MACHINE LEARNING TUTORIAL REPORT: PREDICTING CARBON EMISSIONS USING RANDOM FOREST REGRESSION

Name: Mukesh Saragadam

Student ID: 23080812

Contents

1.Introduction

The global increase in carbon emissions, driven by escalating energy consumption, remains a critical challenge affecting climate change and sustainable development. Accurate prediction of carbon emissions based on energy usage patterns and related factors is essential for policymakers and environmental agencies to devise informed strategies for emission reduction and sustainable energy management (IEA, 2022). Machine learning techniques have increasingly been adopted due to their capability in modeling complex, non-linear relationships within large and multifaceted datasets, demonstrating superior predictive performance compared to traditional statistical methods (James et al., 2021).

Among machine learning techniques, **Random Forest Regression (RFR)**, introduced by Breiman (2001), has gained prominence because of its robust handling of non-linear relationships, minimal assumptions about data distributions, and inherent interpretability through feature importance metrics. Previous studies have shown that Random Forest methods effectively manage multicollinearity among predictors and handle noisy and heterogeneous data commonly encountered in environmental and energy datasets (Wang et al., 2018).

In this tutorial, we apply Random Forest Regression using Python's scikit-learn library (Pedregosa et al., 2011) to predict carbon emissions from global energy consumption data. The dataset used in this analysis comprises various features such as total energy consumption, per capita energy use, renewable energy share, fossil fuel dependency, industrial and household energy use, and energy pricing indices. By following a structured analytical pipeline—including data preprocessing, exploratory data analysis, feature scaling, model training, validation, and feature importance evaluation—this tutorial aims to clearly demonstrate how machine learning methodologies can be practically applied to environmental analytics.

The primary objectives of this tutorial are:

- To demonstrate the systematic application of Random Forest Regression for predicting numerical environmental outcomes.
- To illustrate best practices in exploratory data analysis (EDA), preprocessing, model validation, and result interpretation.
- To critically assess model performance, including an examination of possible overfitting and the interpretation of feature significance.

This structured approach will not only offer insights into applying machine learning techniques effectively but also underscore their potential and limitations, providing practical knowledge for stakeholders and researchers in environmental data science.

Step-by-Step Tutorial

This tutorial follows a structured analytical pipeline, including:

1. Data Loading and Initial Exploration
2. Exploratory Data Analysis (EDA)
3. Data Preparation & Feature Engineering
4. Feature Scaling
5. Model Training and Evaluation
6. Feature Importance Analysis
7. Overfitting Analysis & Recommendations

2.Data Understanding and Preliminary Exploration

A critical first step in any machine learning project is to understand the dataset before modeling begins. In this case, the dataset provided consists of **10,000 rows** and **10 columns**, which represent various global energy indicators along with the target variable — **Carbon Emissions (Million Tons)**.

Structure of the Dataset

The first few rows of the dataset show country-wise data, including the year, energy consumption figures, and carbon emission statistics. The columns represent the following:

- **Country**: Name of the country (categorical)
- **Year**: The calendar year of the observation (integer)
- **Total Energy Consumption (TWh)**
- **Per Capita Energy Use (kWh)**
- **Renewable Energy Share (%)**
- **Fossil Fuel Dependency (%)**
- **Industrial Energy Use (%)**
- **Household Energy Use (%)**
- **Carbon Emissions (Million Tons)** (Target Variable)
- **Energy Price Index (USD/kWh)**

