

School of Built Environment, Engineering and Computing

Leeds Beckett University

**Predictive Model for Net-Zero Emissions: A Data-Driven Approach to U.S. and Global Carbon Forecasting**

By: Tushar Manral

Supervisor Name: Mekala Soosay

Submitted to Leeds Beckett University in partial fulfilment of the requirements for the degree of MSc Data Science

**September 2024**

AI chatbot was not used in this study.

# Candidate’s Declaration

I, Tushar Manral confirm that this dissertation and the work presented in it are my own achievement.

Where I have consulted the published work of others this is always clearly attributed;

Where I have quoted from the work of others the source is always given. With the exception of such quotations this dissertation is entirely my own work;

I have acknowledged all main sources of help;

I have read and understand the penalties associated with Academic Misconduct.

Signed: Tushar Manral

Date: 06/09/2024

Student ID No: 77361020

# Abstract

This research developed a forecasting algorithm to predict future CO2 emissions across the USA. To determine the key factors impacting carbon emissions, the study examined past data on energy use, technological advancements, and legislative Laws. To improve the model’s accuracy and efficiency, feature engineering and hyperparameter optimisation were combined with machine learning approaches including Random Forest, XGBoost, ARIMA, and SARIMA. Among the models, Random Forest was the best model that forecasted CO2 emissions with the highest accuracy as compared to other models with Low Mean Absolute (MAPE) and Root Mean Squared Error (RMSE). Moreover, Random Forest was further tested on the INDIA CO2 emission dataset, demonstrating its flexibility to multiple national settings with varying patterns of energy use, economic structures, and regulatory frameworks. The replication’s success highlighted the model’s adaptability and potential for wider use in nations with different economic and energy uses.

By combining past data with complex predictive techniques, policymakers, business executives, and environmental scientists may all benefit greatly from the research’s important findings. The model provided an effective, data-driven tool for projecting future emissions and aided in the creation of stronger plans and schemes to fulfil the environmental objectives. Furthermore, the study emphasised the model’s importance for global carbon emission mitigation efforts, providing a workable option for countries aiming to meet their Net-zero targets. In the end, by offering useful information and directing better-informed decisions on emissions reduction, our study eventually promoted a sustainable, carbon-free future.

Table of Contents

[Candidate’s Declaration ii](#_Toc176528005)

[Acknowledgements iii](#_Toc176528006)

[Abstract iv](#_Toc176528007)

[List of Figures viii](#_Toc176528008)

[List of Tables viii](#_Toc176528009)

[Equation 10](#_Toc176528010)

[Chapter 1 : Introduction 11](#_Toc176528011)

[1.1 Overview 11](#_Toc176528012)

[1.2 Rational 12](#_Toc176528013)

[1.3 Aim and Objectives 13](#_Toc176528014)

[1.4 Outline 15](#_Toc176528015)

[Chapter 2: Literature Review 16](#_Toc176528016)

[2.1 Global Carbon emissions and their impacts 16](#_Toc176528017)

[2.2 US Carbon Emissions 16](#_Toc176528018)

[2.3. Technology Development 16](#_Toc176528019)

[2.3.1 Coal background in the USA and impacts 16](#_Toc176528020)

[2.3.2. Automobiles revolution in the USA and its impacts 17](#_Toc176528021)

[2.3.3. Petrochemicals industry in the USA and its impacts 17](#_Toc176528022)

[2.3.4 Energy Production and its Environmental Impacts in the USA 17](#_Toc176528023)

[2.3.5 Aviation and its Environmental Impacts in the USA 18](#_Toc176528024)

[2.3.6 Recent Technological Innovations and Sustainable Energy Initiatives in the USA 18](#_Toc176528025)

[2.4 Legislative Laws to Reduce Carbon Emissions 19](#_Toc176528026)

[2.4.1 Clean Air Act Amendments (CAAA) 19](#_Toc176528027)

[2.4.2 Corporate Average Fuel Economy (CAFF) 19](#_Toc176528028)

[2.4.3 Renewable Fuel Standard (RFS) 19](#_Toc176528029)

[2.4.4 American Recovery and Reinvestment Act (ARRA) 19](#_Toc176528030)

[3.1 Data Augmentation Techniques for CO2 emissions classification 20](#_Toc176528031)

[4.1 Key Insights and Transition 22](#_Toc176528032)

[Chapter 3: Methodology 23](#_Toc176528033)

[3.1 Technical Environment and Tools 23](#_Toc176528034)

[3.2 Business Understanding 23](#_Toc176528035)

[3.2 Data Understanding 23](#_Toc176528036)

[3.3 Data Preparation 24](#_Toc176528037)

[3.3.1 EDA (Exploratory Data Analysis) 24](#_Toc176528038)

[3.4 Feature Engineering 24](#_Toc176528039)

[3.5 Modeling 25](#_Toc176528040)

[3.5 Evaluating and Interpretation 25](#_Toc176528041)

[3.5 Deployment and Practical Application 25](#_Toc176528042)

[3.5 Replication of Predictive Model for Cross-National Application 26](#_Toc176528043)

[Chapter 4: Product/Research Design and Implementation 27](#_Toc176528044)

[4.1 Data Collection and Sources 27](#_Toc176528045)

[4.1.1 CO2 emission data 27](#_Toc176528046)

[4.1.2 Importance of Reliable Data 27](#_Toc176528047)

[4.1.3 Data Dictionary and Sample 27](#_Toc176528048)

[4.1.4 Data Description 29](#_Toc176528049)

[4.2 Ethical Considerations 30](#_Toc176528050)

[4.3. Data Pre-processing (ETL) 31](#_Toc176528051)

[4.3.1 Carbon emission data 31](#_Toc176528052)

[4.4 Exploratory Data Analysis (EDA) 34](#_Toc176528053)

[4.4.1 Understanding the Trends in Carbon Emissions 34](#_Toc176528054)

[4.4.2 CO2 Emissions Per Capita and GDP Over Time 34](#_Toc176528055)

[4.4.3 Cumulative CO2 Emissions Over Time 35](#_Toc176528056)

[4.4.4 Per Capita CO2 Emissions by Source Over Time 36](#_Toc176528057)

[4.4.4 CO2 Emission and 5-Year Moving Average 36](#_Toc176528058)

[4.4.5 Correlation Matrix with the Use of Heat Map 37](#_Toc176528059)

[4.4.5 Distribution of Cumulative CO2 Emissions by Source 38](#_Toc176528060)

[4.4.5 Temperature Change from CO2 Over Time 38](#_Toc176528061)

[4.5 Featuring Engineering 39](#_Toc176528062)

[4.6 Model Selection and Development for CO2 Emissions 39](#_Toc176528063)

[4.6.1 Overview of Considering Models 40](#_Toc176528064)

[4.6.2 Hyperparameters for Models 42](#_Toc176528065)

[4.6.3 Performance Matrix for Models 43](#_Toc176528066)

[4.7 Replication of Predictive Model for Cross-National 44](#_Toc176528067)

[4.7.1 Data Pre-processing 44](#_Toc176528068)

[4.7.2 Feature Engineering 44](#_Toc176528069)

[4.7.3 Model Implementation 44](#_Toc176528070)

[4.7.4 Model Training and Evaluation 44](#_Toc176528071)

[4.7.5 Interpretation and Reporting 45](#_Toc176528072)

[Chapter 5: Research Outcomes/Results/Discussion and Evaluation 46](#_Toc176528073)

[5.1 Results 46](#_Toc176528074)

[5.1.1 USA Carbon Emission Model Results 46](#_Toc176528075)

[5.1.2 INDIA Carbon Emission Model Results 47](#_Toc176528076)

[5.2 Discussion 48](#_Toc176528077)

[Chapter 6: Project Management 50](#_Toc176528078)

[6.1 ex 50](#_Toc176528079)

[Chapter 7: Conclusion and Future Work 51](#_Toc176528080)

[7.1 Conclusion 51](#_Toc176528081)

[7.2 Future Work 51](#_Toc176528082)

[References 52](#_Toc176528083)

[Appendices 67](#_Toc176528084)

[Code for ETL 67](#_Toc176528085)

[Code for EDA (Exploratory Data Analysis) 69](#_Toc176528086)

[Predictive Model Code for Random Forest 72](#_Toc176528087)

[Predictive Model Code for XGBoost 74](#_Toc176528088)

[Predictive Model Code for SARIMA 76](#_Toc176528089)

[Replicated Model on INDIA dataset. 78](#_Toc176528090)

[ETL CODE 78](#_Toc176528091)

[EDA (Exploratory Data Analysis) 79](#_Toc176528092)

[Predictive model Code for INDIA dataset 81](#_Toc176528093)

# List of Figures

[Figure 1.- Histogram for the missing values in carbon emissions Dataset 32](#_Toc176543532)

[Figure 2.- Trends in Carbon Emissions by Energy Sources. 33](#_Toc176543533)

[Figure 3.- CO2 Emissions Per Capita and GDP Over Time. 34](#_Toc176543534)

[Figure 4.- Cumulative CO2 emissions by Sources 34](#_Toc176543535)

[Figure 5.- CO2 emission Per Capita by Energy Source 35](#_Toc176543536)

[Figure 6.- CO2 emission with 5-year Moving Average 36](#_Toc176543537)

[Figure 7.- Correlation Matrix of Energy-Related Columns 36](#_Toc176543538)

[Figure 8.- Distribution Of cumulative CO2 Emissions by Source 37](#_Toc176543539)

[Figure 9.- Temperature Change from CO2 Over Time 38](#_Toc176543540)

[Figure 10.- Gantt Chart for Project Management 49](#_Toc176543541)

# List of Tables

[Table 1.- Performance Evaluation of Various Machine Learning Models for CO2 Emissions 20](#_Toc176543511)

[Table 2.- Data Dictionary of CO2 Emission 27](#_Toc176543512)

[Table 3.- Overview of the cleaned Dataset for CO2 Emissions 32](#_Toc176543513)

[Table 4.- performance Matrix of Random Forest Model on USA CO2 Dataset 46](#_Toc176543514)

[Table 5.- performance Matrix of Random Forest Model on INDIA CO2 Dataset 46](#_Toc176543515)

# Equation

[Equation 1.- ARIMA Formula 38](#_Toc176516518)

[Equation 2.- SARIMA Formula 39](#_Toc176516519)

[Equation 3.- Random Forest Formula 39](#_Toc176516520)

[Equation 4.- XGBoost Formula 40](#_Toc176516521)

[Equation 5.- Root Mean Square Root Formula 41](#_Toc176516522)

[Equation 6.- Mean Absolute Percentage Error Formula 41](#_Toc176516523)

# Chapter 1: Introduction

## 

## 1.1 Overview

The rise of greenhouse gases leads to major climate change disturbances that have damaged ecosystems and human society worldwide, making climate change an urgent and major problem. The Intergovernmental Panel on Climate (IPCC) has evaluated the current climate change, which has linked human actions to climate change across the globe (Wang, H. et al., 2022). The impacts of this human-caused climate are manifold, such as seawater levels, unpredictable weather, and concerns regarding the ecosystems and public health (Nolan et al., 2018). To decrease the effects of climate change, the IPCC emphasises the need to achieve zero carbon emissions and eliminate other harmful greenhouse gases from the environment. Many investigations have shown that climate change is not limited to the ecological effects. Still, it has also directly influenced human health through a rise in the intensity and frequency of extreme weather events and indirectly through effects on food nutrition, air quality, disease transmission and mental health (Wang, H. et al., 2022).

It is crucial for countries, especially developed nations like the United States which have significant historical and present impact polluters, to act swiftly to fulfil the need to achieve net-zero emissions globally and mitigate the effects of climate change (Soest, van et al., 2021). Being a major source of greenhouse gas emissions, the USA is essential in spearheading the shift to a sustainable future. To lead international initiatives to tackle climate change and establish an example for other countries to follow, the USA should set aggressive net-zero emission objectives that set effective legislation and regulations (Höhne et al., 2021). Net-zero emissions must be achieved to stop the consecutive rise in global temperatures and lessen the negative effects of climate change on the climate change environment, humans, and the world as a whole (Allen, M. R. et al., 2022). The United States have some obstacles to achieving net-zero emissions as it has extensive reliance on fossil fuels in many different areas of the economy (Hutton et al., 2023). A comprehensive and carried approach is needed to achieve net zero emissions, as this dependence presents significant obstacles to decarbonisation efforts (Kafi et al., 2023). The growing demand for cutting-edge technology for reducing the carbon footprint further complicates the shift, further obstacles to reaching net zero goals include the lengthy lives of current infrastructure and the absence of public support for rapid improvement (Edelenbosch et al., 2023). A comprehensive approach is needed to properly address these issues, including higher emission regulations, carbon reduction campaigns, efficient emissions reduction techniques, and research on carbon capture and utilisation technology (Xue, 2024).

The USA’s transition towards net zero emissions effectively faces difficulties but predictive models can help the USA to predict CO2 emissions which can help them to control emissions and make legislative laws to reduce the carbon emissions. Models like Gradient Boosting Machines (GBM), time series models, Random Forest models and other models offer insightful projections of future conditions (Olanrewaju et al., 2024). Policymakers may simulate other possibilities and assess the possible impacts of various actions on emission reduction and climate change mitigation by implementing complex modelling methods (Zakkour et al., 2021). To accelerate the shift towards the low carbon economy, predictive models assist in determining the carbon emissions from different fossil fuel sources and then the best technological advancement and legislative initiatives can be done (Hutton et al., 2023). Policymakers can record the progress of carbon emission reduction targets, improve tactics, and distribute funding by including predictive modelling in their policy formation effort (Khalifa et al., 2022).

Therefore, this research on predictive modelling and data-driven decision-making can predict carbon emissions from energy sources which has the potential to be extremely helpful for making a new policy and strategic efforts in the direction of reaching net-zero emissions (Lepri et al., 2017). Authorities can build an efficient policy for emission reduction by applying the latest analytics and modelling approaches to provide practical insights into energy consumption patterns, technology improvement, and regulatory factors (Olanrewaju et al., 2024). With the help of this study, policymakers will be enabled to distribute finance more efficiently, make informed choices, and encourage major changes that will contribute to a low-carbon, environmentally friendly future (DeCarolis et al., 2017). Policymakers may effectively navigate the obstacles of transition towards net-zero emissions through the implementation of data-driven tactics, which will help in the worldwide battle against climate change and create a more sustainable future for existing human beings and future generations.

## 1.2 Rational

The current situation of the environment is concerning as climate change has been the most common situation all over the world and to confront climate change and accelerate the global effort to achieve net-zero emissions (Liu, H. et al., 2023). The life-threatening impact of greenhouse gas emissions not only on ecosystems but also the human societies throughout the world, such as climate change, rising sea levels, food-related illnesses, Lyme disease and heart diseases (Zhong, S. and Huang, 2019). Considering the consequences of climate change, this shows the critical need for global cooperation to reduce the negative consequences of carbon emissions (Clements, 2021). Moving towards net zero emissions is very important as the Paris Agreement, a worldwide pact, aims to keep temperatures of the planet Earth far below 1.5°C (Frieler et al., 2017).

The United States infrastructure is highly dependent on fossil fuels, making the nation’s transition to a net-zero economy difficult and needing multiple strategies to reduce emissions. For lawmakers, industrial leaders, environmental organisations, and the general public, predictive modelling may prove to be a very useful instrument to control the nation’s carbon emissions (Piggot et al., 2018). Predictive modelling forecasting of future emissions is essential for the mitigation of the effects of climate change. Numerous research has investigated various approaches and models to anticipate CO2 emissions and, in this research, different types of machine learning models like Random Forest, Boost, and time series forecasting like ARIMA and SARIMA are used and one with the highest training and testing value model will be the best choice for forecasting CO2. These models can mimic potential futures utilising large datasets to evaluate the success of various activities, including shifts in society as a whole fluctuations in the energy marketplace, and advances in technology, this data-driven approach helps to develop feasible and affordable routes towards net-zero emissions (Williams, J. H. et al., 2021). There is so much research available for predictive models for net zero, but they lack practical insights for policymakers due to insufficient previous policy consequences and technical advancements done in the past. By combining historical data, new laws, and technological advancements into a unified framework, this research seeks to narrow this gap. Compared to the other research, this investigation offers an original approach for projecting routes to net-zero emissions by combining the wide range of historical data into a complex prediction framework (Borowiak et al., 2024).

The results of this research will guide the policymakers with practical and data-driven approaches, which is the more promising transition towards a net zero economy. The regulations that are created will facilitate the United States’ shift to net-zero emissions while being more cost-effective. Using historical analysis of energy use, technology advancements, legislative law and prediction modelling, this investigation produces tactical recommendations that can inform policy choices. The study could have an impact on national policies aimed at achieving net-zero emissions while contributing to global climate campaigns, possibly setting a standard for other nations.

## 1.3 Aim and Objectives

1.3.1 Aim

The primary aim of this research is to create a Predictive model for attaining Net-zero emissions in the United States of America (USA) states by examining energy usage, technological developments, and regulatory factors for building a sustainable future.

1.3.2 Objectives

* To thoroughly examine past data about energy use, technological developments, and legislative policies nationwide to pinpoint patterns and trends essential to achieving Net Zero emissions.
* To evaluate, how well earlier technical advancements and policy changes have impacted carbon emissions in the United States, using lessons learned from the past to guide current and future Net Zero initiatives.
* To develop a predictive model, data will be explored to find a suitable algorithm.
  + To identify the suitable machine learning model for predicting the CO2 emission, ensuring feature engineering and appropriate selection of performance metrics.
  + To implement the selected advanced machine learning models and to optimise the hyperparameters with the help of a literature review.
* To verify if the predictive model can be replicated in other nations and how well-suited it is to various energy consumption patterns, economies, technological development, and regulatory frameworks.

