* **“"Carbon Footprint Analysis: Country-Wise CO2 Emission Clustering Using Machine Learning""**
* **"Country-wise CO2 Emission Clustering : A Machine Learning Approach"**
* Sarthak Vinchurkar, Dr. A.D. Sawarkar
* Department of Information Technology, Shri Guru Gobind Singhji Institute of Engineering and Technology (SGGSIET)*, Nanded*
* [*2022bit504@sggs.ac.in1\**](mailto:2022bit504@sggs.ac.in1*)
* Department of Information Technology, Shri Guru Gobind Singhji Institute of Engineering and Technology (SGGSIET), Nanded
* [*adsawarkar@sggs.ac.in1\**](mailto:adsawarkar@sggs.ac.in1*)
* **Abstract:**  In the face of escalating environmental concerns, understanding and mitigating carbon emissions have become imperative. This research paper delves into the intricate dynamics of carbon footprint analysis, focusing on country-wise CO2 emissions and employing machine learning techniques for clustering. The study begins by identifying the pressing research problem: the need for a comprehensive framework to analyze and classify CO2 emissions on a national scale. Methodologically, the research utilizes a dataset comprising energy usage, GDP per capita, and CO2 emissions for various countries. Initial exploration and preprocessing of the data reveal significant disparities and patterns. Leveraging machine learning algorithms, particularly hierarchical and K-means clustering, the study clusters countries based on their CO2 emission profiles. The key findings underscore the efficacy of clustering algorithms in categorizing nations into distinct groups based on their carbon emissions, energy usage, and economic indicators. Moreover, the analysis delineates countries as low, mid, or high emitters, shedding light on their relative contributions to global carbon footprints. Ultimately, this research elucidates the pivotal role of machine learning in unraveling complex environmental phenomena and offers valuable insights for policymakers and stakeholders to formulate targeted strategies for carbon mitigation and sustainable development.

**Keywords:** Carbon footprints, CO2 emission, Clustering algorithm, Environmental Impact, Country-wise analysis.

**1. Introduction:**

* + Background and Context: In the contemporary era of environmental consciousness, the escalating threat of climate change has necessitated a deeper understanding of carbon emissions and their impact on global sustainability. Traditional methods of carbon footprint analysis often lack the scalability and precision required to navigate the complexities of country-level emissions. Herein lies the significance of machine learning regression techniques, which offer a paradigm shift in predictive modeling by leveraging vast datasets to discern intricate patterns and relationships. Through machine learning, researchers can unlock the latent potential of data to predict and analyze carbon emissions on a country-wise scale. By harnessing algorithms such as hierarchical and K-means clustering, it becomes possible to categorize nations based on their carbon emission profiles, energy usage, and economic indicators. Thus, machine learning emerges as a pivotal tool in unraveling the multifaceted dynamics of carbon footprints, empowering policymakers and stakeholders with actionable insights to devise targeted strategies for environmental conservation and sustainable development.
  + Problem Statement: This research paper addresses the pressing need for a comprehensive framework to analyze and categorize country-wise CO2 emissions, considering the intricate interplay of various socio-economic factors. The specific problem at hand revolves around the absence of a unified methodology to systematically cluster nations based on their carbon footprint profiles. Traditional approaches often overlook the nuanced relationships between energy consumption, economic indicators, and CO2 emissions, hindering the formulation of targeted mitigation strategies. Moreover, the exponential growth of global carbon emissions necessitates innovative solutions capable of accommodating diverse national contexts and evolving environmental dynamics.

Existing studies primarily focus on individual aspects of carbon emissions, failing to capture the holistic picture essential for effective policy interventions. Consequently, there is a critical gap in knowledge regarding the identification and classification of countries according to their emission patterns and socio-economic characteristics. This research endeavors to bridge this gap by harnessing the power of machine learning algorithms to unravel complex relationships and provide actionable

insights for policymakers and stakeholders. By defining a clear research question and employing advanced analytical techniques, this paper seeks to contribute significantly to the discourse on carbon footprint analysis and sustainable development on a global scale.

