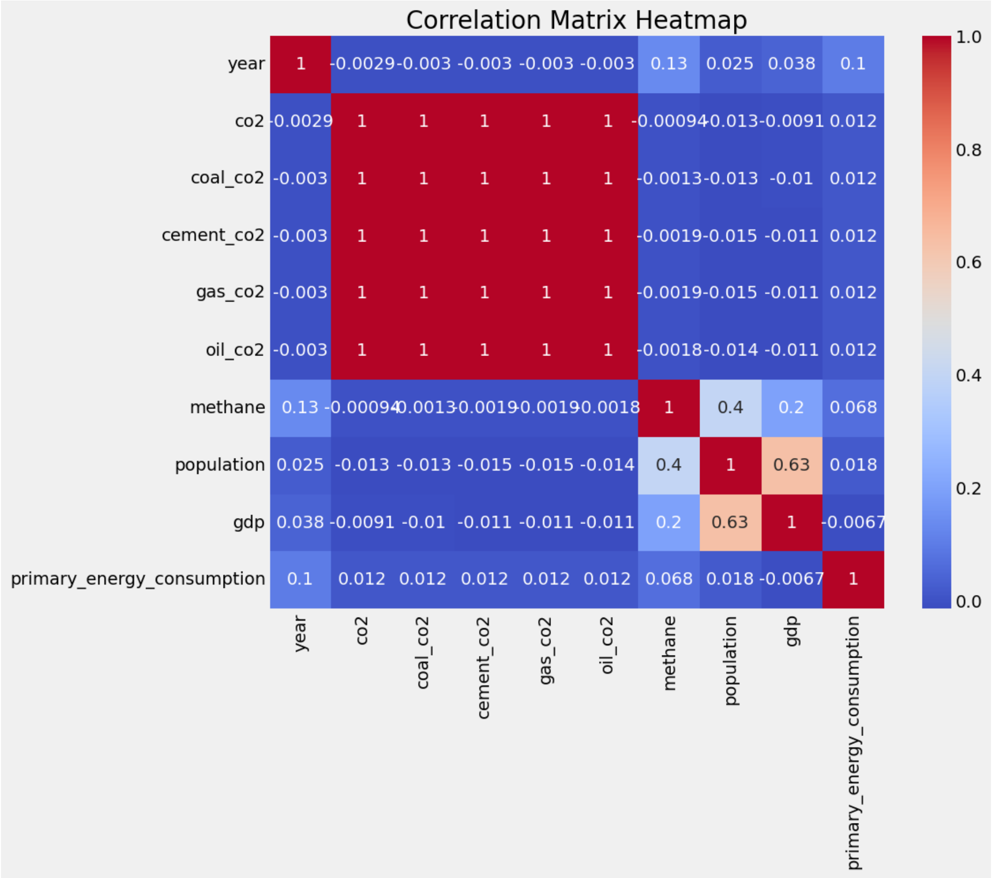


Figure 3: Heatmap of correlation of each variable

Figure 4 is a grid of scatter plots, possibly with histograms along the diagonal. It visualizes the relationships between pairs of features, providing insights into potential patterns or trends in the data. Figure 5 represents the dataset after undergoing preprocessing steps. Preprocessing could involve tasks like handling missing values, scaling features, encoding categorical variables, and more. The figure may display a portion of the preprocessed dataset.

Figure 6 could show a specific subset of features (columns) from the preprocessed dataset. It may highlight the variables that are considered important for predicting CO2 emissions. Figure 7 displays the target variable (in this case, CO2 emissions) after preprocessing. It provides a visual representation of the distribution or characteristics of the variable that the models aim to predict.