These objectives will help to achieve the goal, the research hopes to offer a thorough grasp of the various aspects that are impacting the US transition to Net-zero emissions. This project intends to provide policymakers, industry stakeholders, and the public with actionable insights and strategic recommendations by analysing historical data, assessing previous initiatives, and constructing a prediction model. By providing practical approaches and routes to attaining net-zero emissions in the USA, the end goal is to open the door to a sustainable future.

1.3.3 Hypothesis

The hypothesis for this study is advanced machine learning models will be able to learn and predict CO2 emissions from different sources. These models could be then replicated and applied to other nations with varying energy consumption, economic, and regulatory frameworks, providing information to aid in the creation of plans and policies for efficient emissions reduction.

## 1.4 Outline

This dissertation is organised as follows:

Chapter 2 examines the literature review, and provides a summary of the development of past technologies which are the reason for the CO2 emissions in the USA and also examines the laws which have been made by the US government but eventually failed and also, research that has been done on national and international CO2 emission models.

Chapter 3 defines the approach applied in this work, which includes the use of complex machine learning models for the CO2 emissions algorithm, with the help of CRIPS-DM methodology.

Chapter 4 delivers the study design and execution, detailed model selection, hyperparameter tweaking, and the application of the models to the data.

Chapter 5 discusses the outcomes of the machine learning models, and compares the models to each other to find the best model and the findings.

Chapter 6 addresses project management, including the planning and execution procedure as well as the difficulties faced and how they were solved in time.

Finally, Chapter 7... This section brings the research to a close with overall results, contributions made to the area, and bits of advice for additional study and real-world applications.

# Chapter 2: Literature Review

## 2.1 Global Carbon emissions and their impacts

Global weather has significantly changed due to the immense increase in CO2 levels. Since 1970, emissions have nearly quadrupled, mainly because of growing deforestation and burning fossil fuels such as coal, oil, and natural gas (Moosdorf et al., 2014). The consequences of these acts have impacted the atmosphere, now it contains more CO2, which worsens global warming (Matthews and Caldeira, 2008). The increasing intensity and frequency of weather-changing and rising sea levels, are the output of this, creating unparalleled environmental chaos to envelop the planet. The major risks these ecological impacts pose to biodiversity and human civilisation have made climate change one of the most critical issues facing the modern world. The enormity and critical nature of this problem highlight the necessity of organised global action to reduce CO2 and prepare for climate change.

## 2.2 US Carbon Emissions

The United States is one of the oldest and highest contributors to global carbon emissions, making up over 80% of its energy consumption and carbon dioxide output (Rifai et al., 2018). This significant carbon footprint has deeply affected both the economy and the ecology. Investigation shows that higher levels of greenhouse gases in the atmosphere are impacting global warming and its related effects and that the US has been an important contributor to climate change and rising water levels (Pandey, D. et al., 2011). The United States must greatly impact its carbon footprint when calculating its contribution to global emissions percentage. Research has shown that the construction sector of the United States has emitted 48% of the carbon footprint emitted alone in the United States (Egilmez et al., 2017). The result has shown that the main source of carbon emissions in the US is the industrial sector. Moreover, rapid urbanisation and people have boosted its carbon impact on the environment (Wang, S. et al., 2014).

The carbon footprint of the United States has major repercussions for the industrial structure, energy demands, and international trade have all been found to have an impact on carbon emissions (Chang et al., 2014). Also, it has been emphasised that to achieve a low-carbon economy, it is crucial to decrease carbon emissions while preserving fertile lands (Pei et al., 2021). Financial factors stress the need for policies that balance between preservation of the environment and economic expansion, which is vital for sustainable development (Cai et al., 2022).

## 2.3. Technology Development

### 2.3.1 Coal background in the USA and impacts

The extensive use of coal was an essential component in the Industrial Revolution, which started in the late 18th century and brought about enormous changes in economic and industrial activities. Steam-powered technology evolved during this time, revolutionizing the industrial and transportation sectors with the advent of mechanised processes that substantially enhanced production (Schurr, 1960). Due to reliance on coal as the primary energy source, these advancements were accompanied by a notable rise in carbon emissions. Particle-filled air caused respiratory issues for local and long-term environmental harm in cities like Atlanta and Chicago, which gained notoriety for this characteristic (Stradling and Thorsheim, 1999). The combustion of coal resulted in significant emissions of sulphur dioxide and particulates, establishing problems associated with the expansion of industry. The complex link between environmental degradation and technology development is evident in this period, which has sparked more conversations about sustainable development.

### 2.3.2. Automobiles revolution in the USA and its impacts

The Rise of Ford’s assembly line in 1913 signalled the beginning of the automotive industry’s accent in the early 1900s (Alizon et al., 2009). Car ownership and oil consumption increased because of this development, resulting in vehicles being easier to obtain. Rising C02 emissions and pollution in cities brought about the widespread use of internal combustion engines and increased health concerns. Along with greater consumption of energy and habitat disturbance, the automobile revolution also fuelled the expansion of highways and cities, growing reliance on fossil fuels and their adverse impact. Immense financial and technological advances were brought by the car manufacturers, yet they also resulted in long-lasting environmental problems. This period sheds light on the complex interaction between ecological consequences and industrial progress, showing the wide-range and often unexpected environmental effects of technological development (McCarthy, 2007).

### 2.3.3. Petrochemicals industry in the USA and its impacts

The chemical industry grew rapidly after World War II, especially in synthetic products like plastics (Lewis, 2007). These innovations were praised for their durability and versatility, providing reasonable alternatives for a range of industries. However, plenty of carbon was used in their creation, which increased greenhouse gas emissions (Zheng, J. and Suh, 2019). Widespread use of plastic has resulted in major ecological issues, such as the build-up of garbage that is not biodegradable and contamination by microplastics. As a result of their resistance to natural decomposition, they have poisoned the environment widely, jeopardising animal and public health as well as affecting aquatic ecosystems (MacLeod, 2021). This past period highlights the fragile link between innovation and sustainable development in industrial growth, demonstrating the way technological improvements can have adverse environmental impacts even while they provide benefits short term.

### 2.3.4 Energy Production and its Environmental Impacts in the USA

The United States energy sector was primarily reliant on fossil fuels, mainly coal, to generate power during the first half of the late 20th century (Campbell et al., 2013). Major reasons for pollution and carbon dioxide emissions into the surroundings were coal-fired power plants. Although natural gas releases less CO2 per unit of energy than other possibilities, it has become more common in the second half of the century. However, the introduction of hydraulic drilling or hydraulic fracturing as a method for extracting natural gas has led to environmental issues such as methane leaks, contamination of groundwater, and earthquakes (Gassiat et al., 2013). In addition, nuclear power was an accessible source of low-carbon energy. Nevertheless, there were risks associated with the elimination of radioactive substances and the risk of catastrophic events.

### 2.3.5 Aviation and its Environmental Impacts in the USA

The mid-20th century saw the rise of the aviation sector as a major means of transportation, which had a profound impact on international trade. The introduction of aircraft engines and the development of commercial aircraft greatly improved the efficiency and speed of transit (Guzanek and Borucka, 2021). Yet this breakthrough came at an essential ecological price. The use of jet fuel releases a significant quantity of CO2 and other greenhouse gases, which contributes to global warming (Tan, C. H. et al., 2024). Furthermore, specific pollutants from aeroplanes, such as nitrogen oxides and contrails, have a greater heating impact than CO2 alone. Due to the excessive climatic impact of these high-altitude emissions, aircraft must be a key area of focus for the worldwide drive to lower carbon footprints (Lund et al., 2017). Therefore, this industry reflects the conflict between sustainable development and technological advancements, changing the challenges associated with striking a balance between sustainable development and technical development.

### 2.3.6 Recent Technological Innovations and Sustainable Energy Initiatives in the USA

In the fight against climate change, the USA has come a long way in recent years in embracing cutting-edge technologies and switching to renewable energy sources. Developing renewable energy sources including hydropower, solar power, and wind power has been an important goal to reduce the dependency on fossil fuels for the production of electricity (Jamil et al., 2024). The automobile industry is one of the important contributors to the emissions, and reducing the level of pollution has been made possible in large part due to the development of electric vehicle (EV) innovation and the setting up of a widespread charging network (Manzella, 2017). Furthermore, a large amount of capital has been invested in Carbon Capture Storage (CSS) innovation, which can preserve 90% of emissions to minimise the Carbon emissions from industrial and electrical energy (Li, C. et al., 2024).

## 2.4 Legislative Laws to Reduce Carbon Emissions

### 2.4.1 Clean Air Act Amendments (CAAA)

Concerns about air pollution and its hazardous impact on the atmosphere and human health have risen, the Laws aimed to prevent acid rain by restricting SO2 emissions from burning fossil fuel units (Srinivasan and Tettamanzi, 1997). However, there were so many concerns in the CAAA, including the inability to properly manage and enforce emissions control, which barred them from fulfilling the emission reduction objectives (Attwood et al., 2014). The influence of the improvement on lowering the pollution levels was restricted since limited enforcement mechanisms made it harder to enforce emission control measures properly. The general efficacy of the CAAA has been influenced by political restrictions and legislative process deadlock that prevented the prompt and thorough execution of emissions regulations (Lynch et al., 2000).

### 2.4.2 Corporate Average Fuel Economy (CAFF)

The Law has been put in place to reduce the amount of petroleum used and greenhouse gas emissions from vehicles, following measures of environmental sustainability and energy consumption (Jenn et al., 2016). It’s possible that the CAFF standards had challenges in controlling and maintaining compliance, which kept them from achieving the targeted levels of fuel efficiency and pollution reduction (Zimmerman et al., 2016). Insufficient regulation may have made the rules more challenging to execute successfully, which could have hampered their potential to substantially boost fuel economy and decline emissions (Zimmerman et al., 2016). Achieving greater energy efficiency and emission reductions due to technical constraints made it challenging to completely accomplish the targets set by the CAFF criteria.

### 2.4.3 Renewable Fuel Standard (RFS)

The RFS program is a federal law that mandates replacing or reducing a specific amount of fossil fuel in household heating, jet fuel and transportation fuel with a particular volume of renewable fuel (US EPA, 2015). It is possible that the RFS had issues entirely regulating and maintaining compliance, which prevented the organisation from reaching its goal of biofuel blending and carbon reductions (Lark et al., 2022). Insufficient enforcement penalties may have delayed the successful execution of the RFS, limiting its capacity to cause notable rises in the production and use of biofuel fields (Lade and Bushnell, 2019). The operation and efficacy of the RFS may have been affected by industry lobbying and economic factors, which influenced the biofuel market and mining required compliance (Lade et al., 2018).

### 2.4.4 American Recovery and Reinvestment Act (ARRA)

Major clean energy provisions intended to advance renewable energy, sustainable development and energy efficiency have been included in the American Recovery and Reinvestment ACT (ARRA) 2009. The main motive of the ARRA clean energy plan was to promote sustainable energy practices, reduce carbon emissions, increase employment, and promote economic growth (Carley, 2016). Nevertheless, challenges to its enacting were weak enforcement, corporate influence, deadlock in politics, and limited technology (Aldy, 2013). Regulatory, financial, political, and technical obstacles are highlighted by these issues, which are common to all clean energy rules (Hall and Jennings, 2011). Initiatives for clean energy are failing as planned due to insufficient regulation, industrial campaigns, industrial campaigns, political resistance, and technology imitations. Governing economic political, and technological hurdles are highlighted by these concerns, which are common to all renewable energy rules. Efforts for clean energy have not advanced as planned due to insufficient regulation, industrial advocacy, political obstruction, and technological constraints (Popp et al., 2020).

## 3.1 Data Augmentation Techniques for CO2 emissions classification

Multiple designs for machine learning are used in CO2 emissions forecasting to increase accuracy and adaptability. Multiple studies focussing on different performance measures have used a variety of strategies to execute their models. Table 1 provides a summary of these studies with a focus on the success metrics attained, hyperparameters utilised, and the machine learning models applied.

Wong, (2022) assessed several machine learning architectures, including Support Vector Machine, Random Forest, Decision Tree, Linear Regression, and Extreme Gradient Boosting without providing hyperparameters. The Random Forest model had the Lowest Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) of 24.42 and 26.26, respectively, while the support Vector Machine model had the most errors (MAE:101.11, RMSE:128.74)). The performance metrics for these models varied.

Tan, (2022) employed several models, such as AdaBost, CatBoost, SVM, Random Forest, XGBoost, LightGBM, and Sequential, without defining a hyperparameter for Random Forest. With an RMSE of 62.84 and an MAE of 35.71, the stated performance metrics for these models were consistent. However, there were minor differences in the Mean Absolute Percentage Error (MAPE) among the models.

The following regression models were examined by Michalakopoulos et al., (2023) Gradient Boosting Regressor, Random Forest Regressor, Extra Trees Regressor, Linear Regression, and the Decision Tree. Having an RMSE of 92.68%, the Random Forest Regressor was the most successful, while the Extra Tress Regressor had the highest EMSE of 93.98%. Li et al. (2023) achieved training R-squared (R2) of 0.970, testing R2 of 0.896, and RMSE of 1.35 with the Random Forest Model.

Regression models such as Support Vector Regressor, K-Nearest Neighbours Regressor, Decision Tree Regressor, Random Forest Regressor, Decision Tree Regressor, Random Forest Regressor, and Artificial Neural Network were implemented (Harati et al., 2024). These models performed extremely well, with testing R2 values ranging from 0-9964 to 1.0000, almost perfect. The Random Forest Regressor (RFR) was used by Fujii et al., (2024) on several datasets, including Single Amine, Simplified Descriptor List, and Single and Blended Amines. The performance measures exhibited consistency, with RMSE values ranging from 0.072 to 0.079 and R2 scores about 0.943.

Table 1.- Performance Evaluation of Various Machine Learning Models for CO2 Emissions

|  |  |  |  |
| --- | --- | --- | --- |
| Author | Machine Learning Architecture | Hyperparameters | Performance metrics |
| (Wong,2022)) | Linear Regression | Not stated | MAE:56.76 RMSE:63.86 |
| Decision Tree | MAE:78.23 RMSE:92.11 |
| Random Forest | MAE:24.42 RMSE:26.26 |
| Extreme gradient boosting | MAE:36.85 RMSE:42.56 |
| Support Vector Machine | MAE:101.11 RMSE:128.74 |
| (Tan, H., 2022) | Random Forest | n\_estimators: 100, 200, 500 | RMSE: 62.84 MAE : 35.71 MAPE:0.097 |
| XGBoost | RMSE: 62.84 MAE : 35.71 MAPE: 0.091 |
| LightGBM | RMSE: 62.84 MAE : 35.71 MAPE: 0.1 |
| SVM | RMSE: 62.84 MAE : 35.71 MAPE: 0.13 |
| AdaBoost | RMSE: 62.84 MAE : 35.71 MAPE:0.07 |
| CatBoost | RMSE: 62.84 MAE : 35.71 MAPE:0.11 |
| Sequential | RMSE: 62.84 MAE : 35.71 MAPE:0. 351 |
| (Michalakopoulos et al., 2023) | Linear Regression | Not stated | RMSE: 78.19% |
| Decision Tree | RMSE: 85.06% |
| Extra Trees Regressor | RMSE: 93.98%. |
| Random Forest Regressor | RMSE: 92.68% |
| LightGBM Regressor | RMSE: 90.67% |
| Gradient Boosting Regressor | RMSE: 92.28 |
| (Harati et al., 2024) | Support Vector Regressor | Not stated | Test:0.996 |
| K-Nearest Neighbours Regressor | Test: 1.000 |
| Decision Tree Regressor | Test:0.9966 |
| Random Forest Regressor | Test:0.9964 |
| Artificial Neural Network | Test: 0.9998 |
| (Li, X. et al., 2023) | Random Forest Regressor | Not stated | R2train: 0.970, R2test: 0.896, RMSE: 1.35 |
| (Fujii et al., 2024) | RFR (Single Amine) | Not stated | R2 0.943, RMSE: 0.072-0.073 |
| RFR (Simplified Descriptor List) | R2 0.931, RMSE: 0.079 |
| RFR (Single and Blended Amines) | R2: 0.944, RMSE: 0.073 |

## 4.1 Key Insights and Transition

Due to industrial activity and energy consumption patterns, the United States has a crucial role in affecting global climate change, as shown by the literature. While technological breakthroughs have many positive aspects, their usage, notably in producing petrochemicals and fossil fuels, has also led to environmental damage. Although laws have reduced emissions, there are still issues with their enforcement and technical constraints. Considering these complications, it is evident that more research into more practical approaches to reduce emissions and the creation of technology capable of striking a balance between industrial expansion and environmental sustainability are needed. The use of sophisticated machine learning models has the potential to enhance the precision of emissions projections and steer policy choices in the future.

# 

# Chapter 3: Methodology

This research methodology follows the well-established and defined CRISP-DM (Cross-Industry Standard Process for Data Mining) framework for data mining and predictive modelling. CRISP-DM is a complete guide that ensures that every aspect of the data mining project is carried out methodically, it was developed in the late 1990’s (Saltz, 2021). This framework ensures that every project component from business knowledge to model aspects is carefully studied and implemented, making it especially beneficial for projects requiring enormous datasets and complicated predictive models (Jaggia et al., 2020).