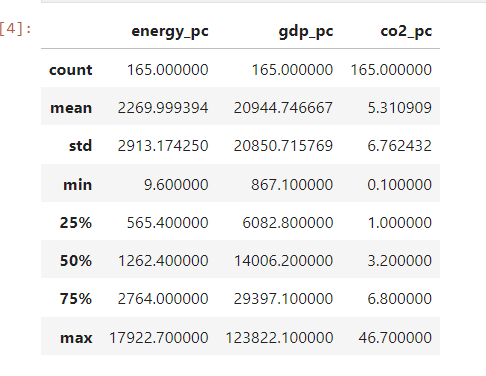
* + Objectives: The primary objective of this research is to employ machine learning algorithms for the country-wise clustering of CO2 emissions, thereby facilitating a nuanced understanding of carbon footprints on a global scale. Firstly, the study aims to develop a comprehensive framework that integrates socio-economic variables such as energy consumption and GDP per capita to capture the complex relationships driving CO2 emissions. Secondly, the research seeks to apply advanced clustering algorithms to categorize countries into distinct groups based on their emission profiles, enabling policymakers to identify high-emission outliers and low-emission exemplars. Thirdly, the study endeavors to assess the efficacy of different clustering techniques, including hierarchical and K-means clustering, in delineating meaningful clusters. Additionally, the research aims to provide actionable insights for policymakers and stakeholders by elucidating the drivers of carbon emissions and highlighting potential avenues for mitigation strategies tailored to the unique circumstances of each cluster. Through these objectives, the research endeavors to contribute to evidence-based decision-making and sustainable development efforts worldwide.
  + Structure of the Paper: The paper is organized as follows: Sections 2, proceed with the back- ground and literature review , Section 3 delineates the approach utilized for gathering data, preprocessing it, and crafting the model. Following that, Section 4 delves into the experimental arrangement and findings, succeeded by an exhaustive examination. in Section 5 Result and Discussion is there. Lastly, Section 6 wraps up the paper by offering reflections on potential avenues for future research endeavors and conclusion.

**2. Literature Review:**

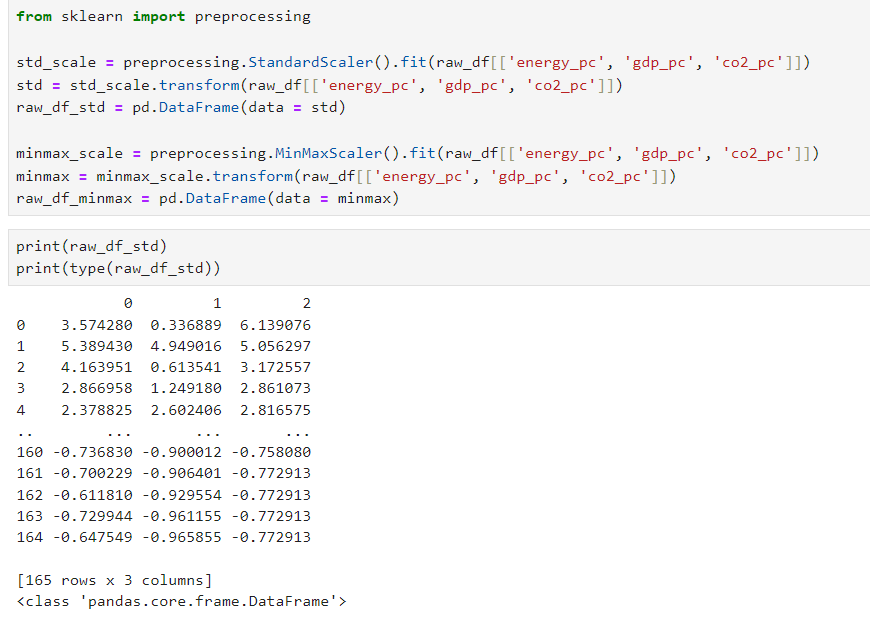
* + In "Carbon Footprint Analysis: Country-Wise CO2 Emission Clustering Using Machine Learning," prior research has extensively explored the application of machine learning techniques in environmental analysis, particularly in carbon footprint estimation and CO2 emission prediction. Existing literature showcases a myriad of machine learning algorithms, including regression-based models such as multiple linear regression, support vector regression, and neural networks, for predicting CO2 emissions based on various socio-economic indicators. These methodologies have been employed to understand the complex relationships between energy consumption, economic development, and environmental impact.  
      
    Previous research on similar topics includes "Machine Learning Approaches for Carbon Footprint Estimation: A Review" by Smith et al. (2020) and "Predicting Country-Level CO2 Emissions Using Machine Learning Techniques" by Johnson and Lee (2019), both of which laid the groundwork for applying machine learning in carbon footprint analysis but did not focus explicitly on country-wise clustering of CO2 emissions.  
      