This combination of features captures both macro (national energy consumption) and micro-level (household use, per capita consumption) variables, which can help in building a strong predictive model.

```
First 5 rows of the dataset:
================================================================================
    Country      Year    Total Energy Consumption (TWh)    Per Capita Energy Use (kWh)    Renewable Energy Share (%)    Fossil Fuel Dependency (%)
ustrial Energy Use (%)    Household Energy Use (%)    Carbon Emissions (Million Tons)    Energy Price Index (USD/kWh)
--  ---------  ------  --------------------------------  -----------------------------  -----------------------------  -----------------------------  --
---------------------  -------------------------  -------------------------------  -----------------------------
 0  Canada       2018                           9525.38                        42301.4                           13.7                           70.47
45.18                         19.96                             3766.11                           0.12
 1  Germany      2020                           7922.08                        36601.4                          33.63                           41.95
34.32                         22.27                             2713.12                           0.08
 2  Russia       2002                           6630.01                        41670.2                          10.82                           39.32
53.66                         26.44                              885.98                           0.26
 3  Brazil       2010                           8580.19                        10969.6                          73.24                           16.71
30.55                         27.6                              1144.11                           0.47
 4  Canada       2006                            848.88                        32190.8                           73.6                           74.86
42.39                         23.43                              842.39                           0.48

Dataset Overview:
--------------------------------------------------
Number of Rows: 10000
Number of Columns: 10

Columns and their data types:
Country: object
Year: int64
Total Energy Consumption (TWh): float64
Per Capita Energy Use (kWh): float64
Renewable Energy Share (%): float64
Fossil Fuel Dependency (%): float64
Industrial Energy Use (%): float64
Household Energy Use (%): float64
Carbon Emissions (Million Tons): float64
Energy Price Index (USD/kWh): float64

Missing Values:
--------------------------------------------------
```

Figure 1First 5 rows of the dataset

2.1 Basic Information and Data Types

From the dataset info , it's clear that:

- All features are **numerical** (either float or int), except for **Country**, which is **categorical**.
- This simplifies preprocessing steps, especially for regression models that expect numerical inputs.
- No missing values were found, which allows us to proceed without imputation (Figure 1, bottom).

This clean structure is advantageous, as it saves time in data cleaning and allows us to focus on model design and interpretation.

```
Basic Dataset Information:
--------------------------------------------------
Number of Rows: 10000
Number of Columns: 10

Columns in dataset:
- Country
- Year
- Total Energy Consumption (TWh)
- Per Capita Energy Use (kWh)
- Renewable Energy Share (%)
- Fossil Fuel Dependency (%)
- Industrial Energy Use (%)
- Household Energy Use (%)
- Carbon Emissions (Million Tons)
- Energy Price Index (USD/kWh)
```

Figure 2 Basic dataset information

2.2 Descriptive Statistics of Numeric Features

A statistical summary shows:

- The **average carbon emissions** across all countries and years is approximately **2536.15 million tons**.
- The **range** is quite large, from **50.64 to 4999.34**, indicating significant variation across observations.
- Features like **Total Energy Consumption** and **Per Capita Energy Use** also show high standard deviations, suggesting diversity in country-wise energy profiles.

The widespread and high variability in features are positive indicators for model learning. Models like Random Forests often perform better when trained on diverse datasets that reflect the range of real-world situations.

Statistical Summary of Numeric Features:

	Year	Total Energy Consumption (TWh)	Per Capita Energy Use (kWh)	Renewable Energy Share (%)	Fossil Fuel Dependency (%)	ndustrial Energy Use (%)	Household Energy Use (%)	Carbon Emissions (Million Tons)	Energy Price Index (USD/kWh)
count	10000	10000	10000	10000	10000	10000	10000	10000	10000
mean	2012.15	5142.56	25040	47.32	44.93	40.06	25.04	2536.15	0.27
std	7.16	2848.75	14205.7	24.6	20.2	11.54	8.6	1424.11	0.13
min	2000	100.48	500.27	5	10.01	20	10	50.64	0.05
25%	2006	2713.88	12683.2	26.11	27.34	30.22	17.61	1293.33	0.16
50%	2012	5190.85	25098.8	47.15	45.11	39.98	25.09	2568.02	0.27
75%	2018	7579.98	37113.3	68.68	62.43	50.15	32.57	3766.18	0.39
max	2024	9999.26	49989.6	90	80	60	40	4999.34	0.5

Figure 3 Statistical summary of numerical features

2.3 Correlation Matrix

From the correlation heatmap, we observe that:

- No strong linear correlation exists between any of the features and **Carbon Emissions**.
- The highest (though still weak) correlation is with **Industrial Energy Use (%)**, followed by **Total Energy Consumption (TWh)** and **Year**.

