Predictive Modeling of CO2 Emissions Based on Economic and Demographic Factors

Abed ALRaouf Sharabati
Arab American University
Faculty of Graduate Studies
Department of Natural, Engineering and Technology Sciences
Ramallah, West Bank, Palestine
a.sharabati@student.aaup.edu

Majdi Owda
Arab American University
Faculty of Artificial Intelligence and Data Science
UNESCO Chair in Data Science for Sustainable Development
Ramallah, West Bank, Palestine
majdi.owda@aaup.edu

Amani Yousef Owda*
Arab American University
Faculty of Graduate Studies
Department of Natural, Engineering and Technology Sciences
Ramallah, West Bank, Palestine
amani.owda@aaup.edu

***Abstract* — This paper develops a predictive model to estimate carbon dioxide (CO_2) emissions based on economic and demographic factors, specifically Gross Domestic Product (GDP) and population. Data from the World Bank Open Data Catalog were collected and processed using machine learning techniques. The Random Forest model was selected for its ability to capture the non-linear relationships between variables. The model's results reveal a strong correlation between economic growth and increased CO_2 emissions, emphasizing the need for stricter environmental policies to mitigate the adverse impacts of economic expansion on climate change. This research provides a data-driven tool to help policymakers assess the influence of economic and demographic factors on CO_2 emissions and implement more effective mitigation strategies.**

Keywords—CO2 Emissions, Gross Domestic Product (GDP), Random Forest, Predictive Modeling.

I. INTRODUCTION

Climate change is one of the most pressing challenges facing the world today, largely fueled by rising CO_2 emissions. These emissions are primarily generated by industrial activity and increasing human energy demands. Over the years, global temperatures have risen steadily. Data from the Copernicus ECMWF (2023) shows that 2023 recorded one of the highest temperature increases compared to pre-industrial levels Figure. 1. To address this issue effectively, it is crucial to examine how economic and demographic factors drive these emissions. Understanding the relationship between economic growth, population, and CO_2 emissions is essential for developing environmental policies and advancing sustainable development goals.

This research focuses on creating a predictive model to estimate CO_2 emissions using GDP and population data as key variables. By employing machine learning methods, specifically Random Forest regression, the study aims to offer a data-driven framework for analyzing and forecasting emission trends. The findings from this approach can assist policymakers in developing targeted strategies to reduce the environmental impact of economic and population growth.

To ensure a thorough analysis, the study focuses on countries that are major contributors to global CO_2 emissions. By uncovering patterns and trends in these regions, the research seeks to provide insights that can inform policy decisions and contribute to global initiatives aimed at reducing emissions.

The paper is organized as follows: Section II provides an overview of existing research on CO_2 emissions and predictive modeling techniques. Section III details the methodology, including data collection, feature selection, and model construction. Section IV presents the results and discusses their implications. Finally, Section V summarizes the key findings and offers recommendations for future research.

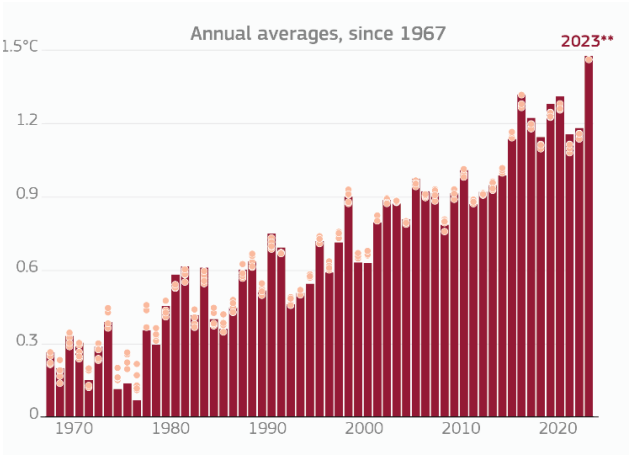


Figure. 1. Global Surface Temperature.

II. LITERATURE REVIEW

This section reviews the current state of the art and highlights the most recent achievements, methodologies, and findings in the field of Predictive Modeling of CO2 Emissions:

A. CO2 Emissions and Climate Change

Climate change is considered a global challenge that is shared by all countries without exception, especially since its main cause is greenhouse gas emissions, especially CO2. Global reports, especially those issued by the Intergovernmental Panel on Climate Change (IPCC), clearly indicate the decisive role played by carbon dioxide emissions in aggravating climate change [1].