## 3.1 Technical Environment and Tools

The entire code to achieve the predictive model has been implemented in the Python programming environment version 3.12.0 with the help of Jupiter Notebook. The main reason to choose the Jupiter notebook for running the Python code as it provides a development environment that facilitates iterative development, debugging and documentation. The main Python libraries used were Pandas for data manipulation, NumPy for numerical computations, and Matplotlib and Seaborn for building detailed visualisations to understand the trends of data (Martins, 2020). For machine learning models Scikit-Learn, XGBoost, and Statsmodels libraries were chosen for the implementation of machine learning models according to their respective areas of expertise. While Random Forest and XGBoost were used for regression tasks, ARIMA and SARIMA models were created for time series analysis using the Statsmodels package (Tunnicliffe Wilson, 2016). Due to their established track record in academic research and their efficacious datasets and intricate model structures, these tools and libraries were selected.

## 3.2 Business Understanding

The CRISP-DM methodology is divided into six stages, starting with the Business Understanding phase, which describes the project’s goal (Wirth and Hipp, 2000). This research’s main objective was to create a predictive model for estimating CO2 emissions in the US to aid the nation’s attempts to attain net-zero emissions. To develop a model that can provide accurate estimates under a range of scenarios, it is necessary to understand the primary drivers of CO2 emissions, which is the focus of this business objectives. Such projections are meant to influence environmental policy and aid with the prevention of global climate change by providing information for government actions trying to reduce emissions.

## 3.2 Data Understanding

Data Understanding is the crucial step as it helps to find the characteristics of the dataset, which was gained from the “OUR WORLD IN DATA” (Ritchie and Roser, 2024). a comprehensive platform providing datasets on CO2 emissions from various energy sources and technology developments. The dataset covered an extensive range of characteristics, including demographic aspects spanning multiple decades, economic metrics like GDP, and emissions from different energy sources. This stage involved a preliminary review of the data to determine its extent, any difficulties, and regions that needed extra care during the data processing process.

## 3.3 Data Preparation

An essential first step in converting the unprocessed material into a format appropriate for analysis and modelling the data preparation (Brownlee, 2020). The dataset was thoroughly cleaned, with missing values being handled using imputation techniques to maintain bias-free results. To keep the model’s predictions from being distorted, outliers were found and dealt with appropriately. Additionally, to enhance model performance and guarantee that the statistical presumptions underlying the models were satisfied, it was also crucial to address skewness in the data. This included normalising skewed distributions of variables especially those related to energy consumption and economic indicators by using transformations such as logarithmic or square root transformations (Osborne, 2002).

## 3.3.1 EDA (Exploratory Data Analysis)

To gain insights and understanding of the data, EDA was carried out after the data had been prepared. Using a variety of visual aids and statistical summaries, EDA entailed looking for patterns, connections, and possible abnormalities in the data which helped to gain the past dynamics of the carbon emissions (Mitchell et al., 2018). With the help of EDA, different kinds of patterns and correlations between variables were utilised to get a better understanding of the structure of the data, time-series plots, correlation metrics, and distribution evolutions. This is to ensure that the analysis is based on a complete comprehensive analysis of the data and to provide guidance for future modelling decisions required at this stage (Claudio-Quiroga et al., 2023).

## 3.4 Feature Engineering

A key element of this research, aimed to increase the machine learning model’s capacity for CO2 emissions forecasting using feature engineering (Ji et al., 2024). The procedure entailed meticulously altering variables to tackle problems like data skewness and non-linearity. For example, to normalise skewed distributions and stabilise variance, logarithmic transformations were performed on several columns, especially those on energy consumption and economic indicators. These adjustments to the data were crucial in preparing the data for the study’s machine learning models, which allowed the algorithms to represent the correlations between the variables.

## 3.5 Modeling

In the Modelling phase, several machine learning techniques were put into practice throughout the modelling phase to handle the problems that the dataset presented. Time series analysis was conducted using ARIMA and SARIIMA models, which were selected based on their capacity to identify patterns and seasonality in the CO2 emissions data (Ji et al., 2024). These models were especially good at managing the data’s temporal characteristics, giving insights into how emissions have changed over time and how they could do so going forward. Random Forest and XGBoost were also used for regression problems because of their reliability while processing high-dimensional data and non-linear correlations, also these models are the most implemented for this task based on the literature review. These models were especially well-suited to represent the intricate relationships among the several variables affecting CO2 emissions. To avoid overfitting and maximise model performance, each model underwent extensive testing and validation using cross-validation techniques to fine-tune the hyperparameters Kohavi (2001), enabling a methodical parameter space exploration to find the best model configurations.

## 3.5 Evaluating and Interpretation

The models were evaluated using performance indicators specific to the objectives of the study throughout the evaluation and interpretation phase. Metrics like Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE) were chosen as they are useful for assessing prediction accuracy, especially when environmental data is included (Han et al., 2023). The average size of the prediction error was measured using the RMSE, which gave a clear indication of how well the model’s predictions agreed with the actual data (Chai and Draxler, 2014). By providing a percentage, MAPE made it simpler to evaluate the relative accuracy of projections at various emissions data scales (Hyndman and Koehler, 2006). Based on these parameters, the top-performing model was determined, and its outcomes were carefully examined to see how they might affect CO2 emissions and policy choices. This study shed important light on how the model may assist decision-making Procedures meant to lower emissions and meet sustainability goals.

## 3.5 Deployment and Practical Application

The created model’s practicality was highlighted throughout the deployment and Practical Application phase. The study demonstrated how policymakers may use the model as a tool to help them make decisions about reducing CO2 emissions. During the whole research, ethical issues were carefully considered, especially those about data privacy and the predictability of the model. The interpretability of the model ensured that no technical stakeholders, such as policymakers, could rely on and make good use of its findings (IPCC,2019). A cross-national validation was also a part of this step, in which the model was used on an INDIA carbon emission dataset which had the same structure of data as the USA dataset (Wang, X. et al., 2022). This validation further established the model’s usefulness as a tool for global climate change mitigation efforts by evaluating its resilience and generalisability in various situations.

## 3.5 Replication of Predictive Model for Cross-National Application

Following the model’s successful creation and examination, a replication phase was carried out to apply the model to the INDIA carbon emission dataset. To guarantee that the replication process matched the primary analysis in terms of data preparation, modelling, and testing, this phase also followed the same CRISP-DM Methodology. The model’s ability and adaptability to various national settings were proven by its successful replication, highlighting its potential as an important instrument in international efforts to mitigate climate change.

# Chapter 4: Product/Research Design and Implementation

## 4.1 Data Collection and Sources

## 4.1.1 CO2 emission data

The Dataset for this research was obtained from “OUR WORLD IN DATA”, a reliable source recognised for complying and displaying USA statistics from a range of subjects such as carbon emissions from energy use, technologies, and other fields. This platform is chosen based on the large historical and up-to-date datasets, which deliver the precision and reliability needed for solid research. As a result, it is one of the most famous platforms for providing historical and current data on energy usage and emissions, which are very useful for developing forecasting models and practical strategies that will lead to net zero emissions.

## 4.1.2 Importance of Reliable Data

It is crucial to guarantee the validity and credibility of outcomes, trustworthy data is necessary to research since it strengthens the entire argument and makes it possible for conclusions. The legitimacy of academic research is dependent on the quality of the data as it can influence the outcomes and make wrong perceptions and interpretations (Trisovic et al., 2021). Ensuring data timeliness, completeness, consistency, correctness, and reliability enhances the quality of research data and reinforces the authenticity of study findings. Experts can strengthen their arguments and recommendations to increase the accuracy of the study findings by maintaining strict standards for data quality (Feder, 2018).

### 4.1.3 Data Dictionary and Sample

During this research, one dataset was used to predict carbon emissions in the USA. The dataset represents the comprehensive historical carbon emissions from 1800 to 2022, it consists of various origins of emissions, like cement production, oil, and coal, and important economic and demographic metrics like GDP and population, which can be seen in Table 2. The dataset also gives detailed measures of emissions such as absolute emissions, per capita emissions, growth rates, temperature change by CO2 and the share of the United States in world emissions. Moreover, it also gives information related to sector-specific emissions and covers the changes in carbon emissions over time, which provides valuable insightful information on the developments done in the past and their effects worldwide.

Table 2.- Data Dictionary of CO2 Emission

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **country** | United States | United States | United States | United States | United States |
| **year** | 1800 | 1801 | 1802 | 1803 | 1804 |
| **iso\_code** | USA | USA | USA | USA | USA |
| **population** | 6000000 | 6113782 | 6229723 | 6347862 | 6468241 |
| **gdp** | NaN | NaN | NaN | NaN | NaN |
| **cement\_co2** | 0 | 0 | 0 | 0 | 0 |
| **cement\_co2\_per\_capita** | 0 | 0 | 0 | 0 | 0 |
| **co2** | 0.253 | 0.267 | 0.289 | 0.297 | 0.333 |
| **co2\_growth\_abs** | NaN | 0.015 | 0.022 | 0.007 | 0.037 |
| **co2\_growth\_prct** | NaN | 5.797 | 8.219 | 2.532 | 12.346 |
| **co2\_including\_luc** | NaN | NaN | NaN | NaN | NaN |
| **co2\_including\_luc\_growth\_abs** | NaN | NaN | NaN | NaN | NaN |
| **co2\_including\_luc\_growth\_prct** | NaN | NaN | NaN | NaN | NaN |
| **co2\_including\_luc\_per\_capita** | NaN | NaN | NaN | NaN | NaN |
| **co2\_including\_luc\_per\_gdp** | NaN | NaN | NaN | NaN | NaN |
| **co2\_including\_luc\_per\_unit\_energy** | NaN | NaN | NaN | NaN | NaN |
| **co2\_per\_capita** | 0.042 | 0.044 | 0.046 | 0.047 | 0.052 |
| **co2\_per\_gdp** | NaN | NaN | NaN | NaN | NaN |
| **co2\_per\_unit\_energy** | NaN | NaN | NaN | NaN | NaN |
| **coal\_co2** | 0.253 | 0.267 | 0.289 | 0.297 | 0.333 |
| **coal\_co2\_per\_capita** | 0.042 | 0.044 | 0.046 | 0.047 | 0.052 |
| **cumulative\_cement\_co2** | 0 | 0 | 0 | 0 | 0 |
| **cumulative\_co2** | 0.253 | 0.52 | 0.81 | 1.107 | 1.44 |
| **cumulative\_coal\_co2** | 0.253 | 0.52 | 0.81 | 1.107 | 1.44 |
| **cumulative\_flaring\_co2** | 0 | 0 | 0 | 0 | 0 |
| **cumulative\_gas\_co2** | 0 | 0 | 0 | 0 | 0 |
| **cumulative\_luc\_co2** | NaN | NaN | NaN | NaN | NaN |
| **cumulative\_oil\_co2** | 0 | 0 | 0 | 0 | 0 |
| **cumulative\_other\_co2** | NaN | NaN | NaN | NaN | NaN |
| **flaring\_co2** | 0 | 0 | 0 | 0 | 0 |
| **flaring\_co2\_per\_capita** | 0 | 0 | 0 | 0 | 0 |
| **gas\_co2** | 0 | 0 | 0 | 0 | 0 |
| **gas\_co2\_per\_capita** | 0 | 0 | 0 | 0 | 0 |
| **oil\_co2** | 0 | 0 | 0 | 0 | 0 |
| **oil\_co2\_per\_capita** | 0 | 0 | 0 | 0 | 0 |
| **share\_global\_cement\_co2** | NaN | NaN | NaN | NaN | NaN |
| **share\_global\_co2** | 0.771 | 0.836 | 0.718 | 1.005 | 1.05 |
| **share\_global\_co2\_including\_luc** | NaN | NaN | NaN | NaN | NaN |
| **share\_global\_coal\_co2** | 0.771 | 0.836 | 0.718 | 1.005 | 1.05 |
| **share\_global\_cumulative\_co2** | 0.032 | 0.063 | 0.094 | 0.124 | 0.156 |
| **share\_global\_cumulative\_co2\_including\_luc** | NaN | NaN | NaN | NaN | NaN |
| **share\_global\_cumulative\_coal\_co2** | 0.032 | 0.063 | 0.094 | 0.124 | 0.156 |
| **share\_global\_cumulative\_oil\_co2** | 0 | 0 | 0 | 0 | 0 |
| **share\_global\_cumulative\_other\_co2** | NaN | NaN | NaN | NaN | NaN |
| **share\_global\_gas\_co2** | NaN | NaN | NaN | NaN | NaN |
| **share\_global\_luc\_co2** | NaN | NaN | NaN | NaN | NaN |
| **share\_global\_oil\_co2** | NaN | NaN | NaN | NaN | NaN |
| **share\_global\_other\_co2** | NaN | NaN | NaN | NaN | NaN |
| **share\_of\_temperature\_change\_from\_ghg** | NaN | NaN | NaN | NaN | NaN |
| **temperature\_change\_from\_ch4** | NaN | NaN | NaN | NaN | NaN |
| **temperature\_change\_from\_co2** | NaN | NaN | NaN | NaN | NaN |
| **temperature\_change\_from\_ghg** | NaN | NaN | NaN | NaN | NaN |
| **temperature\_change\_from\_n2o** | NaN | NaN | NaN | NaN | NaN |

## 4.1.4 Data Description

The dataset of carbon emissions has multiple columns which are directly or indirectly related to CO2 but columns which have been chosen for the in-depth examination by goals of this study:: Year, coal\_Co2, flaring cumulative\_flaring, gas, cumulative\_gas, cumulative\_oil, co2, co2\_per\_capita, Co2\_per\_gdp, coal, co2\_per\_capita, flaring, gas, and oil co2\_per\_capita and temperature\_change\_from\_CO2:

The **Year** column is essential for tracking down past trends throughout the time. Analysing the past year’s data (1800-2022) on carbon emissions may help identify crucial transition times, grasp the effects of multiple factors throughout time, and identify significant variations in energy consumption and emissions. Temporal analysis is essential to start working on predictive models that can present the future forecast of carbon emissions based on historical data.

Coal is the prime source of carbon emission and in the sample dataset coal has been tracked down by the **coal\_co2** and **cumulative\_coal\_co2** columns, respectively, along with their cumulative totals. These columns represent the total historical impacts of coal use and coal consumption. This data is necessary to develop the knowledge to create a plan to cut back on coal usage and move towards more environmentally friendly energy sources.

CO2 emitted by the gas flaring and its cumulative effects are recorded in the **flaring\_co2** and **cumulative\_flaring\_co2** columns. Flaring is one of the major sources of greenhouse gases and is often connected to the manufacturing of gas and petrol. It is essential to understand these types of emissions to discover the Areas of decrease. The combined data provide an overall effect of these operations over time and allow an extensive knowledge of the contribution of flaring to total emissions.

To understand how Natural gas emissions are from the past then examining the **gas\_co2** and **cumulative\_gas\_co2** columns is crucial for the analysis. Natural gas emits less greenhouse gases than coal and oil, but natural gas contributes significantly to greenhouse gas emissions. By examining these data, one may assess the natural gas in the energy mix and find solutions to reduce the natural gas impact on the environment.

It is critical to understand the oil past data which has influenced the carbon emissions in the USA. The **oil\_co2** and **cumulative\_oil\_co2** columns provide information on CO2 emissions from oil consumption and their cumulative totals. Learning the negative impacts of oil is essential to create strategies for reducing oil usage and switching to renewable energy sources. The combination of this data helps assess the historical share of the total greenhouse gases and develops the basis for long-range strategy and policy development.

Overall emission of CO2 is one of the critical metrics which help to identify the total carbon emitted by all sources and assist experts assess the efficacy of diverse policies and initiatives to reduce the carbon emission based on based of the past trends and patterns. This column provides a detailed summary of the complete emissions situation, which is crucial to the study.

The economy's average carbon footprint and carbon intensity have been identified as “co2\_per\_capita” and “co2\_per\_gdp”, respectively. The co2\_per\_capita data shed light on the effectiveness of programs intended to reduce individual emissions, on the other hand, co2\_per\_gdp stats show the carbon efficiency of commerce. Overall, both columns assist in understanding the relationship between emissions, personal actions, and economic expansion and provide a structure for creating targeted policies and initiatives.

## 4.2 Ethical Considerations

The carbon emission for USA data used in this study is collected from the “OUR WORLD IN DATA” an open-source platform renowned for its vast and historical information. The data is anonymised, publicly available for everyone, and includes the years 1800-2022, so there is no chance that it may hurt an individual or anything as it does breach anyone’s information. Strict adherence to ethical principles ensures transparency in the management and interpretation of data in the research. This flexibility is maintained by carefully documenting methods and admitting any limitations or biases, guaranteeing the accuracy and dependability of the study results.

Furthermore, the research goes to considerable lengths to ensure that the data it offers is unbiased and free of bias. The research tries to avoid misunderstanding or neglect of data while contributing significant insights by complying with ethical norms. Since the findings and recommendations are derived from a thorough examination, the study is certain that it will have an advantageous impact on the area without compromising ethical issues.

## 4.3. Data Pre-processing (ETL)

## 4.3.1 Carbon emission data

To ensure data integrity and analytical value, the ETL (Extract, Transform, Load) process applied in this study was carefully designed to prepare the dataset for further examination and all the involved steps have been outlined in Table 3. The systematic input of raw historical USA emissions of carbon data into a Pandas data frame formed the data extraction step. Since the source of the information, which included a variety of emissions metrics, was in CSV format, it was essential to extract useful data before moving on to more complex processing steps so that it would be more structured and managed more easily.