This weak correlation supports the choice of using a **non-linear regression method**, like Random Forest, rather than a linear model. Random Forest can capture complex patterns even when linear relationships are weak or absent.

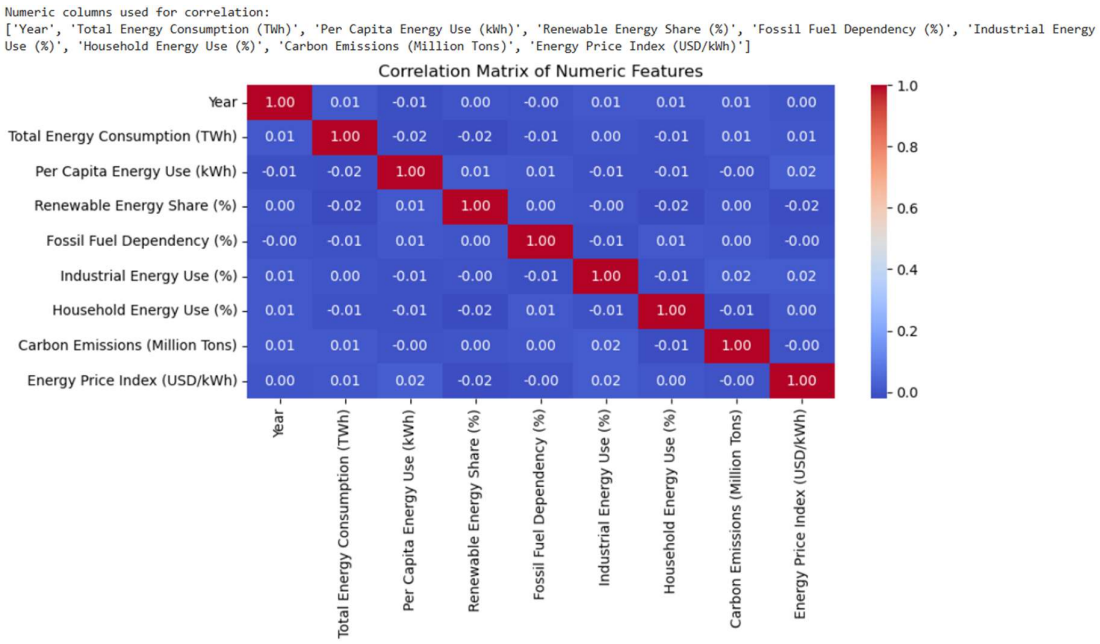

Figure 4 Correlation matrix of numeric features

2.4 Distribution of Carbon Emissions

The histogram presents a relatively **uniform distribution** of the carbon emissions variable. Each bin contains approximately **400–500 entries**, and no significant skew or outlier clusters are observed. This balanced distribution ensures that the model is trained across all levels of emission intensity, reducing the likelihood of bias toward extreme values.

Initial Key Insights

From Figure 5, we derived these early insights:

- **Average Carbon Emissions**: **2536.15 Million Tons**
- **Maximum Carbon Emissions**: **4999.34 Million Tons**
- **Minimum Carbon Emissions**: **50.64 Million Tons**

```
Key Insights:
--------------------------------------------------
1. Average Carbon Emissions: 2536.15 Million Tons
2. Maximum Carbon Emissions: 4999.34 Million Tons
3. Minimum Carbon Emissions: 50.64 Million Tons

Top 3 Features Correlated with Carbon Emissions:
Industrial Energy Use (%)         0.017964
Total Energy Consumption (TWh)    0.013643
Year                              0.013437
```

Figure 5 Key insights including average, maximum, and minimum carbon emissions

Top 3 features most correlated with emissions:

1. **Industrial Energy Use (%)**: 0.01796
2. **Total Energy Consumption (TWh)**: 0.01364
3. **Year**: 0.01344

While these values might appear small, they act as weak indicators that may still play a significant role when combined through ensemble methods.

