**Predicting Greenhouse Gas Emissions Characteristics of the United Kingdom using Deep Learning Algorithm**

By

Chinedu Ukeje

Submitted to

**The University of Roehampton**

In partial fulfilment of the requirements

for the degree of

**MASTER OF SCIENCE IN COMPUTING**

Abstract

## Over the years, Mankind has struggled to expand its reach by creating new and innovative ways to meet its needs. This started from the time of the agricultural revolution to the time of the industrial revolution. The industrial revolution affected our lives because more people from rural areas poured into the city. The migration of people meant an increase in population in these areas. According to the United Nations Populations Fund, the world's population, as of November 2, 2022, has surpassed 8 billion. This increase in population comes with different problems. Environmental issues are one of the major problems affecting modern civilization. Chief among these ecological problems are air pollution and climate change. Greenhouse gases are the major contributors to climate change, and researchers believe CO2 emissions are the primary culprit (Zhong and Haigh, 2013). To ensure that the goal of achieving net-zero emissions by 2050 is adhered to, there is a need to analyse and monitor every emissions data from the past to date and to create a predictive model capable of forecasting future emissions given a set of parameters. Making this information readily available to stakeholders and researchers will ensure they have enough information about the past, present, and future to keep track of this goal.

Signed (apply signature below)

**Declaration**

I hereby certify that this report constitutes my own work, that where the language of others is used, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions, or writings of others.

I declare that this report describes the original work that has not been previously presented for the award of any other degree of any other institution.

**Date:**

**Chinedu Ukeje**

Acknowledgements

I would love to thank my guiding angel, my suprvisor!

Table of Contents

[1. Introduction 7](#_Toc927214740)

[Research Question or Problem statement 8](#_Toc522749298)

[Aims 8](#_Toc1298644285)

[Objectives 8](#_Toc198637622)

[Legal, Social, Ethical and Professional Considerations 8](#_Toc960797214)

[Background 8](#_Toc273743648)

[Report overview 8](#_Toc210207254)

[2. Literature or Technology Review 8](#_Toc1335231035)

[3. Methodology 9](#_Toc1844834773)

[4. Implementation 10](#_Toc1739317238)

[5. Results 11](#_Toc826321870)

[Evaluation 11](#_Toc1085198293)

[Related Work 12](#_Toc55229855)

[6. Conclusion 12](#_Toc253686570)

[Reflection 13](#_Toc1734103406)

[Future Work 13](#_Toc542248490)

[7. References 13](#_Toc1743583894)

[8. Appendices 14](#_Toc649705763)

# **Introduction**

Over the years, Mankind has struggled to expand its reach by creating new and innovative ways to meet its needs. This started from the time of the agricultural revolution to the time of the industrial revolution. The industrial revolution affected our lives because more people from rural areas poured into the city. The migration of people meant an increase in population in these areas. According to the United Nations Populations Fund, the world's population, as of November 2, 2022, has surpassed 8 billion [3] This increase in population comes with different problems. Environmental issues are one of the major problems affecting modern civilization. Chief among these ecological problems are air pollution and climate change. Greenhouse gases are the major contributors to climate change, and researchers believe CO2 emissions are the primary culprit [19]. Some researchers believe that CO2 is not the primary cause of climate change [15], but most think otherwise.

## **1.1.1 Global Warming and the UK**

Global warming, the direct effect of climate change, is now felt globally, making it a hot-button issue among world powers and researchers. The United Kingdom is currently the 6th largest economy in the world. It is believed that the UK could be the fastest-growing economy in the world come 2050, with an average annual growth of 1.9% [18]. This economic growth has attracted highly skilled, talented workers from across the globe and the development of new trade links with other countries, resulting in a population spike in the UK. The mid-year population of the UK for the year 2021 was estimated at 67.0 million, which is an increase of 3.7 million compared with the mid-year population of 2011. In the past decade, the people of England were estimated at 56 million, which is a 6.5% increase and the highest when compared with the other four countries that comprise the UK, with Northern Ireland increasing by 5.0%, Wales at 1.4% and Scotland at 3.5% [14].

## **1.1.2 The UK and Net-Zero**

Since September 2021, the United Kingdom has been actively working to reduce greenhouse gas emissions. It has set legally binding targets to cut emissions under the Climate Change Act 2008. The UK has set out a series of emissions reduction targets to achieve this target. By 2020, the goal was to reduce emissions by at least 34% compared to 1990 levels. By 2030, the target is to reduce emissions by at least 68%, aiming to achieve net-zero greenhouse gas emissions by 2050 [10]. In line with achieving net-zero greenhouse gas emissions by 2050, the UK has significantly reduced greenhouse gas emissions. It has transitioned from coal-fired power generation to cleaner alternatives such as natural gas and renewable energy sources. Policies such as the Carbon Price Floor and Renewable Obligation have also helped incentivize low-carbon technologies [9].

Furthermore, the UK has seen a notable decline in emissions from electricity generation due to a shift away from coal-fired power plants. The country has been increasing its reliance on renewable energy sources, including wind, solar, and offshore wind. In recent years, renewable energy has significantly contributed to the electricity mix. Another sector of the economy of the United Kingdom that contributes to its carbon footprint is its transport sector. The transportation sector is a significant source of greenhouse gas emissions in the UK. Efforts have been made to promote electric vehicles (EVs) by providing grants and incentives to EV buyers. The UK government has also banned the sale of new petrol and diesel cars and vans by 2030, promoting a transition to electric and other zero-emission vehicles. The industrial sector, including manufacturing and heavy industries, contributes to greenhouse gas emissions. The UK has implemented various measures to improve energy efficiency, encourage using low-carbon technologies, and promote sustainable industry practices [2].

Given the above depositions, it is imperative to keep track of greenhouse gas emissions of the United Kingdom going forward. The population of the UK will keep growing in proportion to its continued increasing economy. Greenhouse gas emissions pose a significant problem globally, and the UK is not exempt from this challenge. The UK is among the world's largest emitters of greenhouse gases [4]. The UK government has set ambitious targets to address greenhouse gas emissions which involve green investment, green jobs, skills, and industries, embedding net-zero in government, local climate action, empowering the public and businesses to make green choices, and international leadership and collaboration. The country aims to achieve net-zero emissions by 2050, meaning the total emissions are balanced by removing an equivalent amount of greenhouse gases from the atmosphere [10].

## **1.2 Research Question or Problem statement**

This project intends to answer the following research questions:

1. What are the percentage contributions of CO2 emissions by sectors of each of the countries that make up the UK for the past ten years?
2. What sector of the UK economy contributes the most to temperature changes due to high CO2 emissions?
3. What regions of the UK have suffered the most in terms of temperature change in the past decade?
4. Is there any discernable trend in CO2 emissions of the UK from year to year (taking into consideration the effect Covid-19 had during the periods of 2019-2021)?
5. Is there any significant seasonal difference in CO2 emissions of the UK from year to year (taking into consideration the effect Covid-19 had during the periods of 2019-2021)?
6. What future temperature changes can be forecasted based on past and current emissions data?

## 1.2 Aims

## To develop an effective deep learning model for predicting the United Kingdom's greenhouse gas (GHG) emission characteristics.

## To explore the United Kingdom's mitigation plan for GHG emission characteristics and implement mitigation techniques to help achieve the "Net-zero" initiative.

## To evaluate the performance and effectiveness of the developed GHG emission characteristic predictive and mitigation model.

## 1.3 Objectives

## Collect and curate a dataset of greenhouse gas emissions of the United Kingdom from the Office for National Statistics and other data sources for training and evaluation purposes.

## Investigate and analyse the different contributions of energy supply, residential, transportation, and other sectors of the United Kingdom to greenhouse gas emissions.

## Explore and implement techniques, including forensic analysis, data augmentation, and adversarial training, to enhance the robustness of the greenhouse gas emission characteristics model.

1. Implementation of statistical differencing with stationarity: constant mean, variance and covariance.
2. Develop an evaluation framework that includes appropriate metrics, such as accuracy, precision, recall, and F1-score, RMSE, and R-squared, to assess the performance of the greenhouse gas emissions characteristics model.

## 1.4 Legal, Social, Ethical and Professional Considerations

## Legal, social, ethical, and professional considerations must be considered when conducting academic research. The forensic analysis of greenhouse gas emissions characteristics of the United Kingdom is environmental sciences-related research requiring sensitive datasets that must be processed appropriately and reported considering the effects of legal, social, ethical, and professional ramifications of the problem domain.

## **1.4.1 Legal Considerations**

## When undertaking an academic project, it's essential to be mindful of legal considerations to ensure compliance and ethical conduct. The forensic analysis of greenhouse gas emissions characteristics of the United Kingdom project involves collecting, storing, and analyzing data that has to do with the United Kingdom; hence I made sure that I comply with data protection regulations such as the General Data Protection Regulation (GDPR). I have also obtained informed consent from the Office of National Statistics to use the data. The rest of the data used are open source and freely available on Kaggle and Statista (a subscription account was needed for Statista).