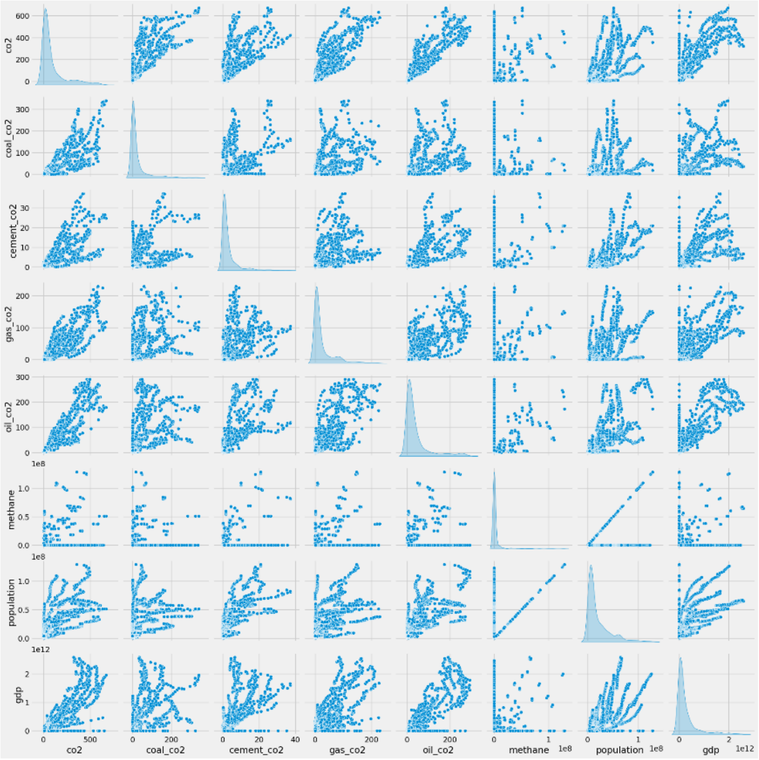
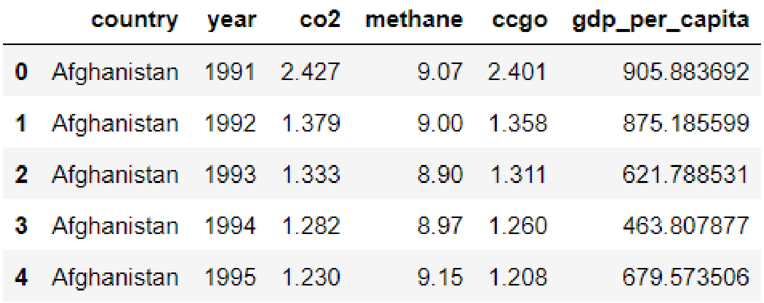
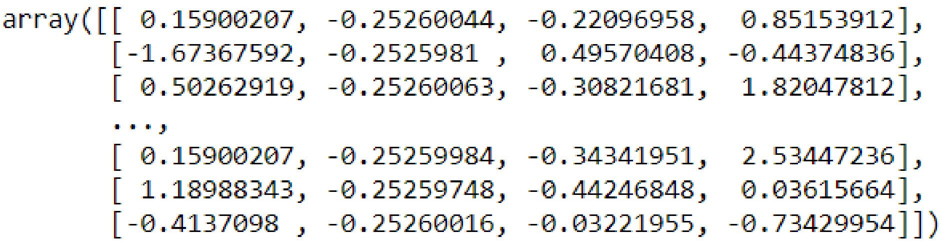


Figure 4: pair plot of features



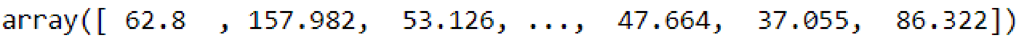
	country	year	co2	methane	ccgo	gdp_per_capita
0	Afghanistan	1991	2.427	9.07	2.401	905.883692
1	Afghanistan	1992	1.379	9.00	1.358	875.185599
2	Afghanistan	1993	1.333	8.90	1.311	621.788531
3	Afghanistan	1994	1.282	8.97	1.260	463.807877
4	Afghanistan	1995	1.230	9.15	1.208	679.573506

