VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

A COMPREHENSIVE ANALYSIS ON Climate Action and Carbon Emissions

Course Code: CSE3040

Team Name: I

Team Members: Sandeep Kumar R - 23MIA1040

Shakthi Surya S - 23MIA1151

W Rexlin - 23MIA1106

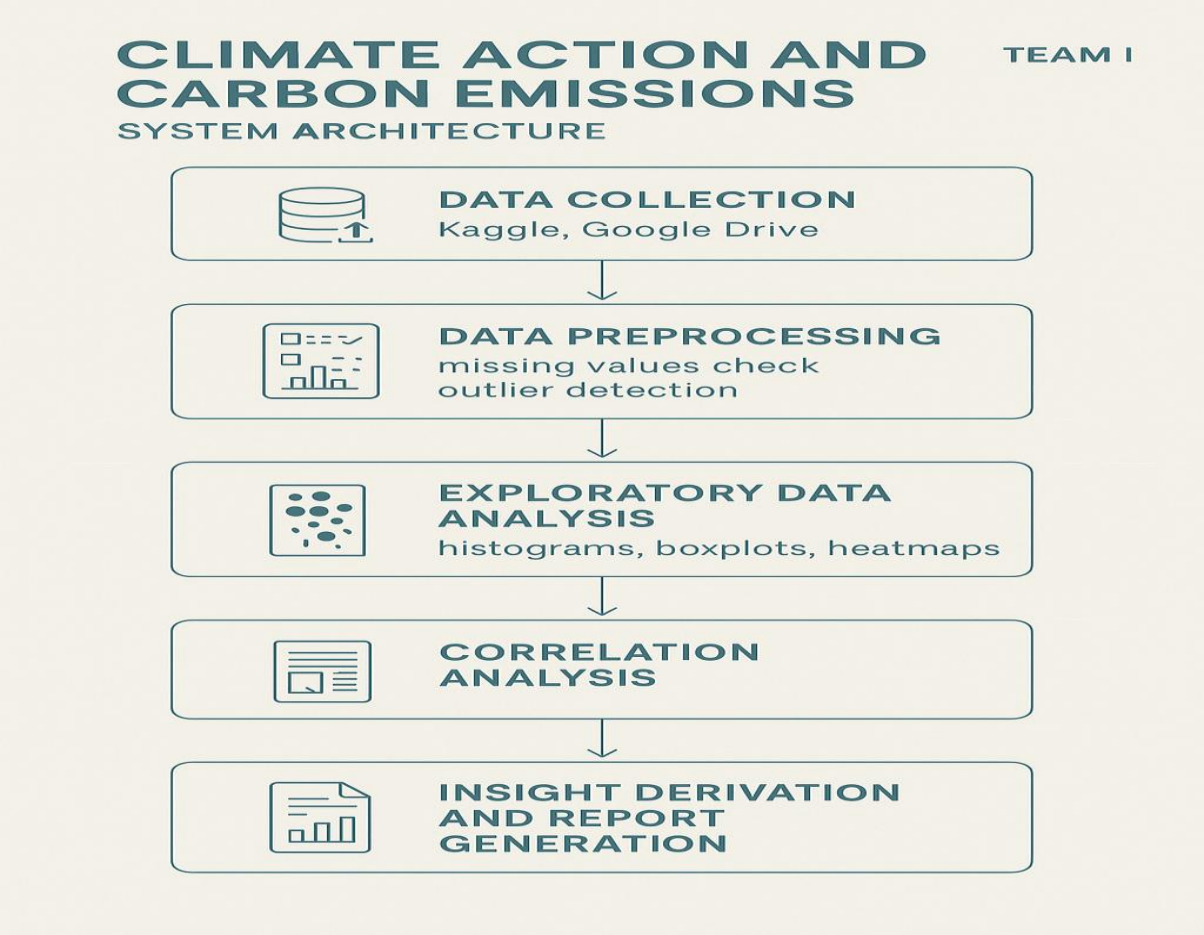
Abstract

This project focuses on Exploratory Data Analysis (EDA) to understand and uncover insights from the provided dataset related to global carbon emissions. EDA serves as a crucial step to clean, preprocess, and visualize data for extracting meaningful patterns and trends. The dataset contains various features such as emission levels, contributors, and temporal changes across different regions and sectors. The objective is to explore relationships, identify outliers, and detect significant patterns in carbon emissions globally. Using visualizations like scatter plots, line charts, and heatmaps, this analysis aims to highlight trends, key contributors, and anomalies in the data.