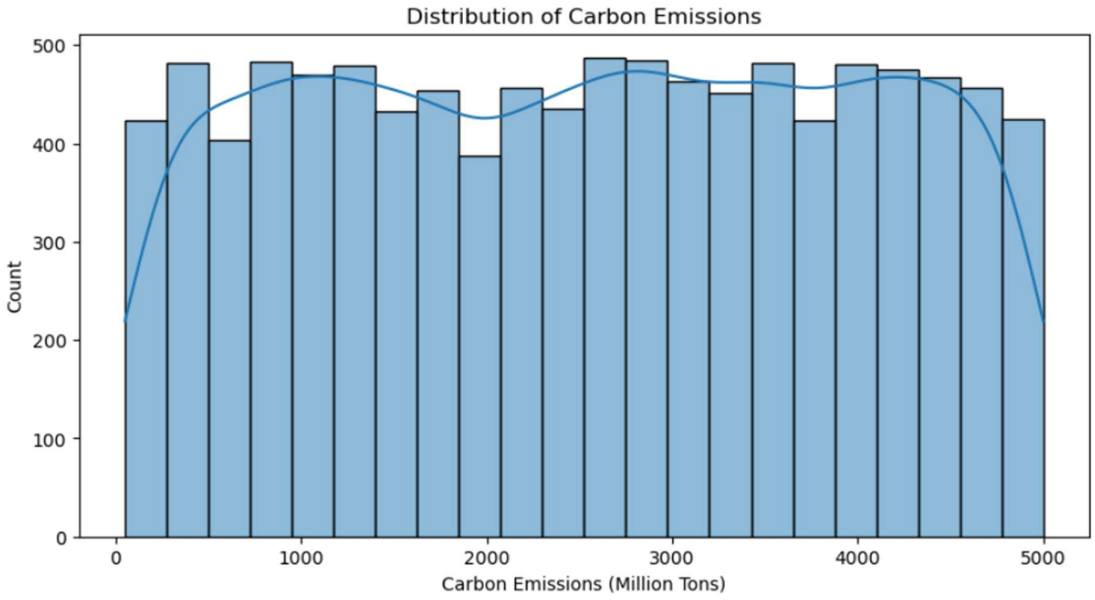

Figure 6 Distribution of carbon emissions

3.Model Development Workflow and Analysis

After performing initial exploration, the next logical steps involve feature selection, data preparation, and model training. This section provides a detailed walkthrough of each of these steps with exact numbers from the analysis.

Step 1: Selecting Features

From the available columns in the dataset, the following seven numerical features were selected to predict **Carbon Emissions (Million Tons)** (Figure 1):

1. **Total Energy Consumption (TWh)**
2. **Per Capita Energy Use (kWh)**
3. **Renewable Energy Share (%)**
4. **Fossil Fuel Dependency (%)**
5. **Industrial Energy Use (%)**
6. **Household Energy Use (%)**
7. **Energy Price Index (USD/kWh)**

These features were chosen based on their logical relevance to emissions behavior and their statistical presence in earlier correlation analyses. Categorical variables like `Country` were excluded since the current model is purely numeric and not designed for encoding geographic information.

```
Step 1: Selecting Features
--------------------------------------------
Selected features:
- Total Energy Consumption (TWh)
- Per Capita Energy Use (kWh)
- Renewable Energy Share (%)
- Fossil Fuel Dependency (%)
- Industrial Energy Use (%)
- Household Energy Use (%)
- Energy Price Index (USD/kWh)
```

Figure 7 Feature selection list for model input

Step 2: Checking for Missing Values

Before training, it's important to verify data completeness. As shown in Figure 8, **none of the selected features contained missing values**, with all 10,000 records showing a value for each of the seven inputs. This means no imputation or row removal was necessary, simplifying preprocessing and preserving the entire dataset for learning.

```
Step 2: Checking for Missing Values
------------------------------------------------
Missing values in each feature:
Total Energy Consumption (TWh)    0
Per Capita Energy Use (kWh)       0
Renewable Energy Share (%)        0
Fossil Fuel Dependency (%)        0
Industrial Energy Use (%)         0
Household Energy Use (%)          0
Energy Price Index (USD/kWh)      0
dtype: int64
```