978-9950-8599-0-6

This increase in atmospheric carbon dioxide levels is mainly due to human behavior, such as burning fossil fuels, industrial activities, and significant dependence on industrial production [2].

B. Economic Impact on CO2 Emissions

There is a close relationship between economic growth, industries, and carbon dioxide emissions. Studies on the Environmental Kuznets Curve (EKC) hypothesis [3] indicate that with the growth of the economic level, environmental pollution increases until the level of per capita income reaches a certain limit, then environmental quality improves, which has been widely and repeatedly proven, as it was found. Studies [4] indicate that high-income countries can achieve better environmental standards due to the availability of appropriate and advanced technologies and the presence of more stringent environmental policies.

C. Population Impact

Population growth is another critical factor affecting overall climate change through CO2 emissions. Ehrlich and Holdren (1971) presented the equation that calculates the environmental impact (I) of humans on the environment as a product of three factors: population (P), affluence (A), and technology (T) [5] gives a strong indication of the three main factors in increasing climate-affecting gas emissions which are population, affluence (GDP per capita), and technology. Studies have explained that an increasing population leads to increased demand for resources and energy, leading to higher carbon dioxide emissions.

(Impact = Population x Affluence x Technology), which

D. Data Visualization and Environmental Studies

Data visualization is crucial in the process of gaining insight into data, especially big data. Studies emphasize that visual data makes the data easier, more useful, accessible, and more informative for legislators and decision-makers.

Visualization reveals trends, patterns, and outliers in the data that are not apparent in the raw data.

Modern visualizations have enhanced the ability to explore data in real time and have significantly integrated data, facilitating the decision-making process [6].

Research [7] utilized exploratory data analysis (EDA) and visualization techniques to uncover patterns in diabetes prevalence among underprivileged communities, which demonstrated how data visualization enhances interpretability and aids in policy formulation, and highlights the role of EDA and visualization in understanding data-driven problems.

[8] proposed an open government data (OGD) framework that enhances data accessibility and transparency to support policymaking. The study highlights how leveraging publicly available datasets can contribute to data-driven decision-making in various domains, including environmental sustainability. In the context of CO_2 emissions, open data sources such as the World Bank and climate monitoring agencies provide valuable insights that enable the development of predictive models for understanding the environmental impact of economic and demographic factors.

E. Predictive Modeling in CO_2 Emission Analysis

Machine learning (ML) techniques, such as Random Forest, Support Vector Machines (SVM), and Deep Learning models, have been increasingly used to uncover complex relationships between variables and improve the accuracy of emission forecasts. For instance, [Author et al., 2023] explored the interplay between solar and wind energy production, coal consumption, GDP, and CO_2 emissions using ML-based predictive models. Their study demonstrated how advanced algorithms can identify non-linear dependencies among these variables, offering data-driven insights into the impact of energy transitions on carbon emissions. The findings suggest that integrating renewable energy data into predictive models enhances the precision of emission forecasts, which is crucial for designing sustainable policies and promoting green energy adoption.

Similarly, in the context of this research, machine learning techniques are employed to predict CO_2 emissions using GDP and population data. By leveraging Random Forest regression, this study aims to improve the accuracy of emission estimates, enabling policymakers to anticipate future trends and implement effective mitigation strategies.

III. METHODOLOGY

This study follows a structured approach to analyzing CO_2 emissions based on economic and demographic factors Figure. 2 . Unlike many previous works that examine all countries globally, this research focuses on a carefully selected set of high-emission nations with contrasting GDP and population characteristics. This targeted scope enables the model to capture patterns that might be hidden in aggregated global datasets, producing insights that are more directly applicable to policymaking in similar national contexts.