## **1.4.2 Social Considerations**

## In addition to legal considerations, academic projects should also consider various social concerns. These considerations focus on the project's impact on individuals, communities, and society. Greenhouse gas emissions can have significant health effects, notably air pollution from fossil fuel combustion, which is mainly the domain of the transport sector. This project will contribute to improving air quality and reducing the health risks associated with pollution through the identification of the sectors of the UK economy with the highest contributions. The result of this project will further reinforce the potential health benefits of transitioning to cleaner energy sources and sustainable practices and further provide analysis results to engage with a diverse range of stakeholders, including community members, policymakers, industry representatives, and advocacy groups.

## **1.4.3 Ethical Considerations**

## Undertaking this type of project involves several ethical considerations beyond legal and social aspects. Ethical considerations include considering confidentiality, privacy, and other ethical issues. Necessary steps have been taken to ensure scientific integrity by conducting rigorous and unbiased research by ensuring that my methodology, data collection process, and analysis are transparent, reliable, and adhere to established scientific standards. Every research material considered is appropriately cited, and permission to use the datasets is obtained through the right channel.

## This project recognizes the impact that greenhouse gas emissions and climate change will have on future generations; hence there is a responsibility to contribute to the body of knowledge aimed at reducing greenhouse gas emissions and mitigating climate change to ensure a sustainable and habitable planet for future generations.

## 1.5 Background

## Greenhouse gas emissions and the United Kingdom have been a subject of significant attention and action in recent years. The United Kingdom has recognized the need to reduce its greenhouse gas emissions to combat climate change and has set various targets and implemented policies to achieve this goal [2]. The United Kingdom was one of the first countries to legislate long-term greenhouse gas reduction targets. The Climate Change Act 2008 established a legally binding target to reduce greenhouse gas emissions by at least 80% below 1990 levels by 2050 [1]. In 2019, the UK became the first major economy to pass a law to bring all greenhouse gas emissions to net-zero by 2050. The United Kingdom has made significant progress in reducing its greenhouse gas emissions. Between 1990 and 2019, the UK's emissions decreased by approximately 43%, while the economy grew by around 75% [10]. This was achieved through a combination of factors, including the decarbonization of electricity generation, increased energy efficiency, and changes in industrial practices. The UK has made substantial strides in decarbonizing its electricity generation. Coal-fired power plants, which used to be a significant source of emissions, have been phased out, and the country has increased its reliance on renewable energy sources such as wind and solar power. In 2020, renewable energy surpassed fossil fuels as the largest source of electricity generation in the UK [9]. The transportation sector is a significant source of greenhouse gas emissions. The UK government has taken steps to promote the adoption of electric vehicles (EVs) and reduce emissions from transportation. This includes financial incentives for EV purchases, investment in charging infrastructure, and plans to ban the sale of new petrol and diesel cars by 2030.

## To ensure that the goal of achieving net-zero emissions by 2050 is adhered to, there is a need to analyse and monitor every emissions data from the past to date and to create a predictive model capable of forecasting future emissions given a set of parameters. Making this information readily available to stakeholders and researchers will ensure they have enough information about the past, present, and future to keep track of this goal.

## 1.6 Report overview

**Chapter 1**- Introduction; In this chapter, the background of the project is discussed in detail, the aims and objectives of the project are stated, and finally, legal, social, ethical, and professional considerations are also highlighted.

**Chapter 2**- Literature review; background research, theoretical approach. This chapter considers different literature by different authors related to the problem domain. Theoretical approach and background research are considered in this section of the project. The review of different technologies is also discussed.

**Chapter 3**- Design and methodology; technological review, methodological review. This chapter discusses the methodology taken to carry out the project and the technological stack that was adopted for the project.

**Chapter 4**- Implementation and result; the answers to the generated research questions in chapter one are evaluated here with discussions, and every step of the methodology to achieve the solution is also discussed.

**Chapter 5**- Conclusion; this part of the project provides a summary of the entire research. It discusses the answers to the research questions and provides a reflection of the result with recommendations and future directions.

In conclusion, a project plan for the project has been developed on Teamwork with a breakdown of the methodology and timeline for each stage of the methodology. The link to this plan is provided in Appendix 3.

# **Literature - Technology Review**

## 2.1 Literature Review

Greenhouse gas emissions refer to releasing gases into the atmosphere, contributing to the greenhouse effect and global warming [8]. These gases trap heat from the sun within the Earth's atmosphere, leading to an increase in global temperatures and associated climate change impacts. The primary greenhouse gases of concern include carbon dioxide (CO2), methane (CH4), nitrous oxide (N2O), and fluorinated gases [1]. Greenhouse gas emissions are released from different sources, which include the following:

* Fossil fuel combustion: Burning of coal, oil, and natural gas for electricity generation, transportation, and industrial processes.
* Deforestation and land-use changes: Clearing of forests, conversion of land for agriculture, and alteration of natural ecosystems.
* Agriculture: Livestock farming, rice cultivation, and the use of synthetic fertilizers that release methane and nitrous oxide.
* Industrial processes: Manufacturing, chemical production, and waste management activities that release various greenhouse gases.

Greenhouse gas emissions contribute to the warming of the Earth's atmosphere and subsequent climate change [15].

A diagram of a car with text

Description automatically generated

**Figure 2.1 greenhouse gas effect.**

The increased concentration of these gases leads to changes in weather patterns, rising global temperatures, melting ice caps, sea-level rise, and extreme weather events such as hurricanes, droughts, and heat waves. Greenhouse gas emissions have steadily increased over the past century, primarily driven by industrialization and population growth [5]. Burning fossil fuels for energy production remains the largest source of CO2 emissions, while methane emissions come from various sources, including livestock and oil and gas operations.

Carrying out research of this magnitude requires the analysis of different large datasets to get insight into past and present greenhouse gas emissions and then go a step further to create a forecasting model that can take as input parameters that are selected from a list of features to forecast future emissions based on past and current data.

Different researchers have done some great work in this problem domain with some level of success. Hamrani et al. 2020, studied the potential of using a deep learning model, classical regression model, and shallow learning model to predict soil greenhouse gas emissions from an agricultural field [12]. To perform the study, they measured CO2 and N2O fluxes with other agronomical and environmental soil data at the site over five years in Quebec, Canada. They reported that Long Short-Term Memory (LSTM) model performed the best, having the highest R coefficient and the lowest root mean square error when compared with other models. According to [14], China, USA, and India have the highest known energy consumption and the highest greenhouse gas emissions. Following the above, they used a machine learning-based time-series model to predict and forecast the CO2 emissions for India for the next ten years based on univariate time-series data from 1980 to 2019. They used five different models. Three statistical models; autoregressive-integrated moving average (ARIMA) model, seasonal autoregressive-integrated moving average with exogenous factors (SARIMAX) model, and the Holt-Winters model, two machine learning model; random forest and linear regression and one deep learning model; LSTM. Once again, they reported that LSTM performed better than the other models considered for the research.

[16] Used SARIMAX to forecast global CO2 emissions for the near future (2022 to 2027), future (2022 to 2054), and far future (2022 to 2072). They took into consideration the effects of Covid-19 while developing the model. They had four different datasets; pre-Covid, start-Covid, trans-Covid, and post-Covid. They reported that the model performed better with the post-Covid dataset; no deep learning model was used for this study. [11] Developed a ‘hybrid model that combines machine learning algorithms and optimization model to forecast greenhouse gas emissions using energy market data’. This study focused on the Iranian energy market data from 1990 to 2018. They combined the power of machine learning models and the accuracy of mathematical programming to improve the models. Nine models were considered: ANN, AR, ARIMA, SARIMA, SARIMAX, RF, SVR, KNN, and LSTM. Their results were compared before and after optimization using the mathematical optimization model. They reported a 31.7% increase when using the PSO algorithm to optimize and a 12.8% increase when using the GWO algorithm.

[7] Forecasted greenhouse gas emissions in Turkey based on electricity production. The forecast was done using deep learning (DL) algorithm, support vector machine (SVM), and artificial neural network (ANN). The study dataset covered the year 1990 to 2018, and they forecasted the period from 2015 to 2018. They stated that the deep learning algorithm performed better than other models that were used to make the forecast. [5] worked on ‘Prediction of greenhouse gas emissions reductions via machine learning algorithms: Toward an artificial intelligence-based life cycle assessment for automotive lightweighting’. Their study was to show if the efforts to replace glass fiber composites with greener light-weighted natural fibers are better and friendlier to the environment. They used a machine learning algorithm to compare both emissions to determine which is better.

[13] developed an ‘Ensemble machine learning for modeling greenhouse gas emissions at different time scales from irrigated paddy fields’. They used a dataset generated over three years of a paddy field in Kunshan, China, and augmented it with the dataset of WSI and FI. Random forest, K-nearest neighbor regression, and gradient boosting regression models were stacked to form an ensemble machine learning model. They reported that the stacked model improved accuracy by 13.36% when compared to other models. [6] Focused on the transportation sector. They worked on ‘Greenhouse gas emission prediction on a road network using deep sequence learning’. A deep learning model was developed to predict link-level greenhouse gas emission rates. They reported that LSTM performed better than all other models that were considered for the study.

