A MINI PROJECT REPORT
ON

PREDICTION AND FORECASTING OF CO2 EMISSIONS WITH EDA USING KNN AND RANDOM FOREST

Submitted in partial fulfillment of the requirement

For the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

By

K. ASHA (21P61A05D8)

Under the esteemed guidance of

DR. PRAVEEN TALARI
ASSOCIATE PROFESSOR
Dept. of CSE

VIGNANA BHARATHI INSTITUTE OF TECHNOLOGY

(A UGC Autonomous Institution, Approved by AICTE, Affiliated to JNTUH,

Accredited by NBA & NAAC) Aushapur (V), Ghatkesar (M), Medchal (dist.)

December - 2024

Counselling Code : VBIT

VIGNANA BHARATHI Institute of Technology ®

(A UGC Autonomous Institution, Approved by AICTE, Accredited by NBA & NAAC-A Grade, Affiliated to JNTUH)

Aushapur (V), Ghatkesar (M), Hyderabad, Medchal – Dist, Telangana – 501 301.

DEPARTMENT
OF
COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the mini project titled "**PREDICTION AND FORECASTING OF CO2 EMISSION WITH EDA USING KNN AND RANDOM FOREST**" submitted by **K. ASHA (21P61A05D8)** in B-tech IV-I semester Computer Science & Engineering is a record of the bonafide work carried out by her.

The results embodied in this report have not been submitted to any other University for the award of any degree.

INTERNAL GUIDE
Dr. PRAVEEN TALARI

HEAD OF THE DEPARTMENT
Dr. RAJU DARA

EXTERNAL EXAMINER

DECLARATION

I, **K. Asha** bearing hall ticketnumber **(21P61A05D8)** hereby declare that the mini project report entitled **"PREDICTION AND FORECASTING OF CO2 EMISSION WITH EDA USING KNN AND RANDOM FOREST"** under the guidance of **Dr. Praveen Talari *Associate Professor***, Department of Computer Science and Engineering**, Vignana Bharathi Institute of Technology, Hyderabad**, have submitted to Jawaharlal NehruTechnological University Hyderabad, Kukatpally, in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering.

This is a record of bonafide work carried out by me and the results embodied in this project have not been reproduced or copied from any source. The results embodied in this project report haven't been submitted to any other university or institute for the award of any other degree or diploma.

K. ASHA (21P61A05D8)

ACKNOWLEDGEMENT

I am extremely thankful to our beloved Chairman, **Dr. N. Goutham Rao,** and Secretary**, Dr. G. Manohar Reddy** who took a keen interest in providing us the infrastructural facilities for carrying out the project work.

I wholeheartedly thank **Dr. P. V. S. Srinivas Professor & Principal**, and **Dr. Raju Dara**, Head of the Department, Computer Science and Engineering for their encouragement, support and guidance in carrying out the project.

I thank my Project Guide and overall Project Coordinator, **Dr. Praveen Talari *Associate Professor***, for providing me with an excellent project and guiding me in completing our mini-project successfully.

I would like to express my indebtedness to the Section coordinators, **Mrs. P. Yamini Devi**, **Dr. K. Laxmaiah**, Department of Computer Science and Engineering**,** for their valuable guidance during the course of the project work.

I would like to express my sincere thanks to all the staff of Computer Science and Engineering, VBIT, for their kind cooperation and timely help during the course of our project.

Finally, I would like to thank my parents and friends who have always stood by me whenever I was in need of them.

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 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 policy-makers 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 policy-makers 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.

Keywords: Regression Analysis, Feature Engineering, Machine Learning and CO2 Emissions, Forecasting, Random Forest model.

Counselling Code : VBIT

VIGNANA BHARATHI Institute of Technology

(A UGC Autonomous Institution, Approved by AICTE, Accredited by NBA & NAAC–A Grade, Affiliated to JNTUH)

VISION

To become, a Center for Excellence in Computer Science and Engineering with a focused Research, Innovation through Skill Development and Social Responsibility.

MISSION

DM-1: Provide a rigorous theoretical and practical framework across ***State-of-the-art*** infrastructure with an emphasis on ***software development***.

DM-2: Impact the skills necessary to amplify the pedagogy to grow technically and to meet ***interdisciplinary needs*** with collaborations.

DM-3: Inculcate the habit of attaining the professional knowledge, firm ethical values, ***innovative research*** abilities and societal needs.

PROGRAM EDUCATIONAL OBJECTIVES (PEOs)

PEO-01: Domain Knowledge: Synthesize mathematics, science, engineering fundamentals, pragmatic programming concepts to formulate and solve engineering problems using prevalent and prominent software.

PEO-02: Professional Employment: Succeed at entry- level engineering positions in the software industries and government agencies.

PEO-03: Higher Degree: Succeed in the pursuit of higher degree in engineering or other by applying mathematics, science, and engineering fundamentals.

PEO-04: Engineering Citizenship: Communicate and work effectively on team-based engineering projects and practice the ethics of the profession, consistent with a sense of social responsibility.

PEO-05: Lifelong Learning: Recognize the significance of independent learning to become experts in chosen fields and broaden professional knowledge.

PROGRAM SPECIFIC OUTCOMES (PSOs)

PSO-01: Ability to explore emerging technologies in the field of computer science and engineering.

PSO-02: Ability to apply different algorithms indifferent domains to create innovative products.

PSO-03: Ability to gain knowledge to work on various platforms to develop useful and secured applications to the society.

PSO-04: Ability to apply the intelligence of system architecture and organization in designing the new era of computing environment.

PROGRAM OUTCOMES (POs)

Engineering graduates will be able to:

PO-01: Engineering knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

PO-02: Problem analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

PO-03: Design/development of solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and cultural, societal, and environmental considerations.

PO-04: Conduct investigations of complex problems: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

PO-05: Modern tool usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modelling to complex engineering activities with an understanding of the limitations.

PO-06: The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

PO-07: Environment and sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

PO-08: Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

PO-09: Individual and team work: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

PO-10: Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

PO-11: Project management and finance: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

PO-12: Life-long learning: Recognize the need for and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

Project Mapping Table:

Topic	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PO1	PO2	PO3	PO4
Development of a Smart Home Energy Management System	✓	✓	✓	✓	✓		✓			✓		✓				

TABLE OF CONTENTS

List of Figures

1. INTRODUCTION

1.1 INTRODUCTION TO THE SYSTEM

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 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 encomp024asses 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. Policy-makers, 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.

1.3 OBJECTIVES

The major objectives of the "Prediction and Forecasting of Co2 Emission with EDA using KNN and RANDOM FOREST" system are described as follows:

- The primary goal is to create machine learning models that accurately predict CO2 emissions based on historical data and relevant features.
- Evaluate Machine learning models (e.g., KNN, Random Forest) to identify the most effective model for CO2 emissions prediction.
- To support long-term planning and decision-making related to environmental policies and emission reduction strategies.
- Conduct comprehensive EDA to understand the data's characteristics, uncover patterns, and identify relationships between CO2 emissions and influencing factors.

1.4 AIM OF THE PROJECT

The aim of the project focused on Predicting and Forecasting of CO2 emissions with Exploratory Data Analysis (EDA) using KNN and Random Forest (Machine learning techniques). Understand and articulate the specific goal of the prediction and forecasting. This could be to predict future CO2 emissions, identify factors contributing to emissions, or forecast emissions for specific regions or industries.

3. 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.

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, China has become a large energy consumer and carbon emitter. 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.

China's CO2 emissions mainly come from electricity generation, industry, construction, transportation, and agriculture. 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.

China's CO2 emission reduction effect and environmental protection policies are expected to significantly impact the global climate [16]. As a signatory to the Paris Agreement, China had committed to achieving a carbon peak by 2030 [17] and achieving carbon neutrality by 2060. However, as a fast-growing carbon polluter, China's commitment holds particular weight, because achieving a carbon peak and carbon neutrality involves technological and economic development, and China's CO2 emissions per unit of gross domestic product (GDP) are still at the highest level in the world.

To achieve its carbon peak and neutrality targets, it is vital to accurately predict China's future CO2 emissions and identify the factors influencing those CO2 emissions, to inform corresponding emission reduction policies.

Studies on CO2 Emission Predictions

Grey prediction models are widely used in CO2 emission forecasting. A multi-variable grey model (GM(1, N) model), based on a linear time-varying parameters discrete grey model (TDGM(1, N)), was established to predict the CO2 emissions from Anhui Province. The key challenge in forecasting CO2 emissions is investigating the dynamic lag relationships. As such, an enhanced dynamic time-delay discrete grey forecasting model was proposed to predict outcomes for systems with dynamic time-lag effects. The model was used to significantly improve the fitting and prediction performance, using CO2 emissions data from 1995 to 2017.

Scenario analysis is widely used to predict long-term CO2 emissions. Zhang et al. (2021) used regression analysis to quantify the impact of three factors on CO2 emissions: Total population, educational attainment, and per capita GDP. Then, CO2 emission predictions for 2018 to 2100 were developed using these three factors across different scenario settings, involving multiple model parameters and explanatory variables. Li et al. (2020) used the Generalized Dividing Index Method to quantify the contribution rate of each factor influencing the CO2 emissions of China's construction industry and predicted the carbon peaks of China's construction industry in three scenarios.

Studies on the Factors Influencing CO2 Emissions

The existing literature has analyzed the effect of industrial structure and economic development on CO2 emissions. Zheng et al. (2019) applied the logarithmic mean Divisia index (LMDI) to estimate seven socioeconomic drivers impacting changes in CO2 emissions in China since 2000. The results showed that industrial structure and energy mix resulted in emissions growth in some regions, but these two drivers led to emission reductions at the national level. Dong et al. (2020) applied the extended STIRPAT decomposition model, Tapio decoupling model, and grey relation analysis to analyze the relationships between CO2 emissions, industrial structure, and economic growth in China from 2000 to 2017. The results showed that the key steps needed for China to reach its carbon peak goal are accelerating the achievement of a clean energy structure, strengthening the strength and speed of industrial structure adjustment, and reducing the dependence on fossil energy. Panel econometric techniques were applied to examine the nexus between CO2 emissions, urbanization level, and industrial structure in North China between 2004 and 2019. The empirical results support the long-term equilibrium relationship between CO2 emissions and industrial structure in North China, indicating that industrial structure significantly impacts CO2 emissions.

