Counselling Code : VBIT

VIGNANA BHARATHI Institute of Technology®

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

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MINOR PROJECT

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

UNDER THE GUIDANCE OF DR.PRAVEEN TALARI
DESIGNATION:ASSOCIATE PROFESSOR
DEPT.OF CSE

BATCH NO:21PC03
YEAR-IV SECTION-C

K.ASHA- 21P61A05D8
K.VAISHNAVI- 21P61A05D9
M.RAGHU VAMSHI- 21P61A05F5

CONTENTS

ABSTRACT

- The project aims to leverage machine learning models for predicting and forecasting CO2 emissions.
- 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.
- Utilizing Exploratory Data Analysis (EDA), the study first identifies key patterns and trends in CO2 emission data. Then, it compares the performance of the K-Nearest Neighbors (KNN) algorithm with the Random Forest algorithm, demonstrating that Random Forest provides better accuracy and robustness.
- 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

- KNN(K NEAREST NEIGHBOR) ALGORITHM
- RANDOM FOREST ALGORITHM
- EXPLORATORY DATA ANALYSIS
- CO2 EMISSION

INTRODUCTION

- CO2 emissions are a major concern in the context of global warming and climate change, as carbon dioxide is a significant greenhouse gas that traps heat in the atmosphere.
- Given the critical impact of CO2 on the environment, there is a pressing need for accurate prediction and forecasting of future CO2 levels.
- Machine learning plays a pivotal role in this context by providing advanced models that can predict future trends in CO2 emissions with greater accuracy compared to traditional methods.
- Additionally, Exploratory Data Analysis (EDA) is a fundamental step in the modeling process, as it helps in understanding data patterns, identifying key features, and preparing the data effectively for building robust predictive models.

LITERATURE SURVEY

TITLE	AUTHOR	YEAR	PROS	CONS
Influencing factors of carbon emissions in China and India	Ahmed, M.; Shuai, C.M.; Ahmed, M.	2022	Uses machine learning; trend analysis.	Limited to two countries; data-intensive.
Carbon emissions from energy consumption in China: Measurement and driving factors	Ma, X.; Wang, C.; Dong, B.; Gu, G.; Chen, R.; Li, Y.; Zou, H.; Zhang, W.; Li, Q.	2019	Comprehensive analysis of emissions.	Broad scope; relies on data accuracy.
Predicting greenhouse gas emissions from agricultural soils	Hamrani, A.; Akbarzadeh, A.; Madramootoo, C.A.	2020	Relevant to agriculture; ML application.	Data quality-dependent; may miss variables.
Short-term prediction of carbon emissions based on the EEMD-PSOBP model	Sun, W.; Ren, C.M.	2021	Innovative modeling approach.	Short-term focus; may lack long-term
Measuring critical influence factors for predicting CO2 emissions in megacities	Zhang, J.X.; Zhang, H.; Wang, R.; Zhang, M.X.; Huang, Y.Z.; Hu, J.H.; Peng, J.Y.	2022	Uses advanced ML (XGBoost) for prediction.	Focused on megacities; may not generalize to smaller cities

PROBLEM STATEMENT

- The increasing levels of CO2 emissions pose significant threats to global climate stability and public health. Despite the availability of historical emissions data, there is a lack of effective predictive models that can provide accurate forecasts of future CO2 emissions.
- This project aims to address this gap by developing and evaluating machine learning models capable of predicting CO2 emissions based on various influencing factors, such as industrial activity, energy consumption, and policy changes.
- By integrating exploratory data analysis (EDA) to preprocess and visualize the data, we will enhance the predictive accuracy and interpretability of the models.

OBJECTIVE

- The objective of this project is to analyze CO2 emission data using Exploratory Data Analysis (EDA) to uncover underlying patterns and insights crucial for effective model development.
- The project involves building and evaluating various machine learning models for predicting CO2 emissions. Additionally, it focuses on comparing the performance of the existing K-Nearest Neighbors (KNN) algorithm with the proposed Random Forest algorithm to determine which approach yields more accurate and reliable predictions.

SYSTEM REQUIREMENTS

Software Requirements

- Coding Language: Python
- IDE: Visual Studio (or) Python IDLE 3.7 version (or) Anaconda 3.7 (or) Jupiter (or) Google colab
- Operating System: Windows 7/8/10, Linux

Hardware Requirements

- Processor : Minimum intel i3
- Ram : Minimum 4 GB
- Hard disk : Minimum 250GB

PROPOSED METHODOLOGY

Step 1. Exploratory Data Analysis (EDA) :

- Data Collection and Data Inspection
- Data Visualization
- Outlier Detection

Step 2. Data Preprocessing :

- Feature Selection
- Handling Missing
- Data Normalization/Scaling Data

Step 3. Existing KNN Model :

- Use KNN Algorithm
- Hyperparameter Tuning and Model Training

Step 4. Perform Proposed RANDOM FOREST Model :

- Use Random Forest Algorithm and Feature Engineering
- Hyperparameter Tuning and Model Training