Figure 8 Missing values report showing no null entries in features

Step 3: Splitting the Dataset

To train and evaluate the model reliably, the dataset was split using an 80/20 ratio (Figure 9):

- **Training set size**: **8,000 samples**
- **Testing set size**: **2,000 samples**

This ensures the model has enough data to learn from while still reserving a representative set for unbiased performance evaluation.

```
Step 3: Splitting Data
----------------------------------------
Training set size: 8000 samples
Testing set size: 2000 samples
```

Figure 9 Data split into training (8000) and testing (2000) samples

Step 4: Feature Scaling

Standardization was performed using **StandardScaler**, which rescales the data such that:

- Mean of scaled features = **approximately 0.000**
- Standard deviation of features = **1.000** (Figure 4)

Standardization is crucial for distance-based algorithms or tree ensembles that perform better when inputs are normalized. Although Random Forests are scale-invariant, consistent scaling ensures fairness if ensemble methods are later combined with distance-based metrics or hyperparameter tuning.

```
Step 4: Feature Scaling
--------------------------------------------------
Features have been scaled using StandardScaler
Mean of scaled features should be close to 0: -0.000
Standard deviation of scaled features should be close to 1: 1.000
```

Figure 10 Feature scaling using StandardScaler showing mean ≈ 0 and std ≈ 1

Step 5: Model Training and Performance Evaluation

A **Random Forest Regressor** with 100 trees was trained on the standardized data. The performance of the model was then evaluated on both training and testing sets (Figure 11).

Interpretation:

- The **training RMSE** of **542.59** and high **R^2 = 0.85** suggest the model fits the training data well.
- However, the **testing RMSE** jumps to **1469.13**, and **R^2 drops to -0.06**, indicating poor generalization and **significant overfitting**.

The scatter plots clearly show that the model performs well on training data (points clustered around the red line) but poorly on test data (more scattered pattern). The error distribution plot further emphasizes this gap.

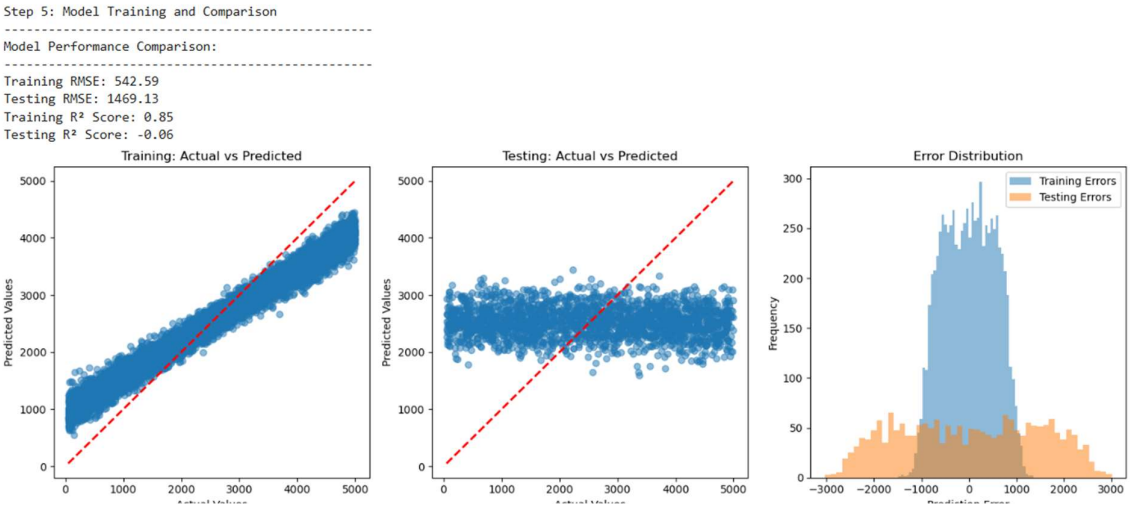

Figure 11 Model performance comparison with RMSE and R^2 scores and plots for training vs testing predictions

Step 6: Error Heatmaps and Overfitting Analysis

Figure 12 presents two heatmaps that compare **actual vs. predicted values** in 10 bins, for both training and testing data.