The interaction between renewable energy and CO2 emissions is also a key research direction.

Dongetal. (2017) used panel unit root, cointegration, causality tests, and the augmented mean group (AMG) estimator to investigate the nexus between per capita CO2 emissions, gross domestic product (GDP), and natural gas and renewable energy consumption in a 1985–2016 sample of BRICS countries (defined as Brazil, Russia, India, China, and South Africa). The results found that natural gas and renewable energy consumption lowers CO2 emissions. Zheng et al. (2021) analyzed the influence of renewable energy generation in China on CO2 emissions using a quantile regression model and path analysis of inter-provincial panel data from 2008 to 2017. The results found that renewable energy has a less direct effect on CO2 emissions, but energy intensity and GDP per capita inhibit CO2 emissions. Abbasi et al. (2022) applied novel dynamic ARDL simulations and Frequency Domain Causality (FDC) models to analyze the environmental factors influencing China's CO2 emissions from 1980 to 2018. The results showed that renewable energy consumption is crucial for achieving sustainable environmental goals, despite there being a short-term detrimental impact on CO2 emissions.

Studies on Using ML for CO2 Emissions

ML applies data and experience to automatically improve computer algorithms using probability theory, statistics, approximation theory, and complex algorithms. For random, high-dimensional, nonlinear, and feedback complex systems, ML uses different core algorithms to identify and predict the patterns of the complex system.

Sun and Ren (2021) combined ensemble empirical mode decomposition (EEMD) and the backpropagation neural network based on particle swarm optimization (PSOBP) to develop short-term univariate predictions of CO2 emissions in China. However, univariate predictions only mine CO2 emissions data and do not consider the influence of other factors on CO2 emissions.

To address this, multivariate time-series forecasting is used to predict CO2 emissions. The least absolute shrinkage and selection operator (LASSO) and principal component analysis (PCA) are used to select features. Then, support vector regression (SVR) and differential evolution-gray wolf optimization (DE-GWO) are used to improve the prediction accuracy of China's CO2 emissions. Results using these approaches show that coal and oil consumption, plate glass, pig iron, and crude steel production are important factors affecting CO2 emissions, and the DE-GWO optimized SVR has an effective level of prediction accuracy. The XGBoost model has also been used to predict the CO2 emissions of expanding megacities. Under the synergistic effect of multiple factors, population, land size, and GDP are still the primary forces driving CO2 emissions. A three-layer perceptron neural network (3-layer PNN) performed well in predicting transportation CO2 emission in Zhuhai, China. The Autoregressive Integrated Moving Average (ARIMA) was used to predict CO2 emissions from four Chinese provinces over three future years using data from 1997 to 2017.

ML is also used to analyze the correlation between feature variables. A different study applied a Causal

Direction from the Dependency (D2C) algorithm to investigate the relationships between coal, solar, wind, and CO2 emissions. Using data from 1990 to 2014 in China, the long short-term memory (LSTM) method was applied to investigate the impact of energy consumption, financial development (FD), GDP, population, and renewable energy on CO2 emissions and forecast emission trends.

The literature review revealed three main points and research gaps. First, many models based on grey predictions have been used to predict CO2 emissions. However, instead of using raw data, grey predictions use a series of generated data that approximate exponential laws. Therefore, grey prediction requires high smoothness in the original data series. When the original data column has poor smoothness, the prediction accuracy of the grey prediction model is not high. Although scenario analysis is suitable for predicting long-term CO2 emissions, the accuracy of prediction results is limited by the ability of decision-makers to develop scenarios and the validity of data.

Second, some studies have investigated the impact of economic growth and industrial structure on CO2 emissions, while others have studied and investigated the relationship between renewable energy, economic growth, and CO2 emissions. However, few studies have put CO2 emissions, industrial structure, and renewable energy consumption into a single research framework to forecast China's CO2 emissions.

Third, some studies have applied ML to predict CO2 emission data in univariate time series, without considering industrial structure and renewable energy, which are closely related to CO2 emissions. Some studies have conducted multivariate time predictions for CO2 emissions in a specific industry. However, few studies have used ML to predict China's CO2 emissions in a multivariable time series.

The three research gaps listed above exhibit where this study finds its objective. Concerning the methodology, we applied shallow learning to predict China's future short-term CO2 emissions without requiring the data samples to meet strict statistical properties. This improved the prediction ability of the model using realistic data. Concerning practicability, we converted the time-series data into supervised learning data with labels and used historical data to predict future data to mitigate the unavailability of sample data. Finally, from an empirical perspective, we integrated renewable energy with the industrial structure and CO2 emissions in a research framework to construct a multivariate forecasting model. This model has a stronger forecasting ability than univariate time series and effectively identified the impact of industrial structure and renewable energy on CO2 emissions.

2.1 EXISTING SYSTEM

K-Nearest Neighbors (KNN) is a simple yet powerful supervised machine learning algorithm used for classification and regression tasks. It's based on the idea that data points with similar features tend to belong to the same class or have similar values in the case of regression. KNN is a distance-based classification algorithm. It assigns a new data point to the majority class of its k-nearest neighbors. The choice of 'k' (the number of neighbors) is a crucial hyperparameter that impacts the model's performance. KNN is an instance-based learning method, meaning it doesn't build a model during training. Instead, it memorizes the entire training dataset and uses it for predictions.

Working Principle:

Step 1: Distance Metric:

- KNN uses a distance metric (typically Euclidean distance, but others like Manhattan, Minkowski, etc., are also possible) to measure the similarity between data points. The algorithm finds the 'k' nearest neighbors with the smallest distances to the new data point.

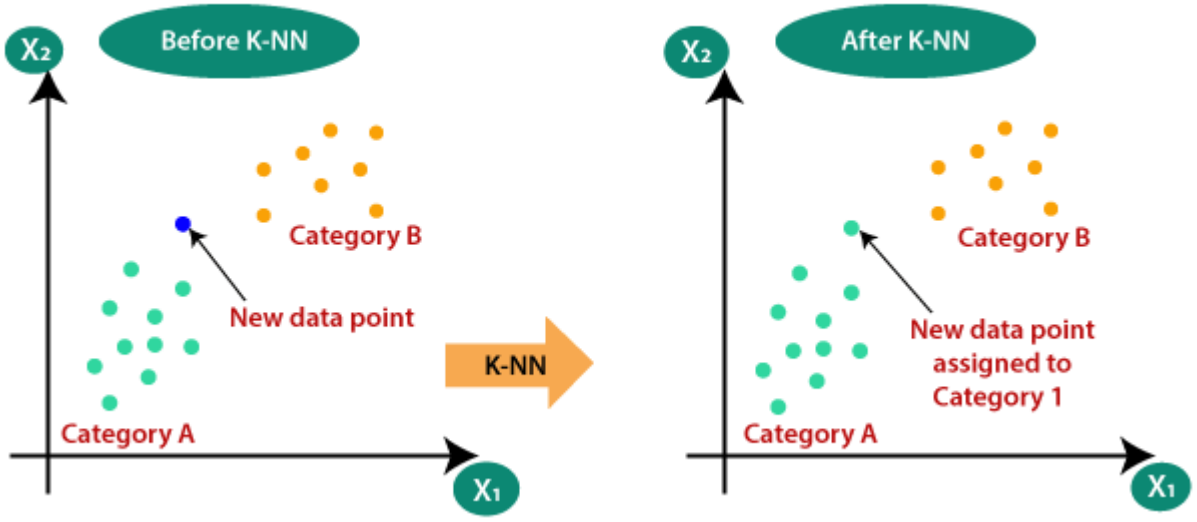

Figure 4.1. KNN initialization

- Voting Mechanism: For classification, KNN uses a majority voting mechanism among its neighbors. The class that occurs most frequently among the neighbors is assigned to the new data point. For regression, it takes the mean (or median) value of the 'k' nearest neighbors as the prediction.

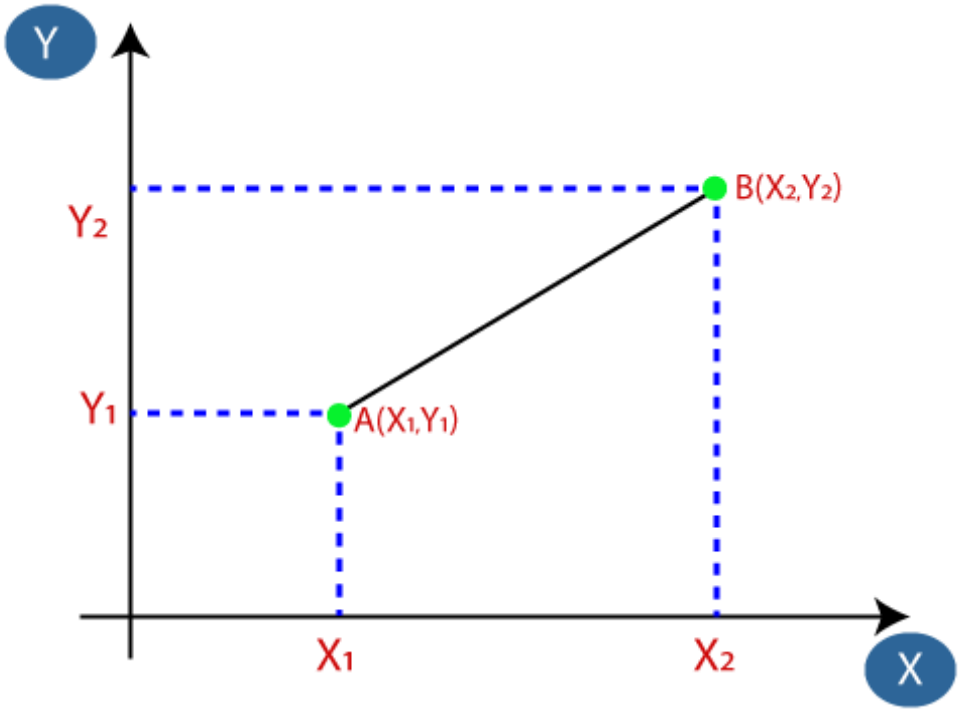

Figure 4.2. Distance measurement in KNN.