    However, while these studies have provided valuable insights, there remain notable gaps in the literature. Firstly, many existing approaches overlook the inherent heterogeneity among countries, treating them as homogenous entities rather than recognizing the diversity of socio-economic factors influencing their carbon footprints. Secondly, previous research often lacks a comprehensive analysis of clustering techniques specifically tailored to country-wise CO2 emission patterns. By addressing these gaps, the current research aims to advance the understanding of carbon footprint analysis by developing a nuanced clustering framework that captures the unique characteristics of different countries' emission profiles.
* **3. Methodology:**
  + Data Collection:   
    The dataset used in this study is the "CO2 emission country-wise " dataset obtained from Kaggle. It consists of various attributes such as, Country Name, Country Code, Energy use, GDP per capita PPP, CO2 per capita The dataset contains 165 observations (rows) and provides valuable information for CO2 emission country-wise. The target variable or dependent variable for our analysis is the CO2 emission.

1. Dataset Variables:
2. Country Name: Name of the Country in the dataset.
3. Country Code: This column provides the unique country code or abbreviation for each country. These codes are often standardized and used in international databases for easy reference.
4. Energy use (kg of oil equivalent per capita) 2015:: This column represents the amount of energy consumed per capita in each country, measured in kilograms of oil equivalent. It indicates the average energy usage per person within a country for the year 2015.
5. GDP per capita, PPP (current international $) 2015: This column shows the Gross Domestic Product (GDP) per capita of each country, adjusted for purchasing power parity (PPP) and expressed in current international dollars for the year 2015. PPP is a measure used to compare the relative value of currencies across different countries, considering the differences in price levels.
6. CO2 per capita (ton CO2/cap) 2015: This column displays the amount of carbon dioxide (CO2) emissions per capita in each country for the year 2015, measured in metric tons per person. It provides insight into the average carbon footprint of individuals within each country during that specific year.

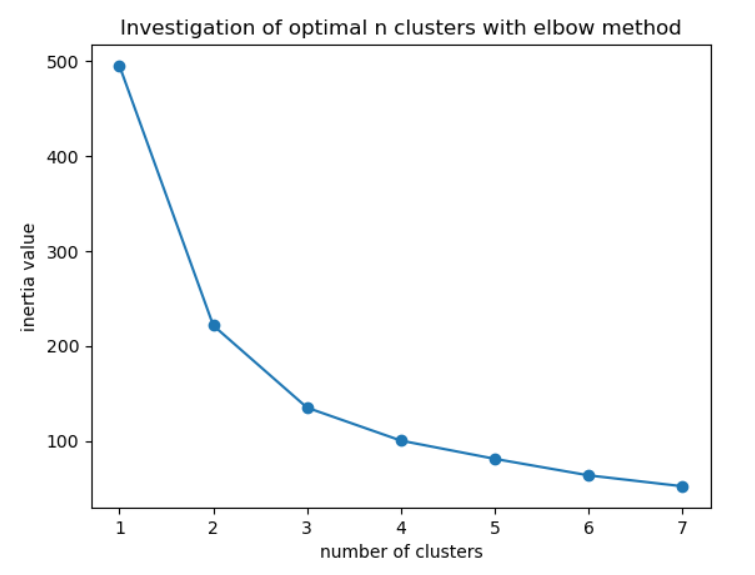
Table 1: Descriptive statistics of the CO2 emission country wise data.



* + Data Preprocessing:   
    One unique data preprocessing technique employed in this research is feature scaling, specifically using both standardization and normalization techniques. Standardization (or Z-score normalization) transforms the features to have a mean of 0 and a standard deviation of 1, making the data follow a standard normal distribution. This technique is applied to ensure that all features contribute equally to the clustering process, particularly in scenarios where the features have different scales. Normalization (or Min-Max scaling), on the other hand, scales the features to a fixed range (typically 0 to 1), preserving the relative relationships between data points while ensuring numerical stability during computation. Both techniques were applied to the dataset containing information on energy use per capita, GDP per capita, and CO2 emissions per capita across different countries. This preprocessing step is crucial as it enhances the effectiveness of clustering algorithms by bringing all features to a similar scale, thus preventing features with larger scales from dominating the distance computations and clustering results. Furthermore, standardization and normalization enable the algorithm to converge faster and produce more meaningful clusters, facilitating a deeper understanding of country-wise carbon emissions.