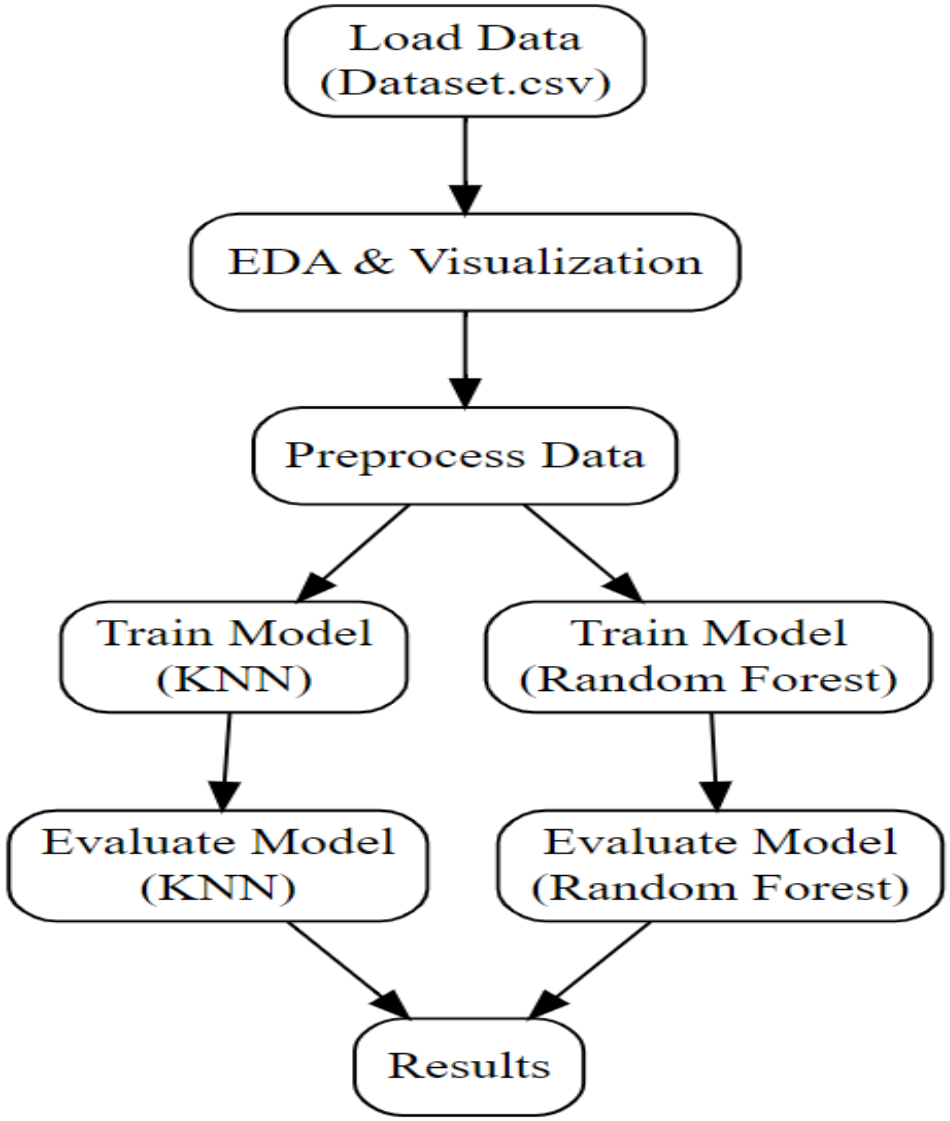
Step 5. Prediction :

- Predict CO2 Emissions.

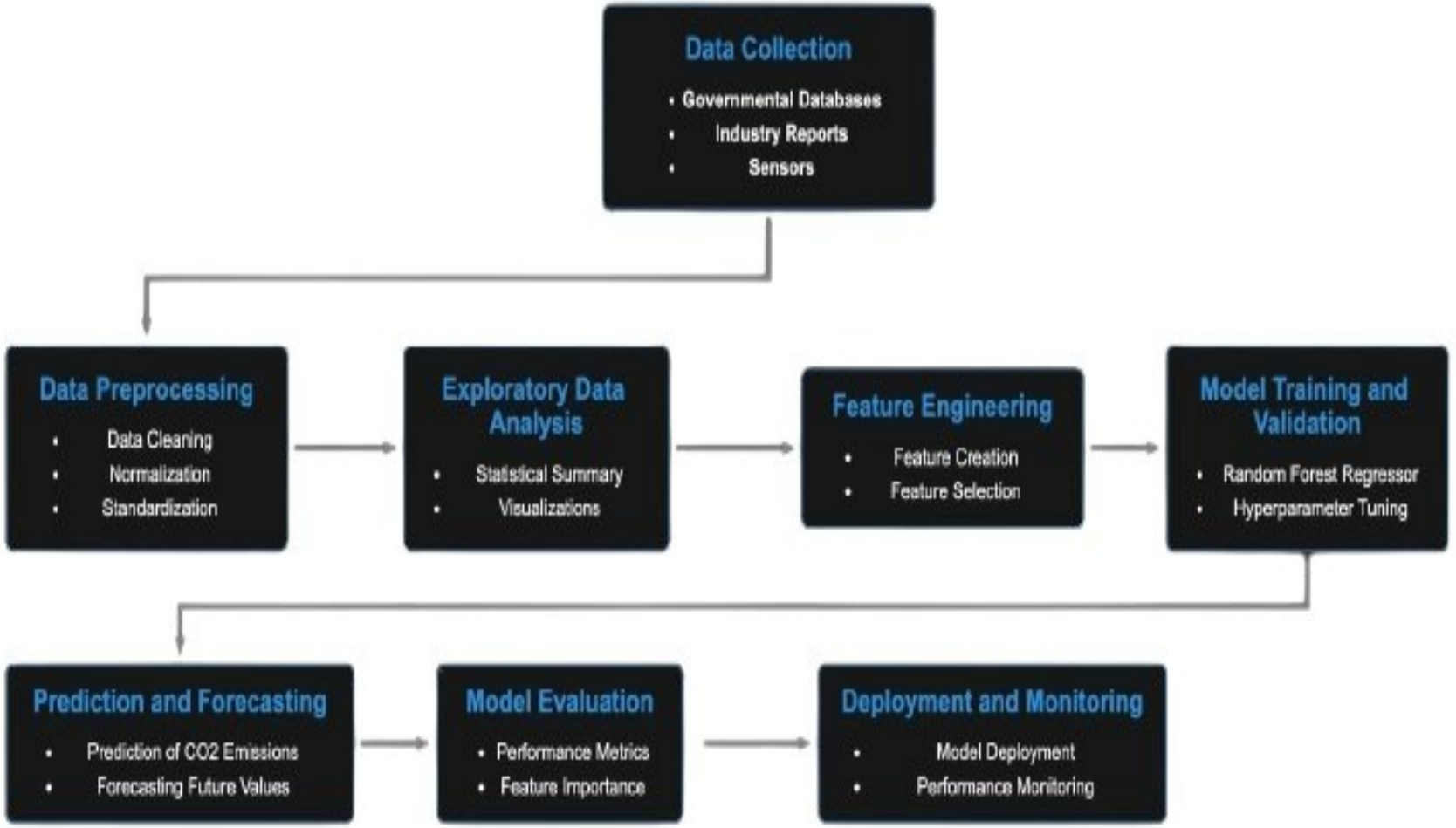
Step 6. Performance Estimation :