Step 2. Hyper parameter 'k':

- Choosing the Right 'k': The choice of 'k' is crucial. A small 'k' makes the model sensitive to noise and outliers but may capture local patterns well. A large 'k' smooths out local variations but can make the model less accurate.
- Methods for Choosing 'k': Cross-validation, grid search, and domain knowledge are common approaches to determine the optimal 'k' value.
- Simplicity: KNN is easy to understand and implement, making it a suitable choice for beginners.
- No Training Phase: It doesn't require a training phase since it memorizes the data, making it suitable for online learning and non-stationary data.
- Non-Parametric: KNN is non-parametric, meaning it makes no assumptions about the underlying data distribution.
- Works for Multiclass Problems: KNN naturally handles multi-class classification problems.

Step 3. Variants:

- Weighted KNN: Assigns different weights to neighbors based on their distance. Closer neighbors have a greater influence on the prediction.
- KNN with Feature Scaling: Feature scaling is essential when using KNN, as it's distance-based. Standardization (scaling features to have mean=0 and standard deviation=1) is often applied.
- KD-Tree and Ball-Tree: These data structures can be used to speed up KNN search for large datasets.

2.2 PROPOSED SYSTEM

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 4.3 shows the proposed system model. The detailed operation is 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 RANDOM FOREST 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 Random Forest model.
- Model Training: Train the proposed Random Forest model using the training data.

Step 5. Prediction:

- Predict CO2 Emissions: Use both the existing and proposed 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 models to assess whether the proposed model provides better predictions.

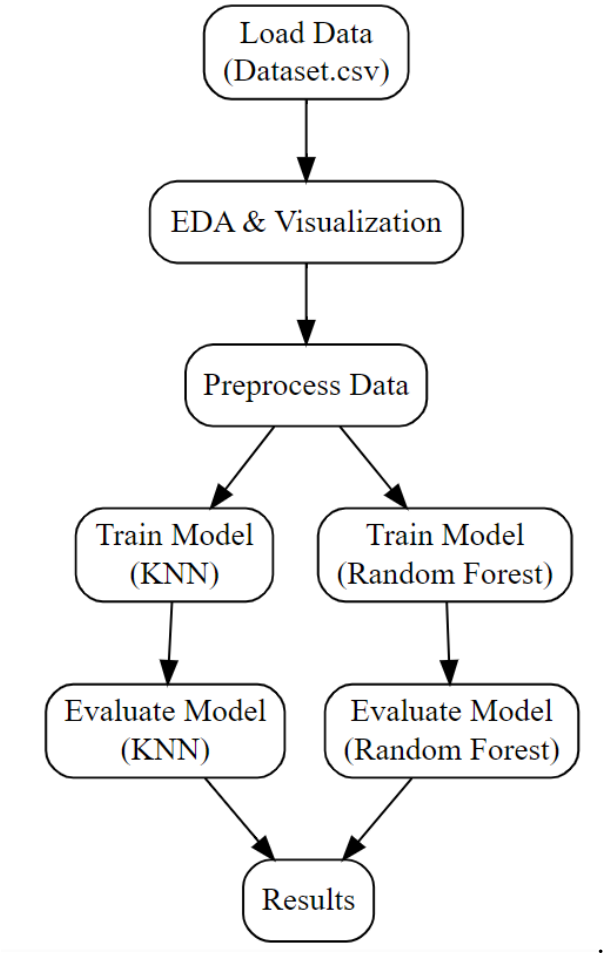

Fig. 4.3: Block diagram of the proposed system.

EDA

The following EDA was performed in this work:

Histogram Plots: Histograms are generated for all numeric features in the dataset to visualize the distribution of each variable. This code snippet creates a figure with a size of 15x15, specifies the axis for plotting, and then generates histograms for each numeric feature in the DataFrame 'df.' These histograms help you understand the distribution of values within each feature.

Countplot for Target Variable: A countplot is created to check if the dataset is balanced or not concerning the 'target' variable. The counterplot visualizes the distribution of the 'target' variable, which is typically used in classification tasks to indicate the class labels. By examining the count of each class, you can assess whether the dataset is balanced or skewed towards certain classes.

Correlation Heatmap: A heatmap is generated to visualize the correlation between different features in the dataset. The code calculates the correlation matrix for all features in the DataFrame 'df' and selects the features with the highest correlation. It then plots a heatmap to display the pairwise correlations between these selected features. The 'annot=True' parameter adds numerical values to the heatmap cells, providing insight into the strength and direction of correlations.

Data Preprocessing

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step in creating a machine-learning model. When creating a machine learning project, it is not always the case that we come across clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put it in a formatted way. So, for this, we use data pre-processing task.

Real-world data generally contains noises, and missing values, and may be in an unusable format that cannot be directly used for machine learning models. Data pre-processing is a required task for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

One-Hot Encoding: Categorical variables are one-hot encoded to convert them into a numerical format suitable for machine learning models. The code uses the pd.get_dummies() function to create binary columns for each category within categorical variables. This transformation allows machine learning algorithms to work with categorical data effectively.

Standardization: Standard Scaler is applied to scale numeric features, ensuring that they have a mean of 0 and a standard deviation of 1. The 'Standard Scaler' from scikit-learn is used to standardize specific numeric features. Standardization is a common preprocessing step to bring features to a similar scale, which can improve the performance of some machine learning algorithms. This transformation is important for several reasons:

- **Equal Scaling**: Standard Scaler scales each feature to have the same scale. This is crucial for algorithms that are sensitive to the scale of features, such as gradient-based optimization algorithms (e.g., in neural networks) and distance-based algorithms (e.g., k-means clustering).
- **Mean Centering**: By subtracting the mean from each data point, Standard Scaler centers the data around zero. This can help algorithms converge faster during training and improve their performance.
- **Normalization**: Scaling by the standard deviation normalizes the data, ensuring that features have comparable variances. This can prevent certain features from dominating others in the modeling process.
- **Interpretability**: Standardized data is more interpretable because it puts all features on a common scale, making it easier to compare the relative importance of features.

Dataset Splitting

In machine learning data pre-processing, we divide our dataset into a training set and a test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model. Suppose we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models. If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance. So we always try to make a machine learning model that performs well with the training set and also with the test dataset.

Training Set: A subset of the dataset to train the machine learning model, and we already know the output.

Test set: A subset of the dataset to test the machine learning model, and by using the test set, the model predicts the output.

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 improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains several 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, it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

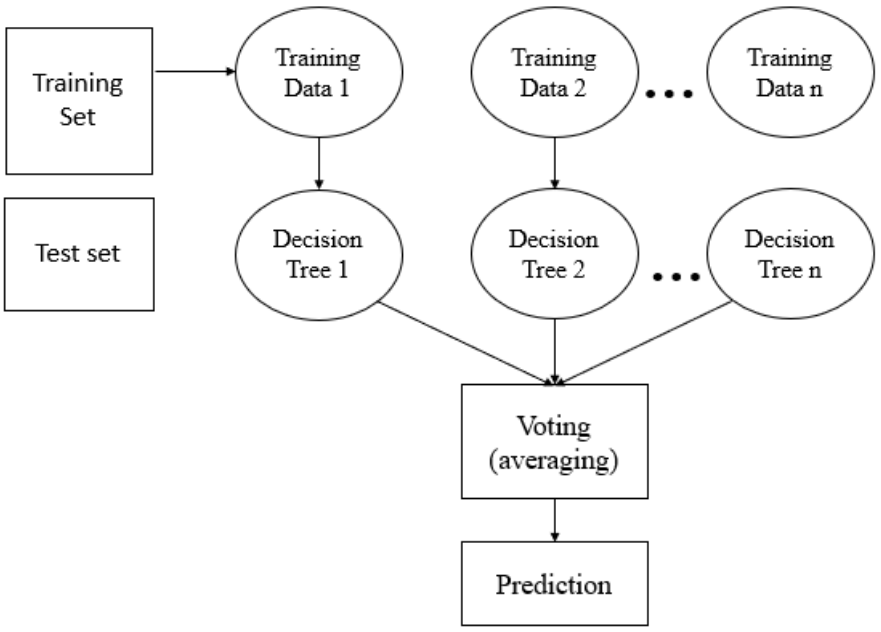

Fig. 4.4: Random Forest algorithm.

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.

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** 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.

Assumptions for Random Forest

Since the random forest combines multiple trees to predict the class of the dataset, 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.

2.3 SCOPE OF THE PROJECT

The scope of this project encompasses the development of a machine learning model (random forest) for predicting and forecasting CO2 emissions, integrating Exploratory Data Analysis (EDA) as a foundational component. Initially, the project involves collecting and preparing comprehensive datasets, including CO2 emission records and related influencing factors such as economic indicators and energy consumption metrics. Through EDA, the project will focus on data cleaning, transformation, and visualization to uncover patterns, trends, and correlations essential for accurate modeling. Feature engineering will follow, aimed at creating and selecting relevant variables that impact CO2 emissions. The core of the project includes selecting and developing appropriate machine learning models—ranging from regression to advanced time-series and deep learning models—while ensuring robust evaluation through metrics such as RMSE and R^2 score. The final model will be employed to generate forecasts, accompanied by uncertainty analysis to assess prediction reliability. The project will also involve the deployment of the model, ongoing monitoring, and the provision of insights and recommendations based on the results, with a focus on effectively communicating findings to stakeholders through detailed reporting and visualizations.

3. ANALYSIS

3.1 FEASIBILITY STUDY

This feasibility study structure ensures a comprehensive analysis of the potential for using machine learning to predict and forecast CO2 emissions. The emphasis on EDA helps to uncover insights that guide model selection and improve accuracy. Be sure to iterate on your findings and adapt your approach based on the specific context and available resources.