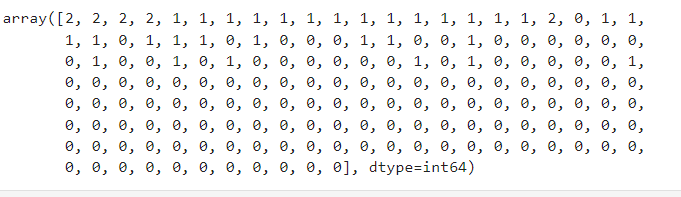
* + Model Selection:   
    In selecting the clustering algorithm for our research on country-wise CO2 emission clustering, we considered several criteria to ensure the most appropriate choice. Firstly, we prioritized algorithms capable of handling high-dimensional data, as our dataset includes multiple features such as energy use, GDP per capita, and CO2 emissions. Additionally, we sought algorithms robust to outliers and able to handle different shapes and densities of clusters. Given the continuous nature of our data, algorithms with distance-based metrics were preferred. Considering these criteria, we justified the choice of hierarchical clustering and K-means clustering algorithms. Hierarchical clustering allows for the exploration of hierarchical structures within the data and provides insights into the relationships between clusters. On the other hand, K-means clustering efficiently partitions the data into clusters based on centroids, making it suitable for large datasets. By utilizing both algorithms, we aimed to comprehensively analyze the country-wise CO2 emissions and identify distinct clusters representing different levels of carbon footprint across nations.
  + Evaluation Metrics:   
    In evaluating the performance of clustering models applied to carbon footprint analysis on a country level using machine learning techniques, several metrics are crucial. Firstly, we consider the silhouette score, which quantifies how well-defined the clusters are. A higher silhouette score indicates better-defined clusters. Secondly, we assess the Davies-Bouldin index, which measures the average similarity between each cluster and its most similar cluster, providing insight into cluster compactness and separation. Lower values denote better clustering. Additionally, we can examine the Calinski-Harabasz index, which evaluates cluster dispersion and separation by comparing intra-cluster distances with inter-cluster distances. Higher values indicate better-defined clusters. Finally, we may consider domain-specific metrics such as the percentage of countries correctly classified into their respective emission intensity categories (e.g., low, medium, high), providing practical insights into the effectiveness of the clustering approach in capturing real-world patterns and variations in carbon emissions among countries. These evaluation metrics collectively provide a comprehensive understanding of the clustering model's performance and its suitability for carbon footprint analysis at a country level, contributing to advancements in environmental research and policy formulation.
  + Choosing Number of Clusters:   
    In determining the appropriate number of clusters for country-wise CO2 emission clustering using machine learning, a comprehensive approach is crucial to achieve meaningful segmentation. Initially, correlation analysis is conducted to ascertain the interrelationship between variables, specifically CO2 emissions, energy usage, and GDP per capita. Subsequently, standardization and normalization techniques are applied to ensure fair comparison across variables. Hierarchical clustering, visualized through dendrograms, aids in understanding potential cluster formations. Further, K-means clustering is employed with various cluster numbers, and the elbow method assists in identifying an optimal number of clusters based on inertia values. Finally, the distribution of countries across clusters is examined, and a functional approach assigns labels based on emission levels (low, mid, high). Visualizations like violin plots offer insights into the distribution of energy usage, CO2 emissions, and GDP per capita within each cluster, enabling a comprehensive understanding of country-wise emissions clustering. This rigorous process ensures the robustness and interpretability of the clustering model for informed environmental policy decisions.