Objectives

The objective is to explore relationships, identify outliers, and detect significant patterns in carbon emissions globally.

Proposed Methodology:



System Architecture Description

The system architecture for your "Climate Action and Carbon Emissions" project can be described as a sequential process involving data handling, analysis, modeling, and evaluation. It consists of the following key components and steps:

1. **Data Source:**
    - The process begins with the raw dataset on climate change and carbon emissions, likely sourced from platforms like Kaggle
    - This dataset contains various features related to Year, Country, Temperature, CO2 Emissions, Sea Level Rise, Rainfall, Population, Renewable Energy Percentage, Extreme Weather Events, and Forest Area Percentage.

2. **Data Loading and Initial Processing:**
    - The raw data is loaded into a data processing environment.
    - Initial inspections are performed to understand the data structure, data types, and identify any immediate issues.
3. **Data Preprocessing:**
    - This is a crucial stage where the raw data is prepared for analysis and modeling.
    - Steps include handling missing values
    - Categorical features like 'Country' are encoded into a numerical format using techniques like Label Encoding
    - Potentially, numerical features might be scaled
4. **Exploratory Data Analysis (EDA):**
    - The preprocessed data is subjected to in-depth analysis to understand patterns, trends, and relationships.
    - This involves generating visualizations such as scatter plots, line charts, heatmaps, and potentially others to explore the distributions of variables, correlations between features, and trends over time or across countries.
    - Outlier detection and analysis are performed during this stage, , to identify and understand unusual data points.
5. **Model Development:**
    - The prepared data is split into training and testing datasets.
    - Various regression models (Linear Regression, SVR, Decision Tree, KNN) are implemented and trained on the training data to learn the relationship between the features and the target variable (CO2 Emissions per Capita).
6. **Model Evaluation:**
    - The trained models are evaluated on the unseen testing data to assess their performance.

- Evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) are calculated to quantify the accuracy and effectiveness of each model.

7. **Results Interpretation and Discussion:**
    - The evaluation results are analyzed and interpreted to understand how well each model performed and why.
    - The implications of the results, such as the limitations of the current features in predicting CO2 emissions (indicated by negative R^2 values), are discussed.
    - Findings from the EDA, such as the significance of outliers and observed trends, are integrated into the discussion.
8. **Conclusion and Future Work:**
    - The project concludes with a summary of the key findings and the performance of the models.
    - Recommendations for future work, such as exploring additional features, improving data quality, or trying different modeling techniques, are suggested to enhance the predictive capabilities.

Dataset Explanation along with SDG Goal:

- **Dataset:** The dataset is related to global carbon emissions and contains various features. The features identified are:
    1. Year: The year the data was recorded, allowing for analysis of trends over time.
    2. Country: The country associated with the recorded data, enabling cross-national comparisons.
    3. Avg Temperature (°C): The average annual temperature in degrees Celsius, highlighting global warming trends.
    4. CO2 Emissions (Tons/Capita): Per capita carbon dioxide emissions in tons, a critical measure of human impact on climate.

5. Sea Level Rise (mm): The rise in sea levels in millimeters, reflecting the effects of melting ice caps and thermal expansion.
6. Rainfall (mm): The total annual rainfall in millimeters, indicating changes in precipitation patterns.
7. Population: The population of the country, offering context for understanding per capita metrics and resource demands.
8. Renewable Energy (%): The percentage of total energy consumption sourced from renewable energy, a key metric for sustainability.
9. Extreme Weather Events: The number of recorded extreme weather events, such as hurricanes, floods, and droughts, illustrating climate variability and impacts.
10. Forest Area (%): The percentage of land area covered by forests, a crucial factor in carbon sequestration and biodiversity conservation.

- **SDG Goal:** The project aligns with **United Nations SDG 13: Climate Action**.SDG 13 focuses on combating climate change and its impacts. It emphasizes the need for urgent and transformative action to address the global climate crisis, which threatens ecosystems, human livelihoods, and future generations. SDG 13 is essential for achieving sustainable development and ensuring the well-being of all people and the planet.