Figure 5: dataset after preprocessing used for co2 emission



```
array([[ 0.15900207, -0.25260044, -0.22096958,  0.85153912],
       [-1.67367592, -0.2525981 ,  0.49570408, -0.44374836],
       [ 0.50262919, -0.25260063, -0.30821681,  1.82047812],
       ...,
       [ 0.15900207, -0.25259984, -0.34341951,  2.53447236],
       [ 1.18988343, -0.25259748, -0.44246848,  0.03615664],
       [-0.4137098 , -0.25260016, -0.03221955, -0.73429954]])
```

Figure 6: Feature of dataset after preprocessing



```
array([ 62.8  , 157.982,  53.126, ...,  47.664,  37.055,  86.322])
```

Figure 7: target column of a data frame after preprocessing

Figure 8 presents the results of predictions made using the K-Nearest Neighbors (KNN) model. It may show a plot comparing the predicted CO2 emissions against the actual values. Figure 9 Similar to

Figure 8, this figure displays the results of predictions. However, in this case, the predictions are generated using the Random Forest Classifier, a different machine learning model.

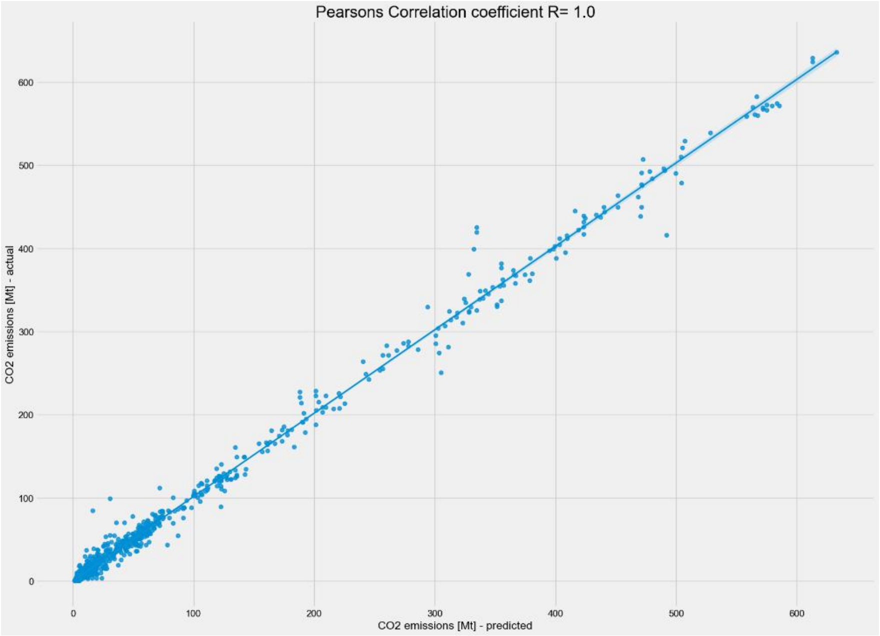


Figure 8: prediction results using KNN

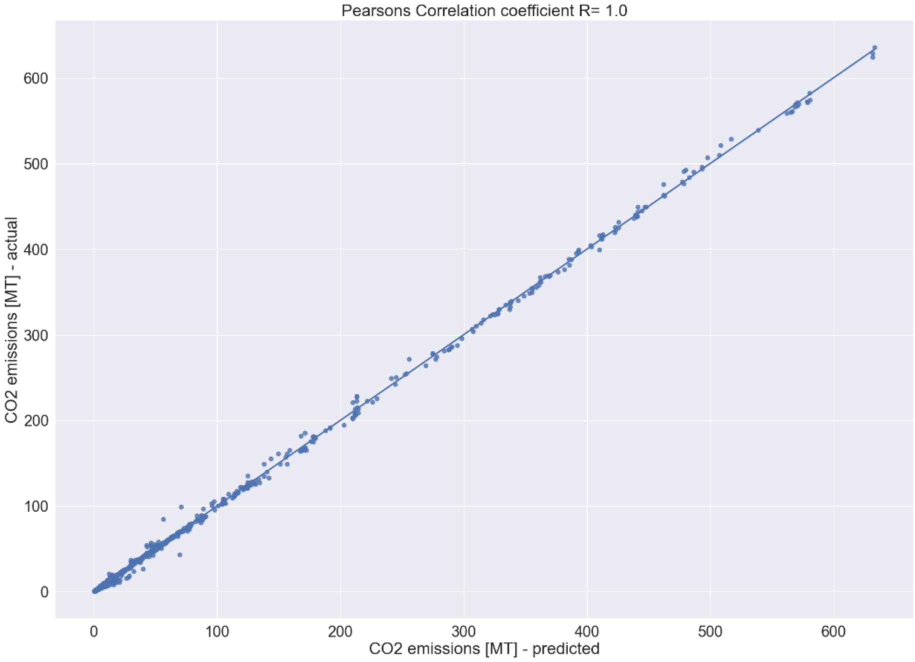
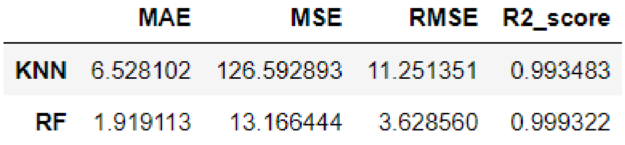