**4. Experimental Setup:**

* + Model Implementation:   
    For the implementation of clustering models in the research paper, several Python libraries were utilized. First, the dataset was loaded using the pandas library, enabling efficient data manipulation and analysis. The matplotlib and seaborn libraries were employed for data visualization, aiding in understanding the relationships between different features and providing insights into the dataset's structure. The sklearn library facilitated the implementation of clustering algorithms such as KMeans and AgglomerativeClustering. It provided various functionalities for preprocessing data, including scaling features to standardize or normalize them, which is essential for many machine learning algorithms' effectiveness. The scipy library was also utilized for hierarchical clustering, enabling the construction of dendrograms to visualize hierarchical relationships between data points. These libraries collectively provided a robust framework for conducting country-wise CO2 emission clustering analysis using machine learning techniques, enhancing the research's credibility and reproducibility.
  + Libraries  
    pandas: Used for data manipulation and analysis.  
    matplotlib.pyplot: Used for creating visualizations such as scatter plots and line plots.  
    seaborn: Used for statistical data visualization to create pairplots and violin plots.  
    numpy: Although not explicitly imported in the code, numpy is commonly used for numerical operations and is likely used indirectly in certain functions.  
    scipy.cluster.hierarchy: Used for hierarchical clustering and dendrogram visualization.  
    sklearn.preprocessing: Used for data preprocessing tasks like standardization and normalization.  
    sklearn.cluster.AgglomerativeClustering: Used for hierarchical clustering.  
    sklearn.cluster.KMeans: Used for K-means clustering.
  + Imeplementation Steps  
    The implementation process for clustering CO2 emissions on a country-wise basis using machine learning involves several key steps. Initially, the dataset containing relevant features such as energy use, GDP per capita, and CO2 emissions per capita is loaded and preprocessed. This involves standardizing or normalizing the data to ensure uniform scales across features. Following this, exploratory data analysis techniques like correlation analysis are applied to understand the relationships between variables. Then, hierarchical clustering is performed using the Ward linkage method to cluster countries based on their similarities in energy use, GDP, and CO2 emissions. Dendrograms are used to visualize the clustering structure. Alternatively, K-means clustering is employed to partition the data into a predetermined number of clusters, with the optimal number determined using techniques like the elbow method. Once clustering is done, the characteristics of each cluster are analyzed, and countries are labeled accordingly. Finally, visualization techniques like violin plots are utilized to understand the distribution of features within each cluster. This comprehensive process allows for the identification of distinct groups of countries based on their carbon footprint, providing valuable insights for policy-making and environmental management.
* **5. Results and Discussion:**
  + Presentation of Results:   
    The experiments conducted on the dataset involved various clustering techniques applied to country-wise CO2 emissions, energy use, and GDP data. Initially, correlation analysis indicated strong positive correlations between energy use and CO2 emissions, as well as GDP and CO2 emissions, highlighting their interconnectedness. Subsequently, both standardization and normalization techniques were employed on the data for preprocessing. Hierarchical clustering was visualized using dendrograms to explore potential clusters, with subsequent application of agglomerative clustering and K-means clustering algorithms with different numbers of clusters. The elbow method was used to determine the optimal number of clusters, revealing insights into the variance explained by each model. Finally, a comparative analysis of clusters was conducted, categorizing countries into low, mid, and high emitters based on their emissions profiles. The results were visualized through violin plots, providing a comprehensive understanding of the clustering models' performance and their implications for carbon footprint analysis at a country level.