One of the most significant phases in the ETL process is transformation, which requires a systematic procedure to clean and improve the quality of the collected data. To maintain the dataset’s durability and reliability, columns with more than 50% missing data were first eliminated as columns with 50% missing values could compromise the integrity of the analysis and result in distorted or false findings and less than columns with missing values have been represented in Figure 1. To maintain the accuracy and completeness of the data, zero was substituted as the mean of each column for the remaining missing data. This methodology was selected because imputing the mean is a commonly used technique that maintains the data’s general distribution and central tendency, minimising the introduction of bias and guaranteeing that the dataset’s statistical features are preserved (Noor et al., 2015). Moreover, an important goal of this phase was to solve the issue of skewness in the data. A logarithmic adjustment was used on important variables, such as coal\_co2, cumulative\_coal\_co2, flaring\_co2, cumulative\_ gas\_co2, co2, gas\_co2\_per\_capita, oil\_co2\_per\_capita and temperature\_change\_from\_co2 as these columns skewness had the undistributed data and had the value more than +1.5 and -1.5 since skewed data distributions might undermine the reliability of statistical analysis and prediction models. By normalising the data distribution, this modification helped to lessen skewness and make the analysis more accurate, also, after completing the transformation the cleaned data can be seen in Table 4.

Ultimately, the cleaned and transformed dataset was back into a CSV file during the loading stage. The final check was made that the data was appropriately formatted for any additional analysis, modelling, or presentation that the study needed, as well as being preserved reliably. The organised ETL procedure was necessary to ensure that the research data was carefully organised and to enhance the reliability and legitimacy of the research outcomes.

Table 3.- ETL Process Summary for USA CO2 emission Dataset

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Extract** | | | | **Transform** | |  | | **Load** | | |
| **Attribute Name** | **Type** | **Key** | **Data Quality Issues** | | **Action Note** | | **Load Attribute** | | **Load Key** | **Notes** |
| Year | Integer | Yes | None | | Retain | | year | | Yes | Primary key for data records |
| coal\_co2 | Float | No | None | | Retain | | coal\_co2 | | No | Retained for analysis |
| cumulative\_coal\_co2 | Float | No | None | | Retain | | cumulative\_coal\_co2 | | No | Retained for analysis |
| flaring\_co2 | Float | No | None | | Retain | | flaring\_co2 | | No | Retained for analysis |
| cumulative\_flaring\_co2 | Float | No | None | | Retain | | cumulative\_flaring\_co2 | | No | Retained for analysis |
| gas\_co2 | Float | No | None | | Retain | | gas\_co2 | | No | Retained for analysis |
| cumulative\_gas\_co2 | Float | No | None | | Retain | | cumulative\_gas\_co2 | | No | Retained for analysis |
| oil\_co2 | Float | No | None | | Retain | | oil\_co2 | | No | Retained for analysis |
| cumulative\_oil\_co2 | Float | No | None | | Retain | | cumulative\_oil\_co2 | | No | Retained for analysis |
| co2 | Float | No | None | | Retain | | co2 | | No | Retained for analysis |
| co2\_per\_capita | Float | No | None | | Retain | | co2\_per\_capita | | No | Retained for analysis |
| co2\_per\_gdp | Float | No | Missing values imputed | | Imputed missing values with mean | | co2\_per\_gdp | | No | Imputed missing values using mean |
| coal\_co2\_per\_capita | Float | No | None | | Retain | | coal\_co2\_per\_capita | | No | Retained for analysis |
| flaring\_co2\_per\_capita | Float | No | None | | Retain | | flaring\_co2\_per\_capita | | No | Retained for analysis |
| gas\_co2\_per\_capita | Float | No | None | | Retain | | gas\_co2\_per\_capita | | No | Retained for analysis |
| oil\_co2\_per\_capita | Float | No | None | | Retain | | oil\_co2\_per\_capita | | No | Retained for analysis |
| temperature\_change\_from\_co2 | Float | No | Missing values imputed | | Imputed missing values with mean | | temperature\_change\_from\_co2 | | No | Imputed missing values using mean |

A graph with blue bars and black text

Description automatically generated with medium confidence

Figure 1.- Histogram for the missing values in carbon emissions Dataset

Table 3.- Overview of the cleaned Dataset for CO2 Emissions

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **year** | 1800 | 1801 | 1802 | 1803 | 1804 | 1805 |
| **coal\_co2** | 0.253 | 0.267 | 0.289 | 0.297 | 0.333 | 0.341 |
| **cumulative\_coal\_co2** | 0.253 | 0.52 | 0.81 | 1.107 | 1.44 | 1.781 |
| **flaring\_co2** | 57.423 | 63.124 | 64.372 | 55.139 | 59.486 | 72.066 |
| **cumulative\_flaring\_co2** | 0.34 | 0.335 | 0.318 | 0.306 | 0.296 | 0.296 |
| **gas\_co2** | 0.18 | 0.196 | 0.198 | 0.169 | 0.18 | 0.217 |
| **cumulative\_gas\_co2** | 4.416 | 4.457 | 4.547 | 4.529 | 4.429 | 4.875 |
| **oil\_co2** | 0.214 | 0.217 | 0.219 | 0.221 | 0.223 | 0.226 |
| **cumulative\_oil\_co2** | 0.211 | 0.216 | 0.218 | 0.210 | 0.220 | 0.223 |
| **co2** | 0.253 | 0.267 | 0.289 | 0.297 | 0.333 | 0.341 |
| **co2\_per\_capita** | 0.042 | 0.044 | 0.046 | 0.047 | 0.052 | 0.052 |
| **co2\_per\_gdp** | 0.708533 | 0.708533 | 0.708533 | 0.708533 | 0.708533 | 0.708533 |
| **coal\_co2\_per\_capita** | 0.042 | 0.044 | 0.046 | 0.047 | 0.052 | 0.052 |
| **flaring\_co2\_per\_capita** | 0.31 | 0.23 | 0.34 | 0.35 | 0.40 | 0.39 |
| **gas\_co2\_per\_capita** | 0.314 | 0.336 | 0.347 | 0.345 | 0.374 | 0.380 |
| **oil\_co2\_per\_capita** | 0.214 | 0.212 | 0.220 | 0.223 | 0.219 | 0.240 |
| **temperature\_change\_from\_co2** | 0.083509 | 0.083509 | 0.083509 | 0.083509 | 0.083509 | 0.083509 |

## 4.4 Exploratory Data Analysis (EDA)

A conceptual framework known as Exploratory Data Analysis (EDA) is based on a core set of ideas and values that are intended to give working researchers insight into data, regardless of where it comes from, and to promote understanding of both probabilistic and non-probabilistic models in a way that prevents incorrect conclusions (Weiner et al., 2012).

## 4.4.1 Understanding the Trends in Carbon Emissions

Figure 2 explains the patterns in carbon dioxide emissions from 1800 to 3033 for three main energy sources: coal, gas, and oil, The years are shown on the x-axis, while the amount of CO2 emissions expressed in million tonnes, is shown on the y-axis. The data indicates an increase in global CO2 emissions, especially following the Industrial Revolution when coal was the main source of emissions at first. Emissions from petrol and oil also increased dramatically over time, especially in the middle of the 20th century. It shows the heavy reliance on fossil fuels and how they have contributed to the atmosphere’s growing CO2 levels. Moreover, Emissions have decreased recently, which may be a sign of a move towards greener energy sources or the effect of environmental laws.

**A graph of a graph showing the amount of carbon dioxide emissions

Description automatically generated**

Figure 2.- Trends in Carbon Emissions by Energy Sources.

## 4.4.2 CO2 Emissions Per Capita and GDP Over Time

Figure 3 represents the per capita and GDP-related changes in CO2 emissions between 1800 to 2022. CO2 emissions are shown on the y-axis, while the x-axis displays the years. The blue line, which displays CO2 emissions per person, shows an overall rising trend over time, which indicates rising energy per person. On the other hand, a high in the middle of the 20th century, the orange line, which indicates CO2 emissions per GDP, gradually declined. The potential for sustainable economic development is highlighted by this pattern, which points towards the increase in energy efficiency and the separation of economic growth from CO2 emissions.

A graph of a graph showing the amount of emissions in the world

Description automatically generated with medium confidence

Figure 3.- CO2 Emissions Per Capita and GDP Over Time.

## 4.4.3 Cumulative CO2 Emissions Over Time

The cumulative CO2 emissions over time from coal, gas, and oil sources are shown in Fig. The period from 1800-2022, shown on the x-axis, is plotted against the cumulative emissions in millions of tonnes on the y-axis. The Figure 4 cumulative format highlights the historical build-up of CO2 in the atmosphere as well as the long-lasting effects of burning fossil fuels over centuries.

A graph of co2 emissions

Description automatically generated

Figure 4.- Cumulative CO2 emissions by Sources

## 4.4.4 Per Capita CO2 Emissions by Source Over Time

The per capita CO2 emissions from oil, gas, and coal are represented in Figure 5 over time. CO2 emissions per capita are plotted on the y-axis, with the x-axis representing the years 1800-2022. According to the figures, emissions from coal climbed first and peaked in the middle of the 20th century, but emissions from oil and gas increased sharply after World War II.

A graph of a graph showing the growth of the co2 emissions

Description automatically generated

Figure 5.- CO2 emission Per Capita by Energy Source

## 4.4.4 CO2 Emission and 5-Year Moving Average

Figure 6 represents the past pattern of CO2 emissions in conjunction with the 5-year moving average. CO2 emissions are displayed in million tonnes on the y-axis, with the years 1800-2022 represented in the x-axis. Nonetheless, the blue line shows the CO2 emissions, which vary significantly, especially in recent decades. The moving average recorded by the red line smooths out these short-term variations to show a long-term rising trend in CO2 emissions. The moving average shows a steady increment in emissions over time, reaching the highest point about 2000, and then showing a little fall, which may indicate the success of initiatives to reduce emissions or adjustments to energy policy.

A graph showing the growth of co2 emissions

Description automatically generated

Figure 6.- CO2 emission with 5-year Moving Average

## 4.4.5 Correlation Matrix with the Use of Heat Map

The correlation matrix of different energy-related factors is displayed in Figure 7 Various energy measures are listed on the x and y axes, and each cell displays the degree and direction of the association between them. Positive correlations are shown by red colours, and negative correlations are indicated by blue shades. Since every variable has perfect correlations with itself, it is not surprising that the diagonal cells show perfect correlations of 1.0. Strong connections between cumulative emissions from coal, gas, and oil are seen in the matrix, suggesting that increases in one frequently follow rises in the others. Understanding the interdependencies between various energy sources and how they affect overall emissions is made easier with the help of Figure 7.

A red and white grid with numbers

Description automatically generated

Figure 7.- Correlation Matrix of Energy-Related Columns

## 4.4.5 Distribution of Cumulative CO2 Emissions by Source

The distribution of total CO2 emissions by coal, petrol and oil is represented in Figure 8 According to the graph, coal accounts for the greatest portion of total emissions 42%, followed by oil 38.9% and gas 19.1%. This distribution demonstrates how coal has contributed the most to historical CO2 emissions, with oil and gas also having a substantial impact on world emissions. A concise overview of the relative effects of various fossil fuels on total CO2 emissions is given in Figure 8.

A diagram of a graph

Description automatically generated

Figure 8.- Distribution Of cumulative CO2 Emissions by Source

## 4.4.5 Temperature Change from CO2 Over Time

Figure 9 represents the connection between the temperature change and carbon emissions over 222 years. The y-axis represents the Temperature change (°C) while the x-axis provides details about the years (1800-2022) provided in the dataset. As we have replaced a few missing values with the average so from 1800 to 1850 there is a constant line which may not provide the correct information but from 1850, there was a constant increase in the temperature from 1851 which represents the link between increasing CO2 levels and rising global temperatures, demonstrating a sharp rise in the 20th century that was accompanied by widespread industrialisation and the combustion of fossil fuels.

A graph showing the temperature of the earth

Description automatically generated

Figure 9.- Temperature Change from CO2 Over Time

## 4.5 Featuring Engineering

One of the most important steps in the data preparation process is feature engineering, which converts the chosen columns into structures or formats that improve the predictive model’s performance (Zheng, A. and Casari, 2018). Even though the relevant features have already been selected in the early stages due to their significance to CO2 emissions, missing values on those features have been filled with their average column value.

## 4.6 Model Selection and Development for CO2 Emissions

The choice of model is a crucial stage in predicting CO2 emissions as it has an immediate impact on the precision and dependability of the results. However, there are three parameters of the predictive model that ensure the accuracy and reliability of predictions.

* **Accuracy**: - A model must produce predictions that closely align with the data that has been observed. Reliable projections and minimised mistakes from accurate models are essential for forecasting future emissions patterns and guiding policy decisions (Bratko, 1997).
* **Interpretability:** - it is critical that a model be able to produce predictions that closely align with the data that has been observed. Reliable projections and minimised error from accurate models are essential for predicting future emission patterns and guiding policy decisions (Bratko, 1997).
* **Computational Efficiency:** - Computational efficiency is a key consideration because of huge datasets and the possible requirement for real-time forecasts. Continuous tracking and forecasting attempts are better served by models that can handle data quickly without losing accuracy (Zhao, T. et al., 2022).

### 4.6.1 Overview of Considering Models

Several models were considered to forecast CO2 emissions, and each was evaluated according to the standards of simplicity, precision, comprehensibility, and computational effectiveness.

* **ARIMA/SARIMA**: - Arima is excellent at using trend forecasting as it models future values using previous values and random error factors, which makes it possible to analyse time series data with confidence (Malik et al., 2020). By adding seasonal components, the SARIMA model expands on this capacity and allows it to take into consideration periodic variations that are frequently found in environmental data, such as CO2 emissions (Williams, A. T. et al., 2023).

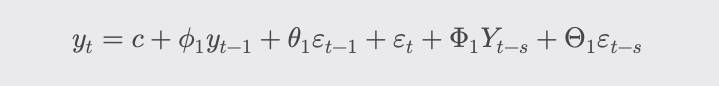
Equation 1.- ARIMA Formula

A black text on a white background

Description automatically generated

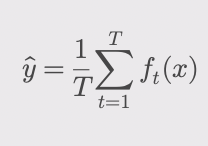
* value at time t, the current observation
* Constant term, representing the baseline level.
* AR (1) term, here is the coefficient for the previous value of the series.
* MA (1) term, where is the coefficient for the previous error.
* Error term (white noise) at time t.

Equation 2.- SARIMA Formula



* : value at time t, the current observation
* Constant term, representing the baseline level.
* : AR (1) term, here is the coefficient for the previous value of the series.
* : MA (1) term, where is the coefficient for the previous error.
* : Error term (white noise) at time t.
* : Seasonal AR(1) term, where is the seasonal AR coefficient and is the value from the previous season.
* : Seasonal MA(1) term, where is the seasonal MA coefficient and is the error from the previous season.
* **Random Forest:** A **r**andom Forest is an ensemble method of combining many decision trees and combining their predictions that successfully simulate intricate, non-linear connections in the CO2 emissions data (Zhang, Y. et al., 2022). This method works well with the high-dimensional datasets commonly seen in environmental research since it manages several variables while improving accuracy and resilience. It is a trustworthy method for emissions forecasting because of its well-established predictive accuracy (Williams, A. T. et al., 2023).

Equation 3.- Random Forest Formula



Where:

* is the predictive value.
* is the number of trees in the forest
* is the prediction of the t-th tree.
* **XGBooST (Extreme Gradient Boosting): -** XGBooST is reorganised for its outstanding performance when dealing with complex datasets, like time series data. This method is a good option for CO2 emission prediction since it reduces overfitting by optimising model performance using a regularisation approach. Nevertheless, XGBoost’s drawback is that it could be harder to interpret, when compared to other models, making its forecasts more difficult to comprehend.

Equation 4.- XGBoost Formula

A mathematical equation with a number and a square

Description automatically generated with medium confidence

Where:

* is the predictive value
* is the number of trees in the forest
* is the prediction of the t-th tree.
* is the learning rate, controlling the contribution of each tree.

## 4.6.2 Hyperparameters for Models

Machine learning models with hyperparameters are those whose values are predetermined and not acquired through training data. In contrast to model parameters, which are changed during training, hyperparameters affect the model’s general structure and functionality (Jin, 2022). Effective hyperparameter tuning is essential as it greatly affects the model’s computational efficiency, performance, and capacity for generalisation.

For this research, four machine learning models have been selected for predicting the CO2 emissions in the USA, each model hyperparameter has been chosen to maximise efficiency. The Random Forest model’s number of trees (n-estimators) was 100. The necessity to manage computing complexity while improving forecasting accuracy led to this decision. In general, more trees increase model stability and lower variation, resulting in strong performance across various data subsets (Tan, H., 2022).

There were 38 boosting rounds (n-estimators) used for the XGBoost model. Through a rigorous process of testing, the number of iterations was established to balance the model’s capacity to learn complicated patterns without overfitting. The model can capture non-linear correlations in the data while preserving generalisability thanks to the reasonable number of boosting rounds (Si and Du, 2020).

ARIMA and SARIMA were used to capture the complex patterns in the CO2 emissions data. With an order of (1,2,1), the ARIMA model was set up to simulate temporal dynamics using a moving average component (q=1), a single lag term (p=1), and second-order differencing (d=2). In addition, the SARIMA model configuration takes into consideration the monthly seasonality present in the data. These hyperparameters were carefully selected to ensure that the models produced precise and understandable CO2 emissions projections, accounting for both seasonal impacts and linear trends, through empirical testing (Kumari and Singh, 2023).

## 4.6.3 Performance Matrix for Models

Performance metrics are crucial instruments for assessing the precision and efficacy of prediction models. By measuring the discrepancy between expected and actual results, these metrics enable researchers to evaluate the efficacy of the models and make necessary modifications (Erickson, B. J. and Kitamura, 2021).

In this research, mean absolute percentage error (MAPE) and root mean square error (RMSE) were used to evaluate the model performance and compare them to select the best model for this study. MAPE and RMSE have been chosen based on the literature review as these are the models used matrix for the evaluation of the models for carbon emissions. RMSE calculate the square root of the average squared differences between the expected and actual values, as it highlights bigger mistakes, it is more susceptible to outliers.