Three key considerations involved in the feasibility analysis are

- Technical Feasibility
- Operational Feasibility
- Economic Feasibility

3.1.1 Technical Feasibility

The technical feasibility of employing Predicting and forecasting CO2 emissions using knn and random forest is robust, supported by access to quality datasets from reputable sources. Exploratory Data Analysis (EDA) is crucial for evaluating data integrity, visualizing relationships, and identifying patterns that inform model selection.

3.1.2 Operational Feasibility

The operational feasibility of implementing machine learning models (knn and random forest) for CO2 emission prediction and forecasting hinges on integrating data collection, processing, and analysis within existing workflows. Organizations must establish efficient processes for data acquisition from reliable sources and ensure timely updates. Training staff on data analysis and machine learning techniques is essential for effective model deployment and maintenance.

3.1.3 Economical Feasibility

To decide whether a project is economically feasible, we have to consider various factors.

- Cost-benefit analysis
- Long-term returns
- Maintenance costs

The economic feasibility of Predicting and Forecasting CO2 emissions using knn and random forest involves a careful cost-benefit analysis. Initial investments may include data acquisition, software licenses, and computational resources, alongside expenses for skilled personnel and training. However, the potential benefits, such as improved regulatory compliance, enhanced decision-making, and reduced carbon footprints, can yield significant long-term savings and environmental impact.

4. HARDWARE AND SOFTWARE REQUIREMENTS

4.1 Hardware Requirements

Minimum hardware requirements are very dependent on the particular software being developed by a given En-thought Python / Canopy / VS Code user. Applications that need to store large arrays/objects in memory will require more RAM, whereas applications that need to perform numerous calculations or tasks more quickly will require a faster processor.

Operating system	:	Windows, Linux
Processor	:	minimum intel i3
Ram	:	minimum 4 GB
Hard disk	:	minimum 250GB

4.2 Software Requirements

The functional requirements or the overall description documents include the product perspective and features, operating system and operating environment, graphics requirements, design constraints and user documentation.

The appropriation of requirements and implementation constraints gives the general overview of the project in regard to what the areas of strength and deficit are and how to tackle them.

- Python IDLE 3.7 version (or)
- Anaconda 3.7 (or)
- Jupiter (or)
- Google Colab

5. SYSTEM DESIGN

The system design for a machine learning model aimed at predicting and forecasting CO2 emissions integrates several key components. It begins with **data collection** from diverse sources such as government databases, industrial reports, and environmental sensors. This data is then processed through a **data preprocessing pipeline** that includes cleaning, normalization, and feature engineering to prepare it for analysis.

5.1 Architecture Diagram

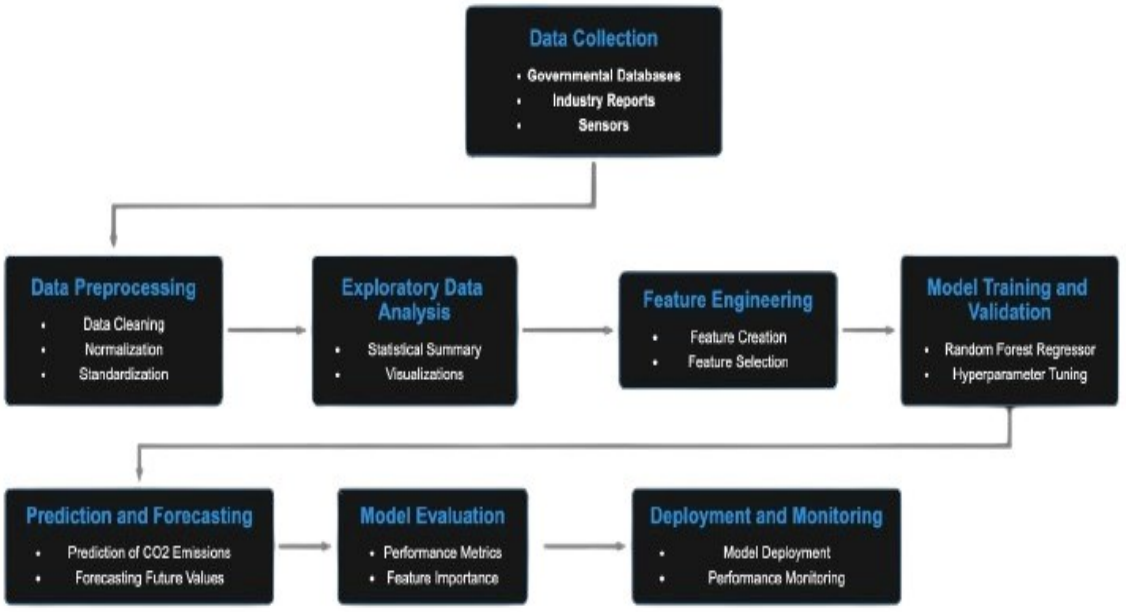

fig :5.1 Architecture diagram of Prediction of CO2 Emission

The architecture for a machine learning model predicting and forecasting CO2 emissions consists of several key components. At the base, Data Sources feed into a Data Ingestion Layer, where data is collected from various sources such as government databases and sensor networks. This data is then processed in the Data Preprocessing Layer, which includes cleaning, normalization, and feature engineering. Next, the Exploratory Data Analysis (EDA) Modulee conducts visualizations and statistical analyses to uncover patterns and correlations, feeding insights back into the preprocessing layer if necessary. The best-performing model is selected and passed to the Model Deployment Layer, which exposes it via APIs for real-time predictions. This architecture is typically orchestrated within a cloud environment for scalability and efficiency.

5.2 Use Case Diagram

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

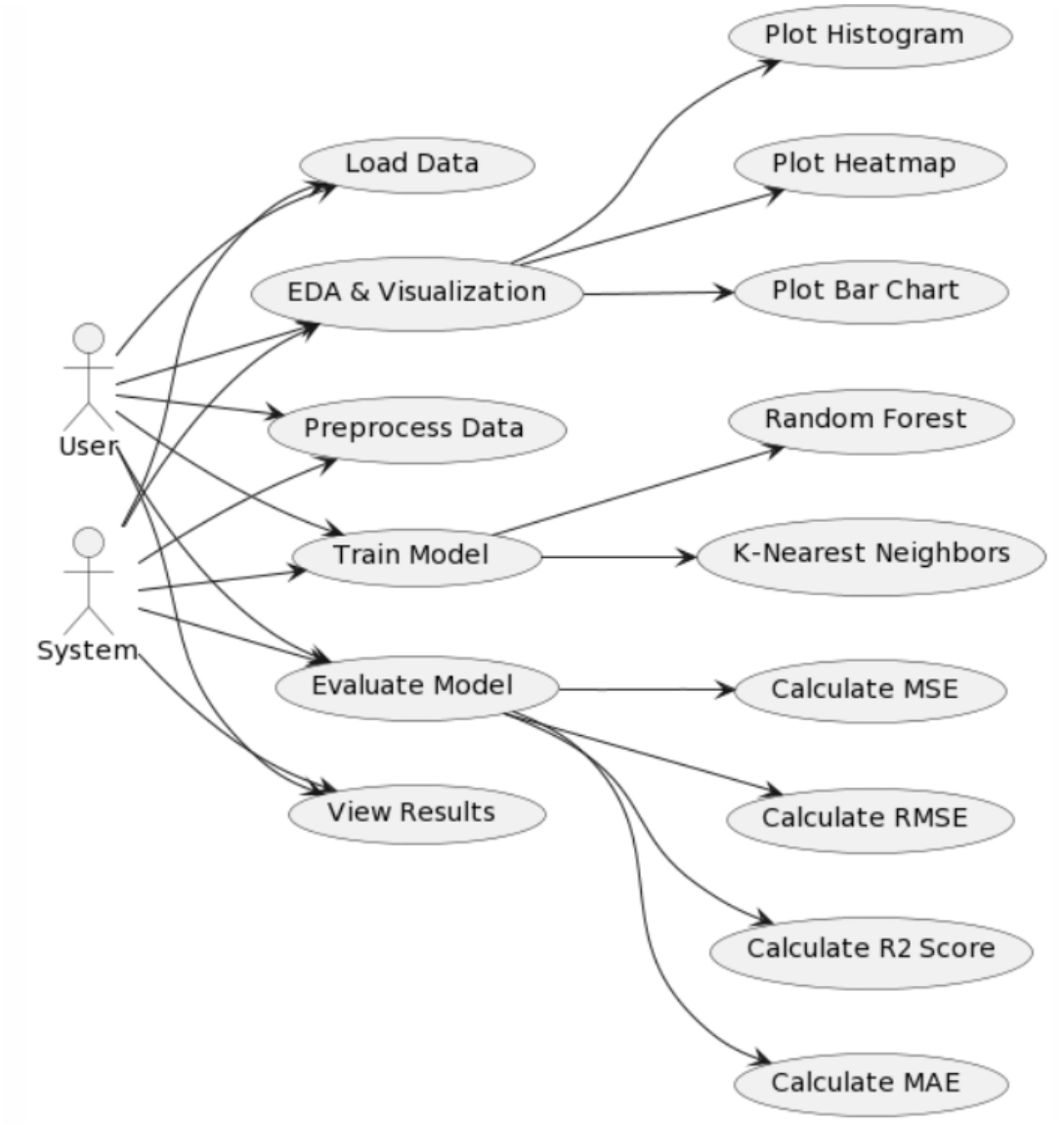

fig :5.2 Use Case Diagram of Prediction of CO2 Emission

5.3 Class Diagram

The class diagram is used to refine the use case diagram and define a detailed design of the system. The class diagram classifies the actors defined in the use case diagram into a set of interrelated classes. The relationship or association between the classes can be either an "is-a" or "has-a" relationship. Each class in the class diagram may be capable of providing certain functionalities. These functionalities provided by the class are termed "methods" of the class. Apart from this, each class may have certain "attributes" that uniquely identify the class.