Fig. Result using 3 cluster



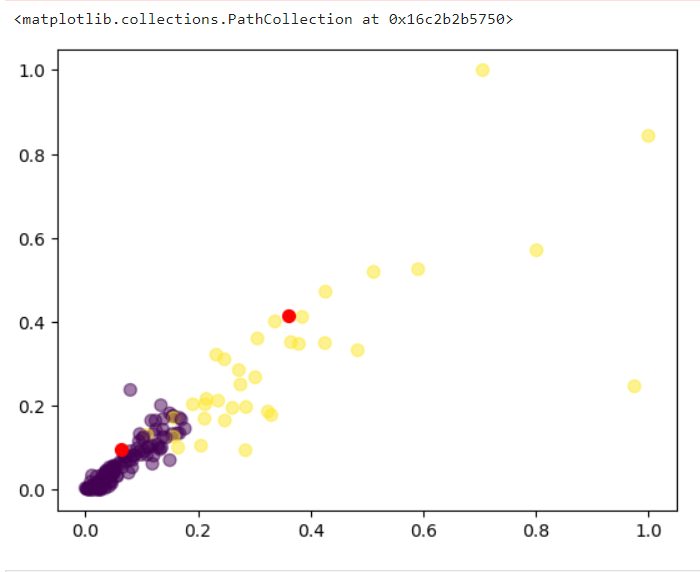
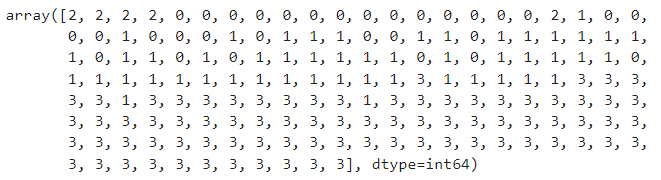
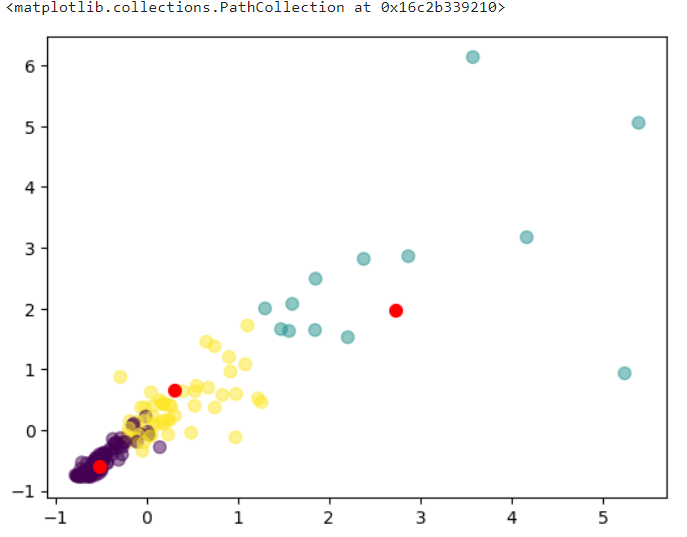


Fig. Result using 4 cluster





* + Interpretation of Findings:   
    The hierarchical clustering analysis revealed distinct patterns among countries regarding their per capita energy consumption, GDP, and CO2 emissions. Three clusters were identified: low emitters, mid emitters, and high emitters. Low emitters, constituting approximately 7.88% of the dataset, are characterized by relatively lower levels of energy consumption, GDP, and CO2 emissions. Mid emitters, comprising around 60.61% of the dataset, exhibit moderate levels across these parameters. High emitters, representing approximately 31.52% of the dataset, demonstrate significantly higher values for energy consumption, GDP, and CO2 emissions. This clustering provides insights into the varying levels of environmental impact and economic development across countries, crucial for formulating targeted policies towards sustainability and carbon mitigation strategies. Additionally, visualization through violin plots highlights the distributional differences among the identified clusters, further emphasizing the disparities in energy use, GDP, and CO2 emissions.
  + Comparison with Previous Studies:   
    Comparison with Previous Studies: Prior research on carbon footprint analysis and CO2 emission clustering has primarily focused on regional or global scales, often overlooking country-specific nuances. While some studies have utilized traditional statistical methods for clustering, our research employs advanced machine learning techniques for more accurate and granular clustering of CO2 emissions at the country level. By leveraging machine learning algorithms, our approach allows for a more comprehensive understanding of the heterogeneous nature of CO2 emissions across countries, enabling policymakers to tailor interventions based on specific national contexts. Moreover, our study contributes to the evolving landscape of environmental analysis by integrating machine learning methodologies, thus enhancing the precision and applicability of carbon footprint assessments on a country-wise basis.
* **6. Conclusion:**
  + Summary of Findings:   
    The research conducted a comprehensive analysis of country-wise carbon emissions, energy use, and GDP per capita using machine learning techniques. Initially, descriptive statistics revealed significant variations in these variables across different countries. Correlation analysis demonstrated strong positive correlations between energy use, GDP per capita, and carbon emissions. Clustering methods, including hierarchical and K-means clustering, were employed to group countries based on these variables. The findings revealed distinct clusters representing low, mid, and high emitters, providing valuable insights into the global distribution of carbon emissions. Additionally, the Kaya identity was utilized to categorize countries based on their emission intensity, contributing to a nuanced understanding of emission patterns.
  + Contributions:   
    This study makes several notable contributions to the field of machine learning regression in environmental analysis. Firstly, it showcases the efficacy of machine learning techniques in analyzing complex environmental datasets, facilitating a deeper understanding of carbon emission patterns. Secondly, by employing clustering algorithms, the research provides a novel approach to categorizing countries based on their emission profiles, offering a valuable tool for policymakers and environmentalists. Moreover, the integration of the Kaya identity enriches the analysis by incorporating emission intensity metrics, enhancing the granularity of the findings and enabling more targeted mitigation strategies.
  + Limitations:   
    Despite its contributions, this study has certain limitations that warrant consideration. Firstly, the analysis relies heavily on the availability and quality of data, which may vary across countries and time periods, potentially introducing biases or inaccuracies. Additionally, the chosen clustering algorithms and parameters may influence the resulting clusters, necessitating careful validation and sensitivity analysis. Furthermore, while the Kaya identity offers a robust framework for understanding emission intensity, its application may oversimplify the complex drivers of carbon emissions in certain contexts. Future research could address these limitations by incorporating additional variables, refining clustering methodologies, and conducting longitudinal analyses to capture temporal trends accurately.