The methodology consists of five main steps: data collection, where relevant datasets were sourced from the World Bank and Climate Watch; data cleaning and preparation, involving filtering, handling missing values, and transforming data for analysis; (EDA) to uncover trends and relationships between CO_2 emissions, GDP, and population; building the predictive model using Random Forest regression; and finally, model evaluation, where performance metrics such as Mean Squared Error (MSE) and R-squared (R^2) were used to assess accuracy

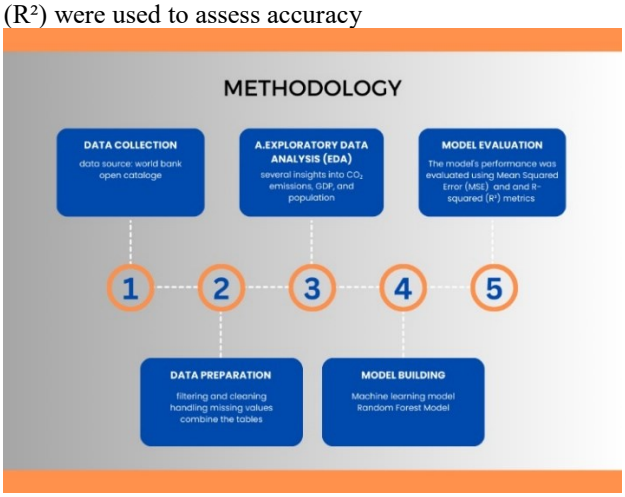


Figure. 2. Methodology Flowchart.

A. Dataset Collection

The dataset used in this study was obtained from the World Bank and Climate Watch, covering key economic and demographic factors influencing CO_2 emissions.

Data Sources:

- CO_2 Emissions per Capita: Data sourced from Climate Watch Historical GHG Emissions (1990-2020), accessed in 2023.
- GDP (current US$): Retrieved from the World Bank national accounts database (latest available version, 2023).
- Population: Extracted from World Bank reports (2023).

This study focuses on the following 10 countries due to their high CO_2 emissions and economic impact: Qatar, Bahrain, Brunei Darussalam, Kuwait, United Arab Emirates, Oman, Australia, Saudi Arabia, Canada, and North America.

B. Data cleaning and pre-processing

1) Loading Data

The datasets for CO2 emissions, GDP, and population were loaded to start the processing.

2) Filtering and Cleaning

The data was filtered to include only the selected countries based on the contributions to global CO2 emissions in 2022: Qatar, Bahrain, Brunei Darussalam, Kuwait, United Arab Emirates, Oman, Australia, Saudi Arabia, Canada, and North America.

Columns not relevant to the analysis were removed.

3) Handling Missing Values

Years with missing values were dropped to ensure the integrity of the analysis. So, the study considers the years from 1990-2020

4) Melt the data and combine the tables

The data sets of CO2 emissions and populations, and GDP were transformed into a pivot table then combined into a single, long-format table for further analysis Figure 3.

	Country Name	Indicator Name	Year	Value
0	United Arab Emirates	CO2 emissions (metric tons per capita)	1990	29.055796
1	Australia	CO2 emissions (metric tons per capita)	1990	15.437183
2	Bahrain	CO2 emissions (metric tons per capita)	1990	20.752003
3	Brunei Darussalam	CO2 emissions (metric tons per capita)	1990	12.447314
4	Canada	CO2 emissions (metric tons per capita)	1990	15.148969

Figure 3. Snapshot from the combined long-format dataset for further analysis.

C. Exploratory Data Analysis (EDA)

The EDA provides several insights into CO_2 emissions, GDP, and population trends over time. These insights help in understanding economic influence and the contribution of population growth to CO2 emissions.

1) Trends in CO_2 Emissions Over Time

Figure 4 presents the time-series trends in CO_2 emissions per capita for the selected countries from 1990 to 2020. The data, sourced from Climate Watch, show that Qatar, Bahrain, and the UAE consistently record the highest CO_2 emissions per capita due to their dependence on fossil fuel–based industries. In contrast, Brunei Darussalam and Oman had lower emissions but showed a gradual increase over time. Meanwhile, North America, including Canada, maintained consistently high emission levels, though with a more stable trend compared to other regions.

A noticeable decline in CO_2 emissions per capita is observed during the 2008 (global financial crisis) and the 2019 (COVID-19 pandemic), reflecting the impact of economic slowdowns on industrial activity and energy consumption.