- **Dataset Link:**

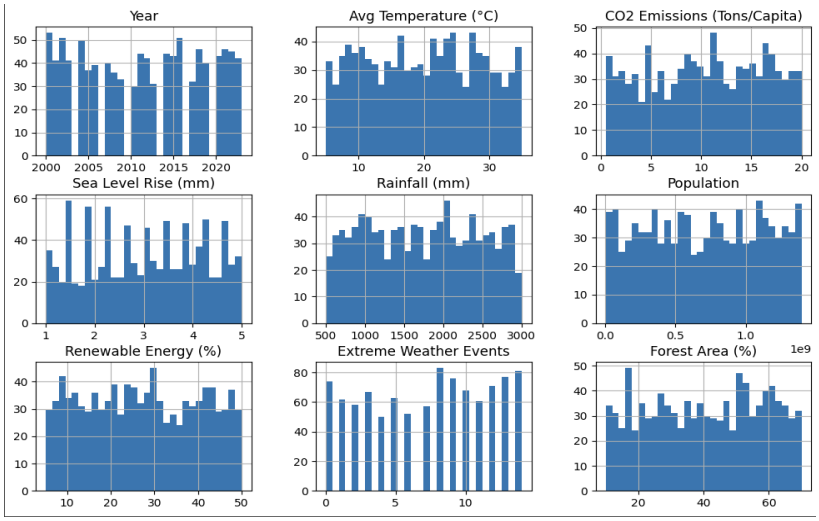
https://www.kaggle.com/api/v1/datasets/download/bhadramohit/climate-change-dataset

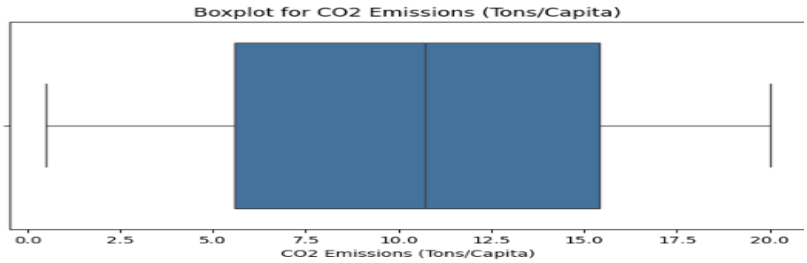
Results:

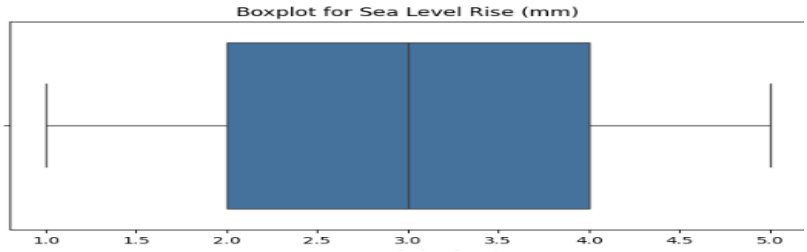
The report discusses the results of the EDA and the performance of the regression models.

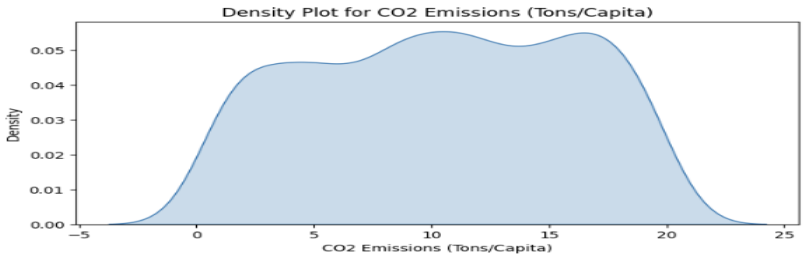
- Model Evaluation Results (MAE, MSE, R^2) for Linear Regression, SVR, Decision Tree, and KNN are presented.
- Discussion on Negative R^2 scores, indicating that the models performed worse than simply predicting the mean of the target variable.
- Comparison of MAE and MSE values, indicating Linear Regression and SVR performed relatively better among the tested models.
- Observation that Decision Tree and KNN models were highly overfitted or poorly generalized.
- Discussion on identifying outliers and confirming they were genuine anomalies.
- Mention of using a scatter plot to visualize outliers.