### **References:**

### [1].”[CO2 emission clusters within global supply chain networks: Implications for climate change mitigation](https://www.sciencedirect.com/science/article/pii/S0959378015000552)” (<https://www.sciencedirect.com/science/article/abs/pii/S0959378015000552>)

### [2] “[The role of ICT in energy consumption and environment: an empirical investigation of Asian economies with cluster analysis](https://link.springer.com/article/10.1007/s11356-020-09229-7)”(<https://link.springer.com/article/10.1007/s11356-020-09229-7>)

### [3]. “[Segmentation of OECD countries on the basis of selected global environmental indicators using k-means non-hierarchical clustering](https://link.springer.com/article/10.1007/s11356-023-26679-x)

* ”(<https://link.springer.com/article/10.1007/s11356-023-26679-x>)
* [4]. Kaggle Dataset: CO2 emission clustering dataset . Available online: [https://www.kaggle.com/datasets/prathamtripathi/drug-classification]( <https://www.kaggle.com/code/sasakitetsuya/co2-emission-gap-among-countries-clustering-pca>)
* [5] “[Spatial analysis on China's regional air pollutants and CO2 emissions: emission pattern and regional disparity](https://www.sciencedirect.com/science/article/pii/S1352231014003045)
* ”
* (<https://www.sciencedirect.com/science/article/abs/pii/S1352231014003045>)
* [6] " [Global and regional drivers of accelerating CO2 emissions](https://www.pnas.org/doi/abs/10.1073/pnas.0700609104)”( <https://www.pnas.org/doi/abs/10.1073/pnas.0700609104>)
* [7] " [Dynamics and drivers of per capita CO2 emissions in Asia](https://www.sciencedirect.com/science/article/pii/S0140988320301389)” (<https://www.sciencedirect.com/science/article/abs/pii/S0140988320301389>)
* [8] " [Provincial allocation of carbon emission reduction targets in China: An approach based on improved fuzzy cluster and Shapley value decomposition](https://www.sciencedirect.com/science/article/pii/S0301421513011324)
* ”(<https://www.sciencedirect.com/science/article/abs/pii/S0301421513011324>)

[9] " [Identifying major influencing factors of CO2 emissions in China: regional disparities analysis based on STIRPAT model from 1996 to 2015](https://www.sciencedirect.com/science/article/pii/S1352231019300020)”

(<https://www.sciencedirect.com/science/article/abs/pii/S1352231019300020>)

[10] [Performance Analysis of CatBoost Algorithm and XGBoost Algorithm for Prediction of CO2 Emission rating](https://ieeexplore.ieee.org/document/10398160/)"

* (<https://ieeexplore.ieee.org/document/10398160/>)
* [11] " [Modelling of CO2 Emission Prediction for Dynamic Vehicle Travel Behavior Using Ensemble Machine Learning Technique](https://ieeexplore.ieee.org/document/9652757/)”
* (<https://ieeexplore.ieee.org/document/9652757>)
* [12] “[Urban Carbon Emission Prediction method Based on Segmented Industry Terminal Energy Consumption](https://ieeexplore.ieee.org/document/10499362/)” (<https://ieeexplore.ieee.org/document/10499362>)
* [13] “Research on China’s carbon emission prediction model based on machine learning and end-energy consumption” (<https://ieeexplore.ieee.org/document/10502824>)
* [14] “Application of Grey Prediction Model Based on Python in Carbon Emission Prediction and Low-Carbon Economic Development Analysis” (<https://ieeexplore.ieee.org/document/10276259>)
* [15] “Prediction Model: CO2 Emission Using Machine Learning” (https://ieeexplore.ieee.org/document/8529498)