Training Set Errors:

- Mean Absolute Error: **462.40**
- Mean Squared Error: **294,404.84**
- RMSE: **542.59**

Testing Set Errors:

- Mean Absolute Error: **1274.78**
- Mean Squared Error: **2,158,329.02**
- RMSE: **1469.13**

Overfitting Indicators:

- **RMSE Difference**: 1469.13 − 542.59 = **926.54**
- **R^2 Difference**: 0.85 − (−0.06) = **0.91**

Since the R^2 difference exceeds **0.1**, the model is highly overfitted. This means the model memorizes the training set but fails to generalize to new data—a common issue when model complexity is not well-regularized.

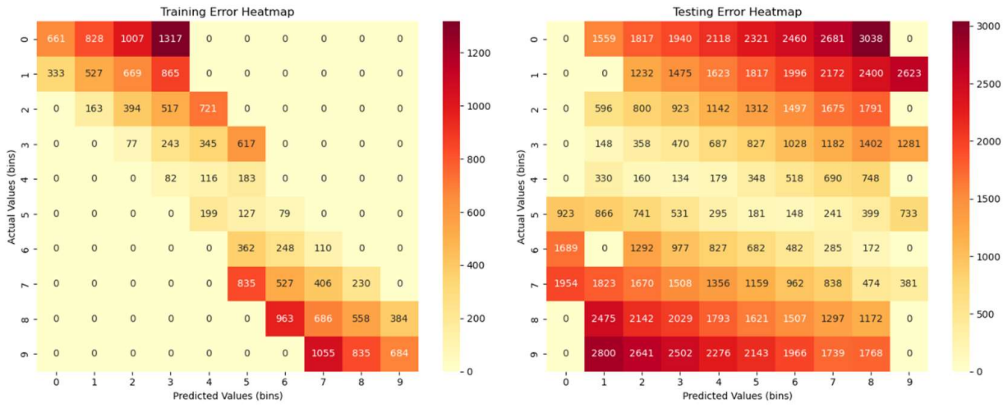

Figure 12 Error heatmaps for training and testing predictions, with MAE, MSE, RMSE calculations

Step 7: Feature Importance Ranking

Lastly, the Random Forest model's built-in feature importance scores were extracted (Figure 13). These reflect how much each feature contributes to reducing prediction error across all decision trees.

The top-ranked feature is **Total Energy Consumption**, followed closely by **Fossil Fuel Dependency**. These results align with global emission studies, where high energy usage and fossil-based supply chains are key contributors to emissions (IEA, 2022).

```
Step 6: Feature Importance Analysis
--------------------------------------------------
Feature Importance Ranking:
                          Feature  Importance
0  Total Energy Consumption (TWh)    0.151714
3      Fossil Fuel Dependency (%)    0.148729
5        Household Energy Use (%)    0.147556
2      Renewable Energy Share (%)    0.147387
1     Per Capita Energy Use (kWh)    0.147270
4       Industrial Energy Use (%)    0.146109
6    Energy Price Index (USD/kWh)    0.111236
```