From the discussions above, there are three main algorithms for forecasting greenhouse gas emissions. Out of these algorithms, it is obvious that deep learning algorithms perform better than other models.

## 2.2 Technology Review

Data analysis and visualization are crucial components of understanding and interpreting complex datasets. They help in extracting meaningful insights, identifying patterns, and communicating findings effectively. Performing data analytics and data visualization requires the use of certain technologies. Currently, there are several technologies used for the different stages of data analysis and data visualization. These technologies are discussed in stages below:

**2.2.1 Problem definition and objectives specification:** in this stage of my project, it was important to properly define the problem to be solved and clearly define the steps to solve them. To perform this task, there was a need to use a project management tool. There are several options to use, like Kissflow, Trello, Asana, Kanban, Teamwork, etc., all of which have their advantages and disadvantages. I opted for Teamwork because it is easy to use and utilizes Kanban-based project management tools. It also has advanced reporting capabilities compared with others like Trello.

**2.2.2 Data collection and preparation:** To perform any form of data analytics, there is a need to collect relevant data to address the problem. The data was collected in Microsoft Excel, which is a spreadsheet program. The spreadsheet program Microsoft Excel was created by Microsoft and is available for Windows, macOS, Android, iOS, and iPadOS. Additionally, it has pivot tables, graphing tools, calculating or calculation capabilities, and the Visual Basic for Applications macro programming language.

It is imperative to ensure the data is accurate, complete, and in a suitable format for analysis. Python is the programming language that is used for this project. The programming language Python is high-level and versatile. The off-side rule is used extensively in its design philosophy, which places a strong emphasis on code readability. Both Python's types and trash collection are dynamic. It is an open-source programming language that has over 500 data analytics libraries. To manipulate and analyze data, the Python programming language has a software package called pandas. It includes specific data structures and procedures for working with time series and mathematical tables. It is free software distributed under the BSD license. It was used in this project to load the datasets and clean the data by handling missing values, outliers, and inconsistencies.

**2.2.3 Exploratory Data Analysis (EDA):** In order to understand the structure and characteristics of the data used for the project, there is a need to perform EDA. This involves examining summary statistics, distributions, and correlations and identifying any initial patterns or insights. As stated above, python is the programming language used for this research work; hence Numpy and Pandas were used to perform this task. Numpy is a library in Python that is based on linear algebra. Pandas is built on top of Numpy, making it easy to perform statistical analysis, data summarization, correlation, and get insights about the dataset.

**2.2.4 Data analysis technique:** To derive insight from the data, appropriate data analysis technique. This can involve methods like regression analysis, hypothesis testing, clustering, classification, or machine learning algorithms, depending on the nature of your problem. In the prediction of greenhouse gas emissions characteristics, a deep learning algorithm was used. Deep learning algorithms has been proven very successful in this regard.

**2.2.5 Data visualization:** In order to communicate insights effectively, it is important to create visual representation of the data in the form of plots, charts, graphs, maps, or dashboards. In the process of selecting the right visualization tool, it is important to keeping in mind the intended audience. Seaborn and Matplotlib was used in python to perform this task while Tableau is used as a visualization tool to create a dashboard.

In conclusion, tools such as Python (with libraries like Pandas, NumPy, and Matplotlib/Seaborn), R (with packages like dplyr, ggplot2), and Tableau are commonly used for data analysis and visualization. They provide a wide range of functions and capabilities to explore, analyze, and present data effectively.

# **Methodology**

This study creates a method of putting the efforts of the government and other stakeholders in context with respect to the net-zero 2050 goal. Most analysis in this problem domain produces chart that are difficult for the average man on the street to understand. This project, on the other hand, performs a region and sectors-wise analysis of the United Kingdom and projects the contributions of each of this region and sector to greenhouse gas emission and how these impacts on climate change. This research analyses greenhouse gas emission data of the United Kingdom using data analysis and visualization tools with the goal of creating a predictive model that is capable of forecasting future emissions based on past and present data. The difference of this research with other works is the creation of a forecasting tool that is available and can be understood by all stakeholders. Figure 3.1 presents the methodology for executing this project.

Dataset

Data Pre-processing

Clean Dataset

Feature Selection

Exploratory Data Analysis

Split Dataset

Train Set

Test Set

Initial Model

Training Evaluation

Final Model

Final Model Evaluation

Data Story and Model Evaluation Result

**Figure 3.1 Methodology**

**3.1 Stage 1: Data Gathering**

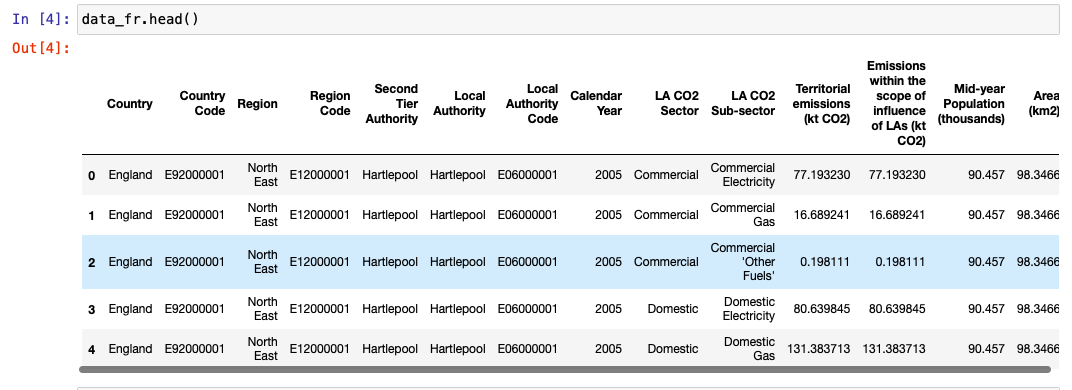
Gathering data for a data analytics project is a critical step in the process. The success of any analytical project will largely depend on the quality, relevance, and comprehensiveness of the data collected. The data used for this project was gotten from the office for National Statistics, United Kingdom. More data was gotten from Statista and Gov.uk to augment the initial data.

A screenshot of a computer

Description automatically generated

**Figure 3.2 greenhouse gas dataset.**

The data collected from this website (figure 3.2) is a time series data that is in a comma separated value (CSV) file. A time series dataset is a set of observations that is recorded in a chronological order. The greenhouse gas emissions data is recorded yearly. Sample of the data is shown in figure 3.3.



**Figure 3.3 greenhouse gas dataset**

The dataset contains fourteen parameters, four of which are numeric while the rest are string data. Figure 3.4 gives information about the dataset. the dataset is 129,388 in total and contains some missing values.

A screenshot of a computer

Description automatically generated

**Figure 3.4 Dataset information**

**3.2 Stage 2: Data Cleaning and Feature Selection**

Before using this data for analysis, the data needs to be cleaned. To perform this task, some libraries need to be imported from python library. Libraries like Pandas, Numpy and Matplotlib. There is a need to remove unwanted commas and spaces from the dataset. This was done using a function called “clean data”.

A screenshot of a computer code

Description automatically generatedA screenshot of a computer

Description automatically generated

**Figure 3.5 (a) Missing value count Figure 3.5 (b) Clean data function**

The data was imported using Pandas which runs on the Numpy framework. The imported data is then cleaned using the written function. A screenshot of this process is shown in figure 3.5 (b). Figure 3.5 (a) shows the information of the dataset after the initial cleaning process. It can be seen from figure 3.5 (a) that even after the initial cleaning task, the dataset still contained several missing values. The next step is to then “clean” the missing values. Since the missing values are small (153), they are filled out using the mean value of the parameter. After the process of removing the missing values (The output from the cleaning process is shown in figure 3.6.), The next step is to drop the columns that are not needed from the dataset.

A screenshot of a computer code

Description automatically generated

**Figure 3.6 dropped columns from the dataset.**

After removing missing values and dropping unnecessary columns, there is a need to remove trend, seasonality and check for stationarity of the dataset (this is discussed in the implementation section of this project).

When the data was imported, the Calendar Year parameter was not in the required format. It was loaded as an int64 data type. To handle a time series data, all time related parameter needs to be in a datetime format. So, the first thing after removing missing values and dropping unnecessary columns was to convert the Calendar Year parameter to a datetime format. The Calendar Year parameter was not the index of the dataset and hence it was set as the index of the dataset to ensure chronological order of the dataset. These steps are shown in figure 3.7.

A screen shot of a computer code

Description automatically generated

**Figure 3.7 Formatting the Calendar year parameter.**

The data in the imported data frame is then converted to a series data structure which makes manipulation of the new dataset very easy. This is also shown in figure 3.7.

Finally, the data is ready for Forecasting. Before developing the machine learning algorithm and using the clean dataset with the algorithm to make a forecast, there is a need to understand the dataset. Before dropping unnecessary columns that will not be used for the forecast, exploratory data analysis was performed on the data to gain insight into the dataset. this is discussed in the next section.

