Southampton Solent University

Faculty of Business, Law and Digital Technologies

MSc Applied AI and Data Science

2022

Simisola Kaothara Olagunju

"Stacking ensemble machine learning algorithm for predicting carbon emissions and reduction strategies"

Supervisor : Dr Muntasir Al-Asfoor
Date of presentation : September 2022

This report is submitted in partial fulfilment of the requirements of Solent University for the degree of MSc Applied AI and Data Science

ACKNOWLEDGEMENTS

I give thanks to God Almighty, the Most Gracious, the Most Merciful, and the source of all wisdom, knowledge, and understanding, for preserving my life to this day.

I would like to convey my sincere gratitude to Dr. Muntasir Al-Asfoor, under whose supervision I have been enormously enriched, for his thoughtful feedback and advice that have allowed me to finish this project.

I owe my deepest gratitude to my husband, Maroof, who has always believed in me and served as a pillar for me during this journey. I appreciate how he helped me through difficult moments with his compassion and excellent sense of humour. None of this would have been possible without your love and endurance. My children, Umar, Asma, Sulaiman, and Maimoonah, deserve my sincere gratitude for their sacrifice and invaluable assistance in making this success possible. I apologise for being grumpier than usual while I was writing this dissertation.

I'm also grateful to my brother, Dr. Alim Sabur Ajibola, for reviewing and criticising my research and helping me to enhance my final research paper. An extra special thanks to my sisters Azeez Qoyumat and Azeez Queen, who completed all the proofreading for my thesis and provided me with excellent feedback. Their support has always been immeasurable for me.

I want to pay my heartiest appreciation to my esteemed mother for her prayers, love, encouragement and support she has provided me throughout my years in the pursuit of academic success. She has instilled in me the value of honesty and hard work which have, and will continue to, guide me in my life endeavours.

In addition, I would like to thank my family and friends for their moral, spiritual, and emotional support throughout this challenging academic year, without which I would not have been able to do anything of value. My genuine gratitude goes out to my classmates Hadizat, Peter, Tinu, and Taiwo for their encouragement when I was doubting my abilities or my work.

I must not fail to express my gratitude to my lecturers and colleagues for their ongoing support and academic counsel, which formed an invaluable component of this dissertation. It gave me a lot of inspiration to see how well my family, my supervisor, and all of you worked together to make sure that my study was successful. In terms of the study, it took a while to reach success. I am grateful to everyone who gave assistance to me when I was studying. You're all appreciated.

ABSTRACT

A major factor causing global warming and climate change is carbon emissions. Since the UK intended to attain net zero carbon emissions by 2050, accurate and steady carbon emissions prediction is helpful for developing emissions reduction plans and attaining carbon neutrality as soon as possible Although many prior studies used one or more models to anticipate carbon emissions, there has been little progress in forecasting accuracy and few studies on UK carbon emissions. Based on UK carbon emissions data from 2005 to 2019, the proposed stacking ensemble model was created by combining three separate base predictive models optimally: random forest, support vector machine and k-nearest neighbour learning models using gradient boosting as meta-regressor. The GridSearchCV optimises the basic model hyperparameters to create a high-performance stacking ensemble model. The data analysis identifies three high carbon emitting sectors as being the industrial, transportation, and domestic sectors. The findings indicate that the proposed model accurately forecasted the carbon emissions, with a very low prediction error value. This suggests that the variation between the expected and actual emissions was minimal. The root mean square error for the suggested model was 0.44. Based on the values of the root mean square error, it can be shown that the stacking ensemble model surpasses the random forest- 0.49, gradient boosting- 0.46, k-nearest neighbour- 0.67, and support vector machine- 2.71 models in terms of accuracy. As a result, the proposed model significantly outperforms other models at forecasting, and the results can provide a useful guide for decision-makers as they develop environmental and climate change policies that are linked to the carbon peaking and neutrality targets. However, this study offers a model that is simple to duplicate and use. The model is effective and adaptable to various environmental data types and can also be used to conduct research on other socioeconomic-related issues.

CONTENTS

FIGURES

TABLES

ACRONYMS

AI	artificial intelligence
ANN	artificial neural network
ARIMA	autoregressive moving average
ARMA	autoregressive moving average
BPNN	back propagation neural network
CO_2	carbon dioxide
CRISP-DM	cross-industry standard process for data mining
CSO	chicken swarm optimization
DE	differential evolution
DGM	discrete grey model
DPSO	double improved particle swarm optimization
DS	data science
EDA	exploratory data analysis
ELM	extreme learning machine
ENN	Elman neural network
EV	electric motor vehicle
FLN	fast learning network
G20	Group of Twenty
GA	generic algorithm
GBR	gradient boosting regressor
GDP	gross domestic product
GHG	greenhouse gases
GM	grey model
GPR	gaussian process regression
GRNN	generalized regression neural network
GRU	gated recurrent unit
IMM	inclusive multiple model
KF	Kalman filter
KNN	k-nearest neighbor
KNNR	k-nearest neighbor regressor
LSSVM	least square support vector machine
LSTM	long short-term memory
MAE	mean absolute error

MAPE	mean absolute percentage error
MGP	multivariate grey model
ML	machine learning
MLP	multi-layer perceptron
MLR	multiple linear regression
MSE	mean square error
PCA	principal component analysis
PSO	particle swarm optimization
R^2	coefficient of correlation
RA	regression analysis
RBF	Radial basis function neural network
RF	random forest
RFR	random forest regressor
RMSE	root mean square error
SOM	self-organising map
SR	stacking regressor
SVM	support vector machine
SVR	support vector regressor
TS	Tabu Search
UK	United Kingdom
CCC	climate change committee

1. INTRODUCTION

1.1 Background

One of the many conditions supporting lives on a planet is the temperature range that is suitable for life. The atmosphere contains specific gases that act as insulation for the Earth's surface from the cold of space. This planet's warmth and habitability are caused by the ability of these gases to absorb infrared radiation from sunlight during the day and radiate it at night. The greenhouse effect is the process of producing and absorbing light at different times of the day and night, and the gases that exhibit this phenomenon are known as greenhouse gases (GHG) (Amarpuri *et al*. 2019).

Carbon dioxide, methane, nitrous oxide, water vapour, and ozone are the main GHGs present in the earth's atmosphere. Despite being essential for life's survival, GHGs have a negative impact on climate change due to their rising atmospheric concentration. The earth's average temperature has sharply increased, rendering some areas uninhabitable (Amarpuri *et al*. 2019).

Over the past few decades, the global greenhouse effect has gotten worse, interrupting the global carbon cycle, which might lead to adverse consequences on global warming, as well as other environmental, physical, and health problems. The long-term advancement of human society and the green economic recovery depend on halting global warming (Ritchie and Roser, 2020; Chen, Qi and Tan 2022).

The Paris Agreement, which was ratified in December 2015, aims to cut GHG emissions and stop global warming. Most nations have set goals for minimising global warming (Qiao *et al*. 2020). The Intergovernmental Panel on Climate Change discovered that over the previous three decades, gas emissions increased as a result of a 0.85°C rise in average land and ocean temperatures (Kadam and Vijayumar 2018).

The UK's GHG emissions in 2021 were 447 $MtCO_2e$, a 47% decrease from levels in 1990. Emissions in 2020 decreased 10% from those in 2019 but increased 4% from those in 2020 as a result of the COVID19 pandemic reaction (CCC 2022).

Due to rising population densities and costly infrastructure that emits these pollutants, our society is progressively becoming more vulnerable to extreme weather events. Economic activities are commonly cited as the main cause of CO_2 emissions. Large-scale CO_2 emissions are inextricably related to a country's economy's rapid expansion (Acheampong and Boateng 2019, Guo *et al*. 2021, Shahbaz *et al*. 2020). An increase in carbon emissions will have a detrimental influence on a number of economic activities, including agriculture, industrial output, population growth, and immigration (Xu *et al*. 2021). Energy generation, industry, trade, transportation, and public and domestic processes are some more economic activities that release carbon dioxide. Other natural factors include respirations and volcanic eruptions (Qader *et al*. 2021).

Studies to date indicate that CO_2 is the predominant greenhouse gas (GHG) in the atmosphere, making up around 99.4% of its parts per billion (ppb). It accounts for over 72% of global warming, excluding water vapour. CO_2 intensity in the atmosphere has increased during the previous 150 years, rising from 280 parts/million to 400 parts/million. Countries around the world have agreed that promoting low-carbon production and lowering CO_2 emissions are the common goals for future sustainable development (Gao *et al*. 2021).

According to UK government website, a long-term alteration of the world's typical weather patterns is climate change and with each of the last three decades warming more than the one before it, 17 of the 18 hottest years ever recorded have happened in the twenty-first century. The UK is already being impacted by increasing temperatures. The most past decade has been 0.8°C warmer than the average between 1961 and 1990. All 10 of the UK's hottest years have occurred since 1990, the last nine since 2002.

The article also mentioned that even if global temperature increases are restricted to 2°C or less, the UK is expected to have impacts. In a world where the temperature rises by 2°C, water levels in the UK may drop by 30% during "dry" seasons and rise by 5-20% during "wet" seasons. Extreme weather conditions are predicted to happen more often in the UK due to rising temperatures.

Due to the world's increasing CO_2 emissions, two serious threats—global warming and climate change—have become increasingly imminent (Nguyen, Huynh and

Nasir 202; Shahbaz *et al*. 2021; Mason, Duggan and Howley 2018). It is undeniable that poor air quality endangers the health of people and has detrimental effects on the environment and the oceans (Qiao *et al*. 2020; Qader *et al*. 2021).

Urbanization and industrialisation have also boosted the world economy while simultaneously causing climate change and global warming (Qader *et al*. 2021). Due to the irreversible effects, international cooperation is required to address these important concerns and maintain sustainable progress (Fang *et al*. 2018).

Environmental issues like climate change, rising sea levels, glacier melt, desertification, unpredictable seasons, severe weather events have become increasingly common in recent years, posing a rising threat to human existence on Earth (Fang *et al*. 2018; Amarpuri *et al*. 2019; Shabri 2022). The likelihood of floods is expected to increase as a result of more regular heavy rainfall incidents, rising sea levels, and an increase in heat waves and droughts in frequency and severity throughout the next century. Wildfires are more likely, and there will be more deaths from heat-related causes.

In order to attain "zero emissions" of carbon dioxide, countries must neutralize all their direct or indirect greenhouse gas emissions produced over a given time period by planting trees, conserving energy, and cutting emissions. Increasingly more nations are now adopting carbon-neutral aims in their national policies (Liu and Zang 2022).

The climate change committee's (CCC) report predicts that achieving net zero CO_2 globally will minimise or stop the consequences of climate change. But if we can limit temperature increases in the future, we can finally reach net-zero global emissions of GHG. If greenhouse gas emissions are decreased globally by 3–4% annually, we still have an opportunity to keep the temperature increase to 1.5°C.

The report pointed out the UK government declared in 2019 a compelled objective to achieve net zero GHG emissions within the UK economy by 2050. As the UK needs to reduce emissions by almost a third to meet the sixth carbon budget by the middle of the 2030s, the current policy would not result in net zero emissions. As a result, delivery must be actively monitored since the plan must be founded on an honest evaluation.

In line with earlier predictions, the UK climate forecasts for 2018 (UKCP18) show a greater possibility of severe weather, rising sea levels, warmer, more wet winters, hotter and drier summers in the UK. If the UK is to decarbonize in a manner that enhances business potential, safeguards energy sources and treats people fairly, significant reforms must be done.

Recent research from Microsoft and PwC indicates that using artificial intelligence (AI) for environmental purposes might reduce global GHG emissions by 1.5 to 4.0% by 2030. As a result of these applications, global greenhouse gas emissions might be reduced by 0.9 to 2.4 gigatons of CO_2e, accelerating the change to a low-carbon society (Herweijer, Combes and Gillham 2018).

Creating climate forecasts and projections is one of the keyways Data Science (DS) is being used to combat climate change. These forecasts and predictions aid in decision-making as well as helping the general public realize the severity of the problem. World leaders may now monitor in real-time how decisions affect carbon emissions thanks to DS and ML. We can gain a greater understanding of the environmental, economic, and social effects of our changing climate thanks to these tools. Data scientists may help turn vast scientific data sets into useful scientific information, which will allow organisations to come up with effective climate change solutions. In a topic as complex as global climate, there are bound to be uncertainties and ambiguities in the Data. DS helps sort through and clarify the ambiguity, so we use the best Data when making decisions (Rauch 2022).

AI's greatest strength is its capacity for experiential learning, enormous data collection from its environment, intuitive connections that people miss, and the recommendation of appropriate actions based on its findings. AI can be used by businesses to monitor emissions, collect data from new sources, determine the degree of certainty of the results, predict emissions to more precisely meet reduction targets, and reduce emissions to save costs (Degot *et al*. 2021).

Forecasts of carbon emissions could be used to predict global warming in the future, provide guidance for policymakers, as well as to estimate the costs of reducing CO_2 and foresee the advantages of limiting temperature rise (Qiao *et al*. 2020; Gao *et al*. 2021; Shabri 2022). The knowledge of the anticipated outcome will encourage businesses to conduct additional research and look for alternate

methods of production to reduce carbon emissions, as well as assist individuals in understanding the impact of their daily actions that generate carbon emissions. People will be inspired to incorporate carbon emission reduction techniques into their lifestyles as a result of becoming more aware of them.

1.2 Problem Statement

Numerous issues caused by climate change, including those affecting agricultural and industrial productivity, population expansion, healthy development, and migration, have a negative impact on people's ability to live normal lives. The natural environment, which is essential to human life, has been severely damaged by the adverse effects of climate change, some of which are irreparable. CO_2 emissions, on the other hand, are the primary factor in the global warming impact that resulted in dramatic climate change. Researchers are more interested than ever in studying CO_2 emissions as the severity of climate issues has drawn human attention to them (Wen and Yuan 2020).

By making forecasts that are more accurate or by streamlining operations across various industries, AI and DS have enormous potential to cut carbon emissions. For example, DS can be used to forecast severe weather, optimize supply chain and monitor peatlands (Niltop 2022). Because a method's accuracy is essential when making predictions, it is important to select the best model that makes such predictions with the fewest errors. A forecast is an effective programming technique and research into CO2 emissions prediction is very important for making policies to accomplish sustainable growth (Shabani et al. 2021).

1.3 Aim

Carbon emissions have proved toxic to environment and destructive in the UK in recent years. As a result, the proposed project dissertation attempts to use stacking ensemble ML algorithm to develop an enhanced forecasting system for carbon emission level.

1.4 Objectives

1. To increase the forecasting accuracy of carbon emissions by using a stacking ensemble ML technique.

2. To evaluate the predicted carbon emission levels across different sectors.

3. Create a web application for carbon emission forecasting so that stakeholders may easily obtain the predicted future emissions.

1.5 Scope

In view of the problem of climate change and global warming which is a result of the presence of GHG mainly CO_2 in the atmosphere, this research forecasts carbon emission using stacking ensemble machine learning algorithm. The research will cover carbon emission data of the twelve regions in UK ranging from 2005 to 2019. The scope will be limited annual emissions data and analysis of the six carbon emitting sectors will be carried out.

1.6 Justification

This project, which uses the stacking ensemble ML algorithm to predict carbon emission in the UK, was driven by the global issue of climate change as a result of GHG emission and the unsettling impacts in our immediate environment. Stakeholders can determine with a high degree of accuracy what techniques to apply in minimising carbon emissions by developing a predictive model of the behaviour that is likely to occur. This research is expected to add to the knowledge base about the relationship between carbon emissions and climate change.

1.7 Research Question

For this study, the research question to be addressed is:

How efficiently can stacking ensemble ML algorithm perform forecast of carbon emissions in the UK?

1.8 Research Project Specification

The research artefact is a web application, which is a digital by-product of the operations the ML algorithm carried out on the supplied data. To predict the values of unknown carbon emission, the online application will use a trained regression model.

1.8.1 Functionality

The web application has three pages that can be navigated.

Table 1.1: Page function

Pages/ Features	Description
Home	This is the main page. It includes a brief problem statement and feature description.
Visualisation	This has graphs and chart which helps to visualize the data used.
Prediction	Users should be able to enter the numerical values of each of the features and make a prediction of the per capita emission.

1.9 Project implementation

Table 1.2: Project implementation schedule

Tasks	June – July				July - September							
	Wk1	Wk2	Wk3	Wk4	Wk5	Wk6	Wk7	Wk8	Wk9	Wk10	Wk11	Wk12
Identify research area												
Introduction												
Background studies												
Formulate research plan												
pilot study due				8th								
Review of related work												
Methodology												
Data collection												
Data analysis												
Result												
Discussion												
Conclusion												
First draft												
Final draft												
Dissertation due												9th

1.10 Structure of the Dissertation

This dissertation is organised as follows:

Chapter 2 provides a detailed review of previous studies regarding the topic of this dissertation. Categories of carbon emissions research, studies on forecasting carbon emission and reduction strategies are examined.

In chapter 3, explains the methodology and the methods used for each phase of the process.

Chapter 4 provides the implementation process of the methodology described.

In chapter 5, the results of the regression analysis are presented and discussion on the performance of stacking regressor and a reference to benchmark studies.

Chapter 6 presents the conclusions and recommendations for future work.

2. LITERATURE REVIEW

2.1 Introduction

The existing literature on carbon emissions is reviewed in this chapter, along with the models that have been utilized to predict carbon emissions in the past. Finally, the chapter reviews the numerous strategies that have been put out to reduce carbon emissions.

2.2 Carbon Emissions Research Categories

Researchers have categorised carbon emission prediction techniques in a variety of ways. Researchers' studies on CO_2 emissions, according to Wen and Yuan (2020), can be generally categorised into two categories: The first category primarily focuses on research on the factors influencing CO_2 emissions, which has two components: various factors and various industries. The second group is primarily concerned with the analysis and prediction of CO_2 emissions.

The two categories of carbon emissions forecasting components, according to Liu and Zang (2022), are multivariate and univariate forecasting techniques. Multivariate approaches mainly consider many influencing factors that have an impact on carbon emissions. These influencing factors, however, are challenging to predict and may lead to a build-up of errors, producing subpar predicting outputs. Univariate forecasting techniques reduce forecasting ambiguity brought on by model hypotheses and multifactor selection by just using historical and current observation time series data. Additionally, multivariate approaches are better suited for making long-term predictions while univariate methods are for making short-term predictions.

According to findings from earlier studies, a wide range of variables affect CO_2 emissions, and the impact mechanism is rather complex. Due to regional differences, the correlation between CO_2 emissions and its affecting factors varies substantially (Wang, Zhong and Yao 2022). Researchers are primarily concerned with the impact of what they view as critical factors on CO_2 emissions when analysing the factors that influence CO_2 emissions (Wen and Yuan 2020).

Researchers primarily looked at how factors such as energy consumption, GDP, population growth, economic growth, FDI, and urbanisation affect CO_2 emissions.