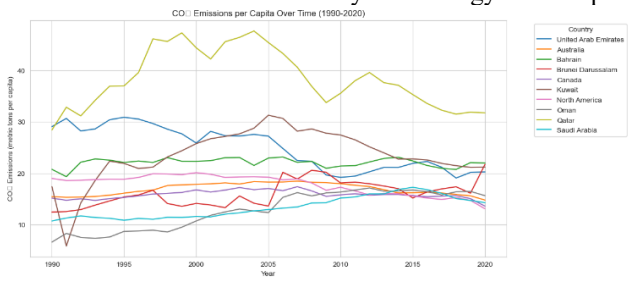


Figure 4. Trends in CO_2 Emissions Over Time.

2) GDP Growth and Its Relationship with CO_2 Emissions

Figure 5 illustrates the relationship between GDP and CO_2 emissions per capita from 1990 to 2020. showed that Countries with higher GDP, like Australia, Canada, and the UAE, showed a strong link to higher CO_2 emissions per capita. Qatar had very high GDP and emissions, highlighting the impact of economic growth. Meanwhile, Brunei Darussalam and Oman, with lower GDP levels, had lower emissions. These trends align with the Environmental Kuznets Curve (EKC) hypothesis [3] indicates that with the growth of the economic level, environmental pollution increases until the level of per capita income reaches a certain limit, then environmental quality improves.

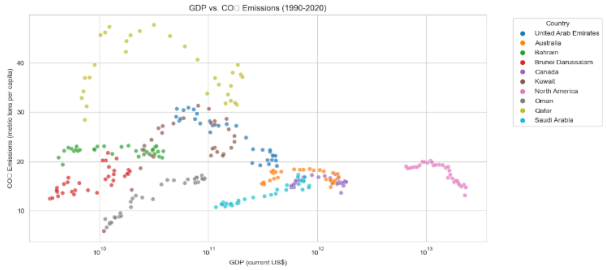


Figure III5. GDP vs CO_2 Emissions.

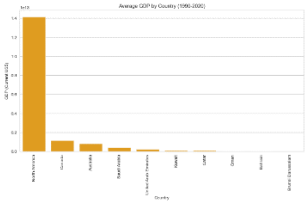


Figure 6 Average GDP per Country.

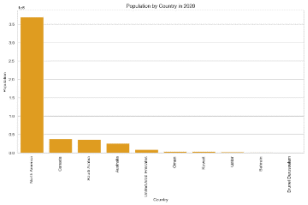


Figure 7 Population by country in 2020.

3) population Growth and Its Relationship with CO_2 Emissions

Figure 7 displays the population sizes of the selected countries in 2020, based on World Bank data. While total emissions tend to scale with population, Figure 8 shows that high per capita emissions are not always associated with large populations. For example, Qatar and the UAE, despite smaller populations, still produce disproportionately high per capita emissions due to energy-intensive economies. Brunei Darussalam and Bahrain saw gradual increases in both population and emissions, suggesting a proportional rise. This analysis indicates that industrial activity influences emissions more than population size alone.

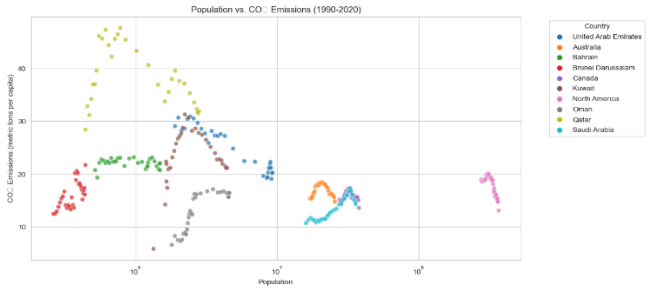


Figure. 8. Population vs CO_2 Emissions.

4) CO_2 Emissions Distribution

Figure 9 shows the distribution of CO_2 emissions per capita between 1990 and 2020. Qatar exhibits the highest variability, reflecting fluctuations caused by economic cycles and industrial changes. Bahrain, the UAE, and Australia maintain high but stable emissions, while Brunei Darussalam and Oman remain at consistently low levels. These trends suggest that economic conditions, policy changes, and industrial expansion play a crucial role in shaping CO_2 emissions levels.