Equation 5.- Root Mean Square Root Formula

A mathematical equation with a square and square with a square and square with a square and a square with a square and a square with a square and a square with a square and a square with a

Description automatically generated

* yi is the actual value for the ith observation.
* ŷi is the predicted value for the ith observation.
* N is the number of observations.
* P is the number of parameter estimates, including the constant.

MAPE calculate the prediction error as a percentage, which makes it simple to read on a variety of scales. The mean absolute difference between the actual and forecasted values about the actual values is computed.

Equation 6.- Mean Absolute Percentage Error Formula

A black and white math equation

Description automatically generated with medium confidence

Here A = Actual, F= Forecast, N= Number of observations, and the vertical bars represent absolute values.

## 4.7 Replication of Predictive Model for Cross-National

## 4.7.1 Data Pre-processing

The pre-processing procedures have followed the same strict approach to guarantee consistency and dependability as the dataset utilised for INDIA is truly like the one employed in the main model. This includes filling in missing values using imputation techniques like mean or median filling. To improve the model performance and integration, further steps like as data normalisation or scaling may be used to pull all features into the same scale. Any modifications to the data that were made to the primary model, including logarithmic adjustments to handle skewed distributions or other methods to deal with outliers.

## 4.7.2 Feature Engineering

Mirrored in the new dataset will be the feature engineering procedure that proved crucial to the primary model’s performance. The process entails generating any extra characteristics that were shown to be noteworthy, including polynomial features, lagged variables that depict temporal dynamics, or interaction terms among variables. The INDIA dataset will incorporate elements utilised in the original research such as interaction terms between energy consumption and economic variables or delayed CO2 emissions estimates. By keeping the insights from the original feature engineering process, this phase guarantees that the model’s predictive power is retained across various datasets.

## 4.7.3 Model Implementation

For the new dataset, the model that had performed the best with the USA dataset has been used. This involved sticking to the same model architecture and hyperparameters, regardless of the model that was determined to be the best, whether it was Random Forest, XGBoost, SARIMA, or ARIMA. The same modelling approaches were successfully applied to the data from INDIA, due to the data from the new nation due to the structural similarities of the datasets, assuring that the model’s performance could be consistently assessed.

## 4.7.4 Model Training and Evaluation

The training and testing of the data separated from the original model was applied to the dataset of INDIA to train the model. The accuracy and robustness of the model were evaluated on the new dataset using performance measures, such as Root Mean Square (RMSE) and Mean Absolute Percentage Error (MAPE), which were previously utilised to test the main model. By directly comparing the outcomes from the two nations, this method gave an understanding of how generalisable the model is. The outcomes of the primary model and the new model’s results were examined and evaluated. In-depth analysis and reporting of any performance disparities or similarities provided important context and dataset-specific information into how effectively the model generalised. When the predictive model was applied to the INDIA carbon emissions data, this comparison showed its advantages and possible areas for improvement for another country as well.

## 4.7.5 Interpretation and Reporting

Lastly, a comparison has been made between the outcomes of the new model and the primary models' outcomes. Any performance parallels or differences were examined and documented. The results shed light on whether the model performed effectively across different datasets and if the primary dataset’s prediction tendencies applied to the new nation.

# Chapter 5: Research Outcomes/Results/Discussion and Evaluation

## 5.1 Results

This chapter describes the evaluation and comparison between the four models applied to forecast CO2 emission for the USA and their results.

## 5.1.1 USA Carbon Emission Model Results

Table 5 represents the remarkable predictions of CO2 emissions by all the models, which has been used in this research. The Random Forest model testing and training MAPE values of 5.97% and 8.44% respectively, the model also achieved 0.0403 and 0.0639 RMSE values in testing and training. Table 5 indicates a little difference between RMSE values for training and testing, indicating that the model has good generalisation capabilities for both training and testing datasets. This little rise in MAPE further demonstrates the accuracy and dependability of the model in estimating CO2 emissions. By contrast, the 38-boost round has been set up in XGBoost as a hyperparameter, which helped the model to perform well. It outperformed Random Forest in training by reaching an RMSE of 0.0348 and MAPE of 2.91% in training. However, the testing score of RMSE 0.0718 and MAPE 10.59%, indicated a somewhat greater error, indicating little overfitting, as mentioned in Table 5. Despite this, the XGBoost model was a strong contender to handle both linear and non-linear interactions.

The ARIMA model with hyperparameter (p=1, d=2, q=1) has performed the worst as compared to the tress-based models. The model recorded a Training RMSE of 0.1192 and a testing RMSE of 0.9279, with training and testing MAPE values of 14.01% and 70.87% respectively. These findings imply that ARIMA had issues applying to the CO2 and was unable to identify the more intricate patterns in the sample. It represents the ability to handle non-linearity is very bad in ARIMA. Moreover, the SARIMA model did not perform well as well with the CO2 emission data set. This model includes parameters like (p=1, d=2, q=1, p=1, D=1, Q=1, S=12. The training and testing RMSE and MAPE values for the SARIMA model, were 0.1278 and 1.0512, respectively. These findings highlight even more how poorly the model performed. These findings highlight the model’s limitation in terms of generalisation, especially its inability to deal with the data’s seasonal and linear trends.

Table 4.- performance Matrix of Random Forest Model on USA CO2 Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Hyperparameters | Training RMSE | Testing RMSE | Training MAPE | Testing MAPE |
| Random Forest | n\_estimators=100 | 0.0403 | 0.0639 | 5.97% | 8.44% |
| XGBoost | n\_estimators=38 | 0.0348 | 0.0718 | 2.91% | 10.59% |
| ARIMA | p=1, d=2, q=1 | 0.1192 | 0.9279 | 14.01% | 70.87% |
| SARIMA | p=1, d=2, q=1, D=1, Q=1, s=12 | 0.1278 | 1.0512 | 15.55% | 101.42% |

## 5.1.2 INDIA Carbon Emission Model Results

Random Forest performed very well with the CO2 emission data set, so random Forest has been applied to the INDIA CO2 emission dataset to verify if the same model can predict and perform well in different countries. When the Random Forest model was replicated on the INDIA dataset, it performed well. The model demonstrated a low error value for both known and unknown data, with a Training RMSE of 7.33. Additionally, the model achieved accuracy in forecasting CO2 emissions with little error as shown in training MAPE of 0.11% and testing MAPE of 0.20% as seen in Table 8 These outcomes demonstrate the model stability and strong adaptability throughout the INDIA dataset.

Table 5.- performance Matrix of Random Forest Model on INDIA CO2 Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Hyperparameters | Training RMSE | Testing RMSE | Training MAPE | Testing MAPE |
| Random Forest | n\_estimators=100 | 14.50 | 7.33 | 0.011% | 0.020% |

## 5.2 Discussion

The capacity of the four models to handle complicated, non-linear interactions within the data was shown in the wide range of performance, each showed when estimating CO2 emissions for the USA dataset.

According to Table 7, the Random Forest model was found to be the most dependable among all the models. Training (0.0403) and Testing (0.0403) RMSE values for the model were low, and this narrow difference represents that it performs well in terms of adaptability without overfitting. Accurate forecasting based on unseen data requires this. The reason for Random Forest exceptional success was its capacity to represent complex non-linear relationships between many variables, Including industrial activity, economic development, and energy use. The model was repeated on the INDIA dataset after performing remarkable performance on the USA dataset. The model flexibility in many geographic settings is confirmed by the low training RMSE of 7.33 and testing MAPE of 0.20% for the INDIA dataset, showing model versatility and application to a variety of datasets.

On the other hand, XGBoost showed signs of overfitting even if it performed admirably overall. While the training RMSE and MAPE of the model were remarkable, at 0.03448 and 2.91%, respectively. Suggesting a somewhat greater error when applied to unknown data. This shows that even while XGBoost is good at finding patterns in the training set, more hyperparameter tweaking could be necessary to improve the model’s ability to generalise to new data. Nevertheless, XGBoost can handle both linear and non-linear interactions, particularly when its hyperparameters are adjusted to minimise overfitting.

The testing value of RMSE and MAPE value of ARIMA was recorded at 0.9279 and 70.87%, respectively. Particularly indicating that the model had difficulty capturing the complex patterns found in the CO2 emissions dataset. The shortcomings of ARIMA become apparent when working with multi-variable datasets where non-linear interactions rather than strictly time-dependent correlations between components dominate the relationships. When abrupt changes or outside effects are present in the underlying data structure, as they frequently are in emissions data because of shifting laws, industrial operations, and energy use over time, Arima is not as effective at predicting future trends as it could be.

Similarly, with a testing RMSE of 1.0512 and a MAPE of 101.42%, SARIMA also underperformed. While SARIMA works well for time-series data that demonstrate significant seasonal patterns, it was unable to adequately capture the simple seasonal cycles seen in the CO2 emissions data. The irregular and multi-dimensional character of emissions data, which can change unpredictably depending on a wide variety of circumstances, was not taken into consideration by the model because it relied on finding recurring seasonal trends.

The adaptability and versatility of the Random Forest model are further supported by its successful replication on the INDIA dataset. The USA is a developed country and India is still in the process of developing. Also, both have different economic systems, energy usage habits, and emission levels. Yet, the Random Forest model has remained very accurate and broadly applicable. The minimal error values for both the training and testing stages demonstrate not only this model is only suitable for the USA but also reliable across different national contexts. Random Forest’s adaptability makes it a useful tool for academics and policymakers to plan carbon reduction plans across different locations and expect emissions with accuracy.

This discovery has significant implications beyond the technical improvements in CO2 emissions predictions. This research greatly advances the current attempts to combat climate change by creating an accurate and reliable prediction model. More accurate emission forecasting helps governments, businesses, and environmental groups make well-informed decisions, carry out focused interventions, and assess how well carbon footprint-reducing policies work. To take proactive action ecological limits are crossed, and predictive models such as the ones investigated in this study can be vital instruments for early identification of possible increases in emissions. Moreover, solid data from precise forecasting helps nations set reasonable goals for reducing, which in turn helps them fulfil international climate agreements like net-zero targets. Over time, the incorporation of these forecasting models into sustainability plans can stimulate the development of environmentally friendly technology, enhance the management of resources, and eventually facilitate the shift to a low-carbon economy. By laying the foundation for more accurate environmental planning.

# Chapter 6: Project Management

This chapter describes the project management strategy applied to this research, showing how the aim of this study was achieved in the time frame.

## 6.1 Objective Alignment and Resource Management

The project's aim was broken down into four smaller parts to ease the tasks at the beginning. These were ranked in order of importance, according to the key events in this research, which were divided into several stages, like data gathering, data preparation, exploratory data analysis, model building, and review. Resources such as Excel and Python have been used to complete objectives quickly and effectively, so the aim can be achieved of this research. The project’s whole duration highlights the significance of establishing the right balance between data analysis, model testing and quick outcomes.

## 6.1 Timeline and Milestones

Figure 10 represents this project’s organised timeline, making it easier to track the due date of the task and job completions. Major completion dates were achieved for tasks including literature review, Methodology, model testing and writing conclusion. Since work was regularly compared to the scheduled time for each section, it helped to prevent any delays. Appropriate changes were implemented when needed to keep the project on track and ensure a high standard of output.

A graph of a diagram

Description automatically generated with medium confidence

Figure .- Gantt Chart for Project Management

# Chapter 7: Conclusion and Future Work

## 7.1 Conclusion

In conclusion, this study has shown the use of complex machine learning models, in particular Random Forest, for forecasting CO2. Specifically, the study looked into prediction models for CO2 emissions was first conducted for the USA and was successfully replicated for the INDIA dataset. This research can assist environmental scientists, regulators, and business executives, a crucial tool to help them anticipate and control carbon emissions more precisely by using these models. The Random Forest model has proven successful in both USA and INDIA datasets, highlighting its versatility and durability as a dependable model for a range of geographical and economic circumstances.

Governments and organisations may use the research findings to set reasonable targets for reducing carbon emissions, evaluate the effects of changing policies, and put proactive steps in place to slow down global warming. Assuring early intervention and targeted action is essential for fulfilling national and international carbon reduction commitments, like those outlined in the international Paris Accord. Predictive models like the ones used in this study can forecast potential spikes in emissions and can replicated in other countries.

## 7.2 Future Work

To improve the prediction model’s accuracy and applicability, future research should concentrate on improving these prediction models by adding more complex data elements, such as economic factors, rates of renewable energy, and regional policy implications. Moreover, extending the use of these models to other nations or areas with distinct energy and economic patterns may yield important insights into how emissions forecasting might be standardised internationally. The combination of data from carbon capture and storage technologies to forecast its long-term efficacy in reducing emissions might potentially be investigated in more detail. Lastly, collaboration across disciplines that combines environmental policy knowledge with machine learning experience will improve the model prediction utility in creating all Comprehensive, data-driven climate solutions.

# References

Aldy, J. E. (2013) Policy MonitorA Preliminary Assessment of the American Recovery and Reinvestment Act’s Clean Energy Package. *Review of Environmental Economics and Policy*, 7 (1) January, pp. 136–155.

Alizon, F., Shooter, S. B. and Simpson, T. W. (2009) Henry Ford and the Model T: Lessons for Product Platforming and Mass Customization. *Design Studies*, 30 (5) September, pp. 588–605.

Allen, M. R., Friedlingstein, P., Girardin, C. A. J., Jenkins, S., Malhi, Y., Mitchell-Larson, E., Peters, G. P. and Rajamani, L. (2022) Net Zero: Science, Origins, and Implications. *Annual Review of Environment and Resources*, 47 (1) October, pp. 849–887.

Arogyaswamy, B. (2022) Climate Change Mitigation Technologies: Prospects and Challenges. *Modern Applied Science*, 16 (4) August, p. p1.

Attwood, A. R., Washenfelder, R. A., Brock, C. A., Hu, W., Baumann, K., Campuzano-Jost, P., Day, D. A., Edgerton, E. S., Murphy, D. M., Palm, B. B., McComiskey, A., Wagner, N. L., Sá, S. S. de, Ortega, A., Martin, S. T., Jimenez, J. L. and Brown, S. S. (2014) Trends in Sulfate and Organic Aerosol Mass in the Southeast U.S.: Impact on Aerosol Optical Depth and Radiative Forcing. *Geophysical Research Letters*, 41 (21), pp. 7701–7709.

Borowiak, A. R., King, A. D., Brown, J. R., Jones, C. D., Ziehn, T., Meinshausen, M. A. and Cassidy, L. (n.d.) Projected Global Temperature Changes after Net Zero Are Small but Significant. [Online]. Available from: <https://www.authorea.com/users/537888/articles/717362-projected-global-temperature-changes-after-net-zero-are-small-but-significant?commit=df8b8e3d26ab4d0e839ce56644b384fedef4a73a> [Accessed 10 August 2024].

Bratko, I. (1997) Machine Learning: Between Accuracy and Interpretability. In: Della Riccia, G., Lenz, H.-J. and Kruse, R. ed., *Learning, Networks and Statistics*, *1997*. Vienna: Springer, pp. 163–177.

Brownlee, J. (2020) *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python*. Machine Learning Mastery.

Cai, C., Qiu, R. and Tu, Y. (2022) Pulling Off Stable Economic System Adhering Carbon Emissions, Urban Development and Sustainable Development Values. *Frontiers in Public Health* [Online], 10 February. Available from: <https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2022.814656/full> [Accessed 31 July 2024].

Campbell, R. J., Folger, P. and Brown, P. (2013) Prospects for Coal in Electric Power and Industry.

Carley, S. (2016) Energy Programs of the American Recovery and Reinvestment Act of 2009. *Review of Policy Research*, 33 (2), pp. 201–223.

Chai, T. and Draxler, R. R. (2014) Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? – Arguments against Avoiding RMSE in the Literature. *Geoscientific Model Development*, 7 (3) June, pp. 1247–1250.

Chang, D.-S., Yeh, L.-T. and Chen, Y. (2014) The Effects of Economic Development, International Trade, Industrial Structure and Energy Demands on Sustainable Development. *Sustainable Development*, 22 (6), pp. 377–390.

Claudio-Quiroga, G., Gil-Alana, L. A. and Maiza-Larrarte, A. (2023) Persistence and Trends in CO2 Emissions in Africa: Is Chinese FDI behind These Features? *Applied Economics*, 55 (30) June, pp. 3498–3513.

Clements, P. (2021) How Should a Global Carbon Budget Be Estimated? [Online]. Available from: <https://www.researchsquare.com/article/rs-336107/v1> [Accessed 10 August 2024].

Contreras, P., Orellana-Alvear, J., Muñoz, P., Bendix, J. and Célleri, R. (2021) Influence of Random Forest Hyperparameterization on Short-Term Runoff Forecasting in an Andean Mountain Catchment. *Atmosphere*, 12 (2) February, p. 238.

DeCarolis, J., Daly, H., Dodds, P., Keppo, I., Li, F., McDowall, W., Pye, S., Strachan, N., Trutnevyte, E., Usher, W., Winning, M., Yeh, S. and Zeyringer, M. (2017) Formalizing Best Practice for Energy System Optimization Modelling. *Applied Energy*, 194 May, pp. 184–198.

Edelenbosch, O., Berg, M. van den, Boer, H. S. de, Chen, H., Daioglou, V., Dekker, M., Doelman, J., Elzen, M. D., Harmsen, M., Hof, A., Mikropoulos, E., Sluisveld, M. van, Stehfest, E., Tagomori, I. S., Zeist, W.-J. van and Vuuren, D. van (2023) Reducing Sectoral Hard to Abate Emissions to Limit Reliance of Carbon Dioxide Removal in 1.5°C Scenarios [Online]. Available from: <https://www.researchsquare.com/article/rs-3182402/v1> [Accessed 9 August 2024].