2.3 Carbon Emission Forecasting

The process of forecasting involves creating predictions based on historical and current data, with trend analysis being the most popular approach. In order to understand the complex relationships between enormous volumes of uncertain data and unpredictable factors, forecasting models are becoming more and more important (Alam and AlArjani 2021).

Numerous prediction techniques, including evolutionary algorithms, grey models (GM), multiple linear regression (MLR), logistic equations, autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA) artificial neural network (ANN), and support vector machines (SVM), have been developed to forecast carbon emissions. However, the magnitude of the training data can have a considerable impact on the forecasting accuracy of artificial intelligence approaches, and statistical methods often need a lot of data that adheres to certain statistical assumptions (Gao *et al*. 2022, Shabri 2022, Ye, Xie and Hu 2021).

Numerous sophisticated models and algorithms have been created in this area, each of which has pros and cons. Using a range of techniques, the following recent research on carbon emission forecasting are presented.

The approaches were classified into two categories by Mason, Duggan, and Howley (2018): statistical techniques and AI techniques. Neural networks and SVM are two AI approaches that offer the benefit of making less hypotheses about the data, making them more adaptable. These are known as non-parametric techniques. Two examples of the statistical methods are ARIMA and linear regression.

Liu and Zang presented a categorization that is comparable but slightly different (2022). They divided carbon emissions forecasting techniques into three categories: statistical models, AI models, and hybrid models. In terms of efficiency and ease of use, statistical models like ARIMA and exponential smoothing models are strong contenders for use in short-term carbon emissions

forecasting. Short-term time series, however, exhibit nonlinear and often changing properties, which goes against the linear assumptions of the statistical model.

2.3.1 Statistical Models

BRICS (Brazil, Russia, India, China, and South Africa) countries' carbon dioxide emissions were studied by Guo *et al.* (2021) using the exponential cumulative grey model (ECGM). Using the particle swarm optimization (PSO) technique, the model found the ideal parameter values by optimising the accumulation method of GM.

In order to estimate CO_2 emissions, Ye, Xie, and Hu (2021) developed a time-delay multivariate grey model (MGM), concentrating on the time delay impact of the influencing factors on CO_2 emissions and addressing the problem of uncertainty with limited sample sizes.

To predict the CO_2 emissions in members of the Asia-Pacific Economic Cooperation (APEC) countries, Qiao, Meng, and Wu (2021) suggested the Discrete Grey Forecasting Model (DGM). Based on the Caputo fractional derivative and the DGM (ß,1) model, they added an accumulation mechanism to the GM.

Using the new information priority concept, Zhou *et al.* (2021) introduced a grey rolling method for analysing the trend and forecasting carbon dioxide emissions in China. The experimental findings reveal that the suggested model exhibits improved modeling and prediction capability as well as a high degree of stability when compared to the other two types of statistical prediction models (the GM and the linear regression) that do not consider the priority of incoming information.

To predict CO_2 emissions in the aviation sector in Shanghai, China, Yang and O'Connell (2020) employed ARIMA. When Holt-Winters and TBATS' forecasting precision were compared to ARIMA's, the comparison revealed that the model is more stable and dependable.

Ding *et al.* (2017) proposed a grey multivariable model combined with the changing trends of driving variables (TDVGM) which is better than the traditional

grey multivariable model that is limited in applying it to real world situations because of its incorrect prediction and poor responsiveness. The model's accuracy in predicting CO_2 emissions from fuel combustion is compared to that of GM, DGM, ARMA, and MLR. Their findings demonstrate that TDVGM outperforms rival models in terms of predicting accuracy.

A MGPM for CO_2 emissions was created by Chiu *et al.* (2020). The proposed model differs from existing MGPMs in several ways, including the feature selection and residual variation, which are thought to increase prediction accuracy. By developing a neural network-based residual model, the estimated values derived from the proposed MGPM are further modified.

2.3.2 AI Models

AI models are superior at eliminating nonlinear series due to their strong nonlinear mapping capabilities. As a result, a variety of AI techniques, such as ANNs and deep learning techniques, have been used to carry out complex forecasting. Despite having excellent forecasting accuracy, AI models are constrained by overfitting and are prone to being stuck in local optima (Liu and Zang 2022). The most popular prediction techniques now are several AI algorithms, which are popular due to their great processing precision and speed (Cui *et al*. 2021).

The long short-term memory LSTM method was developed by Huang, Shen, and Liu (2019) to forecast carbon emissions in China. Principle component analysis (PCA) was performed to identify the four principal components, which decrease the duplication of the input data. The simulation findings demonstrate that LSTM-based carbon emission prediction accuracy is superior to that of back propagation neural network (BPNN) and gaussian process regression (GPR), demonstrating the efficiency of PCA and LSTM in carbon emission prediction.

Forecasts of CO_2 emissions in Gulf nations were compared by Alam and AlArjani (2021). They forecasted CO_2 emissions in the Gulf nations on a yearly basis using the ARIMA, ANN, and holt-Winters exponential smoothing (HWES) forecasting models. The results of the study demonstrate that the ARIMA (1,1,1), Holt-Winters

exponential smoothing (2,1,2), ARIMA (1,1,2), and ARIMA (2,1,2) models are not superior to the artificial neural network model in predicting CO_2 emissions in the Gulf nations.

Based on the enhanced whale algorithm-optimized gradient boosting decision tree, which integrates four optimization techniques and considerably boosts prediction accuracy, Cui *et al*. (2021) created a carbon emission forecast model for China. The effectiveness of the gradient boosting tree prediction model optimised using the enhanced whale algorithm was confirmed using historical data.

Acheampong and Boateng (2019) used ANN to create models for predicting the intensity of carbon emission for Australia, Brazil, China, India, and the United States. For each nation, a feed-forward multi-layer perceptron neural network (FFMLP) was utilised. The networks were trained over a number of iterations using a stochastic gradient descent and backpropagation technique.

To predict carbon emissions in China's Guangdong area, Ren and Long (2021) presented the Fast-Learning Network (FLN) forecasting method, which was optimized by CSO. Three error indications support the CSO-FLN model's superiority (MAE, MAPE, RMSE). The findings demonstrate that the CSO-FLN model's predicting abilities outperform those of FLN and extreme learning machines (ELM).

Shabri (2022) created a model for short-term predicting Malaysia's yearly CO_2 emissions. A nonlinear forecasting model based on time series was built using the Group Method of Data Handling (GMDH) model as one of the Neural Networks, and the Lasso method (Lasso-GMDH) was used to model parameters to increase the GMDH prediction accuracy. Lasso-viability GMDH's usability is confirmed by comparing it with the GM, ANN, and GMDH to predict yearly CO_2 emissions.

Covariance matrix adaptation evolutionary strategy (CMA-ES), an optimization technique, was studied by Mason, Duggan, and Howley (2018) as a method of training neural networks to predict short-term power consumption, power generation, and CO_2 levels in Ireland. They got a result that shows that each component was correctly predicted. On all issues, the models outperformed PSO

and DE, two other evolutionary approaches, in terms of accuracy, speed of convergence, and consistency.

In order to forecast Bahrain's CO_2 emission, Qader *et al*. (2021) used a variety of approaches, including neural network time series nonlinear autoregressive, GPR, and Holt's methods. The findings suggest that the neural network time series nonlinear autoregressive model has done better in this situation.

The enhanced GPR technique for carbon emission forecasting suggested by Fang *et al*. (2018) is based on a modified PSO algorithm that effectively optimises the hyperparameters of the covariance function in the GPR. The effectiveness of the upgraded GPR approach outperformed other conventional forecasting techniques like BPNN and increased the original GPR method's prediction accuracy.

2.3.3 Hybrid Models

It is difficult for the statistical method to obtain high accuracy due to the nonlinear nature of the carbon dioxide emissions data. For improved prediction outcomes, numerous academics started researching intelligent algorithms. More researchers are starting to use metaheuristic algorithms to enhance the algorithms because of the limits of single intelligent algorithms. As a result, many researchers mix two or more models to create a hybrid model (Qiao *et al*. 2020).

According to Liu and Zang (2022), hybrid models have been designed to address the shortcomings of individual forecasting techniques by integrating data pre-processing, optimization, feature selection, and other cutting-edge technologies. No one hybrid model can consistently generate satisfying predicting outcomes since the data properties of different datasets and forecasting processes varies greatly in practical situations. Different forecasting models have different application domains, and hybrid models can combine the benefits of many models (Zhou *et al*. 2021, Jiang *et al*. 2021).

The carbon emissions ensemble forecasting system for (CEEFS) that Liu *et al*. (2022) proposed was used to achieve both point and interval predictions which

consists of data decomposition, phase space reconstruction, model selection, ensemble point prediction, and interval prediction. It has been shown to be successful in improving the accuracy and strength of carbon emissions forecasting. For the reconstructed series prediction, the ARIMA, BPNN, ELM, ENN, DBN, GRU, and LSTM models were used. When compared to other comparison models based on four distinct multi-objective optimization methods, CEEFS produced good quality predicting results. In order to evaluate the final forecasting performance, the coverage width-based criteria (CWC) value is utilised to optimise the forecasting interval's reduction factors.

An intelligent technique for CO_2 emission forecasting in Iran, Canada, and Italy was proposed by Heydari *et al.* (2019) and is created with Generalized Regression Neural Network and Grey Wolf Optimization (GRNN-WO). The findings show that the suggested technique is more accurate than existing single and combined algorithms, such as multi-layer perceptron (MLP), radial basis function neural networks, GRNN-PSO, and GRNN-TS, in terms of short-term renewable energy production and long-term CO2 emission predictions.

Li (2020) proposed KLS, a merger of the Kalman filter (KF), LSTM, and SVM, so as to include time series prediction and regression as well as to address the shortcomings of SVM and LSTM. In order to predict carbon emission as a time series, LSTM was used. This method can address the issues of fake prediction and errors. Next, ridge regression (RR) was used to select variables for SVM to regress, which can address the issue of poor analysis. KF was then used to combine the LSTM and SVM results depending on their variances. According to the results, KLS was more accurate than the other four methods (GM, nonlinear MGM, ARIMA, Kernel-based nonlinear MGM, and BPNN).

Wen and Yuan (2020) employed a BPNN prediction model based on the random forest (RF) and double enhanced particle swarm optimization (DPSO) hybrid prediction models, which adopted measurable methodologies to choose prediction indicators, to study the CO_2 emissions of China's commercial sector. The RF-DPSO-prediction BP's results indicate that it has a lower prediction error than four other models and highlights the significance of RF in enhancing prediction accuracy.

Based on PCA and PSO optimised least squares support vector machine (PSO-LSSVM), Sun, Jin, and Wang (2019) suggested a CO_2 emission forecast model for Hebei Province of China. The influencing factor's dimensions were reduced using PCA, the parameters were solved using PSO, and CO_2 emission predictions were made using LSSVM. When compared to BPNN and LSSVM, the method maintained the capacity to search for the global optimum, enhanced PSO convergence, and had a reduced prediction error.

Wen and Cao (2020) predicted residential energy-related CO_2 emissions in Shanghai, China using the improved chicken swarm optimization (ICSO) algorithm to improve the parameters of SVM to get a hybrid model ICSO-SVM. In order to eliminate data redundancy, PCA extracts four basic components as the SVM's predictive input data. The results demonstrate that ICSO-SVM model performs better than the original CSO-SVM, PSO-SVM, GA-SVM and basic SVM.

In order to forecast carbon emissions under various scenarios, Ma, Jiang, and Jiang (2020) suggested a hybrid multivariate grey model optimized using firefly algorithm (FA-GM). The association rule algorithm was used to identify predictive factors and examine their combined impact on carbon emissions from a time and space viewpoint. FA-GM (1, N) has the best forecasting power in terms of prediction precision and bias when compared to the traditional GM (1, N) model, SVM, multiple regression, ARMAX, GM (1,1), ES, and ARIMA.

In order to improve the conventional LSSVM model with an emphasis on model stability, Qiao *et al*. (2020) proposed a hybrid method that combines the lion swarm optimizer and genetic algorithm. The new algorithm's performance test reveals that it has improved stability and accuracy. The innovative hybrid algorithm has better global optimization ability, faster computational efficiency, better accuracy, and a medium computation speed, as shown by comparison of the forecasting results of the hybrid method with the other eight methods.

Wang, Zhong and Yao (2022) proposed a high-performance hybrid intelligent algorithm model appropriate for China's CO_2 emissions prediction and associated socioeconomic indicator data. The hyperparameters of Least Squares Support Vector Regression (LSSVR) are optimized by the Adaptive Artificial Bee Colony (AABC) algorithm to develop the model. The study's findings demonstrate that the

hybrid algorithm model has higher accuracy and more robustness, with relative error within 5% of the predicted CO_2 emissions.

For predicting the carbon-dioxide emissions in India, Amarpuri *et al.* (2019) adopted a deep learning hybrid model of Convolution Neural Network and Long Short-Term Memory Network (CNN-LSTM). When compared to exponential smoothing, the CNN-LSTM results achieved predictions that were more accurate.

By combining self-organizing map (SOM), singular value decomposition (SVD), artificial neural network (ANN), and adaptive neuro-fuzzy inference system (ANFIS) approaches, Mardani *et al.* (2020) created a hybrid method for forecasting CO_2 emissions. The data were clustered using the SOM clustering algorithm, missing value detection using SVD, and CO_2 emission prediction using ANN and ANFIS based on economic development and energy consumption across Group 20 countries.

A new Inclusive Multiple Model (IMM) was developed by Shabani *et al.* in (2021) to forecast CO_2 emissions in Iran's agricultural sector. To implement the IMM model, the optimal parameters of MLR, GPR and ANN are used as input to another ANN model. When the model was compared with other models under consideration, its high accuracy was validated.

2.4 Carbon Emissions Reduction Strategies

An organisation can lessen their carbon footprint by using IoT sensors in the environment to keep track of carbon emissions, waste production, and energy use, and by processing raw and unstructured data from renewable resources like wind turbines to produce real-time actionable insight (Schoklitsch 2019).

Cossutta, Foo and Tan (2021) suggested integrating low carbon grid with transition from internal combustion to electric motor vehicles (EV) can significantly reduce GHG emissions. Additionally, carbon capture and storage (CCS) and negative emissions technologies (NETs) are used to set off industry's positive emissions. In response to the carbon budgets, the UK power sector is changing the energy generation to less carbon emitting sources, increasing production of renewable

energy and transferring fuel subsidies to low-carbon fuels. Implementing these processes will likelihood of phasing out coal by 2025.

In order to minimise greenhouse gas emissions, international treaties and financial incentive schemes take into account the ability of terrestrial environments and coastal habitat, such as tropical forests to store carbon (Luisetti *et al*. 2019).

Deforestation reduces the number of trees available to absorb carbon dioxide, which results in the release of the carbon that the trees formerly stored back into the atmosphere. More trees can be planted in the UK to lower carbon emissions. Bamboo can be planted because of its quick growth cycle, year-round greenness, and significant CO_2 absorption to slow down global warming. It can also serve a low-carbon substitute for paper and plastic (Borowski, Patuk and Bandala 2022).

The long-term effects of a thoughtfully created climate change education on long-term attitude and behavior were examined by Cordero, Centeno, and Todd (2020). Students' decisions about food, waste, home energy, transportation, and trash all show a significant shift in behavior that was linked to the course. Students' carbon emissions dropped by 3.54 tons/year, compared with an average California resident’s emission of 25.1 ton/year. Additionally, they emphasized that if reductions like those attained in the course could be attained in other classrooms, using education as a climate change mitigation approach would be beneficial and consistent with other mitigation methods.

The UK climate change committee wants to implement technologies that will be requires to meet the 2050 target such as decarbonizing energy generation, reducing energy use in buildings and industries, reducing domestic transport emissions (electric and plug-in hybrid vehicles) and reducing emissions from international aviation and shipping (Stark, Thompson and CCC 2019).

Table 2.1: Recent studies in carbon emission reduction strategies

Author	Country/ Sector	Reduction strategy
Li and Umar 2021	China	Investment in green project reduce short- and long-term carbon emission levels
Langevin and Reyna 2019	US/ Buildings	CO_2 emissions may be reduced by 72%–78% by robust efficiency measures, electrification, and substantial renewable energy penetration.
Ren and Ou 2021	China/ Iron and steel industry	Renewable energy to produce zero carbon electricity and ultra-low carbon technologies (carbon capture, use and storage strategies or hydrogen-based technologies) can reduce CO_2 emissions by 80%-95%
Hao and Ali 2021	G7 countries	Green growth, environmental tax, renewable energy and human capital were found to reduce CO_2 emissions
Akram and Majeed 2020	66 Developing countries	Energy efficiency and renewable energy reduce CO_2 emissions across all quantiles

2.5 Contribution

Numerous studies have examined carbon emissions from a number of perspectives, including the use of various study periods, forecasting techniques, and influencing factors.

Stacking ensemble ML methods have been extensively used to address regression and classification issues in different fields. such as Warfarin dose estimation (Ma

et al. 2018), solar irradiance prediction (Al-Hajj, Assi and Fouad 2019), cheating detection (Zhou and Jiao 2022), crime prediction (Kshatri *et al*. 2021), network intrusion detection (Rajagopal, Kundapur and Hareesha 2020), power consumption anomaly detection (Ouyang *et al*. 2018), short-term load forecasting (Massoaudi *et al*. 2021), hemoglobin estimation (Acharya *et al*. 2019).

According to a thorough examination of the literature, stacking ensemble ML algorithms have never been used to solve the problem of predicting carbon emissions in the UK. In order to close the abovementioned research gap, the stacking ensemble model method is proposed. It combines ML algorithms in the best possible way to predict carbon emissions.

3. METHODOLOGY

3.1 Introduction

This chapter gives an general idea of the research methods that were followed in this study The research design that was selected for this study's objectives was described, along with the justifications for the selection. The methods that were used to collect the data, analyse the data and implement the project will be discussed.

3.2 Research Design

The Cross-Industry Standard Process for Data Mining was used as the quantitative research model for the purposes of this study (CRISP-DM). CRISP-DM, an industry-driven approach, has been the de facto industry standard method outline for data mining since it was first launched in 1999. This technique offers a consistent framework for project planning and management, and because it is cross-industry standard, CRISP-DM may be applied to any DS project, regardless of field. (Martinez-Plumed *et al*. 2019).

As illustrated in figure 3.1, the CRISP-DM model consists of six phases, with arrows designating the most significant and common relationships between phases. There is no set order to the stages. In actuality, projects frequently go back and forth between stages as needed. The CRISP-DM paradigm is adaptable and simple to adjust (Schroer, Kruse and Gomez 2020).