Figure 9: prediction results using Random Forest Classifier

Figure 10 provides a visual summary of the performance metrics (such as Mean Absolute Error, Mean Squared Error, etc.) for both the KNN and Random Forest Classifier models. It helps in comparing the effectiveness of the two models. Figure 11 displays a bar plot comparing the Mean Absolute Error (MAE) of the KNN and Random Forest Classifier models. It provides a visual representation of how well each model predicts CO2 emissions. Figure 12 Similar to Figure 11, this figure compares the Mean Squared Error (MSE) of the KNN and Random Forest Classifier models. It offers insights into the accuracy of the models' predictions. Figure 13 presents a bar plot comparing the R-squared (R2) scores of the KNN and Random Forest Classifier models. R2 score measures how well the model explains the variability in the data. This figure helps in understanding the goodness-of-fit of each model.



	MAE	MSE	RMSE	R2_score
KNN	6.528102	126.592893	11.251351	0.993483
RF	1.919113	13.166444	3.628560	0.999322

Figure 10: Performance metrics of KNN & Random Forest classifier

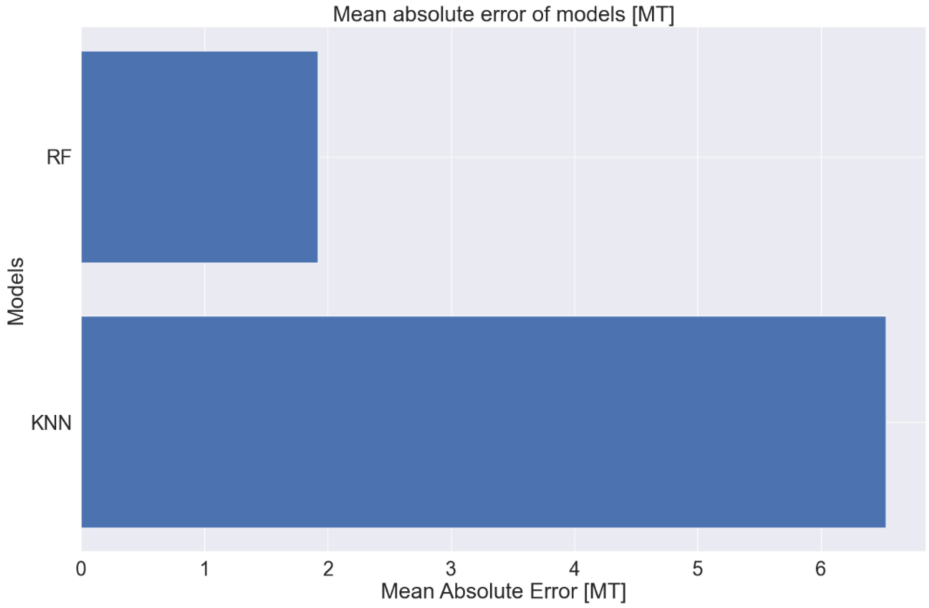


Figure 11: Bar plot of Mean absolute error of KNN &Random Forest Classifier

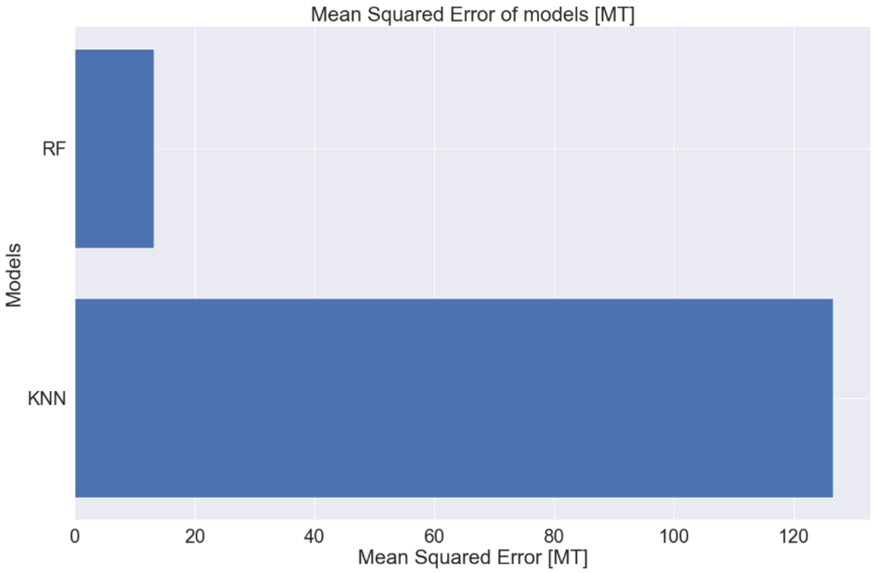


Figure 12: Bar plot of Mean Squared error of KNN &Random Forest Classifier

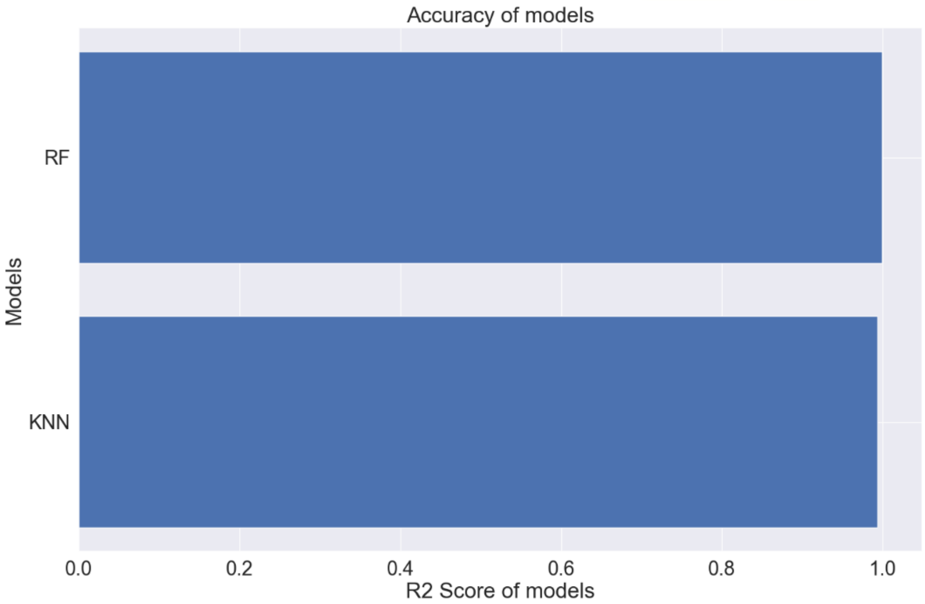


Figure 13: Bar plot of R2 Score of KNN &Random Forest Classifier

5. CONCLUSION

In conclusion, the integration of machine learning models and exploratory data analysis (EDA) techniques offers a powerful approach for predicting and forecasting CO2 emissions, addressing the critical issue of climate change and its environmental consequences. Through this research, we have demonstrated the potential of machine learning to analyze