Egilmez, G., Mohamed Abdul Ghani, N. M. A. and Gedik, R. (2017) Layer Analysis of CO2 Sources in the US Economic Supply Chains: An Input Output LCA Study. *Industrial Management & Data Systems*, 117 (10) January, pp. 2171–2193.

Erickson, B. J. and Kitamura, F. (2021) Magician’s Corner: 9. Performance Metrics for Machine Learning Models. *Radiology: Artificial Intelligence*, 3 (3) May, p. e200126.

Feder, S. L. (2018) Data Quality in Electronic Health Records Research: Quality Domains and Assessment Methods. *Western Journal of Nursing Research*, 40 (5) May, pp. 753–766.

Frieler, K., Lange, S., Piontek, F., Reyer, C. P. O., Schewe, J., Warszawski, L., Zhao, F., Chini, L., Denvil, S., Emanuel, K., Geiger, T., Halladay, K., Hurtt, G., Mengel, M., Murakami, D., Ostberg, S., Popp, A., Riva, R., Stevanovic, M., Suzuki, T., Volkholz, J., Burke, E., Ciais, P., Ebi, K., Eddy, T. D., Elliott, J., Galbraith, E., Gosling, S. N., Hattermann, F., Hickler, T., Hinkel, J., Hof, C., Huber, V., Jägermeyr, J., Krysanova, V., Marcé, R., Müller Schmied, H., Mouratiadou, I., Pierson, D., Tittensor, D. P., Vautard, R., Vliet, M. van, Biber, M. F., Betts, R. A., Bodirsky, B. L., Deryng, D., Frolking, S., Jones, C. D., Lotze, H. K., Lotze-Campen, H., Sahajpal, R., Thonicke, K., Tian, H. and Yamagata, Y. (2017) Assessing the Impacts of 1.5 °C Global Warming – Simulation Protocol of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2b). *Geoscientific Model Development*, 10 (12) November, pp. 4321–4345.

Fujii, T., Sako, M., Ishihama, K., Kohno, Y., Makino, T., Yasuo, N. and Kawauchi, S. (2024) Prediction of CO2 Absorbing Performance of Amine Aqueous Solution Using Random Forest Models. *Gas Science and Engineering*, 129 September, p. 205417.

Gassiat, C., Gleeson, T., Lefebvre, R. and McKenzie, J. (2013) Hydraulic Fracturing in Faulted Sedimentary Basins: Numerical Simulation of Potential Contamination of Shallow Aquifers over Long Time Scales. *Water Resources Research*, 49 (12), pp. 8310–8327.

Guzanek, P. and Borucka, A. (2021) An Analysis of Factors Affecting the Number of Safety Incidents in Civil Aviation. *Safety & Defense*, 7 (2) October, pp. 105–118.

Hall, J. L. and Jennings, E. T. (2011) The American Recovery and Reinvestment Act (ARRA): A Critical Examination of Accountability and Transparency in the Obama Administration. *Public Performance & Management Review*, 35 (1) September, pp. 202–226.

Han, J., Lin, H. and Qin, Z. (2023) Prediction and Comparison of In-Vehicle CO2 Concentration Based on ARIMA and LSTM Models. *Applied Sciences*, 13 (19) January, p. 10858.

Harati, S., Rezaei Gomari, S., Rahman, M. A., Hassan, R., Hassan, I., Sleiti, A. K. and Hamilton, M. (2024) Performance Analysis of Various Machine Learning Algorithms for CO2 Leak Prediction and Characterization in Geo-Sequestration Injection Wells. *Process Safety and Environmental Protection*, 183 March, pp. 99–110.

Höhne, N., Gidden, M. J., Elzen, M. den, Hans, F., Fyson, C., Geiges, A., Jeffery, M. L., Gonzales-Zuñiga, S., Mooldijk, S., Hare, W. and Rogelj, J. (2021) Wave of Net Zero Emission Targets Opens Window to Meeting the Paris Agreement. *Nature Climate Change*, 11 (10) October, pp. 820–822.

Hutton, P. C., Mikhailova, E. A., Lin, L., Hao, Z., Zurqani, H. A., Post, C. J., Schlautman, M. A. and Shepherd, G. B. (2023) Net-Zero Target and Emissions from Land Conversions: A Case Study of Maryland’s Climate Solutions Now Act. *Geographies*, 3 (1) March, pp. 40–59.

Hyndman, R. J. and Koehler, A. B. (2006) Another Look at Measures of Forecast Accuracy. *International Journal of Forecasting*, 22 (4) October, pp. 679–688.

Jaggia, S., Kelly, A., Lertwachara, K. and Chen, L. (2020) Applying the CRISP-DM Framework for Teaching Business Analytics. *Decision Sciences Journal of Innovative Education*, 18 (4), pp. 612–634.

Jamil, F., Shafiq, I., Sarwer, A., Ahmad, M., Akhter, P., Inayat, A., Shafique, S., Park, Y.-K. and Hussain, M. (2024) A Critical Review on the Effective Utilization of Geothermal Energy. *Energy & Environment*, 35 (1) February, pp. 438–457.

Jenn, A., Azevedo, I. M. L. and Michalek, J. J. (2016) Alternative Fuel Vehicle Adoption Increases Fleet Gasoline Consumption and Greenhouse Gas Emissions under United States Corporate Average Fuel Economy Policy and Greenhouse Gas Emissions Standards. *Environmental Science & Technology*, 50 (5) March, pp. 2165–2174.

Ji, Z., Song, H., Lei, L., Sheng, M., Guo, K. and Zhang, S. (2024) A Novel Approach for Predicting Anthropogenic CO2 Emissions Using Machine Learning Based on Clustering of the CO2 Concentration. *Atmosphere*, 15 (3) March, p. 323.

Jin, H. (2022) Hyperparameter Importance for Machine Learning Algorithms [Online]. arXiv. Available from: <http://arxiv.org/abs/2201.05132> [Accessed 28 August 2024].

Kafi, M., Sanaeepur, H. and Ebadi Amooghin, A. (2023) Grand Challenges in CO2 Capture and Conversion. *Journal of Resource Recovery* [Online], 1 (January-December) January. Available from: <https://www.resrecov.com/article\_703896.html> [Accessed 9 August 2024].

Khalifa, A. A., Ibrahim, A.-J., Amhamed, A. I. and El-Naas, M. H. (2022) Accelerating the Transition to a Circular Economy for Net-Zero Emissions by 2050: A Systematic Review. *Sustainability*, 14 (18) January, p. 11656.

Kohavi, R. (2001) A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. 14 March.

Kumari, S. and Singh, S. K. (2023) Machine Learning-Based Time Series Models for Effective CO2 Emission Prediction in India. *Environmental Science and Pollution Research*, 30 (55) November, pp. 116601–116616.

Lade, G. E. and Bushnell, J. (2019) Fuel Subsidy Pass-Through and Market Structure: Evidence from the Renewable Fuel Standard. *Journal of the Association of Environmental and Resource Economists*, 6 (3) May, pp. 563–592.

Lade, G. E., Cynthia Lin Lawell, C.-Y. and Smith, A. (2018) Designing Climate Policy: Lessons from the Renewable Fuel Standard and the Blend Wall. *American Journal of Agricultural Economics*, 100 (2), pp. 585–599.

Lark, T. J., Hendricks, N. P., Smith, A., Pates, N., Spawn-Lee, S. A., Bougie, M., Booth, E. G., Kucharik, C. J. and Gibbs, H. K. (2022) Environmental Outcomes of the US Renewable Fuel Standard. *Proceedings of the National Academy of Sciences*, 119 (9) March, p. e2101084119.

Lepri, B., Staiano, J., Sangokoya, D., Letouzé, E. and Oliver, N. (2017) The Tyranny of Data? The Bright and Dark Sides of Data-Driven Decision-Making for Social Good [Online]. In: Cerquitelli, T., Quercia, D. and Pasquale, F. ed., *Transparent Data Mining for Big and Small Data*. Cham: Springer International Publishing, pp. 3–24. Available from: <https://doi.org/10.1007/978-3-319-54024-5\_1> [Accessed 10 August 2024].

Lewis, R. (2007) World War II Manufacturing and the Postwar Southern Economy. *The Journal of Southern History*, 73 (4), pp. 837–866.

Li, C., Yang, R., Zhao, G., Yue, Y., Zhang, W., Zhang, H., Zhang, Y., Li, S. and Zhou, H. (2024) The Key Role of Hole Transport Layer in Boosting Performance of Two-Dimensional Perovskite Solar Cells. *Solar RRL*, 8 (11), p. 2400201.

Li, X., Zhang, X., Zhang, J., Gu, J., Zhang, S., Li, G., Shao, J., He, Y., Yang, H., Zhang, S. and Chen, H. (2023) Applied Machine Learning to Analyze and Predict CO2 Adsorption Behavior of Metal-Organic Frameworks. *Carbon Capture Science & Technology*, 9 December, p. 100146.

Liu, H., Ampah, J., JIN, C., YAO, M., AFRANE, S., ADUN, H., Fuhrman, J., Ho, D. T. and McJeon, H. (2023) Deployment Expectations of Multi-Gigaton Scale of Carbon Dioxide Removal Could Have Adverse Impacts on Global Climate System [Online]. Available from: <https://www.researchsquare.com/article/rs-3593971/v1> [Accessed 10 August 2024].

Lund, M. T., Aamaas, B., Berntsen, T., Bock, L., Burkhardt, U., Fuglestvedt, J. S. and Shine, K. P. (2017) Emission Metrics for Quantifying Regional Climate Impacts of Aviation. *Earth System Dynamics*, 8 (3) July, pp. 547–563.

Lynch, J. A., Bowersox, V. C. and Grimm, J. W. (2000) Acid Rain Reduced in Eastern United States. *Environmental Science & Technology*, 34 (6) March, pp. 940–949.

MacLeod, M. (2021) *The Global Threat from Plastic Pollution* [Online]. Available from: <https://www.science.org/doi/10.1126/science.abg5433> [Accessed 31 July 2024].

Malik, A., Hussain, E., Baig, S. and Khokhar, M. F. (2020) Forecasting CO2 Emissions from Energy Consumption in Pakistan under Different Scenarios: The China–Pakistan Economic Corridor. *Greenhouse Gases: Science and Technology*, 10 (2), pp. 380–389.

Manzella, A. (2017) Geothermal Energy. *EPJ Web of Conferences*, 148, p. 00012.

Matthews, H. D. and Caldeira, K. (2008) Stabilizing Climate Requires Near-Zero Emissions. *Geophysical Research Letters* [Online], 35 (4). Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1029/2007GL032388> [Accessed 16 August 2024].

McCarthy, T. (2007) *Auto Mania: Cars, Consumers, and the Environment*. Yale University Press.

Michalakopoulos, V., Ilias, L., Kapsalis, P., Mouzakitis, S. and Askounis, D. (2023) Comparison of Machine Learning Algorithms For Predicting CO2Emissions in the Maritime Domain [Online]. In: *2023 14th International Conference on Information, Intelligence, Systems & Applications (IISA)*, *July 2023*. pp. 1–4. Available from: <https://ieeexplore.ieee.org/document/10345936> [Accessed 3 August 2024].

Mitchell, L. E., Lin, J. C., Bowling, D. R., Pataki, D. E., Strong, C., Schauer, A. J., Bares, R., Bush, S. E., Stephens, B. B., Mendoza, D., Mallia, D., Holland, L., Gurney, K. R. and Ehleringer, J. R. (2018) Long-Term Urban Carbon Dioxide Observations Reveal Spatial and Temporal Dynamics Related to Urban Characteristics and Growth. *Proceedings of the National Academy of Sciences*, 115 (12) March, pp. 2912–2917.

Moosdorf, N., Renforth, P. and Hartmann, J. (2014) Carbon Dioxide Efficiency of Terrestrial Enhanced Weathering. *Environmental Science & Technology*, 48 (9) May, pp. 4809–4816.

Nolan, C., Overpeck, J. T., Allen, J. R. M., Anderson, P. M., Betancourt, J. L., Binney, H. A., Brewer, S., Bush, M. B., Chase, B. M., Cheddadi, R., Djamali, M., Dodson, J., Edwards, M. E., Gosling, W. D., Haberle, S., Hotchkiss, S. C., Huntley, B., Ivory, S. J., Kershaw, A. P., Kim, S.-H., Latorre, C., Leydet, M., Lézine, A.-M., Liu, K.-B., Liu, Y., Lozhkin, A. V., McGlone, M. S., Marchant, R. A., Momohara, A., Moreno, P. I., Müller, S., Otto-Bliesner, B. L., Shen, C., Stevenson, J., Takahara, H., Tarasov, P. E., Tipton, J., Vincens, A., Weng, C., Xu, Q., Zheng, Z. and Jackson, S. T. (2018) Past and Future Global Transformation of Terrestrial Ecosystems under Climate Change. *Science*, 361 (6405) August, pp. 920–923.

Noor, N. M., Abdullah, M. M. A. B., Yahaya, A. S. and Ramli, N. A. (2015) Comparison of Linear Interpolation Method and Mean Method to Replace the Missing Values in Environmental Data Set. *Materials Science Forum*, 803, pp. 278–281.

Olanrewaju, O. I. K., Oduro, P. and Babayeju, O. A. (2024) Exploring Capital Market Innovations for Net Zero Goals: A Data-Driven Investment Approach. *Finance & Accounting Research Journal*, 6 (6) June, pp. 1091–1104.

Osborne, J. (2002) Notes on the use of data transformations. *Practical Assessment, Research, and Evaluation* [Online], 8 (1) January. Available from: <https://openpublishing.library.umass.edu/pare/article/id/1455/> [Accessed 2 September 2024].

Pandey, D., Agrawal, M. and Pandey, J. S. (2011) Carbon Footprint: Current Methods of Estimation. *Environmental Monitoring and Assessment*, 178 (1) July, pp. 135–160.

Pei, F., Zhong, R., Liu, L.-A. and Qiao, Y. (2021) Decoupling the Relationships between Carbon Footprint and Economic Growth within an Urban Agglomeration—A Case Study of the Yangtze River Delta in China. *Land*, 10 (9) September, p. 923.

Piggot, G., Erickson, P., Asselt, H. van and Lazarus, M. (2018) Swimming Upstream: Addressing Fossil Fuel Supply under the UNFCCC. *Climate Policy*, 18 (9) October, pp. 1189–1202.

Popp, D., Vona, F., Marin, G. and Chen, Z. (n.d.) The Employment Impact of Green Fiscal Push: Evidence from the American Recovery Act.

Rifai, M. A., Nuryartono, N. and Irfany, M. I. (2018) Carbon Footprint Based on Household Consumption: Case Study on Cocoa Farmer’s Household in Polewali Mandar. *Journal of Sustainable Development*, 11 (6) November, p. p15.

Ritchie, H. and Roser, M. (2024) CO₂ Emissions. *Our World in Data* [Online], January. Available from: <https://ourworldindata.org/co2-emissions> [Accessed 19 August 2024].

Saltz, J. S. (2021) CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps [Online]. In: *2021 IEEE International Conference on Big Data (Big Data)*, *December 2021*. pp. 2337–2344. Available from: <https://ieeexplore.ieee.org/abstract/document/9671634> [Accessed 1 September 2024].

Schurr, S. H. (1960) *Energy in the American Economy, 1850-1975; an Economic Study of Its History and Prospects c.2.* [Online]. HathiTrust. Available from: <https://hdl.handle.net/2027/uc1.b3786242?urlappend=%3Bseq=7> [Accessed 31 July 2024].

Si, M. and Du, K. (2020) Development of a Predictive Emissions Model Using a Gradient Boosting Machine Learning Method. *Environmental Technology & Innovation*, 20 November, p. 101028.

Soest, H. L. van, Elzen, M. G. J. den and Vuuren, D. P. van (2021) Net-Zero Emission Targets for Major Emitting Countries Consistent with the Paris Agreement. *Nature Communications*, 12 (1) April, p. 2140.

Srinivasan, D. and Tettamanzi, A. G. B. (1997) An Evolutionary Algorithm for Evaluation of Emission Compliance Options in View of the Clean Air Act Amendments. *IEEE Transactions on Power Systems*, 12 (1) February, pp. 336–341.

Stradling, D. and Thorsheim, P. (1999) The Smoke of Great Cities: British and American Efforts to Control Air Pollution, 1860–1914. *Environmental History*, 4 (1) January, pp. 6–31.

Summary for Policymakers IPBES Global Assessment.Pdf (n.d.) [Online]. Available from: <https://www.mari-odu.org/academics/2018su\_Leadership/commons/library/Summary%20for%20Policymakers%20IPBES%20Global%20Assessment.pdf> [Accessed 19 May 2024].

Tan, C. H., Yaw, C. T., Koh, S. P., Kadirgama, K., Abed, A. M. and Majdi, H. S. (2024) Sustainable Bio-Jet Fuel Synthesis Techniques for the Aviation Industry. *IOP Conference Series: Earth and Environmental Science*, 1372 (1) July, p. 012036.

Tan, H. (2022) *Carbon Emission Prediction with Macroeconomic Variables and Machine Learning | IEEE Conference Publication | IEEE Xplore* [Online]. Available from: <https://ieeexplore.ieee.org/document/9976625> [Accessed 3 August 2024].

Trisovic, A., Mika, K., Boyd, C., Feger, S. and Crosas, M. (2021) Repository Approaches to Improving the Quality of Shared Data and Code. *Data*, 6 (2) February, p. 15.