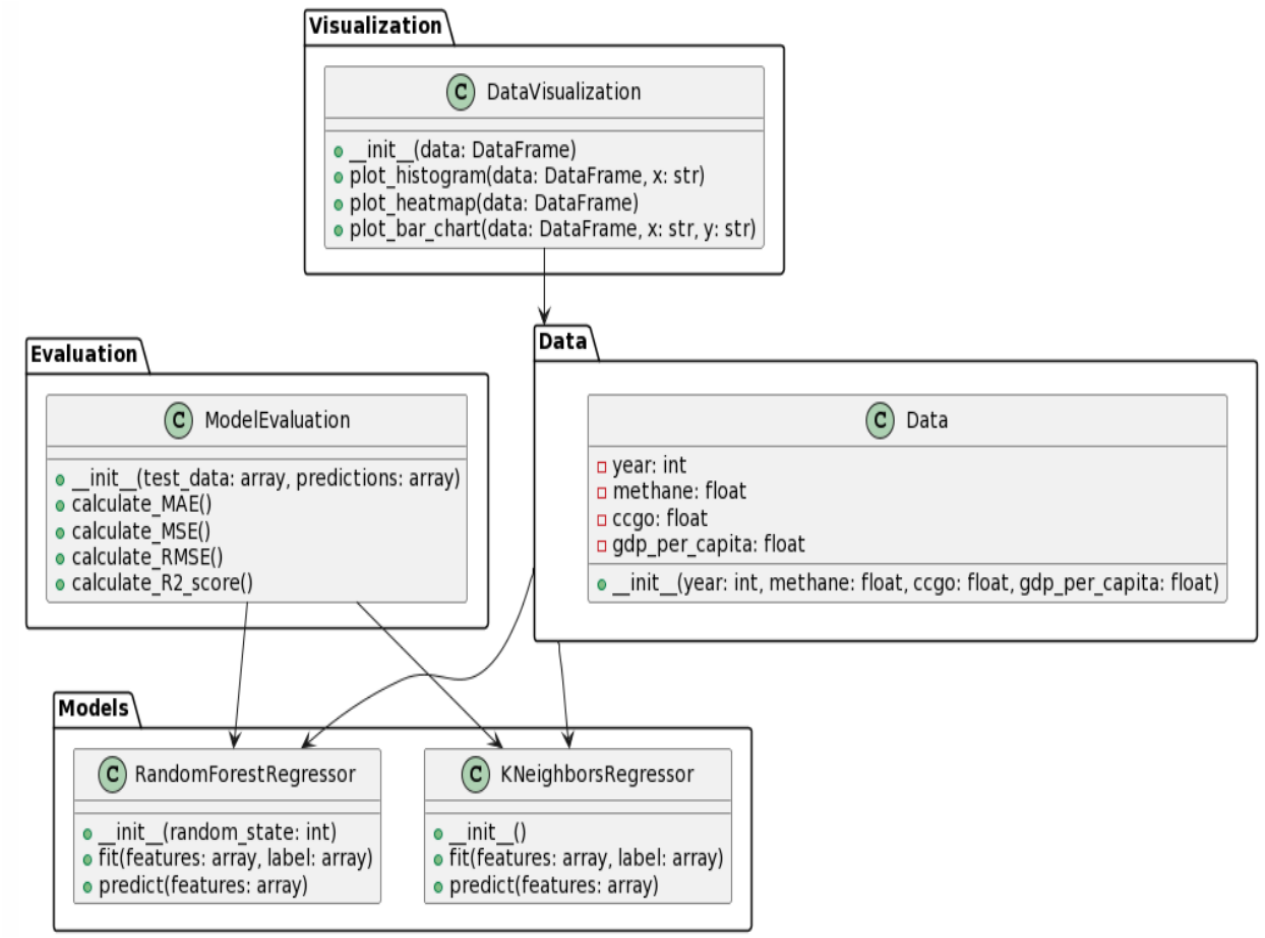

fig :5.3 Class Diagram of Prediction of CO2 Emission

5.4 Activity Diagram

Activity diagrams are graphical representations of Workflows of stepwise activities and actions with support for choice, iteration, and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

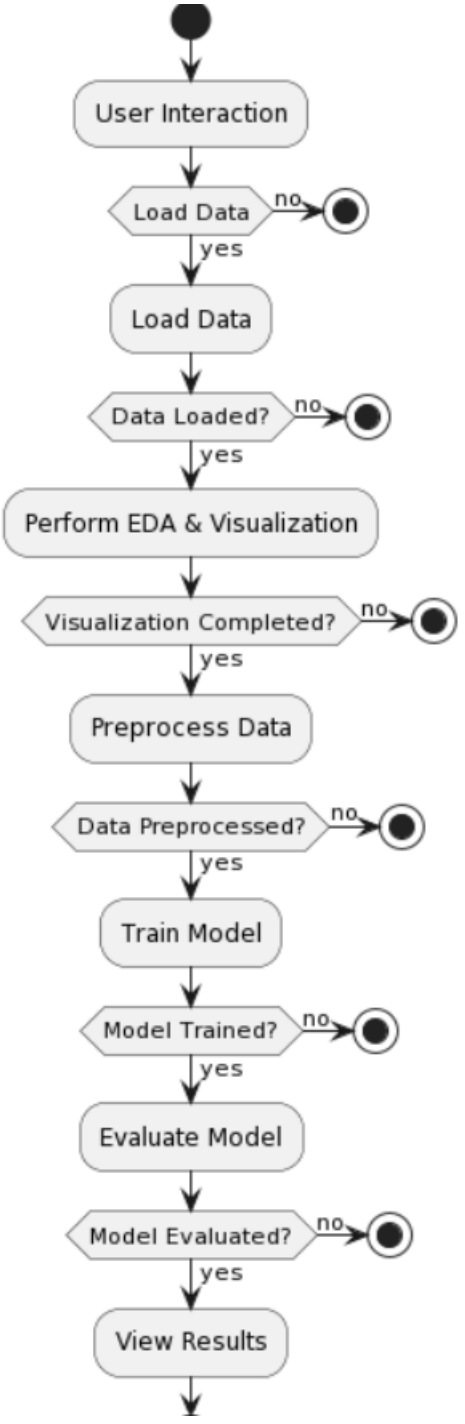

fig :5.4 Activity Diagram of Prediction of CO2 Emission

5.5 Sequence Diagram

A **sequence diagram** in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. A sequence diagram shows, as parallel vertical lines ("lifelines"), different processes or objects that live simultaneously, and as horizontal arrows, the messages exchanged between them, in the order in which they occur. This allows the specification of simple runtime scenarios in a graphical manner.

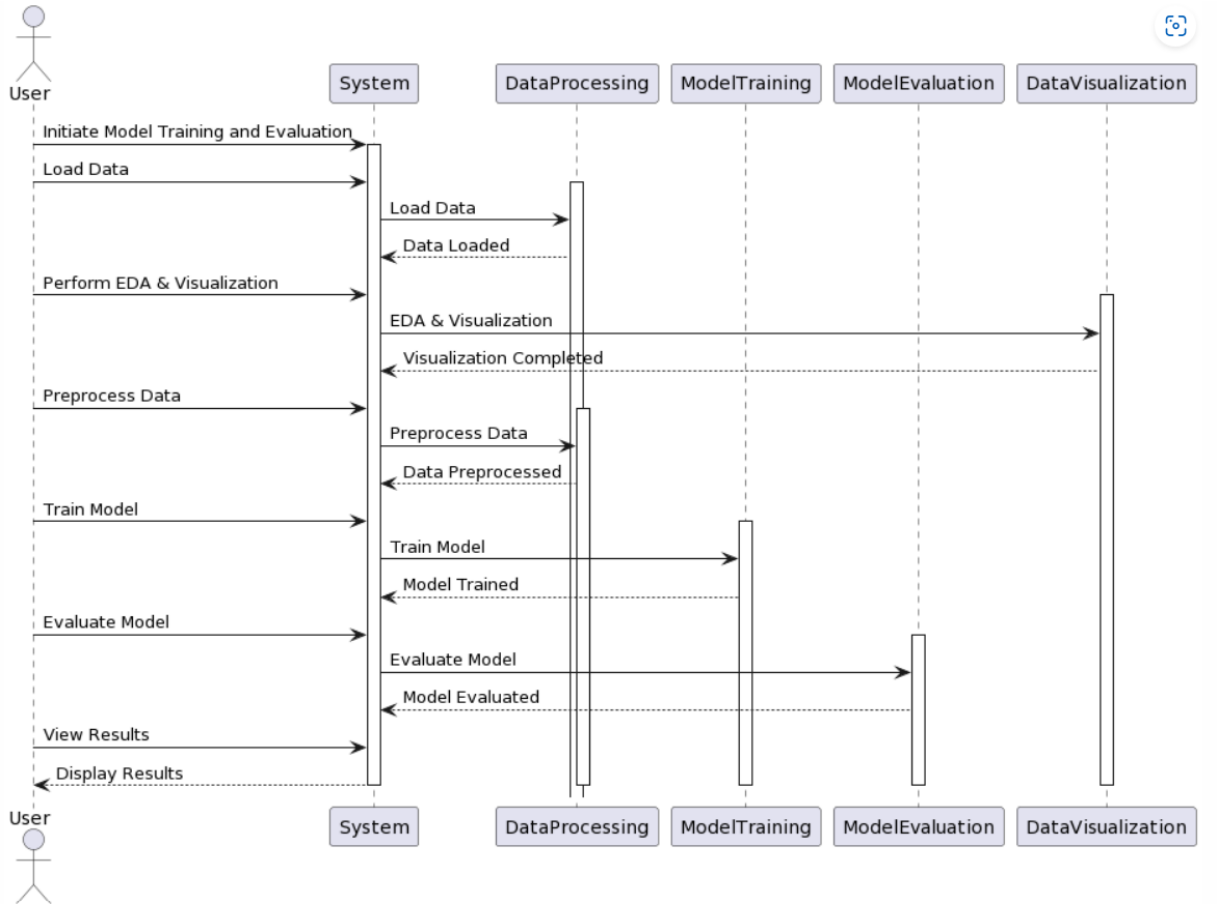

fig :5.5 Sequence Diagram of Prediction of CO2 Emission

6. METHODOLOGY

6.1 MODULES

Output Design

Outputs from computer systems are required primarily to communicate the results of processing to users. They are also used to provides a permanent copy of the results for later consultation. The various types of outputs in general are:

- External Outputs, whose destination is outside the organization
- Internal Outputs whose destination is within organization and they are the
- User's main interface with the computer.
- Operational outputs whose use is purely within the computer department.
- Interface outputs, which involve the user in communicating directly.