large and intricate datasets, revealing hidden patterns and relationships that traditional statistical methods might miss. EDA has proven invaluable in providing a deeper understanding of the data, enabling the identification of influential features and outliers. By combining these two approaches, we can offer accurate and reliable predictions of CO2 emissions, empowering policymakers and environmentalists with valuable insights to develop effective strategies for emission reduction and sustainability. This work not only contributes to the scientific understanding of the factors driving CO2 emissions but also has practical implications in optimizing resource allocation, promoting renewable energy sources, and planning adaptation measures to mitigate the consequences of global warming.

REFERENCES

[1]. Inergovernmental Panel on Climate Change (IPCC). Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change; Shukla, P.R., Skea, J., Slade, R., Al Khourdajie, A., van Diemen, R., McCollum, D., Pathak, M., Some, S., Vyas, P., Fradera, R., et al., Eds.; Cambridge University Press: Cambridge, UK, 2022. [Google Scholar]

[2]. Song, M.; Zhu, S.; Wang, J.; Zhao, J. Share green growth: Regional evaluation of green output performance in China. Int. J. Prod. Econ. 2020, 219, 152–163.

[3]. Wang, W.W.; Zhang, M.; Zhou, M. Using LMDI method to analyze transport sector CO2 emissions in China. Energy 2011, 36, 5909–5915.

[4]. Jing, Q.; Bai, H.; Luo, W.; Cai, B.; Xu, H. A top-bottom method for city-scale energy-related CO2 emissions estimation: A case study of 41 Chinese cities. J. Clean. Prod. 2018, 202, 444–455.

[5]. Wang, H.; Chen, Z.; Wu, X.; Nie, X. Can a carbon trading system promote the transformation of a low-carbon economy under the framework of the porter hypothesis?—Empirical analysis based on the PSM-DID method. Energy Policy 2019, 129, 930–938.