**3.3 Stage 3: Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is a crucial initial step in the data analysis process. It involves visually and statistically exploring the data to gain insights, discover patterns, identify outliers, and check assumptions. EDA helps data analysts and data scientists understand the structure and characteristics of the dataset before applying more advanced modelling techniques.

The first EDA that was performed was to see the distribution of CO2 emissions over time based on the dataset. The dataset contains data from 1990 to 2021. By showing the total emissions of CO2, we can determine the trend of the emissions over time. Figure 3.8 shows the total emissions data by year starting from 1990 to 2021.

A graph with a line going up

Description automatically generated

**Figure 3.8 screenshot of total greenhouse gas emissions from 1990 to 2021**

From figure 3.8, CO2 emissions was 4.73% higher in 2021 than 2020. This can be attributed to Covid-19 virus. The lockdown restrictions ensured reduction in CO2 emissions which is lower than any other year starting from 1990. Even at a 4.73% increase in 2021, it is 5.21% higher than it was in 2019. When compared with 1990, CO2 emissions has been reduced by 47% (809 Mt CO2e).

To answer the first research question, **“What are the percentage contributions of CO2 emissions by sectors of each of the countries that make up the UK for the past ten years?”**, There is a need to show from the dataset the total CO2 emissions of the four countries that makes up the United Kingdom. This is shown in Figure 3.9.

A graph of emissions by countries/regions

Description automatically generatedA graph of emission emissions

Description automatically generated with medium confidence

**Figure 3.9 CO2 emissions by countries in the UK.**

It can be seen from figure 3.9 that England produces the largest CO2 emissions, about 5.35 x 106 (kt CO2), while Norther Ireland produces the lowest CO2 emissions, about 2.46 x 105 (kt CO2). Scotland is the second largest producers of CO2 emissions with 5.90 x 105 (kt CO2) and Wales generate about 4.31 x 105 (kt CO2).

To put this analysis in perspective, figure 3.10 shows the mid-year population broken down by regions. Regions in England have the most population when compared with the other four three countries in the UK. This explains the level of CO2 emissions generated by England. Yorkshire and the Humber, London, North East, West Midlands, North West and South West all have populations greater than Scotland, Northern Ireland and Wales.

A graph of different colored bars

Description automatically generated

**Figure 3.10 Mid-year population by region.**

To answer the second research question, **“what sector of the UK economy contributes the most to temperature changes due to high CO2 emissions?”**, there is a need to show which sector or sectors generates the highest amounts of CO2 emissions. This is because, CO2 is mainly responsible for climate change. Figure 3.11 answers this question.

A graph of different colored lines

Description automatically generatedA graph of different colored bars

Description automatically generated

**Figure 3.11 sector-wise contributions to CO2 emissions in the UK**

Figure 3.11 shows the emission rates of different sectors. These sectors represent the highest consumers of energy in the United Kingdom. They account for over 80% of the total CO2 emissions of the United Kingdom. The massive reduction seen in figure 3.11 from 1990 to 2021 was as a result of the United Kingdom switching from a fossil fuel-based economy to a greener economy, hence the net-zero 2050 target.

We may examine the effects of the early coronavirus pandemic's emissions by using the numbers for 2020. From 2019 to 2020, emissions in the transport sector decreased by about 36%, the largest proportionate decrease of any industry, as opposed to a 2% decrease in 2018.

Between 2019 and 2020, home emissions decreased 12% but still contributes a lot to CO2 emission rate. This is due mainly to Winter heating and the fact that the COVID-19 pandemic in 2020 resulted in many individuals working remotely from home and a lot of individuals losing their jobs. As a result, more people stayed at home, resulting in higher energy consumption for heating but lower emissions from transportation.

Finally for EDA, to answer the third research question, **“what regions of the UK have suffered the most in terms of temperature change in the past decade?”,** We just need to show which area of the UK generates the most CO2 emissions with respect to population.

From figure 3.10, it was shown that England (all the regions in England) has the highest population compared to the other three countries with most of its population situated in London, Yorkshire and the Humber, and Northeast. Figure 3.12 shows these figures for England alone.

A graph with different colored bars

Description automatically generated

**Figure 3.12 Mid-year population of regions in England**

It could be said from figure 3.12 above, that the areas with the highest population are more prone to the effect of CO2 emissions, remember from figure 3.11, it was shown that residential areas contributed more to the generation of CO2 emissions. It is therefore logical to conclude that these areas are more affected by the effect of these emissions. To put this in perspective, figure 3.13 shows the distribution of greenhouse gases generated in the UK.

A screen shot of a graph

Description automatically generated

**Figure 3.13 greenhouse gas emissions of the UK.**

**3.4 Stage 4: Predictive Model Creation**

Creating a predictive model involves the process of building a mathematical or computational algorithm that can make predictions based on data patterns and relationships. One of the hardest tasks in predictive analysis is the determination of what mathematical model to use. In the past, models like ARIMA, SARIMA, SARIMAX and Regressive models have all been used for timeseries forecasting (These models were also considered as alternatives in implementation of this project). Overtime, deep learning has proven to be a powerful approach for forecasting tasks due to its ability to automatically learn complex patterns and representations from data. It excels at capturing non-linear relationships and handling large amounts of data, making it well-suited for time series forecasting problems. Sequence-to-sequence models use an encoder-decoder architecture to handle input sequences and generate output sequences. This approach is useful when the forecasting task requires predicting multiple future time steps (just like the problem we are trying to solve).

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) that has gained popularity for time series forecasting tasks. LSTMs are designed to handle sequential data, making them well-suited for capturing temporal dependencies and patterns in time series. One of the key advantages of LSTM over traditional RNNs is its ability to maintain long-term dependencies in the data. This means it can capture relationships between events that are separated by a considerable time lag, which is essential for many forecasting tasks. This functionality of LSTM makes it ideal for the task of forecasting greenhouse gas emissions of the United Kingdom. The structure of LSTM is shown in figure 3.14.

A diagram of a flowchart

Description automatically generated

**Figure 3.14 the structure of LSTM**

The structure of the LSTM is such that each cell is controlled by three gates that are responsible for the state of the cell. These gates are referred to as input gate, output gate and forget gate respectively. The cells in the LSTM are connected in such a way that they serve as memory modules. In figure 3.14, the cells are , , and  respectively. The formula of LSTM is given below:

(1)

(2)

(3)

(4)

Where:

W = weight matrices

b = bias

Now that the gates in the LSTM cells are mathematically defined, with as the activation function for the cells, the next step is to define how the LSTM network is to predict the time series data. Naturally, a widow size is defined that takes a given data starting from time t – w to a time t such that the LSTM predicts time t + 1, t + 2, t + 3 etc. The size of the window here determines how far into the future the model is to look when making a forecast. To achieve the above, a function is created to structure the dataset into fixed window size. The output of this function is a time series data with fixed window size in a tensor dataset because PyTorch is used to design the LSTM model. This function is shown in figure 3.15.

A screenshot of a computer code

Description automatically generated

**Figure 3.15 dataset creation function.**

The model is declared as a class after importing the PyTorch library. The model consists of an LSTM layer that is used with a fully connected layer. The output of the model is a tuple. The first element in the model is the generated hidden states with each of them representing a time step of the time series input, the input is then parsed to the second element of the model which is the cell memory and hidden states. The LSTM parameter for batch processing is set to true because the input from the formula receives a tensor (window size, time step, features). Finally, the processed time series data is parsed out through the output hidden states with the tanh activation function as a single regression output.

The implementation of this model is discussed in the implementation section of this project work.

**3.5 Project Management**

In this stage of my project, it was important to properly define the problem to be solved and clearly define the steps to solve them. To perform this task, there was a need to use a project management tool. There are several options to use, like Kissflow, Trello, Asana, Kanban, Teamwork, etc., all of which have their advantages and disadvantages. I opted for Teamwork because it is easy to use and utilizes Kanban-based project management tools. It also has advanced reporting capabilities compared with others like Trello.

|  |  |
| --- | --- |
| **Project Report Delivery Schedule**  Note: Reorder the sections in the order that you plan to complete them. | Deadline Date |
| Abstract | Week 1 |
| Declaration | Week 1 |
| Acknowledgements | Week 1 |
| Introduction | Week 1 |
| Literature - Technology Review | Week1 - 2 |
| Methodology | Week 2 - 4 |
| Implementation and Results   * Evaluation * Related Work | Week 4 - 5 |
| Conclusion   * Reflection * Future Work | Week 5 - 6 |
| References | Week 6 |
| Appendices | Week 6 |

Table 1: Project Report Delivery Schedule

|  |  |
| --- | --- |
| **Artefact Delivery Schedule**  Note: Reorder the activities in the order that you plan to complete them. | **Deadline Date** |
| Artefact Planning and Resourcing | Week 1 |
| Artefact Design | Week 1 |
| Artefact Procurement Activities (e.g., data collection, source framework etc.) | Week 2 - 4 |
| Artefact Development, Deployment, Implementation | Week 4 -5 |
| Artefact Evaluation and Testing | Week 4 -5 |
| Artefact Presentation and Demonstration | Week 6 |
| Artefact Screencast | Week 6 |

Table 2: Artefact Delivery Schedule

# **Implementation**

Project implementation and data storytelling are two critical aspects of any data-driven project. Project implementation involves executing the planned activities to achieve the project's objectives, while data storytelling focuses on effectively communicating insights and findings from the data analysis. This section discusses the implementation of the methodology used in solving the problem statement. It discusses the different models that were considered before deciding to use the LSTM model for forecasting future CO2 emissions of the UK.