- **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.

Block diagram of proposed system



SYSTEM ARCHITECTURE



MODULES USED IN THE PROJECT

Data Handling and Manipulation

- Pandas: (For data manipulation and analysis) Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures.
- NumPy: (For numerical operations) NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.

Data Visualization

- Matplotlib: (For creating static plots) Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms.
- Seaborn: For statistical data visualization.

Exploratory Data Analysis (EDA)

- Scikit-learn: (For preprocessing and feature selection) Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python.
- Statsmodels : For statistical tests and models.

Machine Learning

- Scikit-learn: For building and evaluating machine learning models.
- TensorFlow : (For deep learning models) 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.

IMPLEMENTATION

Step 1. Exploratory Data Analysis (EDA) :

Data Collection:

- **Dataset : (Dataset.csv)** Global CO2 Emissions Dataset

- This dataset contains annual emissions and relevant socioeconomic indicators for various countries, enabling analysis and forecasting of CO2 emissions.
- **Rows**: Each row represents data for a specific year and country.
- **Columns**:
  - **Country**: The name of the country.
  - **Year**: The year the data corresponds to.
  - **CO2**: Total CO2 emissions (in metric tons).
  - **Coal CO2**: CO2 emissions from coal combustion (in metric tons).
  - **Cement CO2**: CO2 emissions from cement production (in metric tons).
  - **Gas CO2**: CO2 emissions from natural gas combustion (in metric tons).
  - **Oil CO2**: CO2 emissions from oil combustion (in metric tons).
  - **Methane**: Methane emissions (in metric tons).
  - **Population**: Population of the country in the corresponding year.
  - **GDP**: Gross Domestic Product (in current US dollars).
  - **Primary Energy Consumption**: Total primary energy consumption (in terajoules).

- **Data Inspection** : Examined the dataset's structure, including the number of rows and columns, data types, and any missing values.
- **Data Visualization** : Created 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** : Identified and handled outliers in the dataset, as extreme values can adversely affect machine learning models.

Step 2. Data Preprocessing :

- **Feature Selection** : Chose 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**: Addressed some missing values in the dataset through techniques like imputation or removal of rows/columns with missing data.
- **Normalization/Scaling** : Scaled numerical features to ensure they have similar scales, which can improve the performance of some machine learning algorithms.
- **Data Splitting** : Divided the dataset into training and testing sets for model development and evaluation.

Step 3. Performed KNN Algorithm :

- **Hyperparameter Tuning** : Used 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. Performed RANDOM FOREST Model :

- **Feature Engineering** : Created new features or combinations of features that may improve the prediction of CO2 emissions.
- **Hyperparameter Tuning** : Similar to the existing KNN model, optimized the hyperparameters for the proposed RANDOM FOREST model.
- **Model Training** : Trained the proposed RF model using the training data.

Step 5. Prediction :

- **Predicted CO2 Emissions** : Used both the existing and proposed models to predict CO2 emissions for the testing dataset.

Step 6. Performance Estimation (For Both Models):

- Calculated the MAE to quantify the average absolute difference between predicted and actual CO2 emissions.
- Computed the MSE to measure the average squared difference between predicted and actual emissions.
- Calculated the RMSE by taking the square root of MSE, providing a measure in the original unit (e.g., Mt).
- Determined the R2 score to evaluate how well the model explains the variance in CO2 emissions.
- **Compared** the performance metrics between the existing and proposed models to assess whether the proposed model provides better predictions.

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

	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 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.