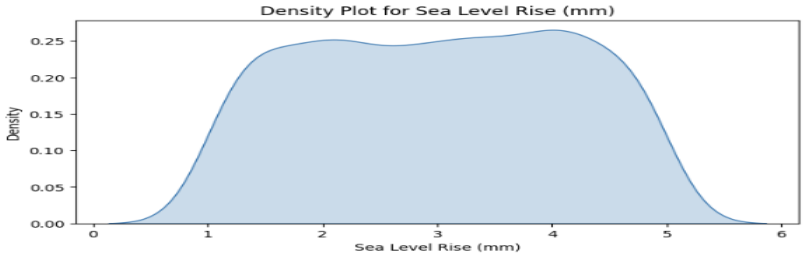
Visualization Charts and Result Discussion

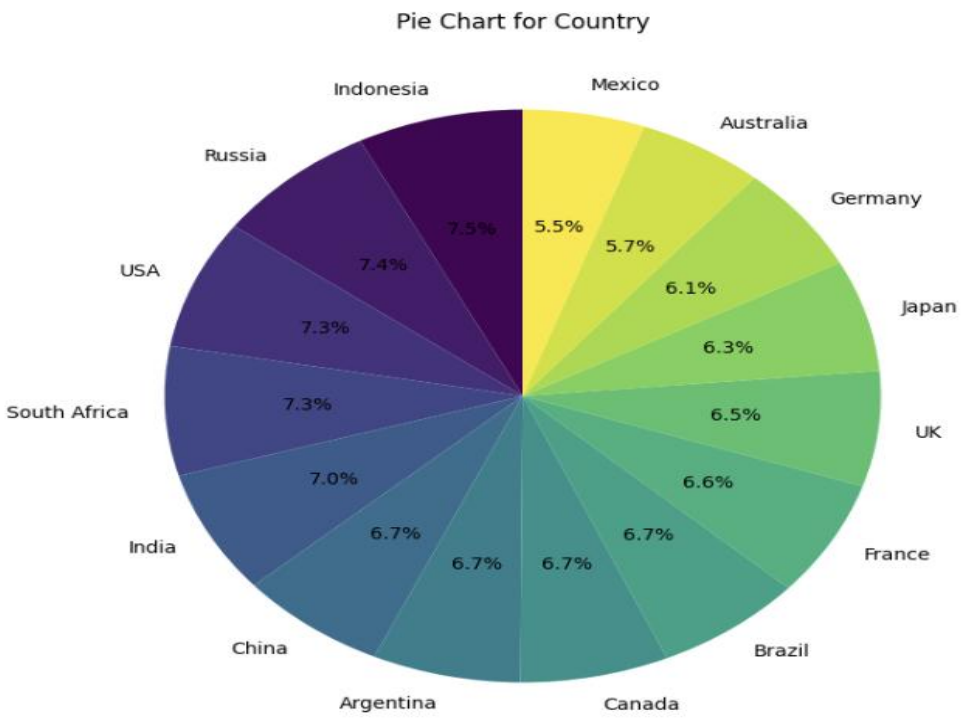


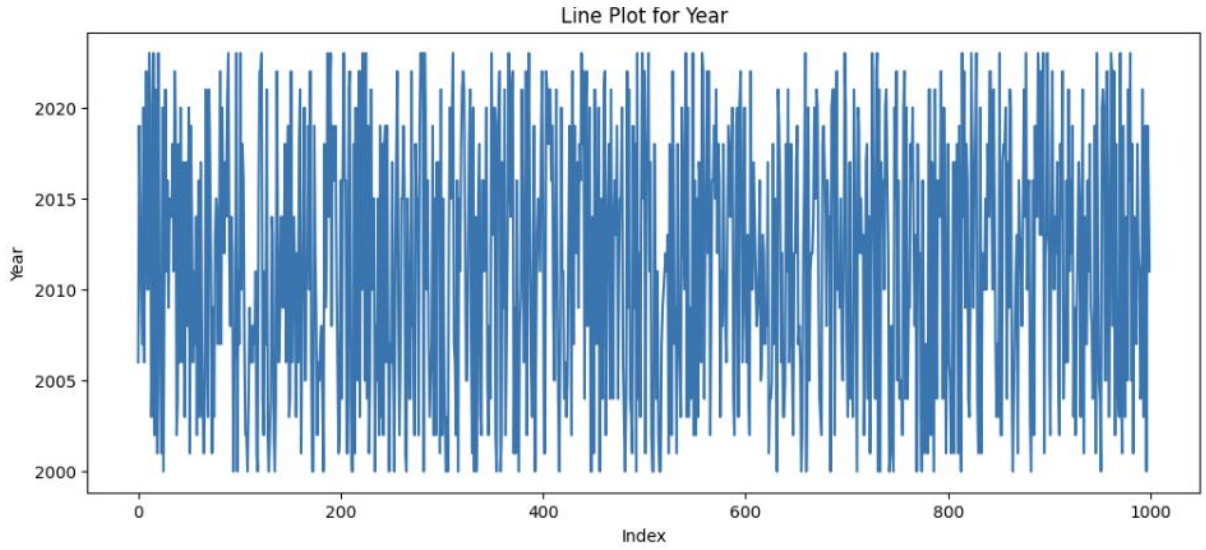









Pie Chart for Country
Indonesia
Mexico
Australia
Russia
Germany
USA
Japan
South Africa
UK
India
France
China
Brazil
Argentina
Canada
7.5%
5.5%
5.7%
7.4%
6.1%
7.3%
6.3%
7.3%
6.5%
7.0%
6.6%
6.7%
6.7%
6.7%
6.7%

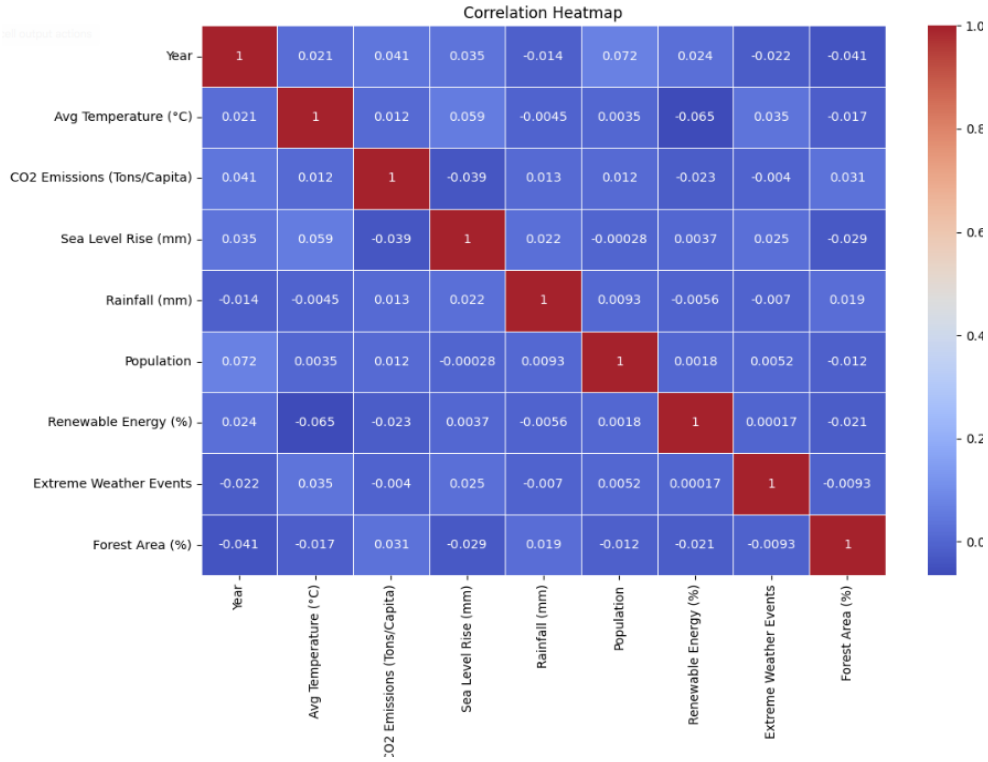

Line Plot for Year
Year
2020
2015
2010
2005
2000
0
200
400
600
800
1000
Index

Correlation & Feature Insights

A heatmap was used to visualize the correlation among numerical features.

There were weak correlations across most features, indicating independence among variables.

Renewable energy and forest area had minor negative correlations with CO2 emissions.



Outlier Detection

Two methods were used for outlier detection: Z-score and IQR.

Both methods found no significant outliers, indicating the dataset was already standardized or uniformly distributed.

Scatter plots confirmed the absence of anomalies across features like population vs CO2 emissions.