**4.1 Data Preparation**

To implement the project, the dataset was imported into Jupyter notebook. Numpy and Pandas Python libraries were used to perform data cleaning. Figure 4.1 shows the dataset information and the missing values information in Jupyter Notebook.

A screenshot of a computer

Description automatically generated

A screenshot of a computer code

Description automatically generated

**Figure 4.1 dataset and missing values information.**

**4.2 Exploratory Data Analysis**

After performing data cleaning task, seaborn heatmap was used to graphically show the dataset and to show that the missing values were fixed (All of which have been discussed in detail in the methodology section of this project). This is shown in figure 4.2.

The next step was to run statistical analysis on the dataset. Pandas has a functionality called “describe” that is used to describe data. This function only works on numeric data and discards any other data type. The function returns the count, minimum, maximum, standard deviation and percentile of the dataset. The output of the function is shown in figure 4.3.

A screenshot of a computer

Description automatically generated

**Figure 4.2 heatmap of cleaned dataset.**

A screenshot of a computer

Description automatically generated

**Figure 4.3 the output of Pandas describe function.**

Figure 4.6 shows the distribution of greenhouse gases in the dataset, it can be seen from the graph that CO2 is one of the major contributors. This project is particular about CO2 emissions and hence, other greenhouse gasses were not considered. The distribution of CO2 emissions sector wise is shown in figure 4.5. In the figure, Energy supply, Transport, Public and Residential are the largest contributor to CO2 emissions in the UK sector wise, while Waste management, Industrial processes, and Agriculture generate very low emissions. Land use and Forestry generates the least CO2 emissions.

A graph with different colored bars

Description automatically generated

**Figure 4.5 CO2 emissions by sectors**

The next step is to extract CO2 emissions data from the dataset. this is shown in figure 4.7. To make the graph more readable, dotted vertical lines are used to identify each year in the graph. Remember the year started from 1990 to 2021. This is shown in figure 4.8.

A screenshot of a computer

Description automatically generated

**Figure 4.6 Distribution of greenhouse gas emissions in the dataset.**

A graph on a white background

Description automatically generated

**Figure 4.7 distribution of total CO2 emissions from of the UK from 1990 to 2021**

A graph with lines and dots

Description automatically generated

**Figure 4.8 total CO2 emissions from of the UK from 1990 to 2021 with dotted grid lines.**

Using a time series dataset for forecasting purpose requires that the dataset is stationary. This is achieved by ensuring the data has a constant mean, variance and covariance. There are many options to achieving stationarity. Some of these options are Decomposition, Moving average, trend filters, polynomial fitting, Box cox transformations and Log transformations.

**4.2 Test of Stationarity of the Dataset**

The first step is to determine if the timeseries data is stationary. To determine this, the Dicky-Fuller test for stationarity was performed. The Dickey-Fuller test is a statistical test used to determine whether a time series data set is stationary or non-stationary. Stationarity is an important concept in time series analysis, as many statistical techniques and models assume that the data is stationary. A stationary time series is one whose statistical properties, such as mean, variance, and autocorrelation, do not change over time. Non-stationary time series, on the other hand, exhibit trends, seasonality, or other patterns that make their statistical properties dependent on the time at which they are observed.

The Dickey-Fuller test specifically focuses on testing for the presence of a unit root in the time series data. A unit root implies that the time series has a stochastic trend and is non-stationary. The Dickey-Fuller test helps us assess whether the series can be transformed to achieve stationarity, usually through differencing the data. The test is formulated as follows:

* Null Hypothesis (H0): The time series has a unit root (non-stationary).
* Alternative Hypothesis (H1): The time series does not have a unit root (stationary).

The test statistic is compared with critical values from statistical tables to determine whether the null hypothesis can be rejected. If the test statistic is less negative than the critical value, the null hypothesis is not rejected, indicating that the data is non-stationary. If the test statistic is more negative than the critical value, you can reject the null hypothesis and conclude that the data is stationary.

The result of performing the Dicky-Fuller test of stationarity is shown in figure 4.9. From the test, it can be seen that the data is stationary because the test statistic is less than 1% of the critical values and hence, we can be 99% confident that this is stationary.

A graph with blue lines and red lines

Description automatically generated

**Figure 4.9 the output of the Dicky-Fuller test**

After testing for stationarity and determining that the data is stationary, the next step in the implementation process is to decompose the data.

**4.3 Dataset Decomposition**

Decomposing time series data is a common technique used to understand and analyze the underlying components of a time series, including trends, seasonal patterns, and residual variations. This decomposition can help in making predictions, identifying anomalies, and gaining insights into the behavior of the data over time. There are several methods for decomposing time series data, with the most common being the additive and multiplicative decomposition approaches. These are discussed below:

1. **Additive Decomposition:** In additive decomposition, the time series is divided into three main components: trend, seasonality, and residuals. The general formula for additive decomposition is:

**Time Series = Trend + Seasonality + Residual**

* + **Trend:** The long-term movement or behavior of the time series data. It represents the overall direction in which the data is moving. Trends can be upward, downward, or flat.
  + **Seasonality:** The repeating, periodic patterns that occur at fixed intervals, such as daily, weekly, or monthly. Seasonal patterns might be driven by external factors like holidays or weather.
  + **Residual:** The random noise or irregular variations in the data that cannot be attributed to the trend or seasonality. It represents the unexplained fluctuations in the time series.

1. **Multiplicative Decomposition:** In multiplicative decomposition, the time series components are multiplied together:

**Time Series = Trend \* Seasonality \* Residual**

Multiplicative decomposition is used when the amplitude of seasonality changes with the level of the time series data. For instance, if the seasonal variations become larger as the values of the time series increase, a multiplicative decomposition might be more appropriate.

The additive approach was adopted for this project. The output is shown in figure 4.10. A look at the output shows that there is no seasonality in the dataset. To answer the Fifth research question, **“Is there any significant seasonal difference in CO2 emissions of the UK from year to year (taking into consideration the effect Covid-19 had during the periods of 2019-2021)?”,** it is obvious from the decomposition of the dataset that there is no seasonality in the dataset. and hence, it can be stated that there is no significant seasonal difference between the total CO2 emissions of one year to the next. It can also be stated that the Covid-19 pandemic had no effect on the seasonality of the dataset. Finally, it can also be seen that there is a downward trend in the dataset.

A graph of a number of data

Description automatically generated

**Figure 4.10 decomposition of the dataset showing Trend, Seasonality and Residue.**

Finally, the last test to perform on the timeseries dataset is to determine if the series is autocorrelated.

**4.4 Test for Data Correlation**

Time series autocorrelation refers to the correlation of a time series with its own past values. It measures how a data point at a given time is correlated with previous data points at various lags. Autocorrelation is a fundamental concept in time series analysis and plays a crucial role in understanding the underlying patterns and dependencies within the data.

The autocorrelation function (ACF) and partial autocorrelation function (PACF) are commonly used tools to analyze autocorrelation in time series data:

1. **Autocorrelation Function (ACF):** The ACF measures the correlation between a time series and its lagged values. It gives you insights into the relationship between a data point and its historical values at different time lags. The ACF is plotted against the lag on the x-axis and the correlation coefficient on the y-axis. In a stationary time series, the ACF typically decreases as the lag increases.
2. **Partial Autocorrelation Function (PACF):** The PACF measures the direct relationship between a data point and its lagged values, while removing the correlation effects of intervening lags. It helps in identifying the order of an autoregressive model (AR) for modeling the data. A significant spike in the PACF at a certain lag suggests that the data point at that lag is directly related to the data point at the current time step.

Figure 4.11 shows the autocorrelation and partial autocorrelation functions output for the dataset.

A graph of a graph of a graph

Description automatically generated with medium confidence

**Figure 4.11 the ACF and PACF plot of the dataset.**

A graph of a graph of co2 emissions

Description automatically generated

**Figure 4.12 the lag plot of the dataset.**

Another way to show the above relationship is through a lag plot. A time series lag plot, also known as a lag scatter plot, is a graphical tool used to visualize the relationship between a data point and its lagged values. It's a simple yet insightful way to assess the autocorrelation and pattern of dependence within a time series. In a lag plot, each data point is plotted against its lagged value(s), typically one or more time steps behind. Figure 4.12 shows the lag plot for the timeseries dataset.

