Durban University of Technology

Data Mining and Machine Learning: A Study of the CO2 Emission Trends in South Africa

By

Ghulam Masudh Mohamed

Student Number: 21712862

A dissertation submitted in fulfilment of the requirement for the degree of Master of Information and Communications Technology

Faculty of Accounting and Informatics, Department of Information Technology, Postgraduate Studies

Supervisor: Dr S.S. Patel (PhD)

Co-Supervisor: Prof N. Naicker (PhD)

2024

DECLARATION

I, *Ghulam Masudh Mohamed*, declare that:

(i) The research reported in this dissertation, except where otherwise indicated, is my original research.

(ii) This dissertation has not been submitted for any degree or examination at any other university.

(iii) This dissertation does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

(iv) This dissertation does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

- Their words have been re-written but the general information attributed to them has been referenced.
- Where their exact words have been used, their writing has been placed inside quotation marks, and referenced.

(v) This dissertation does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the dissertation and in the Reference Section of this dissertation.

_[redacted]

Signature: *Ghulam Masudh Mohamed*

Date: 26/07/2024

Approved by Supervisor: [redacted] 27 July 2024

Approved: Co-Supervisor [redacted]

28 July 2024

ACKNOWLEDGEMENTS

I want to express my heartfelt gratitude to everyone who has supported me throughout this research journey. Without their encouragement and help, completing this study would not have been possible. First and foremost, I am thankful to the almighty Allah for granting me the strength and ability to complete this dissertation. To my supervisor, Dr. Patel, I owe a huge thank you for his dedication and guidance. His expertise and patience were invaluable in navigating this study, and I am fortunate to have developed a meaningful friendship with him.

I am also indebted to my co-supervisor, Prof. Naicker, whose experience and insights significantly shaped this study. His mentorship was instrumental in shaping key aspects of my research. I am incredibly grateful to my friends for their unwavering love and support. They have been my constant source of inspiration. Lastly, I want to thank my beloved family. My father, who has been my pillar of strength, found a different way to motivate me each day. My mother's prayers and encouragement have been a source of strength, driving me towards success. My sister, or should I say "Pooch," is my best friend and role model. Her achievements and encouragement have inspired me to push my limits and complete this dissertation. I am deeply thankful to my grandfather or "Dada," whose unwavering support and encouragement have been a constant source of strength throughout this research journey. He instilled in me the belief that I could overcome any challenge.

"Seek knowledge from the cradle to the grave."

~ Prophet Muhammad (Peace be upon him)

"Acquire knowledge and it will earn you life."

~ Ali ibn Abi Talib (R.A.)

ABSTRACT

This study addresses the pressing global issue of elevated carbon dioxide emissions (CO2E), with a particular focus on South Africa (SA), which ranks amongst the world's top emitters and largest in Africa. By introducing a novel integration of Change-point Analysis (CPA) and Machine Learning (ML) techniques, this research addresses significant gaps in CO2E trend analysis. Unlike previous studies, this research applies CPA methodologies within the distinct context of SA, employing algorithms like cumulative sum (CUSUM) and Bootstrap analysis to pinpoint crucial change-points in CO2E data specific to the country. The Bootstrap analysis determines the confidence levels associated with each detected change.

Additionally, this study sought to validate historical trends and predict future patterns using ML models, with a specific focus on employing the AdaBoost ensemble learning technique. Drawing on insights from a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)-based systematic review, the research selects input variables based on the factors identified as significant contributors to CO2E, ensuring the models capture the relevant variables effectively. The results of the systematic review highlight energy production and economic growth as key drivers of CO2E, thus validating their selection as input data for constructing the CPA and ML models. To conduct this study, secondary data was obtained from the World Bank's Open Data initiative data repository, a common source for environmental research. This selection was justified by a literature review, which highlighted the reliability and applicability of this data source. The CPA results reveal significant change-points in electricity generation, economic growth, and CO2E, with an average confidence level of 94%, indicating the accuracy of this analytical approach. Moreover, the CPA results emphasise the relationship between economic growth, electricity production, and CO2E in SA. Before forecasting future CO2E trends, the effectiveness of the AdaBoost regressor in enhancing model performance was benchmarked against traditional ML algorithms, including Linear regression, Polynomial regression, Bayesian Linear regression and K-Nearest Neighbors (KNN) regression, to determine the most effective technique for forecasting CO2E. The researcher evaluated model performance using key regression ML performance metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), coefficient of determination (R^2) score, and an additional accuracy score introduced by the researcher. Notably, the AdaBoost models demonstrated superior performance, with an average RMSE score of 10,143.17 kilotons (kt), MAE score of 9,642.64 kt, R^2 of 0.90, and accuracy of 96.74%.

The study also revealed that, on average, models that were trained using the AdaBoost algorithm surpassed traditional ML models, in terms of performance. They achieved a reduction in RMSE score by 6,417.29 kt, a decrease in MAE score by 4,358.09 kt, an increase in R^2 score by 0.07 and enhanced accuracy by 0.60%.

Additionally, a comparative analysis of the repeated holdout methods and cross-validation techniques was conducted, with results revealing that repeated holdout had a more significant impact on model performance. After excluding outliers, the average improvement in cross-validation results, due to the repeated holdout method, was a decrease of 783.32 kt for RMSE, a reduction of 1,289.39 kt for MAE, and an increase of 0.88% for accuracy. The extent to which the repeated holdout method improved the performance of ML models that were integrated with cross-validation techniques, was correlated with the initial model performance. For ML models with RMSE and MAE scores equal to or exceeding 15,000 kt, the findings indicate that the repeated holdout methods studied should enhance performance by at least 2,000 kt. Similarly, an improvement of nearly 3% or higher in accuracy was noted, when the cross-validation value for this metric was 94% or lower.

The AdaBoost model, integrated with repeated holdout, was selected as the optimal model, as evidenced by the results, for forecasting CO2E in SA from 2021 to 2027. The forecasted CO2E trends validate that energy production and economic growth are indeed the primary drivers of CO2E in SA, as previously highlighted by the CPA model. This underscores the importance of addressing these factors to effectively mitigate carbon emissions in the country. Moreover, the forecasted results indicate that SA is unlikely to meet the global temperature limit of 1.5 degrees Celsius by 2030, given the trajectory showing a shortfall in achieving the target level of 334 million tonnes (Mt) of CO2E, agreed upon in the Paris Agreement. However, the country did meet its CO2E commitments outlined in the 2030 National Development Plan, showing some progress towards environmental sustainability. Nonetheless, the failure to meet these targets at their lower ranges suggests the need for further efforts to reduce carbon emissions, which is crucial for aligning with the Paris Agreement objectives and achieving a zero net emission rate by 2050. This highlights the importance of ongoing initiatives to enhance environmental policies and practices in SA. Future research should focus on integrating load-shedding dynamics into the analysis to examine and confirm its effects on energy production, economic growth, and CO2E in SA. Additionally, future research should focus on forecasting future change-points for the socio-economic indicators or variables utilised in this study. This can help policymakers anticipate fluctuations and devise proactive strategies, to address

environmental and economic challenges effectively. It is also recommended that future research consider the output of renewable energy production, when analysing CO2E trends.

TABLE OF CONTENTS

LIST OF TABLES

Chapter Two: Literature review

Chapter Three: Meta-analysis of related literature

Chapter Four: Research Methodology

Chapter Five: Results, Analysis and Discussion

LIST OF FIGURES

Chapter Three: Meta-analysis of related literature

Chapter Four: Research Methodology

Chapter Five: Results, Analysis and Discussion

LIST OF ABBREVIATIONS

The following are the abbreviations used in this study:

ADF	Augmented Dickey-Fuller
ADHD	Attention-Deficit Hyperactivity Disorder
AI	Artificial Intelligence
AMG	Augmented Mean Group
ANFIS	Adaptive Neuro-Fuzzy Inference Systems
ANN	Artificial Neural Network
AR	Augmented Reality
ARDL	Autoregressive Distributed Lag
BP	British Petroleum
CADF	Combined Augmented Dickey-Fuller
CCEMG	Common Correlated Effects Mean Group
CIPS	Cross-sectionally augmented Im-Pesaran-Shin
CL	Confidence level
CO2E	Carbon dioxide emissions
CPA	Change-point Analysis
CUSUM	Cumulative Sum
DL	Deep Learning
DOLS	Dynamic Ordinary Least Squares
EKC	Environmental Kuznets Curve
ELM	Extreme Learning Machine
EXC	Exclusion Criterion
FDI	Direct Foreign Investment
FEVDM	Forecast Error Variance Decomposition Method
FGLS	Feasible Generalized Least Squares

FMOLS	Fully Modified Ordinary Least Squares
GDP	Gross domestic product
GHG	Greenhouse gases
GI	Gastrointestinal
GIGO	Garbage In, Garbage Out
GWh	Gigawatt hours
IBS	Irritable Bound Syndrome
IPCC	Intergovernmental Panel on Climate Change
IPS	Im, Pesaran, and Shin
IRF	Impulse Response Function
KF	Kalman Filter
KIF	Key impact factor
KNN	K-Nearest Neighbors
KPSS	Kwiatkowski-Phillips-Schmidt-Shin
kt	Kilotons
kWh	Kilowatt hours
LLC	Levin, Lin, and Chu
LMDI	logarithmic mean Divisia index
LSTM	Long Short-term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MLP	Multi-layer Perceptron
MLR	Multiple Linear Regression
MRE	Mean Relative Error
MSE	Mean Squared Error
Mt	Million tonnes

OLS	Ordinary Least Squares
OPEC	Organization of the Petroleum Exporting Countries
PCSE	Panel-Corrected Standard Errors
PD	Parkinson's Disease
PMG	Pooled Mean Group
PP	Phillips-Perron
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
R^2	Coefficient of Determination
RCT	Randomised controlled trials
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
SA	South Africa
SDA	Structural Decomposition Analysis
SNS	Si-Ni-San
SOM	Self-Organizing Maps
SPRT	Sequential Probability Ratio Test
STIRPAT	Stochastic Impacts by Regression on Population, Affluence, and Technology
SVD	Singular Value Decomposition
SVM	Support Vector Machine
SWOR	Sampling Without Replacement
Sys-GMM	System Generalized Method of Moments
VECM	Vector Error Correction Model
WDI	World Development Indicators
WIOD	World input-output database

CHAPTER ONE: INTRODUCTION AND BACKGROUND

1.1 Introduction

This study aims to use data mining and Machine Learning (ML) techniques to analyse carbon dioxide emissions (CO2E) trends in South Africa (SA). According to the European Union Joint Research Centre (2021), high CO2E are a global issue, with China and the United States of America being the largest contributors, producing 11,680 and 4,535 million tonnes (Mt) of CO2E in 2021, respectively. Mashishi (2021) reports that SA is Africa's largest CO2 emitter and the 12th largest in the world. The environmental consequences of high CO2E affect the planet and its inhabitants, making it necessary to find ways to reduce CO2E (Nunez 2019). These consequences, including climate change-related events, can significantly impact economies by disrupting the production of the agricultural sector, straining infrastructure, and increasing healthcare costs, highlighting the interconnectedness between environmental challenges and the economic well-being of a country (Swaminathan and Muraligopal 2017). Due to its ability to identify patterns and gain insights from large datasets, data mining is a viable technique for analysing CO2E trends in SA. Therefore, this study employs Change-point Analysis (CPA) modelling techniques, representing a data mining method, to identify abrupt changes in CO2E. Additionally, it utilises ensemble ML modelling techniques to validate past trends and forecast future emissions. Adopting an experimental research design approach, the study will use secondary data from the World Bank's Open Data initiative data repository to build the models. The final contribution of this research is to provide insights into current and future trends of CO2E in SA and recommend ways of reducing the carbon footprint in the country, with the overall objective to contribute towards the global efforts to reduce CO2E.

1.2 Background

The issue of high CO2E is a global concern, with severe environmental consequences. The increase in CO2E is causing rapid climate change, which is adversely affecting the planet and all its inhabitants (Letete, Guma and Marquard 2010; Osmanski 2020). These emissions have a direct impact on the economy as well as the quality of life of the population, including their health, livelihoods, and the environment (Nunez 2019; Cairoli n.d). Rising CO2 concentrations cause the Earth's surface temperature to increase, which is a significant factor that drives climate change (Cairoli n.d). The Intergovernmental Panel on Climate Change (IPCC) reported that the Earth’s global temperature will increase by approximately 1.5 degrees Celsius within

the next 100 years, with human behaviour being the main cause for such an increase (U.S. Energy Information Administration 2021; Cairoli n.d). This can lead to a higher frequency of natural disasters such as droughts, floods, wildfires and heatwaves (Hansen, Sato and Ruedy 2012; Morse 2018; Banks 2019). In severe cases, these disasters can result in severe injuries and fatalities (Morse 2018).

Furthermore, an upsurge of CO2E presents significant health risks on multiple fronts. As mentioned previously, these emissions contribute to the likelihood of severe injuries and fatalities. Additionally, increased CO2E adversely affects agriculture by depleting soil fertility, mainly due to more frequent droughts caused by elevated CO2 levels. This makes the soil less nutritious to grow crops in, resulting in lower food production and an increased risk of nutritional deficiencies, such as malnutrition (Morse 2018; Osmanski 2020; Cairoli n.d). Also, the National Oceanic and Atmospheric Administration (2021) reports that the scarcity of a fresh water supply, caused by elevated CO2E levels and climate change, can heighten the risk of waterborne diseases, further highlighting the health risks associated with carbon emissions. Heightened CO2E also increases air pollution, causing additional health issues, especially to those individuals with respiratory conditions like asthma (Morse 2018; Osmanski 2020).

From an economic standpoint, natural disasters triggered by elevated CO2 levels can cause significant damage to property owned by individuals and businesses, leaving the victims in financial distress (Osmanski 2020). For instance, the United States of America incurred $145 billion in damages in 2021, due to natural disasters (Clifford 2022). These elevated levels of CO2E, as previously discussed, have adverse effects on food production, significantly impacting local economies that are reliant on agriculture (Morse 2018). Additionally, the resultant climate change influences fishing and hunting activities, forcing animals and sea-life to migrate to more favourable climates, making it more difficult to gain access to them (Banks 2019; Osmanski 2020). In SA, CO2E are a significant concern, with the country being the largest emitter on the African continent and the twelfth largest in the world (Mashishi 2021). This country also suffered from natural disasters induced by high levels of CO2E, with these severe weather instances resulting in R2.7 billion in financial losses in 2021 (Omarjee 2022). These financial losses not only inflict immediate economic strain on the affected regions, but also contribute to broader economic challenges by reducing overall productivity, diverting resources towards recovery efforts, and potentially increasing government expenditure for relief and reconstruction initiatives.

The South African government is also faced with the challenge of balancing economic growth whilst reducing carbon emissions, as this country's economy heavily relies on energy-intensive industries that contribute to high CO2E, such as the manufacturing sector, making it difficult to transition to a low-carbon economy without disrupting economic growth (Ziervogel *et al*. 2014). Therefore, understanding the trends of CO2E in SA and finding ways to reduce emissions are essential to mitigating the harmful effects of climate change and protecting the economy of the country.

The advent of smart solutions aimed at addressing societal issues, including the promotion of environmental sustainability (Javaid *et al.* 2022), has been facilitated by the emergence of the fourth industrial revolution (4IR) in 2011. Currently, the transition into the ongoing development of the fifth industrial revolution (5IR) further propels the advancement of technological solutions (Doyle-Kent and Kopacek 2020). Enabling techniques involves the likes of data mining and ML and their implementation towards developing smart solutions (Aquilani *et al.* 2020; Sarker 2021). Data mining is suitable for analysing CO2E trends because it allows for the extraction of patterns and valuable insights from large datasets (Kwakwa and Adusah-Poku 2020). Given the complexity and volume of environmental data, data mining techniques, such as statistical methods, can uncover hidden patterns and relationships within the emissions data, facilitating a comprehensive understanding of what drives CO2E in a particular context (Fernandez-Basso, Ruiz and Martin-Bautista 2020). CPA serves as the data mining method used in this study and can be employed to identify and analyse rapid changes in CO2E trends (Arif *et al.* 2017). Unlike many data mining techniques that focus on static data, which is time-invariant (Gan *et al.* 2017), CPA is particularly valuable for analysing time-series data, which describes dynamic phenomena (Horváth and Rice 2014). It helps researchers understand when and where significant changes occur in the trends and behaviour of observed phenomena, providing insights into the evolving nature of CO2E over time. This statistical method is ideal for identifying abrupt changes or turning points in CO2E data. The nature of environmental data often involves shifts in emission patterns due to various factors like policy changes, technological advancements, energy-related policies or changes in governance (Kim and Kim 2012; Brown *et al.* 2018). CPA is specifically designed to detect rapid changes in time-series data, such as shifts in measures of tendency like the mean, offering a precise way to identify critical shifts in CO2E over time (Angeyo *et al.* 2016).

As a subset of artificial intelligence (AI), ML has become an indispensable technology due to its capacity to extract decisions and inferences from data (Sarker 2021). This technique can

analyse data without the need for human intervention (Mohri, Rostamizadeh and Talwalkar 2018). Employing a ML model is vital for validating past CO2E trends identified by CPA and forecasting future trends. ML models that focus on regression use-cases or time-series forecasting, leverage historical data patterns to validate past trends and provide accurate predictions for future emissions, considering the intricate environmental dependencies. Therefore, combining the insights of CPA and ML can lead to a comprehensive understanding of the CO2E trends in SA, which in turn can help develop effective strategies that reduce CO2E, aligning with global efforts to reduce carbon emissions.

1.3 Research problem

Understanding the trends in CO2E is critical for identifying the primary factors that drive carbon emissions and for developing effective mitigation strategies (Tutmez 2006). Despite the use of established statistical models like Stochastic Impacts by Regression on Population, Affluence, and Technology (STIRPAT) and Autoregressive Distributed Lag (ARDL) to analyse CO2E trends (Shahbaz, Bhattacharya and Ahmed 2017; Shuai *et al.* 2018; Yuping *et al.* 2021; Sahoo and Sahoo 2022), these models have limitations, such as assumptions about data stationarity and the need for large sample sizes, which can obscure underlying patterns in CO2E trends and lead to ineffective policies (Huang and Wang 2016; Baloch *et al.* 2021; Li and Guo 2022; Nyeadi 2023). The consequences of not addressing these limitations include inaccurately identifying emission drivers and the formulation of policies that fail to effectively curb CO2E (Vélez-Henao, Vivanco and Hernández-Riveros 2019; Lohwasser, Schaffer and Brieden 2020).

To solve this problem, there is a need for innovative empirical approaches that offer more nuanced insights into CO2E trends (Liu, Guo and Xiao 2019; Nguyen, Huynh and Nasir 2021). Specifically, the integration of CPA and ML presents a novel opportunity to capture abrupt changes in CO2E trends and to forecast future trends with greater accuracy (Aminikhanghahi and Cook 2017; Awe and Adepoju 2020; Wu *et al.* 2020; Bakay and Ağbulut 2021; Ağbulut 2022; Tunde *et al.* 2022). This study addresses the gap in current research by applying CPA and ML techniques to model CO2E in South Africa, a context that has been underexplored, and introduces an ensemble ML method for validating past CO2E trends identified by CPA and predicting future trends (Qiao *et al.* 2020; Li *et al.* 2018). This study further advances the body of knowledge by employing a model-building process that utilises insights from a systematic review to guide the selection of input variables for the respective CPA and ML

models developed in this research. By leveraging these methods, this research aims to provide a more accurate and comprehensive analysis of CO2E dynamics, which is essential for effective policy-making and sustainability efforts (Mays *et al.* 2009; Stoddard *et al.* 2021; Tunde *et al.* 2022).

1.4 Aim and objectives

The aim of this study was to analyse the CO2E trends in SA using data mining and ML approaches.

The research objectives (ROs) outlined to fulfil the research aim, are as follows:

[RO1]: Identify the significant contributors of CO2E.

[RO2]: Identify the most relevant time-series datasets that can be used to implement the CPA and ML algorithms.

[RO3]: Examine the CPA algorithm as a data mining technique, to identify the rapid changes in CO2E in SA.

[RO4]: Validate past trends and predict future trends of CO2E in SA by implementing the appropriate ensemble ML algorithm.

[RO5]: Evaluate the effectiveness of the ML models by using the relevant performance metrics.

1.5 Significance of the research

This research makes the following contributions:

- Introduces a novel combination of CPA and ML techniques, for modelling and analysing CO2E trends over time.
- Applies CPA techniques to analyse CO2E trends for the first-time within an isolated South African context.
- Employs the cumulative sum (CUSUM) and bootstrap analysis CPA algorithms, addressing a gap in the limited research on utilising CPA to analyse CO2E trends.
- The first-time use of ML forecasting techniques to validate past trends and forecast future CO2E trends, within the context of SA.
- Contributes to the advancement of environmental research, by employing ensemble ML methods for modelling CO2E trends.
- Utilises findings from a systematic review, to inform the selection of input variables to build the models in this study. The input variables will be based on the main factors that contribute to CO2E, in order to enhance model robustness and efficacy.

1.6 Scope and delimitations of the study

The scope of this study was focused on analysing the main trends associated with CO2E in SA. It is important to note that the study did not aim to comprehensively cover all factors influencing CO2E in the region. Instead, it focused specifically on the major contributors identified through a systematic review of related and curated literature. The study relied on secondary data obtained from the World Bank's Open Data initiative data repository to construct the models. The selected time-series datasets from this repository cover the period from 1970 to 2022, providing a comprehensive temporal perspective for the analysis.

1.7 Research output

The outcomes of this research, particularly the findings of the systematic review presented in Chapter Three of this dissertation, have been submitted to the *Journal Ilmiah Kursor: Research on Computing and its Applications*. This journal is accredited by the Department of Higher Education and Training (DHET). The submitted manuscript, which is currently under review, is titled: "Identifying Factors that Contribute to CO2 Emissions and Trends in Modelling CO2 Emissions Over Time: A Systematic Review."

1.8 Structure of the dissertation

Chapter One: Introduction and background

This chapter presents the study's background, relevance, research aim, objectives problem statement and research problem. In this chapter, the environmental impacts of increased global and South African CO2E are discussed. The integration of CPA and ensemble ML approaches emerges as a suitable solution, to address the research gaps involving the current techniques used to analyse CO2E trends. This chapter also highlights the outcomes of the research.

Chapter Two: Literature review

This chapter reviews existing literature related to the empirical approaches employed for CO2E modelling. These empirical approaches include: statistical models, CPA algorithms, and ML techniques. This chapter concludes by emphasising the identified gap in the analysis of CO2E trends, highlighted through the review of the diverse empirical techniques previously utilised in the literature.

Chapter Three: Meta-analysis of related literature

This chapter presents a systematic review which was conducted to identify the key factors contributing to CO2E and analyse the trends in CO2E modelling over time. Utilising the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology, 19 papers were selected for analysis after multiple levels of literature screening. The search strategy, academic databases, and exclusion criteria were further discussed in this chapter. The findings of the review were presented through various tables and illustrations, to effectively convey the results.

Chapter Four: Research methodology

This chapter presents the research methodology employed in this study, as well as all the materials and methods utilised in conducting the CPA and ML experiments. In the initial part of this chapter, the comprehensive research process conducted in this study was elaborated. The time-series variables or datasets employed for model construction was then discussed, utilising the World Bank's open data repository as the primary source of data. However, an analytical process was conducted to screen and select the final set of time-series variables or datasets, which was also examined in this chapter. This chapter also thoroughly explores the implementation of the CUSUM and Bootstrap algorithms as a CPA technique, including their underlying architecture and mathematical formulas. Additionally, the chapter delves into the training of the AdaBoost ensemble ML regression model, offering insights into its base architecture. The discussion extends to the application of pre-training hyperparameter tuning techniques and model validation approaches, specifically cross-validation and repeated holdout methods. To assess the performance of the regression model, this chapter also conducts a thorough examination of the relevant performance metrics used. In terms of forecasting CO2E in this context, a structured process was followed using the trained AdaBoost model, which was also thoroughly discussed in this chapter. Additionally, this chapter explores the methods employed to validate the past trends produced by the CPA model, through the examination of the future trends predicted by the ML model. In this validation process, graphical illustrations, such as line graphs, were utilised.

Chapter Five: Results, analysis and discussion

This chapter provides a comprehensive analysis of the results obtained from both the CPA and ML experiments within the context of carbon emissions in SA. The findings, trends, and insights related to the CO2E are discussed, with a focus on the outcomes from the CPA

experiments, followed by the ML experiments. In the context of the ML experiments, this chapter extensively discussed and compared the results obtained from both cross-validation and repeated holdout methods, providing insights into the efficacy of these techniques. This chapter also presented a thorough examination and comparison of results between the ensemble model and traditional regression ML models. Furthermore, a rigorous use of certain statistical measures, such as the mean, and statistical visualisations, such as box and whisker plots, quantified the results, to support the discussion and analysis in this chapter. In-depth tables and illustrations were also utilised, to further quantify the results and enable thorough comparisons between the findings of the experiments. The chapter concludes by examining the validation of the past trends generated by the CPA model, through the future trends predicted by the ML model. This was conducted using the validation techniques mentioned in Chapter Four.

Chapter Six: Summary, conclusions and implications of the study

This chapter presents a summary of the study's findings, contributions, implications, limitations, and suggests areas for future research.

1.9 Chapter summary

This chapter introduces the background, significance, research aim, objectives, problem statement, and research problem of the study, along with an overview of the dissertation structure. The subsequent chapter will provide a literature review that examines the current empirical approaches employed in modelling CO2E.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

This chapter provides an extensive review of the empirical methods employed in the literature to model CO2E. Amongst the various approaches utilised, statistical methods are the most common, with models like the Stochastic Impacts by Regression on Population, Affluence, and Technology (STIRPAT), Autoregressive Distributed Lag (ARDL), Kaya identity- logarithmic mean Divisia index (LMDI), System Generalized Method of Moments (Sys-GMM) and Pooled Mean Group (PMG)-ARDL frequently used. To present a comprehensive overview of the statistical methods employed in the literature, this chapter presents a tabular presentation of each study's geographical focus, methodology and limitations. Thereafter, this chapter explores the potential of CPA and ML techniques, as a novel approach to modelling carbon emissions, since these techniques have not been extensively applied in analysing time-series CO2E data. This chapter also addresses limitations in existing literature, identifying research gaps in the empirical approaches used for CO2E modelling. In this context, the integration of CPA and ML analyses emerges as a promising approach, offering a nuanced understanding of CO2E. This chapter concludes by justifying the novel approach selected in this study, to address the research gaps and suggestions identified in existing research.

2.2 Overview of statistical models in CO2E analysis

2.2.1 STIRPAT models

In the research conducted by Shuai *et al.* (2017), the STIRPAT estimation model was used to examine the impact of economic growth, which is commonly measured by gross domestic product (GDP), population growth and energy production on CO2E in 125 countries worldwide. The related time-series data for the indicators under study, were collected from the World Bank's Open Data initiative data repository. These indicators were examined across different income levels, i.e., high-income, upper-middle-income and lower-middle-income, between 1990 and 2011. Statistical tests including: the Levin, Lin, and Chu (LLC) and Im, Pesaran, and Shin (IPS) unit root tests, the Engle-Granger cointegration test, and the Ordinary Least Squares (OLS) cointegration estimation were conducted, to establish the significance of the relationships between these variables. The study found that economic growth and energy production were the most significant factors contributing to global CO2E. The findings revealed that for high-income countries, energy production played the most significant role in

driving CO2E, whilst for lower-middle-income and upper-middle-income countries, economic growth was the primary factor. The findings emphasise the necessity of developing effective policies and strategies to reduce global CO2E. However, the study was limited by the lack of up-to-date data. Like Shuai *et al.* (2017), Dong, Dong and Dong (2019) also employed the STIRPAT model, which was integrated with various statistical techniques to analyse the significant factors that contribute to global and regional CO2E, and also evaluate the efficacy of both renewable and non-renewable energy sources. These techniques include: the Cross-sectionally augmented Im-Pesaran-Shin (CIPS) unit root test, the Westerlund panel cointegration test, and the Dumitrescu and Hurlin panel causality test to account for cross-sectional dependence and slope heterogeneity, and to examine the causality relationships amongst the variables. The study analysed an unbalanced panel dataset of 128 countries across the world, from the period of 1990 to 2014. In order to develop the model, the authors obtained time-series data from the World Bank's Open Data initiative data repository. The results showed that economic growth was the most significant contributor at the global level, followed by population size and non-renewable energy production. However, the key impact factors (KIFs) at the regional level vary across different regions. The study also indicated that renewable energy could lead to a decline in CO2E globally.

A study conducted by Lin *et al.* (2017) made use of the extended STIRPAT model to analyse the global impact of urbanisation and economic development on CO2E, based on panel data collected between 1991 and 2013, in 53 non-high-income countries. The authors have obtained time-series data related to the examined variables, from the WDI database, which is a subset dataset derived from the World Bank's Open Data initiative data repository. These countries were grouped into non-high-income, upper-middle-income, and lower-middle-income categories. The authors intended to provide a global perspective on CO2E, just like the previous two studies (Shuai *et al.* 2017; Dong, Dong and Dong 2019). CO2E were broken down into nine factors, and the results suggested that urbanisation had a minor effect on CO2E in non-high-income countries, with population and economic growth being the primary drivers. The study, however, encountered limitations due to the lack of recent data and the need to examine other variables against CO2E. Additionally, the extended STIRPAT model was supported by several unit root tests, namely the LLC, IPS, Augmented Dickey-Fuller (ADF), Fisher Chi-square, and Phillips-Perron (PP) unit root tests, which ensured the stationarity and suitability of the variables included in the model.

In the study conducted by Shahbaz, Bhattacharya and Ahmed (2017), the authors investigated the relationship between CO2E, economic growth, population growth, and energy consumption in Australia over a 42-year period, from 1970 to 2012. The WDI database provided the time-series data for the variables under study. The findings suggested that energy consumption had a significant impact on CO2E in Australia and that there is a need for household-level energy efficiency measures to promote sustainable population growth. Furthermore, the study identified population growth as the second-most important factor influencing CO2E. To establish causal relationships amongst the variables, the authors employed the STIRPAT model, similarly to the studies conducted by Shuai *et al.* (2017) and Dong, Dong and Dong (2019). This model was implemented in combination with various statistical tests, such as: the ADF, PP and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) unit root tests, the Vector Error Correction Model (VECM) Granger causality test, and the ARDL bounds cointegration test.

As utilised by previous studies (Shahbaz, Bhattacharya and Ahmed 2017; Shuai *et al.* 2017; Dong, Dong and Dong 2019), the paper by Shuai *et al.* (2018) also made use of the STIRPAT model to determine the KIFs of CO2E. However, the study was conducted exclusively within China. The authors conducted a literature review to collect the data required to construct the model. Statistical techniques were used, including: correlation analysis, Pearson correlation analysis and stepwise regression to establish the strength and direction of the relationships between the KIFs and CO2E. Panel-based data were collected from 1995 to 2014, and the KIFs of CO2E in China were identified as economic growth and the urbanisation rate, with economic growth being the most significant contributor to CO2E. However, it should be noted that the study's findings may not be generalizable to other countries, as it only identified the KIFs of CO2E in China, which is the largest carbon emitter globally. Additionally, the lack of current data limited the scope of the study.

2.2.2 ARDL models

Salahuddin *et al.* (2019) conducted a study to examine the impact of urbanisation and globalisation on CO2E in SA, using annual panel data from 1980 to 2017. The associated time-series data for the analysed variables, were obtained from two sources including: the World Bank's Open Data initiative data repository and the KOF Swiss Economic Institute. To test the stationary nature of the variables, the study employed the Zivot and Andrews and Bai and Perron unit root tests. The study then applied the ARDL cointegration test to determine if a long-term relationship existed between the variables and used the Toda-Yamamoto causality

test to check for causality amongst the variables. By utilising the ARDL model and the statistical techniques mentioned above, the study was able to determine short and long-term coefficients. The results indicated that urbanisation contributed to CO2E in both the short and long-term. However, the study recognized the inadequacy of available data and recommended further research to explore the impact of good governance on CO2E. As utilised in the previous study (Salahuddin *et al.* 2019), Yuping *et al.* (2021) also employed an ARDL model to determine the key drivers of CO2E. The effects of renewable and non-renewable energy consumption, globalisation and economic growth on CO2E were examined. However, the research scope strictly focused on examining CO2E trends in Argentina from 1970 to 2018. The authors have collected time-series data related to the variables under study, from the World Bank's Open Data initiative data repository. Zivot Andrews' unit root tests were conducted to ensure data stationarity, whilst the Maki cointegration test was used to determine long-term relationships between CO2E and the examined variables. Gradual shift causality tests were also employed to investigate causality between the variables. The results showed that renewable energy and globalisation mitigated emissions, whilst non-renewable energy contributed significantly towards an increase. The paper was limited by the lack of the latest available data and highlighted the need for future research to analyse other environmental indicators.

Sahoo and Sahoo (2022) implemented the ARDL model, similarly to the studies carried out by Salahuddin *et al.* (2019) and Yuping *et al.* (2021) to model CO2E. However, using the ARDL model, the authors studied the effect of non-renewable and renewable energy consumption on CO2E in India during the period of 1965 to 2018. The WDI and global British Petroleum (BP) databases provided the time-series data needed to develop the model. The results revealed that non-renewable energy sources for energy production were major contributors to CO2E. To test the relationships amongst the variables, the study utilised the ADF, PP, and Zivot-Andrews' unit root tests, the Engel-Granger and Phillips-Qularis cointegration tests, and the Toda-Yamamoto Granger causality test.

2.2.3 Kaya identity-LMDI models

Ma *et al.* (2019) investigated CO2E in China, by developing an enhanced CO2 decomposition model that integrated the LMDI method with the extended Kaya identity. The model was built by utilising time-series data sourced from the China Energy Statistical Yearbook. This model was used to estimate CO2E and its driving factors in regions that consume excessive amounts of energy. The study spanned from 2005 to 2016, revealing that China contributed to roughly

one-third of the total global CO2E, with its CO2E intensity surpassing the world average. Regarding China's CO2E from energy consumption, the industrial sector was identified as the primary contributor, followed by residential consumption. In addition, the transportation sector was identified as another major contributor. The study recommended implementing measures to decrease energy consumption intensity, enhance the internal operating structure of industries, and refine economic policies to promote low-carbon economic development in China. One limitation of the study was that it lacked the use of latest data. The research carried out by Shen *et al.* (2018) had a similar scope to the preceding study (Ma *et al.* 2019), however Beijing was the exclusive geographical area under study, rather than the whole of China. The authors aimed to develop a framework to determine the factors that drive CO2E at different developmental stages in Beijing, between 1995 and 2014. The study also employed the LMDI model, just like the previous study and additionally utilised the Environmental Kuznets Curve (EKC) theory to identify four development stages. The time-series data required to build the model was obtained from the China Energy Statistical Yearbook. Results showed that economic growth was the primary contributor to CO2E, followed by population scale in the second and third stages. However, the analysis was limited to the period between 1995 and 2014, and the findings were specific to Beijing, which may limit their generalisability to other geographical locations.

2.2.4 Sys-GMM models

Nathaniel *et al.* (2021) analysed the relationship between CO2E and economic growth in 15 African economies, using the Feasible Generalized Least Squares (FGLS) and Panel-Corrected Standard Errors (PCSE) static estimators, as well as the Sys-GMM dynamic estimator. The authors have collected time-series data related to the investigated variables from the WDI database. The static estimators were used to examine the impact of energy use on economic growth, whilst the dynamic estimator was used to investigate the causal relationship between CO2E and economic growth. Together, these estimators can provide more robust and accurate estimates of the relationships between the variables of interest. The models were built using panel data that spanned from 1990 to 2014. Results showed that economic growth and energy production from non-renewable energy sources contributed the most to CO2E. The study by Nguyen, Huynh and Nasir (2021) also utilised a Sys-GMM panel data estimator model, similar to the approach conducted by Nathaniel *et al.* (2021). However, this model was used to examine the impact of economic growth, trade openness, financial development, and Direct Foreign Investment (FDI) on CO2E in the economies of the international Group of Six (G6) countries, i.e., Canada, France, the United Kingdom, Italy, Japan, and the United States. The WDI

database provided data spanning from 1978 to 2014, and the results revealed that economic growth and capital market expansion were the primary drivers of CO2E. The study also implemented the LLC and IPS unit root tests, as well as the Fully Modified Ordinary Least Squares (FMOLS) and the Dynamic Ordinary Least Squares (DOLS) correlation estimators. These tests were used to assess the stationarity of the data, estimate the long-run relationship between variables, and correct any bias and inconsistencies in the model, respectively. It should be noted that the findings of the study were limited to the G6 countries, and that further research using other empirical approaches was needed, to model CO2E more comprehensively.

2.2.5 PMG-ARDL models

The purpose of the research conducted by Mensah *et al.* (2019) was to explore the causal links amongst fossil fuel energy consumption, economic growth, CO2E and oil prices in a panel of 22 African countries during the period 1990 to 2015. The researchers have obtained the relevant time-series data for the examined variables from the WDI database. However, the data related to the oil prices was collected from the annual Organization of the Petroleum Exporting Countries (OPEC) statistical bulletin. The study utilised the PMG-ARDL estimation model and employed a combination of statistical techniques, including the ADF and CIPS unit root tests, the Westerlund cointegration test, and the VECM Granger causality test to investigate the long and short-term dynamic relationships amongst the identified variables and the validity of the proposed model. According to the findings, a causal relationship exists between non-renewable energy consumption and economic growth, as well as CO2E, in both the short and long term. Liu *et al.* (2023) studied the influence of economic growth, urbanisation, and energy usage on CO2E in China between 1995 and 2020. The WDI database provided the time-series data related to the indicators under study. The authors implemented the PMG-ARDL estimation model, just like in the previous study, but with a different variation of statistical techniques integrated, such as: the LLC, IPS, ADF, PP and Fisher Chi-square unit root tests to verify the stationarity of the variables. Pedroni, Kao, and Johansen cointegration tests were used to evaluate the long-term relationships between the variables. The results revealed that energy consumption had a significant effect on CO2E both immediately and in the long run, whereas urbanisation had no noticeable impact. Moreover, long-term environmental degradation was a consequence of economic growth.