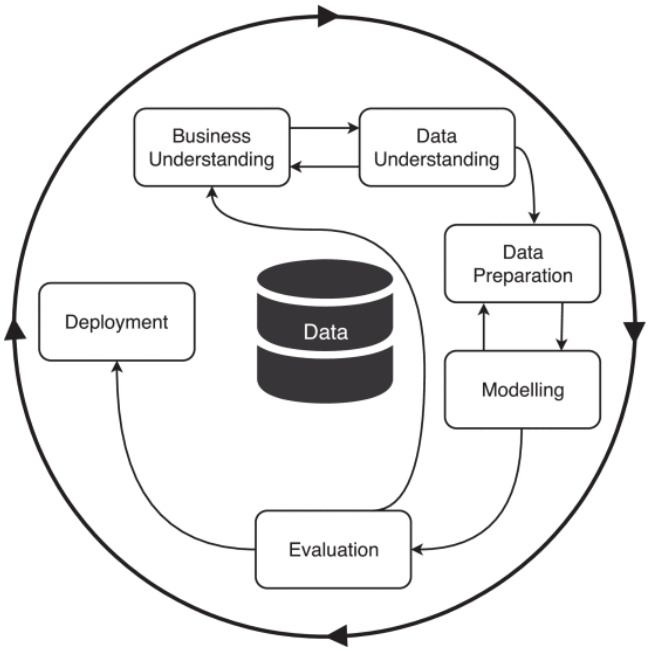


Figure 3.1: CRISP-DM model (adapted from Martinez-Plumed et al. 2019)

Business understanding, data understanding, data preparation, modelling, evaluation and deployment are the main phases of CRIPS-DM.

3.3 Business understanding

It is the first and most important phase of the study in which the research objectives are stated and emphasized and following that, a business strategy is created to accomplish those objectives.

People are collaborating to identify various solutions that can aid in enhancing the environmental situation as part of the fight against climate change, which has already begun. Adopting sustainable living practises is an urgent necessity to solve the problem of global warming. With 1.1% of the world's carbon emissions, the UK was rated 17th (Bolton 2021). Although it may seem gradual, climate change's consequences are becoming increasingly obvious in our environment, and mitigation measures are required. Because carbon emissions everywhere constitute a danger to world growth, every country must take action (Mott, Razo and Hamwey, 2021).

About 80% of the UK's GHG emissions in 2019 were from carbon dioxide, which is the primary GHG. The importance of local governments and regional authority in advancing energy efficiency and lowering carbon dioxide emissions has grown in recent years.

The business plan formulated was to use excel and python for data collection and cleaning, Tableau, matplotlib and seaborn for visualization and five machine learning models for predicting the carbon emissions and choose the best model with the highest predicting accuracy to do future predictions. Then streamlit will be used for deployment and implementation.

3.4 Data understanding

This is the second phase of the CRISP-DM which begins with data collecting, then proceeds to actions to analyse the data, discover data quality problems and acquire initial data insights. During this stage, the analyst may also notice noteworthy subdivisions that may be used to provide ideas for hidden patterns.

3.5 Data Preparation

The third phase of CRISP-DM includes activities required to transform the raw data into a final dataset that can be used as input to the models. All changes were made on a duplicate dataset, leaving the original intact. To ensure compatibility with all algorithms that will be used, data cleaning, data recoding data sorting based on selected variables.

3.6 Modeling

The selection, configuration, and testing of various algorithms as well as the development of the models are tasks that fall within the fourth phase of CRISP-DM. Regression models were selected because the variable to be predicted is a continuous value and thus involves regression analysis (RA). The selected machine

learning algorithms are used on the data to give the desired output. RA is a method for determining the relationship between a dependent variable and one or more independent variables. This strategy may be used to determine the optimal selection of variables for constructing a predictive model (Overbeek *et al*. 2020).

Python and Jupyter Notebook were used in PyCharm to create the model-building code. The libraries used by all the models are scikit-learn or sklearn, pandas, and matplotlib library. In this research, ML regression algorithms such as Random Forest Regressor (RFR), Gradient Boosting Regressor (GBR), K-Nearest Neighbor Regressor (KNNR), Support Vector Regressor (SVR) and Stacking Regressor (SR) are examined.

3.6.1 Random Forest Regressor (RFR)

Breiman first presented the random forest (RF) machine learning technique in 2001. With a voting mechanism and a predetermined number of decision trees, RF enhances learning performance as an ensemble approach. RF displays out-of-bag error estimates, bootstrap sampling, and complete depth decision tree development properties. These characteristics make random forest appropriate for predicting carbon emissions. Following the input of data samples, the RF model initially extracts certain samples using bootstrap sampling before randomly choosing the features of these samples.

With these two rounds of random sampling, RF is less susceptible to overfitting and is more forgiving to noise and outliers (Meng and song 2020). Several classification decision trees are fitted using RFR on various subsamples of the dataset, and the mean or average forecast of each tree is returned. By doing this, overfitting is reduced, and accuracy is increased. The technique produces trees with high variance and little bias by guaranteeing the forest expands up to a user-defined number of trees (Torre-Tojal *et al*. 2022).

3.6.2 Gradient Boosting Regressor (GBR)

Friedman introduced the gradient boosting tree, which primarily resolves the problem of generic loss function optimization. The fundamental concept is to use the loss function's negative gradient to fit the residuals of the last round of base learners, resulting in a steadily decreasing residual estimate for each round. As a result, the output of the base learner's rounds steadily approaches the true value. By ensuring that the loss function lowers as rapidly as possible in each round of training, fitting in the direction of the negative gradient speeds up convergence to a local or global optimal solution (Cui *et al*. 2021).

This estimator accepts the improvement of any discrete loss function and builds an additive model in a stage-by-stage fashion. GBRs are additive models that predict $\hat{y}_i$ in the form shown below for an input $\boldsymbol{x_i}$:

$$\hat{y}_i = F_M(x_i) = \sum_{m=1}^{M} h_m(x_i) \qquad (3.1)$$

Where in the perspective of boosting, the h_m are estimators referred to as weak learners. Decision tree regressors of a fixed size are used as weak learners in gradient tree boosting. The parameter n_estimators' and the constant M are equivalent (Pedregosa *et al*. 2011).

3.6.3 K-Nearest Neighbor Regressor (KNNR)

The K-nearest neighbour (KNN) method estimates similarity-based new data point values in feature sets. This implies a value is allocated to the new point based on how much it fits the points in the training set. When dealing with a regression problem, the ultimate prediction in a regression problem is computed using the average of the data. The k data points (5 by default) that are the closest are chosen after calculating the distance between the new data point and each training point in Euclidean space. As the new data point's ultimate predicted value, the average of these datapoints is determined. The formula below is used to determine the Euclidean distance (Sandhya and Padyana 2021).

$$d(x, y) = \sqrt{\sum(x - y)^2} \quad (3.2)$$

In order to predict numerical outcomes based on similarity metrics, the KNN algorithm keeps track of all training realizations. The simplest form of KNN regression is averaging the numerical target of the K nearest neighbors across all training data (Al-Hajj, Assi and Fouad 2019).

3.6.4 Support Vector Regressor (SVR)

The SVR applies the SVM's underlying idea to regression issues. SVM is a model that categorizes the data points in N-dimensional space and generalises by attempting to identify the best hyper plane. Support data vectors, which are the most effective data components for finding appropriate regression, are used in SVR. SVR are often trained with symmetrical loss functions and penalize high and low wrong values depending on the support vectors (Al-Hajj, Assi and Fouad 2019).

SVR tests to find the optimal line within a limit or predetermined error value. It achieves this by dividing the predicted lines into those that cross through the error border and those that do not. Because the difference between the real and predicted value is greater than the error threshold, or epsilon value, the lines that do not cross the error barrier are not considered. The lines that pass are considered as prospective support vectors to aid in estimating the unknown values (Sandhya and Padyana 2021).

SVR uses the kernel function is used to provide a nonlinear mapping of the input data in order to calculate the regression function linearly in feature space, which is defined as

$$f(x) = w^T \varphi(x) + b \quad (3.3)$$

Where x is the input data, $f(x)$ denotes the regression function, $\varphi(x)$ is the transformation into feature space and w and b are coefficients (Nguyen *et al.* 2021).

3.6.5 Stacking Regressor (SR)

Stacked generalisation was first proposed by Wolpert in 1992 and combines several prediction algorithms in a two-level architecture. It uses a trainable combiner to get the best prediction accuracy and is commonly referred to as a meta-learner. Stacked generalisation is implemented in two steps: The first phase is to generate first-level prediction results using base-learners; the second step is to generate second-level predictions using a meta-learner algorithm on the stack of first-level predictions (Nguyen, Diaz-Rainey and Karuppuarachchi 2021).

In contrast to the "bagging" and "boosting" techniques, that can only stack ML algorithms that are the same, the stacking model can stack different types of algorithms using a meta model to optimise efficiency (Wang 2018).

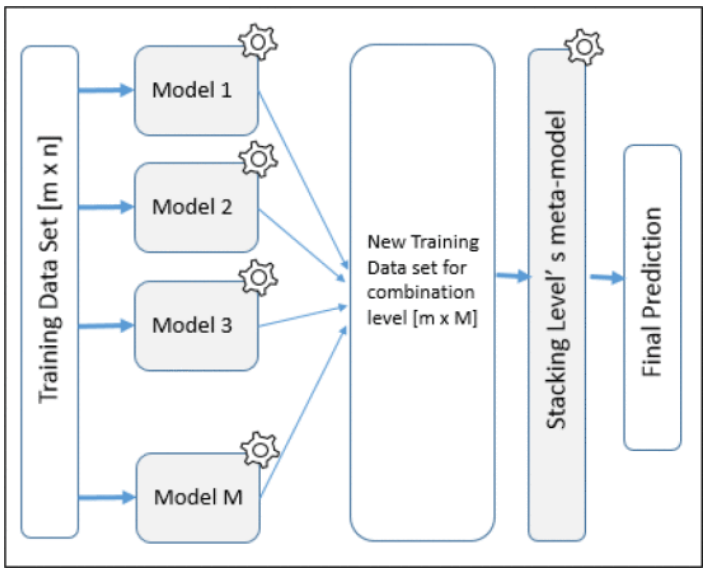


Figure 3.2: Parallel ensemble structure with M models (adapted from Al-Hajj, Assi and Fouad 2019)

3.6.6 Hyperparameter Tuning

To improve the algorithms and provide better predictions while utilizing GridSearchCV, the idea of hyperparameter tuning was applied. In order to select the best hyperparameters for our training data, the GridSearchCV function in Python was used. A function called GridSearchCV it takes a list of possible values

for each hyperparameter, then fits the training data to each possible combination and chooses the model with the highest score (Nguyen *et al*. 2021).

1.7 Evaluation

The goal of the fifth phase of CRISP-DM is to validate any models or information that have been acquired to make sure they satisfy the goals set out at the start of the process. At this phase, the created models' accuracy and robustness are being evaluated. To estimate the efficiency of the models used in this study, the following four performance measures were implemented: R^2 (Coefficient of determination), RMSE (Root Mean square Error), MAE (Mean Absolute Error), and MAPE (Mean Absolute Percentage Error).

In the following formula, n is the total number of data points, Y_1, Y_2 , …, Y_n is the true value, $\hat{Y}_1$, $\hat{Y}_2$, …, $\hat{Y}_n$ is the predicted value and $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_n$ is the mean of the true values.

3.7.1 R^2

This is the coefficient of correlation, which is sometimes referred to as a measure of a regression model's goodness of fit. It measures how well the prediction fits the observation by measuring the ratio of the dependent variable's variation that can be predicted from the independent variables. R^2 scores range from -∞ to 1. The regression model performs better as R2is approaching 1. If R^2 is equal to 0, the model is not outperforming a random model and the regression model is inaccurate if R^2 is negative (Chicco, Warrens and Jurman 2021).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2} \tag{3.4}$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{3.5}$$

3.7.2 Root Mean Square Error (RMSE)

It is the average root-squared deviation between the actual and expected values. It is easier to execute mathematical operations since the loss function stated in terms of RMSE is easily differentiable and is therefore the default metric for many models. Prior to obtaining the averages, RMSE squares the errors. That results in greater penalties for large errors (Chicco, Warrens and Jurman 2021). It performs remarkably well when large errors are unwanted for your model's performance. This shows the variance between the observed and predicted values. Model performance improves with decreasing RMSE values. 0 and +∞ are the greatest and worst values, respectively (Shabani *et al*. 2021).

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2} \tag{3.6}$$

3.7.3 Mean Absolute Error (MAE)

This is the mean of the absolute differences between the model's predicted value and the actual value. The predicted variable unit and the MAE unit are the same. As a result, the MAE cannot evaluate the effectiveness of regression models for various data types. It is robust to outliers. The ideal value is 0 and the worst value is +∞, meaning that the larger the MAE, the more significant the error is (Chicco, Warrens and Jurman 2021).

$$\frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right| \tag{3.7}$$

3.7.4 Mean Absolute Percentage Error (MAPE)

This is the model's predicted value divided by the real value, with the result being the average absolute difference between the real value and the model's predicted value. It enables for comparison across regression models made for different types

of data, similar to how MAE is used but with a percentage difference. The focus on outliers is not narrowed down. With a relatively insightful interpretation in terms of relative inaccuracy, its usage is advised for activities where it is more crucial to be attentive to relative changes than to absolute variations. 0 and +∞ are the greatest and worst values, respectively (Chicco, Warrens and Jurman 2021).

$$\frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i-\hat{Y}_i}{Y_i}\right| \qquad (3.8)$$

1.8 Deployment

This is the process's last phase. It involves applying created models for end users, with the goal of using modelling to organise knowledge so that it may be included into the decision-making process.

Streamlit: This open-source app framework is specifically utilised in DS and ML applications. The code in streamlit is simple to implement and integrate with other applications. No coding of a backend, definition of pathways, administration of HTTP requests, connection to a frontend, or creation of HTML, CSS, or JavaScript is required (Shukla, Maheshwari and Johri 2021). Streamlit was used in conjunction with several Python libraries including Pandas, NumPy, Matplotlib, Joblib, and seaborn offered on PyCharm (Python Integrated Development Environment).

3.9 Conclusion

The justification for selecting CRISP-DM as the methodology for the research project is presented along with a brief discussion of the six CRISP-DM stages and their relevance to the study.

4. IMPLEMENTATION

4.1 Introduction

The model implementation process used to accomplish the research project's goals is covered in this chapter. Figure 4 illustrates the implementation flow diagram followed in the project to realize the future prediction results.

The aim of the project is to predict carbon emissions in the UK and identify the regions and industries with rising carbon emissions. When this information is made public, it can aid in the application of policies to reduce emissions and fulfil the legally binding goal set by the UK government's Climate Change Act to achieve net zeros emissions across UK economy by 2050 (UK government 2019).

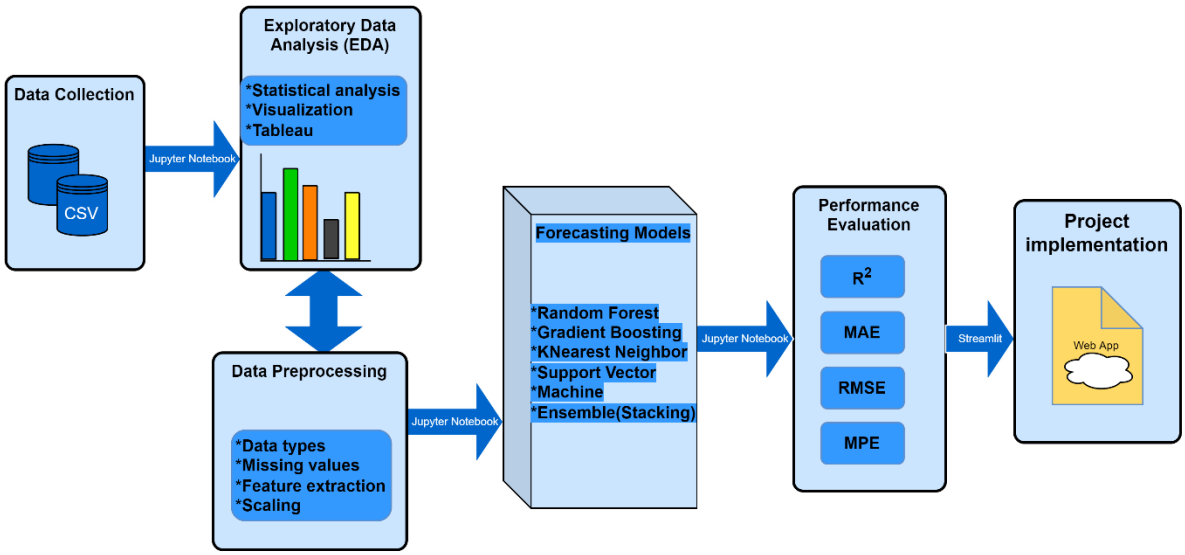


Figure 4.1: Implementation flow diagram

4.2 Data Collection

The dataset for the study project was obtained from the UK government website (gov.uk), which has made the dataset publicly available for reuse. This National Statistics publication contains the most recent data of territorial carbon dioxide (CO_2) emissions for Local Authority (LA) areas for 2005-2019, as well as a report that explains the basis for the estimates, summarizes the key findings, and explains some of the challenges to think about while using the data. The metric

for per capita emissions is tonnes (t), the population is thousands, and the area is square kilometers (km^2).

Table 4.1: Data variables

Number	Variable Name	Variable Description	Variable Type
1	Region/Country	12 regions in UK	Categorical
2	Second Tier Authority	District/ metropolitan area	Categorical
3	Local Authority	County council	Categorical
4	Code	Local authority code	Categorical
5	Year	Year of collecting emission (15)	Numerical
6	Industry Total	Total emissions in the industry sector	Numerical
7	Commercial Total	Total emissions in the commercial sector	Numerical
8	Public Sector Total	Total emissions in the public sector	Numerical
9	Domestic Total	Total emissions in the domestic sector	Numerical
10	Transport Total	Total emissions in the transport sector	Numerical
11	LULUCF Net Emissions	Total emissions in the land use and land use change sector	Numerical
12	Grand Total	Total emissions of all sectors	Numerical
13	Population	Population per region	Numerical
14	Per_Capita_Emission (t)	Emissions per person	Numerical (Dependent)
15	Area(km^2)	Area covered in each region for emission collection	Numerical

16	Emissions_Per_km^2 (kt)	Total emissions for per square km	Numerical

Table 4.2: Carbon emitting sectors and subsectors

Sectors	Subsectors
Industry	Industry electricity, industry gas, industry other fuels, large industrial installations, agriculture
Commercial	Commercial electricity, commercial gas, commercial other fuels
Domestic	Domestic electricity, domestic gas, domestic other fuels
Public sector	Public sector electricity, public sector gas, public sector other fuels
Transport	Road transport (A roads), Road transport (motorways), Road transport (minor roads), diesel railways
LULUCF	Net Emissions: Forest land, Net Emissions: cropland, Net Emissions: grassland, Net Emissions: wetlands, Net Emissions: settlements, Net Emissions: harvested wood products

4.3 Exploratory Data Analysis

Exploratory data analysis (EDA) is an investigatory technique that uses summary statistics (numerical tools) and visualization techniques (graphical tools) to get to know data, discover patterns, identify potential relationships between variables, and understand what you can learn from them (Komorowski *et al*. 2016). The EDA

for this research project was divided into two components. The first part was done in Microsoft Excel using raw data.