Output Definition

The outputs should be defined in terms of the following points:

- Type of the output
- Content of the output
- Format of the output
- Location of the output
- Frequency of the output
- Volume of the output
- Sequence of the output

It is not always desirable to print or display data as it is held on a computer. It should be decided as which form of the output is the most suitable.

Input Design

Input design is a part of overall system design. The main objective during the input design is as given below:

- To produce a cost-effective method of input.
- To achieve the highest possible level of accuracy.
- To ensure that the input is acceptable and understood by the user.

Input Stages

The main input stages can be listed as below:

- Data recording

- Data transcription
- Data conversion
- Data verification
- Data control
- Data transmission
- Data validation
- Data correction

Input Types

It is necessary to determine the various types of inputs. Inputs can be categorized as follows:

- External inputs, which are prime inputs for the system.
- Internal inputs, which are user communications with the system.
- Operational, which are computer department's communications to the system?
- Interactive, which are inputs entered during a dialogue.

Input Media

At this stage choice has to be made about the input media. To conclude about the input media consideration has to be given to;

- Type of input
- Flexibility of format
- Speed
- Accuracy
- Verification methods
- Rejection rates
- Ease of correction
- Storage and handling requirements
- Security
- Easy to use
- Portability

Keeping in view the above description of the input types and input media, it can be said that most of the inputs are of the form of internal and interactive. As Input data is to be the directly keyed in by the user, the keyboard can be considered to be the most suitable input device.

Error Avoidance

At this stage care is to be taken to ensure that input data remains accurate form the stage at which it is recorded up to the stage in which the data is accepted by the system. This can be achieved only by means of careful control each time the data is handled.

Error Detection

Even though every effort is made to avoid the occurrence of errors, still a small proportion of errors is always likely to occur, these types of errors can be discovered by using validations to check the input data.

Data Validation

Procedures are designed to detect errors in data at a lower level of detail. Data validations have been included in the system in almost every area where there is a possibility for the user to commit errors. The system will not accept invalid data. Whenever an invalid data is keyed in, the system immediately prompts the user and the user has to again key in the data and the system will accept the data only if the data is correct. Validations have been included where necessary.

The system is designed to be a user friendly one. In other words the system has been designed to communicate effectively with the user. The system has been designed with popup menus.

User Interface Design

It is essential to consult the system users and discuss their needs while designing the user interface:

User Interface Systems Can Be Broadly Classified As:

- User initiated interface the user is in charge, controlling the progress of the user/computer dialogue. In the computer-initiated interface, the computer selects the next stage in the interaction.
- Computer initiated interfaces

In the computer-initiated interfaces the computer guides the progress of the user/computer dialogue. Information is displayed and the user response of the computer takes action or displays further information.

User Initiated Interfaces

User initiated interfaces fall into two approximate classes:

- Command driven interfaces: In this type of interface the user inputs commands or queries which are interpreted by the computer.
- Forms oriented interface: The user calls up an image of the form to his/her screen and fills in the form.

The forms-oriented interface is chosen because it is the best choice.

Computer-Initiated Interfaces

The following computer – initiated interfaces were used:

- The menu system for the user is presented with a list of alternatives and the user chooses one; of alternatives.
- Questions – answer type dialog system where the computer asks question and takes action based on the basis of the users reply.

Right from the start the system is going to be menu driven, the opening menu displays the available options. Choosing one option gives another popup menu with more options. In this way every option

leads the users to data entry form where the user can key in the data.

Error Message Design

The design of error messages is an important part of the user interface design. As user is bound to commit some errors or other while designing a system the system should be designed to be helpful by providing the user with information regarding the error he/she has committed.

This application must be able to produce output at different modules for different inputs.

Performance Requirements

Performance is measured in terms of the output provided by the application. Requirement specification plays an important part in the analysis of a system. Only when the requirement specifications are properly given, it is possible to design a system, which will fit into required environment. It rests largely in the part of the users of the existing system to give the requirement specifications because they are the people who finally use the system. This is because the requirements have to be known during the initial stages so that the system can be designed according to those requirements. It is very difficult to change the system once it has been designed and on the other hand designing a system, which does not cater to the requirements of the user, is of no use.

The requirement specification for any system can be broadly stated as given below:

- The system should be able to interface with the existing system
- The system should be accurate
- The system should be better than the existing system
- The existing system is completely dependent on the user to perform all the duties.

6.2 INTRODUCTION TO TECHNOLOGIES USED

Python

Python is an open source, high-level, interpreted, interactive and object-orientedprogramming language. Python is designed to be highly readable. It supports object-oriented style or technique of programming that encapsulates code within objects. Python is processed at runtimeby the interpreter. It is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers and games.

Python is certainly one of the best languages when working with Machine Learning and AImodels as it has many built-in libraries which can be used directly without much implementation and code.

TensorFlow

TensorFlow is a free and open-source software library for data flow and differentiable programming across a range of tasks. It is a symbolic math library and is also used for machine learning applications such as neural networks. It is used for both research and production at Google.

TensorFlow was developed by the Google Brain team for internal Google use. It was released under the Apache 2.0 open-source license on November 9, 2015.

6.3 LIBRARIES USED

NumPy

NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.

It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary datatypes can be defined using NumPy which allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

Pandas

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. Python was majorly used for data munging and preparation. It had very little contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data load, prepare, manipulate, model, and analyze. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.

Matplotlib

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter Notebook, web application servers, and four graphical user interface toolkits. Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, error charts, scatter plots, etc., with just a few lines of code. For examples, see the sample plots and thumbnail gallery.

For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties etc.. via an object-oriented interface or via a set of functions familiar to MATLAB users.

Scikit – learn

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use.

7. RESULTS AND PERFORMANCE EVALUATION

Figure 7.1 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 7.2 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 7.3 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 7.1: sample dataset used for co2 emission

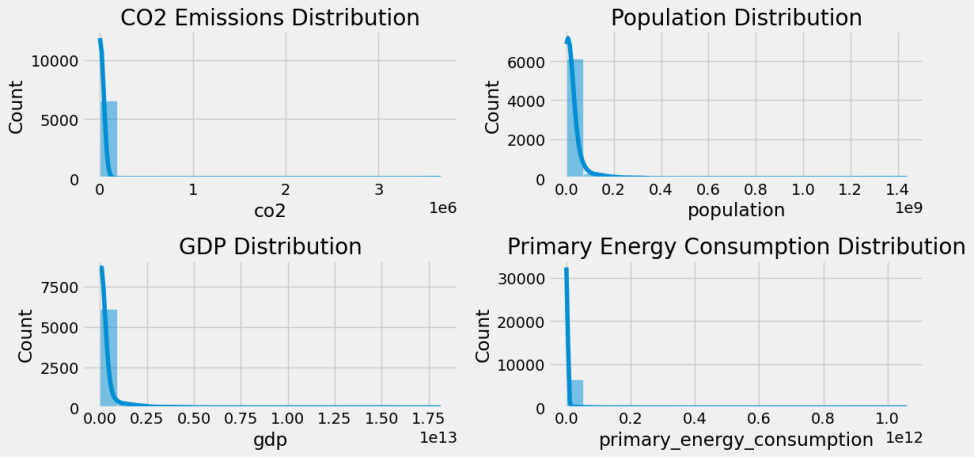

Figure 7.2: This subplot displays the distribution of CO2 emissions

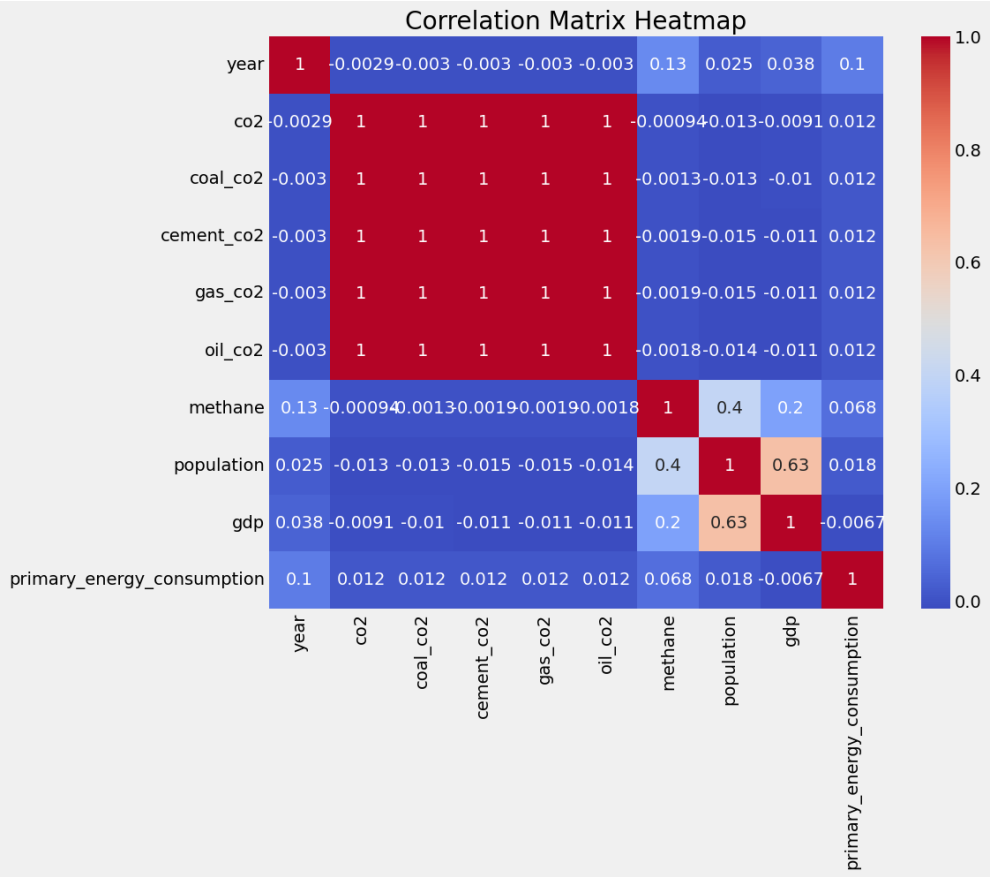

Figure 7.3: Heatmap of correlation of each variable

Figure 7.4 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 7.5 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 7.6 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 7.7 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.