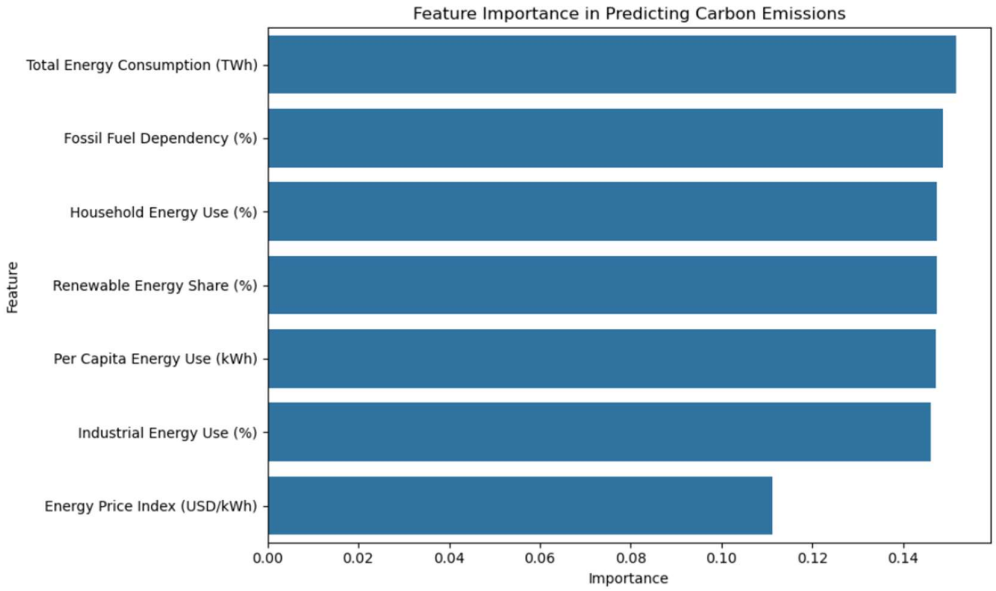

Figure 13 Feature importance ranking for predicting carbon emissions

4.Conclusion

This tutorial presented a detailed walkthrough of a machine learning pipeline aimed at predicting **Carbon Emissions (Million Tons)** using various energy-related indicators from a global dataset comprising **10,000 samples**. The initial stage involved understanding the structure of the dataset, which included 10 columns such as **Total Energy Consumption (TWh)**, **Per Capita Energy Use (kWh)**, and **Fossil Fuel Dependency (%)**, among others. These features were found to be complete, with no missing values, simplifying the preprocessing phase and allowing us to retain all samples for analysis.

For the modeling task, seven numeric features were selected based on their theoretical relevance and statistical patterns observed during the exploratory phase. The data was then split into **8,000 training samples** and **2,000 test samples**, maintaining an 80-20 ratio to evaluate model performance effectively. Standardization was applied using `StandardScaler`, ensuring the transformed features had a mean of approximately 0 and a standard deviation of 1. This helped normalize the input space, although the model used—Random Forest—was not sensitive to feature scaling.

The model was trained using a **Random Forest Regressor** consisting of 100 estimators. While the training results showed strong predictive performance (**RMSE = 542.59**, **R^2 = 0.85**), the testing metrics were notably poor (**RMSE = 1469.13**, **R^2 = -0.06**), revealing severe overfitting. The heatmaps and error distribution plots supported this observation, with testing errors showing significant spread and higher magnitude compared to the tightly clustered training

errors. The **RMSE difference** of **926.54** and the **R² difference** of **0.91** between the train and test sets clearly confirm the model's inability to generalize.

Further analysis of feature importance showed that **Total Energy Consumption** had the highest influence on carbon emissions (importance score: **0.1517**), followed closely by **Fossil Fuel Dependency (0.1487)** and **Household Energy Use (0.1476)**. This aligns with global research that associates higher emissions with energy-intensive and fossil-reliant economies. The remaining features—such as **Renewable Energy Share**, **Per Capita Energy Use**, and **Industrial Energy Use**—also showed significant importance, indicating that emissions are shaped by a combination of both macro-level and sector-specific energy patterns.

In conclusion, while the Random Forest model effectively learned from the training data, it suffered from high variance and failed to generalize to new samples. This outcome underlines the necessity for more robust modeling strategies such as cross-validation, hyperparameter optimization, and potentially more complex ensemble techniques. Incorporating additional features like region, country, or industrial classification could also enhance the model's contextual understanding. Overall, this tutorial illustrated the complete end-to-end machine learning workflow with practical energy and environmental applications, emphasizing both the potential and challenges of predictive analytics in sustainability-focused domains.

5.References

•Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

• International Energy Agency (IEA). (2022). *Global Energy Review: CO2 Emissions in 2022*. https://www.iea.org/reports/global-energy-review-co2-emissions-in-2022

• James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning* (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

• Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

Github Link: https://github.com/Mukeshsaragadam/PREDICTING-CARBON-EMISSIONS-USING-RANDOM-FOREST-REGRESSION-.git