2.2.6 Additional statistical approaches

Liu, Guo and Xiao (2019) investigated the underlying factors that contributed to the growth of consumption-based greenhouse gas (GHG) emissions in 40 countries around the world. The study made use of the structural decomposition analysis (SDA) model, which was built on data sourced from the world input-output database (WIOD), within the period of 1995 to 2009. The findings generated by this model suggested that the main driving force behind the increase in GHG emissions, was the rapid growth of the global economy, i.e., economic growth. China and India were identified as the main contributors to CO2E growth, highlighting the need for effective policies and strategies to stimulate CO2E reduction on a global scale. However, the study's scope was limited by the lack of up-to-date data, and it was recommended to use other empirical approaches to investigate CO2E in the future. The study by Sharif *et al.* (2019) aimed to examine the relationship between non-renewable and renewable energy consumption and CO2E, utilising panel data from 74 countries across the globe from 1990 to 2015. The WDI database provided the time-series data for the analysed variables. The study employed the FMOLS model approach, which was supplemented by various statistical tests, such as the CIPS unit root test, the Westerlund bootstrap cointegration test, and the Dumitrescu and Hurlin causality test. These tests helped to ensure that the variables were stationary, established long-term equilibrium relationships, and investigated causal relationships between the variables, respectively. The findings indicated that non-renewable energy consumption was a primary cause of CO2E. The study concluded by offering meaningful policy implications for policymakers and governments.

Anwar *et al.* (2022) conducted an accounting approach using the forecast error variance decomposition method (FEVDM) and the impulse response function (IRF) to assess the influence of renewable energy consumption, urbanisation, financial development, agriculture, and economic growth on CO2E. CO2E modelling was conducted on 15 Asian economies during the period of 1990 to 2014, with the time-series data related to the examined variables obtained from the World Bank's Open Data initiative data repository. To examine the long-term relationships amongst the variables, the authors employed the LLC unit root test and the Engle-Granger cointegration test, which were crucial for determining the stationarity of the variables and establishing the existence of a long-term relationship amongst them. The results showed that urbanisation and economic growth led to an increase in CO2E, whilst the consumption of renewable energy reduced CO2E. However, the study did not find any significant impact on CO2E from agriculture. Chen *et al.* (2022) examined the relationship between energy

consumption, urbanisation, economic growth, population growth, and CO2E in the BRICS economies, i.e., Brazil, Russia, India, China, and SA, between 1990 and 2019. The authors have gathered time-series data linked to the variables under study from the WDI database. The study employed the Common Correlated Effects Mean Group (CCEMG) and the Augmented Mean Group (AMG) regression estimators, along with the Combined Augmented Dickey-Fuller (CADF) and CIPS unit root tests, the Westerlund cointegration test, and the Dumitrescu and Hurlin causality test. The methods utilised in the study were chosen to account for cross-sectional dependence and heterogeneous slopes in the data, and to investigate the causal relationships between the variables under investigation. The authors identified energy consumption and economic growth as the main contributors to CO2E, whilst population growth and urbanisation had negligible effects on CO2E. The study has its limitations, as it only had access to data up to 2019, and the findings were limited to the BRICS countries.

Shan *et al.* (2020) performed a study to calculate the most recent CO2E inventories for China and its provinces for the years 2016 and 2017, based on the IPCC Accounting scope model. This model was built using historical time-series CO2E data obtained from the Carbon Emission Accounts & Datasets (CEADs) and China CO2 emission and energy inventory databases, during the period of 1997 to 2015. The inventories revealed that CO2E was primarily generated from energy production and energy consumption for industrial production, across 47 economic sectors. A limitation of the study was that it lacked data for nine of the provinces. *Table 2.1* provides a comprehensive presentation of the statistical models employed in the literature to analyse CO2E trends, based on several criteria, including the country or region, methodology used, and study limitations or constraints.

Table 2.1: Analysis of literature that utilised statistical techniques to model CO2E, based on country, methodology and study constraints

Author/s	**Country**	**Methods**	**Limitations/constraints**
(Lin *et al.* 2017)	53 countries worldwide: 13 European countries, seven North American countries, 16 Asian Countries, seven South American countries and 10 African countries	Extended STIRPAT model with the following unit root tests: - LLC - IPS - ADF - Fisher Chi-square - PP	- Lack of latest data - The need to study the impact of other variables towards CO2E
(Shuai *et al.* 2017)	125 countries worldwide	STIRPAT model with: - LLC & IPS unit root tests	Lack of latest data

		- Engle Granger cointegration tests - OLS cointegration estimation	
(Shahbaz, Bhattacharya and Ahmed 2017)	Australia	STIRPAT model with: - ADF, PP and KPSS unit root tests - ARDL bounds test cointegration analysis - VECM Granger causality tests	Findings limited to Australia
(Shuai *et al.* 2018)	China	STIRPAT model with: - Correlation Analysis - Pearson Correlation - Stepwise regression	- Lack of latest data - Findings are pertinent to China only
(Shen *et al.* 2018)	Beijing (China)	- Kaya-LMDI decomposition model - EKC theory to identify CO2E at different income levels	- Findings pertinent only in Beijing - Limited data
(Ma *et al.* 2019)	China	Kaya-LMDI decomposition model	Lack of latest data
(Liu, Guo and Xiao 2019)	40 countries worldwide: seven Asian countries, three North American countries, one Oceanic country, one South American country, 27 European countries and one entity to represent the rest of the world	SDA model	- Lack of latest data - Need for the use of other empirical approaches
(Dong, Dong and Dong 2019)	128 countries worldwide	STIRPAT model with: - CIPS - Westerlund panel cointegration tests - Dumitrescu and Hurlin panel causality test	Lack of latest data
(Sharif *et al.* 2019)	74 countries worldwide	FMOLS approach with: - CIPS unit root tests - Westerlund cointegration tests - Dumitrescu and Hurlin causality tests	Lack of latest data

(Mensah *et al.* 2019)	22 African countries	PMG-ARDL estimation model with: - ADF and CIPS unit root tests - Westerlund cointegration tests - VECM Granger causality tests	Findings limited to the African continent
(Salahuddin *et al.* 2019)	South Africa	ARDL model with: - Zivot Andrews' and Bai and Perron unit root tests - ARDL bounds cointegration tests - Toda-Yamamato causality tests	- Lack of latest data - The need to investigate the influence of good governance on CO2E
(Shan *et al.* 2020)	China	IPCC Accounting Scope	Lack of data for nine of the provinces
(Nathaniel *et al.* 2021)	15 African countries: Algeria, Cameroon, Egypt, Ethiopia, Sudan, Gabon, Togo, Kenya, South Africa, Morocco, Senegal, Tunisia, Ghana, Nigeria, and Congo Republic	- Static estimators: FGLS and PCSE models - Dynamic estimators: Sys-GMM	Lack of latest data
(Nguyen, Huynh and Nasir 2021)	G6 countries: Italy, France, Japan, Canada, United Kingdom and the United States	Sys-GMM panel data estimation with: - IPS and LLC unit root tests - FMOLS & DOLS correlation estimators	- Findings only limited to six countries - Need for the use of other empirical approaches
(Yuping *et al.* 2021)	Argentina	ARDL model with: - Zivot Andrews' unit root tests - Maki cointegration tests - Gradual shift causality tests	- Lack of availability for latest data - Need for analysis to be conducted on other indicators of environmental quality
(Anwar *et al.* 2022)	15 Asian countries	Accounting approach, FEVDM and IRF with: - LLC unit root tests - Engle-Granger cointegration tests	Findings limited to the Asian continent
(Sahoo and Sahoo 2022)	India	ARDL approach with: - ADF, PP and Zivot-Andrews' unit root tests	Findings limited to India

		- Engel Granger and Phillips-Qularis cointegration tests - Toda-Yamamoto Granger causality tests	
(Chen *et al.* 2022)	BRICS countries: Brazil, Russia, India, China, and South Africa	CCEMG and AMG regression estimators with: - CADF and CIPS unit root tests - Westerlund cointegration tests - Dumitrescu and Hurlin causality tests	- Lack of latest data - Findings only limited to the BRICS countries
(Liu *et al.* 2023)	China	PMG-ARDL estimation model with: - LLC, IPS, ADF, PP and Fisher Chi-square unit root tests - Pedroni, Kao, and Johansen cointegration tests	Findings constrained to China

This table presents a clear overview of the research characteristics, in order to gain insights into the trends and patterns in the literature. The reviewed studies examined CO2E on a global scale, with some examining worldwide CO2E whilst others focused on specific countries, regions, or continents. China, India, and the United States, which are the top three countries in the world that produce the most CO2 (Blokhin, Smith and Perez 2021), were frequently studied, as demonstrated in the studies by Shuai *et al.* (2017), Liu, Guo and Xiao (2019), and Dong, Dong and Dong (2019). Similarly, SA, the highest CO2-emitting country in Africa (Blokhin, Smith and Perez 2021), was a commonly modelled country on this continent, as depicted in the research conducted by Salahuddin *et al.* (2019), Nathaniel *et al.* (2021), and Chen *et al.* (2022). Studying high-CO2-emitting countries provides insights into the factors that contribute to CO2E and how they change over time. This is because these nations often have diverse industrial, economic, and energy profiles, reflecting a range of contributing factors to CO2E. By examining these countries, researchers can identify patterns, trends, and influential variables that contribute to CO2E, helping to understand the dynamics and changes in emissions over time.

This table also includes the various statistical models that were used across the CO2E research, making it possible to identify any gaps where additional analytical approaches can be used to model carbon emissions. Furthermore, the presentation of the study limitations or constraints in this table provides a transparent evaluation of each study's findings, highlighting areas for potential improvement and future research.

2.3 Overview of CPA methods in CO2E analysis

Regarding the CPA models utilised in research, Awe and Adepoju (2020) employed a recursive Bayesian algorithm to investigate change-points in the relationship between CO2E and energy production in Nigeria and SA, during the period of 1970 to 2010. The results revealed a significant shift in the association between energy production and CO2E in Nigeria, with the model's slope demonstrating its lowest point in the late 1980s and its highest in the early 2000s. For SA, the model showed a similar pattern, with the lowest slope trajectory in the late 1980s to mid-1990s and the steepest incline around the early 2000s. These change-points can be attributed to several factors, including population growth, economic development, industrial expansion, changes in governance, human activities, and vehicle emissions in these nations. The data used to build the model, was sourced from the WDI database. Tunde *et al.* (2022) also studied CO2E in Nigeria, utilising the Man-Kendall approach to investigate the trends between the transportation sector and CO2E from 1971 to 2014. Block bootstrapping techniques were integrated, to help identify abrupt changes in the time-series data. The analysis aimed to identify any monotonic patterns, determine their direction and magnitude, and pinpoint potential change-points in the dataset. However, no distinct breaks were found in transportation-related emissions. The study's key finding suggests that CO2E across various sectors have consistently increased, posing significant environmental and health risks. This highlights the importance of discouraging activities contributing to this major GHG. Furthermore, the data for the study was also sourced from the WDI database.

The study by Wang *et al.* (2017) introduced a method for calculating cumulative CO2E from industrial energy consumption data in three regions of Eastern China i.e., Shanghai, Jiangsu, and Zhejiang, for the years 1995 to 2014. The authors employed the Hodrick-Prescott filter to analyse CO2E fluctuations, as well as a grey correlation-based change-point detection algorithm to find key shifts in the data. Despite varying industrial structures, these regions maintain consistent cumulative CO2E cycles during this period, according to the study. These emission patterns correspond to different stages of energy policies, highlighting their persistent

role in emission regulation. The research aims to promote low-carbon production and sustainable energy development, and it suggests refining energy policies based on energy consumption and CO2E data. The data for the study was primarily sourced from the China Energy Statistical Yearbook covering energy consumption data from 1995 to 2014.

2.4 Overview of ML techniques in CO2E analysis and forecasting

A hybrid ML approach known as the (KLS) method was introduced by Li (2020), which combined time-series prediction and variable regression techniques to forecast CO2E in China. This method combines the Kalman filter (KF) with the long short-term memory (LSTM) and SVM ML algorithms. The LSTM algorithm predicts CO2E as a time-series output result, whilst the SVM algorithm makes use of 10 selected variables, determined by ridge regression, to regress CO2E, and thereafter the KF integrates the results of both algorithms. This hybrid model was trained on data from 1965 to 2005 and its prediction performance was tested on data from 2006 to 2014, with the results demonstrating the efficacy of integrating both algorithms, rather than training individual models. The study also predicts China's CO2E for 2030, projecting a peak in 2024, followed by a gradual decrease. The data for the models was sourced from the China Statistical Yearbook and the World Bank data repository. This novel approach achieved a Mean Squared Error (MSE) score of 0.0039 and a Mean Absolute Error (MAE) score of 0.061. Li *et al.* (2018) also trained an adaptation of the SVM algorithm, called the SVM-Extreme Learning Machine (ELM) model, to forecast energy consumption and CO2E trends in Beijing. The authors conducted a comprehensive analysis of major carbon emissions sources, such as the usage of petrol, coal and natural gas, to generate power in the region. The SVM kernel function was employed to enhance the weight values in the ELM algorithm's hidden layer. The Grey prediction theory was then utilised to forecast the regions' energy consumption from 2017 to 2030. Thereafter, the SVM-ELM model was trained on CO2E and energy consumption data from 2000 to 2016 and demonstrated superior accuracy compared to the normal SVM algorithm. By using the predicted results of the Grey Prediction Theory (1, 1) algorithm as input, the authors were able to predict CO2E in the region up to 2030, highlighting the substantial impact energy consumption has on CO2E. In addition, the data in the study was sourced from the China Energy Consumptions Statistics Yearbook. The results of the ELM-SVM model indicated optimum performance, with a Root Mean Squared Error (RMSE) score of 0.1234, a coefficient of determination (R^2) score of 0.9978 and a Mean Relative Error (MRE)

score of 1.62. Whilst the forecasting model showed high accuracy, the authors suggested the exploration of new techniques to enhance the performance of ML models.

Qiao *et al.* (2020) also adapted the SVM model to forecast emission trends in 12 countries, including SA, from 2018 to 2025. The study proposed merging the genetic algorithm and the lion swarm optimizer to improve the least squares SVM model, in order to predict CO2E. The model was built on CO2E data belonging to the 12 countries under study, within the period of 1965 to 2017. Using the model for forecasting emissions from 2018 to 2025, the average MAE score was 37.30 Mt, and the Mean Absolute Percentage Error (MAPE) score was 0.483%. For SA, the MAE score was 7.81 Mt, and the MAPE score was 0.365%. The data to train the model was sourced from the "67th Statistical Review of World Energy."

The study by Ağbulut (2022) also trained a SVM model, in addition to a Deep Learning (DL) Artificial neural network (ANN) model to forecast CO2E trends in Turkey's transportation sector. The GDP per capita, population, energy demand resulting from transportation sector activities, vehicle kilometres and year of manufacture, served as the input parameters for the models. Strong correlations were observed amongst these variables, with the results of the study indicating high prediction accuracy. For all ML algorithms employed, R^2 scores ranged between 0.8639 and 0.9235. In addition, the study projected a considerable increase in energy demand and CO2E, with a 3.7% and 3.65% increase respectively, in Turkey's transportation sector by 2050, suggesting the need for revised energy investments and policy measures. The dataset covers the years from 1970 to 2016, with data sources from the World Bank open data repository, Turkish Statistical Institute, and Turkish General Directorate of Highways.

Mardani *et al.* (2020) employed ML and DL techniques to analyse the relationships between CO2E, energy consumption and economic growth in G20 countries during the period of 1962 to 2016. The approach includes Singular Value Decomposition (SVD) for data prediction, Self-Organizing Maps (SOM) for data clustering, and ANN and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) for CO2E prediction. Evaluating real-world data from the WDI dataset, the method demonstrated good predictive performance, achieving a MAE score of 0.065. Future research is encouraged to integrate ML with other empirical methods and explore additional variables beyond energy production and economic growth in G20 nations. The data for the examined variables was obtained from the WDI database.

The study by Hosseini *et al.* (2019) predicted Iran's future CO2E and their adherence to the Paris Agreement by the year 2030. This agreement obliges Iran to decrease its GHG emissions within a specified time frame. Using ML and DL techniques, the Multiple linear regression (MLR) and Multi-layer perceptron (MLP) models were trained to forecast annual CO2E based on the following time-series variables: CO2 intensity, population, non-renewable energy production, per capita energy and per capita GDP. Both models exhibit R^2 scores above 0.99, with the MLR model demonstrating slightly lower prediction performance than the MLP model. The paper outlines two scenarios for projecting Iran's CO2E from 2015 to 2030, drawing on historical data from 1975 to 2015, which was obtained from the World Bank's Open Data repository. The findings indicate that Iran will struggle to comply with the international accords, highlighting the need to revise its energy structure to enhance sustainability. In terms of future research, the authors proposed exploring additional methods to improve the efficiency of ML techniques.

2.5 Related Work

The objective of the research by Xia *et al.* (2024) was to examine the temporal fluctuation characteristics of Hefei's fine particle matter (PM2.5) concentration and the variables that influenced it between 2013 and 2021. The study used mixed frequency data on socio-economic development and air pollution to do this. An *LGBNN* technique was formed by combining a multi-layer neural network model with the Light Gradient Boosting Machine (LightGBM) ensemble learning algorithm and an adaptive change-point detection (ACPD) model, based on a self-encoding-decoding network. Significant change-points in PM2.5 concentration were identified using the ACPD model, which also revealed periodic oscillation patterns and phased properties. The importance of policy considerations was highlighted by feature selection analysis using the LGBNN model, which was followed by air pollution, socio-economic development and meteorological factors. Additionally, the LGBNN model showed exceptional prediction accuracy and the capacity to represent intricate variable interactions.

Jiang, Zhu and Tian (2023) conducted a study to enhance PM2.5 concentration interval prediction whilst tackling the issue of high volatility and uncertainty in pollutant time-series data. A refined model for interval prediction was therefore built. The PM2.5 time-series was broken down by the model into two separate parts: a trend term and a fluctuation term. The ELM method was used to forecast the trend term, which showed the long-term evolution of PM2.5 levels. Meanwhile, the Lower Upper Bound Estimation (LUBE) interval prediction

approach was utilised to analyse the fluctuation term, which captured the short-term fluctuations and irregularities in the PM2.5 data, in order to offer a more thorough knowledge of the variability within the time-series. For improved efficiency and stability, LUBE was enhanced using an Interval Perturbation-based Adjustment Strategy (IPAS) and the iterated cumulative sums of squares (ICSS) change-point detection technique. Maximal Information Coefficient (MIC) and Partial Autocorrelation Function (PACF) were used in the model's feature selection process to combine meteorological data with other air pollution indicators. By integrating the trend and fluctuation term forecasts, the final prediction intervals (PIs) were formed. The model demonstrated better performance and stability than conventional models when verified using daily PM2.5 data from Wuhan, China.

The study by Martinez *et al.* (2023) employed a recurrent neural network (RNN) based on a LSTM architecture to separate anomalous from normal diffusion in single-particle trajectories. Using the transition density function (TDF), the model precisely located change-points in the mean-squared displacement (MSD) curve that resemble the Einstein linear regime. The self-diffusion coefficient was supported by this segmentation method, which also forecasted ephemeral behaviour impacted by chemical and physical variables, like salt concentrations. The study highlighted the model's capacity to produce consistent change-point predictions by showcasing its accuracy in a variety of synthetic and experimental environments. The use of window average filtering improved the model's performance despite certain drawbacks, such as oscillations close to change-points.

The research by Gupta, Wadhvani and Rasool (2022) focused on improving the detection of change-points in time-series data to facilitate real-time applications. A three-phase design that combines Change-Point Detection (CPD) with adaptive pre-processing was demonstrated in the suggested technique. In order to smooth the data and minimise noise and outliers, recursive singular spectrum analysis was used during the pre-processing stage. Through the analysis of reconstruction loss errors, a deep learning-based autoencoder neural network was employed in the CPD phase to identify change-points. This combination method enhanced the ability to identify changes in auto-correlation, mean, variance, and frequency. Because the data was pre-processed and analysed in real-time, the system surpassed current CPD techniques and produced change-point identification that were more precise and efficient.

2.6 Contribution from this research

Statistical models have frequently been employed to model CO2E over time, with the STIRPAT (Shahbaz, Bhattacharya and Ahmed 2017; Shuai *et al.* 2018; Dong, Dong and Dong 2019) and ARDL models (Yuping *et al.* 2021; Sahoo and Sahoo 2022) being amongst the most utilised in research. In SA, the STIRPAT (Lin *et al.* 2017; Shuai *et al.* 2017; Abelha *et al.* 2020; Yılmaz and Şensoy 2022) and ARDL (Salahuddin *et al.* 2019; Iorember *et al.* 2021) models are frequently employed, with a recent study by Mensah *et al.* (2019) expanding upon the ARDL model, by utilising the PMG-ARDL model. The Sys-GMM model was another model adopted in the literature by Nathaniel *et al.* (2021), to analyse CO2E trends in SA.

However, there is still a need to employ other empirical approaches in addition to these models (Liu, Guo and Xiao 2019; Nguyen, Huynh and Nasir 2021). This is important because diverse empirical approaches can offer complementary perspectives, helping to validate findings and enhance the robustness of conclusions (Mays *et al.* 2009). Additionally, new empirical approaches might capture nuances or trends that previous methods might overlook. By capturing nuances and trends overlooked by traditional methods, these approaches can contribute to more informed and effective mitigation strategies, to reduce CO2E (Stoddard *et al.* 2021). Therefore, this research has introduced a novel approach of combining the insights derived from both CPA and ML techniques. This combination aims to leverage the strengths of both methods, offering a more comprehensive and accurate understanding of CO2E dynamics.

Whilst the STIRPAT model, ARDL regression model, LMDI model, Sys-GMM model and PMG-ARDL estimation model have significantly contributed to the understanding of CO2E trends, they are not without their limitations. These models often require large sample sizes and make certain assumptions about stationarity or continuity within data series, which can sometimes obscure actual patterns in CO2E (Huang and Wang 2016; Baloch *et al.* 2021; Li and Guo 2022; Nyeadi 2023). Moreover, certain models, like the STIRPAT model, assume linear relationships between variables under specific constraints, which are not true in all instances (Vélez-Henao, Vivanco and Hernández-Riveros 2019; Lohwasser, Schaffer and Brieden 2020), potentially oversimplifying the complex interactions that drive CO2E trends. Given the intricate and multifaceted nature of CO2E drivers, models that tend to oversimplify relationships hinder their ability to provide significant and reliable insights into the complexities of emission trends. If the interrelationships amongst the variables are not properly addressed, a model might attribute the changes in CO2E to specific factors incorrectly, resulting

in distorted conclusions about the primary drivers of CO2E. This can lead to the formulation of ineffective policies, as the main drivers of emissions are misrepresented. This, in turn, hampers efforts to implement targeted and impactful interventions to curb CO2E.

As such, the integration of CPA and ML has emerged as a promising approach to gain insights into CO2E trends. CPA identifies abrupt changes in time-series data, which is crucial in understanding the trends of CO2E in SA (Aminikhanghahi and Cook 2017). This approach can detect structural breaks in time-series data without needing prior knowledge about the number or location of these breaks (Monyeki, Naicker and Obagbuwa 2020). It allows researchers to uncover shifts in CO2E patterns that other methods might miss due to their assumptions about stationarity or continuity within the data series (Baltagi, Kao and Liu 2017). Unlike other techniques, CPA does not rely heavily on large sample sizes making it more feasible for smaller-scale studies (Xie, Li and Xiong 2014). Therefore, this model offers a more nuanced view of carbon emissions dynamics, addressing the limitations of previous methods and enhancing the accuracy of trend analysis in a rapidly changing context (Killick and Eckley 2014).

ML is an effective tool for validating past trends and accurately predicting future trends, enabling a full understanding of the key CO2 emitters in the country, as ML algorithms can learn and adapt to changing patterns over time (Bakay and Ağbulut 2021). Combining CPA and ML to analyse CO2E trends can provide a more accurate and comprehensive understanding of the multifaceted nature of CO2E in SA. Moreover, innovative empirical techniques have the potential to capture trends that might be overlooked by previous methods. This approach allows for a more robust and data-driven analysis of CO2E trends, which is critical for effective policy-making and decision-making towards achieving sustainability goals (Tunde *et al.* 2022).

Although CPA has been used to analyse CO2E trends (Awe and Adepoju 2020; James and Menzies 2022; Tunde *et al.* 2022), it has not been extensively researched, and there is a lack of application in an isolated South African context. This gap led to the selection of the cumulative sum (CUSUM) algorithm and bootstrap analysis for this study, emphasising the lack of CPA algorithms utilised in this specific context. Furthermore, to the best of current knowledge, there has not been extensive utilisation of ML techniques to forecast CO2E trends in SA, with Qiao *et al.* (2020) building a Support vector machine (SVM) model to forecast carbon emissions in 12 countries throughout the world. Whilst Köne and Büke (2010) utilised regression analysis to forecast the CO2E in the top 25 emitting countries in the world. Wu *et*

al. (2015) and Wu *et al.* (2020) on the other hand, employed grey forecasting models, to forecast carbon emissions in the BRICS countries, namely Brazil, Russia, India, China, and South Africa.

Notwithstanding the use of ML techniques to forecast CO2E trends in global research (Li *et al.* 2018; Ağbulut 2022), an ensemble method has not been frequently used. This approach could potentially improve the performance of previous models, as Li *et al.* (2018) and Hosseini *et al.* (2019) both suggested the exploration of new empirical techniques to enhance model performance. Therefore, this research aims to address the gap in existing studies that analyse trends in CO2E, by using a novel combination of CPA and ML to provide a more accurate and comprehensive analysis. This includes the first-time use of CPA and ML forecasting in an isolated South African context, as well as the implementation of a novel ensemble ML method for validating past trends and forecasting future trends.

The novelty of integrating CPA with ML techniques, has been further validated through a review of extended literature. Previous research has highlighted the effectiveness of combining CPA and ML to analyse various socio-economic and environmental indicators (Gupta, Wadhvani and Rasool 2022; Jiang, Zhu and Tian 2023; Martinez *et al.* 2023; Xia *et al.* 2024). However, these studies have predominantly focused on different environmental variables and contexts, often neglecting the in-depth modelling of CO2E. In contrast, this study uniquely concentrates on quantifying CO2E and examining the specific factors influencing these emissions within a defined context. By employing CPA to identify significant past trends in CO2E and leveraging regression-based ML algorithms to validate these past trends and forecast future trends, this study not only extends the application of CPA and ML, but also addresses a gap in the literature where CO2E has not been extensively modelled. Consequently, the findings of this study not only confirm the validity of the CO2E trend analysis, but also establish a novel and comprehensive framework for future research in this area, setting a precedent for the combined use of CPA and ML to explore CO2E trends and related factors. In this regard, to the best of knowledge, the use of CPA to identify historical CO2E trends, followed by the application of ML techniques to validate these trends and forecast future patterns, has not been extensively explored in existing literature.

Furthermore, this study will use the findings from the systematic review conducted by the author, which is presented in Chapter Three of this dissertation, which identified the major factors driving CO2E. The review will inform the selection of input variables for the models

developed in this research. Accurately identifying the primary drivers of CO2E is crucial for building effective models that can inform policy and decision-making (Reich 2010). Incorporating these factors into the models, will enhance the accuracy and effectiveness of the analysis in this research. By accurately capturing the influence of these key emission factors, the impact of the specific variables on CO2E can then be analysed (Debone, Leite and Miraglia 2021). In terms of the ensemble ML model, utilising the significant factors of CO2E as input features helps in discerning relationships between the features and the CO2 predicted in the regression model. This knowledge enables the model to precisely capture the impact of each variable or factor on CO2E. The approach of justifying the selection of specific variables in the model-building process for CO2E modelling has not been thoroughly explored in the existing literature.

The identification of the main factors driving CO2E and their relative importance is crucial for effective policy and decision-making, to reduce carbon emissions (Reich 2010). By focusing on the significant contributors, it is possible to develop targeted strategies that have a greater impact on reducing the effects of high CO2E on the environment (Kebede 2017; Debone, Leite and Miraglia 2021). Additionally, this review will help to understand the existing methods used to model CO2E over time, which can inform the development of future interventions. This is achieved by analysing the strengths and limitations of existing techniques, which helps to identify future research gaps.

2.7 Chapter summary

This chapter examined various empirical approaches that were used in the literature, namely statistical models, CPA methods and ML techniques that were employed to analyse CO2E trends. The existing literature on CO2E modelling reveals that statistical models and techniques were commonly utilised, to investigate the factors influencing CO2E in different countries and regions. The most frequently implemented statistical model was the STIRPAT regression model, which has been utilised in multiple studies (Shahbaz, Bhattacharya and Ahmed 2017; Shuai *et al.* 2017; Shuai *et al.* 2018; Dong, Dong and Dong 2019). Additionally, Lin *et al.* (2017) adapted the STIRPAT model to create an extended version that incorporates additional variables that extend the existing factors utilised by STIRPAT, such as political and cultural factors, to understand the intricate link between human activities and their impact on the environment. (Wu *et al.* 2021). Similarly, the ARDL regression model has also emerged as a

prevalent choice, as evident by the research undertaken by Salahuddin *et al*. (2019), Yuping *et al.* (2021), and Sahoo and Sahoo (2022). In addition to the two previously mentioned models, the Kaya identity – LMDI statistical model (Shen *et al.* 2018; Ma *et al.* 2019), the Sys-GMM data model (Nathaniel *et al.* 2021; Nguyen, Huynh and Nasir 2021), and the PMG-ARDL model (Mensah *et al.* 2019; Liu *et al.* 2023) have also been considered by researchers.

Although numerous studies have utilised statistical models to analyse CO2E trends, they have encountered various constraints. Some studies were limited by the lack of the latest data available, as reflected in the research by Lin *et al.* (2017) and Shuai *et al.* (2018). The challenge of limited access to the latest data can significantly impact the accuracy and comprehensiveness of CO2E analysis. Without up-to-date data, the trends and patterns in CO2E may not accurately reflect the current state of affairs. Outdated data could result in misleading conclusions and hinder the ability to identify real-time changes and developments in emission trends. This could result in decision-makers and policymakers developing mitigation strategies based on outdated information, potentially leading to ineffective or inappropriate actions to address CO2E. The need for deeper investigations into the roles of additional variables or indicators in affecting the environment was also identified (Lin *et al.* 2017; Salahuddin *et al*. 2019; Yuping *et al.* 2021).

Contextual limitations also arose, as certain studies focused on specific regions which makes the findings not universally applicable (Shuai *et al.* 2018; Chen *et al.* 2022). This contextual limitation can limit the generalisability of a study's conclusions and insights. It becomes challenging to draw broader conclusions about CO2E trends when the analysis is constrained to a specific geographical area. This can hinder the development of comprehensive strategies and policies that address CO2E at a global scale and across different contexts. Different regions have unique social, environmental and economic characteristics that influence CO2E patterns. Strategies that work effectively in one region may not be suitable for another due to varying population densities, energy sources and industries (Meng *et al.* 2011). This reinforces the significance of the South African contextual study presented in this dissertation. The next chapter will present a meta-analysis and systematic review of carbon emissions research, aiming to identify the significant factors contributing to CO2E and examine the trends in modelling CO2E over time.

CHAPTER THREE: META-ANALYSIS OF RELATED LITERATURE

3.1 Introduction

This chapter presents a systematic review that aims to: (i) identify the major contributors to CO2E and (ii) analyse the trends in modelling CO2E over time. Understanding the primary drivers of CO2E is crucial for building accurate models that can inform policy and decision-making (Reich 2010; Liao *et al.* 2019). Models that accurately incorporate these drivers can help policymakers and stakeholders develop targeted strategies that focus on mitigating the impact of the main drivers on overall CO2E, in order to address climate change and achieve sustainability goals (Liao *et al.* 2019; Debone, Leite and Miraglia 2021).

The emphasis on identifying the main factors influencing CO2E in this systematic review is crucial in fulfilling RO1 of this study, which aims to determine the significant contributors to CO2E. By synthesising the findings from the review, it makes it possible to gain insights into the key drivers that have been recognized in previous studies. This knowledge informs and guides the model-building process in this study. The implementation of the CPA model and ensemble ML model, will therefore be specifically tailored to capture and analyse the identified factors, enhancing the precision and effectiveness of the models. This implies that the main factors of CO2E will serve as input features for both models. In the ensemble ML model, using key CO2E factors as input features, makes it possible to accurately capture the relationship of each feature, with the predicted CO2E in the regression model, emphasising the influence of each input variable on carbon emissions. This factor-driven model building allows for a more targeted exploration and modelling of these emission factors, within the context of SA.

Furthermore, examining trends in modelling CO2E over time provides insight into current models' effectiveness and identifies the need for future interventions (Zhu, Duan and Fan 2015; Le Quéré *et al.* 2020). However, a comprehensive analysis of each selected paper, including a detailed examination of their methodologies, was already presented in the previous chapter, which focused on the literature review. This chapter focuses primarily on identifying the significant drivers of CO2E and any trends in carbon emissions research.

This systematic review, conducted within the framework of PRISMA, aimed to identify the factors that contribute to CO2E and analyse the trends in carbon emissions modelling from 2017 to 2023. Out of the 61,903 articles identified, 19 were selected for analysis after applying the PRISMA exclusion criteria. This chapter further elaborates on the implementation of the

PRISMA methodology, emphasising the following steps: the search strategy, the academic database search terms for literature retrieval, and the exclusion criteria employed to remove irrelevant papers from the search results.

The chapter also includes a meta-analysis which is primarily focused on examining the general characteristics of the selected studies, including the publication year, authors, research scope, study periods, and time-series CO2E datasets. In the subsequent sections, meta-analysis data are presented alongside graphical representations that illustrate publications by year, publications by journal index and time-series CO2E data sources, amongst others.

3.2 Utilising PRISMA guidelines for systematic reviews

The PRISMA methodology has been utilised across the literature to enhance the quality of systematic reviews, in various fields and disciplines. This methodology is a tool used in systematic reviews, to screen and select articles from large academic databases, based on pre-defined inclusion criteria, thereby refining search results to a select few articles for analysis and review (Sarkis-Onofre *et al.* 2021).

In the health sciences, particularly the Psychiatry field, a systematic review was conducted by Şalvarlı and Griffiths (2021) utilising PRISMA guidelines to identify personality traits associated with gaming disorders. The review of 21 selected articles extracted by PRISMA identified 24 distinct traits linked to gaming disorders. In a systematic review using PRISMA, Strahler Rivero *et al.* (2015) investigated video gaming as a treatment for Attention-deficit hyperactivity disorder (ADHD) patients. After applying inclusion criteria, 14 papers were identified for review, and the findings suggested that this treatment can enhance the cognitive abilities of patients. In the pulmonary field, Hasani *et al.* (2020) conducted a PRISMA-aligned systematic review to identify the most common symptoms amongst COVID-19 patients. After analysing 30 selected articles, fever was identified as the most widespread symptom. In the systematic review by Quarato *et al.* (2017), the PRISMA method was followed to determine the effects of air pollution on public health. Based on the findings from the 15 screened papers, exposure to air pollutants increases the risk of contracting pulmonary and allergic illnesses.

When exploring other health sciences fields, a systematic review by Lee and Lim (2017) utilised PRISMA to investigate acupuncture as a treatment for individuals suffering from Parkinson's disease (PD). Following the screening of 14 papers for review, the findings indicated that this treatment is more effective than standard treatment alone. Ling *et al.* (2015)

evaluated the effectiveness of traditional Chinese medicine using the Si-Ni-San (SNS) medical formula on gastrointestinal (GI) diseases, such as Irritable bound syndrome (IBS). The systematic review driven by PRISMA, identified 83 articles for review, with medication derived from this formula providing some relief to patients. Yasin *et al.* (2020) performed a systematic review related to health workers. The authors utilised PRISMA to identify the factors influencing job satisfaction amongst nurses. The review screened 38 papers for analysis based on the inclusion criteria, with the workplace environment and employer type being the most cited factors.

In the field of education, particularly in leveraging technology to enhance the learning or teaching process, Pérez, Daradoumis and Puig (2020) examined the use of chatbots as teaching assistants in education. The systematic review implemented PRISMA to review 80 selected articles, with the findings indicating that educational chatbots are currently being utilised and possess the potential to improve student learning. Manzano-León *et al.* (2021) conducted a systematic review using PRISMA, to investigate the effectiveness of gamification in improving student learning and performance. Upon applying the inclusion criteria, 14 papers were identified for review, with the findings demonstrating that gamification is a viable approach to learning.

From the perspective of enhancing medical education, Tang *et al.* (2020) performed a systematic review following the PRISMA guidelines, to analyse the impact of augmented reality (AR) as a learning tool in medical education. The findings of the 36 screened papers indicate that there is a lack of evidence to support the adoption of AR as a learning tool. In a systematic review utilising PRISMA, Khashaba (2020) examined the effectiveness of collaborative formative assessments conducted online with peers as a learning tool for physiology students. The analysis of the eight identified papers demonstrated that students responded positively to this learning approach. Jayakumar *et al.* (2015) presented a PRISMA-aligned systematic review to examine the effectiveness of electronic learning (e-learning) as a method of teaching in surgical education. There were 38 articles selected for review after meeting the inclusion criteria, with e-learning being identified as a practical teaching approach. Männistö *et al.* (2020) examined a similar learning method, as the authors conducted a systematic review using PRISMA, to assess the effectiveness of online collaborative learning as a learning method in nursing education. There were five randomised controlled trials (RCT) studies selected for review after applying the inclusion criteria, with this learning method being identified as an effective learning approach.