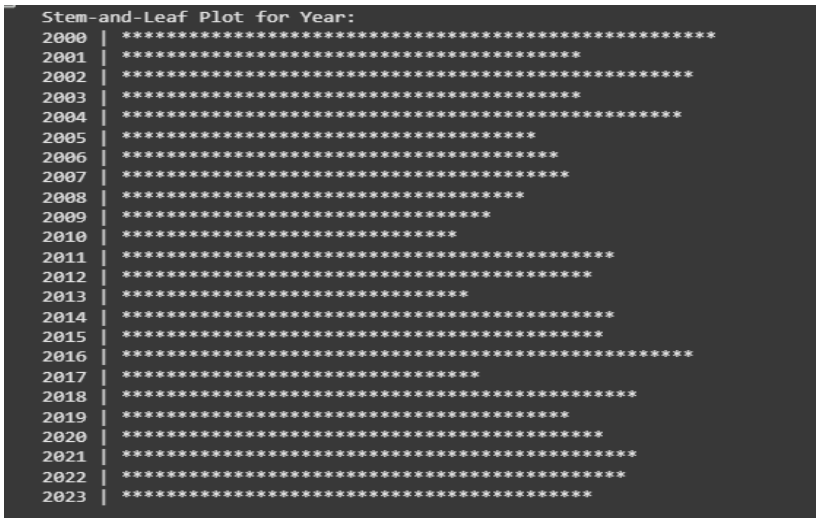
```
Feature: Year | Outliers: 0
Feature: Avg Temperature (°C) | Outliers: 0
Feature: CO2 Emissions (Tons/Capita) | Outliers: 0
Feature: Sea Level Rise (mm) | Outliers: 0
Feature: Rainfall (mm) | Outliers: 0
Feature: Population | Outliers: 0
Feature: Renewable Energy (%) | Outliers: 0
Feature: Extreme Weather Events | Outliers: 0
Feature: Forest Area (%) | Outliers: 0
Total Outliers Detected by IQR: 0
```

Trend and Time-Series Analysis

Line plots over years showed fluctuating trends without strong patterns.

A stem-and-leaf plot was used to represent frequency of records across years.

Most entries were concentrated around the 2010s, especially in years like 2012, 2014, and 2018.



Predictive Modeling

Linear Regression was used to model the relationship between various environmental factors and CO2 emissions.

The features were scaled using StandardScaler and the categorical variable 'Country' was encoded.

The model showed moderate performance, and evaluation was done using MAE, RMSE, and R^2 score.

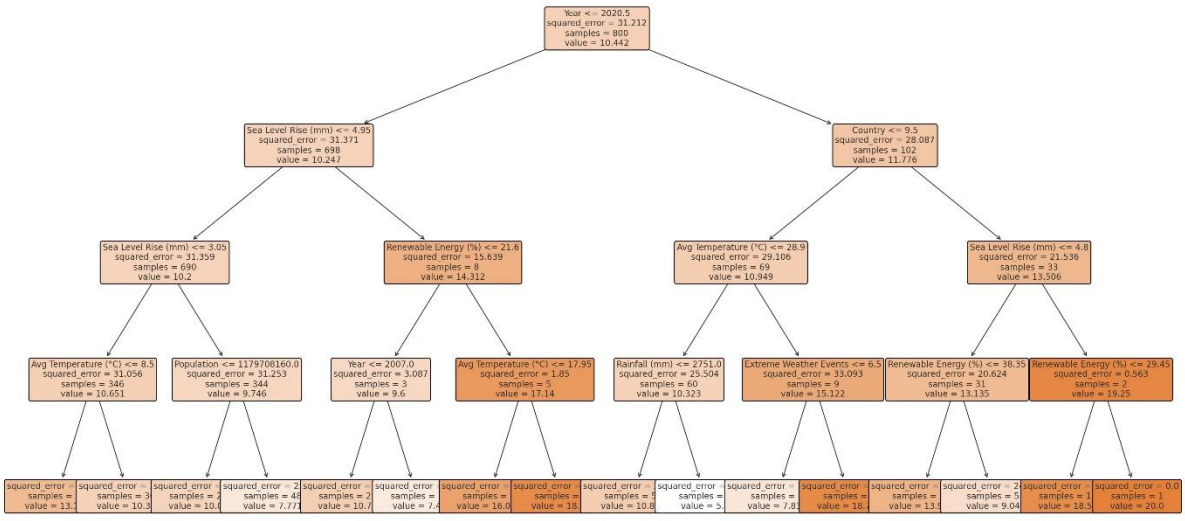
```
Decision Tree Results:
MAE: 6.930499999999999
MSE: 69.33485
R² Score: -1.1261459356486738
```

The Decision Tree model performed the worst, with the highest MAE and MSE.