Tunnicliffe Wilson, G. (2016) Time Series Analysis: Forecasting and Control,5th Edition, by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung, 2015. Published by John Wiley and Sons Inc., Hoboken, New Jersey, Pp. 712. ISBN: 978-1-118-67502-1. *Journal of Time Series Analysis*, 37 March, p. n/a-n/a.

US EPA, O. (2015) *Overview of the Renewable Fuel Standard Program* [Online]. Available from: <https://www.epa.gov/renewable-fuel-standard-program/overview-renewable-fuel-standard-program> [Accessed 1 August 2024].

Wang, H., Luo, J., Zhang, M. and Ling, Y. (2022) The Impact of Transportation Restructuring on the Intensity of Greenhouse Gas Emissions: Empirical Data from China. *International Journal of Environmental Research and Public Health*, 19 (19) January, p. 12960.

Wang, S., Fang, C., Ma, H., Wang, Y. and Qin, J. (2014) Spatial Differences and Multi-Mechanism of Carbon Footprint Based on GWR Model in Provincial China. *Journal of Geographical Sciences*, 24 (4) August, pp. 612–630.

Wang, X., Wang, S. and Zhang, Y. (2022) The Impact of Environmental Regulation and Carbon Emissions on Green Technology Innovation from the Perspective of Spatial Interaction: Empirical Evidence from Urban Agglomeration in China. *Sustainability*, 14 (9) January, p. 5381.

Weiner, I. B., Schinka, J. A. and Velicer, W. F. (2012) *Handbook of Psychology, Research Methods in Psychology*. John Wiley & Sons.

Williams, A. T., Sperl, R. E. and Chung, S. M. (2023) Anomaly Detection in Multi-Seasonal Time Series Data. *IEEE Access*, 11, pp. 106456–106464.

Williams, J. H., Jones, R. A., Haley, B., Kwok, G., Hargreaves, J., Farbes, J. and Torn, M. S. (2021) Carbon-Neutral Pathways for the United States. *AGU Advances*, 2 (1), p. e2020AV000284.

Wong, F. (2022) Carbon Emissions Allowances Trade Amount Dynamic Prediction Based on Machine Learning [Online]. In: *2022 International Conference on Machine Learning and Knowledge Engineering (MLKE)*, *February 2022*. pp. 115–120. Available from: <https://ieeexplore.ieee.org/document/9763576> [Accessed 3 August 2024].

Xue, L. (2024) Historical and Future Global Warming under Different RCP Emission Scenarios. *Applied and Computational Engineering*, 63 (1) May, pp. 34–44.

Zakkour, P. D., Heidug, W., Howard, A., Stuart Haszeldine, R., Allen, M. R. and Hone, D. (2021) Progressive Supply-Side Policy under the Paris Agreement to Enhance Geological Carbon Storage. *Climate Policy*, 21 (1) January, pp. 63–77.

Zhang, Y., Liu, X., Lei, L. and Liu, L. (2022) Estimating Global Anthropogenic CO2 Gridded Emissions Using a Data-Driven Stacked Random Forest Regression Model. *Remote Sensing*, 14 (16) January, p. 3899.

Zhao, T., Zheng, Y. and Wu, Z. (2022) Improving Computational Efficiency of Machine Learning Modeling of Nonlinear Processes Using Sensitivity Analysis and Active Learning. *Digital Chemical Engineering*, 3 June, p. 100027.

Zheng, A. and Casari, A. (2018) *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O’Reilly Media, Inc.

Zheng, J. and Suh, S. (2019) Strategies to Reduce the Global Carbon Footprint of Plastics. *Nature Climate Change*, 9 (5) May, pp. 374–378.

Zhong, S. and Huang, C. (2019) Climate Change and Human Health: Risks and Responses. *Chinese Science Bulletin*, 64 (19) February, pp. 2002–2010.

Zimmerman, N., Wang, J. M., Jeong, C.-H., Wallace, J. S. and Evans, G. J. (2016) Assessing the Climate Trade-Offs of Gasoline Direct Injection Engines. *Environmental Science & Technology*, 50 (15) August, pp. 8385–8392.

# Appendices

## Code for ETL

energy\_related\_columns = ['year',

'co2\_per\_unit\_energy',

'co2\_including\_luc\_per\_unit\_energy',

'coal\_co2',

'cumulative\_coal\_co2',

'flaring\_co2',

'cumulative\_flaring\_co2',

'gas\_co2',

'cumulative\_gas\_co2',

'oil\_co2',

'cumulative\_oil\_co2',

'co2',

'co2\_per\_capita',

'co2\_per\_gdp',

'coal\_co2\_per\_capita',

'flaring\_co2\_per\_capita',

'gas\_co2\_per\_capita',

'oil\_co2\_per\_capita',

'temperature\_change\_from\_co2'

]

# Create a new data frame with the selected columns

energy\_data = data[energy\_related\_columns]

# Display the first few rows of the dataframe

print(energy\_data.head())

missing\_values = energy\_data.isnull().sum()

# Plotting the missing value histogram

plt.figure(figsize=(12, 6))

missing\_values.plot(kind='bar', color='skyblue')

plt.title('Histogram of Missing Values per Selected Column')

plt.xlabel('Columns')

plt.ylabel('Number of Missing Values')

plt.xticks(rotation=45, ha='right')

plt.grid(axis='y')

plt.show()

# Remove columns with more than 50% missing values

threshold = 0.5

energy\_data = energy\_data.loc[:, energy\_data.isnull().mean() < threshold]

# Display the remaining columns after filtering

remaining\_columns = energy\_data.columns.tolist()

print("Remaining columns after removing those with more than 50% missing values:")

print(remaining\_columns)

# Display the filtered data

print("Filtered data after removing columns with more than 50% missing values:")

print(energy\_data.head())

# Checking the Skewness of the data

energy\_data.skew()

# Calculate the mean for co2\_per\_gdp and temperature\_change\_from\_co2

co2\_per\_gdp\_mean = energy\_data['co2\_per\_gdp'].mean()

temperature\_change\_from\_co2\_mean = energy\_data['temperature\_change\_from\_co2'].mean()

# Impute missing values for co2\_per\_gdp and temperature\_change\_from\_co2 using the mean

energy\_data['co2\_per\_gdp'] = energy\_data['co2\_per\_gdp'].fillna(co2\_per\_gdp\_mean)

energy\_data['temperature\_change\_from\_co2'] = energy\_data['temperature\_change\_from\_co2'].fillna(temperature\_change\_from\_co2\_mean)

# Check the result

print("Remaining missing values after imputation:")

print(energy\_data.isnull().sum())

# Visualising the Skewness of the columns

def visualize\_skewness(df, title):

plt.figure(figsize=(15, 20))

for i, column in enumerate(df.columns, 1):

plt.subplot(6, 3, i)

sns.histplot(df[column].dropna(), kde=True)

plt.title(f'Distribution of {column}')

plt.suptitle(title, fontsize=16)

plt.tight\_layout(rect=[0, 0, 1, 0.96])

plt.show()

plt.figure(figsize=(15, 20))

for i, column in enumerate(df.columns, 1):

plt.subplot(6, 3, i)

sns.boxplot(y=df[column])

plt.title(f'Box plot of {column}')

plt.suptitle(title, fontsize=16)

plt.tight\_layout(rect=[0, 0, 1, 0.96])

plt.show()

# Visualize skewness before transformation

visualize\_skewness(energy\_data, "Original Data Distribution and Skewness")

# Handling the Skewness of the data

from sklearn.preprocessing import PowerTransformer

# Define a function to handle skewness

def handle\_skewness(df):

pt = PowerTransformer(method='yeo-johnson') # Yeo-Johnson works for both positive and negative values

for col in df.columns:

if df[col].skew() > 1 or df[col].skew() < -1: # If the column is highly skewed

df[col] = pt.fit\_transform(df[[col]])

return df

# Apply the skewness handling function

energy\_data\_transformed = handle\_skewness(energy\_data.copy())

# Code for EDA (Exploratory Data Analysis)

# CO2 emission over time

custom\_ticks = [1800, 1850, 1900, 1950, 2000, 2022]

plt.figure(figsize=(12, 6))

plt.plot(energy\_data['year'], energy\_data['co2'], label='Total CO2')

plt.plot(energy\_data['year'], energy\_data['coal\_co2'], label='Coal CO2')

plt.plot(energy\_data['year'], energy\_data['gas\_co2'], label='Gas CO2')

plt.plot(energy\_data['year'], energy\_data['oil\_co2'], label='Oil CO2')

plt.xlabel('Year')

plt.ylabel('CO2 Emissions (Million Tonnes)')

plt.title('CO2 Emissions Over Time')

plt.legend()

plt.xticks(custom\_ticks)

plt.show()

# Per Capita and Per GDP Emissions Over Time

plt.figure(figsize=(12, 6))

plt.plot(energy\_data['year'], energy\_data['co2\_per\_capita'], label='CO2 per Capita')

plt.plot(energy\_data['year'], energy\_data['co2\_per\_gdp'], label='CO2 per GDP')

plt.xlabel('Year')

plt.ylabel('CO2 Emissions')

plt.title('CO2 Emissions Per Capita and Per GDP Over Time')

plt.legend()

plt.xticks(custom\_ticks)

plt.show()

# Corelation Matrix with the use of Heat Map

correlation\_matrix = energy\_data.corr()

plt.figure(figsize=(20,5))

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm', center=0)

plt.title('Correlation Matrix of Energy Related Columns')

plt.show()

# Co2 emission by Source Over Time

plt.figure(figsize=(12, 6))

plt.plot(energy\_data['year'], energy\_data['coal\_co2'], label='Coal CO2 Emissions')

plt.plot(energy\_data['year'], energy\_data['gas\_co2'], label='Gas CO2 Emissions')

plt.plot(energy\_data['year'], energy\_data['oil\_co2'], label='Oil CO2 Emissions')

plt.xlabel('Year')

plt.ylabel('CO2 Emissions')

plt.title('CO2 Emissions by Source Over Time')

plt.legend()

plt.grid(True)

plt.show()

# Historical Comparision: Cumilative Emissions Over Time

plt.figure(figsize=(12, 6))

plt.plot(energy\_data['year'], energy\_data['cumulative\_coal\_co2'], label='Cumulative Coal CO2')

plt.plot(energy\_data['year'], energy\_data['cumulative\_gas\_co2'], label='Cumulative Gas CO2')

plt.plot(energy\_data['year'], energy\_data['cumulative\_oil\_co2'], label='Cumulative Oil CO2')

plt.xlabel('Year')

plt.ylabel('Cumulative CO2 Emissions (Million Tonnes)')

plt.title('Cumulative CO2 Emissions Over Time')

plt.legend()

plt.xticks(custom\_ticks)

plt.show()

# CO2 emission Per GDP Over Time

plt.figure(figsize=(12, 6))

plt.plot(energy\_data['year'], energy\_data['co2\_per\_gdp'], label='CO2 Emissions Per GDP')

plt.xlabel('Year')

plt.ylabel('CO2 Emissions Per GDP')

plt.title('CO2 Emissions Per GDP Over Time')

plt.legend()

plt.grid(True)

plt.xticks(custom\_ticks)

plt.show()

# Per capita CO2 emission by Source Over Time

plt.figure(figsize=(12, 6))

plt.plot(energy\_data['year'], energy\_data['coal\_co2\_per\_capita'], label='Coal CO2 Per Capita')

plt.plot(energy\_data['year'], energy\_data['gas\_co2\_per\_capita'], label='Gas CO2 Per Capita')

plt.plot(energy\_data['year'], energy\_data['oil\_co2\_per\_capita'], label='Oil CO2 Per Capita')

plt.xlabel('Year')

plt.ylabel('CO2 Emissions Per Capita')

plt.title('Per Capita CO2 Emissions by Source Over Time')

plt.legend()

plt.grid(True)

plt.xticks(custom\_ticks)

plt.show()

# Distribution of Cumulative CO Emissions by Source

plt.figure(figsize=(10, 7))

cumulative\_emissions = [

energy\_data['cumulative\_coal\_co2'].iloc[-1],

energy\_data['cumulative\_gas\_co2'].iloc[-1],

energy\_data['cumulative\_oil\_co2'].iloc[-1]

]

labels = ['Cumulative Coal CO2', 'Cumulative Gas CO2', 'Cumulative Oil CO2']

plt.pie(cumulative\_emissions, labels=labels, autopct='%1.1f%%', startangle=140, colors=['blue', 'orange', 'green'])

plt.title('Distribution of Cumulatiove CO2 Emissions by Source')

plt.show()

# Distribution of Cumulative CO2 Emissions by Source

plt.figure(figsize=(10, 7))

cumulative\_emissions = [

energy\_data['cumulative\_coal\_co2'].iloc[-1],

energy\_data['cumulative\_gas\_co2'].iloc[-1],

energy\_data['cumulative\_oil\_co2'].iloc[-1]

]

labels = ['Cumulative Coal CO2', 'Cumulative Gas CO2', 'Cumulative Oil CO2']

plt.pie(cumulative\_emissions, labels=labels, autopct='%1.1f%%', startangle=140, colors=['blue', 'orange', 'green'])

plt.title('Distribution of Cumulatiove CO2 Emissions by Source')

plt.show()

# CO2 Emissions and 5-year Moving Average

data['co2\_moving\_avg'] = data['co2'].rolling(window=5).mean()

# Plot moving average

plt.figure(figsize=(15, 7))

plt.plot(data['year'], data['co2'], marker='o', linestyle='-', label='CO2 Emissions')

plt.plot(data['year'], data['co2\_moving\_avg'], color='red', label='5-Year Moving Average')

plt.title('CO2 Emissions and 5-Year Moving Average')

plt.xlabel('Year')

plt.ylabel('CO2 Emissions')

plt.legend()

plt.xticks(custom\_ticks)

plt.grid(True)

plt.show()

# Temperature Change from CO2 over Time

plt.figure(figsize=(10, 6))

plt.plot(energy\_data['year'], energy\_data['temperature\_change\_from\_co2'], marker='o', linestyle='-', color='b')

# Adding titles and labels

plt.title('Temperature Change from CO2 Over Time')

plt.xlabel('Year')

plt.ylabel('Temperature Change (°C)')

plt.xticks(custom\_ticks)

# Displaying the plot

plt.grid(True)

plt.show()

## Predictive Model Code for Random Forest

from sklearn.ensemble import RandomForestRegressor

from sklearn.model\_selection import GridSearchCV, train\_test\_split

from sklearn.metrics import r2\_score, mean\_squared\_error, mean\_absolute\_percentage\_error

import numpy as np

# Load the cleaned dataset

df\_cleaned = pd.read\_csv('cleaned\_data.csv')

# List of columns to predict (excluding the 'year' column and other non-target columns)

columns\_to\_predict = ['coal\_co2', 'cumulative\_coal\_co2', 'flaring\_co2', 'cumulative\_flaring\_co2',

'gas\_co2', 'cumulative\_gas\_co2', 'oil\_co2', 'cumulative\_oil\_co2',

'co2', 'co2\_per\_capita', 'co2\_per\_gdp', 'coal\_co2\_per\_capita',

'flaring\_co2\_per\_capita', 'gas\_co2\_per\_capita', 'oil\_co2\_per\_capita',

'temperature\_change\_from\_co2']

# Set up hyperparameters for tuning

param\_grid = {

'n\_estimators': [50, 100, 200],

'max\_depth': [5, 10, 20],

'min\_samples\_split': [2, 5, 10]

}

# Initialize lists to store overall metrics for all target columns

train\_r2\_list = []

test\_r2\_list = []

train\_rmse\_list = []

test\_rmse\_list = []

train\_mape\_list = []

test\_mape\_list = []

# Iterate through each column to predict

for target\_col in columns\_to\_predict:

print(f"Tuning for {target\_col}...")

# Define features (X) and target variable (y)

X = df\_cleaned.drop(columns=[target\_col])

y = df\_cleaned[target\_col]

# Split the data into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Initialize RandomForestRegressor

rf\_model = RandomForestRegressor(random\_state=42)

# GridSearchCV to find the best hyperparameters

grid\_search = GridSearchCV(estimator=rf\_model, param\_grid=param\_grid, cv=3, n\_jobs=-1, scoring='r2')

grid\_search.fit(X\_train, y\_train)

# Best model after hyperparameter tuning

best\_rf\_model = grid\_search.best\_estimator\_

# Make predictions on both training and testing sets

y\_train\_pred = best\_rf\_model.predict(X\_train)

y\_test\_pred = best\_rf\_model.predict(X\_test)

# Calculate R², RMSE, and MAPE for training and testing

r2\_train = r2\_score(y\_train, y\_train\_pred)

rmse\_train = np.sqrt(mean\_squared\_error(y\_train, y\_train\_pred))

mape\_train = mean\_absolute\_percentage\_error(y\_train, y\_train\_pred)

r2\_test = r2\_score(y\_test, y\_test\_pred)

rmse\_test = np.sqrt(mean\_squared\_error(y\_test, y\_test\_pred))

mape\_test = mean\_absolute\_percentage\_error(y\_test, y\_test\_pred)

# Append results

train\_r2\_list.append(r2\_train)

train\_rmse\_list.append(rmse\_train)

train\_mape\_list.append(mape\_train)

test\_r2\_list.append(r2\_test)

test\_rmse\_list.append(rmse\_test)

test\_mape\_list.append(mape\_test)

# Calculate overall average scores

overall\_train\_r2 = np.mean(train\_r2\_list)

overall\_test\_r2 = np.mean(test\_r2\_list)

overall\_train\_rmse = np.mean(train\_rmse\_list)

overall\_test\_rmse = np.mean(test\_rmse\_list)

overall\_train\_mape = np.mean(train\_mape\_list)

overall\_test\_mape = np.mean(test\_mape\_list)

# Print overall scores

print("\nOverall Training Performance After Tuning:")

print(f"Average R²: {overall\_train\_r2}")

print(f"Average RMSE: {overall\_train\_rmse}")

print(f"Average MAPE: {overall\_train\_mape}")

print("\nOverall Testing Performance After Tuning:")

print(f"Average R²: {overall\_test\_r2}")

print(f"Average RMSE: {overall\_test\_rmse}")

print(f"Average MAPE: {overall\_test\_mape}")

## Predictive Model Code for XGBoost

import pandas as pd

import xgboost as xgb

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import r2\_score, mean\_squared\_error, mean\_absolute\_percentage\_error

import numpy as np

# Load the cleaned dataset

df\_cleaned = pd.read\_csv('cleaned\_data.csv')

# List of columns to predict (excluding the 'year' column)

columns\_to\_predict = ['coal\_co2', 'cumulative\_coal\_co2', 'flaring\_co2', 'cumulative\_flaring\_co2',

'gas\_co2', 'cumulative\_gas\_co2', 'oil\_co2', 'cumulative\_oil\_co2',

'co2', 'co2\_per\_capita', 'co2\_per\_gdp', 'coal\_co2\_per\_capita',

'flaring\_co2\_per\_capita', 'gas\_co2\_per\_capita', 'oil\_co2\_per\_capita',

'temperature\_change\_from\_co2']

# Initialize lists to store overall metrics for all target columns

train\_r2\_list = []

test\_r2\_list = []

train\_rmse\_list = []

test\_rmse\_list = []

train\_mape\_list = []

test\_mape\_list = []

# Loop through each column to predict using XGBoost

for target\_col in columns\_to\_predict:

print(f"Training XGBoost model with 38 boosting rounds for {target\_col}...")