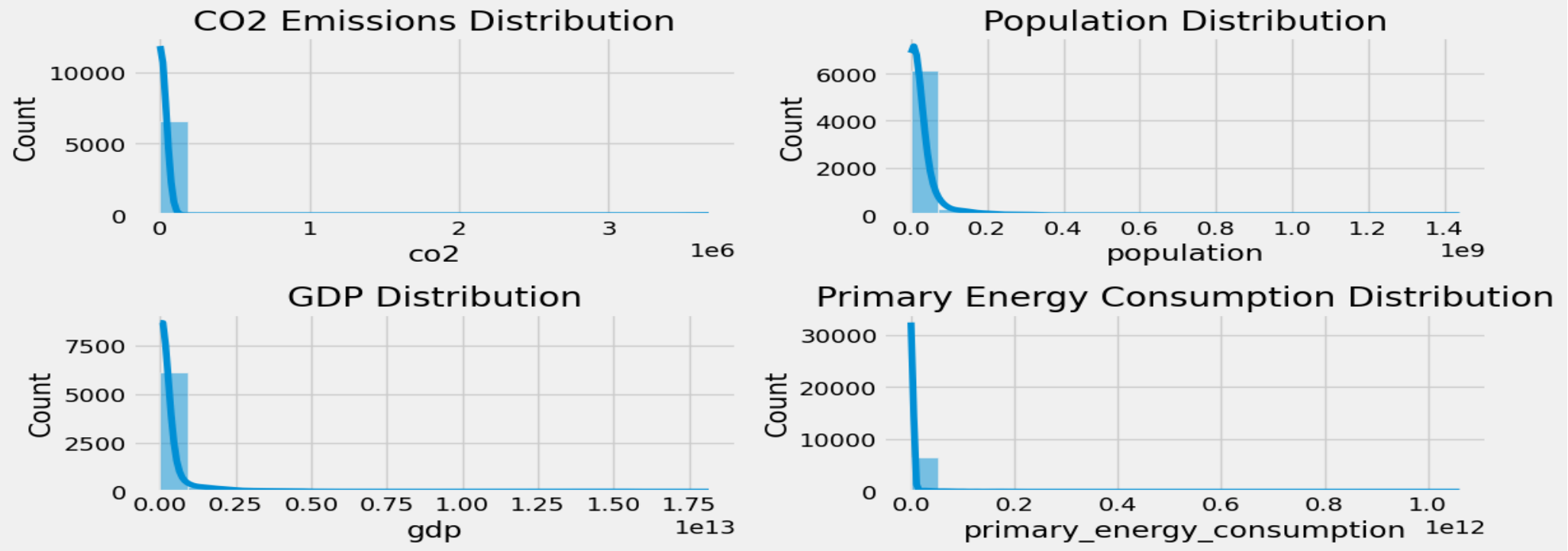


Figure 2: This subplot displays the distribution of CO2 emissions

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.