From figure 4.11 and 4.12, there is a level of autocorrelation in the dataset. This level of correlation is very strong and serves the purpose of this project because it can be said that there is a correlation between past and present data points in the dataset which answers the fourth research question: **“Is there any discernable trend in CO2 emissions of the UK from year to year (taking into consideration the effect Covid-19 had during the periods of 2019-2021)?”**. What this means is that there is a downward trend in CO2 emissions of the UK from year to year. It is obvious that there is a sharp dip in CO2 emissions towards the end of the graph, but this is due to the Covid-19 pandemic spanning 2019 to early 2021. Even with the effect of the pandemic, there is a noticeable downward trend of CO2 emissions overall through 1990 to 2021 as can be seen in figure 4.10.

**4.5 Model Selection**

After preprocessing and cleaning the dataset, the cleaned dataset is split into a train and test set. The train set is used to train a selected model while the test set is used to evaluate the selected model. Now that the problem, objectives and the inference of this project have all been clearly defined, it is time to select a model to use for forecasting CO2 emissions of the UK. Model selection is a critical step in machine learning and statistical modeling, where we choose the most appropriate algorithm or model for a given problem or dataset. The goal is to find a model that can generalize well to unseen data and make accurate predictions or inferences. The next step in the implementation process is to consider different models and finally select the best to perform the required forecast.

**4.5.1 Auto Regression Model**

The first model to consider for this project is the Auto Regression model. An autoregression model, often referred to as an autoregressive (AR) model, is a statistical time series forecasting model that uses past observations of a time series to predict future values. Autoregressive models are commonly used in time series analysis and econometrics. They are part of a broader class of models called autoregressive integrated moving average (ARIMA) models. The order of an autoregressive model, denoted as "p," represents the number of lagged (previous) observations of the time series that are used as predictors for the current value. For example, in a first-order autoregressive model (AR(1)), only the immediate previous value is used; in a second-order autoregressive model (AR(2)), the two most recent values are used, and so on. The general mathematical representation of an autoregressive model of order p is:

**Y(t) = c + ϕ₁ \* Y(t-1) + ϕ₂ \* Y(t-2) + ... + ϕp \* Y(t-p) + ε(t)**

Here, Y(t) represents the current value of the time series, Y(t-i) represents the lagged values up to lag p, ϕ₁, ϕ₂, ..., ϕp are the autoregressive coefficients, c is a constant, and ε(t) is white noise, representing random error. Estimating the autoregressive coefficients (ϕ₁, ϕ₂, ..., ϕp) and the constant (c) is typically done using statistical methods like maximum likelihood estimation (MLE).

Once the autoregressive model is estimated and validated, it can be used for forecasting future values of the time series. Forecasting can be done recursively, where the model generates predictions for one step ahead and then uses those predictions for subsequent steps. The model makes the following assumptions:

* Autoregressive models assume that the time series is stationary or can be transformed into a stationary series. Stationarity implies that the statistical properties of the time series do not change over time.
* The white noise term ε(t) is typically assumed to have a mean of zero and constant variance.

Figure 4.13 shows the code for the autoregression model as defined for this project.

A screenshot of a computer program

Description automatically generated

**Figure 4.13 the declaration of the Auto Regression model.**

Autoregressive models can be extended to include seasonal components (seasonal autoregressive models, SAR), integrated components (as in ARIMA models), or other additional factors to capture complex time series patterns. This is discussed in the next section.

**4.5.2 ARIMA Model**

The second model to consider for this project is the ARIMA model. ARIMA stands for AutoRegressive Integrated Moving Average. It's a popular time series forecasting model that combines autoregressive (AR) and moving average (MA) components with differencing to handle non-stationary data. ARIMA models are widely used for making predictions and analyzing time series data.

1. **AutoRegressive (AR) Component:** The autoregressive component involves modeling the relationship between the current value of the time series and its past values. An AR(p) model uses the previous p values to predict the current value. Mathematically, an AR(p) model can be represented as:

**y(t) = c + φ₁\*y(t-1) + φ₂\*y(t-2) + ... + φₚ\*y(t-p) + ε(t)**

Where:

* + y(t) is the value at time t.
  + φ₁, φ₂, ..., φₚ are the autoregressive coefficients.
  + c is a constant term.
  + ε(t) is the white noise error term at time t.

1. **Integrated (I) Component:** The integrated component involves differencing the time series to make it stationary. Stationary data has constant mean and variance over time, which simplifies modeling. The "integrated" part of ARIMA comes from this differencing process. The order of differencing, denoted as d, indicates how many times the data needs to be differenced to achieve stationarity.
2. **Moving Average (MA) Component:** The moving average component models the relationship between the current value of the time series and past forecast errors (residuals). An MA(q) model uses the q most recent forecast errors to predict the current value. Mathematically, an MA(q) model can be represented as:

**y(t) = c + θ₁\*ε(t-1) + θ₂\*ε(t-2) + ... + θₚ\*ε(t-q) + ε(t)**

Where:

* + ε(t-1), ε(t-2), ..., ε(t-q) are the past forecast errors.
  + θ₁, θ₂, ..., θₚ are the moving average coefficients.

The ARIMA model combines these three components into a single framework:

* p: The number of autoregressive lags.
* d: The order of differencing.
* q: The number of moving average lags.

ARIMA models can be quite effective for forecasting time series data, especially when the data exhibits trends, seasonality, or other complex patterns. However, finding the appropriate values for p, d, and q requires experimentation, often involving methods like grid search or more advanced techniques like SARIMA (Seasonal ARIMA). The summary and output of fitting the model is shown in figure 4.13.

A screenshot of a computer

Description automatically generated

A graph with a line

Description automatically generatedA blue line graph with numbers

Description automatically generated

A black numbers and a white background

Description automatically generated

**Figure 4.14 the output of training the ARIMA model.**

**4.5.3 Exponential Smoothening Model**

The third model to consider for this project is Exponential smoothening model. Exponential smoothing is a time series forecasting method that is used to forecast data points by giving more weight to recent observations while gradually decreasing the weight of older observations. It is particularly useful for time series data that exhibit trend or seasonality. Exponential smoothing comes in several variations, and one of the most commonly used methods is the simple exponential smoothing model. The forecasted value for a specific time period (t) is the primary output of the exponential smoothing model. This forecast is based on the weighted average of past observations. The smoothing parameter, denoted as α (alpha), is a constant value between 0 and 1 that determines the weight assigned to the most recent observation. A higher α places more emphasis on recent data, while a lower α places more emphasis on older data. To start the forecasting process, an initial forecast (F1) is typically required. It can be based on the first observation in the time series or set to an arbitrary value. The simple exponential smoothing model is represented by the following equations:

* + **Initialization: F1 = Y1 (the first observed value)**
  + **Forecasting: Ft = α \* Yt + (1 - α) \* Ft-1**

Here, Ft represents the forecasted value at time t, Yt represents the observed value at time t, and Ft-1 represents the previous forecasted value. The choice of the smoothing parameter α is critical. Smaller values of α make the forecast less responsive to recent changes, while larger values make it more responsive. α is often chosen through trial and error or through automated optimization techniques. Figure 4.15 shows the code for the Exponential smoothening model as defined for this project.

A screenshot of a computer program

Description automatically generated

**Figure 4.15 the declaration of the Exponential smoothening model.**

Exponential smoothing is a widely used technique in time series forecasting due to its simplicity and effectiveness in capturing short-term patterns. However, it may not perform well on time series with complex long-term trends or irregular patterns, for which more advanced methods may be required. A more advanced model, the LSTM, is discussed in the next section.

**4.5.4 LSTM Model**

The last model, which was discussed in the methodology section, is a deep learning model. The deep learning model consists of an LSTM layer that is used with a fully connected layer. The output of the model is a tuple. The first element in the model is the generated hidden states with each of them representing a time step of the time series input, the input is then parsed to the second element of the model which is the cell memory and hidden states. The LSTM parameter for batch processing is set to true because the input from the formula receives a tensor (window size, time step, features). Finally, the processed time series data is parsed out through the output hidden states with the tanh activation function as a single regression output. The definition of the LSTM class is shown in figure 4.16.

A screenshot of a computer code

Description automatically generated

**Figure 4.16 the declaration of the LSTM model.**

After building the four models above, they were trained and evaluated on the training set. The root mean squared error (RMSE) of the models are discussed in the next section.

# **Results**

Discussing machine learning results is a crucial step in the data science process. Effective result discussion provides insights into the model's performance, its strengths and weaknesses, and the potential implications of the findings. This project has created four different predictive models for forecasting total CO2 emissions of the United Kingdom given past and present dataset. Exploratory data analysis was carried out to gain insight of the dataset and to make sense of the dataset. The designed models (discussed in the implementation section) were trained with the training dataset. The results from both the models are shown in figure 5.1.

A graph of a training output

Description automatically generatedA graph of a graph with a line and a line

Description automatically generated with medium confidence

A graph of a graph showing the growth of a training output

Description automatically generatedA graph of a training output

Description automatically generated

**Figure 5.1 the training output from the four models.**

To understand and compare the models, there is a need to evaluate their outputs. The root mean squared error (RMSE) is the matric that was used for this purpose and is discussed in the evaluation section.