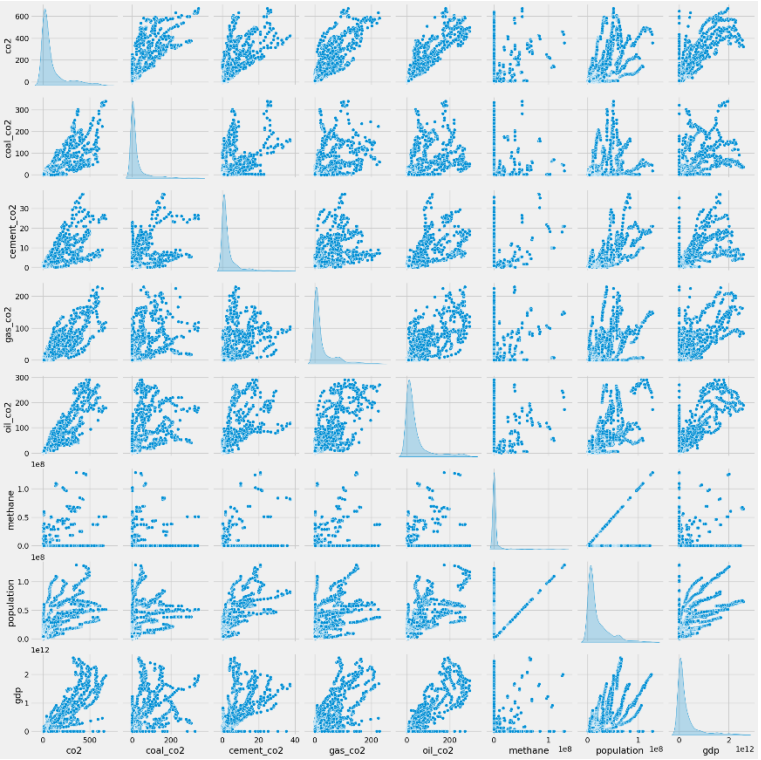

Figure 7.4: pair plot of features

	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 7.5: 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 7.6: Feature of dataset after preprocessing

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

Figure 7.7: target column of a data frame after preprocessing

Figure 7.8 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 7.9 Similar to Figure 7.8, 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.

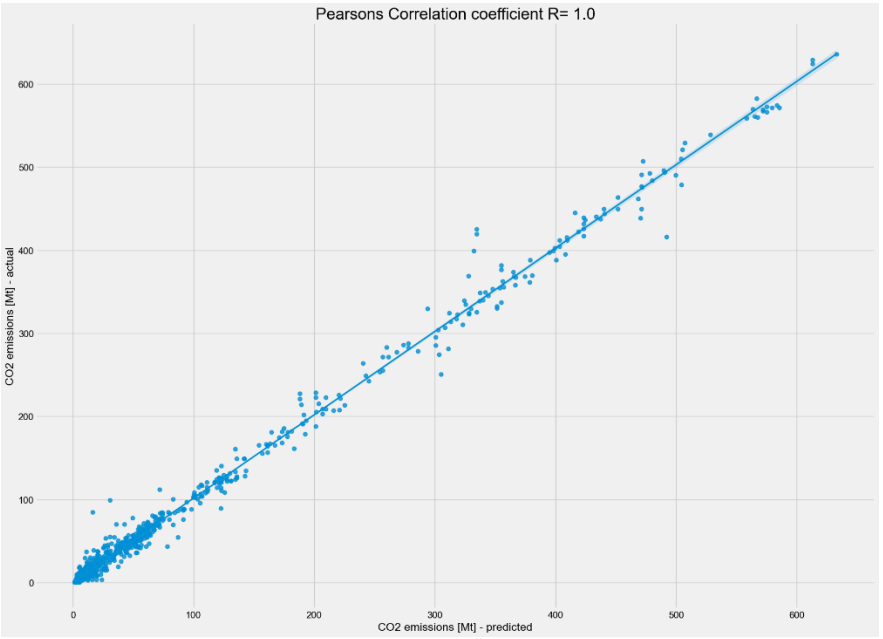

Figure 7.8: prediction results using KNN

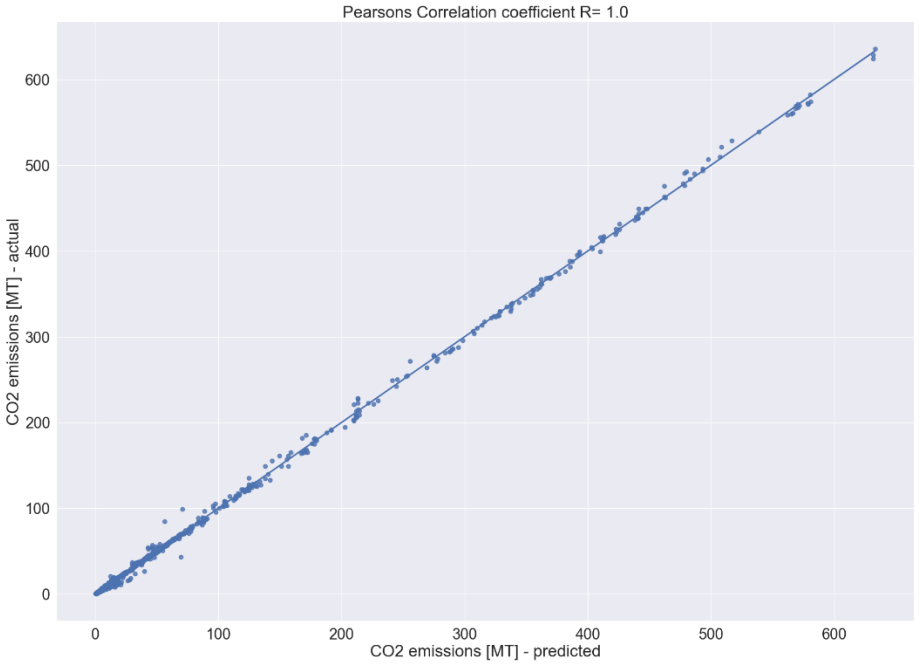

Figure 7.9: prediction results using Random Forest Classifier

Figure 7.10 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 7.11 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 7.12 Similar to Figure 7.11, 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 7.13 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	126.59289	11.25135	0.993483
RF	1.919113	13.166444	3.62856	0.999322