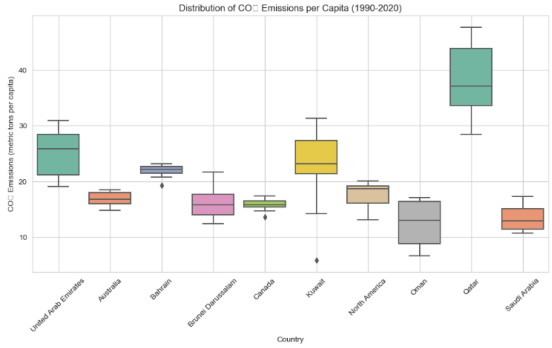


Figure 9. Distribution of CO_2 Emissions per capita.

5) GDP vs. Population

Figure 10 plots GDP against population size using a log scale. It reveals that economic strength is not necessarily tied to population, with small-population countries like Qatar and the UAE maintaining high GDP levels. Australia and Canada, with larger populations, exhibit steady GDP growth, indicating the role of economic diversity in sustainability. Meanwhile, Brunei Darussalam and Oman have both low GDP and population, reflecting gradual economic expansion.

Figure 10. GDP vs. Population.

IV. Feature Engineering:

Considering that we have different countries with high differences between GDP and populations, log transformations to GDP and Population were applied, also the GDP per capita was used. These feature engineering techniques were used to improve the predictive accuracy of the model. By combining GDP per capita with log-transformed GDP and population, the model can account for both proportional and absolute effects, offering a richer interpretation of how economic and demographic drivers affect CO_2 emissions.

1) GDP per Capita

To account for economic output per person, GDP per capita was added and calculated using the formula:

GDP per Capita = GDP\Population

This feature helps compare countries with different population sizes more accurately. It better explains how economic activity per capita influences CO_2 emissions.

2) Log Transformations

As noted in **Error! Reference source not found.** and **Error! Reference source not found.** GDP and population have highly skewed distributions, where a few countries have huge values compared to others. To handle this, logarithmic transformations were applied to normalize the data and reduce skewness.

Log GDP = log (GDP)

Log Population = log (population)

V. Model Building

The ML model was built to predict the variations in CO2 emission for the coming years.

1) Selecting the Model

As shown in Figure III5 and Figure. 8, there is a non-linear relationship between GDP, Population, and CO2 emissions, so the Random Forest (RF) model was chosen for the complexity of the relationship. Traditional regression models, often used in similar studies, assume a strictly linear relationship between variables, which does not reflect the complexity observed in the data. The Random Forest algorithm was chosen for its robustness to multicollinearity and ability to model complex, non-linear interactions between GDP, population, and emissions.

RF is well-suited for handling complex interactions between features and provides robust predictions by averaging

multiple decision trees. Additionally, XGBoost, a more advanced gradient boosting model, was tested to compare performance.

- Target variable: CO2 emissions.
- Independent variables: GDP and population.

2) *Data Preparation*

The dataset was divided into 80% training data and 20% testing data to evaluate model performance, The model parameters were set to default values initially, with 100 trees in the forest.

3) *Hyperparameter Tuning*

To optimize the model, a GridSearchCV approach was used to fine-tune key parameters.

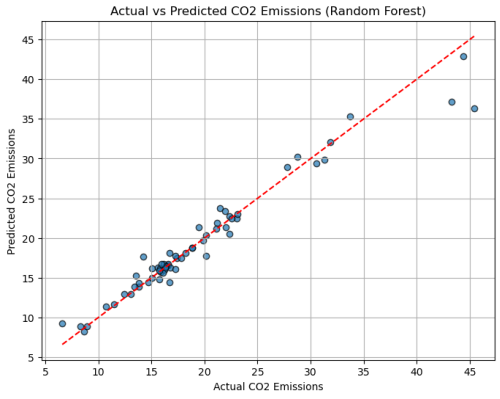
VI. Model Evaluation

The model performance results indicate that the Random Forest Regressor outperformed XGBoost in predicting CO_2 emissions based on GDP and population data. Figure 11 compares the performance of the Random Forest and XGBoost models in predicting CO_2 emissions.

	Model	MSE	RMSE	R² Score
0	Random Forest	3.077210	1.754198	0.950746
1	XGBoost	5.094041	2.256998	0.918465

Figure. 11. Model Performance Evaluation.