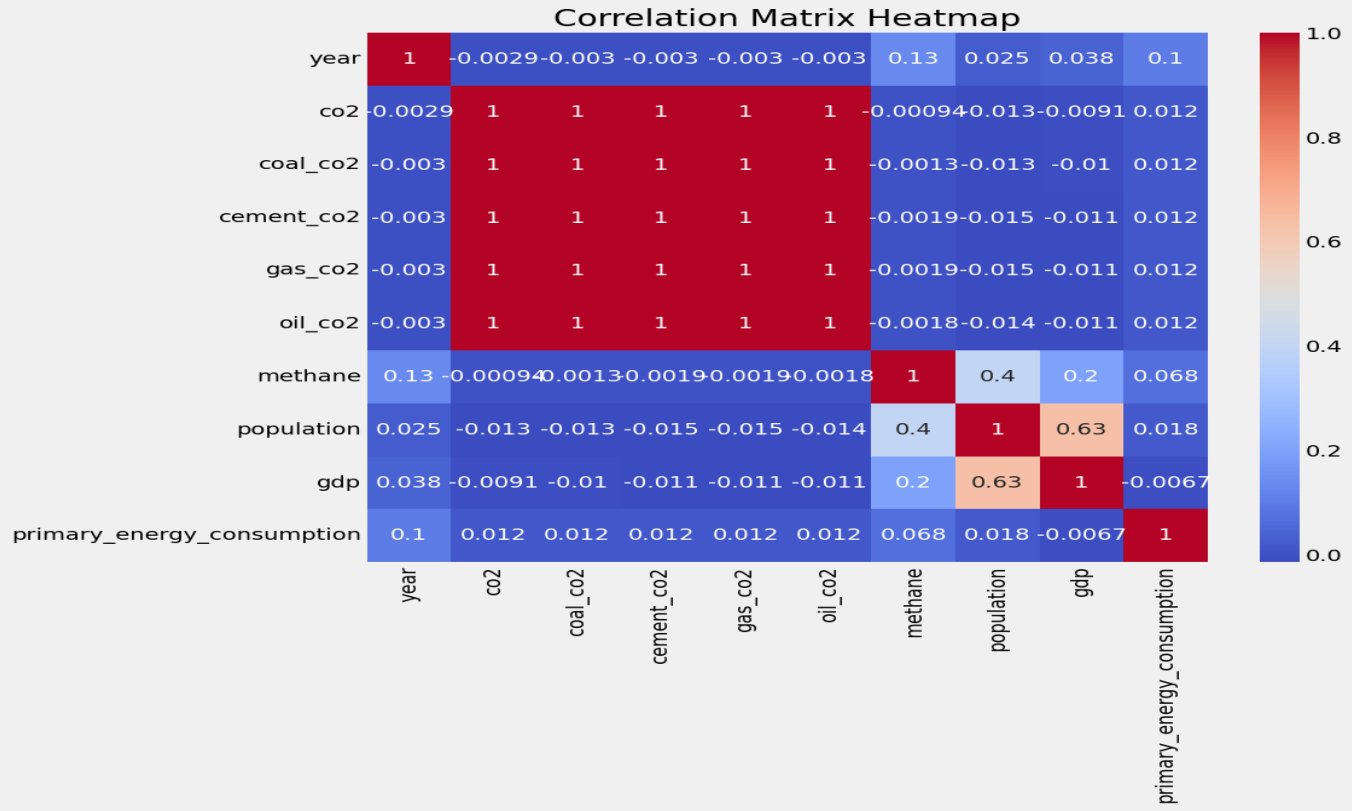


Figure 3: Heatmap of correlation of each variable

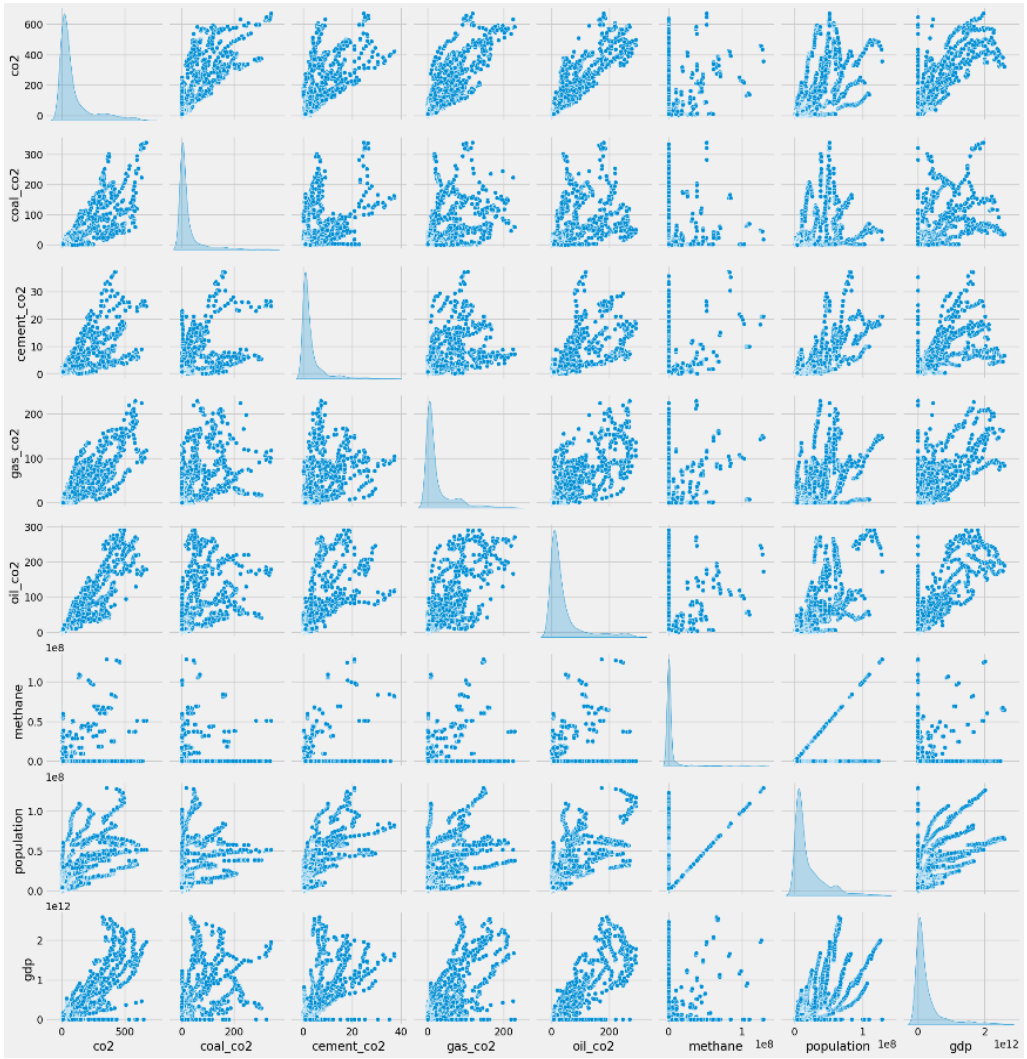


Figure 4: pair plot of features

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

	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

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.

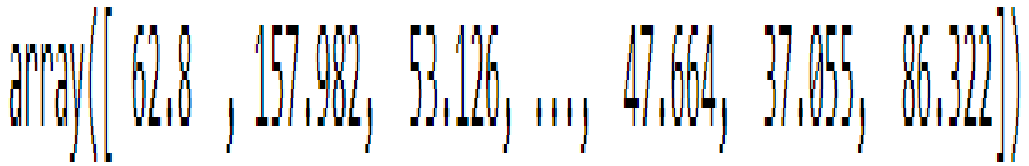


Figure 7: target column of a data frame after preprocessing

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.