The raw data gives annual CO_2 emissions in kilo tonnes of carbon dioxide from 2005 to 2019 from six sectors and twelve regions. The dataset represents territorial emissions i.e., emissions that occur inside the border of the UK. The dataset has 6345 rows and 41 columns that include the NAN values.

After thorough examination, it was found that the raw data had been stored as an excel file, making it challenging to use in a Jupyter notebook. The data was subjected to various filters, column grouping, and conditional formatting. There were several rows that lacked a code. It was discovered that some emissions are not assigned to any one region.

The platform used for coding was PyCharm, which incorporates Python and Jupyter notebook. The second part of the EDA was done in python. Pandas library was used for the following:

1. To load the dataset, the necessary libraries for analysis and modelling were imported. The head() method was used to view the data head.

```
#load your data into the dataframe
my_data = read_csv('COM726.csv')
```

Figure 4.2: Loading data

1. Using the function shape(), the data's shape was verified to determine the number of rows and columns.
2. The data description was used to visualize its statistics such as the total count, standard deviation, minimum, maximum and the percentiles using the function describe().

	Region_Country	Second_Tier_Authority	Local_Authority	Code	Year	Industry_Total	Commercial_Total	Public_Sector_Total	Domestic_Total	Transport_
count	5685	5685	5685	5685	5685.000000	5685.000000	5685.000000	5685.000000	5685.000000	568
unique	12	151	379	379	NaN	NaN	NaN	NaN	NaN	
top	South East	Scotland	Darlington	E06000005	NaN	NaN	NaN	NaN	NaN	
freq	960	480	15	15	NaN	NaN	NaN	NaN	NaN	
mean	NaN	NaN	NaN	NaN	2012.000000	297.525928	141.302674	50.575374	332.576640	33
std	NaN	NaN	NaN	NaN	4.320874	697.539149	150.151810	54.754418	220.761836	22
min	NaN	NaN	NaN	NaN	2005.000000	0.800000	1.600000	0.200000	2.800000	
25%	NaN	NaN	NaN	NaN	2008.000000	85.200000	61.800000	19.800000	199.000000	18
50%	NaN	NaN	NaN	NaN	2012.000000	154.700000	98.500000	33.400000	271.500000	29
75%	NaN	NaN	NaN	NaN	2016.000000	276.800000	163.300000	58.800000	391.700000	42
max	NaN	NaN	NaN	NaN	2019.000000	10076.700000	1942.900000	534.000000	2292.400000	184

Figure 4.3: Variable statistics

3. The data information was checked to know the column names and the total number, non-null count and the data types. This enabled us to check if the column names need to be changed and to see if the data types are appropriate for this research using the function info().
4. Missing values were checked for in the data. Missing values occur when no information is provided for one or more items in the data. The missing values were represented with Nan and isna() and isnull() functions in Pandas were used to check for the missing values.
5. We checked for duplicates and unique values in the data using duplicated() function and unique () function in pandas respectively.
6. The data anomaly i.e., skewness and the kurtosis of the data was checked using kurtosis() and skew() function respectively.

4.3.1 Univariate analysis

It is the simplest type of analysis, because the data only pertains to one variable It is not interested in causes or relationships, and its primary goal is to interpret the data and uncover patterns within it (Sial *et al*. 2021). For each variable, a histogram was plotted to check the quality of the data and determine if it was

skewed. Box plots were used to look for outliers that might affect prediction accuracy.

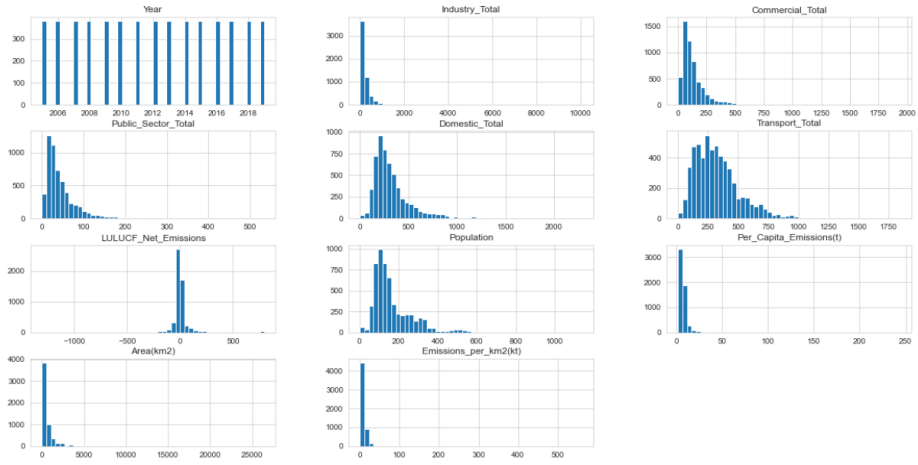


Figure 4.4: Histogram plot for data distribution

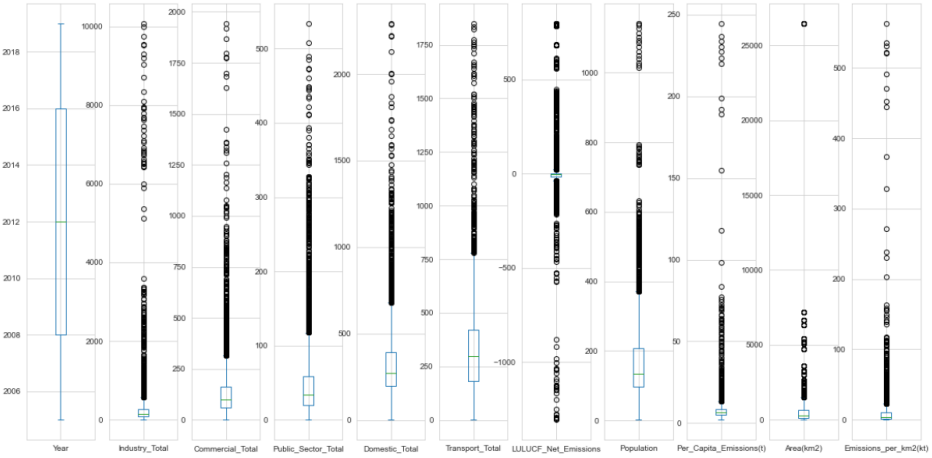


Figure 4.5: Box plot for outliers in each variable

Each variable's distribution and skewness are displayed in Figure 4.4. To be able to make an accurate forecast, the majority of the variables must be normalized because they do not all have a gaussian distribution.

Visualization using box plots is the method used to spot outliers, as seen in figure 4.5. The outliers are the data points that are outside the boxplot's whiskers. Figure 4.5 shows that the area has the least outliers and the LULUCF has the most.

4.2 Bi-Variate analysis

This form of study investigates two variables at the same time. This type of data analysis is involved in causes and relationships, and the analysis is carried out to discover the connection between the two variables (Sial *et al*. 2021). A box plot was used to investigate the relationship between Per Capita Emission and Region Country, while a line plot was used to assess the relationship between Per Capita Emission and Year.

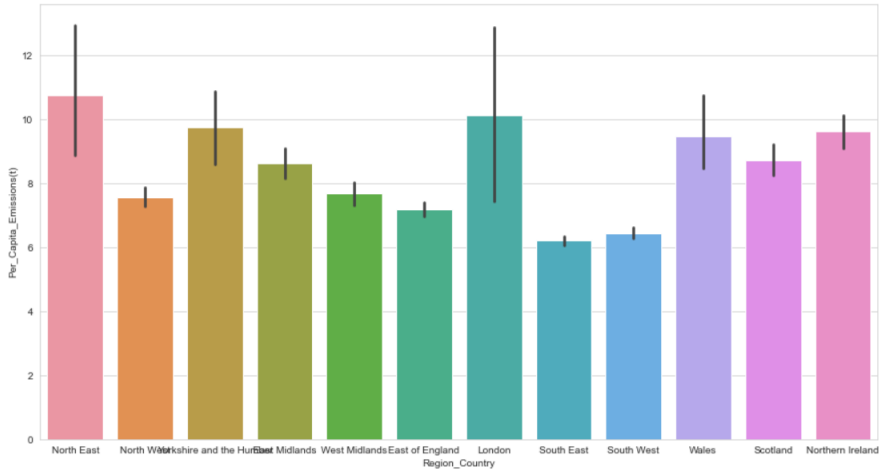


Figure 4.6: Bar plot of regions by per capita emissions

Figure 4.6 shows that North East region is the leading emitter per head followed by London, Yorkshire and the Humber and Wales while South East has the lowest per capita emissions.

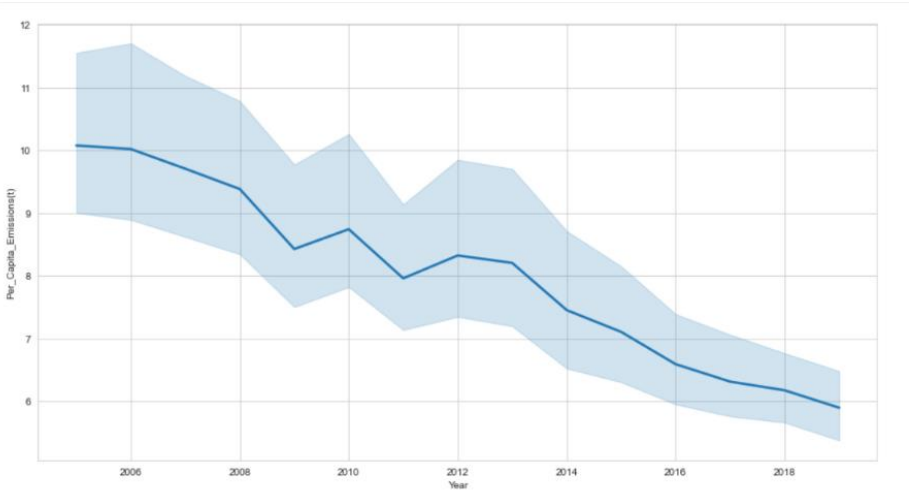


Figure 4.7: Per capita emission from 2005 to 2019

Figure 4.7 show that the per capita emissions in the UK exhibit a decreasing trend from 2.1(t) in 2005 to 244.6(t) in 2019. There was a slight increase and decrease between 2009 and 2011.UK can achieve its net zero carbon emissions target she follows her carbon budget as This is represented in overall necessary carbon reductions of 46% from 2009 to 2030 (CCC 2010), and the per capita emission continues to decrease at this rate.

4.3.3 Multivariate analysis

This sort of analysis examines three or more variables at once and helps to summarize the data (Sial *et al*. 2021). The correlation among the variables was carried out using the corr() function in Pandas and visualized with seaborn heatmap. Other multivariate analyses were carried out in tableau.

The data needs to be cleaned by changing some data types, removing the missing values, remove the outliers, normalizing the data and extracting the relevant features.

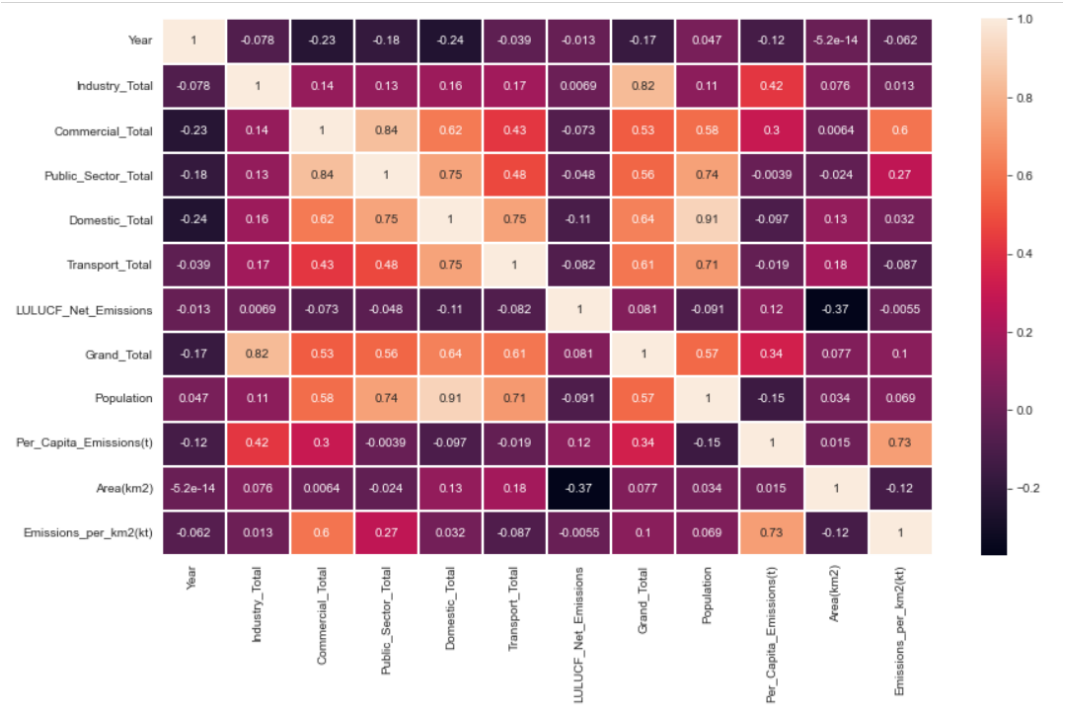


Figure 4.8: Variable correlation heatmap

4.4 Data Cleaning

The data was downloaded as an excel file. All the rows that were saved as unallocated in the region/ country column were removed. The grouping was removed for each sector and the subsectors were removed leaving only the total for each sector. All the frozen panes were disabled, and the conditional formatting removed. The rows for title, colour code, and region-total were deleted. Some column names were changed before the cleaned dataset was exported as a CSV file to be used in Jupyter notebook for further analysis and modelling.

4.5 Data Pre-processing

The data needs to be preprocessed to remove impurities. The following preprocessing steps were carried out:

1. The Year column was changed from float 64 datatype to integer 64.
2. Missing data needs to be filled or removed because most machine learning algorithms will give an error if you pass a Nan value to it (Harrison 2019). The column with missing values was removed with drop() Pandas function and the rows with missing values were removed with dropna() Pandas function.
3. Outliers are data points that differ from the rest of the data in the dataset. After visualizing the boxplot which shows the presence of outliers, the outliers were removed with an automatic outlier remover, isolation forest (Staerman *et al*. 2019) and Quantile-based Flooring and Capping (Roy, Gosh and Goswami 2021). The flooring was done for the lower values with 1st percentile and capping for the higher values with 99th percentile.
4. From the histogram plot and data anomaly check i.e., skewness and kurtosis (Cain, Zhang and Yuan 2017), the data needed to be normalized to get a standard distribution to improve our model's accuracy. The normalization was carried out with RobustScaler() function of sklearn because it is an estimation that is robust to outliers (Kwak and Kim 2017).

```
scaler = RobustScaler().fit(X_train)
train_sc = scaler.transform(X_train)

test_sc = scaler.transform(X_test)
```

Figure 4.9: Robust scaler

5. Feature selection was implemented using the recursive feature elimination which is a wrapper method and random forest regressor feature importance which is an embedded method because it has its own in-built mechanism (Gnana, Balamurugan and Leavline 2016). It used the model's accuracy to identify which attributes contribute much more to the target attribute prediction.

```
#recursive feature elimination
rfe = RFE(regress_model, n_features_to_select=5, step=1)
rfe = rfe.fit(X, y)

print("Number of features", rfe.n_features_)
print("Selected Features", rfe.support_)
print("Feature Ranking", rfe.ranking_)
```

Figure 4.10: Recursive feature elimination

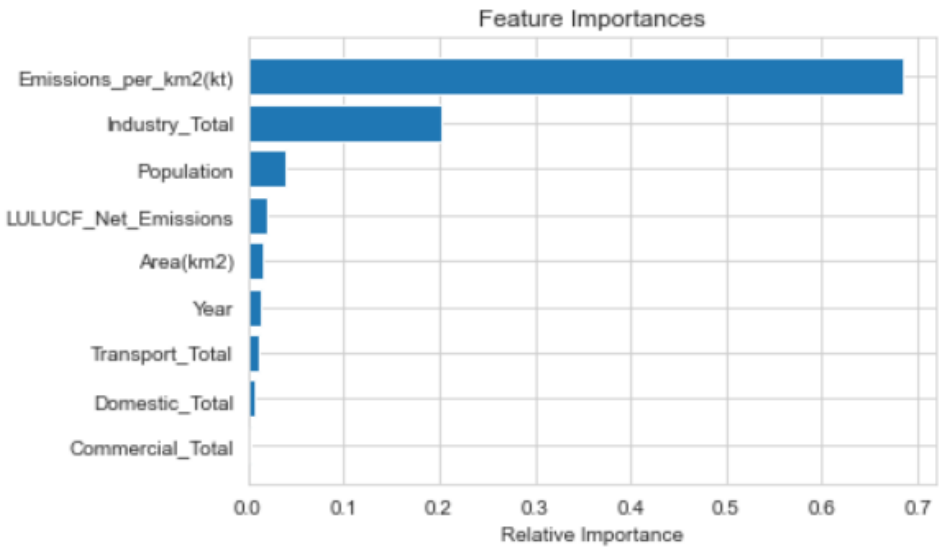


Figure 4.11: Feature importance using random forest regressor

4.6 Model Construction and evaluation

The train test split function of sklearn was used to split the data into train and test split. These machine learning models are trained on 75% of data to learn the patterns in the data and evaluated on the remaining 25% of the data.

Single RA algorithms used include RFR, GBR, KNNR and SVR. Their best hyperparameters were selected with GridSearchCV as shown in Table 4.3. After selecting the best hyperparameters, the models were fit on the training data and predictions were made using the test data.

For SR model, the first level models (RFR, KNNR and SVR) result were used as inputs (meta-features) of the second level model GBR which is then trained to generate the final results that will be used to make predictions. Compared to other algorithms, GBR is excellent for a meta-learner since it is resilient to outliers, missing data, and a large number of related and unrelated variables (Meharie *et al*. 2021).