When investigating other areas of education, Parola *et al.* (2023) employed the PRISMA guidelines in a systematic review, to assess the effectiveness of digital games in providing guidance on career selection. The findings of the seven selected papers reveal that digital games can be utilised as a tool for career advice. Abelha *et al.* (2020) presented a systematic review guided by PRISMA, that analysed 69 screened articles, in order to examine the effects of higher education institutes on ensuring workplace readiness. These authors report a discrepancy between the skills of graduates and the requirements of employers. Erschens *et al.* (2019) investigated the occurrence of burnout amongst health students. The systematic review utilised the PRISMA method to review 58 selected articles, with the findings indicating that burnout is a common issue amongst this population.

In the environmental sciences, Mardani *et al.* (2019) conducted a systematic review following the PRISMA method, to explore the linkage between economic growth and CO2E. Subsequent to the analysis of 175 selected papers, the findings suggest that policy development is needed to reduce CO2E without impeding economic growth. Joseph and Mustaffa (2023) applied PRISMA in a systematic review, to analyse the current techniques for managing CO2E in the construction sector. The review screened 99 articles, with the findings demonstrating the need to develop consistent carbon policies and practices for construction.

In the field of social sciences, a systematic review by Awan *et al.* (2022) used the PRISMA method to analyse the trends of sharing fake news on social media, during the Coronavirus (COVID-19) pandemic. The review of 20 screened articles identified self-promotion, impulsive sharing of instant news without verification and socialisation as the primary trends.

3.3 Methods

This study employed the PRISMA method, which is a guideline designed to enhance the quality of systematic reviews and meta-analyses by providing a checklist of items to be reported (Moher *et al.* 2015). The purpose of using PRISMA in a systematic review is to enhance the quality of the review, by making sure there is a rigorous paper selection process and providing a transparent and reproducible account of the study (Page *et al.* 2021). In this systematic review, the PRISMA method was utilised to identify and select relevant studies from the last six years pertaining to the topics of: “the main factors contributing to CO2 emissions” and “trends in CO2 emissions modelling over time”, by using standardized and transparent processes. This ensured that the review produced unbiased and comprehensive findings (Liberati *et al.* 2009). The PRISMA process involved several crucial steps, including defining the systematic review

objectives, conducting detailed database searches, screening, and selecting the relevant studies against multiple levels of eligibility criteria, assessing the quality of the screened studies, and presenting the results. These steps ensured the validity and reliability of the systematic review, as advocated by Galama and Scholtens (2021).

3.3.1 Search strategy

The search strategy was implemented across multiple databases, including: ScienceDirect, Google Scholar, Scopus, Web of Science, and GreenFILE. These databases are relevant to the objectives of this review, as GreenFILE is a database that specifically focuses on topics related to sustainability and the environment (Elton Bryson Stephens Company n.d), whilst the other four databases cover a broader range of topics related to carbon emissions, sustainability, and the environment.

3.3.2 Search terms

To ensure a comprehensive analysis of the factors contributing to CO2E and trends in CO2E modelling over time, this study utilised specific inclusion criteria. In order to find studies that identify CO2 contributors, the search strategy involved multiple search strings, including "key factors of carbon emissions," "key contributors of carbon emissions," "factors that contribute to carbon emissions," "factors that lead to carbon emissions," "major contributors of carbon emissions," "carbon emissions by sector," "key factors of CO2 emissions," "key contributors of CO2 emissions," "factors that contribute to CO2 emissions," "factors that lead to CO2 emissions," "major contributors of CO2 emissions," and "CO2 emissions by sector." Additionally, to find studies on CO2 emissions modelling over time, the search strategy included the following search strings: "carbon footprint modelling over time," "carbon footprint modelling and time series analysis," "CO2 emissions modelling over time," and "CO2 emissions modelling and time-series analysis." The searches were conducted between 2017 and 2023 to ensure that the most recent and up-to-date studies were included in the review.

3.3.3 Exclusion criteria

The exclusion criteria for this review were applied at four levels, to ensure that only relevant and high-quality papers were included in the review. At the first level, duplicates were removed, as indicated in *Table 3.1* below.

Table 3.1: Exclusion criteria at the first level

ID	Exclusion Criterion (EXC)
EXC1_1	Duplicate records were removed

At the second level, articles that were not written in the English language were omitted from the search. Books, conference papers, theses, reports, commentaries or book reviews, were excluded. Articles older than 2017 were also excluded from this review. The exclusion criteria at this level are presented in *Table 3.2* below.

Table 3.2: Exclusion criteria at the second level

ID	Exclusion Criterion (EXC)
EXC2_1	Articles that were not written in the English language
EXC2_2	Books, conference papers, theses, reports, commentaries or book reviews
EXC2_3	Articles older than 2017

At the third level of exclusion criteria, articles published between 2017 and 2020 with fewer than 50 citations, according to Google Scholar, were omitted. Articles published in journals ranked lower than quartile two (Q2), according to the Scimago Journal Rank (2022), were excluded. The citation and journal ranking criteria ensure that the selected articles are from reputable sources and have received sufficient attention in the academic community. This helps to enhance the credibility and impact of the findings of this review (Delgado-Rodríguez and Sillero-Arenas 2018). Furthermore, the third level of exclusion criteria also involved screening the abstract, findings and conclusion sections of articles screened at this level, to determine their eligibility for this review. Papers that only investigated CO2 reduction methods or studied the impact of CO2E were removed. The researcher also excluded studies that examined the effects of a single, pre-defined factor on CO2E without justifying the selection of that factor or variable. Papers that did not work with time-series data or model the CO2 produced by specific sectors, organisations, or individuals, by making use of input features or predictors, were also excluded. *Table 3.3* below presents the exclusion criteria applied at this stage.

Table 3.3: Exclusion criteria at the third level

ID	Exclusion Criterion (EXC)
EXC3_1	Articles published between 2017 and 2020 with fewer than 50 citations
EXC3_2	Articles published in journals that have an index ranking lower than Q2

EXC3_3	Articles that were aimed at discussing CO2 reduction methods
EXC3_4	Articles that discussed the impacts of CO2E were omitted
EXC3_5	Articles that only focused on the effects of CO2E on a pre-defined sector
EXC3_6	Articles that did not focus on CO2 modelling on time-series data
EXC3_7	Articles that predicted the CO2 produced by specific entities, using input predictors were removed

The fourth level involved a deeper analysis of the papers to evaluate their eligibility. Studies that simply carried out forecasting on time-series data, without identifying and analysing the trends of the time-based variables were omitted. The researcher also excluded studies that explored how CO2E was a causality for other events or had an influence on other sectors, as one of the objectives of this study was to identify the main contributors to CO2E. Only studies that investigated CO2E trends over time, using related time-based variables, were included. It is important to note that for the second to fourth level of exclusion criteria, all requirements must be met cumulatively and not individually. Ultimately, it was at the discretion of the researcher to determine the eligibility of a given paper, towards meeting the objectives of this study. After applying all the levels of exclusion criteria, a total of 19 papers were identified for this review. The exclusion criteria at the final level are demonstrated in *Table 3.4* below.

Table 3.4: Exclusion criteria at the fourth level

ID	Exclusion Criterion (EXC)
EXC4_1	Articles that performed simple predictions on time-series data
EXC4_2	Articles that investigated how CO2E were a causality for other events or had an influence on other sectors

3.3.4 Visualisations

This chapter consisted of several visuals to present and summarize the findings of the reviewed studies. These visuals include a meta-analysis table, a timeline of study periods, and several other tables that illustrate the publication distribution by methodology and CO2E contributors. The use of these visuals helps to make the information more accessible and understandable to readers. They provide a quick and concise overview of the findings, making it easier for readers to comprehend and compare the results of the different studies. This enables a more comprehensive understanding of the factors driving CO2E and trends in modelling CO2E over time, facilitating the development of accurate models. This, in turn, helps policymakers identify effective mitigation strategies to counteract the impact of increased CO2E.

3.4 Results

Figure 3.1 illustrates the flow of literature retrieval using the PRISMA methodology, which provides a systematic approach for reviewing and selecting articles for systematic reviews. This figure further depicts the implementation of the PRISMA methodology to select 19 relevant papers for this review, from an initial pool of 61,903 articles obtained through database searches. The implementation of this methodology in a systematic review was outlined by many studies, such as those by Saib *et al.* (2022) and Rajkoomar *et al.* (2022). The screening process involved four levels of exclusion criteria. At the first level, 52,566 articles were excluded, resulting in 9,337 articles that progressed to the second level. Of these, 8,923 articles were excluded, leaving 414 articles to move on to the third level. At this level, 343 articles were omitted, leaving 71 articles to be assessed at the final level of screening. After a thorough analysis, 52 articles were further excluded, and the remaining 19 papers were identified for this review. *Table 3.5* below presents the results of the meta-analysis conducted on selected papers in this review, providing an overview of the studies.

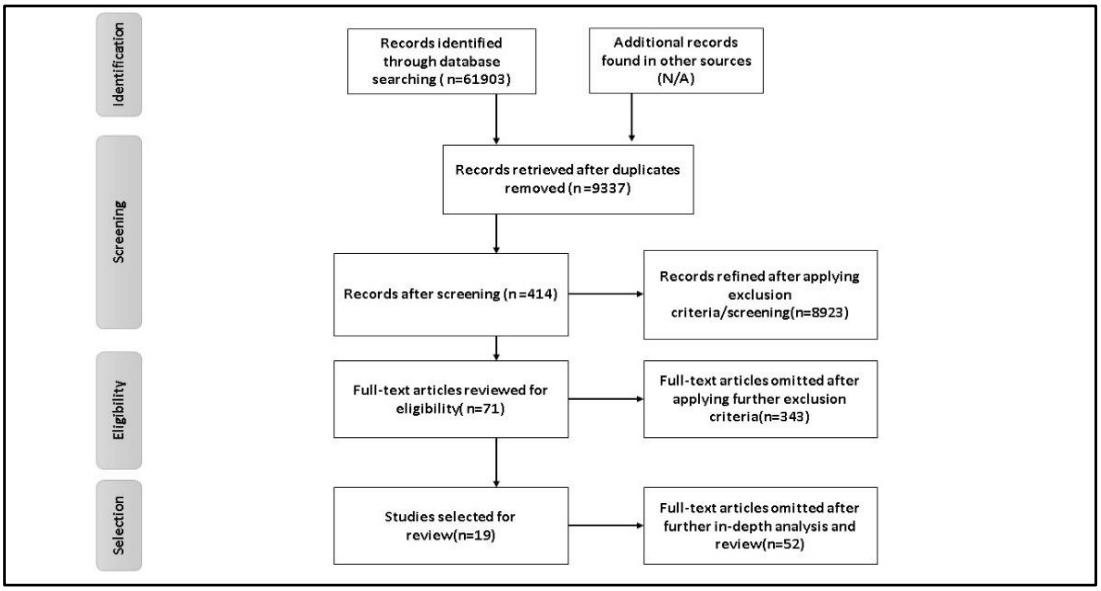


Figure 3.1: Flow of literature retrieval using the PRISMA methodology, adapted from Page and Moher (2017)

Table 3.5: Meta-Analysis and Overview of the selected papers in this systematic review

Study Identity	Year	Author/s	Research scope	Time period	Data source and description
S1	2017	(Lin *et al.* 2017)	Examined the patterns of how urbanisation and economic growth have affected the CO2E of 53 countries globally between 1991 and 2013.	1991-2013	- World Development Indicators (WDI) 2016 - Time-series data related to the CO2E of the 53 countries studied
S2	2017	(Shuai *et al.* 2017)	Identified the key contributors of CO2E across 125 countries during the period of 1990 to 2011	1990-2011	- World Bank database 2016 - Panel data (time-series variables) related to the CO2E of 125 countries
S3	2017	(Shahbaz, Bhattacharya and Ahmed 2017)	Found the key CO2 emission determinants in Australia during the period of 1970 to 2012	1970-2012	- WDI 2013 - Panel data (time-series variables) related to the CO2E of Australia
S4	2018	(Shuai *et al.* 2018)	Found the key CO2 emission determinants in China during the period of 1995 to 2014	1995-2014	Potential key determinants in China were found in the period of 1995 to 2014, through a literature review and fed into the model
S5	2018	(Shen *et al.* 2018)	Determined the major contributors of CO2E in Beijing during the period of 1995 to 2014	1995-2014	- China Energy Statistical Yearbooks (1996–2015) - Panel data (time-series variables) related to the CO2E of Beijing
S6	2019	(Ma *et al.* 2019)	Discovered the driving factors behind increased CO2E in China during the period of 2005 to 2016	2005-2016	- China Energy Statistical Yearbook - Data is based on fossil fuel energy consumption
S7	2019	(Liu, Guo and Xiao 2019)	Intended to find the causes of growth of greenhouse gas emissions in 40 countries worldwide during the period of 1995 to 2009	1995-2009	World input-output database (WIOD)
S8	2019	(Dong, Dong and Dong 2019)	Intended to find the key factors that contribute towards CO2E across 128 countries during the period of 1990 to 2014	1990-2014	- World Bank database 2017 - Panel data (time-series variables) related to the CO2E of 128 countries worldwide
S9	2019	(Sharif *et al.* 2019)	Investigated the relationship between non-renewable energy, renewable energy and CO2E in 74 countries worldwide during the period of 1990 to 2015	1990-2015	- WDI - Panel data (time-series variables) related to the CO2E of 74 nations worldwide
S10	2019	(Mensah *et al.* 2019)	Studied the causal relationship between CO2E, economic growth and power	1990-2015	- WDI

			consumption in 22 African countries from 1990 to 2015.		- Annual Organization of the Petroleum Exporting Countries (OPEC) oil prices - Panel data (time-series variables) related to the CO2E of the 22 African countries studied
S11	2019	(Salahuddin *et al.* 2019)	Analysed the relationship between globalisation, urbanisation and CO2E in South Africa during the period of 1980 to 2017	1980-2017	- World Bank database 2018 - KOF Swiss Economic Institute, 2018 - Panel data (time-series variables) related to the CO2E of South Africa
S12	2020	(Shan *et al.* 2020)	Found the CO2E in China, in the years 2016 to 2017, by using historical CO2E data during the period of 1997 to 2015	1997-2015	- Carbon Emission Accounts & Datasets (CEADs) - China CO2 emission and energy inventory (2016 to 2017)
S13	2021	(Nathaniel *et al.* 2021)	Analysed the relationship between economic growth and CO2E in 15 selected African economies during the period of 1990 to 2014	1990-2014	- WDI - Panel data (time-series variables) related to the CO2E of 15 African countries
S14	2021	(Nguyen, Huynh and Nasir 2021)	Identified the key CO2 emission contributors within the G6 countries during the period of 1978 to 2014	1978-2014	- WDI - Panel data (time-series variables) from the G6 countries
S15	2021	(Yuping *et al.* 2021)	Intended to examine the relationship between globalization, renewable energy, economic growth and non-renewable energy on CO2E in Argentina during the period of 1970 to 2018	1970-2018	- World Bank database 2020 - Annual panel data (time-series variables) related to the CO2E of Argentina
S16	2021	(Anwar *et al.* 2022)	Investigated how urbanisation, agriculture, renewable energy consumption, and economic growth impact CO2E in 15 Asian countries from 1990 to 2014.	1990-2014	- World Bank database 2019 - Panel data (time-series variables) related to the CO2E of the 15 Asian countries
S17	2022	(Sahoo and Sahoo 2022)	Studied the relationship between non-renewable energy, renewable energy and CO2E in India during the period of 1965 to 2018	1965-2018	- WDI 2018 - Global British petroleum (BP) statistics 2019 - Disaggregated data related to CO2E of India
S18	2022	(Chen *et al.* 2022)	Studied how energy consumption, population growth, urbanisation, and economic growth impacted CO2E, in the BRICS countries during the period of 1990 to 2019	1990-2019	- WDI - Panel data (time-series variables) related to the CO2E of the BRICS countries
S19	2023	(Liu *et al.* 2023)	Analysed the relationship between economic growth, urbanisation and energy consumption on CO2E in China during the period of 1995 to 2020	1995-2020	- WDI - Panel data (time-series variables) related to the CO2E of China

This table offers detailed information on the 19 selected articles, including their study identity (S), year, author, research scope, time period, and data source. The research scope outlined in column four enables the studies to be categorized according to their objectives. The first category, which comprises of studies: S3 to S6, S12, and S14, aims to identify the primary contributors to CO2E in a specific country or region. The second category, consisting of: S1, S9 to S11, S13, and S15 to S19, investigates the correlation between economic and environmental factors and CO2E, either at a regional or global level. The third category, including S2, S7, and S8, aims to identify the causes of CO2E growth at a global level.

Figure 3.2 displays the number of publications by year, presenting a graphical representation of the CO2E research conducted over time. This illustration shows the distribution of the 19 selected articles in terms of publication year. Majority of the articles selected were published in 2019, with six publications in that year, followed by 2021 and 2017 with four and three articles, respectively. Both 2018 and 2022 had two publications each. However, the number of publications decreases further in 2020 and 2023, with only one publication in each of these years. *Figure 3.3* displays the distribution of publications across various journal indexes, highlighting the number of articles categorised under each journal index. This figure illustrates the article distribution according to journal quartile index, with 13 of the selected articles in Q1 journals, and the remaining six in Q2 journals. This observation indicates that the papers selected for this systematic review have been published in esteemed journals with high impact factors.

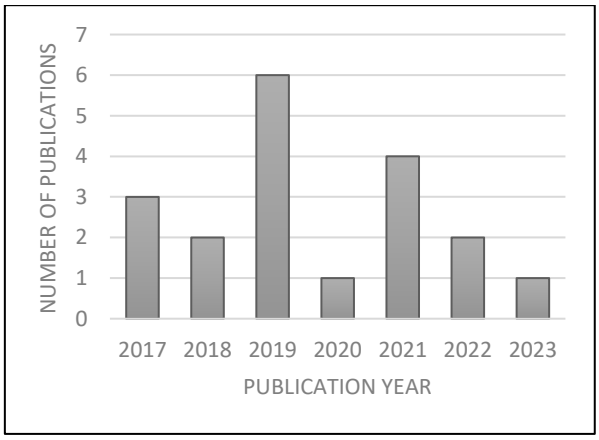


Figure 3.2: Publications by year, modified from Saib *et al.* (2022)

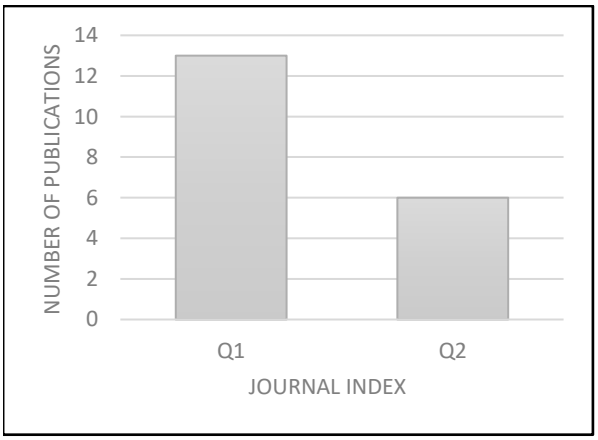


Figure 3.3: Publications by journal index

This implies that this review has utilised high-quality papers to fulfil its objectives (Kraus, Breier and Dasí-Rodríguez 2020).

Figure 3.4 was created from column five of *Table 3.5*, provides a timeline of the study periods covered by the selected papers, with the overall range spanning from 1965 to 2020. This figure illustrates the range of years included in the analysis and provides a visual representation of the temporal scope of the literature on CO2E. The majority of the studies, such as those by S2 and S9, cover a period from the 1990s to the 2010s, indicating a consistent focus on CO2E over the past few decades. Some of the studies, such as those by S3, S15, and S17, have a long timeframe of 20+ years, allowing for a comprehensive analysis of emission trends and patterns over time. Furthermore, many of the studies share overlapping timeframes, enabling data comparison and cross-referencing. Conducting data comparisons involves examining similar data points or variables across different studies conducted during the same period. Cross-referencing, on the other hand, is the process by which researchers compare data, especially when those studies span the same or comparable time periods, with the aim of confirming and validating the information. This approach aids researchers in ensuring the reliability and consistency of their findings, enabling the identification of CO2E patterns or trends that may exhibit greater robustness, when observed across multiple studies.

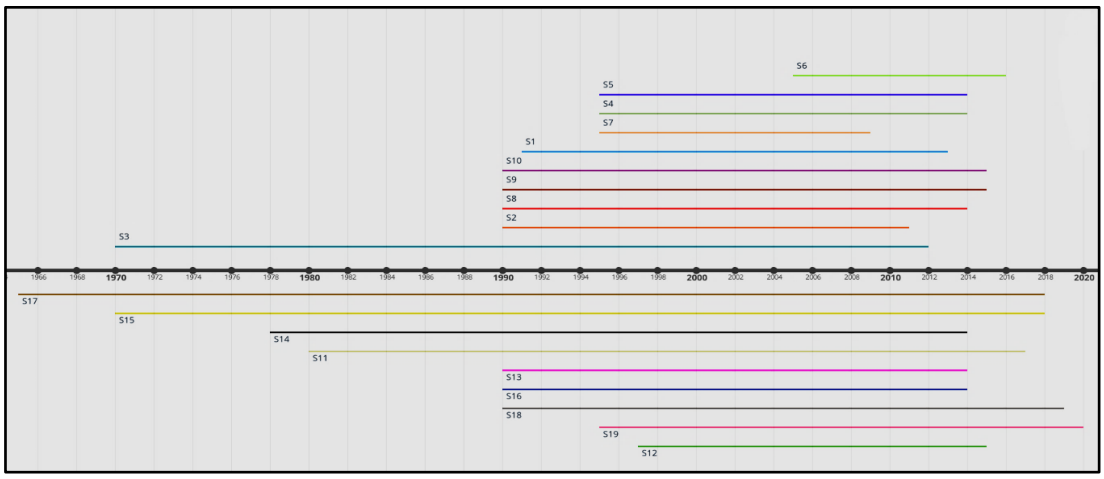


Figure 3.4: Timeline of study periods for selected papers from 1965 to 2020

Table 3.6 provides an overview of the contributors to CO2E based on the findings from the selected papers. This table presents the CO2 contributors identified in the selected studies, S1 to S19. These findings indicate that there were various factors identified as contributors to CO2E, including economic growth, energy production or consumption, population growth, urbanisation, the transportation sector, and market expansion. This table also provides information on the frequency with which different factors are identified as contributors to CO2E across the selected studies, as indicated in column three. Furthermore, this table provides an overview of the most frequently mentioned sources of CO2E, allowing for an understanding

of the relative importance of each contributor as highlighted in the selected literature. Economic growth and energy production or consumption are the most commonly identified contributors, with each factor appearing in 12 publications respectively, but not necessarily in the same publications.

Table 3.6: CO2E contributors by number of publications

Factor	**Article references**	**Total number of publications**
Economic growth	S1, S2, S4, S5, S7, S8, S10, S13, S14, S16, S18, S19	12
Energy production or energy consumption	S2, S3, S6, S8, S9, S10, S12, S13, S15, S17, S18, S19	12
Population growth	S1, S3, S5, S8	4
Urbanisation	S4, S11, S16	3
Transportation sector	S6	1
Market expansion	S14	1

Figure 3.5 below depicts the time-series datasets that were utilised by the reviewed studies to model CO2E. This figure was derived from the last column of *Table 3.5*. Amongst the 19 studies analysing CO2E trends, three datasets stood out in terms of prevalence. The World Development Indicators (WDI), which is a subset of the World Bank's Open data repository, was used in nine (47%) of the studies, followed by additional datasets obtained from the World Bank database, which was used in five (26%) of the studies, and two studies (10%) used the China Energy Statistical Yearbooks to obtain data. Collectively, these top three datasets were employed in approximately 84% of the studies, emphasising their predominant role in conducting research on CO2E trends.

The extensive use of the WDI and additional datasets from the World Bank in previous research validates their applicability in analysing CO2E trends in this study. These datasets are recognized for their comprehensive coverage of socio-economic and environmental indicators, which is crucial for understanding the complex dynamics of carbon emissions (Fantom and Serajuddin 2016). Given that this research involves a data mining approach to analyse emission

trends, the richness of these datasets, allows for a thorough exploration of the factors influencing CO2E.

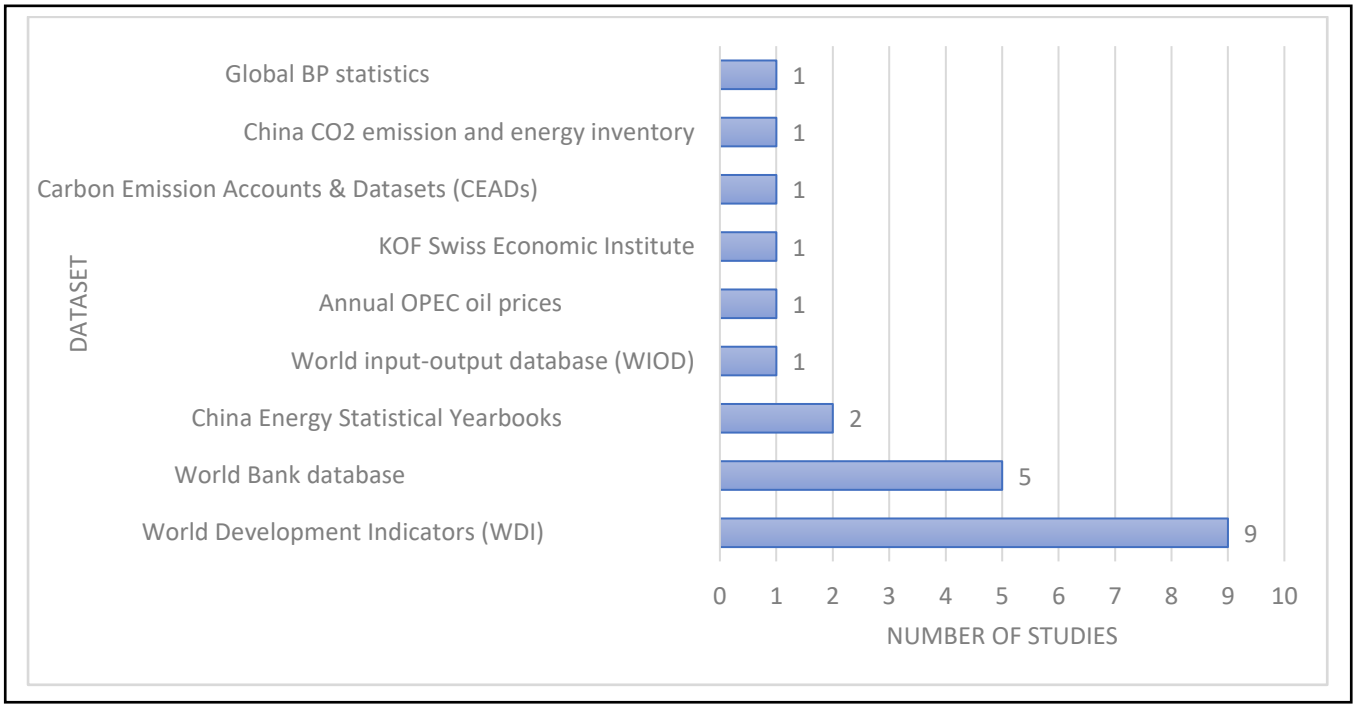


Figure 3.5: Distribution of publications based on dataset utilisation

The WDI, in particular, provides a wide array of variables, including economic, energy, and environmental indicators, offering a rounded perspective on the factors influencing CO2E (Torchio, Lucia and Grisolia 2020). This makes this dataset applicable for this study, as it enables a thorough analysis of the intricate relationships between economic activities, energy consumption, and environmental factors in the South African context.

3.5 Discussion

Literature has consistently demonstrated the effectiveness of PRISMA as a reliable tool for conducting systematic reviews across diverse fields and disciplines, including health sciences and environmental sciences, amongst others. The use of PRISMA will be critical in achieving the objectives of this chapter, as it provides a systematic approach to reviewing and synthesising literature. By utilizing PRISMA, this review will be able to effectively identify and analyse all relevant factors contributing to CO2E and trends in CO2E modeling over time. This will ensure that the review is conducted transparently and rigorously, producing reliable and unbiased research findings (Liberati *et al.* 2009).

In respect to the main drivers of CO2E, the literature consistently highlights economic growth and energy production as pivotal drivers of global CO2E, as reinforced by various studies, such as research authored by Nathaniel *et al.* (2021), Liu *et al.* (2023) and Mensah *et al.* (2019). The insights gathered from understanding the key contributors to CO2E make it possible to develop accurate models capable of informing policy formulation and making key decisions towards achieving sustainability goals. Constructing models that efficiently integrate these major contributors aids policymakers and stakeholders in devising useful strategies to curb CO2E and its environmental repercussions (Zhang *et al.* 2020). In addition to these main factors, population growth (Lin *et al.* 2017; Shahbaz, Bhattacharya and Ahmed 2017; Shen *et al.* 2018; Dong, Dong and Dong 2019) and urbanisation rates (Shuai *et al.* 2018; Salahuddin *et al.* 2019; Anwar *et al.* 2022) emerged as additional contributors, albeit with a lesser influence. The relationship between CO2E and economic growth can be attributed to the rise in both energy production and energy consumption (Boamah *et al.* 2020). As countries develop and their economies grow, there is a greater need for energy to power industrial activities, the transportation sector, and household needs (Medlock and Soligo 2001). This leads to a rise in the combustion of fossil fuels, which constitutes a significant contribution to CO2E (Johnsson, Kjärstad and Rootzén 2019). Moreover, energy production itself contributes significantly to CO2E, as the burning of fossil fuels such as coal and gas, releases large amounts of CO2 into the atmosphere (Cai, Sam and Chang 2018; Johnsson, Kjärstad and Rootzén 2019).

In terms of the study period covered by the selected papers analysed in this review, the overall range spanned from 1965 to 2020, with each individual paper covering a different time period within this range. However, the majority of the studies, such as those by Dong, Dong and Dong (2019) and Mensah *et al.* (2019), covered a period from the 1990s to the 2010s. This temporal scope provides researchers with a wealth of historical data and insights, providing a foundation for understanding long-term trends in CO2E under a specific context, in order to develop mitigation strategies. For the analysis of CO2E trends in SA, these study periods offer a valuable context for benchmarking, comparisons, and understanding how global or regional patterns may have influenced emissions in SA, during these time periods.

Furthermore, given the relatively recent study periods, it appears that research on CO2E analysis is still ongoing. The decades-long emphasis on CO2E highlights how crucial it is to continue researching and understanding the dynamics of carbon emissions. The recent research on CO2E signifies that it remains a relevant contemporary issue, underscoring the ongoing importance of

CO2E research in addressing environmental challenges and informing policies for a sustainable future.

3.6 Chapter summary

The purpose of this chapter was to conduct a systematic review to identify the main factors that contribute to CO2E and to examine the trends in modelling global CO2E over time. The researcher utilised the PRISMA framework and employed four levels of exclusion criteria, to select 19 articles from a pool of 61,903 publications obtained through database searches, to conduct a meta-analysis and track global carbon emissions research from 2017 to 2023. The findings of this systematic review reveal that economic growth and energy production are the primary drivers or contributors to CO2E. Accurately incorporating these drivers into models can facilitate the identification of effective strategies to reduce CO2E and mitigate their impact on the environment. This informs the subsequent application of the CPA and ensemble ML models, in terms of the variables selected. The CPA model, designed to identify significant changes or change-points, can now focus on capturing shifts in CO2E related to economic growth and energy production. Meanwhile, the ensemble ML model, tasked with validating past trends and forecasting future ones, will specifically target these significant drivers.

In addition, the models used in the reviewed studies have consistently focused on analysing CO2E over the past few decades, covering a period from the 1990s to the 2010s, indicating a continuous research focus on this critical environmental issue over the past few decades. The findings also indicate that the WDI and additional datasets from the World Bank's open data repository were commonly utilised datasets in CO2E modelling, as observed in most of the reviewed papers. This prevalence indicates that researchers in the field recognize the reliability, comprehensiveness, and relevance of these datasets for analysing the unique trends related to CO2E. Consequently, these datasets are considered potential data sources for this study, based on their established usage in prior research.

Additionally, this review emphasises the need to address the joint contribution of economic growth and energy production or consumption towards CO2E. The implementation of appropriate strategies, such as a greater commitment from world leaders to commit to the use of clean energy sources, to mitigate the rise of CO2E, is needed. This can help reduce CO2E without negatively impacting the positive effects of energy production and economic growth on a country's sustainability. The subsequent chapter will present the research methodology that this study employed to achieve its objectives.

CHAPTER FOUR: RESEARCH METHODOLOGY

4.1 Introduction

This chapter delves into the research methods and materials employed to achieve the research objectives of this study. It begins by providing an overview of the research process, offering readers a quick understanding of the study's focus and methodology. The research design techniques, data acquisition process and utilised datasets are then discussed. Following this, the justification for the selection of variables is elaborated, including the analytical process and meta-analysis employed for screening and finalising the variable set. The utilisation of CPA and ML in this study is justified, and detailed discussions on the employed CPA techniques, namely CUSUM and Bootstrap analysis, are provided. The researcher benchmarked the AdaBoost regressor, the ML technique selected for this study, against four traditional ML algorithms: Linear Regression, Polynomial Regression, Bayesian Linear Regression, and K-Nearest Neighbors (KNN) Regression. Therefore, the architectures and workflows of these algorithms are detailed in this chapter. Afterward, the architecture and workflow of the AdaBoost regression algorithm was discussed, and the training process for this model was elaborated. The process of enhancing model performance through hyperparameter tuning and validation techniques is explored, supported by relevant diagrams representing their workflows. The performance metrics for evaluating the regression model are covered, and an overview of the CO2E forecasting process is presented. This chapter concludes by discussing the relationship between CPA and ML, highlighting how ML validates past trends identified by the CPA model, through graphical analysis.

4.2 An overview of the research process

Figure 4.1 illustrates the research process followed in this study, which depicts the steps taken to achieve this study’s objectives. This research process began by conducting a PRISMA-aligned systematic review to identify the key factors contributing to CO2E. These factors will serve as the input variables for building the CPA and ML models. By incorporating the review's findings into the selection of input variables, more robust models that can accurately provide insights into South Africa's CO2E trends may be developed. The next step is to acquire the necessary data to build the models. Secondary data from the World Bank's Open Data initiative data repository will be used, consisting of time-series datasets and variables related to CO2E in SA. The author’s systematic review has revealed that economic growth and energy

production are the main drivers of CO2E. As a result, this study will utilise datasets related to energy production, economic growth, and CO2E in SA.

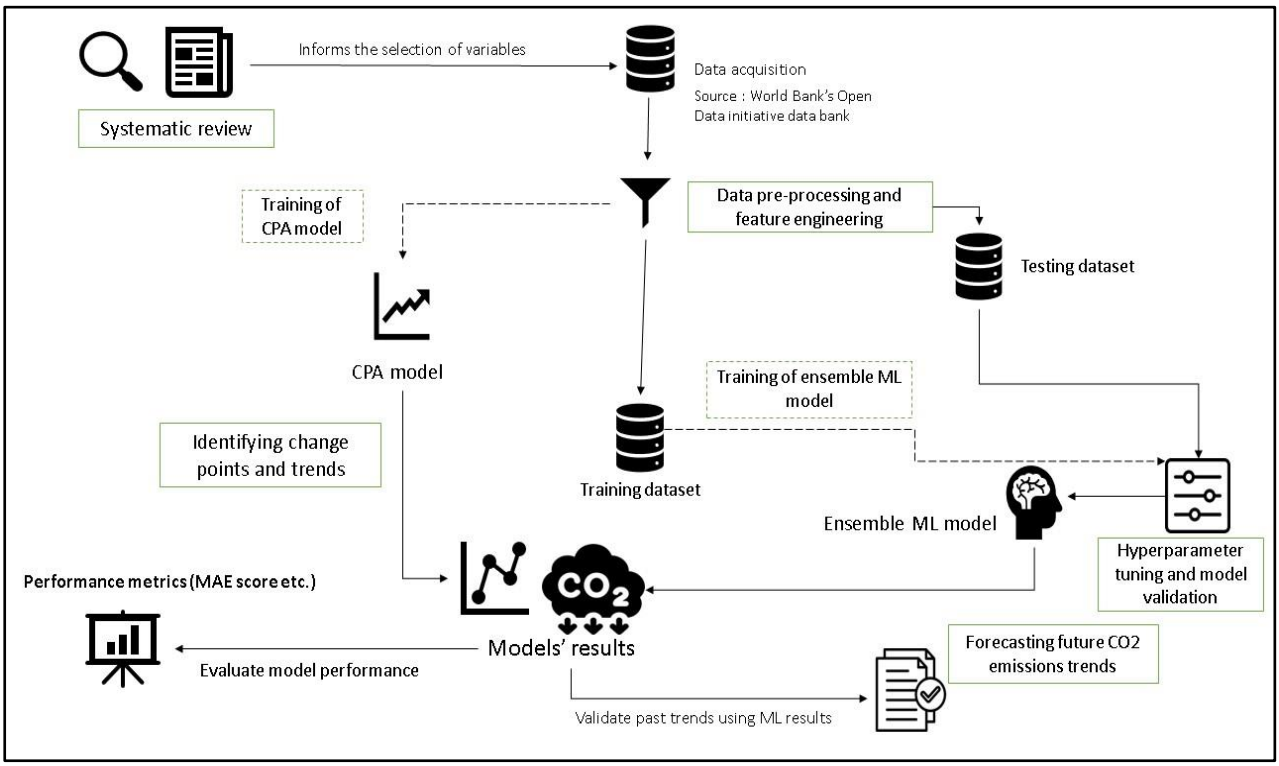


Figure 4.1: High-level overview of the research process, adapted from Mohamed, Patel and Naicker (2023)

However, an analytical process will be carried out to determine the most suitable data sources for the final models. Any acquired data will be pre-processed through feature engineering to ensure accuracy and completeness, including handling missing values and correcting format issues. The CPA model will then be used to identify abrupt changes and trends in the CO2E data. This analysis will be followed by training an ensemble ML regression model to forecast CO2E. Prior to training the ML model, hyperparameter tuning techniques utilising cross-validation will be employed to optimise the parameters, to improve the performance of the model. Additionally, repeated holdover techniques were applied as an alternative approach, to further enhance model performance. A subsequent comparative analysis will identify the more effective approach between these two methods in the context of this study. To evaluate ML model performance within this context, common performance metrics such as the MAE score, RMSE score and R^2 score will be utilised. Furthermore, the author will include an accuracy score that they calculated, as an additional metric, since accuracy metrics are not traditionally employed in ML regression scenarios. Lastly, the forecasted CO2E trends will then be utilised to validate the past trends, through graphical analysis.