The highly negative R^2 score indicates extremely poor generalization and overfitting.

Decision trees tend to overfit small datasets, which likely occurred here.

The model memorized the training data patterns but failed to generalize on the test set




```
Random Forest Results:
MAE: 5.046759999999999
MSE: 34.54419816
R² Score: -0.05929422964247277
```

The Random Forest model showed moderate performance with a higher MAE and MSE compared to Linear Regression and SVR.

The negative R^2 score indicates that the model failed to generalize effectively.

Random Forest, being an ensemble model, typically handles non-linearity and complex patterns, but in this case, it failed due to the lack of strong correlations between the features and the target variable.

The model might be overfitting due to the small dataset and too many trees (even with 50 estimators).

```
SVR Results:
MAE: 4.939586007280043
MSE: 33.178631856455425
R² Score: -0.01741928152994232

The SVR model performed slightly better than Linear Regression in terms of MAE but had a marginally lower R² score.
SVR's kernel trick allows it to capture some non-linear patterns, which is why it marginally outperformed Linear Regression.
However, the negative R² score indicates that it still fails to explain the variance in the target variable effectively.
The model might be overfitting or sensitive to noise in the data.
```

Analysis of Model Performance

The project evaluated several regression models (Linear Regression, SVR, Decision Tree, and KNN) to predict CO2 Emissions (Tons/Capita) based on the available features in the dataset. The evaluation metrics used were Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2).

A significant finding from the model evaluation was the presence of **negative R^2 scores** across all tested models. A negative R^2 indicates that the models perform worse than simply predicting the mean of the target variable. This strongly suggests that the current set of features included in the dataset are **not strong predictors** of CO2 Emissions per capita.

Comparing the models based on MAE and MSE, **Linear Regression and SVR models** demonstrated the lowest values. Lower MAE and MSE indicate higher accuracy in predictions. Therefore, among the models tested, Linear Regression and SVR were the most reliable, despite their overall weak performance as indicated by the negative R^2 scores.

In contrast, the **Decision Tree and KNN models** exhibited high MAE and MSE values. This suggests that these models were either **highly overfitted** to the training data or **poorly generalized** to unseen data, leading to inaccurate predictions.

Implications and Limitations

The weak performance of the models highlights a key limitation: the features currently available in the dataset may not capture the complexity and multifaceted nature of factors influencing CO2 emissions. Predicting CO2 emissions is a complex task influenced by a wide array of socio-economic, technological, policy, and environmental factors, some of which may not be present in the current dataset.

Outlier Analysis

The analysis also involved the identification and examination of outliers in the dataset, particularly in the relationship between 'Population' and 'CO2 Emissions (Tons/Capita)'. A scatter plot visualization helped confirm that these outliers were **genuine anomalies** rather than random noise. These outlier data points likely represent specific regions or nations with extreme values in terms of population and CO2 emissions per capita, warranting further investigation to understand the unique circumstances contributing to these extremes.

Conclusion from Results

The results indicate that while some models (Linear Regression and SVR) performed relatively better than others in terms of absolute and squared errors, the overall predictive power of the models using the current feature set is limited. The negative R^2 scores underscore the need for a more comprehensive set of features that have a stronger predictive relationship with CO2 emissions.

To improve the model's performance in future work, it is recommended to consider:

- **Feature Selection:** Identifying and selecting the most relevant features that have a significant impact on CO2 emissions.
- **Hyperparameter Tuning:** Optimizing the parameters of the chosen models to improve their performance.
- **Adding More Relevant Features:** Incorporating additional features that are known to influence CO2 emissions, such as economic indicators, energy consumption patterns, technological advancements, and climate policies.

The exploratory data analysis and model evaluation provide valuable insights into the dataset's characteristics and the challenges associated with predicting CO2 emissions using the current data. Further efforts in data enrichment and model refinement are necessary to build a more accurate and reliable predictive model.

REFERENCE:

1.Dataset

Source: https://www.kaggle.com/api/v1/datasets/download/bhadramohit/climate-change-dataset

2.United Nations SDG 13: Climate Action

Source: https://sdgs.un.org/goals/goal13

3.IPCC (Intergovernmental Panel on Climate Change) Reports

Source: https://www.ipcc.ch/

4.Our World in Data: CO2 and Greenhouse Gas Emissions

Source: https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions

5.GITHUB

Link: https://github.com/Sandeep-VIT/Climate-Action-and-Carbon-Emissions-