- Mean Squared Error (MSE): Random Forest (3.08) has a lower error than XGBoost (5.09), meaning its predictions are closer to actual values.
- Root Mean Squared Error (RMSE): Random Forest (1.75) has a smaller RMSE, indicating smaller deviations between actual and predicted CO_2 emissions.
- R^2 Score: Random Forest achieved 0.95, meaning it explains 95% of the variance in CO_2 emissions, while XGBoost explains 91.8%.



h

Figure. 12. Actual vs Predicted CO2 Emissions (Random Forest).

Additionally, a feature importance analysis was conducted using the Random Forest model to assess the relative contribution of each variable (GDP, population, GDP per capita, log GDP, and log population) to CO_2 emissions prediction. This step provided insights into which factors exert the greatest influence on model outputs, supporting a more interpretable and policy-relevant analysis.

VII. Outcomes and Results

The results show that the RF model performed better than the XGBoost model in predicting CO_2 emissions per capita. The Random Forest model achieved a high R^2 score of 0.95, meaning it explains 95% of the variation in CO_2 emissions per capita based on GDP and population data. It also had a low RMSE of 1.75, indicating that its predictions are close to the actual values.

The XGBoost model, while still effective, had a slightly lower R^2 score of 0.92 and a higher RMSE of 2.26, meaning its predictions were not as accurate as those from the Random Forest model. A plot in the Figure. 12 compares actual vs. predicted CO_2 emissions per capita, confirming that the Random Forest model's predictions closely follow the real values. These results highlight that GDP and population have a strong impact on CO_2 emissions. Countries with higher GDP tend to have higher emissions, but the relationship is not strictly linear, which is why a non-linear model like RF performed well.

As shown in Figure the feature importance results indicate that GDP per capita is the strongest predictor of CO_2 emissions, followed by log population and population size. GDP and log GDP contributed less to the model's predictions. This suggests that emissions are more closely linked to the intensity of economic activity per person and population scale than to total GDP alone.

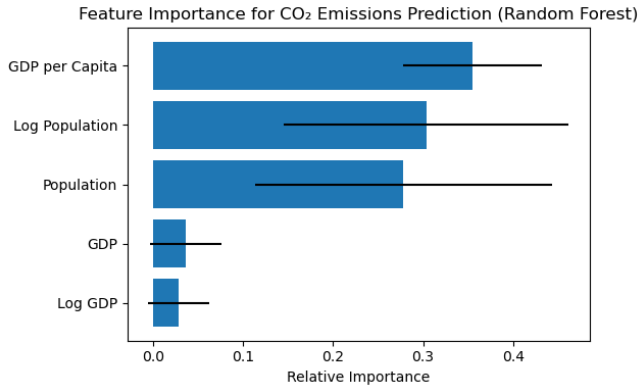


Figure 13 Feature Importance for CO_2 Emissions Prediction (Random Forest)

VIII. Discussion and Suggestions

There is a clear relationship between CO2 emissions and the size of the domestic product and population, which indicates that the climate is greatly influenced by human behavior. It has become clear that countries with a remarkably high GDP

can reduce carbon dioxide emissions by establishing strict legislation and using better technological means.
Developing countries may witness a rapid increase in emissions due to industrialization and the lack of strict regulations. Decision makers must avoid this effect by investing in green technology and alternative energy and increasing public awareness of the importance of considering the problem of climate change, whether through individual or collective practices, in addition to increasing international cooperation and transferring Knowledge and practices that have a positive impact from countries that have recovered from the emissions problem to countries that are still facing the problem.
The methodology developed in this study can be applied as a rapid-assessment tool for policymakers. By focusing on high-emission, economically diverse countries, the model can simulate potential emissions outcomes under varying economic growth or population change scenarios. This enables decision-makers to prioritize interventions — such as carbon pricing, industrial efficiency measures, or technology adoption — in contexts where emissions are most responsive to economic factors.

IX. Conclusions & Future Work

This study confirms the relationship between increasing population, rising income, and higher CO_2 emissions, emphasizing the role of machine learning models in enabling decision-makers to take actions toward emission reduction. It highlights the necessity of analyzing environmental data to ensure informed decisions that effectively mitigate climate change. Future research should expand the study to include more countries for a broader understanding of global CO_2 emission trends and consider additional influencing variables such as industrial and energy consumption. Moreover, exploring more advanced models beyond Random Forest and XGBoost, such as Support Vector Machines (SVM), and deep learning models could enhance prediction accuracy and further support data-driven policymaking.