# Define features (X) and target variable (y)

X = df\_cleaned.drop(columns=[target\_col])

y = df\_cleaned[target\_col]

# Split the data into training and testing sets (80% train, 20% test)

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Initialize XGBoost Regressor with 38 boosting rounds

xgb\_model = xgb.XGBRegressor(n\_estimators=38, max\_depth=5, learning\_rate=0.1, random\_state=42)

# Train the XGBoost model

xgb\_model.fit(X\_train, y\_train)

# Make predictions on both training and test sets

y\_train\_pred = xgb\_model.predict(X\_train)

y\_test\_pred = xgb\_model.predict(X\_test)

# Calculate R², RMSE, and MAPE for training and testing

r2\_train = r2\_score(y\_train, y\_train\_pred)

rmse\_train = np.sqrt(mean\_squared\_error(y\_train, y\_train\_pred))

mape\_train = mean\_absolute\_percentage\_error(y\_train, y\_train\_pred)

r2\_test = r2\_score(y\_test, y\_test\_pred)

rmse\_test = np.sqrt(mean\_squared\_error(y\_test, y\_test\_pred))

mape\_test = mean\_absolute\_percentage\_error(y\_test, y\_test\_pred)

# Append the results to the lists

train\_r2\_list.append(r2\_train)

train\_rmse\_list.append(rmse\_train)

train\_mape\_list.append(mape\_train)

test\_r2\_list.append(r2\_test)

test\_rmse\_list.append(rmse\_test)

test\_mape\_list.append(mape\_test)

# Calculate overall average scores for training and testing

overall\_train\_r2 = np.mean(train\_r2\_list)

overall\_test\_r2 = np.mean(test\_r2\_list)

overall\_train\_rmse = np.mean(train\_rmse\_list)

overall\_test\_rmse = np.mean(test\_rmse\_list)

overall\_train\_mape = np.mean(train\_mape\_list)

overall\_test\_mape = np.mean(test\_mape\_list)

# Print overall scores

print("\nOverall Training Performance (XGBoost with 38 Boosting Rounds):")

print(f"Average R²: {overall\_train\_r2}")

print(f"Average RMSE: {overall\_train\_rmse}")

print(f"Average MAPE: {overall\_train\_mape}")

print("\nOverall Testing Performance (XGBoost with 38 Boosting Rounds):")

print(f"Average R²: {overall\_test\_r2}")

print(f"Average RMSE: {overall\_test\_rmse}")

print(f"Average MAPE: {overall\_test\_mape}")

## Predictive Model Code for SARIMA

from statsmodels.tsa.statespace.sarimax import SARIMAX

from sklearn.metrics import mean\_squared\_error, mean\_absolute\_percentage\_error

import numpy as np

import pandas as pd

# Load the cleaned dataset

df\_cleaned = pd.read\_csv('cleaned\_data.csv')

# Ensure that the data is sorted by time (assuming 'year' column)

df\_cleaned = df\_cleaned.sort\_values(by='year')

# List of columns to predict (excluding the 'year' column)

columns\_to\_predict = ['coal\_co2', 'cumulative\_coal\_co2', 'flaring\_co2', 'cumulative\_flaring\_co2',

'gas\_co2', 'cumulative\_gas\_co2', 'oil\_co2', 'cumulative\_oil\_co2',

'co2', 'co2\_per\_capita', 'co2\_per\_gdp', 'coal\_co2\_per\_capita',

'flaring\_co2\_per\_capita', 'gas\_co2\_per\_capita', 'oil\_co2\_per\_capita',

'temperature\_change\_from\_co2']

# Initialize lists to store RMSE and MAPE for each column

sarima\_rmse\_train\_list = []

sarima\_mape\_train\_list = []

sarima\_rmse\_test\_list = []

sarima\_mape\_test\_list = []

# Loop through each column to predict using SARIMA

for target\_col in columns\_to\_predict:

print(f"Training SARIMA(1,2,1)(1,1,1,12) model for {target\_col}...")

# Select the target column as a time series and check if it's formatted correctly

target\_series = df\_cleaned[['year', target\_col]].set\_index('year')

# Check for missing values and handle them (e.g., forward fill or interpolation)

target\_series = target\_series.fillna(method='ffill')

# Split the data into training and testing sets (80% train, 20% test)

train\_size = int(len(target\_series) \* 0.8)

train, test = target\_series[:train\_size], target\_series[train\_size:]

try:

# Fit SARIMA model with specified order (1,2,1)(1,1,1,12)

sarima\_model = SARIMAX(train, order=(1, 2, 1), seasonal\_order=(1, 1, 1, 12))

sarima\_fit = sarima\_model.fit(disp=False)

# Make predictions on the training set

train\_predictions\_sarima = sarima\_fit.fittedvalues

# Make predictions on the test set

test\_predictions\_sarima = sarima\_fit.forecast(steps=len(test))

# Evaluate model performance on the training set

train\_rmse\_sarima = np.sqrt(mean\_squared\_error(train, train\_predictions\_sarima))

train\_mape\_sarima = mean\_absolute\_percentage\_error(train, train\_predictions\_sarima)

# Evaluate model performance on the testing set

test\_rmse\_sarima = np.sqrt(mean\_squared\_error(test, test\_predictions\_sarima))

test\_mape\_sarima = mean\_absolute\_percentage\_error(test, test\_predictions\_sarima)

# Append the results to the lists

sarima\_rmse\_train\_list.append(train\_rmse\_sarima)

sarima\_mape\_train\_list.append(train\_mape\_sarima)

sarima\_rmse\_test\_list.append(test\_rmse\_sarima)

sarima\_mape\_test\_list.append(test\_mape\_sarima)

except Exception as e:

print(f"Error with {target\_col}: {e}")

# Skip this column and proceed with the next

# Calculate overall average scores for training and testing

average\_train\_rmse\_sarima = np.mean(sarima\_rmse\_train\_list)

average\_train\_mape\_sarima = np.mean(sarima\_mape\_train\_list)

average\_test\_rmse\_sarima = np.mean(sarima\_rmse\_test\_list)

average\_test\_mape\_sarima = np.mean(sarima\_mape\_test\_list)

# Output performance metrics

print(f"\nOverall SARIMA(1,2,1)(1,1,1,12) Performance on Training Set")

print(f"Average RMSE (Training): {average\_train\_rmse\_sarima}")

print(f"Average MAPE (Training): {average\_train\_mape\_sarima}")

print(f"\nOverall SARIMA(1,2,1)(1,1,1,12) Performance on Testing Set")

print(f"Average RMSE (Testing): {average\_test\_rmse\_sarima}")

print(f"Average MAPE (Testing): {average\_test\_mape\_sarima}")\_mape\_arima}")

## Replicated Model on INDIA dataset.

### ETL CODE

# Selecting the Features from the dataset

energy\_related\_columns = ['year',

'co2\_per\_unit\_energy',

'co2\_including\_luc\_per\_unit\_energy',

'coal\_co2',

'cumulative\_coal\_co2',

'flaring\_co2',

'cumulative\_flaring\_co2',

'gas\_co2',

'cumulative\_gas\_co2',

'oil\_co2',

'cumulative\_oil\_co2',

'co2',

'co2\_per\_capita',

'co2\_per\_gdp',

'coal\_co2\_per\_capita',

'flaring\_co2\_per\_capita',

'gas\_co2\_per\_capita',

'oil\_co2\_per\_capita'

]

# Create a new dataframe with the selected columns

energy\_data = data[energy\_related\_columns]

# Display the first few rows of the dataframe

print(energy\_data.head())

from scipy.stats import boxcox

high\_skew\_columns = [

'co2\_including\_luc\_per\_unit\_energy',

'coal\_co2\_per\_capita',

'flaring\_co2\_per\_capita',

'gas\_co2\_per\_capita',

'oil\_co2\_per\_capita'

]

# Step 2: Apply Box-Cox transformation to reduce skewness

new\_skewness = data.copy()

for column in high\_skew\_columns:

# Shift data to be positive if necessary

if (data[column] <= 0).any():

shifted\_data = data[column] - data[column].min() + 1

else:

shifted\_data = data[column]

# Apply Box-Cox transformation

new\_skewness[column], \_ = boxcox(shifted\_data)

# Step 3: Check skewness after transformation

transformed\_skewness = new\_skewness.skew()

print(transformed\_skewness)

still\_high\_skew\_columns = [

'coal\_co2',

'cumulative\_coal\_co2',

'flaring\_co2',

'cumulative\_flaring\_co2',

'gas\_co2',

'cumulative\_gas\_co2',

'oil\_co2',

'cumulative\_oil\_co2',

'co2',

'co2\_per\_capita'

]

# Step 2: Apply further transformations (square root or cube root) to reduce skewness

new\_skewness\_additional = new\_skewness.copy()

for column in still\_high\_skew\_columns:

# Check if column values are non-positive and apply transformation accordingly

if (new\_skewness[column] <= 0).any():

# Apply cube root transformation

new\_skewness\_additional[column] = np.cbrt(new\_skewness[column] - new\_skewness[column].min() + 1)

else:

# Apply square root transformation

new\_skewness\_additional[column] = np.sqrt(new\_skewness[column])

# Step 3: Check skewness after additional transformations

additional\_transformed\_skewness = new\_skewness\_additional.skew()

print(additional\_transformed\_skewness)

## EDA (Exploratory Data Analysis)

custom\_ticks = [1900, 1950, 2000, 2022]

plt.figure(figsize=(12, 6))

plt.plot(energy\_data['year'], energy\_data['co2'], label='Total CO2')

plt.plot(energy\_data['year'], energy\_data['coal\_co2'], label='Coal CO2')

plt.plot(energy\_data['year'], energy\_data['gas\_co2'], label='Gas CO2')

plt.plot(energy\_data['year'], energy\_data['oil\_co2'], label='Oil CO2')

plt.xlabel('Year')

plt.ylabel('CO2 Emissions (Million Tonnes)')

plt.title('CO2 Emissions Over Time')

plt.legend()

plt.xticks(custom\_ticks)

plt.show()

plt.figure(figsize=(12, 6))

plt.plot(energy\_data['year'], energy\_data['co2\_per\_capita'], label='CO2 per Capita')

plt.plot(energy\_data['year'], energy\_data['co2\_per\_gdp'], label='CO2 per GDP')

plt.xlabel('Year')

plt.ylabel('CO2 Emissions')

plt.title('CO2 Emissions Per Capita and Per GDP Over Time')

plt.legend()

plt.xticks(custom\_ticks)

plt.show()

plt.figure(figsize=(12, 6))

plt.plot(energy\_data['year'], energy\_data['coal\_co2'], label='Coal CO2 Emissions')

plt.plot(energy\_data['year'], energy\_data['gas\_co2'], label='Gas CO2 Emissions')

plt.plot(energy\_data['year'], energy\_data['oil\_co2'], label='Oil CO2 Emissions')

plt.xlabel('Year')

plt.ylabel('CO2 Emissions')

plt.title('CO2 Emissions by Source Over Time')

plt.legend()

plt.grid(True)

plt.show()

plt.figure(figsize=(12, 6))

plt.plot(energy\_data['year'], energy\_data['cumulative\_coal\_co2'], label='Cumulative Coal CO2')

plt.plot(energy\_data['year'], energy\_data['cumulative\_gas\_co2'], label='Cumulative Gas CO2')

plt.plot(energy\_data['year'], energy\_data['cumulative\_oil\_co2'], label='Cumulative Oil CO2')

plt.xlabel('Year')

plt.ylabel('Cumulative CO2 Emissions (Million Tonnes)')

plt.title('Cumulative CO2 Emissions Over Time')

plt.legend()

plt.xticks(custom\_ticks)

plt.show()

plt.figure(figsize=(12, 6))

plt.plot(energy\_data['year'], energy\_data['co2\_per\_gdp'], label='CO2 Emissions Per GDP')

plt.xlabel('Year')

plt.ylabel('CO2 Emissions Per GDP')

plt.title('CO2 Emissions Per GDP Over Time')

plt.legend()

plt.grid(True)

plt.xticks(custom\_ticks)

plt.show()

plt.figure(figsize=(10, 7))

cumulative\_emissions = [

energy\_data['cumulative\_coal\_co2'].iloc[-1],

energy\_data['cumulative\_gas\_co2'].iloc[-1],

energy\_data['cumulative\_oil\_co2'].iloc[-1]

]

labels = ['Cumulative Coal CO2', 'Cumulative Gas CO2', 'Cumulative Oil CO2']

plt.pie(cumulative\_emissions, labels=labels, autopct='%1.1f%%', startangle=140, colors=['blue', 'orange', 'green'])

plt.title('Distribution of Cumulative CO2 Emissions by Source')

plt.show()

# Calculate moving average

data['co2\_moving\_avg'] = data['co2'].rolling(window=5).mean()

# Plot moving average

plt.figure(figsize=(15, 7))

plt.plot(data['year'], data['co2'], marker='o', linestyle='-', label='CO2 Emissions')

plt.plot(data['year'], data['co2\_moving\_avg'], color='red', label='5-Year Moving Average')

plt.title('CO2 Emissions and 5-Year Moving Average')

plt.xlabel('Year')

plt.ylabel('CO2 Emissions')

plt.legend()

plt.grid(True)

plt.show()

## Predictive model Code for INDIA dataset

import pandas as pd

from sklearn.ensemble import RandomForestRegressor

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import r2\_score, mean\_squared\_error, mean\_absolute\_percentage\_error

import numpy as np

# Load the cleaned dataset

df\_india\_cleaned = pd.read\_csv('cleaned\_ind.csv')

# List of columns to predict (excluding 'year' and non-target columns)

columns\_to\_predict = [

'co2\_per\_unit\_energy',

'co2\_including\_luc\_per\_unit\_energy',

'coal\_co2',

'cumulative\_coal\_co2',

'flaring\_co2',

'cumulative\_flaring\_co2',

'gas\_co2',

'cumulative\_gas\_co2',

'oil\_co2',

'cumulative\_oil\_co2',

'co2',

'co2\_per\_capita',

'co2\_per\_gdp',

'coal\_co2\_per\_capita',

'flaring\_co2\_per\_capita',

'gas\_co2\_per\_capita',

'oil\_co2\_per\_capita'

]

# Checking if all columns exist in the dataset

missing\_columns = [col for col in columns\_to\_predict if col not in df\_india\_cleaned.columns]

if missing\_columns:

print(f"The following columns are missing from the dataset: {missing\_columns}")

# Split data into features (X) and target (y) with 'co2' as the target variable

X = df\_india\_cleaned.drop(columns=['co2'])

y = df\_india\_cleaned['co2']

# Split data into training and testing sets (80% training, 20% testing)

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Initialize the RandomForestRegressor model

rf\_model = RandomForestRegressor(n\_estimators=100, random\_state=42)

# Train the model

rf\_model.fit(X\_train, y\_train)

# Make predictions on the training and testing sets

y\_train\_pred = rf\_model.predict(X\_train)

y\_test\_pred = rf\_model.predict(X\_test)

# Calculate performance metrics for both training and testing sets

train\_r2 = r2\_score(y\_train, y\_train\_pred)

test\_r2 = r2\_score(y\_test, y\_test\_pred)

train\_rmse = np.sqrt(mean\_squared\_error(y\_train, y\_train\_pred))

test\_rmse = np.sqrt(mean\_squared\_error(y\_test, y\_test\_pred))

train\_mape = mean\_absolute\_percentage\_error(y\_train, y\_train\_pred)

test\_mape = mean\_absolute\_percentage\_error(y\_test, y\_test\_pred)

# Display results

print(f"Training R²: {train\_r2}")

print(f"Testing R²: {test\_r2}")

print(f"Training RMSE: {train\_rmse}")

print(f"Testing RMSE: {test\_rmse}")

print(f"Training MAPE: {train\_mape}")

print(f"Testing MAPE: {test\_mape}")

## Video Link: - https://youtu.be/2ZEtjcVPGJ8