4.3 Research design

Quantitative research involves the use of statistical, mathematical, and computational tools to draw conclusions from numerical data (Chih-Pei and Chang 2017; International Research n.d). This research is quantitative because it employs data mining and ML techniques to identify trends and patterns in CO2E data, which are numerical in nature. These techniques are also quantitative which supports the selection of this research method (Kotsiantis, Zaharakis and Pintelas 2007). The studies conducted by Tunde *et al.* (2022) and Li *et al.* (2018) used the quantitative approach, which involved statistical and ML techniques to analyse CO2E trends. The use of various model performance metrics, such as the MAE score, which is quantitative in nature, further supports the suitability of the quantitative approach for this study.

The experimental research design approach is suitable for this study, because similar research has used this methodology to analyse CO2E trends. Awe and Adepoju (2020) utilised experimental research to identify rapid trends in CO2E using the CPA model to carry out experiments, whilst Ağbulut (2022) employed ML experiments to forecast CO2E. Experimental research involves manipulating the independent variable to evaluate its effect on the dependent variable (Formplus 2021). Therefore, this design approach is appropriate because the data mining and ML techniques employed in this study, including the CPA and ML models, will be manipulated to analyse CO2E trends, and the performance of these models will serve as the dependent variable.

4.4 Data acquisition

4.4.1 Datasets

The time-series data required for building the models in this study will be obtained from the World Bank's Open Data initiative data repository. The World Bank is a reputable international organisation recognised for its stringent data collection and quality assurance procedures (The World Bank n.d). Through the World Bank's Open Data initiative, researchers can gain access to an extensive selection of time-series datasets, including socio-economic indicators, such as CO2E, from various trusted sources, including government agencies and international organisations (Dethier 2007; The World Bank 2012). Additionally, the repository's dedication to regular updates and thorough data validation processes, further enhances the data's reliability (Albert *et al.* 2023). By using data from this repository, the study's findings gain validity and credibility, allowing for a thorough analysis and understanding of the CO2E patterns in SA.

The following time-series databases from the World Bank, will be utilised to acquire the required data for building the models:

- Sustainable Energy for all (The World Bank 2018)
- Country Climate and Development Report (The World Bank 2022)
- World Development Indicators (The World Bank 2023).

The World Data Bank's Open Data collection emerges as an efficient data source for this research, especially its databases, such as the WDI. This decision was motivated by the results of the systematic review conducted in Chapter Three, which demonstrated that these sources were predominantly utilised in research to model CO2E trends over time. Therefore, these data sources are considered ideal, given that this study also aims to analyse the trends in CO2E, through time-series analysis.

Apart from the systematic review's findings, the selection of the World Data Bank and its databases, which contain time-series datasets, is further justified, as they provide essential data for understanding CO2E trends. Its extensive datasets provide insight about emission patterns, enabling a more in-depth analysis (Desai *et al.* 2022). This is especially important for CPA, since the historical background provided by these sources helps identify significant changes in CO2E. In terms of ML forecasting, these data sources hold importance for validating past trends and predicting future trends. They ensure that the study's analysis of CO2E is well-supported and based on reliable and accurate data (Mishra 2011).

4.4.2 Justification of the time-series variables examined

The selection of variables for this study, was informed by the findings of the systematic review, which was conducted in Chapter Three. This review identified economic growth and energy production as the primary contributors to CO2E. The increase in energy production and consumption captures the relationship between CO2E and economic growth. Energy is required to power industrial processes, transportation, and household requirements, as countries expand economically. As a result, there is an increase in the burning of fossil fuels, which contributes significantly to CO2E. In terms of energy production, South Africa's dependence on coal to product energy results in significant CO2E, as burning coal releases large amounts of CO2 into the atmosphere, when generating electricity (Joshua and Alola 2020). Therefore, by utilising these contributors as input variables for the models, the research

ensures precision, relevance, and effectiveness in its analysis. This helps to provide a more targeted understanding of the major factors shaping CO2E patterns in SA.

Furthermore, the systematic review also consisted of research articles that analysed the CO2E trends in SA (Lin *et al.* 2017; Shuai *et al.* 2017; Mensah *et al.* 2019; Nathaniel *et al.* 2021; Chen *et al.* 2022). However, not all of those studies were conducted in an isolated South African context. Nonetheless, those studies also identified economic growth and energy production as the significant drivers of CO2E. The fact that the factors driving CO2E in SA, independently point to the same factors contributing to global CO2E, serves as a form of validation for the variables employed in this study. It reinforces the robustness and credibility of the chosen variables, adding strength to this research methodology (Sperber 2004). This cross-validation of factors also ensures that the models are directly applicable to the unique context of SA. This enhances the relevance of this research for policymakers and stakeholders in the country.

Economic growth and energy production have been identified as the main factors that contribute to CO2E in SA, in additional studies as well (Seymore, Inglesi-Lotz and Blignaut 2014; Dhansay *et al.* 2022; Espoir, Sunge and Bannor 2023). The cross-validation of the systematic review's results with the findings from other studies strengthens the robustness of the selected variables, highlighting economic growth and energy production as the major contributors to CO2E in SA. This consistency across multiple studies confirms the validity of the selected variables. All of which helps to strengthen the reliability and overall findings of this study (Hofer and Piccinin 2009), as the key influence of economic growth and energy production on carbon emissions in SA is confirmed and precisely captured.

4.4.3 Variable selection analysis

Table 4.1 below presents the initial set of variables selected for this study, which depicts the start of the analytical process conducted, to determine the final variables that were used in the CPA and ensemble ML models. This analytical process ensures that the most relevant data are selected, which helps to ensure the accuracy and reliability of the study's outcomes (Josselin and Le Maux 2017). Additionally, this process enforces the principle of "Garbage In, Garbage Out" (GIGO), which emphasises how crucial it is to utilise high-quality input data, since the quality of the data has a significant impact on the accuracy of the results (Kordos 2018; Wartner *et al.* 2019). This study ensures that the input data is useful and relevant, by carefully screening variables through an analytical process, which improves the validity of the data mining and ML

models. This meticulous process will help to safeguard against potential misrepresentations or inaccuracies in the results (Hellerstein 2008).

Table 4.1: Time-series variables meta-analysis

Variable name	Description	Time-period	# of data points	Unit of measure	Database source
Total electricity output in Gigawatt hours (GWh)	Total Gigawatt hours (GWh) produced by all of South Africa's power plants	1990-2015	26	GWh	Sustainable Energy for All
Electricity net generation	Electricity generation from various power plants in South Africa, including both utility and non-utility sources	1980-2020	41	Billion kilowatt hours (kWh)	Country Climate and Development Report
CO2 emissions (kt)	CO2 emissions resulting from the combustion of fossil fuels in South Africa	1995-2020	26	Kilotons (kt)	WDI
Electricity production from coal sources (% of total)	Percentage of electricity produced from all types of primary and secondary coal sources in South Africa	1980-2015	36	% contribution	WDI
Electricity production from nuclear sources (% of total)	Percentage of nuclear power plant-produced electricity in South Africa	1984-2014	31	% contribution	WDI
Electricity production from oil sources (% of total)	Percentage of electricity generated from petroleum sources and crude oil in South Africa	1976-2015	40	% contribution	WDI
GDP growth (annual %)	South Africa’s GDP growth rate expressed as a percentage per year at market prices, using the constant local currency	1980-2022	43	% growth rate	WDI

The *Total electricity output (GWh)* variable provides data for a relatively shorter time span, from 1990 to 2015, with 26 data points. Whilst the variable focuses on the total number of Gigawatt hours (GWh) generated by all power plants in SA, it may not capture the broader range of data and trends related to electricity generation. In addition, the limited time span might not adequately represent the long-term trends and patterns required for forecasting future CO2E trends in SA. The *Electricity net generation* variable, on the other hand, spans a longer time-period, which is from 1980 to 2020, with 41 data points, providing a more comprehensive dataset. This variable includes both utility and non-utility sources from electricity and

combined heat and power plants, giving a broader view of electricity generation in SA. The extended time span enhances the potential for capturing long-term trends and detecting significant changes in CO2E.

Therefore, the *Electricity net generation* variable appears to be the better choice for building the CPA and ML models. It covers a longer time-period and has a higher number of data points, which offers a more robust dataset, which can improve the accuracy of analysing CO2E trends and enhance the forecasting of future trends. Additionally, the inclusion of both utility and non-utility sources provides a more holistic view of electricity generation, which is likely to be highly influential in analysing CO2E trends. In terms of the unit of measurement, the *Total electricity output* variable is measured in GWh, which is a widely recognized and standardised unit used to quantify electrical energy generation (Maroney 2019). This unit of measurement is ideal for analysing CO2E trends, due to its standardised and easily comparable format. However, even though it measures in billion kilowatt hours (kWh), the *Electricity net generation* variable offers a greater number of data points and spans a longer time period. If any issues arise in the analysis, this variable's data could be converted to GWh to allow for consistent comparisons.

The selection of the *Electricity production from coal sources (% of total)* as a key variable is strongly justified by the comparative analysis depicted in *Figure 4.2*. The contribution of each energy source was calculated by utilising their percentage contributions towards the total electricity production from the *Electricity production from coal sources (% of total)*, *Electricity production from nuclear sources (% of total)*, and *Electricity production from oil sources (% of total)* variables, respectively. Each of these percentage variables represents the proportion of total net electricity generation in SA. To quantify their actual electricity generation in billions of kWh, each percentage value out of 100 was multiplied by the *Electricity net generation* variable. This calculation is further illustrated in *Equation 4.1* below. Subsequently, the contributions in billions of kWh from coal, nuclear, and oil sources to the total amount of net electricity generation, can be directly compared after applying this calculation.

$$Contribution\ of\ each\ source\ (in\ billion\ kWh) = \frac{Percentage\ contribution\ of\ each\ source}{100} \times Electricity\ net\ generation\ (in\ billion\ kWh)$$

Equation 4.1: Calculation of the contribution of each energy source towards the total electricity generation

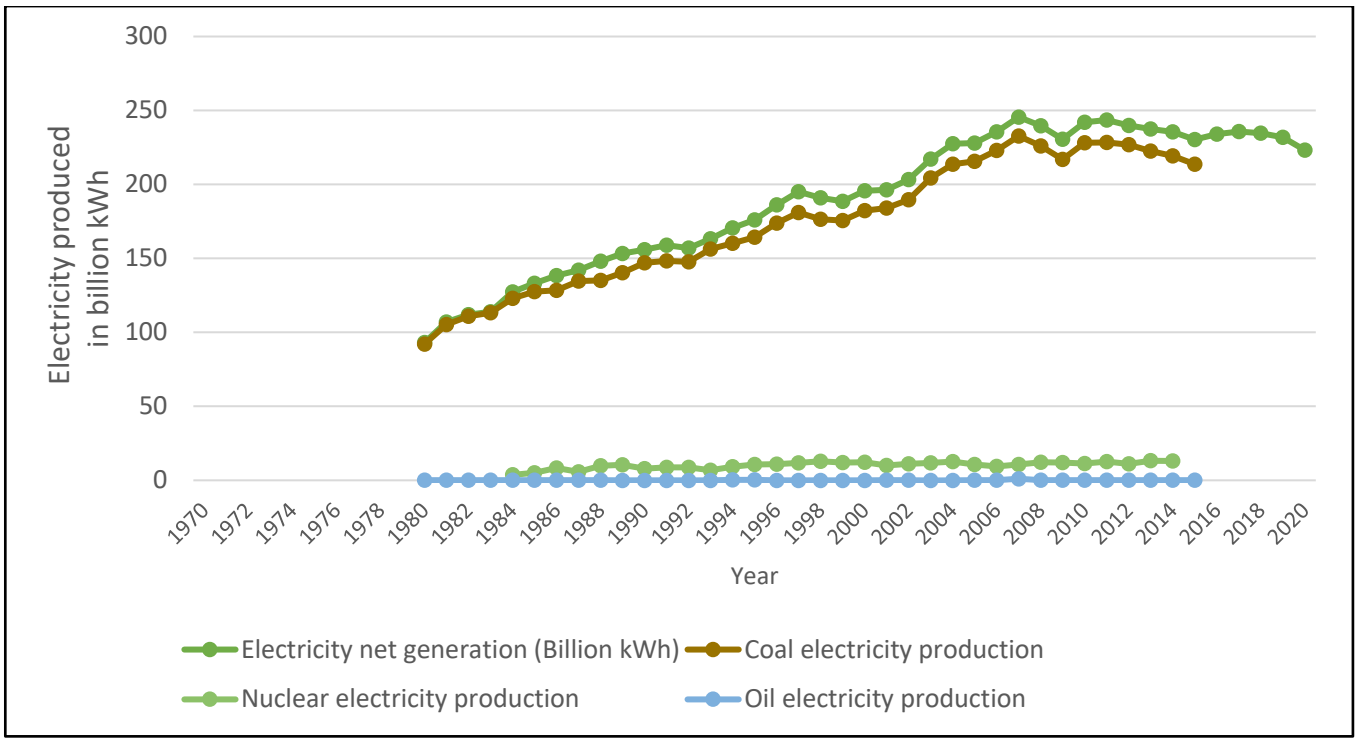


Figure 4.2 : Electricity produced in SA by each source

The results demonstrate that electricity production from coal sources, is the highest contributor to the *Electricity net generation* variable. This is of paramount importance to the research because it reiterates coal's dominant role in producing electricity in SA (Menyah and Wolde-Rufael 2010), and consequently, its substantial contribution to CO2E. Given the large amounts of CO2E emitted by burning coal, understanding its importance as a key source for energy production, is essential for developing efficient mitigation techniques and making well-informed energy policy decisions. By examining the composition of the *Electricity net generation* variable, insights are gathered into the energy sources that drive CO2E, which can help facilitate the development of sustainable energy policies and practices. Furthermore, the results imply a correlation between the trends of the total electricity generated and the energy generated from coal sources. This is because South Africa's energy structure is dependent on coal sources. The similar trends shared between these variables indicate that using either the *Electricity net generation* variable or the *Electricity production from coal sources (% of total)* variable is interchangeable. Therefore, it is reasonable to use the *Electricity net generation* variable to examine the effect of energy production on CO2E, as the majority of the total energy is derived from coal either way.

The variable *CO2 emissions (kt)* holds significant importance in this research, as it represents the key CO2E dataset for comparing against energy production and economic growth. This

provides essential insights into the relationship between carbon emissions and these contributing factors. Furthermore, this variable quantifies CO2E produced by the burning of fossil fuels in SA. The unit of measurement for this variable is kt, which is commonly used to quantify CO2E on a large scale (Haseeb, Wattanapongphasuk and Jermsittiparsert 2019). This provides a comprehensive representation of the carbon dioxide produced from fossil fuel burning, during the time-period of 1995 to 2020. As there are 26 data points available, the dataset spans a considerable duration, allowing for a robust analysis of CO2E trends and patterns in SA.

The *GDP growth (annual %)* is another key variable in this study, contributing to the analysis of CO2E trends in SA. Its unit of measurement, which is a percentage growth rate, allows for a standardised comparison of GDP growth in SA, over the period of 1980 to 2022, including 43 data points. By analysing the annual percentage growth rate of GDP at market prices based on the fixed local currency, it is possible to gain insights into the economic growth trends of the country. This variable plays a crucial role in understanding the relationship between economic growth and energy production, shedding light on how GDP growth affects CO2E in SA, over the specified time frame.

4.5 CPA techniques for analysing CO2E trends

CPA, sometimes referred to as Change-Point Detection, is a statistical technique representing the data mining approach applied in this study. This technique locates points or periods in time-series data where a notable shift or deviation from the established pattern has occurred (Arif *et al.* 2017; Parpoula and Karagrigoriou 2022). It works by detecting any significant changes in the underlying statistical properties of the data, such as fluctuations in the mean (Camci 2010). Additionally, CPA is a particularly useful technique for deriving crucial insights from meteorological data, which includes variables such as ozone levels, precipitation and concentrations of CO2 (Chelani 2011; Arif *et al.* 2017). Hence, in the context of this research, CPA can be used to identify significant changes or breakpoints in the carbon emissions data, indicating shifts in emission patterns or trends over time. These identified shifts can be very effective in identifying the factors that influence South Africa's CO2E. Thereafter, policymakers may use this information to target specific time-periods or circumstances that are associated with the significant shifts in emissions, which can help to develop focused and efficient mitigation strategies.

In order to identify abrupt changes in the time-series data, this study utilised the CUSUM and Bootstrap CPA algorithms. Control charts were traditionally used to identify changes in time-series data (Taylor 2000; Abbas and Fried 2020). A control chart is a graphical tool used to monitor variations in a process over time (Le Thien *et al.* 2023). In the context of analysing CO2E trends, it can be employed to detect shifts or abnormalities in the emission patterns, which helps to identify significant changes that may need further analysis. However, according to James *et al.* (2020), the analysis generated by control charts is not consistently accurate. The CPA techniques utilised in this study, stand out as a superior method for detecting shifts in time-series data when compared to control charts. Unlike control charts that update with new data, CPA is applied after all the data has been collected (Tartakovsky *et al.* 2006; Monyeki, Naicker and Obagbuwa 2020). This approach ensures a more comprehensive analysis by considering the entire dataset as a cohesive unit. In the case of CO2E, where trends may be influenced by various factors over time, examining the complete dataset allows for a thorough understanding of the intricate carbon emission patterns and fluctuations (Sharma 2011). This allows for targeted interventions and the development of more effective mitigation strategies to reduce CO2E in the country, which are tailored to the specific dynamics revealed through the analysis of such patterns.

CPA also excels in identifying subtle and long-term changes that control charts might overlook, showcasing its power in characterizing changes with confidence levels and intervals (Taylor 2000, 2018; Monyeki, Naicker and Obagbuwa 2020). The importance of confidence levels in characterizing the identified changes lies in the enhanced reliability and robustness of the findings (Han *et al.* 2020). Confidence levels provide a measure of the certainty or uncertainty associated with the identified shifts in carbon emission patterns. This is particularly crucial when dealing with environmental data, where multiple factors can contribute to variations in CO2E. By incorporating confidence levels, this study gains the ability to distinguish between changes that are statistically significant and those that may be natural changes or fluctuations. This precision is critical for understanding the reliability of the identified patterns and making well-informed interpretations of the data. When a change in CO2E is detected with a high confidence level, it signifies a more reliable and substantial shift within the emissions data. By having a thorough understanding of these reliable patterns, strategies to mitigate CO2E may be more precisely adjusted to address the identified shifts. Essentially, the high confidence levels support the study's ability to offer more reliable and accurate insights, providing a strong basis

for the development of effective policies that will have a significant influence on reducing the CO2E in the country.

In comparison to control charting, CPA is also resilient to outliers, and provides greater control over the total error rate when working with large historical datasets (Taylor 2018; Hussain *et al.* 2019; Monyeki, Naicker and Obagbuwa 2020). The ability of CPA to maintain stability and minimise errors, despite the presence of outliers in the data, contributes to the reliability and accuracy of the CPA model. This resilience is crucial for discerning genuine shifts in CO2E trends from random fluctuations or irregularities in the data. As a result, the improved insights of the CPA model can help develop more efficient strategies to reduce carbon emissions. Additionally, CPA is easier to use and more flexible, making it preferable in scenarios where precise change detection is crucial (Cho 2016; Taylor 2018). The application of CPA becomes essential when detecting abrupt changes, especially in the case of CO2 fluctuations that occur instantly. Unlike control charts, which may lag in detecting such abrupt changes (Monyeki, Naicker and Obagbuwa 2020), CPA, specifically utilising CUSUM and Bootstrap techniques, excels in detecting sudden shifts (Arif *et al.* 2017), aligning well with the rapid nature of CO2 fluctuations. This capability is crucial for understanding the dynamic and rapidly changing landscape of CO2E in SA, providing a real-time perspective that traditional control charts may delay in capturing. Consequently, this timeliness is essential for policymakers and stakeholders to respond promptly to emerging emission trends, allowing for the development of agile and effective strategies to address the specific challenges of reducing CO2E.

4.5.1 CUSUM and Bootstrap analysis

The CUSUM and Bootstrap analysis algorithms were selected to analyse CO2E trends and detect change-points in the time-series data. The choice of these algorithms to perform CPA in this research is advantageous due to the specific challenges posed by other CPA algorithms. Several alternative algorithms, such as the Bayesian CPA or Sequential Probability Ratio Test (SPRT), have limitations that make them less suitable for this study (Georgescu 2012; Tartakovsky, Nikiforov and Basseville 2014). For instance, the Bayesian CPA algorithm relies heavily on prior knowledge or assumptions about the data, which can be challenging to define accurately in the context of CO2E trends (Tartakovsky and Moustakides 2010; Aminikhanghahi and Cook 2017). Additionally, the SPRT algorithm requires a clear specification of the distribution of the underlying data, which may not be easily achievable due

to the complex and non-linear nature of CO2E (Ba and McKenna 2014; Xu, Mei and Moustakides 2021).

In contrast, the CUSUM algorithm is well-suited for detecting gradual shifts and small changes in CO2E, allowing for a more comprehensive analysis of the trends (Wachs 2022). Its ability to adapt to various data patterns and identify subtle variations makes it highly applicable to the investigation of emissions over time (Fortea-Sanchis and Escrig-Sos 2019). The Bootstrap analysis algorithm complements CUSUM by effectively handling irregular data distributions and also providing robust and reliable results through the calculation of confidence levels for each detected change (Buzun and Avanesov 2017). The combined strengths of these algorithms can enhance the reliability and applicability of this study's findings, enabling a more comprehensive understanding of the dynamics in CO2E over a specified period.

In this study, the aforementioned CPA techniques were used to detect any changes and trends in the CO2E data in SA. Arif *et al.* (2017) and Taylor (2018) illustrated the application of the CPA techniques selected for this study, which served as a guiding framework during their implementation. The subsequent questions are addressed in the application of this method, as previously demonstrated by Monyeki (2021) who also utilised a similar methodology to analyse crime patterns in SA:

i. "Did any change/s take place?"

ii. "When did they/it happen?"

iii. "To what extent is the researcher confident in the accuracy of the changes observed?"

(Taylor 2018).

The results of the CUSUM algorithm are presented in a CUSUM chart. A CUSUM chart is a type of control chart used in statistical analysis to detect small shifts in the process mean over time (Wachs 2010). It accumulates the differences between individual data points and a reference value, which is typically the mean (Ryu, Wan and Kim 2010). A CUSUM chart often illustrates an anomaly in the process mean when its cumulative sum shows a noticeable upward or downward trend that surpasses predefined boundaries or deviates from established patterns (Taylor 2018). Meanwhile, Efron (1992) formulated the bootstrap algorithm, which is a widely used resampling method for assessing the confidence levels in a statistical distribution. In this context, the bootstrap algorithm is utilised to determine the confidence level of each detected

change. Furthermore, recursive algorithms are used in CPA to detect numerous shifts (Arif *et al.* 2017). The dataset is then divided into subsets, each with distinct means, through the use of iterative algorithms. This method produces a series of approximated change-points, each associated with their own confidence level (Taylor 2000; Arif *et al.* 2017). A backward elimination process is then utilised to reduce any false positive detections.

Utilising a time-series dataset, the cumulative sum of the data points is computed and traced, to create the CUSUM charts (Taylor 2000). To begin this process, the individual data points are represented by $x_1, x_2, \ldots, x_n$ where n refers to the number of observations in the dataset. The original dataset is subject to a three-step process, in order to detect any change-points, where $D_0 = \{ x_1, \cdots, x_n \}$ of size n ($n_0 = |D_0|$) represents the dataset. The first step involves calculating the average ($\overline{x}$) of $x_1, x_2, \ldots, x_n$, which is done by summing up the values and dividing the total by the number of data points, as depicted in *Equation 4.2*.

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Equation 4.2: Calculation of the mean in a time-series dataset

The next step is to initialise the cumulative sum to zero. Thus, assign S_0 to the value zero, denoted as $S_0 = 0$. Thereafter, the cumulative sum is calculated using *Equation 4.3* below.

$$S_i = S_i - 1 + (x_i - \overline{x}), i = 1,2,3, \ldots . n$$

Equation 4.3: Formula to calculate the cumulative sum

This formula represents the calculation of the cumulative sum (S_i) at each step in a sequence. In this context, this formula is applied at each data point in a time-series dataset. It is computed by adding the difference between each data point (x_i) and the mean ($\overline{x}$), to the previous cumulative sum ($S_i - 1$). This process is repeated for each data point in the dataset, from $i = 1$ to $i = n$. The cumulative sum illustrates the contribution of each data point to the overall sum, considering its deviation from the mean. This calculation is visually represented in a CUSUM chart. In essence, this chart captures how the cumulative sum evolves as each data point is fed into the formula, considering the difference between the data point and the mean (Taylor Enterprises n.d). Therefore, this technique is fundamental in performing CPA for detecting any shifts or changes in the trends of the time-series data. A change-point is identified when there's a notable shift or deviation in the cumulative sum, indicating a change in the underlying

statistical properties of the data, typically referring to the mean (Taylor Enterprises n.d). In the context of this study, this shift suggests a change or alteration in the pattern of CO2E or the other related emission factors. A sudden increase or decrease in the cumulative sum signifies a significant deviation from the expected pattern, pointing to a change in the trends (Rogerson 2003). Additionally, this CPA technique works iteratively, by applying the cumulative sum calculation across the entire dataset, to detect multiple change-points. This calculation answers question *i.* of the CPA approach, as changes or change-points in the data can be identified.

Thereafter, the confidence levels for the identified changes need to be computed, as it is crucial to establish the reliability and significance of these shifts in the data. Bootstrap analysis is employed to calculate these confidence levels, providing a measure of certainty associated with the observed changes (Brownstone and Valletta 2001). However, prior to initiating the bootstrap analysis, it is imperative to compute an approximation of the magnitude of the changes. This magnitude serves as a crucial boundary for the CUSUM chart, defining the extent of the observed shifts in the data. The formula to calculate the magnitude is illustrated in *Equation 4.4*.

$$S^{i}_{diff} = \max S_i - \min S_i = S_{max} - S_{min}, i = 0, \ldots, \mathrm{n}$$

Equation 4.4: Formula for calculating the magnitude of change in the time-series dataset

The above formula calculates the difference between the maximum (max) and minimum (min) values of the cumulative sum (S_i) over the entire dataset. This difference, denoted as S^{i}_{diff}, represents the range or spread of the cumulative sum values. Basically, this helps to determine the variation in the cumulative sum throughout the dataset. This range is crucial for creating a restriction boundary in the chart, which is necessary for identifying changes or change-points in the dataset. A smaller range denotes more data stability, whereas a larger range might suggest that the cumulative sum has more notable variations (Siegel 2012). Moreover, the boundary created using this range provides a reference for assessing the significance of the detected changes, during the subsequent bootstrap analysis.

As mentioned previously, the calculation of the magnitude of change is essential for identifying change-points (Aminikhanghahi and Cook 2017). When a change-point occurs, there is typically a significant shift in the cumulative sum values. By determining the range (max − min) of the cumulative sum, a reference range that characterises the normal variability in the data is established. If the cumulative sum values surpass this reference range, it suggests that a

change or change-point has occurred. The reason for this is that, under normal conditions, the computed range offers an indication of the anticipated shifts in the cumulative sum. Any values exceeding this range indicate deviations from the expected pattern, signalling potential changes in the underlying statistical properties of the data.

After the magnitude of change has been calculated, N bootstrap iterations are run on the dataset D_0 (Arif *et al.* 2017). An individual bootstrap iteration executes in the following manner:

i. A bootstrap sample comprising of n data points from the time-series dataset D_0, is symbolised as x_j (j = 1, 2, 3, ..., n). The sampling without replacement (SWOR) technique is applied to produce this sample, by randomly rearranging the initial n values (Singh *et al.* 2009).

ii. The cumulative sum for the bootstrap dataset is subsequently calculated, using the formula presented in *Equation 4.3*. This value is denoted as S_j.

iii. The magnitude of change across the entire bootstrap sample is then computed, as illustrated in *Equation 4.5* below.

$$S^{\,j}_{diff} = \max S_j - \min S_j = S_{max} - S_{min}, j = 0, \ldots, \mathrm{n}$$

Equation 4.5: Formula for computing the magnitude of change in the bootstrap dataset

iv. The following procedure is executed when the magnitude of change in the original dataset exceeds that of the bootstrap dataset.

To initiate the procedure, the confidence level of the change must be calculated. Let N denote the total amount of successfully implemented bootstrap samples, and let K depict the quantity of bootstraps such that $S^{\,i}_{diff} > S^{\,j}_{diff}$. The confidence level ($CL$), expressed as a percentage, is defined as the proportion of cases where the previous condition is true, relative to the total number of bootstrap samples, as depicted in *Equation 4.6.*

$$CL = \frac{\Sigma_{i=1}^{N}(S^{i}_{diff} > S^{j}_{diff})}{N} \times 100 = \frac{K}{N} \times 100$$

Equation 4.6: Formula for calculating the CL in detecting change-points

This formula calculates the CL as a percentage, representing the probability of detecting a change-point, based on the observed cumulative sum differences in the bootstrap samples. The condition ($S_{diff}^{i} > S_{diff}^{j}$) identifies instances where a change-point is observed. The formula sums up these instances across all bootstrap samples (from 1 to N), as indicated by the summation symbol. For a simplified representation of the summation, K serves as a variable representing the total number of instances where the condition is satisfied, and N is the total amount of bootstrap samples. To obtain the confidence level on a percentage scale, K was divided by N, and the result was multiplied by 100. This calculation addresses question *iii.* of the CPA approach, as the confidence level provides a measure of accuracy to the researcher, for the occurrence of changes or change-points in the data.

The application of bootstrapping results in a distribution-free technique that relies solely on the independent error structure assumption (Albrecher *et al.* 2019). Essentially, this method does not require any specific assumptions about the fundamental distribution of the data; rather, its validity is based on the independence of errors. Because of this, it is a reliable and adaptable method, especially in cases where it is unclear or difficult to define the distributional assumptions. The data points, denoted by the symbols X_1, X_2, ... in this structure, must be dispersed according to a time order. This independent error structure serves as the foundation for both CPA and control charting, which both use the mean-shift paradigm. *Equation 4.7* provides a further illustration of this, emphasising the independence of errors resulting from the bootstrap process.

$$X_i = m_i + e_i$$

Equation 4.7: Formula that represents the mean-shift model and independent error assumption in bootstrap analysis

In this formula, m_i refers to the mean value at time i. Apart from a tiny set of values of i known as change-points, usually $m_i = m_i - 1$. The random error connected to the i-th value is represented by e_i. The inclusion of a random error term, e_i, signifies the inherent variability in the data at the i-th time point. This error term is assumed to be independent, with a normal distribution, an identical distribution across observations, and a zero mean. The application of this formula adheres to an independent error structure, signifying that the random errors (e_i) are presumed to be independent and distribution-free, at distinct time points. Additionally, although specific distributional characteristics are assigned to e_i for simplicity, the overall approach, guided by bootstrapping, also remains distribution-free. This combination ensures

flexibility and robustness in the statistical analysis, without making firm assumptions about the distribution of the underlying data. In this context, the presence of an independent error structure indicates that each data point's associated errors, are unrelated and statistically independent. This means that the occurrence or magnitude of errors at one time point provides no information about errors at other time points.

Upon detecting a change, an estimation of when the change took place is promptly computed utilising the CUSUM estimator. This formula, expressed in *Equation 4.8* below, provides a quantitative measure for pinpointing the occurrence of the detected change.

$$|S_m| = \max |S_i|, i = 0, \ldots, n$$

Equation 4.8: Calculation of the time at which the change took place

In this equation, S_m in the CUSUM chart corresponds to the farthest point from zero. The final point prior to the change is estimated to be at point m, whilst the subsequent point, denoted as m+1, estimates the initial point following the change (Taylor 2018; Monyeki 2021). This equation states that S_m is equal to the maximum absolute cumulative sum S_i. Therefore, the point where the cumulative sum reaches its maximum magnitude is denoted by S_m. In the context of CPA, a change-point is considered to occur at the time corresponding to just after point S_m, representing a significant deviation in the cumulative sums. The change-point is where the observed values exhibit a notable shift, either an increase or decrease, signalling a change in the underlying process. By identifying when S_m is equal to the maximum cumulative sum, the CUSUM chart helps pinpoint the time when this significant change in the data distribution or process occurred. This is achieved by referring to point m+1, as this represents the first point subsequent to the change-point, as the actual change occurs just slightly after point m (Arif *et al.* 2017).

Following a change-point, the Mean Squared Error (MSE) emerges as the subsequent estimator, providing an alternative approach to pinpointing when the change occurred. This estimator is defined in *Equation 4.9.*

$$\mathrm{MSE(m)} = \min \mathrm{MSE}(i), i=1\ldots|\mathrm{D}|$$

Equation 4.9: Calculation of the time at which the change took place by utilising the MSE

In this formula, MSE (m) represents the MSE for a specific time point m, and the equation seeks to find the minimum MSE across all possible time points i in the dataset D. An indication of when the change most likely occurred is at the time point m, which corresponds to the smallest MSE value (Yan *et al.* 2018). In simpler terms, the formula helps to pinpoint the specific moment in the dataset where the MSE is minimised, and that corresponding time point (m) represents when the change took place. These calculations, namely the CUSUM and MSE estimators, addresses question *ii.* of the CPA approach, by identifying when the change-points occurred.

By utilising a recursive algorithm, CPA identifies various significant change-points within the time-series data (Albrecher *et al.* 2019), whilst the bootstrapping technique is applied to calculate the confidence levels and limits associated with these changes. A backward elimination procedure is then executed to omit any insignificant change-points once the change-points and their corresponding confidence levels have been identified. Removing a change-point causes the neighbouring change-points and their levels of confidence to be re-evaluated, which effectively lessens the likelihood of false detections (Arif *et al.* 2017).

4.5.2 Implementation of the CPA techniques

The Change-Point Analyzer tool which was developed by Taylor Enterprises (n.d), was selected to implement the chosen CPA algorithms in this study. Its selection was further justified by its utilisation in previous studies that also analysed time-series data (Arif *et al.* 2017; Taylor 2018; Monyeki, Naicker and Obagbuwa 2020). This CPA tool's user-friendly interface makes it usable to researchers without extensive programming knowledge, which is typically required for implementing CPA in Jupyter notebook, using the Python programming language (Sharma, Swayne and Obimbo 2016; Law, Endert and Stasko 2020). Additionally, the Change-Point Analyzer tool provides a variety of statistical tests specifically designed for detecting change-points (Taylor Enterprises n.d). This range of tests increases the tool's versatility and ensures its suitability for different types of data (Taylor Enterprises n.d). In contrast, using other tools such as MATLAB, may require more extensive learning and training to effectively employ its capabilities and perform CPA (Vidyullatha *et al.* 2016). However, this tool's simplicity and ease of interpretation contributes to its applicability (Taylor Enterprises n.d). Researchers can easily understand and interpret the results generated by the tool, helping to draw meaningful conclusions from the analysis. Conversely, the previously mentioned

methods to implement CPA may require significant programming knowledge, resulting in a steeper learning curve and increased complexity of use.

4.6 ML techniques for forecasting future CO2E trends

After detecting the change-points, an ensemble ML model will be trained to validate the historical CO2E trends identified by the CPA model and forecast future trends. ML, which is represented as a subset of AI, enables computers to improve over time and learn from past experience, without the need to programme explicitly (Miller *et al.* 2018). ML algorithms facilitate intelligent decision-making and prediction capabilities, by analysing patterns found in data (El Bouchefry and de Souza 2020). There are different categories of ML approaches, each specialised for a particular problem case, such as reinforcement learning, unsupervised learning and supervised learning (Emmert-Streib and Dehmer 2022). The supervised learning approach was selected for this study. In supervised learning, an underlying algorithm examines the link between the input characteristics and the matching output column by using a labelled dataset to train a model (Borovicka *et al.* 2012). There are two types of supervised learning, namely classification, where the model predicts a categorical label, and regression, where the model predicts a continuous numerical value (Sen, Hajra and Ghosh 2020). In this study, supervised regression is employed, which focuses on predicting and understanding the CO2E in SA. The application of regression is particularly suitable for this research, as CO2E data inherently contains continuous numerical values (Ross and Willson 2017).