Table 4.3: Hyperparameter tuning using GridSearchCV

Model	Selected hyperparameter (param_grid)	Best hyperparameter
RFR	n_estimators: [50,100,200,300] max_features: ['auto', 'sqrt', 'log2'] max_depth: [None,2,4,6,8]	n_estimators: 200 max_features: 'auto' max_depth: None
GBR	n_estimators: [50,100,200,300,400] max_features: ['auto', 'sqrt', 'log2'] max_depth: [1,2,3,4]	n_estimators: 400 max_features: 'sqrt' max_depth: 3
KNNR	leaf size: [5,10,15,20,25,30] n_neighbors: [5,10,15,20] p: [1,2]	Leaf size: 15 n_neighbors: 5 p: 1

SVR	C: [1,5,10,20,40] epsilon: [0.05,0.1,0.3]	C: 1 epsilon: 0.3

According to Table 3.4, the GridSearchCV searched the list of parameters provided to identify the optimal hyperparameter for each model. Each model is then adjudged to be the best model using the best parameters. The constructed models were each evaluated using R2, MAE, RMSE and MAPE. The evaluation results were compared to be able to select the best performing model.

```
#evaluate
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
mape = mean_absolute_percentage_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
#the r2 for the train dataset
r2_train = r2_score(y_train, y_pred_train)

#printing the model evaluation values
print('Mean absolute error: {:.2f}'.format(mae))
print('Mean squared error: {:.2f}'.format(mse))
print('Root mean squared error: {:.2f}'.format(rmse))
print('Mean absolute percentage error:, {:.2f}'.format(mape))
print('r2 score:{:.2f}'.format(r2))
print('R2 Score Train data:{:.2f}'.format(r2_train))
```

Figure 4.12: Model evaluation

4.7 User Interface

The web application is created and deployed using Python's Streamlit module. The required libraries were imported and several streamlit components were used to create the user interface for predicting per capita carbon emissions.

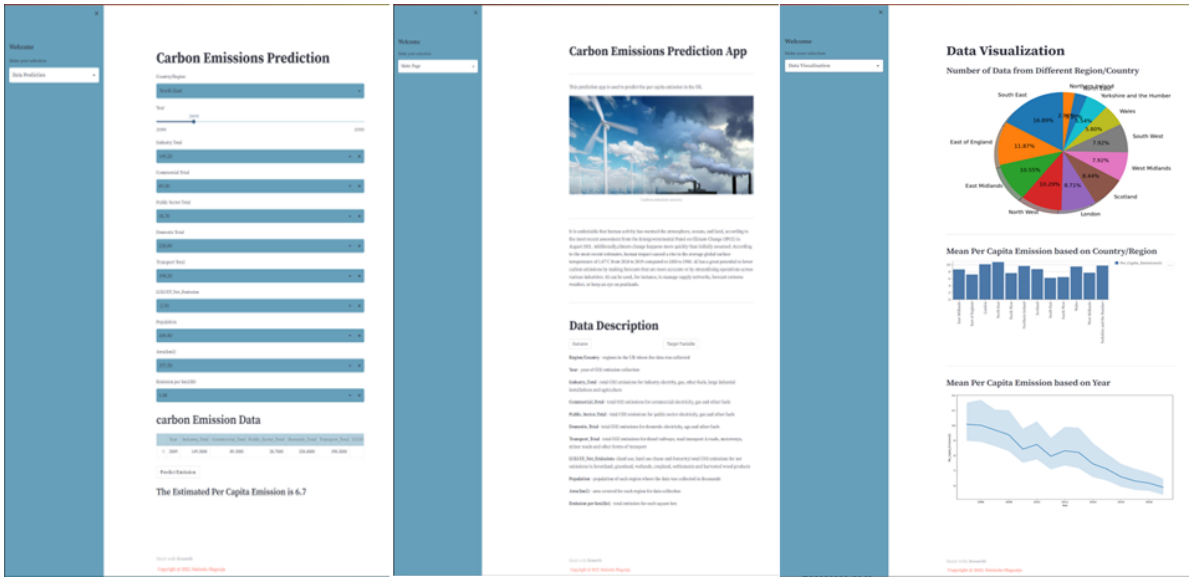


Figure 4.13: Web application for carbon emission prediction

4.8 Conclusion

The Implementation chapter provides an overview of the implementation process, data cleaning, model construction procedures, libraries utilised, and functions used to obtain predictions from the models.

5. RESULT AND DISCUSSION

5.1 Introduction

This chapter includes an in-depth data analysis presentation, together with the results and discussion of this research. The findings are divided into three sections: data robustness, tableau analysis, and analysis of model prediction accuracy.

5.2 Data robustness

We employ different feature selection methods for reducing the number of predictors as the last robustness check before training the ML learners. Three techniques, namely recursive feature removal, random forest feature importance, and Pearson correlation-based selection, were used for this (figure 4.10, figure 4.11 and figure 4.8). It was discovered that both algorithms make judgments that are very comparable. It's interesting to note that the prediction accuracy is almost the same when utilising all the predictors as compared to a limited predictor set when feeding ML algorithms. This demonstrates how our model accurately alleviate worries about redundancy.

The flooring and capping using percentiles method and the isolation forest method of outlier detection and removal were compared. According to the examination of the two approaches, capping and flooring using percentiles was more effective at identifying and eliminating outliers from the dataset.

5.3 Tableau Analysis

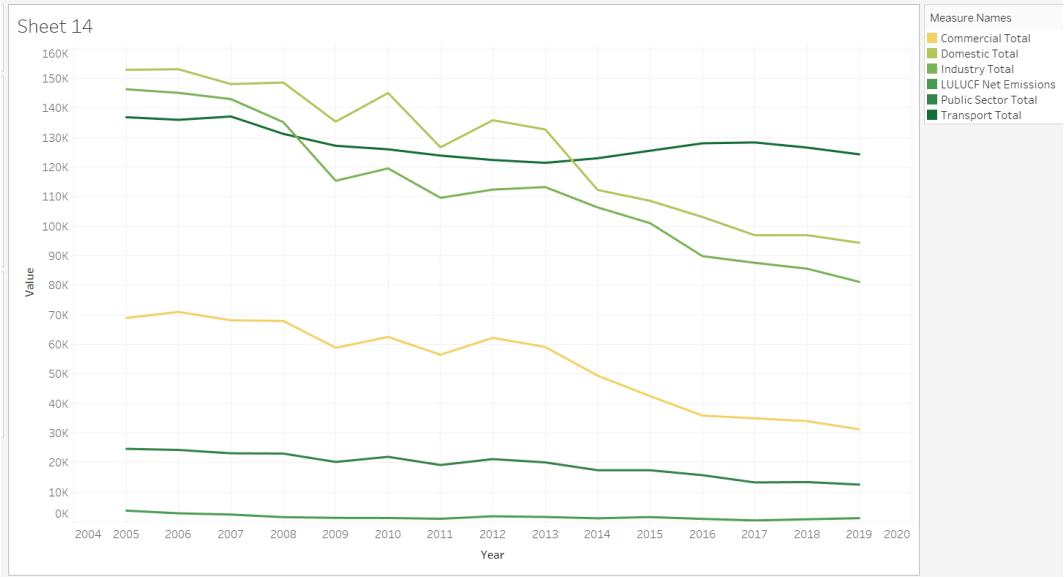


Figure 5.1: Emissions from all sectors in UK

The total carbon emissions from all sectors from 2005 to 2019 are represented in Figure 5.1. In 2019, LULUCF emissions are lowest at 1197 ($MtCO_2$) while transportation sector has the greatest emissions of 124303 ($MtCO_2$). 36% of total carbon dioxide emissions come from the transportation sector.

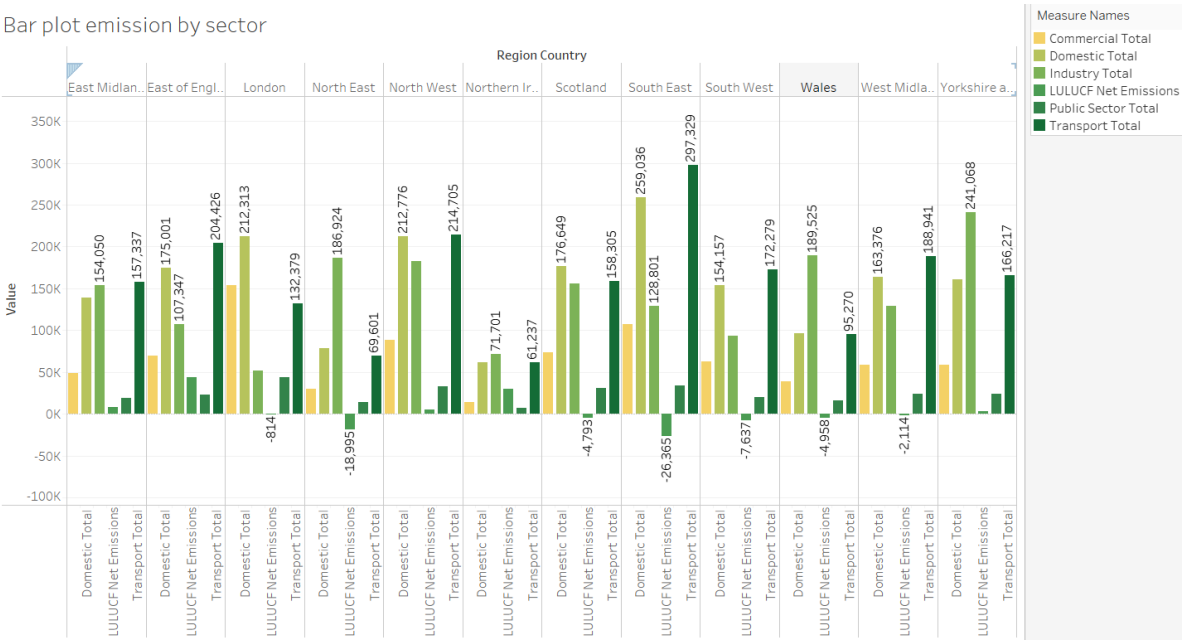


Figure 5.2: Carbon emission by sector and region

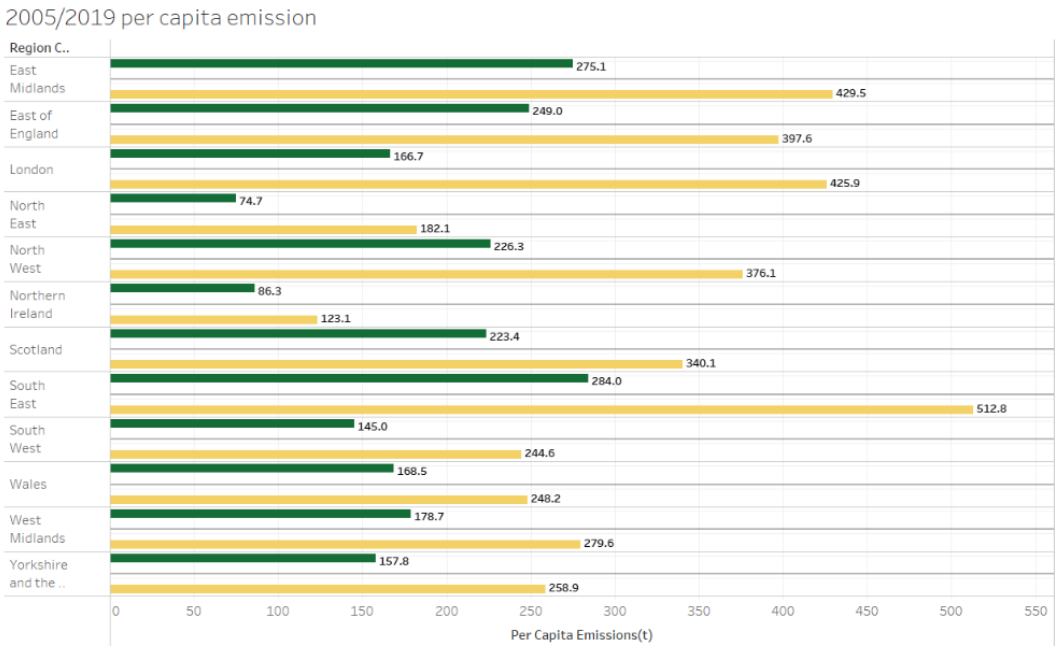


Figure 5.3: Comparing 2005 and 2019 emissions

Emissions per capita enable comparisons across regions with different population. Figure 4.8 shows that in 2019, the South-East, East Midlands, and London had the greatest per-capita emissions, while Northern Ireland had the lowest. North East had the lowest per capita emissions in 2005, while East Midlands, South-East, and North-West had the most. The South-East, the East Midlands, and London have higher emissions per person than other areas because of their dense populations and significant industrial and transportation emissions (Figure 4.7, appendix B). This is indicative of the substantial industrial base and dense population of these regions. Between 2005 and 2019, Northern Ireland's carbon emissions changed the least, by 29%. London saw the greatest percentage change between 2005 and 2019, at 60.85%. Figure shows three high carbon emitting sectors as industry sector, transport sector and domestic sector.

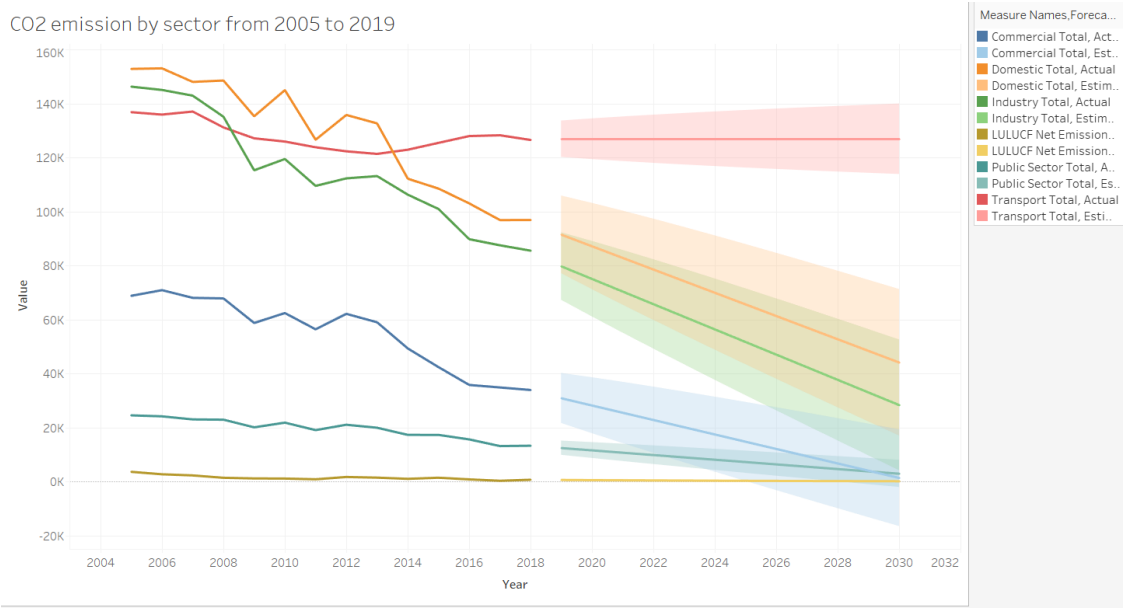


Figure 5.4: Carbon emission forecast by sector from 2020 to 2030

The forecast of carbon-emitting sectors from 2019 to 2030 (figure 5.1) reveals that LULUCF net emissions and the transportation sector are expected to remain almost constant; the overall change is not significant, public-sector emissions are declining at a slow rate, while emissions from the domestic, commercial, and industrial sectors have a strong steady decline. By 2030, emissions from the transport sector will be the greatest (126917 $MtCO_2$) and the lowest from LULUCF (264 $MtCO_2$). The commercial total showed the highest percentage change from 2019 to 2030, at 95.3% (appendix D).

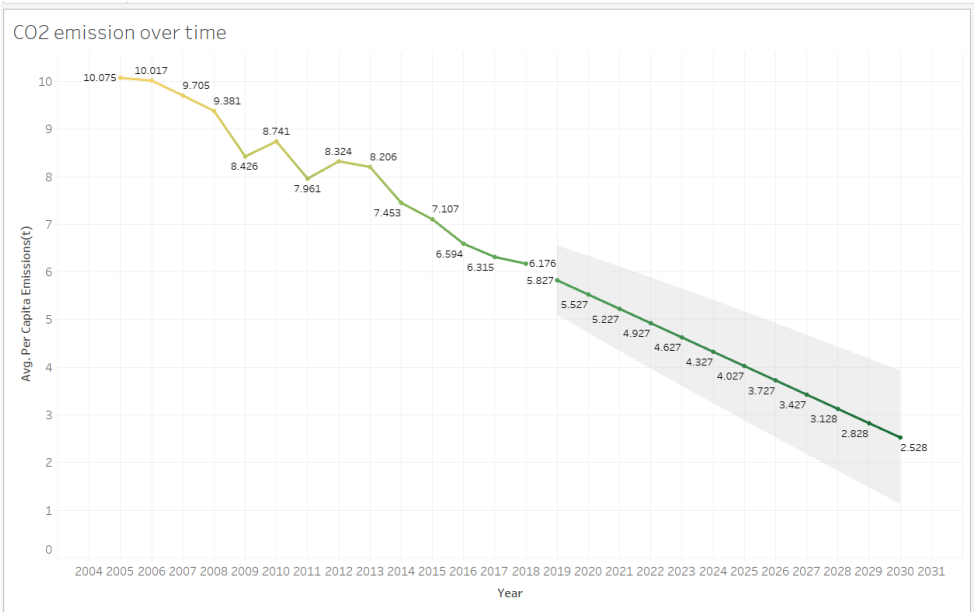


Figure 5.5: Forecast of UK's carbon emissions from 2019 – 2030

A distinct declining trend can be seen in figure 5.5. The average per capita emission is predicted to be 2.528 (t) by 2030, showing a trend toward decreasing emissions over time. This forecasted scenario shows that the UK will be able to fulfil its commitments under the Paris Agreement. By the end of 2030, implementing this approach will result in a 74.5% decrease.

5.4 Analysis of model prediction accuracy.

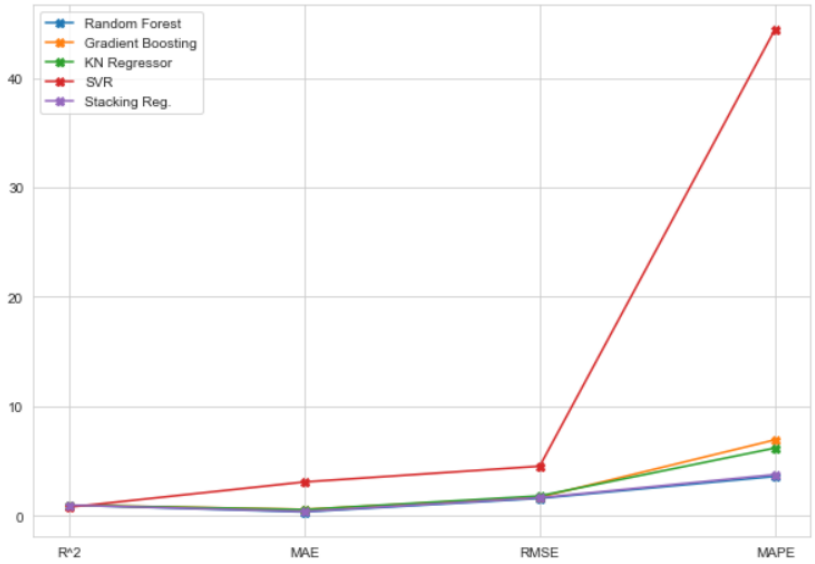


Figure 5.6: Models performance metrics graph

Table 5.1: Performance metrics of different models with outliers

Models	R^2	MAE	RMSE	MAPE
RFR	0.97	0.35	1.62	0.04
GBR	0.97	0.34	1.59	0.04
KNR	0.96	0.53	1.87	0.06
SVR	0.78	3.00	4.52	0.44
SR (RFR, KNR, SVR, GBR)	0.97	0.38	1.67	0.04
SR (RFR, KNR, GBR)	0.97	0.44	2.77	0.04

Table 5.2: Performance metrics of different models without outliers

Models	R^2	MAE	RMSE	MAPE
RFR	0.99	0.25	0.49	0.03
GBR	0.99	0.25	0.46	0.03
KNR	0.98	0.44	0.67	0.06
SVR	0.69	1.43	2.71	0.18
SR (RFR, KNR, SVR, GBR)	0.99	0.29	0.52	0.04
SR (RFR, GBR, KNR)	0.99	0.29	0.44	0.03

The performance accuracy of the models with outliers is shown in Table 5.1, while the performance accuracy without outliers is shown in Table 5.2. The results were studied in relation to the impact of eliminating outliers from the emission data. We maintained the outliers in our dataset and used the same process to train the single algorithms. The prediction accuracy with outliers and without outliers were compared. It was discovered that the predictions produced by the base ML algorithms and the SR had significantly improved. This validates the decision to eliminate outliers from the model.