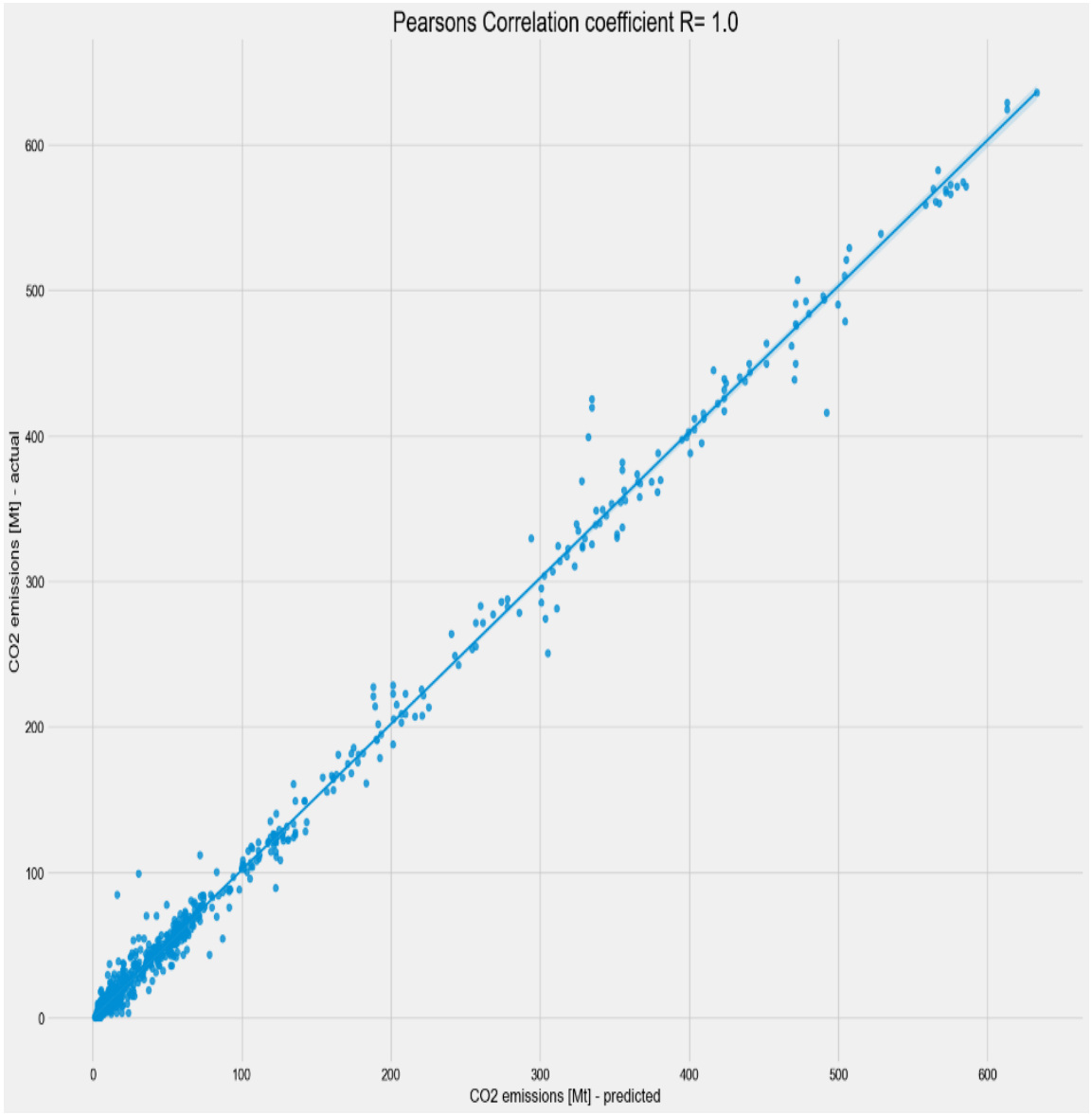


Figure 8: prediction results using KNN

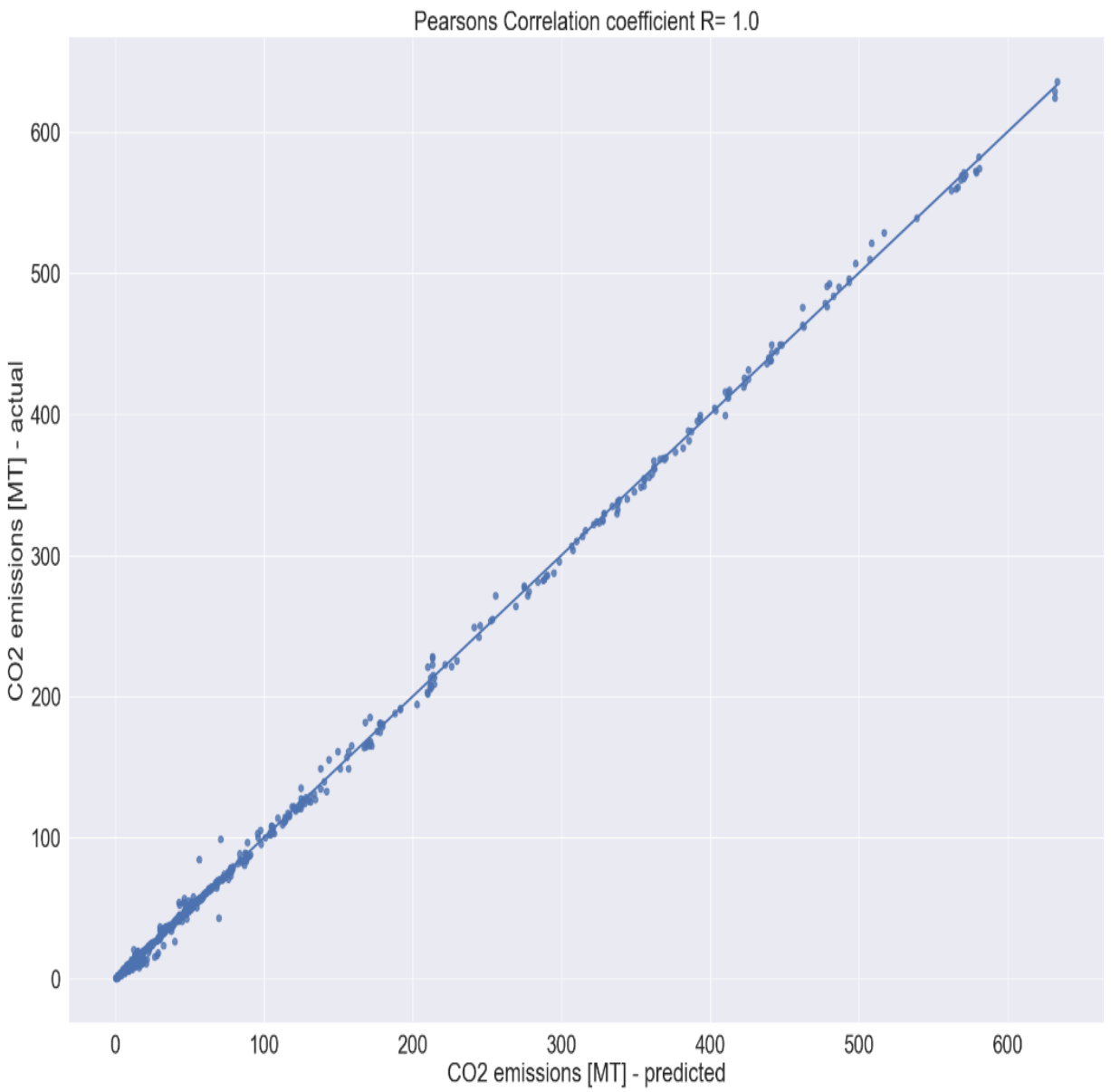


Figure 9: prediction results using Random Forest Classifier

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

	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