ML algorithms essentially learn from data repeatedly, improving their models over time, to enhance performance and adjust to changing environments. Because of its versatility, ML is an effective technique for gathering information, identifying trends, and reaching well-informed conclusions in a variety of contexts. In this context, the intricate relationships and non-linear trends prevalent in CO2E, can be easily detected with the use of ML, which is characterised by algorithms that possess the capabilities to recognise complex patterns in the data (Kumar, Mahalanobis and Juday 2005). In addition to offering a comprehensive understanding of evolving emission patterns, neural networks and ensemble learning approaches provide real-time adaptability to environmental shifts. In this study, ML is employed to first validate the trends identified by the CPA, providing a robust validation of the notable changes in CO2E. This technique is adept at managing a wide range of factors, as it can thoroughly analyse economic data and energy production measures, to support the CPA results. Secondly, ML is essential for predicting future patterns in emissions within the South African context. These

algorithms utilise historical data to forecast emissions, which helps with proactive decision-making and policy development (Liu *et al.* 2022).

The ability of ML techniques to modify model hyperparameters and handle missing values, demonstrates their technical edge (Sayeed, Ahmad and Peng 2022). Imputation methods enable ML models to efficiently handle partial datasets, guaranteeing that the analysis is based on a more precise depiction of the available data (White, Royston and Wood 2011). By fine-tuning the parameters of prediction models, it is possible to gain a more precise understanding of the complexities associated with CO2E trends. Additionally, the scalability potential of ML is advantageous when it comes to handling increasing data volumes (Rengachary *et al.* 2023). The efficiency with which ML algorithms handle large environmental datasets makes the analysis of long-term historical records of CO2E in SA feasible.

4.6.1 Overview of traditional ML algorithms

The AdaBoost regressor was benchmarked against several traditional ML algorithms, including Linear Regression, Polynomial Regression, Bayesian Linear Regression, and KNN Regression, to evaluate the most effective method for forecasting CO2E. In the following subsections, these four ML algorithms are defined and explained along with their architectural frameworks.

4.6.1.1 Linear regression

Linear regression is a fundamental statistical method used in ML for predicting a target variable based on one or more input variables (Mali 2024). It assumes a linear relationship between the independent or input variables and the dependent or target variable (Sravani and Bala 2020). The primary goal of linear regression is to identify the best-fitting linear equation that minimises the prediction error (Mali 2024). Linear regression can be divided into two main types: simple linear regression and multiple linear regression (Yilu 2022). Simple linear regression involves a single independent variable and a dependent variable, where the relationship between these variables is modelled using a straight line. For instance, the relationship between CO2E and a single predictor such as energy production can be analysed using simple linear regression. Multiple linear regression, on the other hand, involves multiple independent variables predicting a single dependent variable, such as CO2E predicted using economic growth and energy consumption. The general form of a linear regression equation is represented in *Equation 4.10* below (Rong and Bao-Wen 2018).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

Equation 4.10: Linear regression model representation

In this equation, y is the dependent variable, β_0 is the intercept of the regression line, $\beta_1, \beta_2, \cdots, \beta_n$ are the coefficients of the independent variables $x_1, x_2, \cdots, x_n$, and ϵ represents the error term accounting for the deviation of observed values from predicted values (Rong and Bao-Wen 2018). The intercept (β_0) represents the expected value of y when all predictors are zero, serving as the baseline level of the dependent variable. The coefficients ($\beta_1, \beta_2, \cdots, \beta_n$) indicate the change in the dependent variable for a one-unit change in the respective independent variable, holding all other variables constant. Each coefficient reflects the strength and direction of the relationship between the independent and dependent variables (Rong and Bao-Wen 2018). The error term (ϵ) captures the differences between the observed and predicted values, accounting for the variability in y not explained by the linear relationship with the predictors or input variables.

The objective of linear regression is to estimate the coefficients ($\beta_0, \beta_1, \beta_2, \cdots, \beta_n$) such that the sum of the squared differences between the observed and predicted values, known as the Residual Sum of Squares (RSS), is minimised (Yilu 2022). This estimation is typically performed using the OLS method. The visual representation of the linear regression algorithm's architecture and equation is depicted in *Figure 4.3* below.

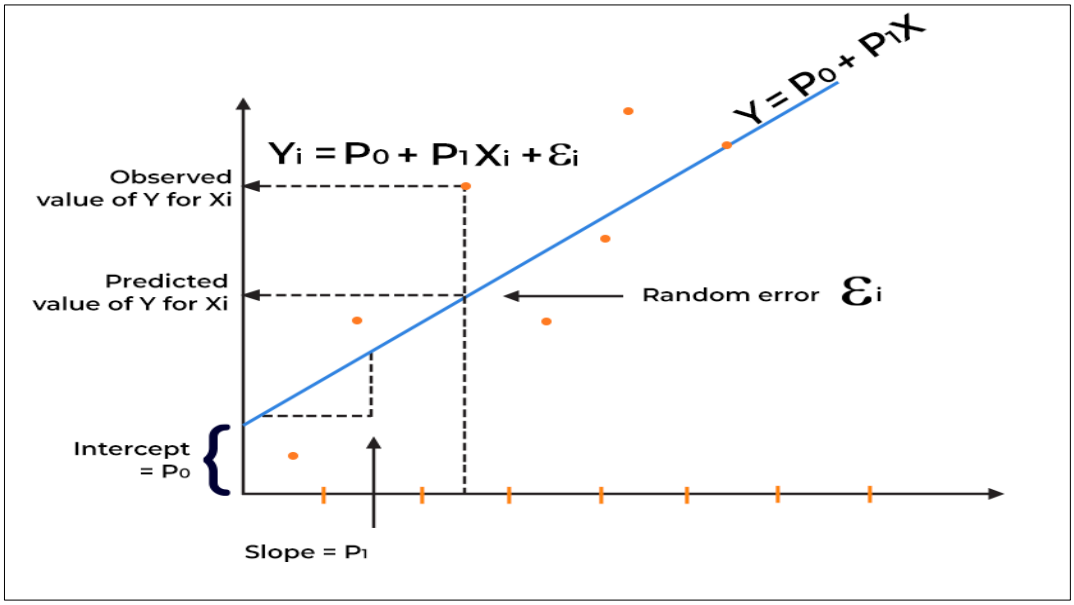


Figure 4.3: Linear regression algorithm architecture and equation, illustrating the linear relationship between input variables and the predicted output, obtained from Kanade (2023)

Linear regression is widely used due to its simplicity, ease of interpretation, and effectiveness for many real-world applications (Kanade 2023). Additionally, linear regression serves as a baseline model against which more complex ML models, such as AdaBoost regression, are benchmarked, as done in this study. This allows for the evaluation of whether advanced models provide significant improvements in predictive accuracy over the simple linear model (Sravani and Bala 2020). Furthermore, the research conducted by Kapoor *et al.* (2022) and Li and Sun (2021) successfully utilised linear regression models to predict CO2E, demonstrating its effectiveness in capturing the relationships between emissions and influencing factors. This supports the use of linear regression in this study for modelling CO2E, ensuring robust and comparable results.

4.6.1.2 Polynomial regression

Polynomial regression is a type of regression analysis where the relationship between the independent variables and the dependent variable is modelled as an nth degree polynomial (Maulud and Abdulazeez 2020). Unlike linear regression, which assumes a straight-line relationship between the input variables and the output, polynomial regression can capture the non-linear relationships by introducing polynomial terms of the predictor variables (Mısır and Akar 2022). This approach is particularly useful when the data exhibits a non-linear dependence, as it allows for more flexibility in modelling complex data patterns (Mısır and Akar 2022).

In polynomial regression, the original features or independent variables are transformed into polynomial features of a specified degree (Agrawal 2024). This allows the model to fit more complex, non-linear relationships in the data. When the relationship between the variables is not adequately represented by a straight line, polynomial regression is an effective method to capture the non-linear patterns present in the data (Simplilearn 2023; Agrawal 2024). *Equation 4.11* presents the general form of the polynomial regression equation (Mısır and Akar 2022; Agrawal 2024).

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \epsilon$$

Equation 4.11: Polynomial regression model representation

In this equation, y represents the dependent variable, which is the outcome the researcher aims to predict. The term x stands for the independent variable, which is the input feature. The coefficients $\beta_0, \beta_1, \beta_2, \cdots, \beta_n$ are the parameters of the model that need to be estimated from the

data. These coefficients determine the weight and influence of each polynomial term on the prediction. The terms x^2, x^3, ..., x^n are the polynomial features, where n is the degree of the polynomial, allowing the model to capture more complex, non-linear relationships (Mısır and Akar 2022; Agrawal 2024). The error term ϵ accounts for the variability in the data that the model does not explain. This polynomial regression equation thus transforms the original input variable x into multiple polynomial features, enabling the model to fit a wide range of non-linear data patterns effectively. *Figure 4.4* illustrates the architecture and equation of the polynomial regression algorithm.

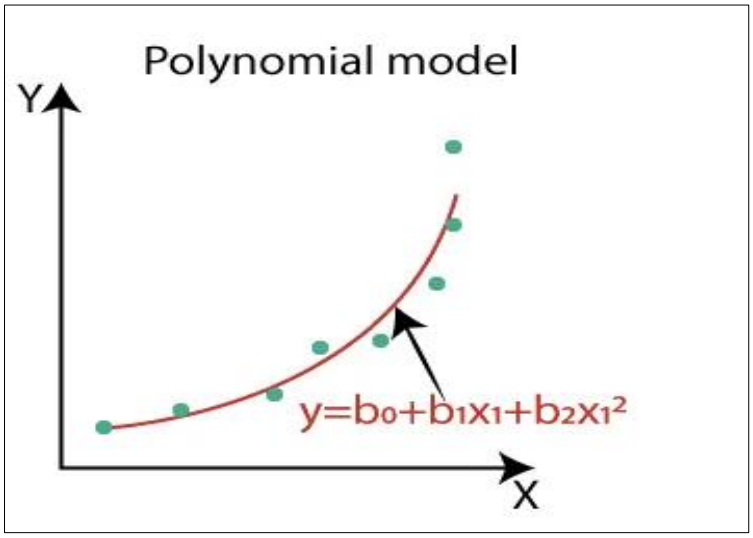


Figure 4.4: Polynomial regression architecture and equation representing a non-linear relationship between input variables and the predicted output, as depicted by Mahata (2024)

Given that CO2E trends may exhibit non-linear relationships with influencing factors such as economic growth, industrial activity, and energy production, polynomial regression can effectively capture these complex patterns. By applying polynomial regression, this study can identify and model the non-linear interactions between CO2E and their predictors.

4.6.1.3 Bayesian Linear regression

Bayesian linear regression is a statistical method in ML that extends traditional linear regression by incorporating Bayesian inference (Koehrsen 2018). Unlike OLS regression, which provides point estimates for the regression coefficients, Bayesian linear regression estimates the posterior distribution of the coefficients (Kong *et al.* 2020). This approach allows for incorporating prior knowledge about the parameters and updating these probabilistic estimates with observed data, leading to a more flexible and robust model (Baldwin and Larson 2017).

In Bayesian linear regression, the coefficients are treated as random variables with their own probability distributions. The method begins with prior distributions for these parameters, reflecting any initial assumptions before observing the data (Koehrsen 2018; Kong *et al.* 2020). When the data is observed, the Bayes' theorem is then applied to update the prior distributions to posterior distributions (Baldwin and Larson 2017; Kong *et al.* 2020). The posterior distribution combines the prior information with the likelihood of the observed data, providing a comprehensive understanding of the parameters' uncertainty (Kong *et al.* 2020). The relationship between the dependent variable y and the independent variables X is presented in *Equation 4.12* below (Koehrsen 2018; Kong *et al.* 2020).

$$y = X\beta + \epsilon$$

Equation 4.12: Bayesian linear regression model representation

In this equation, y represents the observed outcomes, X is the design matrix of input features, β is the vector of regression coefficients, and ϵ denotes the error terms, typically assumed to follow a normal distribution with mean 0 and variance σ^2. The regression coefficients β are treated as random variables with prior distributions, commonly assumed to follow a multivariate normal distribution, as depicted in *Equation 4.13* (Koehrsen 2018; Kong *et al.* 2020).

$$\beta \sim N(\mu_0, \Sigma_0)$$

Equation 4.13: Prior distribution of regression coefficients in Bayesian Linear Regression

In this notation, $\mu_0 \; and \; \Sigma_0$ are the prior mean and covariance matrix, respectively. The likelihood of the observed data given the parameters is expressed in *Equation 4.14* below (Koehrsen 2018; Kong *et al.* 2020).

$$P(y \mid X, \beta, \sigma^2) = N(y \mid X\beta, \sigma^2 I)$$

Equation 4.14: Likelihood function in Bayesian Linear Regression

Bayes' theorem is employed to revise the prior distributions of the model coefficients using the observed data, resulting in the posterior distribution of these coefficients, as illustrated in *Equation 4.15* (Koehrsen 2018; Kong *et al.* 2020).

$$P(\beta \mid y, X) = \frac{P(y \mid \beta, X) * P(\beta \mid X)}{P(y \mid X)}$$

Equation 4.15: Bayes' theorem for calculating the posterior distribution of coefficients in Bayesian regression

This posterior distribution reflects both the prior distribution and the likelihood of the observed data, providing a comprehensive estimate of the regression coefficients and their uncertainty. Additionally, the selection of Bayesian Linear Regression in this study is justified by Chen, Mihara and Wen (2018) and Koca Akkaya and Akkaya (2023), who demonstrated its effectiveness for CO2E modelling in their respective studies.

4.6.1.4 K-Nearest Neighbors (KNN) regression

KNN regression is a non-parametric ML algorithm used for predicting continuous outcomes based on the similarity between data points (Singh 2024). Unlike parametric models that assume a specific form for the relationship between features and the target variable, KNN regression makes predictions by considering the k-nearest neighbors to a given input sample (Al-Shehri *et al.* 2017; Song *et al.* 2017). In other words, a new data point is assigned an output based on how closely it matches the training set's data points. This algorithm identifies the k closest data points in the training set to a given input sample by calculating distances, commonly using metrics such as the Euclidean distance (Song *et al.* 2017; Beskopylny *et al.* 2022). To add on, KNN regression operates on the principle that similar data points should have similar target values.

The first step in KNN regression is to measure the distance between the input sample and all other samples in the training dataset (Song *et al.* 2017; Singh 2024). A common distance metric used in KNN is the Euclidean distance (Beskopylny *et al.* 2022). *Equation 4.16* presents the formula to compute the Euclidean distance between two data points x_i and x_j (Song *et al.* 2017; Singh 2024).

$$\mathrm{d}(x_{i,}, x_j) = \sqrt{\sum_{m=1}^{M} (x_{i,\mathrm{m}} - x_{j,\mathrm{m}})^2}$$

Equation 4.16: Formula for calculating the Euclidean distance between two data points

In this formula, $\mathrm{d}(x_{i,}, x_j)$ represents the distance between data points x_i and x_j, $x_{i,\mathrm{m}}$ and $x_{j,\mathrm{m}}$ are the m-th features of points x_i and x_j, respectively, and M denotes the total number of

features. After calculating these distances, the algorithm selects the k nearest neighbors based on the smallest distances (Al-Shehri *et al.* 2017). This step is critical as the value of k influences the model's performance. The final step is to predict the target value for the input sample by averaging the target values of the k nearest neighbors (Song *et al.* 2017; Beskopylny *et al.* 2022; Singh 2024). The prediction for the target value $\hat{y}(x)$ is given in *Equation 4.17* below (Song *et al.* 2017; Singh 2024).

$$\hat{y}(x) = \frac{1}{k}\sum_{i=1}^{k} y_i$$

Equation 4.17: K-Nearest Neighbors prediction formula for the target value

In this equation, $\hat{y}(x)$ is the predicted target value for the input sample, and y_i represents the target values of the i-th nearest neighbors. To add on, the selection of KNN for modelling CO2E in this study is justified, as Chimphlee and Chimphlee (2023) and Glavas (2024) also employed this algorithm in their respective studies to model CO2E. *Figure 4.5* presents a visual representation of the workflow of the KNN algorithm.

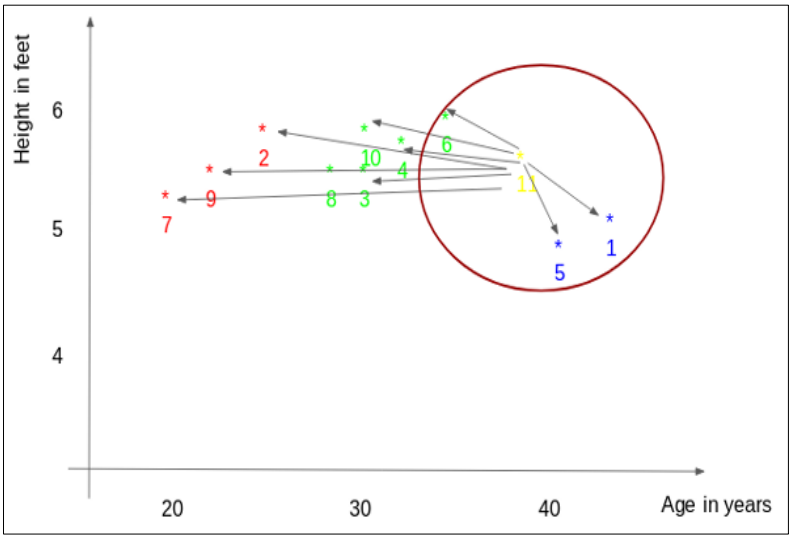


Figure 4.5: Workflow of the KNN algorithm, sourced from Singh (2024)

4.6.2 Overview of ensemble learning and AdaBoost regression

Ensemble learning is a ML approach that deviates from traditional methods by integrating the predictions of several models, to enhance overall performance (Daumé 2017). This technique stands out for its ability to mitigate overfitting, boost stability and enhance predictive accuracy (Guan 2021). By integrating diverse models, ensemble methods capture various aspects of

underlying data patterns, making them more effective and resilient. The key advantage of the ensemble approach lies in the low likelihood of identical mistakes being made by every model that is utilised (Hambali *et al.* 2019).

In this study, the application of ensemble methods offers several advantages. It contributes to the validation of past trends identified by the CPA model and forecasts future trends in CO2E. Through the collaboration of different models, ensemble learning enhances the overall reliability and accuracy of predictions (Özöğür-Akyüz, Windeatt and Smith 2015). Additionally, it considers the intricacy of the data and assists the model in effectively generalising unseen data. This makes ensemble learning a valuable tool for understanding the intricate and dynamic patterns of CO2E trends, ensuring a comprehensive analysis for effective decision-making.

To develop an ensemble regression model for forecasting future CO2E trends, this study will employ the AdaBoost regression algorithm, a technique that is a type of boosting method. Boosting is an ensemble learning technique that builds a powerful predictive model by merging several weak learners, which are usually decision trees (Martinez and Gray 2014). It operates sequentially, with each weak learner concentrating on the errors made by its precursors, thereby improving overall accuracy. AdaBoost is a specific boosting algorithm that assigns weights to each weak model based on its performance, with better-performing models receiving greater attention (Gu and Angelov 2021). It is an improvement on the boosting technique that builds a model with higher prediction capabilities by combining several weak learning models. The AdaBoost ensemble approach has garnered significant interest in the field of ML techniques, due to its low error rates and efficient performance, particularly with noisy datasets. (Han and Sim 2008; Shahraki, Abbasi and Haugen 2020). This method also has the benefit of requiring fewer input parameters and little to no prior knowledge about the poor learner.

Therefore, the AdaBoost Regressor is well-suited for this study, as it combines multiple weak regression models to create a robust ensemble (Gao *et al.* 2010). Also, it adapts to instances with larger prediction errors, which effectively can help capture the complex environmental relationships and improve the accuracy of validating the identified carbon emission trends in SA (Dawari 2022). Additionally, this robust ensemble can enhance the accuracy of forecasting future CO2E trends within this context. Improving the accuracy of forecasting CO2E is crucial, as it enables policymakers, researchers, and stakeholders to make more informed decisions and

develop effective strategies, to mitigate CO2E and address environmental challenges in SA. *Figure 4.6* below illustrates the generic architecture and workflow of the AdaBoost algorithm.

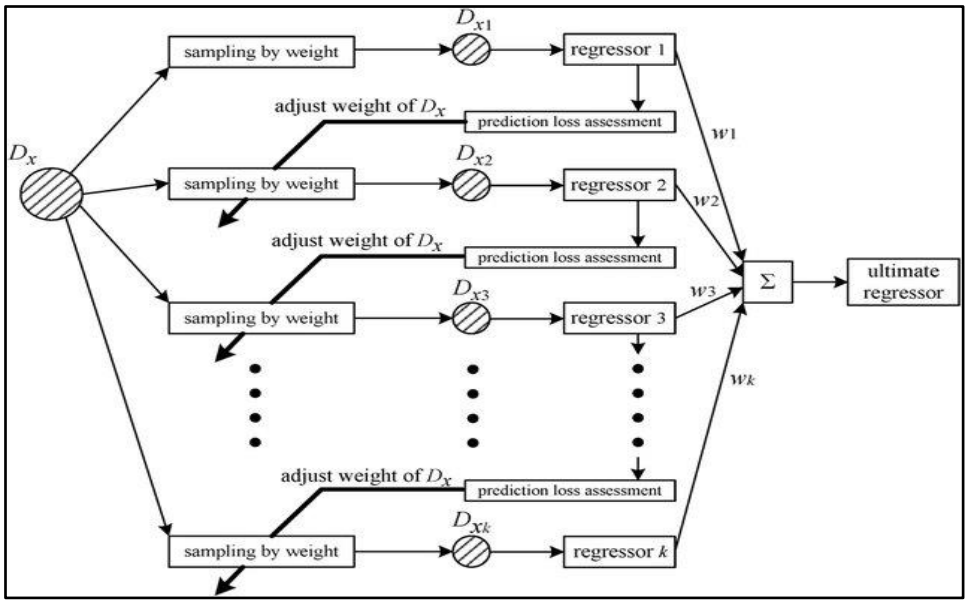


Figure 4.6: AdaBoost Regression algorithm workflow, as depicted by Jin *et al.* (2020)

As demonstrated in the figure above, the AdaBoost algorithm functions in a sequence of iterations (Jin *et al.* 2020). This algorithm starts by processing the input dataset, as denoted by D_x. In D_x, every data point is initially given the same weight, which affects the likelihood that it will be sampled. After that, the first regressor is trained using the training set D_{x1}, which is produced by sampling a subset from D_x. A prediction loss assessment is then utilised, with the intention of evaluating the trained regressor and determining a weight (W) to assign to it. The weight value assigned to the first regressor is denoted by W1. This assessment's primary aim is to modify the weights for the initial dataset (D_x). The greater the prediction error for a certain instance of the initial dataset, the higher the weight allocated to that instance. According to the weight adjustment method, a data point with a high prediction error indicates a weakness in the regressor, which will require greater attention in the following round. Each instance of D_x has a different probability of being chosen for the subsequent round, based on changes in the weights of the dataset. It is therefore possible to train a specific number of regressors (k), by iteratively applying the aforementioned process. These trained regressors often perform poorly, but this algorithm's effectiveness is demonstrated by combining these poor regressors into a weighted combination, to create a strong regressor with better prediction capabilities.

In essence, to improve prediction accuracy, the AdaBoost regressor employs an iterative process, essentially functioning as a meta-estimator. The first step in this approach is to fit a

regression model to the initial dataset (D_x). In subsequent rounds, the regression model is fitted to further subsets of the same dataset, and the weights of the occurrences are modified in accordance with the error rate of the present predictions (Min and Luo 2016). The goal of this iterative process is to produce a strong ensemble model that outperforms the performance of the individual k regressors. As illustrated in *Figure 4.3* above, this iterative process samples k training sets, denoted as D_xk, from the main dataset D_x, based on the number of predefined trained regressors (k). This is because each individual regressor is trained on a distinct subset of the original dataset. Consequently, k weights, represented by Wk, are added to the weighted sum, which is also associated with the number of defined trained regressors (Jin *et al.* 2020).

4.6.3 Training the AdaBoost regression model

The AdaBoost regression model was trained utilising the Jupyter Notebook and Python programming language. Jupyter Notebook is an intuitive and interactive tool that provides the smooth integration of code, explanatory text, and visualisations, to help in training ML models (De Marchi and Soille 2018). Its versatility facilitates an exploratory and iterative approach, crucial in the initial stages of data analysis. On the other hand, Python, with its extensive frameworks and libraries to conduct ML, provides a convenient environment for implementing and experimenting with various algorithms, resulting in its widespread use in ML research (Bilina and Lawford 2012). Furthermore, libraries such as Pandas facilitate effective data processing and analysis, whilst scikit-learn, a popular library, offers reliable tools for ML modelling (Stančin and Jović 2019). The supportive online community and extensive documentation surrounding Python, make it the perfect option for researchers who intend to conduct research in ML (Millman and Aivazis 2011).

The training data for the model underwent preparation through the utilisation of a DataFrame, which is a tabular data structure within the Pandas library offered by Python, that facilitates two-dimensional data organisation (Krikler *et al.* 2019). To facilitate simple processing and analysis, the data is arranged into columns and rows using a DataFrame (Embarak 2018). Pandas is widely used for data manipulation and cleaning due to its simplicity and efficiency. It is a crucial part of the data pre-processing stage of ML, as it offers strong capabilities for handling incomplete data, combining datasets, and filtering out irrelevant data (Rajagopalan 2021). The columns in the DataFrame represented the input features and output label of the model. Specifically, the input features included the year under observation, *Electricity net generation (Billion kWh)*, and *GDP growth (annual %)* columns, whilst the output label was

the *CO2 emissions (kt)* column. Appropriately assigning these columns appropriately ensures that the data is structured correctly for training the regression model. The study focused on socio-economic indicators in SA from 1995 to 2020, resulting in 26 data points to analyse the trends in CO2E. Despite having more data points for electricity generation and economic growth, all the variables were required to align with the number of data points available for the CO2E variable. This is done to prevent any mismatch in data points between the variables. This alignment is crucial for the model to make accurate predictions, emphasising the importance of consistency in the dataset, as the model can only make predictions with balanced data points for the input columns and output column.

For model evaluation, the dataset was split into training and testing sets using an 80-20 split ratio. This signifies that 80% of the data was utilised for training the model, whereas the remaining 20% was employed for testing its performance. By training the model on a subset of the data and assessing its generalisation on unseen data, this commonly used procedure aids in evaluating the model's performance. The 80-20 split was chosen due to its balance between affording a sufficient amount of data for training, ensuring the model captures underlying patterns, and setting aside a sizeable portion for testing, allowing for a thorough evaluation of its forecasting capabilities using new data (Tian *et al.* 2014; Kebonye 2021). The **train_test_split** function from the scikit-learn library was employed for this purpose. As previously mentioned, this library is a widely employed ML library in Python that provides efficient tools for data analysis and modelling. Within scikit-learn, **model_selection** is a sub-module that offers a comprehensive suite of tools for model assessment and selection, including the **train_test_split** function, which divides the dataset into segments of training and testing data (Hao and Ho 2019).

The AdaBoost regression model was then created using the **AdaBoostRegressor** class from the **sklearn.ensemble** module. The ensemble module within scikit-learn specifically focuses on ensemble learning methods, making it an ideal resource for implementing algorithms like AdaBoost (Pedregosa *et al.* 2011). The **AdaBoostRegressor** class encapsulates the functionality needed to solve regression tasks utilising the AdaBoost algorithm (Hastie *et al.* 2009). Following the training process, the regression model is capable of predicting CO2E on the scale of kt. To individually assess their effects on model performance, hyperparameter tuning and model validation techniques were separately employed during the training process. These techniques involve either adjusting the model's parameters to find the optimal

configuration or assessing its performance on subsets of validation data. The details of these techniques are discussed in the subsequent section.

4.6.4 Hyperparameter tuning and model validation techniques

This study compares the effectiveness of two techniques aimed at enhancing model performance, including hyperparameter tuning using k-fold cross-validation and repeated holdout model validation. Hyperparameter tuning is the process of systematically searching for the best combination of hyperparameter values for a ML model, to improve the predictive accuracy of the model (Shahhosseini, Hu and Pham 2022). These hyperparameters are configuration settings external to the model, that cannot be discovered by analysing the data. The development of dependable ML models requires the precise tuning of hyperparameters (Shatnawi *et al.* 2022). The process of tuning involves optimising these hyperparameters to significantly enhance the model's performance (Bardenet *et al.* 2013; Bergstra, Yamins and Cox 2013). The correct choice of hyperparameter values can result in faster convergence, greater generalisation to unseen data, enhanced flexibility for new data, and increased model accuracy (Binder *et al.* 2020). Conversely, poorly selected hyperparameters can lead to either underfitting or overfitting, which makes it more difficult for the model to identify the underlying trends in the data (Grosse *et al.* 2020). The enhanced performance is crucial for this study, as accurate models are essential for reliably identifying and forecasting CO2E trends. The correct selection of hyperparameters enables the ML models to capture intricate patterns within socio-economic indicators, leading to more trustworthy insights. Improved model accuracy, achieved through hyperparameter tuning, enhances the overall effectiveness of this research, by providing a solid basis for informed policy decisions related to carbon emissions.

To improve the process of manual tuning, several methods of automating the selection of hyperparameters have been devised, such as k-fold cross-validation (Cerqueira, Torgo and Mozetič 2020). The dataset is divided into k folds, or subsets, when utilising this validation technique (Mabuni and Babu 2021). The model is then trained k times, each time using a different fold as the test set, with the training set consisting of the remaining data (Rodriguez, Perez and Lozano 2009). This process helps assess the model's performance across different subsets, providing a more robust estimate of its generalisation ability. K-fold cross-validation assesses a model's performance repeatedly on various subsets of the original dataset, which is a complementary technique to hyperparameter tuning. Within this cross-validation framework, the model undergoes training and evaluation multiple times, each with a different set of

hyperparameter values. This process explores a range of hyperparameter configurations, allowing for a thorough examination of the model's behaviour under different scenarios. The model is then assessed on each fold using a specific hyperparameter set, and the performance metrics are computed. This iterative process ensures that the model's performance is scrutinised across diverse subsets of the data, providing a more illustrative evaluation.

The performance metrics obtained from each fold are then averaged, producing a cross-validated metric for a particular set of hyperparameter values. By averaging the performance metrics over these iterations, K-fold cross-validation provides a weighted influence of each hyperparameter configuration, reflecting its impact on the overall performance of the model (Rodriguez, Perez and Lozano 2009). The main objective is to identify the hyperparameter configuration that yields the best cross-validated performance (Morales-Hernández, Van Nieuwenhuyse and Rojas Gonzalez 2023). In essence, this process concurrently assesses the effects of several hyperparameter value combinations, across multiple folds of the dataset. This concurrent checking ensures that the chosen hyperparameter set, consistently produces optimal performance across the different subsets, contributing to the selection of parameters that enhance the model's generalisation and overall effectiveness (Tsirikoglou *et al.* 2017). *Figure 4.7* further illustrates the hyperparameter tuning with cross-validation process utilised in this study.

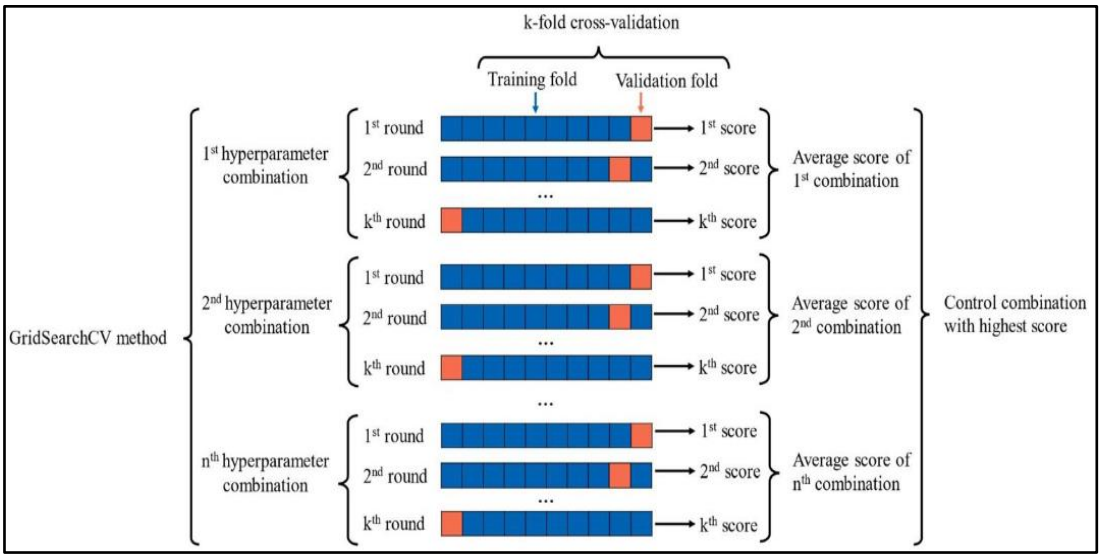


Figure 3.7: Optimising hyperparameters through K-fold cross-validation, obtained from Shatnawi *et al.* (2022)

The **GridSearchCV** class from the scikit-learn library was employed, to perform an exhaustive search over a predefined hyperparameter grid. Grid Search Cross-Validation (GridSearchCV)

is a technique used for hyperparameter tuning in ML (Ghawi and Pfeffer 2019). Its purpose is to systematically search through a predefined hyperparameter grid, evaluating the model's performance for each combination of hyperparameters, using cross-validation (Varoquaux *et al.* 2015). The hyperparameter values are initialised beforehand within the parameter grid, and the **GridSearchCV** class exhaustively tests all possible combinations, selecting the one that yields the best model performance based on a specified scoring metric (Bergstra and Bengio 2012). The number of estimators and the learning rate were the two hyperparameters under consideration. In order to establish the cross-validation technique for a more robust model evaluation, the **RepeatedKFold** class from scikit-learn was used, with five splits and three repeats. Lastly, the best hyperparameter configuration was determined through the search process, and the model with the optimal configuration was fitted to the entire training dataset.

The researcher employed the repeated holdout validation technique as the second technique to enhance model performance. Repeated holdout is a model validation technique where the dataset is randomly divided into training and testing sets multiple times, with each iteration using a different split (Raschka 2018). This process is repeated for a specified number of times, providing an ensemble of evaluation results. The repeated holdout approach adds randomness to the training and testing data splits, to assist in thoroughly evaluating the performance of a model (Lee, Liong and Jemain 2018). It addresses issues related to the variability in evaluating model performance that can arise from a single train-test split. Through the process of averaging performance measurements across several iterations, this technique yields a more dependable assessment of a model's capacity for generalisation. In the context of this study, repeated holdout can help by providing a more reliable evaluation of the models. The randomisation in the data splitting process ensures that the models are evaluated across diverse subsets, providing a robust indication of their ability to generalise on unseen data. This is crucial in ensuring the reliability of past trends and forecasting future trends, enhancing the overall credibility of this study's findings. The repeated holdout process is demonstrated in *Figure 4.8* below.

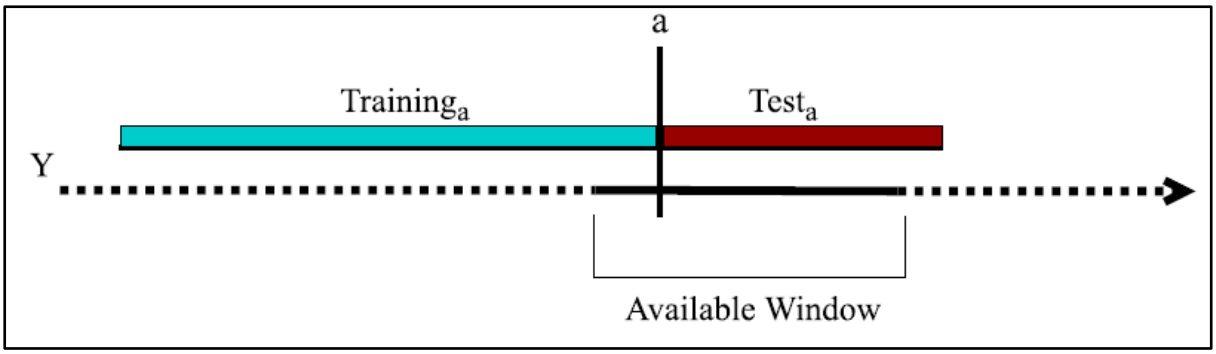


Figure 4.8: An overview of the Repeated holdout process, sourced from Cerqueira, Torgo and Mozetič (2020)

A single iteration of this technique, as depicted in the figure above, selects a random point (a) from the accessible sampling window. This point serves as the boundary between the training set and the testing set. Subsequently, observations preceding this point are utilised for training, whilst observations following this point are employed for testing. For this study, the AdaBoost model was initialised with specific hyperparameters, and arrays were created to store the various regression performance metrics. In the repeated holdout loop, which was executed for six iterations, the dataset was systematically split into training and testing sets for each iteration. Subsequently, the AdaBoost model was trained on the training data, and predictions were generated for the corresponding test data. The performance metrics were then calculated and stored for each iteration. Finally, the average metrics across all six iterations were computed and printed. The previously discussed functions from the scikit-learn library, necessary for both splitting the data and creating the AdaBoost regressor, were the essential components needed for implementing the model in this iterative process.