For each model, the results are analysed, as shown in figure 5.3. To avoid a biased comparison between the different models, the best hyperparameters models were selected with GridSearchCV. To get the best performing SR model, two different combinations of base learners, (RFR, KNR, SVR) and (RFR, KNR) were used as input to the meta learner (GBR).

For all the performance metrics, SVR has the worst values. For single algorithms, RFR and GBR, the R^2, MAE and MAPE values are the same except for RMSE where RFR has a value of 0.49 and GBR has 0.46.

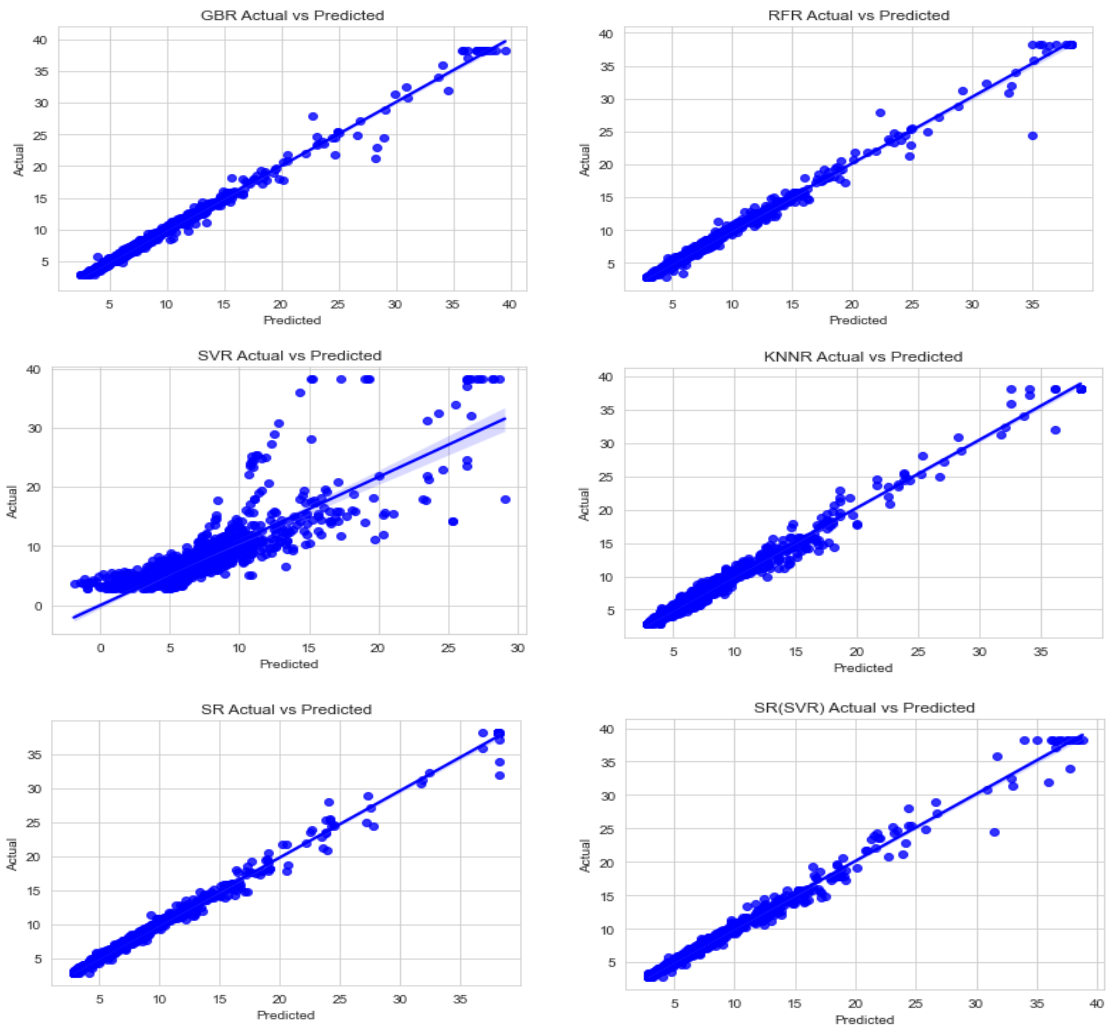


Figure 5.7: Plot of Actual versus predicted for all the models

Figure 5.4 displays the scatter plots of the actual and predicted values for the models developed by RFR, GBR, KNR, SR (RFR, KNR, SVR, GBR), and SR (RFR, KNR, GBR). The concentration of points near the centre line (actual value = predicted value) and even distribution of points on either side of the line indicates a strong correlation between the model's predictions and its actual values. The SVR model's scatter distributions display large dispersion, which denotes a poor correlation between the model's predicted result and the actual value.

The R^2 value of the SVR is 0.69, while the RFR, GBR, KNR, SR (RFR, KNR, SVR, GBR), and SR (RFR, KNR, GBR) are 0.99, 0.99, 0.98, 0.99, and 0.99, respectively, as given in table 5.2. Other measures also show that the suggested model in this research has more precision than the single models, demonstrating its superiority.

5.5 Discussion

The goal of this dissertation is to create an improved carbon emission level forecasting system using the stacking ensemble ML technique. Five regression models (RFR, GBR, KNNR, SVR and the SR) were used to establish the better prediction accuracy of the proposed stacking ensemble ML model. The regressors were developed and tested. Due to the inclusion of SVR, which has a low forecasting accuracy, SR (RFR, KNR, SVR, GBR) was unable to outperform SR (RFR, KNR, GBR). RFR, GBR and SR had the best R^2 value of 0.99. Based on the RMSE comparison, SR (RFR, KNR, GBR) has the lowest error value of 0.44 and is most appropriate for forecasting carbon emissions. Overall SVR is the worst performing model in terms of R2, MAE, RMSE and MAPE.

Consequently, the results indicate that the SR (RFR, KNR, GBR) model performs marginally better than the GBR and RFR models, which are supported by the preceding findings. Conclusions concerning the performance should be taken with caution, owing to the little variations across the models. In general, the SR (RFR, KNR, GBR) model is more accurate in predicting instances of carbon emission. GBR had the quickest execution time, whereas Stacking Regressor had the slowest predicted execution speed. It should be emphasised that the SR (RFR, KNR, GBR) model used most of its time during the training and model-development phases. Once the model has been created, the SR (RFR, KNR, and GBR) model takes about the same amount of time to test as the GBR and RFR model.

It is generalizable from the statistical results that the stacking ensemble model performs better than all individual learning methods in terms of predicting carbon emissions. Stacking ensemble ML approaches have never performed worse than choosing the most accurate single algorithm, according to Breiman (1996).

This excellent result of the SR algorithm is consistent with many other studies that produced various machine learning ensemble methods for practical regression problems (Agrawal*et al*. 2019; Al-Rakhami*et al*. 2019; Al-hajj, Assi and Fouad 2019; Nti, Adekoya and Weyori 2020; Tugay and Oguducu 2020; Mehari*et al*. 2021; Jebamalar and Kamalakannan 2021; Nguyen, Diaz-Rainey and Kuruppuarachchi 2021).

However, it is difficult to compare the outcomes of the suggested SR model with those of the ensemble learning algorithms that the abovementioned authors have researched. This is due to the fact that some requirements were considered in the research that were different from those considered in earlier studies, such as the carbon emissions dataset, selected feature, model hyperparameters, number of base-learners and application field (Ma *et al*. 2018, Xie *et al*. 2022). Banister (2019) observed that the transportation sector is crucial to attaining a low-carbon future since the UK's 2050 net zero objective is moving the direction of fulfilling the Paris Agreement.

Agbulut (2022) emphasised the effectiveness of machine learning (ML) models in forecasting carbon emissions, while Hosseini *et al*. (2019) demonstrated the population's significance as one of the important variables in predicting carbon emissions.

5.6 Summary

The results indicate a strong relationship between per capita emissions and important independent factors. This shows that the independent variables chosen are crucial for predicting carbon emissions. The accuracy of the models could be further improved by providing daily carbon emissions data which can forecast sequentially into the future. Due to lack of data on daily carbon emissions, seasonality, trends and cycle could not be analysed.

For the purpose of actively observing the process of reducing carbon emissions and reaching carbon neutrality, accurate prediction of CO_2 emissions useful and as a result, the SR model provides a straightforward and reliable model for predicting UK carbon emission. Policymakers may employ emissions reduction techniques and create achievable action schemes to minimise economic losses and maximise environmental benefits by accurately predicting the results of carbon emissions.

5. CONCLUSION AND FUTURE WORK

5.1 Introduction

The important research findings are summarized in this chapter as they relate to the research's aim and main research question, and their importance and contribution are discussed. It will also examine the research's limitations and suggest areas for future work.

5.2 Conclusion

Understanding global warming and climate change is essential for protecting the environment. The direct source of the global warming impact that resulted in extreme climate change is CO_2 emissions. Carbon emission predictions can be used to estimate future global warming, quantify the costs of lowering CO_2 emissions and reflect on the actual benefits of limiting temperature rise.

This research work aimed at forecasting carbon emission in UK using stacked ensemble ML algorithm by experimenting with several regression models to show the superiority of SR model based on the annual data from 2005 to 2019. The data on carbon emissions were analysed using four separate regressor models, namely RFR, GBR, KNNR, and SVR, as well as one stacked ensemble model, SR. The SR model used an optimal combination of the individual regressor models.

The data was normalised using robust scaler because of the presence of outliers in the data. The search for an optimal set of predictors was effectively solved by adopting the GridSearchCV algorithm to tune the hyperparameters so as to get the best parameters for each model used, as a result, the precision of the results of the carbon emissions study is substantially improved.

The results show that SR (RFR, KNR, GBR) model which is the best combination for stacking regressor outperforms the four models with the lowest RMSE value of 0.44 Its higher performance is verified across all the evaluation metrics which can positively indicate that the SR (RFR, KNR, GBR) model has improved accuracy and

generalization ability for carbon emissions prediction. The best model which is SR (RFR, KNR, GBR) model was deployed using streamlit to create a web application.

Further findings show that the stacking regressor model SR (RFR, KNR, GBR) shows the smallest deviation between actual and predicted carbon emissions when compared with the other individual regressor models.

According to the Tableau forecast results of UK's carbon emissions from 2019 to 2030, UK's carbon emissions will steadily decrease in the future. Carbon emissions are likely to fall to 2.58 tonnes per capita emissions by 2030. The results show that emissions in domestic sector, commercial sector and industrial sector showed a strong downward trend; LULUCF and transport sector showed a continuously constant trend in carbon emissions and public sector has a slow downward trend in carbon emissions.

The developed prediction model may also be used to analyse a wide range of large-scale carbon emissions data, including complicated and non-linear data with missing value, and outlier. The adoption of the newly suggested model package offers the UK's climate change committee significant advantages thanks to its improved predictive performance, including the ability to prepare more precise initial budgets, resource allocations, and project budget in the case of carbon emissions.

Even though the proposed stacking regression machine learning has achieved good results, several limitations remain that should be considered. There was lack of previous research studies on carbon emission in the UK which allows for further analysis and the data used was an annual data which was unable to capture seasonality, cycle and trends. This research clearly illustrated the relationship between the predictors and per capita emissions, but it also raises the issue of whether there are enough influencing elements to improve carbon emission prediction.

5.3 Recommendations and future work

Future research could examine a variety of important factors such as GDP and energy consumption while reducing the usage of categorical variables to enhance the performance of single algorithms in order to construct more accurate and robust models.

Expanding the research to include daily carbon emissions for future forecasting allows for the acquisition of a bigger dataset on carbon emissions. Additionally, a more thorough exploratory data analysis may be performed on the dataset to uncover additional insights and patterns.

Additionally, machine learning has emerged as a crucial technique in the study of CO_2 emission predictions. In order to increase accuracy and improve forecasting performance, the research can be improved by examining the performance of more sophisticated hybrid models. The method of this research's flowchart might be used by other high-CO_2 emitting nations like China, Russia, Japan, and Germany to create reliable prediction models for aiding decision-making when formulating environmental (climate change) policies.

The government should strongly endorse a framework that is favourable to sustainable growth as the UK nears the end of its third carbon budget and begins its fourth one the following year in order to successfully reduce CO_2 emissions. For citizens to actively participate in the reduction of CO_2 emissions via their own behaviours, the government must help people transition to a green economy and a low-carbon lifestyle.

6. REFERENCE LIST/ BIBLIOGRAPHY

A. SHABRI, 2022. Forecasting the annual carbon dioxide emissions of Malaysia using Lasso-GMDH neural network-based. - *2022 IEEE 12th Symposium on Computer Applications & Industrial Electronics (ISCAIE).* pp.123-127

ACHARYA, S. *et al.,* 2019. Non-invasive estimation of hemoglobin using a multi-model stacking regressor. *IEEE journal of biomedical and health informatics,* 24(6), 1717-1726

ACHEAMPONG, A.O. and E.B. BOATENG, 2019. Modelling carbon emission intensity: Application of artificial neural network. *Journal of Cleaner Production,* 225, 833-856

AĞBULUT, Ü, 2022. Forecasting of transportation-related energy demand and CO2 emissions in Turkey with different machine learning algorithms. *Sustainable Production and Consumption,* 29, 141-157

AKRAM, R. *et al.,* 2020. Heterogeneous effects of energy efficiency and renewable energy on carbon emissions: Evidence from developing countries. *Journal of Cleaner Production,* 247, 119122

AL-HAJJ, R., A. ASSI and M.M. FOUAD, 2019. Stacking-based ensemble of support vector regressors for one-day ahead solar irradiance prediction. *2019 8th International Conference on Renewable Energy Research and Applications (ICRERA).* IEEE, pp.428-433

ALKHEDER, S. and A. ALMUSALAM, 2022. Forecasting of carbon dioxide emissions from power plants in Kuwait using United States Environmental Protection Agency, Intergovernmental panel on climate change, and machine learning methods. *Renewable Energy,* 191, 819-827

AMARPURI, L. *et al.,* 2019. Prediction of CO 2 emissions using deep learning hybrid approach: A Case Study in Indian Context. *2019 Twelfth International Conference on Contemporary Computing (IC3).* IEEE, pp.1-6

BANISTER, D., 2019. The climate crisis and transport. *Transport Reviews,* 39(5), 565-568

BIENEFELD, C. *et al.,* 2022. On the Importance of Temporal Information for Remaining Useful Life Prediction of Rolling Bearings Using a Random Forest Regressor. *Lubricants,* 10(4), 67

BOLTON, P., 2021. *UK and global emissions and temperature trends* [viewed Aug 9, 2022]. Available from: https://commonslibrary.parliament.uk/uk-and-global-emissions-and-temperature-trends/

BOROWSKI, P.F., I. PATUK and E.R. BANDALA, 2022. Innovative Industrial Use of Bamboo as Key “Green” Material. *Sustainability,* 14(4), 1955

BREIMAN, L., 1996. Stacked regressions. *Machine Learning,* 24(1), 49-64

CAIN, M.K., Z. ZHANG and K. YUAN, 2017. Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior research methods,* 49(5), 1716-1735

CHAUDHRY, S.M. *et al.,* 2020. The impact of carbon emissions on country risk: Evidence from the G7 economies. *Journal of environmental management,* 265, 110533

CHEN, H., S. QI and X. TAN, 2022. Decomposition and prediction of China's carbon emission intensity towards carbon neutrality: From perspectives of national, regional and sectoral level. *The Science of the total environment; Sci Total Environ,* 825, 153839

CHEN, R. *et al.,* 2015. Forecasting holiday daily tourist flow based on seasonal support vector regression with adaptive genetic algorithm. *Applied Soft Computing,* 26, 435-443

CHICCO, D., M.J. WARRENS and G. JURMAN, 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science,* 7, e623

CHIU, Y. *et al.,* 2020. A Multivariate Grey Prediction Model Using Neural Networks with Application to Carbon Dioxide Emissions Forecasting. *Mathematical Problems in Engineering,* 2020, e8829948

Climate change a. [viewed Jun 20, 2022]. Available from: https://www.gov.uk/guidance/climate-change

Climate Change Committee b. [viewed Jul 3, 2022]. Available from: https://www.theccc.org.uk/

COCCO MARIANI, V. *et al.,* 2019. Pressure prediction of a spark ignition single cylinder engine using optimized extreme learning machine models. *Applied Energy,* 249, 204-221

COMMITTEE ON CLIMATE CHANGE, 2010. The fourth carbon budget: reducing emissions through the 2020s.