**5.1Evaluation**

Machine learning model evaluation is a critical step in the model development process. It involves assessing the performance and generalization ability of a trained model using various metrics and techniques. The evaluation process helps determine how well the model is likely to perform on unseen data and whether it meets the desired criteria for deployment.

The developed models were evaluated using the Root Mean Squared Error (RMSE) matric. Root Mean Squared Error (RMSE) is a commonly used metric for measuring the accuracy of a predictive model, particularly in regression analysis. It provides an estimate of how closely the predicted values match the actual values in a dataset. RMSE is particularly useful when there is a need to penalize larger errors more heavily, as it squares the differences between predicted and actual values.

The formula for calculating RMSE is as follows:

Where:

* n is the number of data points in the dataset.
* yirepresents the actual (observed) value of the target variable for the i*i*th data point.
* y^irepresents the predicted value of the target variable for the i*i*th data point.

Key points about RMSE:

* **Range and Interpretation:** RMSE is measured in the same units as the target variable. It gives you an idea of the average magnitude of the prediction errors.
* **Squaring of Errors:** Squaring the errors ensures that larger errors have a more significant impact on the RMSE value, making it sensitive to outliers.
* **Lower Values are Better:** A lower RMSE indicates better model performance, as it means the predicted values are closer to the actual values.

When interpreting RMSE:

* A RMSE of 0 indicates that the model's predictions perfectly match the actual data, which is rarely achievable in practice.
* A lower RMSE indicates better predictive accuracy, but the "goodness" of RMSE depends on the specific problem and the range of the target variable.

The LSTM model has an RMSE value of 13.693. This was achieved after several hyperparameter tunning, while the Exponential smoothening, ARIMA, and Auto regression models had RMSE’s of 25.624, 19.365 and 18.040 respectively. From the result of the root mean squared error, the LSTM model performs better at forecasting the dataset the remaining models. The result is shown in the table below and the graph in figure 5.2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MATRIC | EXPONENTIAL | ARIMA | AUTOREG | LSTM |
| RMSE | 25.624 | 19.365 | 18.040 | 13.693 |

**Figure 5.3 the comparison of LSTM and ARIMA**

It is obvious that the LSTM performed better than the other models in this case. This is a direct result of tunning the hyperparameters while training the LSTM.

## **5.2 Related Work**

After evaluating the models, it was discovered that LSTM performed better than the other models hence, it was used to forecast CO2 emissions for twelve years into the future. The result of this forecast is shown in figure 5.4.

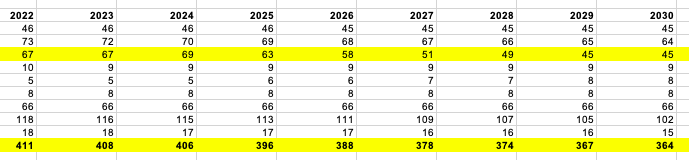
A graph showing the growth of a model

Description automatically generatedA table of numbers and symbols

Description automatically generated

**Figure 5.4 the twelve years forecast of the LSTM model.**

This result was compared with similar forecast made by the Department for Business, Energy & Industrial Strategy (shown in figure 5.5). Table 5.1 shows the comparison of these forecasts and the graph in figure 5.6 puts it in perspective.



**Figure 5.5 DBEIS forecast.**

Table 5.1 forecast comparison

|  |  |  |
| --- | --- | --- |
| **YEAR** | **Model** | **DBEIS** |
| 2022 | 422 | 411 |
| 2023 | 385 | 408 |
| 2024 | 392 | 406 |
| 2025 | 387 | 396 |
| 2026 | 355 | 388 |
| 2027 | 354 | 378 |
| 2028 | 337 | 374 |
| 2029 | 324 | 367 |
| 2030 | 315 | 364 |
| 2031 | 303 | 360 |
| 2032 | 289 | 354 |
| 2033 | 268 | 349 |

**Figure 5.6 graph of comparing both forecasts.**

It is obvious that the model performed very well showing only little differences starting from 2025 to 2033 when compared.

# **Conclusion**

The UK has committed to achieving a series of emissions reduction targets. By 2020, the goal was to reduce emissions by at least 34% compared to 1990 levels. By 2030, the target is to reduce emissions by at least 68%, and the ultimate aim is to achieve net-zero greenhouse gas emissions by 2050. he UK has made significant progress in reducing greenhouse gas emissions. It has transitioned from coal-fired power generation to cleaner alternatives such as natural gas and renewable energy sources. Additionally, policies such as the Carbon Price Floor and Renewable Obligation have helped incentivize low-carbon technologies.

## The United Kingdom's (UK) CO2 emissions and climate change efforts are significant topics due to the country's commitment to addressing environmental challenges and reducing its carbon footprint. The UK has historically been one of the largest emitters of CO2 in Europe due to its reliance on fossil fuels for energy production, industry, and transportation.Over the years, the UK has taken various steps to reduce emissions, including transitioning to cleaner energy sources and implementing policies to promote energy efficiency.

The Climate Change Act of 2008 established a legally binding commitment for the UK to reduce its greenhouse gas emissions by at least 80% by 2050 compared to 1990 levels. This target has since been updated to achieve net-zero emissions by 2050. The UK has made significant progress in transitioning its energy mix toward renewable sources such as wind, solar, and hydroelectric power. Offshore wind farms in the UK are among the largest and most efficient in the world. The UK has committed to phasing out unabated coal power generation by 2024. Coal-fired power plants have been closing or converting to alternative fuels as part of efforts to reduce emissions. The UK government has set a target to ban the sale of new petrol and diesel cars by 2030, promoting the adoption of electric vehicles (EVs). Charging infrastructure for EVs is being expanded across the country.

The UK participates in the European Union Emissions Trading System (EU ETS), which places a cap on emissions from industries and allows trading of emission allowances. The UK hosted the 26th UN Climate Change Conference of the Parties (COP26) in 2021, where global leaders discussed and made commitments to tackle climate change. The UK is also focusing on climate adaptation measures to cope with the effects of climate change, such as rising sea levels and extreme weather events. The UK is investing in CCS technology to capture CO2 emissions from industrial processes and power plants, with the goal of reducing emissions.

## **6.1 Reflection**

The UK government and various organizations promote public awareness and education on climate change, encouraging individuals to take action to reduce their carbon footprint. The UK's efforts to reduce CO2 emissions and combat climate change are part of a broader global movement to address the environmental challenges posed by greenhouse gas emissions. These efforts are crucial in mitigating the impacts of climate change and transitioning to a more sustainable and low-carbon future.

This project has investigated the CO2 emissions characteristics of the UK with respect to its constituent countries and sectors. Data was collected and preprocessed to make them analysis ready and suitable for forecasting. A deep learning model was created to forecast future emission rates of the UK. The model is evaluated using the Root Mean Squared Error (RMSE) matric. RMSE is a commonly used metric for evaluating the performance of a regression model. It measures the average deviation between the predicted values of the model and the actual (observed) values in the dataset. RMSE is a popular choice because it penalizes larger errors more heavily, making it sensitive to outliers and better suited for assessing the accuracy of predictions.

## **6.2 Future Work**

When considering future work related to CO2 emissions characteristics, involves looking at further research and analysis to deepen our understanding of the factors influencing CO2 emissions, their patterns, and potential strategies for reduction. The following are the area that needs further research:

1. There is a need to conduct a more extensive temporal analysis to identify trends and fluctuations in CO2 emissions over longer time periods. This could involve analyzing data from different decades or historical records.
2. It is important to explore spatial patterns of CO2 emissions at regional, national, or global scales. Identify regions with high emissions and investigate the underlying factors contributing to these differences.
3. Dive deeper into specific sectors contributing to CO2 emissions, such as energy production, transportation, industry, and agriculture. Examine the characteristics of each sector's emissions and explore mitigation strategies.
4. Investigate the relationship between economic indicators (GDP, industrial output, etc.) and CO2 emissions. Examine how economic growth and emissions are interconnected.
5. Study the transition from fossil fuels to renewable energy sources and its impact on CO2 emissions. Analyze the effectiveness of policies promoting clean energy adoption.
6. Explore the role of technological advancements, such as carbon capture and storage (CCS) and sustainable transportation technologies, in reducing CO2 emissions.
7. Examine the effectiveness of different environmental policies and regulations in curbing CO2 emissions. Assess the challenges and opportunities in policy implementation.
8. Investigate how societal behaviors, consumer preferences, and lifestyle choices impact emissions. Analyze the potential for behavior change campaigns to reduce carbon footprints.
9. Study the influence of urbanization and land use changes on CO2 emissions. Analyze emissions from urban areas and explore sustainable urban planning practices.
10. Compare CO2 emissions characteristics between different countries or regions. Identify best practices and lessons that can be learned from successful emission reduction strategies.
11. Investigate the implications of CO2 emissions characteristics on climate change. Analyze how different emission profiles contribute to global temperature rise and related impacts.
12. Examine the vulnerability of specific regions or communities to the impacts of CO2 emissions. Explore adaptation strategies to mitigate potential risks.
13. Collaborate with experts from fields such as economics, sociology, environmental science, and policy to gain a holistic understanding of CO2 emissions characteristics.