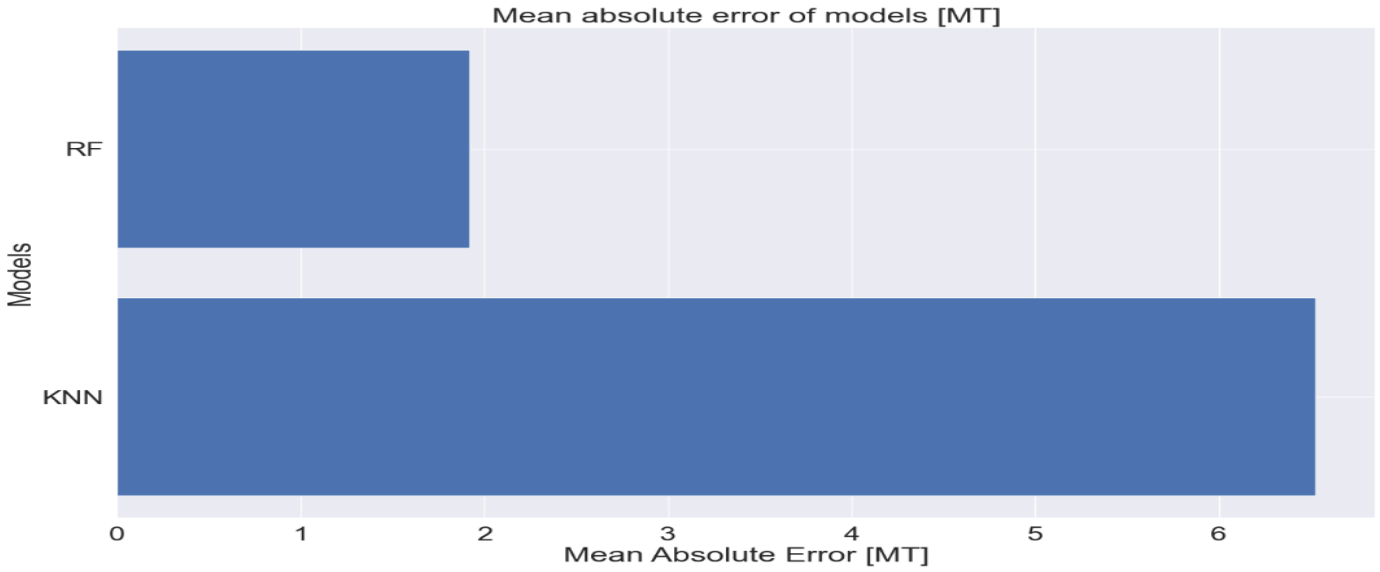


Figure 11 displays a bar plot comparing the Mean Absolute Error (MAE) of the KNN and Random Forest Classifier models. It provides a visual representation of how well each model predicts CO2 emissions.