CUI, X. *et al.,* 2021. Forecasting of Carbon Emission in China Based on Gradient Boosting Decision Tree Optimized by Modified Whale Optimization Algorithm. *Sustainability,* 13, 12302

DEGOT, C. *et al.,* 2021. Reduce Carbon and Costs with the Power of AI. *BCG.[ed], URL: https://www.bcg.com/publications/2021/ai-toreduce-carbon-emissions,*

FANG, D. *et al.,* 2018. A novel method for carbon dioxide emission forecasting based on improved Gaussian processes regression. *Journal of Cleaner Production,* 173, 143-150

FARUQUE, M.O. *et al.,* 2022. A comparative analysis to forecast carbon dioxide emissions. *Energy Reports,* 8, 8046-8060

GAO, M. *et al.,* 2021. A novel fractional grey Riccati model for carbon emission prediction. *Journal of Cleaner Production,* 282, 124471

GNANA, D.A.A., S.A.A. BALAMURUGAN and E.J. LEAVLINE, 2016. Literature review on feature selection methods for high-dimensional data. *International Journal of Computer Applications,* 136(1), 9-17

GUO, J. *et al.*, 2021. Forecasting carbon dioxide emissions in BRICS countries by exponential cumulative grey model. *Energy Reports,* 7, 7238-7250

H. T. HO, 2018. Forecasting of CO2 Emissions, Renewable Energy Consumption and Economic Growth in Vietnam Using Grey Models. - *2018 4th International Conference on Green Technology and Sustainable Development (GTSD)*. pp.452-455

HAO, L. *et al.*, 2021. Green growth and low carbon emission in G7 countries: How critical the network of environmental taxes, renewable energy and human capital is? *Science of The Total Environment,* 752, 141853

HARRISON, M., 2019. *Machine learning pocket reference: working with structured data in python*. O'Reilly Media

HERWEIJER, C., B. COMBES and J. GILLHAM, 2018. How AI can enable a sustainable future. *Microsoft and PWC: London, UK,*

HEYDARI, A. *et al.*, 2019. Renewable Energies Generation and Carbon Dioxide Emission Forecasting in Microgrids and National Grids using GRNN-GWO Methodology. *Energy Procedia,* 159, 154-159

HOSSEINI, S.M. *et al.*, 2019. Forecasting of CO2 emissions in Iran based on time series and regression analysis. *Energy Reports,* 5, 619-631

HUANG, Y., L. SHEN and H. LIU, 2019. Grey relational analysis, principal component analysis and forecasting of carbon emissions based on long short-term memory in China. *Journal of Cleaner Production,* 209, 415-423

JEBAMALAR, J.A. and T. KAMALAKANNAN, 2021. Enhanced stacking ensemble model in predictive analytics of environmental sensor data. *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*. IEEE, pp.482-486

JIANG, P. *et al.*, 2021. A combined forecasting system based on statistical method, artificial neural networks, and deep learning methods for short-term wind speed forecasting. *Energy,* 217, 119361

KOMOROWSKI, M. *et al.*, 2016. Exploratory data analysis. *Secondary analysis of electronic health records,* , 185-203

KSHATRI, S.S. *et al.*, 2021. An empirical analysis of machine learning algorithms for crime prediction using stacked generalization: An ensemble approach. *IEEE Access,* 9, 67488-67500

KWAK, S.K. and J.H. KIM, 2017. Statistical data preparation: management of missing values and outliers. *Korean journal of anesthesiology,* 70(4), 407-411

L. AMARPURI *et al.*, 2019. Prediction of CO2 emissions using deep learning hybrid approach: A Case Study in Indian Context. - *2019 Twelfth International Conference on Contemporary Computing (IC3)*. pp.1-6

LANGEVIN, J., C.B. HARRIS and J.L. REYNA, 2019. Assessing the Potential to Reduce U.S. Building CO2 Emissions 80% by 2050. *Joule,* 3(10), 2403-2424

LI, Y., 2020. Forecasting Chinese carbon emissions based on a novel time series prediction method. *Energy, Science and Engineering,*

LI, Z. *et al.,* 2021. Determinants of Carbon Emission in China: How Good is Green Investment? *Sustainable Production and Consumption,* 27, 392-401

LIU, Z. *et al.,* 2022. Ensemble system for short term carbon dioxide emissions forecasting based on multi-objective tangent search algorithm. *Journal of environmental management,* 302, 113951

MA, J. *et al.,* 2020. Identification of high impact factors of air quality on a national scale using big data and machine learning techniques. *Journal of Cleaner Production,* 244, 118955

MA, X., P. JIANG and Q. JIANG, 2020. Research and application of association rule algorithm and an optimized grey model in carbon emissions forecasting. *Technological Forecasting and Social Change,* 158, 120159

MA, Z. *et al.,* 2018. Ensemble of machine learning algorithms using the stacked generalization approach to estimate the warfarin dose. *PloS one,* 13(10), e0205872

MARDANI, A. *et al.,* 2020. A multi-stage method to predict carbon dioxide emissions using dimensionality reduction, clustering, and machine learning techniques. *Journal of Cleaner Production,* 275, 122942

MASON, K., J. DUGGAN and E. HOWLEY, 2018. Forecasting energy demand, wind generation and carbon dioxide emissions in Ireland using evolutionary neural networks. *Energy,* 155, 705-720

MASSAOUDI, M. *et al.,* 2021. A novel stacked generalization ensemble-based hybrid LGBM-XGB-MLP model for short-term load forecasting. *Energy,* 214, 118874

MEHARIE, M.G. *et al.,* 2021. Application of stacking ensemble machine learning algorithm in predicting the cost of highway construction projects. *Engineering, Construction and Architectural Management,*

MENG, M. and C. SONG, 2020. Daily photovoltaic power generation forecasting model based on random forest algorithm for north China in winter. *Sustainability,* 12(6), 2247

MODISE, R.K., K. MPOFU and O.T. ADENUGA, 2021. Energy and Carbon Emission Efficiency Prediction: Applications in Future Transport Manufacturing. *Energies,* 14(24), 8466

MOHAREKAR, T.T. *et al.,* DETECTION AND CLASSIFICATION OF PLANT LEAF DISEASES USING CONVOLUTION NEURAL NETWORKS AND STREAMLIT.

MOTT, G., RAZO, C. and HAMWEY, R., 2021. *Carbon emissions anywhere threaten development everywhere* [viewed Aug 9, 2022]. Available from: https://unctad.org/news/carbon-emissions-anywhere-threaten-development-everywhere

NGUYEN, D.K., T.L.D. HUYNH and M.A. NASIR, 2021. Carbon emissions determinants and forecasting: Evidence from G6 countries. *Journal of environmental management,* 285, 111988

NGUYEN, Q., I. DIAZ-RAINEY and D. KURUPPUARACHCHI, 2021. Predicting corporate carbon footprints for climate finance risk analyses: A machine learning approach. *Energy Economics,* 95, 105129

NGUYEN, R. *et al.,* 2021. Predicting PV Power Generation using SVM Regression. *2021 IEEE Green Energy and Smart Systems Conference (IGESSC).* IEEE, pp.1-5

NILTOP, A., 2022. How to Reduce the AI Carbon Footprint as a Data Scientist. In: *statworx*®. -02-02T15:00:59+00:00 [viewed Aug 18, 2022]. Available from: https://www.statworx.com/en/content-hub/blog/how-to-reduce-the-ai-carbon-footprint-as-a-data-scientist/

NTI, I.K., A.F. ADEKOYA and B.A. WEYORI, 2020. A comprehensive evaluation of ensemble learning for stock-market prediction. *Journal of Big Data,* 7(1), 1-40

OUYANG, Z. *et al.,* 2018. Multi-view stacking ensemble for power consumption anomaly detection in the context of industrial internet of things. *IEEE Access,* 6, 9623-9631

OVERBEEK, M.V. *et al.,* 2021. Covid-19 Prediction in Indonesia Capital Buffer with K Nearest Neighbor Regression. *2021 5th International Conference on Informatics and Computational Sciences (ICICoS).* IEEE, pp.238-243

P. KADAM and S. VIJAYUMAR, 2018. Prediction Model: CO2 Emission Using Machine Learning. *- 2018 3rd International Conference for Convergence in Technology (I2CT).* pp.1-3

PLOTNIKOVA, V., M. DUMAS and F. MILANI, 2020. Adaptations of data mining methodologies: a systematic literature review. *PeerJ Computer Science,* 6, e267

QADER, M.R. *et al.,* 2021. Forecasting carbon emissions due to electricity power generation in Bahrain. *Environmental Science and Pollution Research,* 29, 17346–17357

QIAO, W. *et al.,* 2020. A hybrid algorithm for carbon dioxide emissions forecasting based on improved lion swarm optimizer. *Journal of Cleaner Production,* 244, 118612

QIAO, Z., X. MENG and L. WU, 2021. Forecasting carbon dioxide emissions in APEC member countries by a new cumulative grey model. *Ecological Indicators,* 125, 107593

QIU, W., W.J. MURPHY and A. SUTER, 2020. Kurtosis: a new tool for noise analysis. *Acoust Today,* 16(4), 39-47

R. AL-HAJJ, A. ASSI and M. M. FOUAD, 2019. Stacking-Based Ensemble of Support Vector Regressors for One-Day Ahead Solar Irradiance Prediction. *- 2019 8th International Conference on Renewable Energy Research and Applications (ICRERA).* pp.428-433

RAJAGOPAL, S., P.P. KUNDAPUR and K.S. HAREESHA, 2020. A stacking ensemble for network intrusion detection using heterogeneous datasets. *Security and Communication Networks,* 2020

RAO, M., 2021. *Machine Learning in Estimating CO_2 Emissions from Electricity Generation*. IntechOpen

RAUCH, S., 2022. *How Data Science is Driving Innovation in Climate Change Research* [viewed Aug 18, 2022]. Available from: https://www.simplilearn.com/how-data-science-is-driving-innovation-in-climate-change-research-article

REN, F. and D. LONG, 2021. Carbon emission forecasting and scenario analysis in Guangdong Province based on optimized Fast Learning Network. *Journal of Cleaner Production,* 317, 128408

REN, L. *et al.,* 2021. A review of CO2 emissions reduction technologies and low-carbon development in the iron and steel industry focusing on China. *Renewable and Sustainable Energy Reviews,* 143, 110846

RITCHIE, H., M. ROSER and P. ROSADO, 2020. CO_2 and Greenhouse Gas Emissions. *Our World in Data,*

ROY, M.S. *et al.,* 2021. Comparative Analysis of Machine Learning Methods to Detect Chronic Kidney Disease. *Journal of Physics: Conference Series*. IOP Publishing, pp.012005

SAHA, S., A. HAQUE and G. SIDEBOTTOM, 2022. Towards an ensemble regressor model for anomalous isp traffic prediction. *arXiv preprint arXiv:2205.01300,*

SALVIA, M. *et al.,* 2021. Will climate mitigation ambitions lead to carbon neutrality? An analysis of the local-level plans of 327 cities in the EU. *Renewable and Sustainable Energy Reviews,* 135, 110253

SANDHYA, V. and A. PADYANA, 2021. Machine Learning based Crop Yield Prediction on Geographical and Climatic Data. *2021 Sixth International Conference on Image Information Processing (ICIIP)*. IEEE, pp.186-191

SCHOKLITSCH, H., *Council Post: Climate Change And Big Data: Investing For A Solution* [viewed Aug 18, 2022]. Available from: https://www.forbes.com/sites/forbesfinancecouncil/2019/09/06/climate-change-and-big-data-investing-for-a-solution/

SCHRÖER, C., F. KRUSE and J.M. GÓMEZ, 2021. A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science,* 181, 526-534

SHABANI, E. *et al.,* 2021. A novel approach to predict CO2 emission in the agriculture sector of Iran based on Inclusive Multiple Model. *Journal of Cleaner Production,* 279, 123708

SHAHBAZ, M. *et al.,* 2021. Time-varying impact of financial development on carbon emissions in G-7 countries: Evidence from the long history. *Technological Forecasting and Social Change,* 171, 120966

SHAHBAZ, M., M.A. NASIR and D. ROUBAUD, 2018. Environmental degradation in France: The effects of FDI, financial development, and energy innovations. *Energy Economics,* 74, 843-857

STAERMAN, G. *et al.,* 2019. Functional isolation forest. *Asian Conference on Machine Learning.* PMLR, pp.332-347

STARK, C., M. THOMPSON and CLIMATE CHANGE COMMITTEE, 2019. Net Zero The UK's contribution to stopping global warming.

SUN, W., H. JIN and X. WANG, 2019. Predicting and Analyzing CO2 Emissions Based on an Improved Least Squares Support Vector Machine. *Polish journal of environmental studies,* 28(6), 4391-4401

TEG, A. and A. ALARJANI, 2021. A Comparative Study of CO2 Emission Forecasting in the Gulf Countries Using Autoregressive Integrated Moving Average, Artificial Neural Network, and Holt-Winters Exponential Smoothing Models. *Advances in Meteorology,* 2021

TORRE-TOJAL, L. *et al.,* 2022. Above-ground biomass estimation from LiDAR data using random forest algorithms. *Journal of Computational Science,* 58, 101517

TUGAY, R. and S.G. OGUDUCU, 2020. Demand prediction using machine learning methods and stacked generalization. *arXiv preprint arXiv:2009.09756,*

UKCP headline findings c. [viewed Jun 20, 2022]. Available from: https://www.metoffice.gov.uk/research/approach/collaboration/ukcp/summaries/headline-findings

V. TANANIA, S. SHUKLA and S. SINGH, 2020. Time Series Data Analysis And Prediction Of CO2 Emissions. - *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence).* pp.665-669

WANG, P., Y. ZHONG and Z. YAO, 2022. Modeling and Estimation of CO2 Emissions in China Based on Artificial Intelligence. *Computational Intelligence and Neuroscience,* 2022, e6822467

WANG, R., 2018. Significantly improving the prediction of molecular atomization energies by an ensemble of machine learning algorithms and rescanning input space: A stacked generalization approach. *The Journal of Physical Chemistry C,* 122(16), 8868-8873

WEI, S., W. YUWEI and C. ZANG, 2018. Forecasting CO2 emissions in Hebei, China, through moth-flame optimization based on the random forest and extreme learning machine. *Environmental Science and Pollution Research,* 25, 28985–28997

WEN, L. and Y. CAO, 2020. Influencing factors analysis and forecasting of residential energy-related CO2 emissions utilizing optimized support vector machine. *Journal of Cleaner Production,* 250, 119492

WEN, L. and X. YUAN, 2020. Forecasting CO2 emissions in Chinas commercial department, through BP neural network based on random forest and PSO. *Science of The Total Environment,* 718, 137194

WU, L. *et al.*, 2015. Modelling and forecasting CO2 emissions in the BRICS (Brazil, Russia, India, China, and South Africa) countries using a novel multi-variable grey model. *Energy,* 79, 489-495

XU, H. *et al.,* 2021. Forecasting Chinese CO2 emission using a non-linear multi-agent intertemporal optimization model and scenario analysis. *Energy,* 228, 120514

YANG, H. and J.F. O'CONNELL, 2020. Short-term carbon emissions forecast for aviation industry in Shanghai. *Journal of Cleaner Production,* 275, 122734

YE, L., N. XIE and A. HU, 2021. A novel time-delay multivariate grey model for impact analysis of CO2 emissions from China's transportation sectors. *Applied Mathematical Modelling,* 91, 493-507

ZHANG, F. *et al.,* 2016. Urban link travel time prediction based on a gradient boosting method considering spatiotemporal correlations. *ISPRS International Journal of Geo-Information,* 5(11), 201

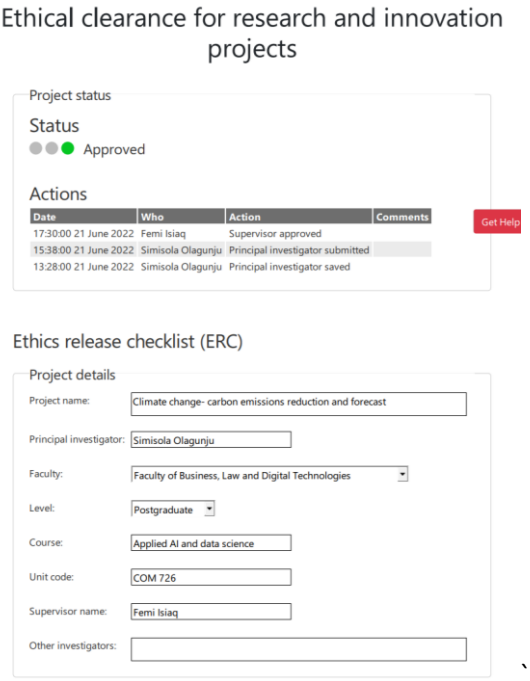
ZHOU, T. and H. JIAO, 2022. Exploration of the stacking ensemble machine learning algorithm for cheating detection in large-scale assessment. *Educational and Psychological Measurement,* , 00131644221117193

ZHOU, W. *et al.*, 2021. Forecasting Chinese carbon emissions using a novel grey rolling prediction model. *Chaos, Solitons & Fractals,* 147, 110968

ZHU, B. *et al.*, 2019. Achieving the carbon intensity target of China: A least squares support vector machine with mixture kernel function approach. *Applied Energy,* 233-234, 196-207

7. APPENDICES

7.1 Appendix A: Ethics Form



Ethical clearance for research and innovation projects

Project status

Status

Approved

Actions

Date	Who	Action	Comments
17:30:00 21 June 2022	Femi Isiaq	Supervisor approved	
15:38:00 21 June 2022	Simisola Olagunju	Principal investigator submitted	
13:28:00 21 June 2022	Simisola Olagunju	Principal investigator saved	

Get Help

Ethics release checklist (ERC)

Project details

Project name: Climate change- carbon emissions reduction and forecast

Principal investigator: Simisola Olagunju

Faculty: Faculty of Business, Law and Digital Technologies

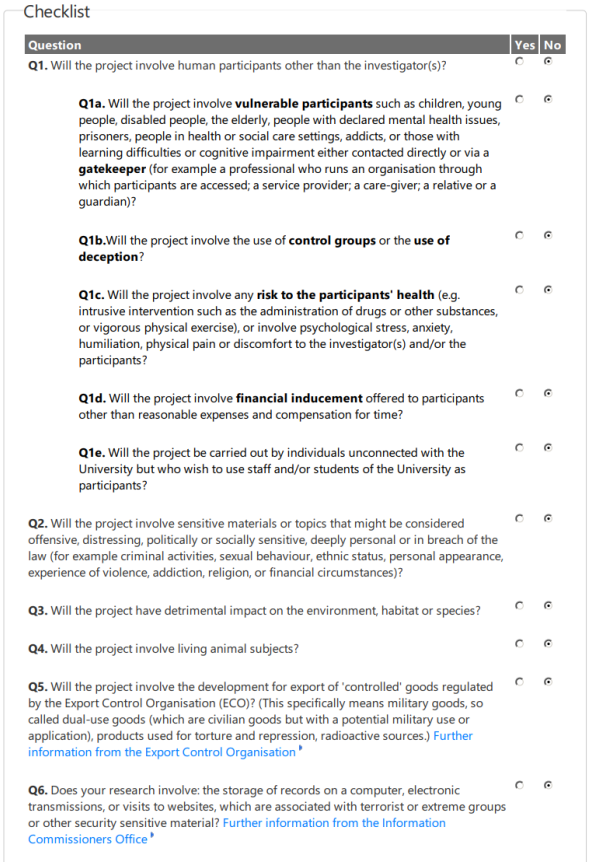
Level: Postgraduate

Course: Applied AI and data science

Unit code: COM 726

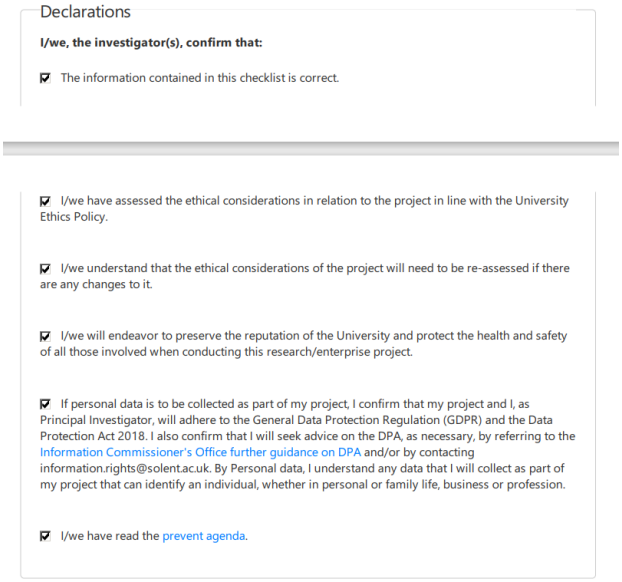
Supervisor name: Femi Isiaq

Other investigators:



Checklist

Question	Yes	No
Q1. Will the project involve human participants other than the investigator(s)?	○	●
Q1a. Will the project involve **vulnerable participants** such as children, young people, disabled people, the elderly, people with declared mental health issues, prisoners, people in health or social care settings, addicts, or those with learning difficulties or cognitive impairment either contacted directly or via a **gatekeeper** (for example a professional who runs an organisation through which participants are accessed; a service provider; a care-giver; a relative or a guardian)?	○	●
Q1b. Will the project involve the use of **control groups** or the **use of deception**?	○	●
Q1c. Will the project involve any **risk to the participants' health** (e.g. intrusive intervention such as the administration of drugs or other substances, or vigorous physical exercise), or involve psychological stress, anxiety, humiliation, physical pain or discomfort to the investigator(s) and/or the participants?	○	●
Q1d. Will the project involve **financial inducement** offered to participants other than reasonable expenses and compensation for time?	○	●
Q1e. Will the project be carried out by individuals unconnected with the University but who wish to use staff and/or students of the University as participants?	○	●
Q2. Will the project involve sensitive materials or topics that might be considered offensive, distressing, politically or socially sensitive, deeply personal or in breach of the law (for example criminal activities, sexual behaviour, ethnic status, personal appearance, experience of violence, addiction, religion, or financial circumstances)?	○	●
Q3. Will the project have detrimental impact on the environment, habitat or species?	○	●
Q4. Will the project involve living animal subjects?	○	●
Q5. Will the project involve the development for export of 'controlled' goods regulated by the Export Control Organisation (ECO)? (This specifically means military goods, so called dual-use goods (which are civilian goods but with a potential military use or application), products used for torture and repression, radioactive sources.) Further information from the Export Control Organisation	○	●
Q6. Does your research involve: the storage of records on a computer, electronic transmissions, or visits to websites, which are associated with terrorist or extreme groups or other security sensitive material? Further information from the Information Commissioners Office	○	●



Declarations

I/we, the investigator(s), confirm that:

☑ The information contained in this checklist is correct.