The hyperparameter tuning and model validation techniques were not only employed to improve model performance but were also aimed at accommodating the non-stationary characteristics inherent in CO2E data. Data that is non-stationary suggests that its statistical characteristics, such as its variance and mean, fluctuate with time (Levendis 2018). In the case of CO2E, the underlying patterns and trends can evolve, due to various factors such as policy changes, technological advancements, or economic shifts (Kim and Kim 2012). The non-stationarity of COE2 data poses a challenge for developing accurate and robust predictive models. By employing hyperparameter tuning with cross-validation, the model can adapt to changing patterns and optimise its parameters, to cater for all the different fluctuations in the data (Xia *et al.* 2018). This helps ensure that the model is flexible enough to capture the evolving trends in CO2E.

On the other hand, repeated holdout validation is crucial for assessing the model's generalisation performance over time (Schorfheide and Wolpin 2012). Non-stationary data may exhibit variations and patterns that are not consistent across different time periods. As a result, the repeated holdout approach, through its random selection of training and testing sets, offers a more thorough evaluation by replicating many scenarios and reflecting the unpredictability inherent in non-stationary data. These techniques ensure that the models are not only accurate but also adaptable to the evolving dynamics of CO2E, enhancing the reliability of the predictions over time.

4.6.5 Performance metrics for model evaluation

In evaluating the performance of the AdaBoost model for predicting CO2E, the following regression performance metrics were utilised: RMSE, MAE and R^2 scores (Tatachar 2021; Padhma 2023). In addition, the researcher introduced a unique accuracy score for assessing the model's performance, as accuracy is typically associated with classification problems (Aurelio *et al.* 2019). Each metric serves a distinct purpose in assessing the model's accuracy and goodness of fit.

4.6.5.1 Root mean squared error (RMSE)

RMSE is a measure of the average magnitude of the errors, between the predicted and actual values (Plevris *et al.* 2022). The formula calculates the square root of the average of the squared differences between each predicted (y_i) and actual ($\hat{y}_i$) CO2E value from the testing dataset (Tatachar 2021). This metric was employed to measure the average magnitude of prediction errors. The formula for calculating the RMSE is represented in *Equation 4.10,* as depicted by Tatachar (2021) and Padhma (2023).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Equation 4.10: Formula for calculating the RMSE

In a regression scenario, RMSE is crucial as it considers both the magnitude and direction of errors, providing a comprehensive view of predictive accuracy. A lower RMSE indicates that the model's predictions are closer to the actual values (Padhma 2023).

4.6.5.2 Mean absolute error (MAE)

MAE calculates the average absolute differences between predicted and actual CO2E values (Plevris *et al.* 2022). The formula sums the absolute differences between each predicted (y_i) and actual ($\hat{y}_i$) value and divides by the number of observations (Tatachar 2021). This metric is calculated using *Equation 4.11* below, as previously demonstrated by Tatachar (2021) and Padhma (2023), which measures the average absolute magnitude of prediction errors.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Equation 4.11: Calculation of the MAE value

This metric is suitable for regression scenarios where understanding the average magnitude of errors without emphasising outliers is essential. A lower MAE suggests better model performance, as it indicates smaller average absolute errors (Padhma 2023). This means that a lower value for this metric denotes less deviation between the actual and projected values.

4.6.5.3 Coefficient of determination (R^2) score

R^2 measures the proportion of the variance in the CO2E data that is captured by the model (Chicco, Warrens and Jurman 2021). The R^2 score is expressed in *Equation 4.12* below, as illustrated by Tatachar (2021) and Chicco, Warrens and Jurman (2021).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Equation 4.12: Formula for computing the R^2 score

The formula compares the sum of squared differences between the predicted (y_i) and actual ($\hat{y}_i$) values, to the sum of squared differences between the actual values and their mean ($\bar{y}$) (Chicco, Warrens and Jurman 2021). The proportion of the dependent variable's variance that can be predicted based on the independent variable, is evaluated using this metric (Plevris *et al.* 2022). Since this metric produces a percentage value, the score to expect can range from zero to one. In regression, the R^2 score is a vital metric, with a higher score closer to one suggesting a better fit. A higher value of this metric indicates that a greater amount of the data variation can be explained by the model (Chicco, Warrens and Jurman 2021). In this study, a

higher R^2 score indicates that the model effectively captures the variation in CO2E, providing reliable predictions.

4.6.5.4 Accuracy

Equation 4.13 calculates the percentage accuracy, by subtracting the MAPE from 100. MAPE is the absolute percentage difference between the predicted (y_i) and actual ($\hat{y}_i$) values in a testing dataset (Chicco, Warrens and Jurman 2021; Padhma 2023).

$$Accuracy = 100 - (\frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \times 100)$$

Equation 4.13: Calculation of the accuracy score

In this formula, n represents the number of observations in the test set. The formula evaluates the accuracy of the model by considering the relative MAPE errors. It also measures how accurately the model predicts CO2E, in relation to the MAE. A higher percentage accuracy indicates better model performance. In standard ML practice, MAPE is commonly used in regression scenarios (Plevris *et al.* 2022). However, in this study, the MAPE formula was manipulated, to create a percentage accuracy metric, as indicated in the above equation.

For CO2E forecasting, achieving lower values for RMSE and MAE, and higher values for the R^2 and accuracy scores is desirable, as this reflects a model that produces predictions that closely align with the actual values. These metrics collectively serve as essential benchmarks for assessing the accuracy and explanatory capabilities of the AdaBoost regression model, in the context of analysing CO2E trends in SA.

In this study, the performance metrics were implemented using the **metrics** submodule from the sci-kit learn library. This submodule provides a set of functions to evaluate the performance of ML models (Gevorkyan, Demidova and Kulyabov 2020). Specifically, the **mean_squared_error**, **mean_absolute_error**, and **r2_score** functions were employed, to implement these performance metrics. To calculate the RMSE, the MSE was first computed using the **mean_squared_error** function. Since sci-kit learn does not have a built-in function for the RMSE score, the square root function from the **math** library was then applied to the MSE result. The **math** library provides basic mathematical functions in Python (Saha 2015). Additionally, the NumPy (np) library was used to formulate the calculation for accuracy, which included the MAPE. The NumPy library is an effective tool for array manipulation and

numerical operations in Python (McKinney 2022). This customised accuracy metric was derived using the **np.abs** and **np.mean** functions, to compute the absolute percentage errors and then calculate the mean of these errors.

4.7 An overview of the CO2E forecasting process

Following the evaluation of the AdaBoost regression model by utilising the relevant performance metrics, the trained model was then required to forecast future CO2E in SA. Given that the model relies on electricity production and economic growth as input features, predicting future carbon emissions entailed obtaining data for these input features. Rather than employing random input data, a systematic approach was adopted using different layers to forecast CO2E, as demonstrated in *Figure 4.9* below.

Layer 1 involved training separate AdaBoost regressors, denoted as AdaBoost model 1 and AdaBoost model 2, by using historical electricity production and economic growth data, respectively. AdaBoost model 1 predicted future electricity production, whilst AdaBoost model 2 forecasted economic growth, within a specified timeframe. The outputs from both models of Layer 1 moved to Layer 2, where they served as input features for the previously trained AdaBoost model.

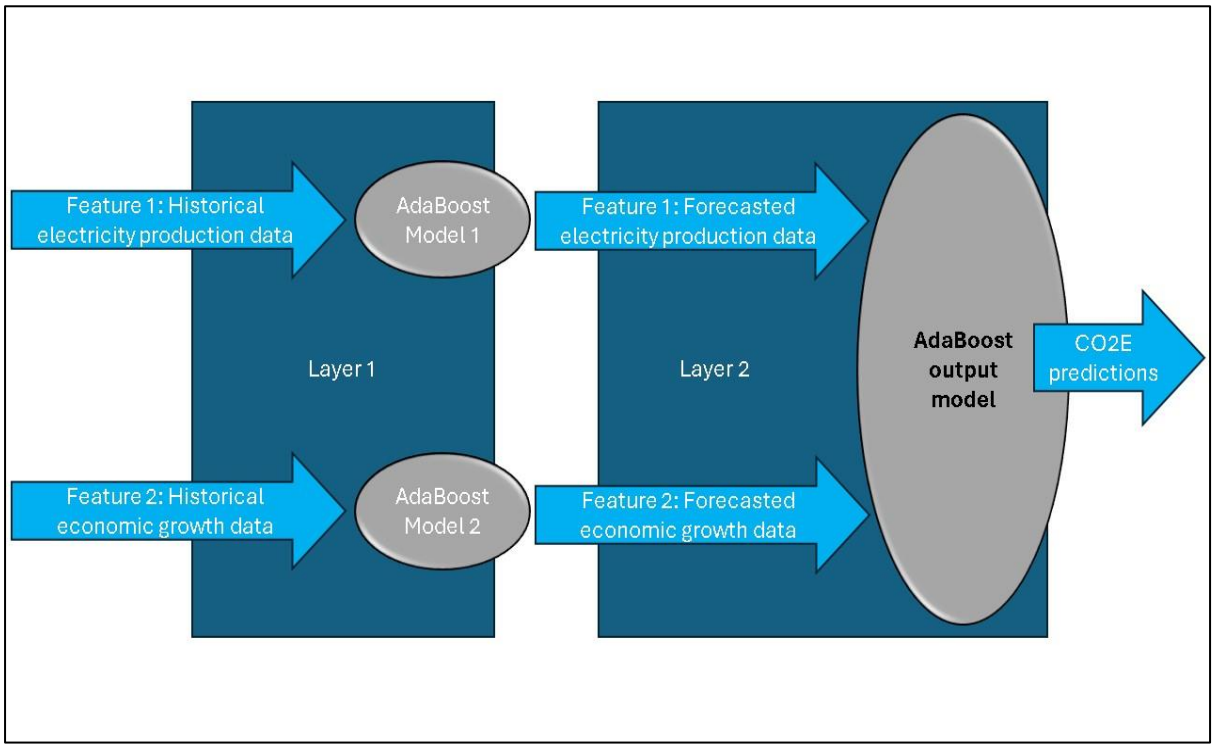


Figure 4.9: The layered approach utilised for forecasting CO2E

The advantage of using this systematic layered approach to generate input data for forecasting CO2E, as opposed to using random input values, lies in the methodical consideration of the relevant factors. In this structured approach, the input features were forecasted individually by the specialised AdaBoost regressors. This allows for a more nuanced understanding and prediction of each contributing factor separately. The benefit to this study is that this comprehensive and structured approach, captures the intricate relationships between electricity production, economic growth and CO2E, providing more accurate and informed predictions compared to an ad-hoc input approach.

4.8 The relationship between CPA and ensemble ML

In this study, CPA and ML play complementary roles. CPA was employed to identify significant changes or breakpoints in the CO2E trends. It helps pinpoint instances where there are abrupt shifts in emission patterns. On the other hand, ensemble ML models, such as AdaBoost, were utilised to validate the past trends generated by the CPA model, and forecast future CO2E trends based on historical data, specifically electricity production and economic growth. The combination of CPA and ML allows for a comprehensive analysis, integrating insights from both statistical change detection and predictive modelling, to understand and anticipate CO2E trends in SA.

In the validation process, the past trends identified by the CPA model were subjected to confirmation through line graphs. These line graphs visualise the historical trends generated by the CPA model, alongside the trends predicted by the AdaBoost model. This visual validation aids in detecting any deviation or concurrence in the trends, contributing to a comprehensive understanding of the accuracy and reliability of the models. The selection of line graphs for validation is justified by their intuitive and accessible nature, providing a clear representation of trends over time (Acartürk 2014). The visual comparison allows for a quick assessment of how well the ML model aligns with the historical trends pinpointed by the CPA model. This validation approach ensures a more thorough validation of past trends and relationships identified in this research (Wang *et al.* 2020). Visualisations offer an intuitive overview, by quickly highlighting any patterns or discrepancies in the trends.

4.9 Chapter summary

This chapter presented the research methods and materials employed to achieve the study's objectives. It began with an overview of the research process, providing quick insights into the study's purpose and workflow. The research design, which opted for quantitative methods and an experimental approach, was tailored to the numerical nature of socio-economic indicators, like CO2E. The data was sourced from the World Bank's Open Data repository, utilising several time-series databases, such as Sustainable Energy for All, Country Climate and Development Report and WDI. The systematic review conducted in Chapter Three, identified economic growth and energy production as the primary contributors to CO2E, and additional studies further supported these findings. Consequently, time-series data related to these major CO2E contributors was then obtained for further analysis in this study, as discussed in this chapter.

In this chapter, the researcher conducted an analytical process that included a meta-analysis, calculations, and figures to meticulously screen and select the variables for this study. Specifically, a comparative analysis of various energy sources, including coal, nuclear and oil, was also carried out to select the most influential variables. Energy produced from coal sources was chosen due to South Africa’s reliance on coal to produce electricity. The final selected variables for this study, included GDP growth, electricity production and CO2E in SA. This analytical process ensured that the GIGO principle was enforced, to ensure data quality. The chapter also delves into the choice and justification of the Change-Point Analyzer tool, which was utilised to implement the CUSUM and Bootstrap algorithms. The architecture and mathematical formulas for these techniques were also discussed. This chapter further discussed the selection of ML and ensemble learning to forecast CO2E trends, specifically AdaBoost regression, with a detailed discussion on its techniques, equations, and architecture provided. However, prior to discussing the AdaBoost algorithm, the chapter presented the architecture and equations of four traditional ML algorithms, including Linear Regression, Polynomial Regression, Bayesian Linear Regression, and KNN Regression, which serve as a benchmark for evaluating the performance of the AdaBoost algorithm in the next chapter.

The process of training the regression model was also outlined, utilising the Jupyter Notebook environment and various Python libraries such as Pandas and scikit-learn. The techniques for enhancing the model performance, such as hyperparameter tuning and model validation through K-fold cross-validation and repeated holdout, were also explored. These procedures

were implemented using specific classes belonging to the sci-kit learn library, such as **GridSearchCV** for hyperparameter tuning and **RepeatedKFold** for configuring the data splits during the cross-validation process. This chapter demonstrated the utilisation of various performance metrics, including RMSE, MAE, accuracy, and R^2 Score, to evaluate model performance. These metrics were implemented using the scikit-learn **metrics** submodule.

The forecasting process was also discussed, which consisted of a layered approach, employing AdaBoost regressors to forecast economic growth and energy production in the first layer. Thereafter, the predictions from this layer were passed to the second layer, where they served as input features for the trained AdaBoost regressor, to forecast CO2E. This chapter concluded by highlighting the synergy between CPA and ML, emphasising how the predictions made by the AdaBoost model, validated the past trends identified by the CPA model, using graphical analysis. The validation process incorporated line graphs for effectively visualising the trends. The subsequent chapter will present and discuss the results obtained from the CPA and ML experiments.

CHAPTER FIVE: RESULTS, ANALYSIS AND DISCUSSION

5.1 Introduction

This chapter presents and discusses the results of both the CPA model and the ML models utilised in this study. The results from the CPA techniques, particularly the CUSUM and Bootstrap methods, are presented to depict change-points in the data and their respective confidence levels. Graphical illustrations are employed to visualise the identified change-points and a table is used to present the confidence levels associated with each change, obtained through the Bootstrap analysis. Initially, each variable under study is examined separately using these visuals, followed by a collective discussion of the CPA results to derive insights.

Subsequently, the ML results are analysed, encompassing all performance metrics discussed in the previous chapter. The results of the trained models that were built with various ML algorithms are presented and compared in a tabular format. Additionally, statistical measures are employed to determine the best-performing model. Furthermore, the results of the repeated holdout and cross-validation techniques are compared, to determine the more efficient technique for improving model performance. This comparison utilises tables containing the statistical properties of the evaluation metric results for both techniques, to quantify the results. Graphs portraying statistical distributions, such as box and whisker plots, are also used to further quantify and visualise the better-performing technique and illustrate its impact on improving the other method, as demonstrated in the study by Mohamed, Patel and Naicker (2023).

The chapter proceeds to measure the change or improvement resulting from the better-performing method, by evaluating the statistical characteristics of the enhancement in each evaluation metric. To aid this process, statistical distribution graphs were also employed to visualise these changes. As a result, implementation guidelines are then formulated to specify when it is appropriate to use the better-performing technique, to enhance model performance over other methods. Lastly, forecasted CO2E predictions for a justified time period are presented to ascertain the future trajectory of CO2E in the country. These forecasts validate the past trends generated by the CPA model, and the chapter concludes with a synthesis of the CPA and forecasted ML trends.

5.2 CPA results

Prior to analysing the results of the CPA model, it is important to understand how to interpret them. A table and two different types of graphs were used to present the CPA findings. The first graph is a modified version of a control chart, with background changes indicating a potential change-point (Arif *et al.* 2017). This illustration also comprises of control limits, which serve as boundaries indicating the expected variability in the data. In this graph, the y-axis represents the quantified variable under study, such as CO2E, electricity production or GDP growth, whilst the x-axis displays the time range in years (Monyeki, Naicker and Obagbuwa 2020). The second graph illustrates the CUSUM chart, which identifies the precise occurrence of a change-point.

Figures 5.1 to *5.8* showcase the CPA results of each variable under study. Each variable or socio-economic indicator is presented through a CPA graph, resembling a modified control chart, displaying the CPA results along with control limits and background changes (Taylor 2018). Following the CPA graph, a corresponding CUSUM chart illustrates the timing of the identified change-points, confirming their occurrence. Additionally, a Bootstrap table is provided, indicating the confidence level associated with each change (Arif *et al.* 2017). As a result, for each variable analysed, two graphs and a Bootstrap table are included for a thorough examination.

In the modified control chart, change-points are observable, with the blue region indicating a background change, denoting the occurrence of a change-point (Arif *et al.* 2017). The addition of two lines to the individual chart introduces control limits, represented by a red line indicating upper and lower bounds. Outliers, which are the points surpassing these limits, indicate the maximum expected variation assuming no change has taken place (Taylor Enterprises n.d). The points exceeding these boundaries signify a change. However, these control limits are based on a normal distribution and may not suit all datasets. If every data point is included within this range, it indicates that the model fully accounts for the variability in the data. As previously mentioned, any shifts in the shaded background from yellow to blue, indicate changes in the data.

Figure 5.1 below presents the CPA results of electricity net generation from power plants in SA, spanning the period from 1980 to 2020. The analysis reveals significant changes in electricity production over time, pinpointing the specific years where notable shifts occurred.

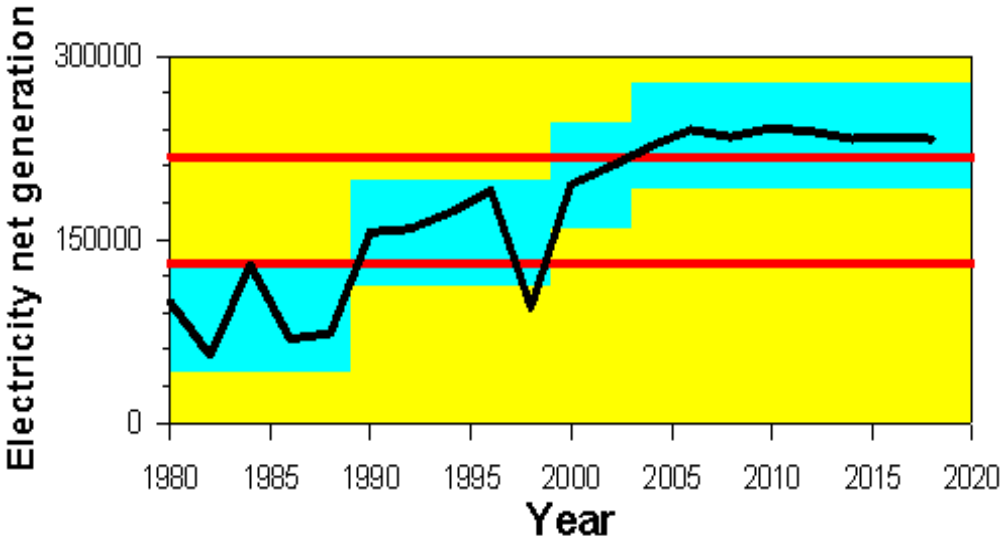


Figure 5.1: CPA of Electricity net generation in SA from 1980 to 2020

The first change-point occurred in 1990, indicating a sharp increase in electricity production, following a discernible increase observed from 1986 to 1990. Subsequent to this change-point, production continued to rise, reaching another peak in 1994. This increase can be attributed to the end of apartheid in SA, which led to economic reforms and increased industrialisation (Simon 2001). With the dismantling of apartheid policies, there was a surge in investments, particularly in the energy sector, to meet the growing demands of a transitioning economy (Nowak 2005). Additionally, the country began to focus on expanding its energy infrastructure, including the construction of new power plants and the improvement of existing ones, to support industrial growth and provide electricity to previously underserved communities (Baker 2017).

The next change-point emerged in 2000, illustrating another increase in electricity production. However, there was a notable decline in production from 1996 leading up to this change-point. Following the second change-point in 2000, electricity production experienced a rapid peak until 2004, when yet another change-point occurred, signifying another increase. The increase in electricity production around 2000 can be linked to several factors, including economic growth, population expansion, and increased demand for electricity in SA (Ritchie, Roser and Rosado 2022). Since 2000, SA has experienced a 30% increase in energy demand, validating the observed trends in electricity production and highlighting the nation's growing need for sustainable energy solutions (Kohler 2014). During this period, SA experienced a period of economic stability and growth, which fuelled higher energy consumption, particularly in

industries such as mining, manufacturing, and telecommunications (South African Government 2002).

Post-2004, electricity production maintained a relatively steady trend, with a slight drop observed in the period of 2007 to 2008. To validate the change-point detected in 2004, Eskom, which is South Africa's primary electricity utility, launched a significant capacity expansion program to address the growing demand for electricity in the country (Eskom n.d). This program involved the construction and commissioning of new power plants and the expansion of existing ones to increase electricity generation capacity (Rafindadi and Ozturk 2017). However, the load-shedding crisis in SA primarily caused a drop in electricity production from 2007 to 2008 (Thompson 2023). Nonetheless, the data remained consistently above the upper control limit, indicating a constant trajectory in electricity generation. Since 2008, South Africa's electricity production has maintained a relatively constant trajectory with minimal growth. This constant trend is largely attributed to the ongoing load-shedding cycles, persisting to the present day, which have hindered significant expansion in production (Herbst and Lalk 2015). Despite efforts to address infrastructure challenges and expand generation capacity, the history of the load-shedding crisis has continued to impact the reliability and stability of the power grid (Herbst and Lalk 2015).

The cumulative sum at each time point or data point is depicted in the CUSUM chart, where a change-point is denoted by the background of the chart shifting from yellow to blue (Monyeki, Naicker and Obagbuwa 2020). An upward trend suggests a time period where the data points are above the overall average, whilst a downward trend indicates that the data points are below the average. Furthermore, straight-line trends denote periods without change, whilst abrupt shifts in direction signal a change in values (Taylor Enterprises n.d). *Figure 5.2* presents the CUSUM chart, which illustrates the electricity net generation from power plants in SA for the period of 1980 to 2020. The chart effectively visualises the cumulative sum over time, aiding in the identification of change-points and trends in electricity production.

The first change-point occurred in 1990, indicated by the change in background from yellow to blue. Following this change-point, the graph depicts a slight increase in the cumulative sum until 1996, suggesting a period of relatively consistent growth in electricity production. The next change-point is observed in 2000, coinciding with an abrupt increase in electricity production.

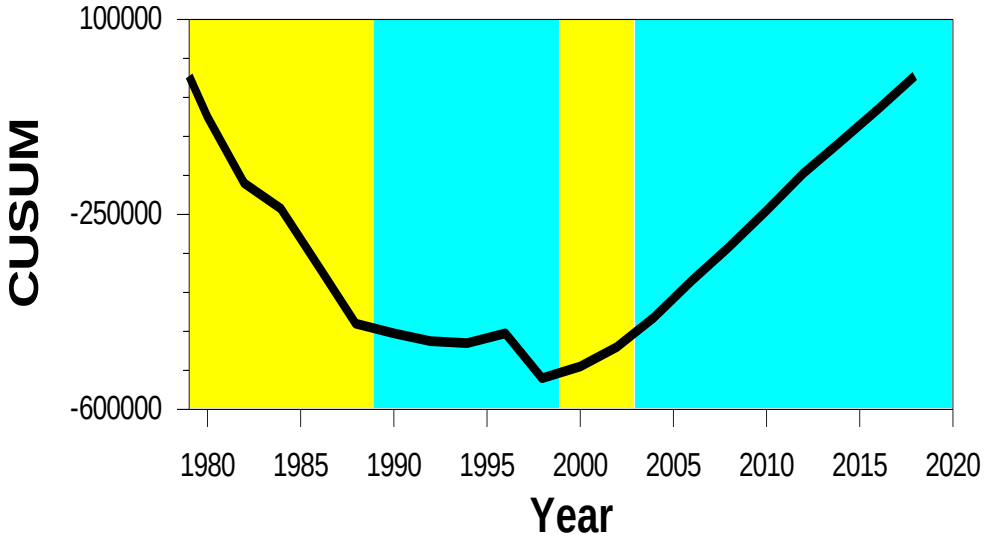


Figure 5.2: CUSUM plot of electricity net generation in SA from 1980 to 2020

However, preceding this change-point, there has been a noticeable decline in production since 1996, as evidenced by the slight drop in the cumulative sum. Following the second change-point in 2000, the cumulative sum experiences a sustained period of rapid growth, extending from 2004 which is the next change-point, to the end of the time period. This is observed by the continuous upward trend following the change-point, which indicates that the data is consistently above the overall average. This upward trend indicates a sustained increase in electricity production exceeding the overall average, suggesting a period of robust growth and heightened energy demand. Notably, the CUSUM chart validates and confirms the change-points identified in the previous CPA graph. The CUSUM chart mirrors the shifts observed in the CPA graph, demonstrating consistency and reinforcing the reliability of the identified change-points. This alignment between the two charts provides confidence in the accuracy of the CPA and enhances the interpretation of the data.

The results of the Bootstrap analysis are displayed in a tabular format. Each change-point is accompanied by a confidence level or percentage, signifying the certainty that the change actually transpired (Monyeki, Naicker and Obagbuwa 2020). Additionally, a confidence interval for the timing of the change-point is provided, indicating the accuracy of pinpointing the time in which the change occurred (Taylor 2018). For all confidence intervals, a 95% level of confidence is used. Furthermore, the table provides further details regarding each change, including the average value of the variable before the change, and the initial average value following the change.

The table also includes a classification for each change, denoted by its level, indicating its significance. Changes can be categorised into multiple levels, ranging from level-1 to level-5 (Monyeki 2021). In this proposed methodology, the detection process iterates through multiple passes of the data, to identify changes at different levels. The initial iteration detects a level-1 change, the most prominent and easily discernible change-point in the CUSUM chart. Subsequent passes through the data can detect changes at level-2, and so on. Consequently, higher-level changes, such as level-2 to level-5, signify a greater likelihood of a change occurring (Arif *et al.* 2017). Depending on the number of changes found and the number of data iterations used to find them, the specific level for a particular change-point may differ. *Table 5.1* presents the bootstrap results of the electricity net generation in SA.

Table 5.1: Bootstrap results showing significance of changes in electricity net generation in SA

Confidence Level for Candidate Changes = 50%, Confidence Level for Inclusion in Table = 90%, Confidence Interval = 95%, Bootstraps = 1000, Without Replacement, MSE Estimates

Year	Confidence Interval	Conf. Level	From	To	Level
1990	(1986, 1994)	92%	85937	155450	2
2000	(1992, 2000)	98%	155450	203140	2
2004	(2004, 2004)	92%	203140	235730	4

The CUSUM chart analysis of electricity net generation from 1980 to 2020 revealed several significant change-points, each accompanied by confidence levels and classification levels. The first change-point occurred in 1990 with a confidence level of 92%, indicating high certainty that the change transpired. Prior to the change, electricity production was recorded at 85.937 billion kWh, increasing to 155.450 billion kWh immediately following the change. This change was classified as a level-2 change, signifying a significant change from the previous trend, but not the most prominent change-point detected during the initial iteration.

The second change-point, identified in 2000, had a confidence level of 98%, indicating even higher certainty. The electricity produced before the change was 155.450 billion kWh, rising to 203.140 billion kWh after the change. This change-point was also a level-2 change, indicating another significant deviation, but not the most prominent change-point detected in the initial iteration. The change-point that occurred in 2004, had a confidence level of 92%, indicating substantial confidence in the occurrence of the change. Prior to the change, electricity production stood at 203.140 billion kWh, increasing to 235.730 billion kWh afterwards. This change-point was; however, a level-4 change, indicating a higher level of

significance compared to level-2 changes. These quantitative values are essential in understanding the magnitude of the changes in electricity production at each change-point. They provide insight into the scale of the shifts in production and help assess the significance of the changes. The confidence levels further enhance the reliability of the findings, indicating the degree of certainty in the detected change-points.

Figure 5.3 illustrates the results of the CPA conducted on electricity production from all primary and secondary coal sources in SA from 1980 to 2015.

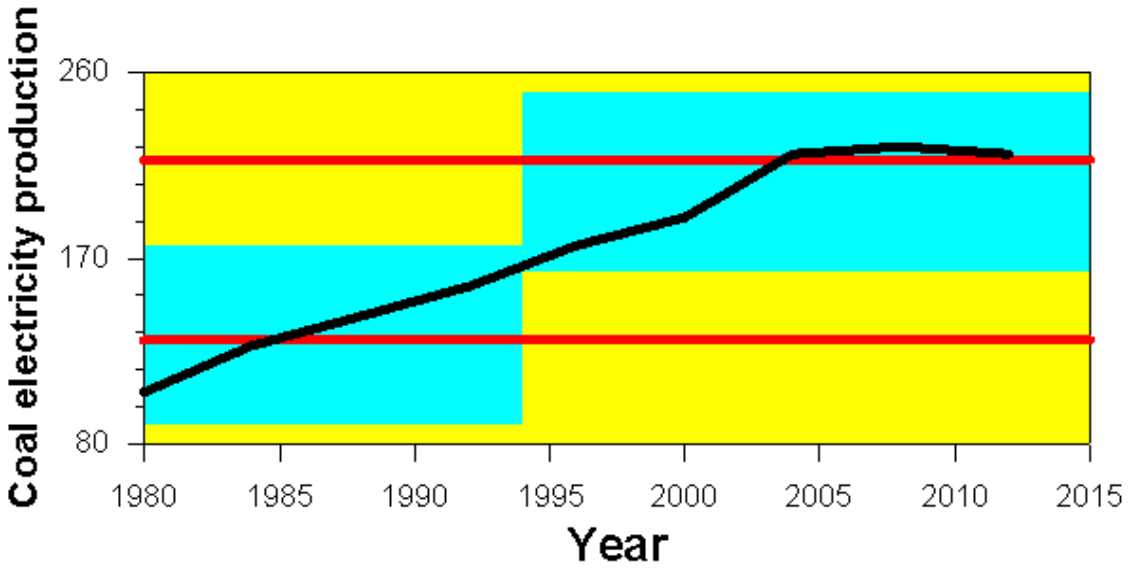


Figure 5.3: CPA of coal electricity production in SA from 1980 to 2015

The first change-point occurred in 1996, indicating a significant increase in coal electricity production. Prior to this change-point, there was a rapid rise in production from 1980, culminating in a peak at the onset of 1996. The surge in coal electricity production in 1996 was driven by a combination of factors. The post-apartheid era brought a heightened need for equal energy distribution, prompting an increase in supply to meet the growing demands and foster economic development (Mngomezulu 2016). South Africa's heavy reliance on coal, driven by its cost-effectiveness when compared to other energy sources like oil and gas, further fuelled this growth (Dikgwatlhe 2018). Additionally, the country's rich coal reserves and established coal industry infrastructure, solidified coal's role as a primary energy source, driving the upsurge in coal electricity production, during this period (Burton, Caetano and McCall 2018; Dikgwatlhe 2018). Following this change-point, production continued to peak until the end of the study period in 2015, suggesting a sustained period of high coal electricity output.

After 2004, the electricity produced exhibited a continuous trend until 2008, during which time the production remained relatively stable. Notably, the data prior to 2008 remained consistently above the upper limit of the CPA graph, indicating a significant deviation from the expected trend. Following the slight drop in production in 2008, the graph shifted closer to the upper limit, suggesting a convergence towards the expected level of variability.

The CUSUM chart depicting electricity production from all primary and secondary coal sources in SA from 1980 to 2015 is presented in *Figure 5.4* below.

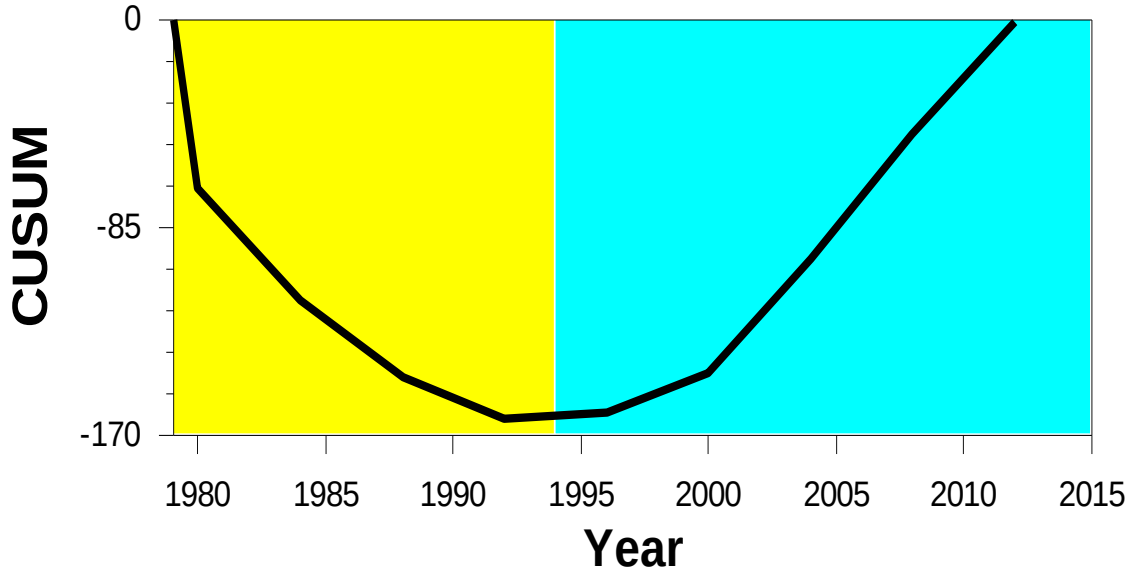


Figure 5.4: CUSUM plot of coal electricity production in SA from 1980 to 2015

The first change-point occurred in 1996, as evidenced by the blue region of the chart. Following this change-point, the graph continued to show an upward trend, indicating a sustained increase in coal electricity production until the end of the study period. This is noted by the ongoing upward trend following the change-point, which signifies that the data is above the overall average during this time period. This upward trend depicts a significant upsurge in coal-fired electricity production, highlighting the crucial role of coal in meeting the country's energy needs. Moreover, it is noteworthy that this CUSUM chart serves to validate and reconfirm the change-points identified in the previous CPA graph. The shifts observed in the CPA graph should be consistent with those in the CUSUM chart. In this case, the validation of the identified change-points using these two methods enhances the credibility and reliability of the findings (Ma, Grant and Sofronov 2020). Consistency between the two analytical approaches strengthens the confidence in the identified change-points and provides a more robust understanding of the trends in coal electricity production in SA.

Table 5.2 below presents the bootstrap results for the electricity produced from coal sources in SA.

Table 5.2: Bootstrap results showing significance of changes in coal electricity production in SA

Confidence Level for Candidate Changes = 50%, Confidence Level for Inclusion in Table = 90%, Confidence Interval = 95%, Bootstraps = 1000, Without Replacement, MSE Estimates

Year	Confidence Interval	Conf. Level	From	To	Level
1996	(1996, 1996)	93%	133.43	206.72	1

The CUSUM chart analysis conducted on electricity production from coal sources in the period of 1980 to 2015 revealed a significant change-point in 1996, with a confidence level of 93%. This high confidence level indicates a strong sense of certainty about the occurrence of the change. Prior to the change, electricity production was measured at 133.43 billion kWh, increasing to 206.72 billion kWh immediately following the change. This change-point was classified as a level-1 change. In the context of bootstrap analysis, a level-1 change represents the most prominent change-point detected in the initial iteration of the analysis (Arif *et al.* 2017). It signifies a significant deviation from the previous trend and is easily discernible.

The CPA results of GDP growth in SA from 1980 to 2022 are depicted in *Figure 5.5* below. Given the nature of GDP growth, which is not a linear trajectory, this graph exhibits multiple peaks and drops, reflecting the dynamic nature of economic trends over time (Ibrahim and Alagidede 2018). Following a steady increase from 1992 to 1997, the first change-point in 2000 marked a slight increase in GDP growth. Prior to this change-point, there was a period of robust economic expansion, which was characterised by sustained growth rates. However, a slight drop in GDP growth occurred in 1998, just before the identified change-point, indicating a shift in the economic landscape. Subsequent to this change-point, GDP growth exhibited a modest upward trajectory until around 2008, demonstrating resilience and stability in the economy during that time period.

The period of steady growth in the mid-1990s was largely influenced by South Africa's transition to democracy in 1994, which brought in a surge of economic reforms and increased foreign investment (Aron and Kingdon 2007; Faulkner and Loewald 2008). The government implemented policies aimed at liberalising the economy, attracting foreign investment, and promoting economic growth (Hirsch, Levy and Nxele 2021).