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

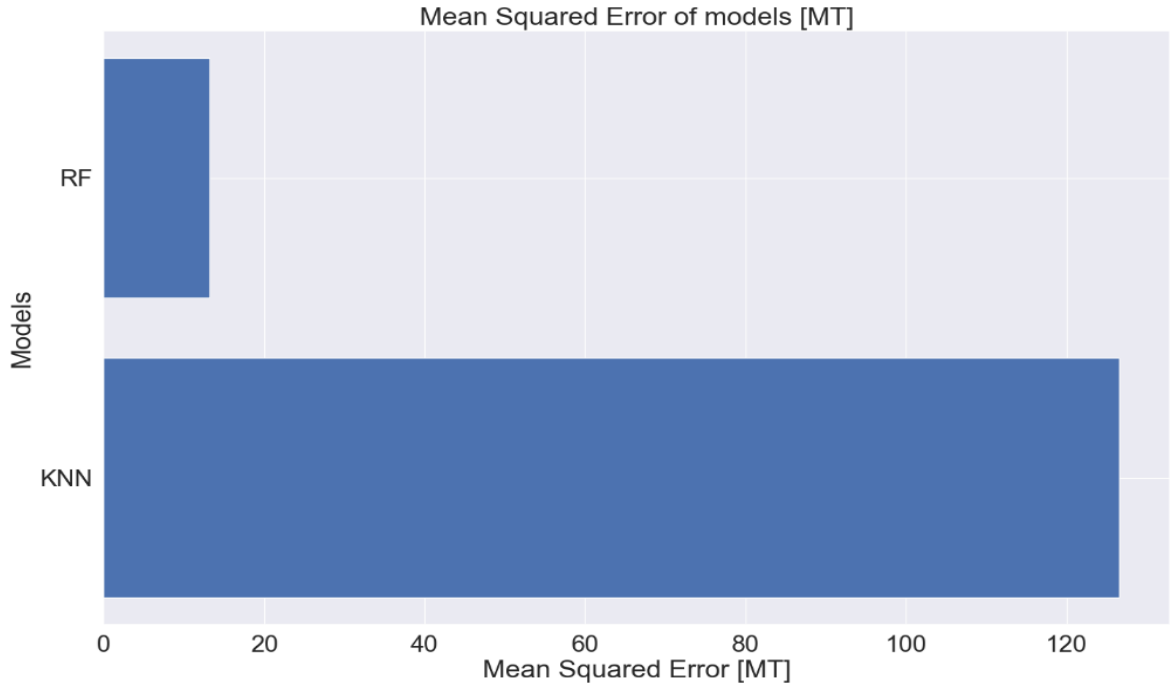


Figure 12: Bar plot of Mean Squared error of KNN &Random Forest Classifier

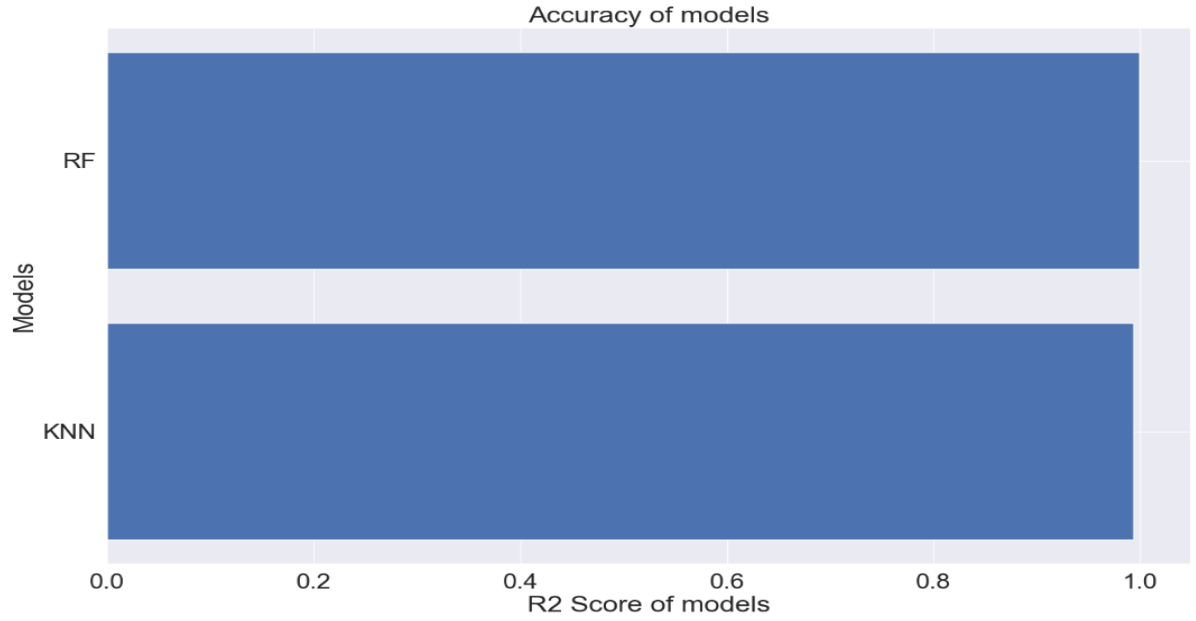


Figure 13: Bar plot of R2 Score of KNN &Random Forest Classifier

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.

REFERENCES

1. 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.

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

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

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

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

9. 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.

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

11. 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.

12. 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.

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

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. 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.

THANK YOU