Figure 7.10: Performance metrics of KNN & Random Forest classifier

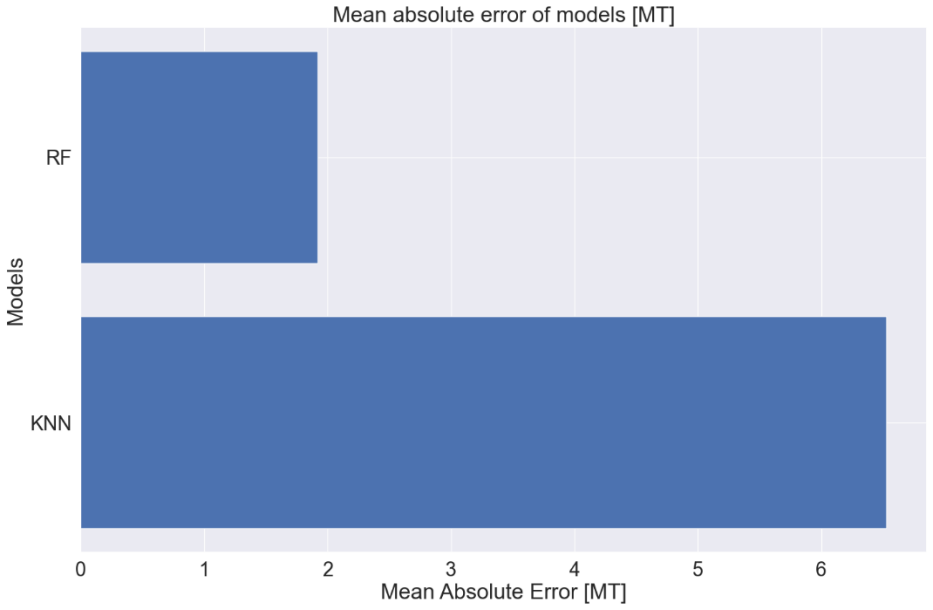

Figure 7.11: Bar plot of Mean absolute error of KNN &Random Forest Classifier

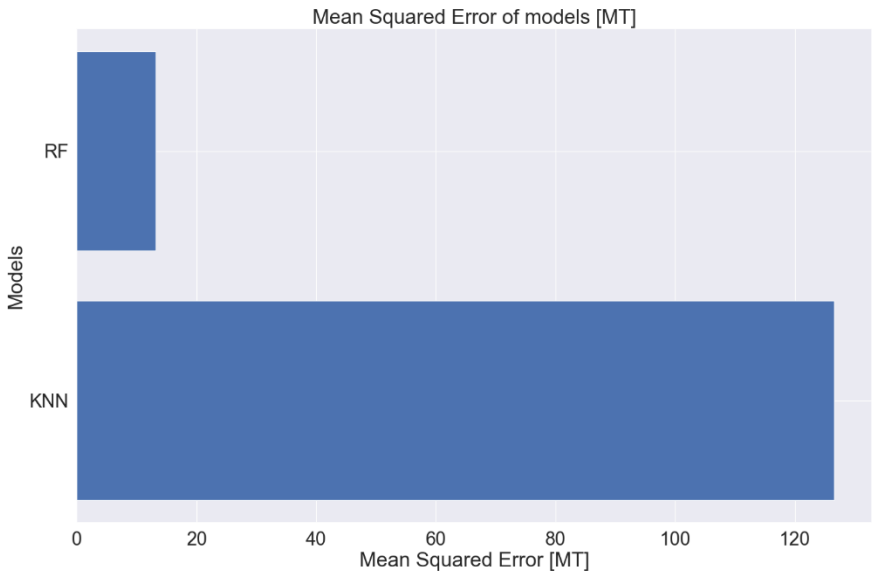

Figure 7.12: Bar plot of Mean Squared error of KNN &Random Forest Classifier

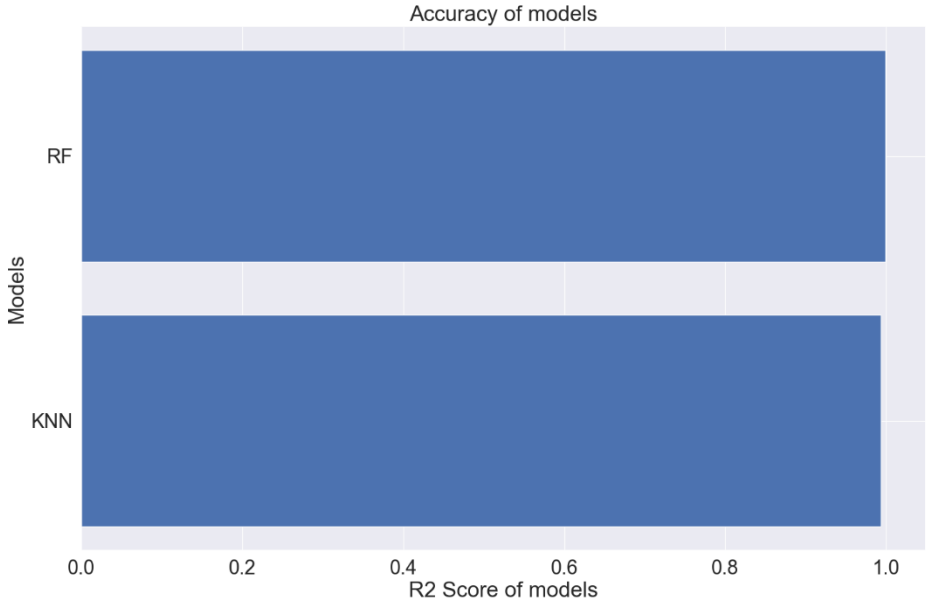

Figure 7.13: Bar plot of R2 Score of KNN &Random Forest Classifier

8. 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.

Our study specifically compared the performance of K-Nearest Neighbors (KNN) and Random Forest algorithms for predicting CO2 emissions. The results showed that the Random Forest algorithm outperformed KNN in terms of accuracy, making it a more reliable choice for such complex datasets. This finding underscores the importance of selecting appropriate machine learning models for achieving accurate predictions.

EDA has proven invaluable in providing a deeper understanding of the data, enabling the identification of influential features and outliers. By combining EDA with the more accurate Random Forest algorithm, we can offer precise 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.

Future Scope

Looking ahead, the future scope of this research is promising. Firstly, the refinement and improvement of machine learning models can enhance prediction accuracy and reliability. Researchers can explore advanced deep learning techniques and ensemble methods to further optimize CO2 emission forecasts. Additionally, incorporating real-time data sources and satellite imagery for monitoring and updating emissions data will make the models more dynamic and responsive to changing environmental conditions. Furthermore, the integration of socio-economic and policy-related variables can provide a more comprehensive understanding of the drivers of CO2 emissions, facilitating the development of targeted and effective climate policies. Collaboration between researchers, governments, and environmental organizations will be crucial in collecting high-quality data and implementing the findings into actionable policies. Ultimately, the future scope of this research lies in its potential to drive meaningful change in our efforts to combat climate change and protect our planet for future generations.

9. REFERENCES

[1]. Intergovernmental 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.

[16]. Fan, J.-L.; Da, Y.-B.; Wan, S.-L.; Zhang, M.; Cao, Z.; Wang, Y.; Zhang, X. Determinants of carbon emissions in 'Belt and Road initiative' countries: A production technology perspective. Appl. Energy 2019, 239, 268–279.

[17]. Net, X. Statement by H.E. Xi Jinping President of the People's Republic of China At the General Debate of the 75th Session of The United Nations General Assembly. Available online: https://baijiahao.baidu.com/s?id=1678546728556033497&wfr=spider&for=pc (accessed on 25 June 2022).

[18]. Xiong, P.P.; Xiao, L.S.; Liu, Y.C.; Yang, Z.; Zhou, Y.F.; Cao, S.R. Forecasting carbon emissions using a multi-variable GM (1,N) model based on linear time-varying parameters. J. Intell. Fuzzy Syst. 2021, 41, 6137–6148.

[19]. Ye, L.; Yang, D.L.; Dang, Y.G.; Wang, J.J. An enhanced multivariable dynamic time-delay discrete grey forecasting model for predicting China's carbon emissions. Energy 2022, 249, 123681.

[20]. Zhang, F.; Deng, X.Z.; Xie, L.; Xu, N. China's energy-related carbon emissions projections for the shared socioeconomic pathways. Resour. Conserv. Recycl. 2021, 168, 105456.

[21]. Li, B.; Han, S.W.; Wang, Y.F.; Li, J.Y.; Wang, Y. Feasibility assessment of the carbon emissions peak in China's construction industry: Factor decomposition and peak forecast. Sci. Total Environ. 2020, 706, 135716.

[22]. Zheng, J.L.; Mi, Z.F.; Coffman, D.; Milcheva, S.; Shan, Y.L.; Guan, D.B.; Wang, S.Y. Regional development and carbon emissions in China. Energy Econ. 2019, 81, 25–36.

[23]. Dong, B.Y.; Ma, X.J.; Zhang, Z.L.; Zhang, H.B.; Chen, R.M.; Song, Y.Q.; Shen, M.C.; Xiang, R.B. Carbon emissions, the industrial structure and economic growth: Evidence from heterogeneous industries in China. Environ. Pollut. 2020, 262, 114322. [PubMed]

[24]. Siqin, Z.Y.; Niu, D.X.; Li, M.Y.; Zhen, H.; Yang, X.L. Carbon dioxide emissions, urbanization level, and industrial structure: Empirical evidence from North China. Environ. Sci. Pollut. Res. 2022, 29, 34528–34545. [PubMed]

[25]. Dong, K.Y.; Sun, R.J.; Hochman, G. Do natural gas and renewable energy consumption lead to less CO2 emission? Empirical evidence from a panel of BRICS countries. Energy 2017, 141, 1466–1478.

[26]. Zheng, H.Y.; Song, M.L.; Shen, Z.Y. The evolution of renewable energy and its impact on carbon reduction in China. Energy 2021, 237, 121639.

[27]. Abbasi, K.R.; Shahbaz, M.; Zhang, J.J.; Irfan, M.; Alvarado, R. Analyze the environmental sustainability factors of China: The role of fossil fuel energy and renewable energy. Renew. Energy 2022, 187, 390–402.

[28]. Sun, W.; Ren, C.M. Short-term prediction of carbon emissions based on the EEMD-PSOBP model. Environ. Sci. Pollut. Res. 2021, 28, 56580–56594.

[29]. Shi, M.S. Forecast of China's carbon emissions under the background of carbon neutrality. Environ. Sci. Pollut. Res. 2022, 29, 43019–43033.

[30]. Zhang, J.X.; Zhang, H.; Wang, R.; Zhang, M.X.; Huang, Y.Z.; Hu, J.H.; Peng, J.Y. Measuring the critical influence factors for predicting carbon dioxide emissions of expanding megacities by XGBoost. Atmosphere 2022, 13, 599.

[31]. Lu, X.Y.; Ota, K.R.; Dong, M.X.; Yu, C.; Jin, H. Predicting transportation carbon emission with urban big data. IEEE Trans. Sustain. Comput. 2017, 2, 333–344.

[32]. Ning, L.Q.; Pei, L.J.; Li, F. Forecast of China's carbon emissions based on ARIMA method. Discret. Dyn. Nat. Soc. 2021, 2021, 1441942.

[33]. Magazzino, C.; Mele, M.; Schneider, N. A machine learning approach on the relationship among solar and wind energy production, coal consumption, GDP, and CO2 emissions. Renew. Energy 2021, 167, 99–115.

[34]. Ahmed, M.; Shuai, C.M.; Ahmed, M. Influencing factors of carbon emissions and their trends in China and India: A machine learning method. Environ. Sci. Pollut. Res. 2022, 29, 48424–48437.

[35]. Ullah, I.; Liu, K.; Yamamoto, T.; Al Mamlook, R.E.; Jamal, A. A comparative performance of machine learning algorithm to predict electric vehicles energy consumption: A path towards sustainability. Energy Environ. 2021.

[36]. Cortes, C.; Vapnik, V. Support-vector networks. Mach. Learn. 1995, 20, 273–297.

[37]. Hamrani, A.; Akbarzadeh, A.; Madramootoo, C.A. Machine learning for predicting greenhouse gas emissions from agricultural soils. Sci. Total Environ. 2020, 741, 140338.

[38]. Breiman, L. Random forests. Mach. Learn. 2001, 45, 5–32.

[39]. Ullah, I.; Liu, K.; Yamamoto, T.; Zahid, M.; Jamal, A. Prediction of electric vehicle charging duration time using ensemble machine learning algorithm and Shapley additive explanations. Int. J. Energy Res. 2022, 46, 15211–15230.

[40]. Ullah, I.; Liu, K.; Yamamoto, T.; Shafiullah, M.; Jamal, A. Grey wolf optimizer-based machine learning algorithm to predict electric vehicle charging duration time. Transp. Lett. Int. J. Transp. Res. 2022.