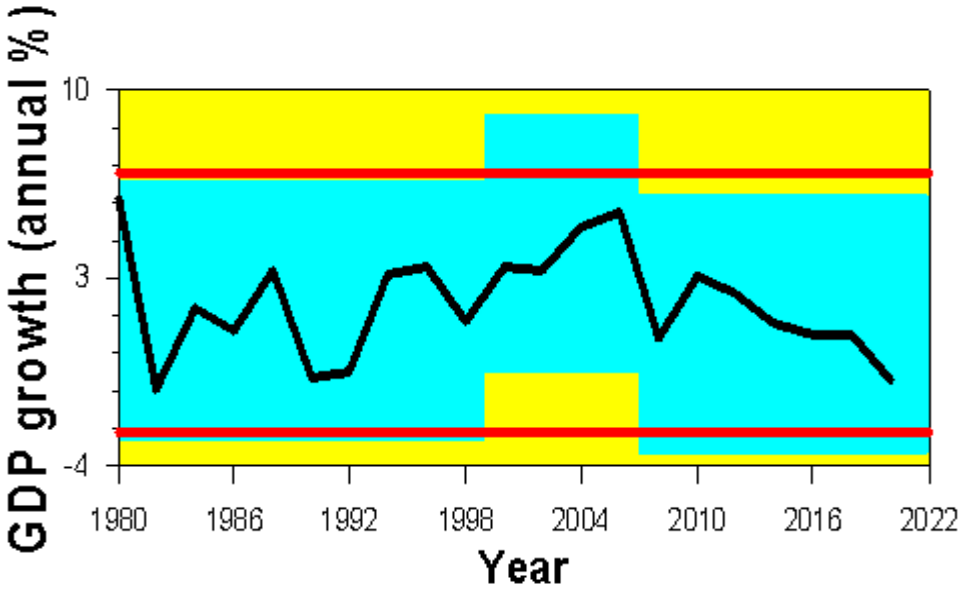


Figure 5.5: CPA of GDP growth in SA from 1980 to 2022

However, the slight drop in GDP growth in 1998 may be attributed to external factors, such as the Asian financial crisis, which impacted global markets and affected South Africa's export-oriented industries (Siriwardana and Dollery 2002). Despite this, the economy rebounded, leading to modest growth until around 2008.

The next significant change-point occurred in 2008, coinciding with a sudden drop in GDP growth in the country. This drop persisted for the remainder of the time period under analysis. The sharp decline in GDP growth following the 2008 change-point suggests a significant disruption or decline in the economy. This significant change-point can be linked to the global financial crisis that began around 2007 to 2008. SA, like many other countries, experienced the repercussions of this crisis, including declining demand for exports, reduced investment inflows, and tightening credit conditions (Rena and Msoni 2014). The country's mining sector, a key contributor to GDP, was particularly affected by the falling commodity prices and decreased demand (Mafukata 2017).

Additionally, domestic factors such as electricity shortages, exacerbated by frequent load-shedding, had a detrimental impact on economic activity in the country (Ateba, Prinsloo and Gawlik 2019). Load-shedding, which became increasingly common from around 2008 onwards, disrupted business operations, reduced economic productivity and, dampened investor confidence (Walsh, Theron and Reeders 2021; Hlongwane, Mahapa and Nthebe 2023). Additionally, SA experienced severe drought conditions during this period, particularly

in the agricultural sector, which further hampered economic growth (Archer *et al.* 2019). The combined effect of these several factors contributed to the prolonged period of subdued GDP growth observed, after the 2008 change-point, reflecting the challenges faced by the economy during this time. *Figure 5.6* illustrates the CUSUM chart of GDP growth in SA from 1980 to 2022.

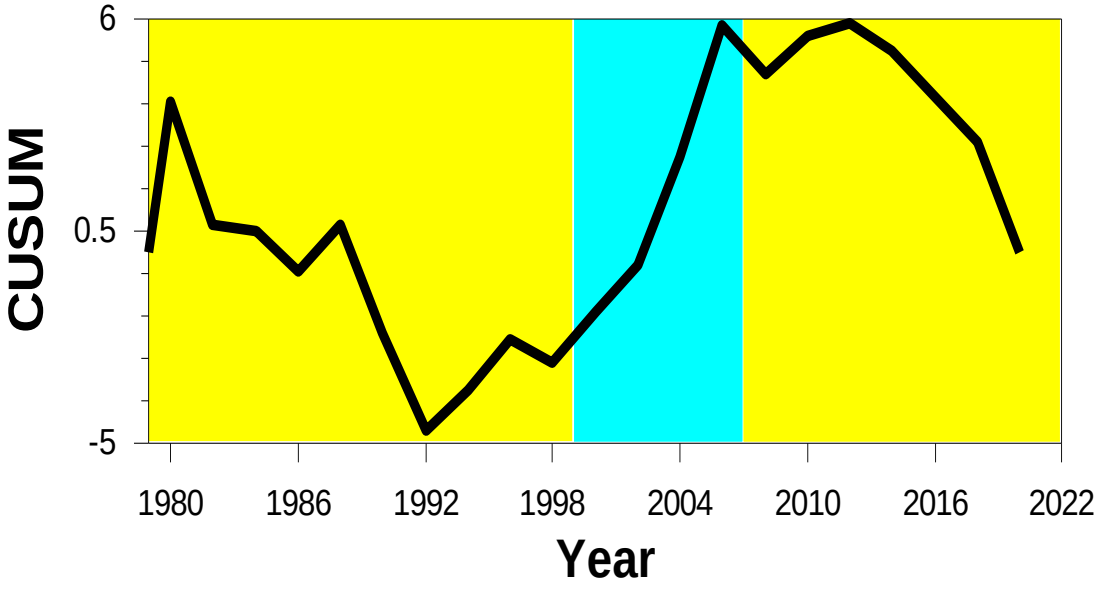


Figure 5.6: CUSUM plot of GDP growth in SA from 1980 to 2022

The first change-point, identified in 1996, is indicated by a change in the blue region of the graph. Following this change-point, the CUSUM chart shows a continuous upward trend, indicating a sustained increase in GDP growth. When the CUSUM graph is on an upward trend, it signifies that the data points are consistently above the overall average, suggesting a period of economic expansion and positive GDP growth. The second significant change-point occurred in 2008, as depicted by a change in the background colour of the graph. Following this change-point, the CUSUM chart shows a continuous downward trend, indicating a sharp decline in GDP growth. When the CUSUM graph is on a downward trend, it suggests that the data points are consistently below the overall average, reflecting a period of economic decline. Furthermore, the CUSUM chart serves to validate and reconfirm the change-points identified in the previous CPA graph. The alignment of change-points between the two methods strengthens the credibility and validity of the results. This consistency enhances confidence in the identified inflection points and underlying trends in GDP growth, providing valuable insights for policymakers, economists, and analysts.

The bootstrap analysis of GDP growth in SA from 1980 to 2022 revealed significant change-points, each accompanied by its respective confidence and classification level. This analysis is presented in *Table 5.3* below.

Table 5.3: Bootstrap results showing significance of changes in GDP growth in SA

Confidence Level for Candidate Changes = 50%, Confidence Level for Inclusion in Table = 90%, Confidence Interval = 95%, Bootstraps = 1000, Without Replacement, MSE Estimates

Year	Confidence Interval	Conf. Level	From	To	Level
2000	(1984, 2000)	92%	1.8107	4.2932	4
2008	(2008, 2012)	92%	4.2932	1.2641	5

The first change-point occurred in 2000, with a confidence level of 92%. Prior to this change, GDP growth stood at 1.81%, but increased to 4.29% following the change. This change was classified as a level-4 change. In bootstrap analysis, a level-4 change-point typically denotes a substantial shift in the data, indicating a significant change from the previous trend. Similarly, the change-point that occurred in 2008, also had a confidence level of 92%. Before this change, GDP growth was 4.29%, but dropped sharply to 1.26% after the change. This change was classified as a level-5 change. In bootstrap analysis, a level-5 change indicates a more significant shift from the previous pattern, when compared to a level-4 change. It indicates a substantial and abrupt shift in the data, often reflecting a significant economic event or structural change.

The CPA results of CO2E in SA from 1995 to 2020 reveal several notable change-points, as demonstrated in *Figure 5.7* below. This analysis suggests fluctuations in CO2E over the analysed period, with significant changes observed at various points in time. The first change-point occurred in 2001, signalling an upward shift in CO2E after a period of steady growth throughout the 1990s. This change-point can be linked to various factors in SA, such as transitioning from the apartheid regime (Smith 2023). The effects of apartheid left the country with vast social and economic disparities, justifying the need for rapid industrialisation and urban development to address these challenges (Turok 2012; Smith 2023). These initiatives lead to increased energy consumption and, consequently, higher CO2E. Additionally, the growing demand for electricity, primarily generated from coal-fired power plants, contributed to rising emissions during this period (Winkler and Marquand 2009; Pretorius, Piketh and Burger 2015).

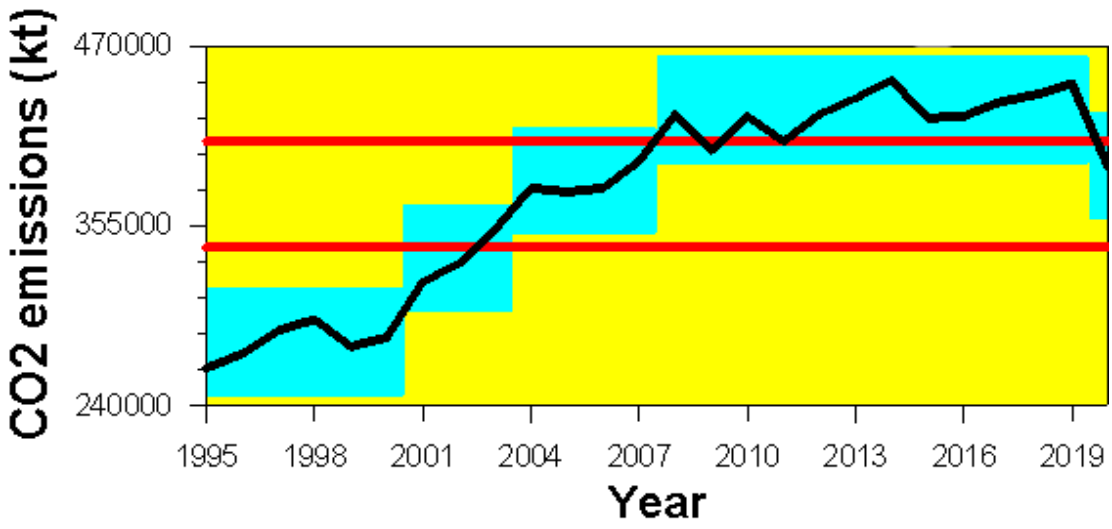


Figure 5.7: CPA of CO2E in SA from 1995 to 2020

As the economy expanded and energy-intensive industries such as mining and manufacturing flourished, the demand for fossil fuels, particularly coal, increased, driving up CO2E (Spalding-Fecher, Williams and van Horen 2000; Hanto *et al.* 2022).

Subsequently, emissions continued to increase until the next change-point in 2004. Following this second change-point, CO2E remained on a steady upward trajectory. This change-point correlates with the ongoing industrial and economic growth in SA. During the mid-2000s, the country's mining sector experienced significant expansion, driven by high global demand for minerals and commodities (Miller, Saunders and Oloyede 2008; Sorensen 2011). This growth, coupled with increased energy consumption in this sector, further fuelled CO2E during this time period.

Another significant change occurred in 2008, marked by a rapid peak in emissions. This change-point can be attributed to several factors, including the global economic downturn and domestic challenges. The global financial crisis that began in 2007 to 2008, led to a slowdown in economic activity worldwide, impacting South Africa's export-oriented industries and reducing demand for its commodities (Maisonnave *et al.* 2009; Rena and Msoni 2014). Consequently, the decline in export revenues and investment inflows put pressure on the economy, prompting the government to implement proactive measures to support growth (Madubeko 2010). As a result, there was a temporary resurgence in industrial activity and energy consumption, leading to the peak in CO2E following the economic crisis (Sadorsky 2020).

This peak was followed by a slight drop in CO2E, although it remained consistently above the upper limit throughout the 2010s. Furthermore, throughout the 2010s period, CO2E exhibited periodic fluctuations with slight drops and increases. This subsequent period of sustained high emissions can be attributed to ongoing economic development and infrastructure investment in SA. Despite efforts to transition to cleaner energy sources and improve energy efficiency, the country's heavy reliance on coal for electricity generation and industrial processes continued to drive emissions (Henneman *et al.* 2016; Mirzania *et al.* 2023). Additionally, the ongoing load-shedding issue experienced in SA during the 2010s (Steenkamp 2016), exacerbated by aging infrastructure and capacity constraints, further contributed to CO2E. This was because businesses and households turned to backup generators, which are typically powered by diesel or petrol, during power outages. It's worth noting that whilst load-shedding temporarily reduces electricity consumption and emissions during power outages, the reliance on backup generators can lead to increased emissions from fossil fuel combustion, particularly diesel, offsetting some of the environmental benefits (Jakhrani *et al.* 2012; Borofsky 2021). This dynamic interplay between load-shedding, electricity consumption patterns and carbon emissions, contributes to the multiple occurrences of peaks and drops in CO2E observed, following the introduction of load-shedding in the country.

Notably, after 2019, there was a noticeable decline in CO2E, resulting in a drop below the upper limit. This decline may be attributed to a variety of factors, including policy interventions aimed at reducing emissions and technological advancements in renewable energy and energy efficiency. SA has committed to reducing its GHG emissions under international agreements, such as the Paris Agreement, and has implemented measures to transition towards a more sustainable and low-carbon economy (Burton, Marquard and McCall 2019; Sampene *et al.* 2021). Additionally, the increasing adoption of renewable energy sources, such as wind and solar power, coupled with improvements in energy efficiency practices may have contributed to the decline in emissions observed in recent years (Sampene *et al.* 2021). This shift to a climate-resilient and green economy is consistent with the nation's resolve to meet its CO2E objectives under the 2030 National Development Plan (Qu *et al.* 2023).

However, it's essential to consider the impact of load-shedding on carbon emissions, particularly as load-shedding occurrences increased rapidly from 2019 onwards (Naidoo 2023). Whilst load-shedding can temporarily reduce electricity consumption and emissions during power outages (Sguazzin 2023), it also prompted a shift in behaviour towards more energy-

conscious practices and a greater emphasis on energy conservation. This inadvertently contributed to a decrease in CO2E in SA, showcasing the potential for behavioural changes to positively impact environmental sustainability efforts. However, it's important to note that load-shedding, although beneficial in reducing emissions, is not an optimal or sustainable method for achieving long-term emissions reductions, as it has a severe effect on economic growth (Goldberg 2015). *Figure 5.8* presents the CUSUM chart of CO2E in SA from 1995 to 2020. This chart reveals significant change-points, as marked by the shifts in the background colour.

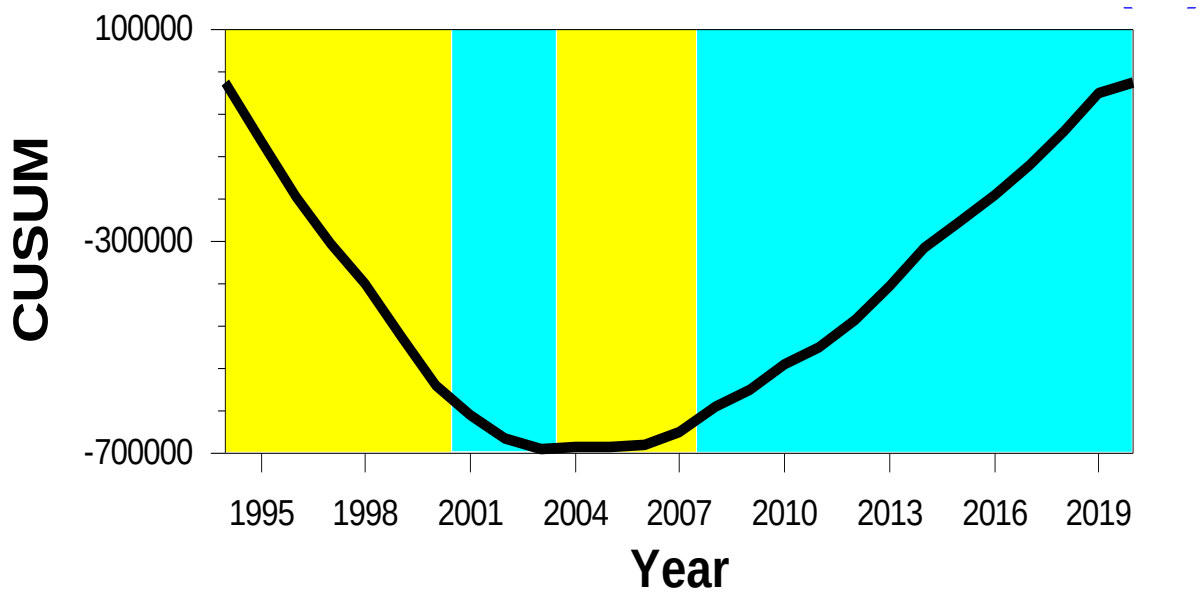


Figure 5.8: CUSUM plot of CO2E in SA from 1995 to 2020

The first change-point occurred in 2001, as demonstrated by the change in the background colour from yellow to blue. This change signifies a shift in the trend of CO2E data, possibly indicating a change in the underlying factors affecting emissions. The second change-point occurred in 2004, as indicated by another change in background colour. Moreover, the change-point in 2008, marked by yet another shift in background colour, demonstrates an upward trend in emissions until the end of the study period. This trend indicates that emissions fell above average during this period. Additionally, the alignment of change-points between the CUSUM chart and the previous CPA graph, validates and reaffirms the identified change-points. This validation ensures that the conclusions drawn from the analysis are robust and reliable, providing a more accurate understanding of the trends and patterns in CO2E over time.

Table 5.4 below depicts the bootstrap analysis of CO2E in SA from 1995 to 2020. This analysis revealed notable change-points, each paired with its corresponding confidence level and classification level.

Table 5.4: Bootstrap results showing significance of changes in CO2E in SA

Confidence Level for Candidate Changes = 50%, Confidence Level for Inclusion in Table = 90%, Confidence Interval = 95%, Bootstraps = 1000, Without Replacement, MSE Estimates

Year	Confidence Interval	Conf. Level	From	To	Level
2001	(2001, 2001)	90%	280970	334900	2
2004	(2004, 2004)	95%	334900	383620	1
2008	(2008, 2010)	100%	383620	426500	2

The first change-point occurred in 2001 with a confidence level of 90%, indicating a relatively high level of certainty in the observed change. Prior to this change-point, carbon emissions were recorded at 280,970 kt, increasing to 334,900 kt immediately following the change. This change was classified as a level-2 change, indicating a significant change in the observed trend. The second change-point, which occurred in 2004, was detected with a confidence level of 95%, indicating a higher level of certainty. Prior to this change, the CO2E was 334,900 kt, rising to 383,620 kt after the change. This change-point was classified as a level-1 change, representing the most prominent change-point detected in the initial iteration of the analysis. The change-point that occurred in 2008, had a confidence level of 100%, indicating the highest level of certainty in the observed change. Prior to this change, CO2E was 383,620 kt, increasing to 426,500 kt afterwards. This change-point was classified as a level-2 change, similar to the first change-point, signifying a noteworthy shift in the observed trend.

Table 5.5 provides a comprehensive summary of the change-points identified through the CPA model, across the various key variables in South Africa's socio-economic landscape. Additionally, this table demonstrates the current trajectory of each variable subsequent to the identified change-points.

Table 5.5: Summary of potential shifts for each socio-economic indicator or variable

Variable	**Change-points**	**Latest shift**
Electricity net generation	1990, 2000, 2004	Constant trajectory
Electricity production from coal sources	1996	Constant trajectory
GDP growth	2000, 2008	Downward
CO2 emissions	2001, 2004, 2008	Downward

5.2.1 Discussion on CPA results

The CPA reveals several key trends in SA's electricity production and GDP growth. One significant observation is the close relationship between economic indicators and energy consumption patterns. During phases of economic expansion and industrial development, there is a noticeable surge in electricity production, as the energy sector ramps up production to meet growing demands. The growing demand for electricity underscores the importance of energy production in supporting economic activities and meeting the needs of a growing population, highlighting the interdependence between energy supply and economic development. This observation is further evident in the CPA graphs. For instance, the increase in economic growth from 1992 to 1997, and the change-point that occurred in 2000, correspond to increases in electricity net generation. Consequently, this suggests a possible correlation between economic growth and electricity demand. As the economy expands, driven by factors like increased investment, industrial activity, and population growth, there is a corresponding rise in the demand for electricity to power industries, homes, and infrastructure (Rahman 2021).

However, the surge in electricity production often corresponds with an increase in CO2E, given the reliance on fossil fuels, particularly coal, for electricity generation. Conversely, economic downturns or external shocks often lead to declines in electricity production and emissions, as seen in the aftermath of the global financial crisis. Moreover, the persistent challenge of load-shedding, exacerbated by aging infrastructure and capacity constraints, has impeded significant growth in electricity production and contributed to fluctuations in GDP growth. This trend has been evident in recent years, with electricity production showing minimal growth whilst GDP growth declines. However, the shift in non-renewable energy production can also be attributed to South Africa's transition to adopt cleaner energy sources. Simultaneously, the country has faced various economic challenges, including high unemployment rates, global economic downturns, fluctuations in commodity prices, severe droughts, and load-shedding, all of which have impacted economic output (Mandeya and Ho 2021; Naidoo 2021; Lawlor 2023). The cumulative effect of these economic challenges has posed significant obstacles, counteracting the positive strides made in transitioning to sustainable energy sources.

Load-shedding, which is one of the economic challenges SA faces, adds another layer of complexity to these trends. The dependence on coal-fired power generation, coupled with electricity shortages, underlines the vulnerability of South Africa's energy sector to systematic disruptions, highlighting the need for investment in diversified energy sources and improved

infrastructure resilience. Furthermore, load-shedding disrupts electricity supply and consumption patterns, leading to fluctuations in both electricity production and GDP growth (Akpeji *et al.* 2020). This dynamic re-emphasises the need for robust infrastructure upgrades and renewable energy transitions, to ensure long-term sustainability and economic resilience. Additionally, the interplay between energy consumption and GDP growth underscores the importance of addressing energy security and efficiency, to foster sustainable economic development, whilst minimising any environmental impacts.

The CPA findings also shed light on the complex dynamics driving CO2E in SA. These findings reveal that electricity production and GDP growth are the primary drivers of CO2E in SA. In respect to the linkage between GDP growth and carbon emissions, the transition from apartheid and the subsequent economic reforms throughout the years, played a pivotal role in shaping emission trends, with rapid industrialisation and urban development driving up energy consumption. Consequently, the increased energy production from coal sources has resulted in elevated carbon emissions within the country. This pattern is also noticeable in the CPA results, as periods of rapid GDP growth are also associated with increases in CO2E, which is particularly perceptible around the change-points in 2000 and 2004. This indicates a link between economic development, energy production, and CO2E. As industries expand and energy-intensive sectors like mining and manufacturing grow, there is a rise in fossil fuel consumption, primarily coal, leading to higher CO2E levels (Musa, Maijama'a and Yakubu 2021). This underscores the challenge of balancing economic growth with environmental sustainability.

In addition, the reliance on coal-fired power plants, particularly in energy-intensive sectors such as mining and manufacturing, further exacerbated the growth in CO2E. Additionally, external factors such as the global financial crisis and domestic challenges like load-shedding, have influenced carbon emission patterns. These factors are associated with reduced economic activity, resulting in a decrease in energy demand and, consequently, a reduction in CO2E. This highlights the sensitivity of energy production to fluctuations in carbon emissions. Despite efforts to transition to cleaner energy sources and improve energy efficiency, the heavy dependence on coal and ongoing infrastructure challenges have sustained high emissions levels. However, recent declines in emissions suggest progress towards a more sustainable and low-carbon economy, driven by policy interventions and technological advancements in renewable energy and energy efficiency. Nonetheless, the impact of load-shedding on CO2E

underscores the need for a comprehensive approach to energy management and conservation efforts.

It is worth noting that the statistical models presented in Chapter Two of this study have unequivocally identified economic growth and electricity production as the principal drivers of CO2E in SA, as demonstrated by the findings of the systematic review conducted in Chapter Three (Lin *et al.* 2017; Shuai *et al.* 2017; Mensah *et al.* 2019; Nathaniel *et al.* 2021; Chen *et al.* 2022). Hence, the alignment between the findings from these statistical models, the systematic review and the results from the CPA suggests a robust consistency in identifying economic growth and electricity production, as the major contributors to CO2E in SA. This convergence underscores the significance of these factors in driving carbon emission trends in the country, reinforcing the importance of addressing both economic development and cleaner energy production strategies, to effectively mitigate carbon emissions and foster sustainable environmental practices. Consequently, SA has already made efforts to reduce GHG emissions and promote renewable energy adoption, as demonstrated by its commitment to the 2030 National Development Plan. This reflects the country's broader commitment to environmental stewardship and climate resilience. The increasing adoption of renewable energy sources and energy efficiency measures presents opportunities for economic diversification, job creation, and long-term economic growth, whilst mitigating environmental risks and promoting sustainable development.

5.3 ML model performance comparison

This section illustrates the ML results derived from the various trained models employed in this study. Based on the notion that ensemble approaches improve model performance of traditional ML techniques, the AdaBoost algorithm was chosen for this study, as justified in Chapter Four. Consequently, to empirically validate this hypothesis, the researcher employed four traditional ML algorithms alongside the AdaBoost regressor for benchmarking and determining the most effective technique for predicting CO2E. These include: Linear regression, Polynomial regression, Bayesian Linear regression and KNN regression. This comparative approach aimed to provide a comprehensive evaluation of model performance and confirm the effectiveness of ensemble methods in enhancing predictive accuracy. The results pertaining to the relevant performance metrics, which were also extensively discussed in Chapter Four of this study, were presented for examination. Through the systematic analysis of these performance metrics across the diverse models, a comparative analysis is facilitated,

offering insights into the efficacy and robustness of each algorithm, in the context of this study's objectives.

Table 5.6 below offers a comprehensive overview of the results obtained from the various ML models employed in this study. It includes values for each performance metric, such as the RMSE score, MAE score, R^2 score and accuracy. This allows for a thorough comparison of each model's effectiveness. Additionally, the effects of employing cross-validation and repeated holdout techniques on model performance are analysed. This comparative evaluation provides insights into the relative strengths and weaknesses of each technique, aiding in the selection of the most suitable approach under the conditions of this study.

Table 5.6: Cross-validation and Repeated holdout evaluation metrics

Algorithm	**RMSE score**		**MAE score**		**R^2 score**		**Accuracy**	
	Cross-validation technique	***Repeated holdout method***	***Cross-validation technique***	***Repeated holdout method***	***Cross-validation technique***	***Repeated holdout method***	***Cross-validation technique***	***Repeated holdout method***
AdaBoost regressor	10,215.45	10,070.89	9,129.07	10,156.2	0.942	0.85	96.61%	96.87%
Linear regression	15,492.6	15,254.39	12,602.0	12,075.55	0.924	0.77	96.52%	96.75%
Polynomial regression	17,754.47	16,064.77	15,384.95	13,214.53	0.90	0.68	95.5%	96.36%
Bayesian Linear regression	16,202.0	15,141.19	13,514.0	12,342.76	0.918	0.75	96.3%	96.71%
KNN regressor	22,029.60	14,472.64	21,149.0	11,722.99	0.847	0.83	94.16%	96.81%

The results of the ML models, as presented in the table above, provide insights into their performance in predicting CO2E. Utilising cross-validation techniques, the AdaBoost regressor emerges as the top-performing model, boasting the lowest RMSE score of 10,215.45 kt and a MAE score of 9,129.07 kt. These metrics quantify how closely the predictions made by the model align with the actual values. Hence, lower values for these metrics indicate that the predictions are closer to the actual values, signifying higher accuracy in forecasting CO2E. In addition, its high R^2 score of 0.942 indicates the model's ability to explain approximately 94.2% of the variance in the CO2E data. Moreover, achieving an accuracy of 96.61%, the AdaBoost regressor demonstrates robust predictive capability, making it an optimal choice for forecasting CO2E in SA, given the conditions of this study. In terms of employing the repeated holdout method, the AdaBoost regressor continues to outperform other models, possessing the

lowest RMSE score of 10,070.89 kt and a MAE score of 10,156.52 kt. It maintains a high R^2 score of 0.85 and an accuracy of 96.87%, reaffirming its reliability in CO2E prediction.

Table 5.7 provides a comprehensive overview of the average performance metrics across all ML algorithms utilised in the study. By averaging the performance metrics for each algorithm, it allows for a comparison of their effectiveness in generating models for CO2E forecasting. This analysis aids in identifying which ML algorithm consistently produces the best-performing models, based on the evaluation metrics.

Table 5.7: Average performance of each evaluation metric across all the ML algorithms

	RMSE mean	**MAE mean**	**R^2 mean**	**Accuracy mean**
AdaBoost regressor	10,143.17	9,642.64	0.90	96.74%
Linear regression	15,373.5	12,338.78	0.85	96.63%
Polynomial regression	16,909.62	14,299.74	0.79	95.93%
Bayesian Linear regression	15,671.6	12,928.38	0.83	96.5%
KNN regressor	18,251.12	16,436.0	0.84	95.49%

Based on the table above, the AdaBoost regressor stands out as the best-performing algorithm, with an average RMSE score of 10,143.17 kt, MAE score of 9,642.64 kt, R^2 score of 0.90 and accuracy of 96.74%. This indicates that the AdaBoost algorithm consistently produced models with the lowest errors and highest accuracy, when compared to the other models. Similarly, when considering Linear regression, Polynomial regression, Bayesian Linear regression, and KNN regression, AdaBoost surpasses them in terms of both RMSE and MAE, with the lowest values across these metrics. This demonstrates the effectiveness of the AdaBoost algorithm in training models that can predict CO2E under the constraints of this study. To further quantify the results, on average, the models trained with the AdaBoost algorithm improved the performance of the traditional ML models, by reducing the RMSE score by 6,417.29 kt, decreasing the MAE score by 4,358.09 kt, increasing the R^2 score by 0.07, and improving the accuracy by 0.60%.

After determining the optimal performing model, the subsequent assessment involves discerning the effects of cross-validation and repeated holdout on model performance, as previously mentioned. *Table 5.8*, which was derived from *Table 5.6*, provides an overview of the statistical properties of all the evaluation metrics. This comparison offers valuable insights

into the effectiveness of each technique in assessing model performance. By analysing these statistical properties, researchers can discern patterns and variations in performance metrics, facilitating a comprehensive evaluation of the techniques employed.

Utilising the data from *Table 5.8*, box and whisker plots were generated, as depicted in *Figures 5.9* to *5.11*. These visual representations showcase the statistical distribution of each evaluation metric, allowing for a direct comparison between the cross-validation and repeated holdout results. Through these plots, researchers can gain a deeper understanding of the variability and central tendency of each metric under different evaluation techniques. This comparative analysis helps determine the robustness and reliability of each technique for improving model performance. As a result, this analysis informs the selection of the most suitable technique between cross-validation and repeated holdout for future studies or applications under similar constraints.

Table 5.8: Statistical properties of evaluation metrics

		RMSE score	MAE score	R^2 score	Accuracy
Mean	Cross-validation	16,338.82	14,355.80	0.91	95.82%
	Repeated holdout	14,200.78	11,902.41	0.78	96.7%
First Quartile	Cross-validation	12,854.03	10,865.54	0.87	94.83%
	Repeated holdout	12,271.77	10,939.59	0.72	96.54%
Median	Cross-validation	16,202.0	13,514.0	0.92	96.3%
	Repeated holdout	15,141.19	12,075.55	0.77	96.75%
Third Quartile	Cross-validation	19,892.04	18,266.98	0.93	96.57%
	Repeated holdout	15,659.58	12,778.65	0.84	96.84%
Interquartile ranges (IQRs)	Cross-validation	7,038.01	7,401.44	0.06	1.74%
	Repeated holdout	3,387.82	1,839.05	0.13	0.3%
Standard Deviation	Cross-validation	3,811.02	3,957.52	0.03	0.92
	Repeated holdout	2,125.97	1,002.69	0.06	0.17

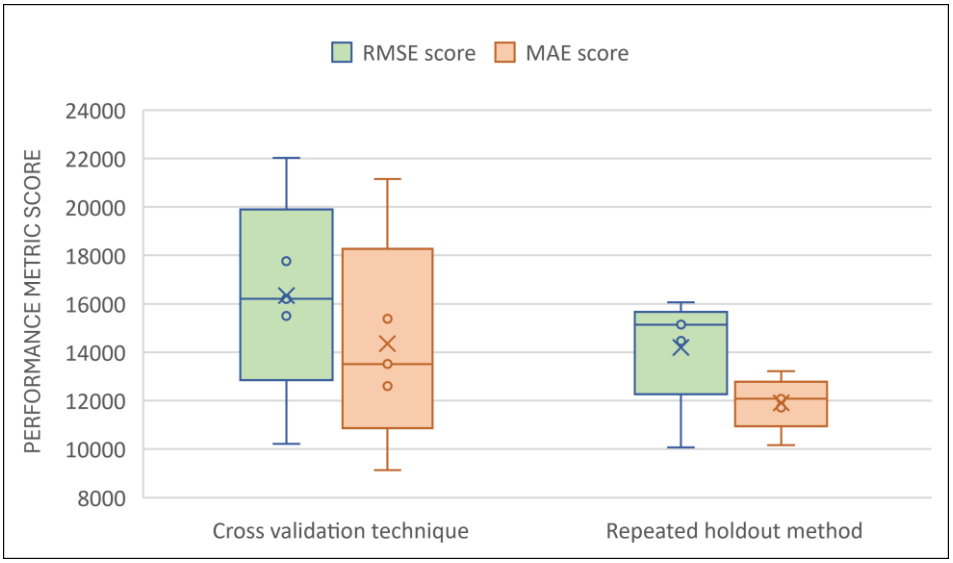


Figure 5.9: Statistical distribution of the RMSE and MAE metrics for both Cross-Validation and Repeated Holdout methods

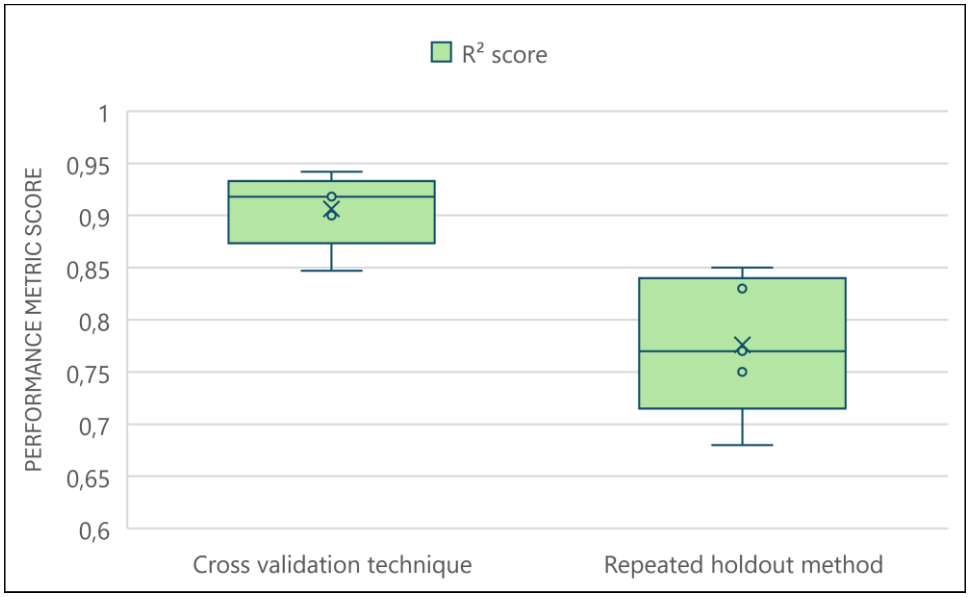


Figure 5.10: Statistical distribution of the R^2 metric for both Cross-Validation and Repeated Holdout methods

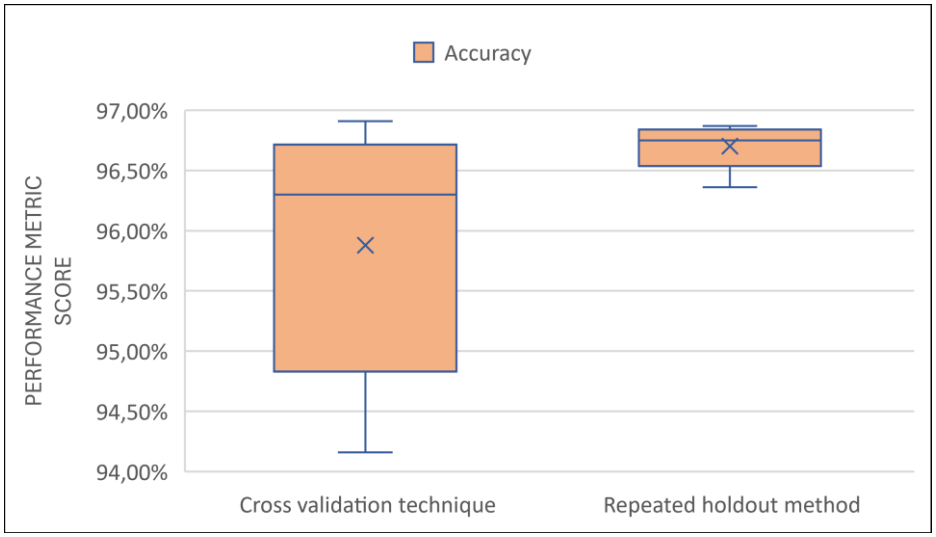


Figure 5.11: Statistical distribution of the accuracy metric for both Cross-Validation and Repeated Holdout methods

From the findings presented in *Table 5.8* and *Figures 5.9* through *5.11*, the repeated holdout method generally resulted in lower RMSE scores across all evaluation metrics, when compared to cross-validation. This indicates that it had a more significant effect on reducing prediction errors. Likewise, the repeated holdout method consistently yielded lower MAE scores across all evaluation metrics compared to cross-validation, suggesting that it had a more pronounced impact on reducing the magnitude of prediction errors. The only exception was that the first quartile value of the MAE metric in cross-validation, was marginally lower than that of repeated holdout. In terms of the accuracy score, the repeated holdout method generally resulted in slightly higher accuracy percentages across all evaluation metrics in comparison to cross-validation. This suggests that it had a slightly more substantial influence on improving overall model accuracy.