It is important to note that all these areas of research identified above for future work represents a particular research topic that can be broken down into smaller research topics.

I intend to publish the findings of this project in a journal. Some of the journals that has been considered are: Renewable and Sustainable Energy Reviews, Science of The Total Environment, Energy & Environmental Science and Environmental Science & Technology.

# **References**

1. 'Achieving net zero emissions with machine learning: the challenge ahead', (2022) *Nature Machine Intelligence,* 4(8), pp. 661-662. Available at: <https://doi.org/10.1038/s42256-022-00529-w.>
2. 'Greenhouse gas emissions, UK - Office for National Statistics', (2022) Available at: <https://www.ons.gov.uk/economy/environmentalaccounts/bulletins/greenhousegasintensityprovisionalestimatesuk/2021.>
3. *8 Billion Lives, Infinite Possibilities.* Available at: <https://www.unfpa.org/swp2023> (Accessed: Jul 12, 2023).
4. 'Atmospheric emissions: greenhouse gas emissions intensity by industry - Office for National Statistics', Available at: <https://www.ons.gov.uk/economy/environmentalaccounts/datasets/ukenvironmentalaccountsatmosphericemissionsgreenhousegasemissionsintensitybyeconomicsectorunitedkingdom.>
5. Akhshik, M., Bilton, A., Tjong, J., Singh, C.V., Faruk, O. and Sain, M. (2022) 'Prediction of greenhouse gas emissions reductions via machine learning algorithms: Toward an artificial intelligence-based life cycle assessment for automotive lightweighting', *Sustainable Materials and Technologies,* 31, pp. e00370. Available at: <https://doi.org/10.1016/j.susmat.2021.e00370.>
6. Alfaseeh, L., Tu, R., Farooq, B. and Hatzopoulou, M. (2020) 'Greenhouse gas emission prediction on road network using deep sequence learning', *Transportation Research Part D: Transport and Environment,* 88, pp. 102593. Available at: <https://doi.org/10.1016/j.trd.2020.102593.>
7. Bakay, M.S. and Ağbulut, Ü (2021) 'Electricity production based forecasting of greenhouse gas emissions in Turkey with deep learning, support vector machine and artificial neural network algorithms', *Journal of Cleaner Production,* 285, pp. 125324. Available at: <https://doi.org/10.1016/j.jclepro.2020.125324.>
8. Dan Bishop 'Predicting Greenhouse Gas Emissions with Machine Learning | by danbishop | Medium', Available at: <https://medium.com/@datawithdan_/predicting-greenhouse-gas-emissions-with-machine-learning-f27560ac101b.>
9. Department for Energy Security and Net Zero and Department for Business, Energy & Industrial Strategy *UK greenhouse gas emissions statistics.* Available at: <https://www.gov.uk/government/collections/uk-greenhouse-gas-emissions-statistics> (Accessed: Jul 14, 2023).
10. Department for Environment, Food and Rural Affairs (2023) *UK and England's carbon footprint to 2020.* Available at: <https://www.gov.uk/government/statistics/uks-carbon-footprint> (Accessed: Jul 14, 2023).
11. Emami Javanmard, M. and Ghaderi, S.F. (2022) 'A Hybrid Model with Applying Machine Learning Algorithms and Optimization Model to Forecast Greenhouse Gas Emissions with Energy Market Data', *Sustainable Cities and Society,* 82, pp. 103886. Available at: <https://doi.org/10.1016/j.scs.2022.103886.>
12. Hamrani, A., Akbarzadeh, A. and Madramootoo, C.A. (2020) 'Machine learning for predicting greenhouse gas emissions from agricultural soils', *Science of The Total Environment,* 741, pp. 140338. Available at: <https://doi.org/10.1016/j.scitotenv.2020.140338.>
13. Jiang, Z., Yang, S., Smith, P. and Pang, Q. (2023) 'Ensemble machine learning for modeling greenhouse gas emissions at different time scales from irrigated paddy fields', *Field Crops Research,* 292, pp. 108821. Available at: <https://doi.org/10.1016/j.fcr.2023.108821.>
14. Kumari, S. and Singh, S.K. (2022) 'Machine learning-based time series models for effective CO2 emission prediction in India', *Environmental Science and Pollution Research,* Available at: <https://doi.org/10.1007/s11356-022-21723-8.>
15. L. Amarpuri, N. Yadav, G. Kumar and S. Agrawal (2019) *Prediction of CO2 emissions using deep learning hybrid approach: A Case Study in Indian Context.* pp. 1.
16. Meng, Y. and Noman, H. (2022) 'Predicting CO2 Emission Footprint Using AI through Machine Learning', *Atmosphere,* 13(11) Available at: <https://doi.org/10.3390/atmos13111871.>
17. Office for National Statistics, U.K. 'Population estimates for the UK, England, Wales, Scotland and Northern Ireland - Office for National Statistics', Available at: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2021.>
18. PwC Press *UK could remain a top-10 global economy in 2050 - despite Brexit.* Available at: <https://www.pwc.co.uk/press-room/press-releases/regions/northern-ireland/world-in-2050.html> (Accessed: Jul 13, 2023).
19. Zhong, W. and Haigh, J. (2013) 'The greenhouse effect and carbon dioxide', *Weather,* 68, pp. 100-105. Available at: <https://doi.org/10.1002/wea.2072.>

# **Appendices**

A screenshot of a computer

Description automatically generated

**Figure 8.1 Dataset information**

A screenshot of a computer code

Description automatically generated

**Figure 8.2 Missing value count**

A screenshot of a computer

Description automatically generated

**Figure 8.3 Clean data function**

A screenshot of a computer code

Description automatically generated

**Figure 8.4 dropped columns from the dataset.**

A screen shot of a computer code

Description automatically generated

**Figure 8.5 Formatting the Calendar year parameter.**

A graph with a line going up

Description automatically generated

**Figure 8.6 screenshot of total greenhouse gas emissions from 1990 to 2021**

A graph of emissions by countries/regions

Description automatically generatedA graph of emission emissions

Description automatically generated with medium confidence

**Figure 8.7 CO2 emissions by countries in the UK.**

A graph of different colored bars

Description automatically generated

**Figure 8.8 Mid-year population by region.**

A graph of different colored lines

Description automatically generatedA graph of different colored bars

Description automatically generated

**Figure 8.9 sector-wise contributions to CO2 emissions in the UK**

A graph with different colored bars

Description automatically generated

**Figure 8.10 Mid-year population of regions in England**

A screen shot of a graph

Description automatically generated

**Figure 8.11 greenhouse gas emissions of the UK.**

A screenshot of a computer

Description automatically generated

**Figure 8.12 heatmap of cleaned dataset.**

A screenshot of a computer

Description automatically generated

**Figure 8.13 the output of Pandas describe function.**

A graph with different colored bars

Description automatically generated

**Figure 8.14 CO2 emissions by sectors**

A graph on a white background

Description automatically generated

**Figure 8.15 distribution of total CO2 emissions from of the UK from 1990 to 2021**

A graph with lines and dots

Description automatically generated

**Figure 8.16 total CO2 emissions from of the UK from 1990 to 2021 with dotted grid lines.**

A graph with blue lines and red lines

Description automatically generated

**Figure 8.17 the output of the Dicky-Fuller test**

A graph of a number of data

Description automatically generated

**Figure 8.18 decomposition of the dataset showing Trend, Seasonality and Residue.**

A graph of a graph of a graph

Description automatically generated with medium confidence

**Figure 8.19 the ACF and PACF plot of the dataset.**

A graph of a graph of co2 emissions

Description automatically generated

**Figure 8.20 the lag plot of the dataset.**

A screenshot of a computer program

Description automatically generated

**Figure 8.21 the declaration of the Auto Regression model.**

A screenshot of a computer

Description automatically generated

**Figure 8.22a the output of training the ARIMA model.**

A graph with a line

Description automatically generatedA blue line graph with numbers

Description automatically generated

A black numbers and a white background

Description automatically generated

**Figure 8.22b the output of training the ARIMA model.**

A screenshot of a computer program

Description automatically generated

**Figure 8.23 the declaration of the Exponential smoothening model.**

A screenshot of a computer code

Description automatically generated

**Figure 8.24 the declaration of the LSTM model.**

A graph of a training output

Description automatically generatedA graph of a graph with a line and a line

Description automatically generated with medium confidence

A graph of a graph showing the growth of a training output

Description automatically generatedA graph of a training output

Description automatically generated

**Figure 8.25 the training output from the four models.**

**Figure 8.26 the comparison of LSTM and ARIMA**

A graph showing the growth of a model

Description automatically generatedA table of numbers and symbols

Description automatically generated

**Figure 8.27 the twelve years forecast of the LSTM model.**

A table with numbers and a yellow stripe

Description automatically generated

**Figure 8.28 DBEIS forecast.**

**Figure 8.29 graph of comparing both forecasts.**

A screenshot of a computer

Description automatically generated

**Figure 8.30 teamwork project management tool.**