References

[1] Pachauri, Rajendra K, and Allen, Myles R and Barros, Vicente R and Broome, John and Cramer, Wolfgang and Christ, Renate and Church, John A and Clarke, Leon and Dahe, Qin and Dasgupta, Purnamita and others, Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change, Ipcc, 2014.

[2] United Nation, "www.un.org".*What Is Climate Change?.*

[3] Mahmood, Haider and Furqan, Maham and Hassan, Muhammad Shahid and Rej, Soumen, "The environmental Kuznets Curve (EKC) hypothesis in China: A review," *Sustainability,* vol. 15, 2023.

[4] Dinda, Soumyananda, "Does environment link to economic growth?," *Human Security and Climate Change, mai,* 2005.

[5] York, Richard and Rosa, Eugene A and Dietz, Thomas, "STIRPAT, IPAT and ImPACT: analytic tools for unpacking the driving forces of environmental impacts},," *Ecological economics,* 2003.

[6] Keim, Daniel A and Mansmann, Florian and Thomas, Jim, "Visual analytics: how much visualization and how much analytics?," *Acm Sigkdd Explorations Newsletter,* 2010.

[7] Owda, Majdi and Owda, Amani Yousef and Fasli, Maria, "An Exploratory Data Analysis and Visualizations of Underprivileged Communities Diabetes Dataset for Public Good," *IEEE,* pp. 581--585, 2023.

[8] Fasli, Maria and Owda, Amani Yousef and Abbasi, Tufail and Owda, Majdi and Stergioulas, Lampros and Neupane, Bhanu, "Open Government Data (OGD) Framework for Sustainable Development," *IEEE,* 2023.

[9] Hannah Ritchie and Max Roser, "CO_2 emissions," *OurWorldInData.org,* 2020.

[10] T. Munzner, Visualization analysis and design, CRC press, 2014.

[11] J. Holdren, "A brief history of IPAT," *the journal of population and sustainability,* 2018.

[12] Change, The Intergovernmental Panel on Climate, "The Intergovernmental Panel on Climate Change," [Online]. Available: https://www.ipcc.ch/.

[13] "The Copernicus Climate Change Service," 2023. [Online]. Available: https://climate.copernicus.eu/.

[14] emissions, Biophysical and economic limits to negative CO2, Smith, Pete and Davis, Steven J and Creutzig, Felix and Fuss, Sabine and Minx, Jan and Gabrielle, Benoit and Kato, Etsushi and Jackson, Robert B and Cowie, Annette and Kriegler, Elmar and others, Nature Publishing Group UK London, 2016.

[15] Friedlingstein, Pierre and O'sullivan, Michael and Jones, Matthew W and Andrew, Robbie M and Hauck, Judith and Olsen, Are and Peters, Glen P and Peters, Wouter and Pongratz, Julia and Sitch, Stephen and others, "Global carbon budget 2020," *Earth System Science Data Discussions,* 2020.

[16] Hannah Ritchie, Pablo Rosado and Max Roser, "CO_2 and Greenhouse Gas Emissions," *Our World in Data,* 2023.

[17] Keim, Daniel A and Mansmann, Florian and Thomas, Jim, "Visual analytics: how much visualization and how much analytics?," *Acm Sigkdd Explorations Newsletter,* 2010.

[18] Roser, Hannah Ritchie and Max, "Our World in Data," 2020. [Online]. Available: https://ourworldindata.org/co2-emissions.

[19] World Bank Open Data, "Population, total," 2022. [Online]. Available: https://data.worldbank.org/indicator/SP.POP.TOTL.

[20] Wirkd Bank Open Data, "CO2 emissions (metric tons per capita)," 2023. [Online]. Available: https://data.worldbank.org/indicator/EN.ATM.CO2E.PC.

[21] World Bank Open Data, "GDP (current US$)," 2023. [Online]. Available: https://data.worldbank.org/indicator/NY.GDP.MKTP.CD.