☑ I/we have assessed the ethical considerations in relation to the project in line with the University Ethics Policy.

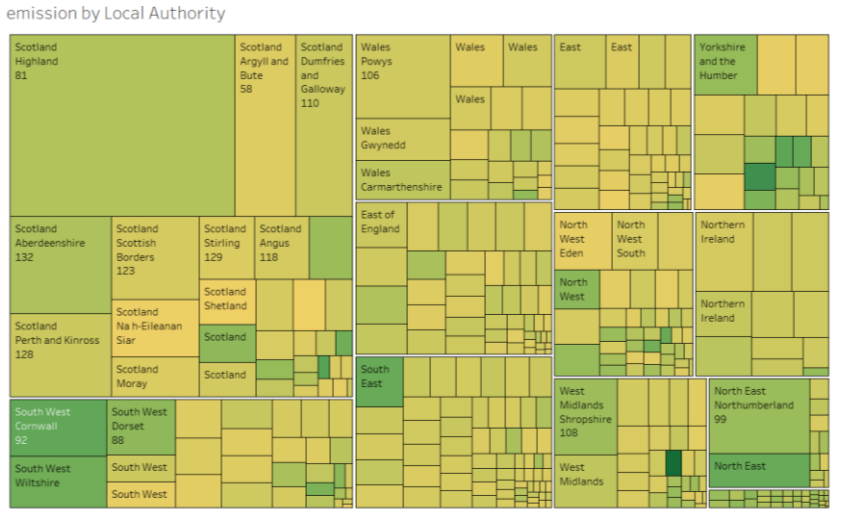
☑ I/we understand that the ethical considerations of the project will need to be re-assessed if there are any changes to it.

☑ I/we will endeavor to preserve the reputation of the University and protect the health and safety of all those involved when conducting this research/enterprise project.

☑ If personal data is to be collected as part of my project, I confirm that my project and I, as Principal Investigator, will adhere to the General Data Protection Regulation (GDPR) and the Data Protection Act 2018. I also confirm that I will seek advice on the DPA, as necessary, by referring to the Information Commissioner's Office further guidance on DPA and/or by contacting information.rights@solent.ac.uk. By Personal data, I understand any data that I will collect as part of my project that can identify an individual, whether in personal or family life, business or profession.

☑ I/we have read the prevent agenda.

7.2 **Appendix B:** Tableau Visualization



7.3 **Appendix C:** Forecast Model summary for carbon emissions in Tableau

All forecasts were computed using exponential smoothing.

Sum of Commercial Total

Model			Quality Metrics					Smoothing Coefficients		
Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Additive	Additive	None	4,718	4,197	0.94	8.0%	247	0.500	0.000	0.000

Sum of Domestic Total

Model			Quality Metrics					Smoothing Coefficients		
Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Additive	Additive	None	7,332	6,739	0.92	5.3%	259	0.482	0.000	0.000

Sum of Industry Total

Model			Quality Metrics					Smoothing Coefficients		
Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Additive	Additive	None	6,373	4,906	0.84	4.2%	255	0.500	0.000	0.000

Sum of LULUCF Net Emissions

Model			Quality Metrics					Smoothing Coefficients		
Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Multiplicative	Multiplicative	None	522	418	0.87	28.5%	185	0.029	0.000	0.000

Sum of Public Sector Total

Model			Quality Metrics					Smoothing Coefficients		

Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Additive	Additive	None	1,340	1,165	0.80	6.0%	212	0.490	0.000	0.000

Sum of Transport Total

Model			Quality Metrics					Smoothing Coefficients		
Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Additive	None	None	3,436	2,833	1.40	2.2%	234	0.500	0.000	0.000

Avg. Emissions per km2(kt)

Model			Quality Metrics					Smoothing Coefficients		
Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Additive	Additive	None	0.479	0.449	0.86	4.9%	-11	0.500	0.000	0.000

Avg. Per Capita Emissions(t)

Model			Quality Metrics					Smoothing Coefficients		
Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Additive	Additive	None	0.368	0.326	0.81	3.9%	-18	0.500	0.000	0.000

7.4 Appendix D: Forecast summary

Options Used to Create Forecasts

Time series: Year

Measures: Sum of Commercial Total, Sum of Domestic Total, Sum of Industry Total, Sum of LULUCF Net Emissions, Sum of Public Sector Total, Sum of Transport Total

Forecast forward: 12 periods (2019 – 2030)

Forecast base on: 2005 – 2018

Ignore last: 1 period (2019)

Seasonal pattern: None (Searched for a seasonal pattern recurring every 1 Periods)

Sum of Commercial Total

Initial	Change From Initial	Seasonal Effect			Contribution		
2019	2019 – 2030	Length	High	Low	Trend	Season	Quality
30,962 ± 29.9%	-95.3%	None			100.0%	0.0%	Poor

Sum of Domestic Total

Initial	Change From Initial	Seasonal Effect			Contribution		
2019	2019 – 2030	Length	High	Low	Trend	Season	Quality
91,521 ± 15.7%	-51.7%	None			100.0%	0.0%	Poor

Sum of Industry Total

Initial	Change From Initial	Seasonal Effect			Contribution		
2019	2019 – 2030	Length	High	Low	Trend	Season	Quality
79,795 ± 15.7%	-64.4%	None			100.0%	0.0%	Poor

Sum of LULUCF Net Emissions

Initial	Change From Initial	Seasonal Effect			Contribution		
2019	2019 – 2030	Length	High	Low	Trend	Season	Quality
707 ± 56.9%	-62.6%	None			100.0%	0.0%	Poor

Sum of Public Sector Total

Initial	Change From Initial	Seasonal Effect			Contribution		
2019	2019 – 2030	Length	High	Low	Trend	Season	Quality
12,544 ± 20.9%	-75.9%	None			100.0%	0.0%	Ok

Sum of Transport Total

Initial	Change From Initial	Seasonal Effect			Contribution		
2019	2019 – 2030	Length	High	Low	Trend	Season	Quality
126,917 ± 5.3%	0.0%	None			0.0%	0.0%	Poor

7.5 **Appendix E:** User interface code

```
import ...

hide_menu = """
<style>
footer:after{
    content: 'Copyright @ 2022: Simisola Olagunju' ;
    display:block;
    position:relative;
    color:tomato;
    padding:5px;
    top:3px
}
<style>
"""
st.markdown(hide_menu,unsafe_allow_html=True)
loaded_model = joblib.load('rf_model.sav')

# Create a page dropdown
st.sidebar.header('Welcome')
page = st.sidebar.selectbox("""
Make your selection""", ["Main Page", "Data Visualisation", "Data Prediction"])


if page == "Main Page":

    ### INFO
    st.title('Carbon Emissions Prediction App')
    st.markdown("""---""")
    st.write(
        '''
```

```
    st.markdown("""---""")
    st.write(
        '''
        # Data Description
        ''')
    # data description
    col1, col2 = st.columns(2)

    button1 = col1.button("features")
    if (button1):
        st.write('**Region/Country** - regions in the UK where the data was collected')
        st.write('**Year** - year of CO2 emission collection')
        st.write('**Industry_Total** - total CO2 emissions for industry electrity, gas, other fuels, large indusrial installations and agriculture')
        st.write('**Commercial_Total** - total CO2 emissions for commercial electricity, gas and other fuels')
        st.write('**Public_Sector_Total** - total CO2 emissions for public sector electricity, gas and other fuels')
        st.write('**Domestic_Total** - total CO2 emissions for domestic electricity, ags and other fuels')
        st.write('**Transport_Total** - total CO2 emissions for diesel railways, road transport A roads, motorways, minor roads and other forms of transport')
        st.write('**LULUCF_Net_Emissions** -(land use, land use chane and foresrtry) total CO2 emissions for net emissions in forestland, grassland, wetlands, cropland, settlemer
        st.write('**Population** - population of each region where the data was collected in thousands')
        st.write('**Area(km2)** - area covered for each region for data collection')
        st.write('**Emission per km2(kt)** - total emission for each square km')

    button2 = col2.button("Target Variable")
    if (button2):
        st.write('**Per_Capita_Emissions(t)** - tonnes of CO2 emission per person')

if page == "Data Visualisation":

    ### INFO
    st.title('Data Visualization')
```

```
    Transport_Total = st.number_input('Transport Total', key=4)
    LULUCF_Net_Emissions = st.number_input('LULUCF Net Emission', key=5)
    Population = st.number_input('Population', key=6)
    Area = st.number_input('Area(km2)', key=7)
    Emissions_per_km2 = st.number_input('Emission per km2(kt)', key=8)

    user_report_data = {
        'Year': Year,
        'Industry_Total': Industry_Total,
        'Commercial_Total': Commercial_Total,
        'Public_Sector_Total': Public_Sector_Total,
        'Domestic_Total': Domestic_Total,
        'Transport_Total': Transport_Total,
        'LULUCF_Net_Emissions': LULUCF_Net_Emissions,
        'Population': Population,
        'Area' : Area,
        'Emissions_per_km2' : Emissions_per_km2
    }

    report_data = pd.DataFrame(user_report_data, index=[0])
    return report_data

user_data = user_report()
st.header(' carbon Emission Data')
st.write(user_data)

emission = loaded_model.predict(user_data)
ok = st.button('Predict Emission')
if ok:
    st.subheader(f'The Estimated Per Capita Emission is {np.round(emission[0], 2)}')
```

7.6 **Appendix F:** Implementation code

```
X = my_data.drop('Per_Capita_Emissions(t)', axis=1) #  X= my_data.iloc[:,:-1]
y = my_data['Per_Capita_Emissions(t)'] # y= my_data.iloc[:,-1]
```

```
#declare a linear model
regress_model = LinearRegression()

#recursive feature elimination
rfe = RFE(regress_model, n_features_to_select=5, step=1)
rfe = rfe.fit(X, y)

print("Number of features", rfe.n_features_)
print("Selected Features", rfe.support_)
print("Feature Ranking", rfe.ranking_)
```

```
Number of features 5
```

```
ffe = f_regression(X,y)
variab = []
for i in range(0, len(X.columns)-1):
    if ffe[0][i] >=10:  #check variables importance over 10 percent
        variab.append(X.columns[i])

print(variab)
```

```
['Year', 'Industry_Total', 'Commercial_Total', 'Domestic_Total', 'LULUCF_Net_Emissions', 'Population']
```

```
rf_model = RandomForestRegressor(random_state=0)
rf_model.fit(X, y)
t_features = X.columns
relevance= rf_model.feature_importances_
imp = np.argsort(relevance)[-9:] #Top 10 features
plt.title('Feature Importances')
plt.barh(range(len(imp)), relevance[imp],align='center')
plt.yticks(range(len(imp)), [t_features[i] for i in imp])
plt.xlabel('Relative Importance')
plt.show()
```

```
def cap_data(df):
    for col in df.columns:
        print("capping the ",col)
        if (((df[col].dtype)=='float64') | ((df[col].dtype)=='int64')):
            percentiles = df[col].quantile([0.01,0.99]).values
            df[col][df[col] <= percentiles[0]] = percentiles[0]
            df[col][df[col] >= percentiles[1]] = percentiles[1]
        else:
            df[col]=df[col]
    return df

final_df=cap_data(my_data)
```

```
iforest = IsolationForest(random_state=0)
y_pred = iforest.fit_predict(X_train)
```

```
X_train_iforest, y_train_iforest = X_train.iloc[(y_pred != -1), :], y_train.iloc[(y_pred != -1)]
```

```
print(X_train_iforest.shape, y_train_iforest.shape)
```

```
rfr=RandomForestRegressor(random_state=0)
forest_grid = {
    'n_estimators': [50,100,200,300],
    'max_features': ['auto', 'sqrt', 'log2'],
    'max_depth' : [None,2,4,6,8]
}
CV_rfr = GridSearchCV(estimator=rfr, param_grid=forest_grid, cv= 5)
CV_rfr.fit(X_train, y_train)
CV_rfr.best_params_
```

```
{'max_depth': None, 'max_features': 'auto', 'n_estimators': 200}
```

```
rf_model = RandomForestRegressor(max_depth=None, max_features= 'auto', n_estimators=200, random_state=0)
#fit the model to the training set
model1 = rf_model.fit(X_train, y_train)
#predict the model
y_pred = model1.predict(X_test)
#predict model on train data
```

```
gbr=GradientBoostingRegressor(random_state=0)
gradient_param = {'n_estimators': [50,100,200,300,400],
                  'max_features': ['auto', 'sqrt', 'log2'],
                  'max_depth' : [1,2,3,4]
}
CV_gbr = GridSearchCV(estimator=gbr, param_grid=gradient_param, cv= 5)
CV_gbr.fit(X_train, y_train)
CV_gbr.best_params_
```

```
{'max_depth': 3, 'max_features': 'sqrt', 'n_estimators': 400}
```

```
#initialise the model
gb_model = GradientBoostingRegressor(max_depth= 3, max_features= 'sqrt', n_estimators= 400, random_state=0)
#fit gb model to the training set
model2 = gb_model.fit(X_train, y_train.ravel())  #use ravel to avoid warning about 2d array
#predict the model
y_pred = model2.predict(X_test)
#predict model on train data
```

```
knr=KNeighborsRegressor()
knr_param = {
    'leaf_size': [5,10,15,20,25,30],
    'n_neighbors': [5,10,15,20],
    'p' : [1,2]
}
CV_knr = GridSearchCV(estimator=knr, param_grid=knr_param, cv= 5)
CV_knr.fit(X_train, y_train)
CV_knr.best_params_
```

```
{'leaf_size': 15, 'n_neighbors': 5, 'p': 1}
```

```
#initialise the model
knn = KNeighborsRegressor(leaf_size= 15, n_neighbors= 5, p= 1)
#fit the to the training set
model5 = knn.fit(X_train, y_train.ravel())  #use ravel to avoid warning about 2d array
#predict the model
y_pred = model5.predict(X_test)
#predict model on train data
```

```
svr=LinearSVR(random_state=0)
svr_param = {
    'C': [1,5,10,20,40],
    'epsilon': [0.05,0.1,0.3],
}
CV_svr = GridSearchCV(estimator=svr, param_grid=svr_param, cv= 5)
CV_svr.fit(X_train, y_train)
CV_svr.best_params_
```

```
{'C': 1, 'epsilon': 0.3}
```

```
#initialise the model
svr_model = LinearSVR(C= 1, epsilon= 0.3, random_state=0)
#fit gb model to the training set
model6 = svr_model.fit(X_train, y_train)
#predict the model
y_pred = model6.predict(X_test)
#predict model on train data
```

```
estimators = [('rf', RandomForestRegressor(max_depth=None, max_features= 'auto', n_estimators=200, random_state=0)),
              ('svr', LinearSVR(C= 1, epsilon= 0.3, random_state=0)),
              ('knr', KNeighborsRegressor(leaf_size= 15, n_neighbors= 5, p= 1))]

final_estimator = GradientBoostingRegressor(max_depth= 3, max_features= 'sqrt', n_estimators= 400, random_state=0)
reg = StackingRegressor(
    estimators=estimators,
    final_estimator=final_estimator)
stackmod = reg.fit(X_train, y_train)
#predict the model
y_pred = stackmod.predict(X_test)
#predict model on train data
y_pred_train = stackmod.predict(X_train)
#evaluate
```

```
estimators = [('rf', RandomForestRegressor(max_depth=None, max_features= 'auto', n_estimators=200, random_state=0)),

              ('knr', KNeighborsRegressor(leaf_size= 15, n_neighbors= 5, p= 1))]

final_estimator = GradientBoostingRegressor(max_depth= 3, max_features= 'sqrt', n_estimators= 400, random_state=0)
reg2 = StackingRegressor(
    estimators=estimators,
    final_estimator=final_estimator)
stackmod2 = reg2.fit(X_train, y_train)
#predict the model
y_pred = stackmod2.predict(X_test)
#predict model on train data
y_pred_train = stackmod2.predict(X_train)
#evaluate
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
mape = mean_absolute_percentage_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
#the r2 for the train dataset
r2_train = r2_score(y_train, y_pred_train)

#printing the model evaluation values
print('Mean absolute error: {:.2f}'.format(mae))
print('Mean squared error: {:.2f}'.format(mse))
print('Root mean squared error: {:.2f}'.format(rmse))
print('Mean absolute percentage error:, {:.2f}'.format(mape))
print('r2 score:{: 2f}' format(r2))
```

K