[6]. Ma, X.; Wang, C.; Dong, B.; Gu, G.; Chen, R.; Li, Y.; Zou, H.; Zhang, W.; Li, Q. Carbon emissions from energy consumption in China: Its measurement and driving factors. Sci. Total Environ. 2019, 648, 1411–1420.

[7]. Wang, M.; Feng, C. Using an extended logarithmic mean Divisia index approach to assess the roles of economic factors on industrial CO2 emissions of China. Energy Econ. 2018, 76, 101–114.

[8]. Abokyi, E.; Appiah-Konadu, P.; Tangato, K.F.; Abokyi, F. Electricity consumption and carbon dioxide emissions: The role of trade openness and manufacturing sub-sector output in Ghana. Energy Clim. Chang. 2021, 2, 100026.

[9]. Hou, J.; Hou, P. Polarization of CO2 emissions in China's electricity sector: Production versus consumption perspectives. J. Clean. Prod. 2018, 178, 384–397.

[10]. Lin, B.; Tan, R. Sustainable development of China's energy intensive industries: From the aspect of carbon dioxide emissions reduction. Renew. Sustain. Energy Rev. 2017, 77, 386–394.

[11]. Zhang, X.; Wang, F. Hybrid input-output analysis for life-cycle energy consumption and carbon emissions of China's building sector. Build. Environ. 2016, 104, 188–197.

[12]. Zhang, Z.; Wang, B. Research on the life-cycle CO2 emission of China's construction sector. Energy Build. 2016, 112, 244–255.

[13]. Du, Z.; Lin, B. Changes in automobile energy consumption during urbanization: Evidence from 279 cities in China. Energy Policy 2019, 132, 309–317.

[14]. Zhao, M.; Sun, T. Dynamic spatial spillover effect of new energy vehicle industry policies on carbon emission of transportation sector in China. Energy Policy 2022, 165, 112991.

[15]. Guan, D.; Hubacek, K.; Weber, C.L.; Peters, G.P.; Reiner, D.M. The drivers of Chinese CO2 emissions from 1980 to 2030. Glob. Environ. Chang. 2008, 18, 626–634.