Collectively, the analysis reveals that the median and mean values of all the RMSE and MAE metrics were consistently lower for the repeated holdout results compared to cross-validation. In terms of the accuracy score, the statistical properties of the repeated holdout results were higher in comparison to cross-validation. This suggests that, on average, the performance of models trained using the repeated holdout method was superior. Additionally, the IQRs and standard deviation values for these evaluation metrics were consistently lower for the repeated holdout results, when compared to cross-validation. This indicates that the variability in the performance metrics amongst the models trained using repeated holdout was reduced, implying greater consistency and stability in model performance.

For instance, transitioning from the cross-validation technique to the repeated holdout method, resulted in the standard deviation of RMSE scores reducing from 3,811.02 kt to 2,125.97 kt. The standard deviation of the MAE scores decreased from 3,957.52 kt to 1,002.69 kt. Similarly, the standard deviation of the accuracy scores was reduced from 0.92 to 0.17. This reduction signifies a decrease in the spread or variability of the performance metrics, reinforcing the notion that the repeated holdout method yielded more reliable and consistent results compared to cross-validation.

Conversely, the analysis indicates that the values of all the statistical properties for the R^2 score were better for the cross-validation results compared to repeated holdout. This suggests that, on average, the models trained using cross-validation exhibit better overall fit to the data, when compared to those trained using repeated holdout. Whilst the R^2 score can provide insights into the proportion of variance explained by a regression model, they alone do not determine the model's reliability or appropriateness (Fernando 2023). Factors such as inherent variability in the data and the risk of overfitting can influence R^2 values (Frost n.d). Hence, it is essential to consider a range of evaluation metrics to comprehensively assess model performance. When comparing cross-validation and repeated holdout techniques, although the cross-validation results may depict higher R^2 values, the superior statistical properties exhibited by the other evaluation metrics in the repeated holdout results, highlight their greater effectiveness in enhancing model performance.

Following the analysis of the ML results thus far, the findings unveil two prospective techniques that could be employed to augment ML model performance in future studies, which face similar constraints as this research. These techniques include ensemble learning and repeated holdout methods. This aligns with the recommendations put forth by Li et al. (2018) and Hosseini et al. (2019), who advocated for exploring different methods to enhance ML techniques in CO2E prediction.

Now that the repeated holdout method has been proven superior to the cross-validation technique in enhancing model performance, the next phase entails quantifying the magnitude of this improvement. This comparison will be made in relation to the results obtained from the models trained using cross-validation. *Table 5.9* below illustrates the extent to which the repeated holdout method has improved the evaluation metrics, relative to the cross-validation technique.

This table initially presents the individual improvements in performance for each evaluation metric within each trained model. Subsequently, it outlines the average improvement for each metric, across all the models studied. The purpose of this analysis is to provide a comprehensive understanding of the efficacy of the repeated holdout method in enhancing model performance. Notably, the R^2 score has been excluded from this table, as the cross-validation technique outperformed repeated holdout for this specific evaluation metric.

Table 5.9: Effect of Repeated Holdout on evaluation metrics

Algorithm	RMSE Change	MAE Change	Accuracy Change
AdaBoost regressor	-144.56	-1,027.13	+0.26%
Linear regression	-238.21	-526.45	+0.23%
Polynomial regression	-1,689.70	-2,170.42	+0.86%
Bayesian Linear regression	-1,060.81	-1,171.24	+0.41%
KNN regressor	-7,556.96	-9,426.01	+2.65%
Average	**-2138.05**	**-3323.53**	**+0.88%**

For the RMSE and MAE scores, a change was denoted with the minus sign, as an improvement in these metrics requires lower values. Conversely, for accuracy, a change was indicated with a plus sign, as an improvement for this metric necessitates an increase in value. Amongst all the models studied, the KNN regressor exhibits the most significant improvement in performance metrics due to the repeated holdout method. Specifically, the performance of each evaluation metric under the cross-validation technique has seen the following improvements: the RMSE decreased by 7,556.96 kt, the MAE decreased by 9,426.01 kt, and the accuracy score increased by 2.65%.

In terms of the overall improvement across all models, the repeated holdout method consistently improved upon the cross-validation results. On average, the RMSE score has reduced by 2,138.05 kt, the MAE score has decreased by 3,323.53 kt, and the accuracy score has improved by 0.88%. It is worth noting that despite the repeated holdout method demonstrating overall better improvement compared to the cross-validation results, the MAE score of the AdaBoost model exhibited better performance, when utilising cross-validation. This is discernible in *Table 5.9* by the grey-shaded background.

Figures *5.12* and *5.13* present box and whisker plots illustrating the statistical distribution of the change induced by the repeated holdout method, on the evaluation metrics of the cross-validation technique. These figures were constructed based on the data provided in *Table 5.9*. The purpose of generating these graphical representations is to visually depict the extent and variability of the improvements achieved, by employing the repeated holdout method over cross-validation.

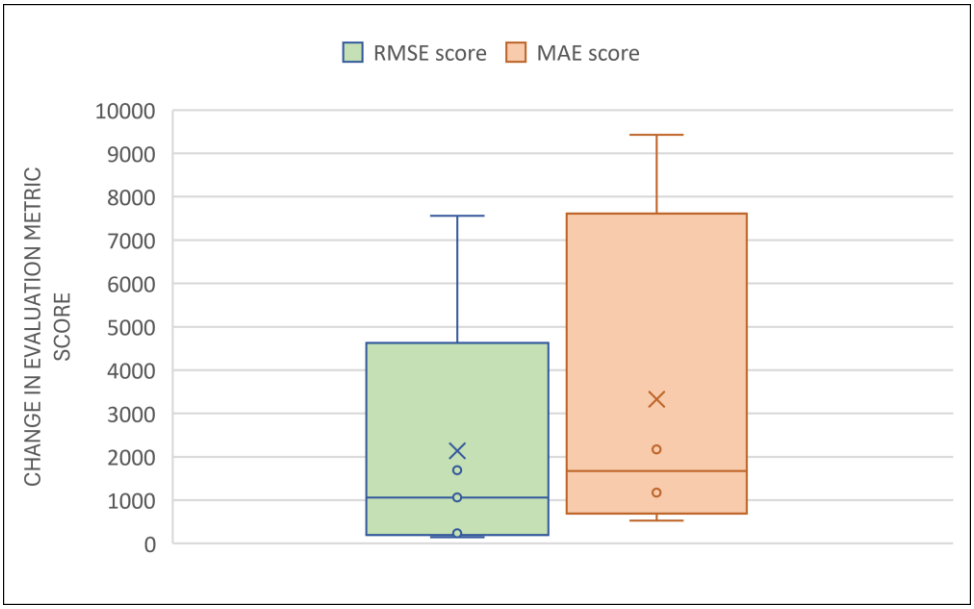


Figure 5.12: Statistical distribution of variations in scores for evaluation metrics

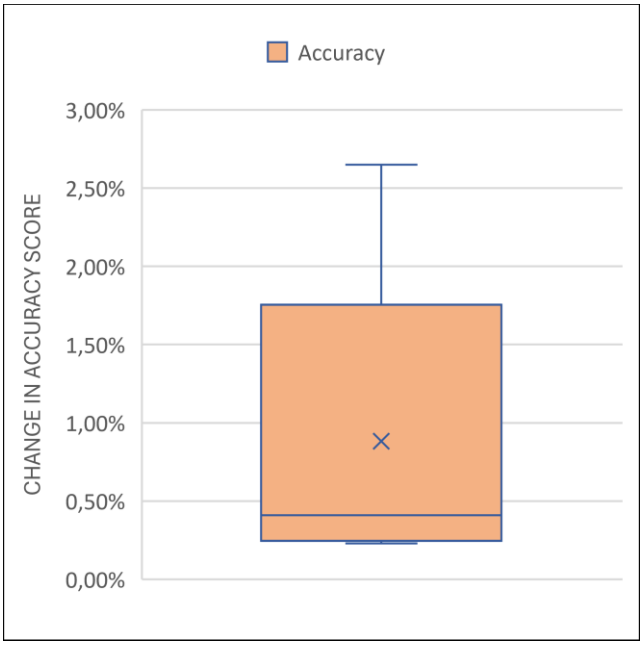


Figure 5.13: Statistical distribution of variations in accuracy scores

In *Figure 5.12*, the variability is notably high in the changes of both evaluation metrics. This suggests that quantifying the expected change in the RMSE and MAE metrics may not be as precise. However, this variability can be attributed to the significant changes observed in the RMSE and MAE scores of the KNN regressor. Such variability is indicative of the complex dynamics involved in modelling CO2E data. Nevertheless, upon excluding these outliers, the average improvement of the cross-validation results by the repeated holdout method was 783.32 kt for the RMSE score, and 1,289.39 kt for the MAE score, which remains notably substantial.

Furthermore, upon examining the distribution of changes depicted in *Figures 5.12* and *5.13*, it is evident that all the evaluation metrics exhibit a positive skew. This skewness may arise from the initially favourable performance attained by the models trained using cross-validation. With these metrics already demonstrating commendable results, there is limited scope for improvement without the risk of compromising the model's generalisation ability, through the model being overfit on the training dataset.

Prompted by the aforementioned observation, the subsequent analysis entailed examining the relationship between the cross-validation evaluation metric scores and the changes in these evaluation metric scores, following the application of the repeated holdout method. *Figures 5.14* and *5.15* below present scatterplots depicting this relationship. *Figure 5.14* showcases the relationship between the cross-validation RMSE and MAE scores and the respective changes in these evaluation metrics, whilst *Figure 5.15* illustrates the relationship between the cross-validation accuracy score and the change in this metric.

The trend lines illustrated in *Figure 5.14* indicate a robust positive correlation, as evidenced by all the R^2 scores exceeding 0.7. This suggests that as the RMSE and MAE scores increase, the enhancement of these metric scores also increases. These findings indicate that the repeated holdout techniques employed in this study significantly improve ML models that exhibit weaker performance for these evaluation metrics. However, the improvement is minimal for models already demonstrating strong performance.

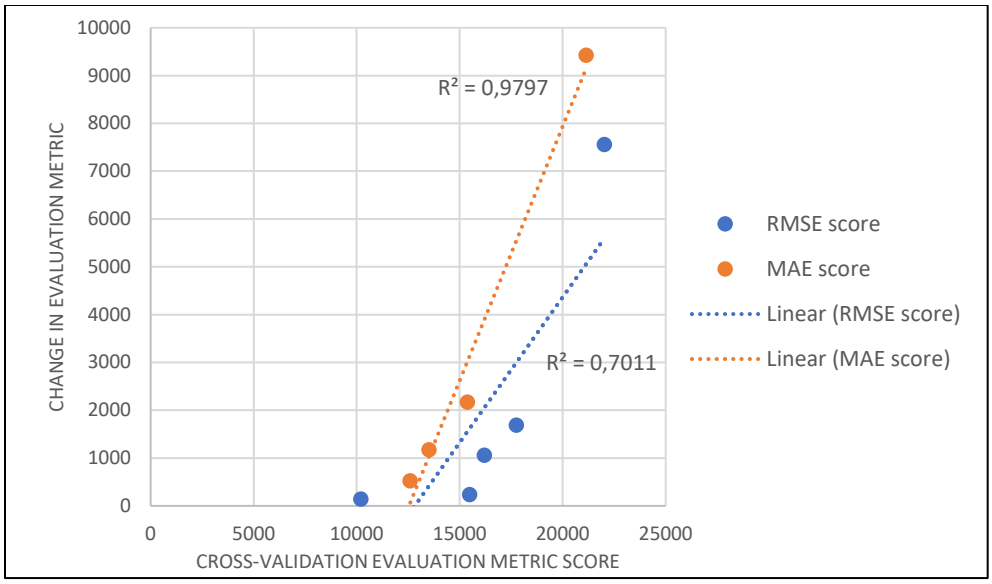


Figure 5.14: Relationship between the evaluation metric score obtained through cross-validation and the corresponding change in the evaluation metric score

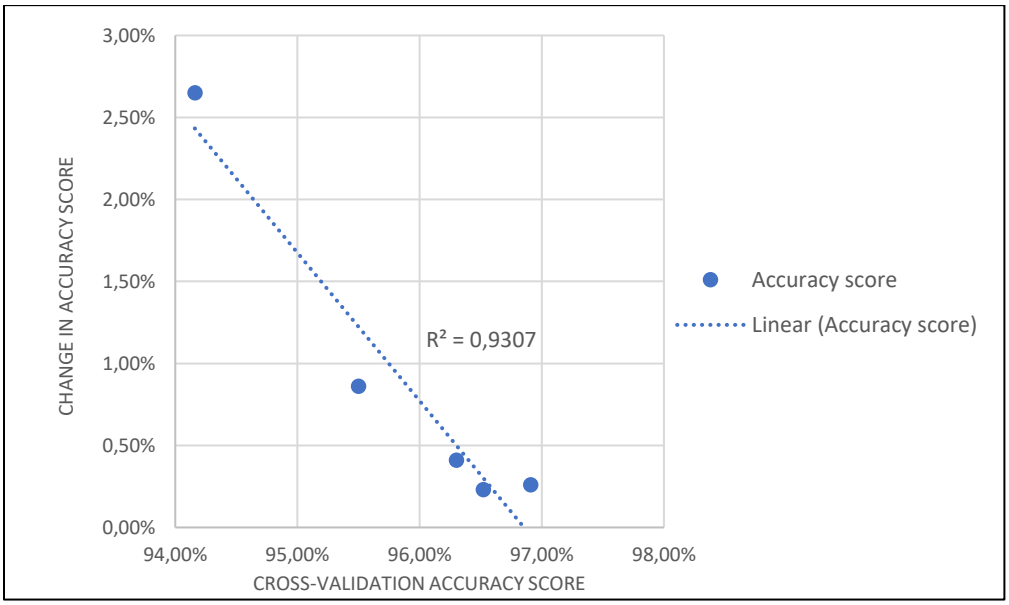


Figure 5.15: Relationship between the cross-validation accuracy score and the corresponding change in the accuracy score

The trend line presented in the figure above signifies a robust negative correlation, as also indicated by the R^2 score surpassing 0.7. This suggests that as the accuracy score increases, the enhancement of this metric decreases. These findings indicate that the repeated holdout techniques employed in this study marginally improve ML models that already demonstrate high accuracy scores.

Conversely, this technique is more effective in improving models that exhibit a weaker accuracy performance.

Upon examining *Figures 5.14* and *5.15*, it becomes evident that a significant improvement of 2,000 kt or greater in the RMSE and MAE scores is observed, when the cross-validation values of these metrics are equal to or exceed 15,000 kt. Similarly, for the accuracy score, an improvement of nearly 3% or more is observed, when the cross-validation value of this metric falls at or is below 94%. These findings provide valuable insights into the types of ML models for CO2E forecasting that stand to benefit the most from the repeated holdout techniques employed in this study. Furthermore, these insights serve as an implementation guideline that offers researchers a clear direction on when to implement repeated holdout techniques, to enhance ML models that initially adhere to other methodologies.

5.4 Validation of past trends and forecasting future CO2E trends

The AdaBoost regressor, which was integrated with the repeated holdout method, was selected as the primary model for forecasting CO2E in this study. This decision was supported by the fact that the AdaBoost regressor consistently produced the most optimal models, amongst the various ML algorithms evaluated. Additionally, the repeated holdout method was chosen for integration with the AdaBoost regressor due to its demonstrated effectiveness in enhancing model performance compared to cross-validation. By combining the strengths of the AdaBoost algorithm with the repeated holdout technique, the study aims to achieve more accurate and reliable forecasts of CO2E in SA.

Table 5.10 presents the predictions made by the AdaBoost model within the time period of 2021 to 2027. As previously discussed in Chapter Four, the input features, namely electricity production and GDP growth, were also forecasted for this time period, to serve as the features for this model. Furthermore, the choice of this time period was based on the observation from the CPA results, where a change-point tends to occur approximately every 6 years. Consequently, predictions were made for a 7-year period to encompass the possibility of encountering one or more change-points. This approach enables the assessment of the accuracy of the model's predictions before the occurrence of a change-point, providing valuable insights into its predictive capability over time. Additionally, this also helps to evaluate the model's performance in capturing any underlying

trends and deviations in future CO2E. It is worth noting that the economic growth rate in South Africa for 2023, stood at 0.6% (Statistics South Africa 2023). The AdaBoost regressor predicted GDP growth of 0.72% for the same year. This observation underscores the accuracy of the model in capturing real-world patterns, as evidenced by its ability to closely align with live data from the actual economy. Such alignment highlights the reliability and effectiveness of the model in making predictions based on observed patterns. *Figure 5.16* illustrates the forecasted electricity production in SA from 2021 to 2027.

Table 5.10: Predicted CO2E in SA from 2021 to 2027

Year	Electricity production (billion (kWh	GDP growth (annual %)	CO2 emissions (kt)
2021	229.403	4.81%	416,006.59
2022	200.712	1.9%	404,683.04
2023	193.396	0.72%	395,601.34
2024	227.884	2.21%	422,041.72
2025	226.041	1.91%	418,106.1
2026	207.657	2.9%	390,163.08
2027	214.678	2.7%	394,369.7

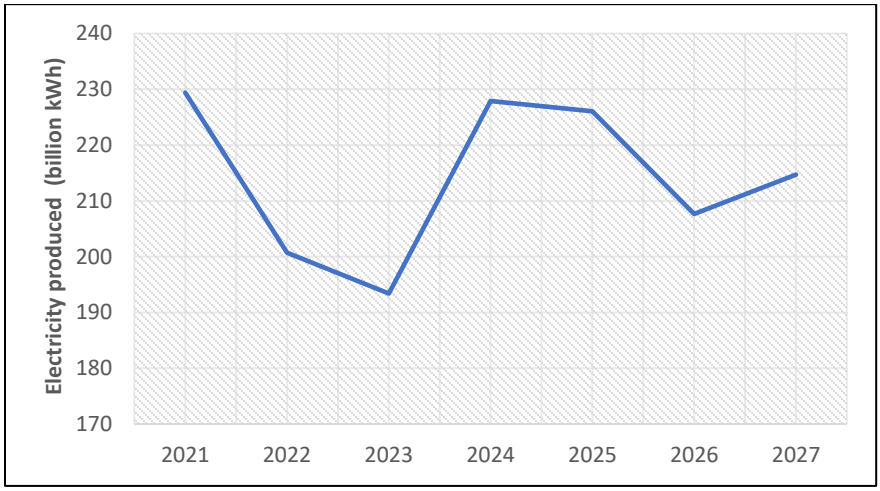


Figure 5.16: Electricity production in SA from 2021 to 2027

The CPA results revealed a consistent trajectory in electricity net generation over the latter part of the study period, reflecting steady production levels with minor fluctuations. The stability in electricity production aligns with the findings of the CPA model, which identified change- points in electricity production corresponding to economic shifts and policy interventions. This validation underscores the reliability and robustness of the CPA model in identifying trends and change-

points in electricity production. It confirms that the fluctuations and shifts observed in the forecasted data are consistent with the historical dynamics captured by the CPA.

Furthermore, the forecasted fluctuations in electricity production reaffirm the importance of considering historical trends and patterns when making future projections. By validating the CPA results, the forecasted information provides confidence in the accuracy of the model's predictions and enhances the understanding of the underlying factors driving electricity production trends over time. The forecasted GDP growth in SA within the period of 2021 to 2027 is depicted in *Figure 5.17* below.

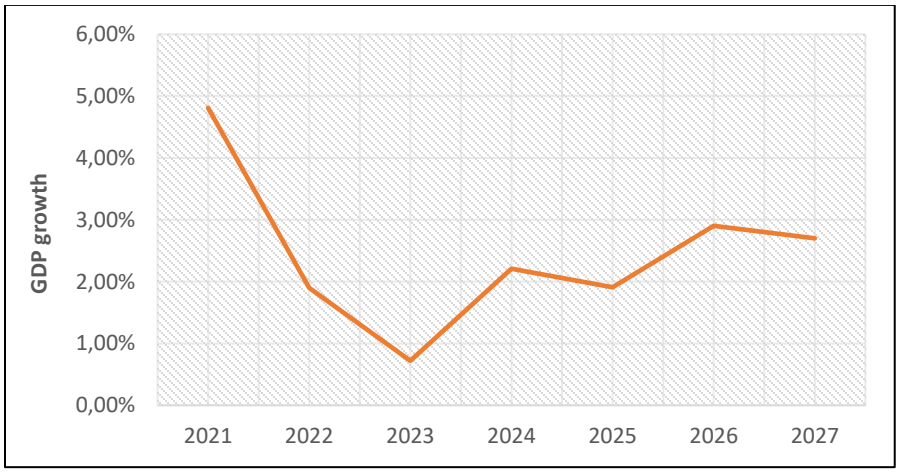


Figure 5.17: GDP growth in SA from 2021 to 2027

The forecasted downward trajectory in GDP growth is consistent with the findings of the CPA model, reflecting the ongoing economic challenges and fluctuations observed in SA. This alignment validates the CPA results, affirming the accuracy of the model in capturing historical trends and projecting future trajectories. Furthermore, this trend underscores the importance of addressing structural constraints in the energy sector and implementing policies to stimulate economic growth and resilience.

Moreover, the unexpected increase in GDP growth from 2023 to 2027, offers valuable insights into potential shifts or factors driving economic performance. This upward trajectory may signal improvements in economic conditions, such as increased investment, industrial activity, mitigation strategies to curb load-shedding, or government interventions to stimulate growth. However,

further analysis is warranted to understand the underlying drivers of this upsurge and its implications for future economic prospects.

Figure 5.18 presents the forecasted CO2E in SA during the period of 2021 to 2027. The forecasted downward trajectory in CO2E aligns with the trends identified in the CPA results, validating the model's ability to capture historical patterns accurately. This consistency suggests that the CPA model effectively identifies key inflection points and trends in CO2E, enhancing confidence in its predictive capabilities.

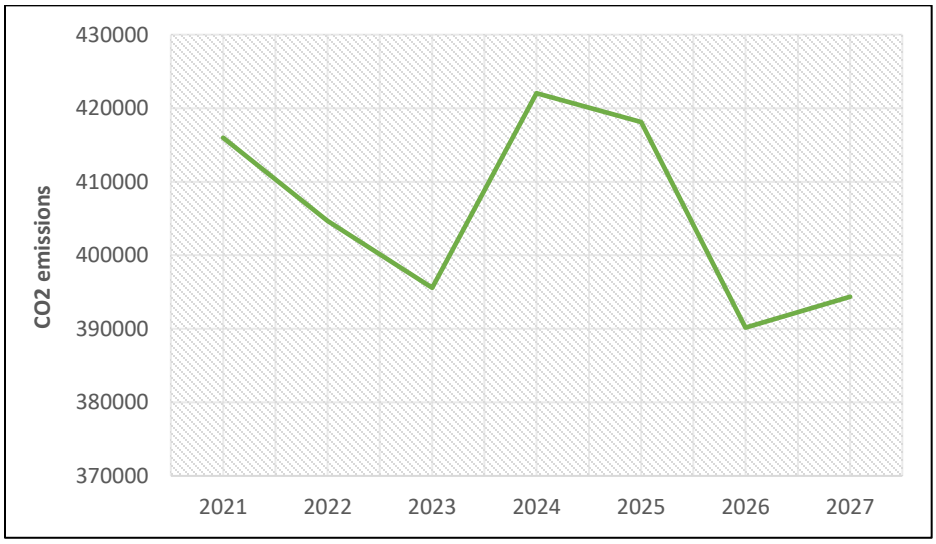


Figure 5.18: CO2E in SA from 2021 to 2027

The sustained decline in CO2E until 2023 mirrors historical trends, reflecting efforts to address environmental concerns and transition towards cleaner energy sources. This trend underscores the importance of implementing sustainable energy policies and reducing reliance on fossil fuels, to mitigate climate change and environmental degradation.

However, the unexpected peak in CO2E in 2024 followed by a subsequent drop until 2027, raises concerns about potential factors driving this fluctuation. Possible explanations could include changes in industrial activity, energy consumption patterns, or policy interventions affecting emissions levels. Further analysis is required to understand the underlying drivers of this peak and its implications for future emissions trajectories.

The forecasted trends in electricity net generation, GDP growth, and CO2E offer valuable insights into the interrelationships between these variables and their implications for South Africa's sustainable development. The consistency between the CPA trends and the forecasted trends, underscores the robustness of the predictive models in capturing historical patterns and projecting future trajectories accurately. This alignment enhances confidence in the forecasted data and reinforces the significance of electricity production and GDP growth as key determinants of CO2E in SA.

The downward trajectory in CO2E until 2023, followed by a significant peak in 2024 and a subsequent decline till 2027, reflects the complex dynamics driving carbon emission trends in the country. These peaks and declines closely emulate the trends of the forecasted electricity production data, underscoring non-renewable energy production as a significant driver of CO2E in the country. For instance, the surge in emissions in 2024, coinciding with an increase in electricity production from fossil fuels during the same time period, highlights the continued reliance on carbon-intensive energy sources, such as coal, and the challenges of transitioning to cleaner alternatives. The subsequent decline in CO2E and electricity production suggests efforts to address environmental concerns and adopt sustainable energy practices, albeit with fluctuations influenced by economic and policy factors.

The relationship between GDP growth and CO2E is nuanced, with instances where GDP growth corresponds to increases in emissions, but not in a one-to-one ratio. Whilst economic expansion often drives energy demand and carbon emissions in energy-intensive sectors, the transition to cleaner energy sources and energy efficiency measures can decouple GDP growth from emissions growth. This underscores the importance of pursuing sustainable economic development strategies that prioritise environmental conservation and mitigate carbon emissions.

In terms of the economy, the trends indicate the need for policies that promote sustainable growth and investment in renewable energy infrastructure. By aligning economic objectives with environmental sustainability goals, SA can enhance its competitiveness in the global market and attract investment in clean technologies. From a political standpoint, the findings emphasise the importance of proactive government intervention in addressing energy security, climate change, and infrastructure challenges. Policies to mitigate the impact of load-shedding on GDP are

essential, as interruptions in electricity supply disrupt economic activity. Although load-shedding has been shown to reduce CO2E, its negative impact on the economy and environmental risks makes it an unviable long-term solution.

To align with the global imperative of limiting the rise in average global temperature to below 1.5 degrees Celsius, SA must significantly reduce its carbon emissions to approximately 334 Mt of CO2E by the year 2030 (Climate Analytics 2022). The forecasted emissions for 2025 and 2027 suggest that SA is on track to meet its targets outlined in the 2030 National Development Plan, which require emissions levels ranging from 350 to 420 Mt of CO2E by 2030 (United Nations Development Programme 2021). However, there remains a crucial need to aim for the lower end of these targets, to achieve the ambitious 1.5-degree Celsius goal set forth in the Paris Agreement (Climate Action Tracker 2020). Furthermore, the forecasted CO2E for 2025 places SA closer to the lower end of their target range of 398 to 510 Mt of CO2E for that year (United Nations Development Programme 2021), which reflects significant progress towards achieving emission reduction objectives. Although SA demonstrates progress in meeting its NDP goals, it is essential to prioritise efforts towards achieving the lower ranges of these emissions goals, to effectively contribute towards global climate mitigation efforts. Therefore, upholding stringent emission reduction measures and accelerating the transition to renewable energy sources is imperative to ensuring sustained success and alignment with international climate objectives.

Additionally, the trend of fluctuating emissions suggests that efforts to transition away from fossil fuels may be underway, as indicated by the decrease in emissions in 2026 despite fluctuations in electricity production. However, further reductions in fossil fuel reliance are needed to align with global targets, as by 2030, the percentage of primary energy derived from fossil fuels must decrease to 67% worldwide (Rogelj *et al.* 2018).

This shows that significant progress has been made; however, further efforts are needed to curtail CO2E to meet the 2030 emissions target, which is essential to attaining the 1.5 degrees Celsius temperature limit, as outlined in the Paris Agreement. Moreover, further efforts are imperative to meet revised mitigation targets and attain net-zero emissions by 2050, as stated in the 2030 National Development Plan (Rogelj *et al.* 2018). Achieving these aims demands sustained political

commitment, investment in clean energy technologies and international collaboration, to effectively address global climate challenges.

5.5 Chapter summary

This chapter focused on understanding the complex patterns of CO2E in SA, by analysing both the CPA and ML models employed in this study. By dissecting historical data and predictive models, this chapter uncovers key insights into the factors driving CO2E in the region. The CPA results were showcased through the utilisation of a modified control chart and CUSUM graph, illustrating the detected change-points, which were further accompanied by a Bootstrap analysis table outlining the respective confidence levels for each change. The CPA results revealed significant shifts in variables like electricity net generation, GDP growth, and CO2E. Notable change-points occurred in 1990, 2000, and 2004 for electricity generation, 2000 and 2008 for GDP growth, and 2001, 2004 and 2008 for CO2E. These findings, supported by an average confidence level of 94%, underscore the accuracy of identifying change-points using this analytical approach. Furthermore, the CPA results highlight the interconnectedness of GDP growth, electricity production, and CO2E.

The fluctuations in economic activity and energy demand significantly influence carbon emission trends. As economies expand, there is a corresponding increase in energy demand to fuel industrial activities, resulting in higher levels of electricity production and, consequently, elevated CO2E. This observation is particularly pronounced in SA due to its heavy reliance on coal energy production, which is a significant emitter of CO2E from fossil fuel combustion. Furthermore, this intrinsic relationship between economic growth, energy production, and carbon emissions underscores the significant role of GDP growth and electricity generation in shaping CO2E trends. However, external factors such as load-shedding introduce a dynamic element to this analysis. Whilst load-shedding may temporarily reduce emissions, it also disrupts energy supply chains and has adverse economic impacts, such as business downtime. Thus, sustainable energy solutions are still necessary to achieve long-term goals, without compromising economic stability.

In respect to the ML results, the performance of AdaBoost regression was compared against several traditional algorithms, including Linear regression, Polynomial regression, Bayesian Linear regression, and KNN regression, to benchmark and determine the most effective approach for

CO2E prediction. Notably, the AdaBoost models emerged as the most optimal for forecasting CO2E, boasting robust evaluation metric scores, with an average RMSE score of 10,143.17 kt, MAE score of 9,642.64 kt, R^2 score of 0.90 and accuracy of 96.74%. Moreover, the analysis revealed that, on average, models trained with the AdaBoost algorithm outperformed traditional ML models. They reduced the RMSE score by 6,417.29 kt, decreased the MAE score by 4,358.09 kt, increased the R^2 score by 0.07, and improved accuracy by 0.60%.

In terms of hyperparameter tuning and model validation techniques, the comparative analysis favoured the repeated holdout method over cross-validation for improving model performance. This preference was supported by superior values in the statistical properties of the RMSE, MAE, and accuracy evaluation metrics in the repeated holdout method, when compared to cross-validation. Further reinforcing this observation, the box and whisker plots demonstrated that the mean and median values of these metrics were consistently better with the repeated holdout method. Although the R^2 metric exhibited better results with cross-validation, the repeated holdout method showcased significant improvements in the RMSE, MAE and accuracy scores.

To quantify the change in which repeated holdout improved the cross-validation results, statistical measures of the changes and box and whisker plots were also utilised. Without outliers removed, on average, the RMSE score decreased by 2,138.05 kt, the MAE score decreased by 3,323.53 kt, and the accuracy score improved by 0.88%. The variability in the changes of the RMSE and MAE evaluation metrics is significantly high, indicating the influence of underlying outliers that may affect results. Upon excluding these outliers, the average improvement of the cross-validation results by the repeated holdout method was 783.32 kt for the RMSE score, and 1,289.39 kt for the MAE score, which remains notably substantial.

An implementation guideline was formulated to determine when to apply the repeated holdout method, over other techniques. The analysis revealed that a significant improvement of 2,000 kt or greater in the RMSE and MAE scores is observed, when the cross-validation values of these metrics equal or exceed 15,000 kt. Similarly, for the accuracy score, an improvement of nearly 3% or more is observed, when the cross-validation value of this metric falls at or is below 94%. The changes observed in each evaluation metric serve as the criteria of the guideline, indicating when to employ repeated holdout methods, to enhance model performance. Additionally, the formulation of the guideline was achieved through the use of scatterplots.

This chapter concludes by validating the past trends of the CPA model through the forecasts generated by the ensemble learning technique, specifically the AdaBoost regressor integrated with the repeated holdout method. This model configuration was chosen due to the consistent optimal performance demonstrated by the ensemble learning and repeated holdout method, across all the models employed with these techniques. Furthermore, the validation process was conducted utilising line graphs. The results reveal that the forecasted CO2E trends validate the past trends generated by the CPA model. By comparing the historical data with the forecasted trends, it becomes evident that the CO2E trajectory remains consistent over time, affirming the reliability of the analytical methods employed.

SA is still facing some challenges in reducing carbon emissions, to meet the global temperature threshold of 1.5 degrees Celsius, as committed to in the Paris Agreement (Schleussner *et al.* 2022). This is underscored by the trajectory of forecasted emissions, indicating that SA is unlikely to meet the targeted CO2E level of 334 Mt by 2030, to achieve the temperature threshold. This is further evidenced by South Africa's failure to meet the lower range of the CO2E commitments outlined in the 2030 National Development Plan, as indicated by the forecasted data. However, whilst the target range has been met, signifying progress in achieving greater environmental sustainability, it is imperative to note that to remain below the temperature threshold of 1.5 degrees Celsius by 2030 and achieve a zero net emission rate by 2050, the CO2E targets need to be consistently met at the lower ranges. Despite efforts to transition to cleaner energy sources and mitigate environmental impact, through the commitment to international accords, the reliance on fossil fuels, such as coal, to produce energy continues to drive CO2E levels. Addressing this issue requires a multifaceted approach, including investment in renewable energy, the adoption of sustainable practices across industries, and the implementation of stringent environmental policies. Additionally, mitigating the adverse effects of external factors like load-shedding is crucial to maintaining economic stability, whilst striving for emissions reduction. The subsequent chapter will provide an overview of the study's summary, conclusion and its implications.

CHAPTER SIX: SUMMARY, CONCLUSIONS AND IMPLICATIONS OF THE STUDY

6.1 Introduction

This chapter provides a comprehensive overview of the study's findings, outlining key conclusions, implications, and potential avenues for future research. Furthermore, it also addresses how each research objective was attained.

6.2 Summary of conclusions

This section provides a summary of the study's findings and discusses how each research objective was accomplished. Each chapter either achieved or helped towards attaining a specific objective.

6.2.1 To identify the significant contributors of CO2E [RO1]

The first objective was accomplished in Chapters Three and Four of this study. In Chapter Three, a PRISMA-aligned systematic review was conducted to understand the primary drivers of CO2E, which is essential for constructing accurate models that inform policy and decision-making. From a pool of 61,903 publications, 19 articles were selected and analysed, following the application of exclusion criteria, revealing energy production and economic growth as the primary drivers. These factors were then utilised as input variables in both the CPA and ML models employed in this study. Chapter Four further validated these selections through cross-validation with additional studies, reaffirming economic growth and energy production as key determinants of CO2E in SA.

6.2.2 To identify the most relevant time-series datasets that can be used to implement the CPA and ML algorithms [RO2]

The second objective of this study was fulfilled in Chapters Three and Four. In Chapter Three, the systematic review identified the World Bank's Open Data repository as the most common data source for carbon emissions research, affirming its applicability for this study. The selection of the specific time-series datasets or variables was justified in Chapter Four, where an analytical process was undertaken, to determine the final variables used in the CPA and ensemble ML models. This process ensured the selection of the most pertinent data, which was of high-quality, focusing on

variables related to electricity production, economic growth, and CO2E in SA, thereby enhancing the accuracy and reliability of the study's outcomes.

6.2.3 To examine the CPA algorithm as a data mining technique, to identify the rapid changes in CO2E in SA [RO3]

This objective was achieved in Chapters Four and Five of this study. Chapter Four provided a detailed explanation of the CPA techniques utilised, namely the CUSUM and Bootstrap analysis, elucidating their mathematical formulas and technical workflows. Additionally, the Change-Point Analyzer tool, employed for implementing these techniques, was discussed. In Chapter Five, the application of these CPA techniques was demonstrated, showcasing modified control charts and CUSUM charts to pinpoint the exact change-points in the data. A Bootstrap analysis table depicted the corresponding confidence levels for each change-point, with an average confidence level of 94%, affirming the accuracy of this analytical approach. The CPA results underscored the interconnectedness of GDP growth, electricity production, and CO2E. This elucidated how economic expansion drives energy demand for industrial activities, subsequently leading to increased CO2 levels due to fossil fuel combustion, particularly coal. South Africa’s reliance on coal-fired energy highlights electricity production and economic growth as prominent factors that contribute to CO2E.

6.2.4 To validate past trends and predict future trends of CO2E in SA by implementing the appropriate ensemble ML algorithm [RO4]

The fourth objective of this study was attained in Chapters Four and Five. Chapter Four elaborated on the implementation of the ensemble ML algorithm, specifically AdaBoost, discussing its architecture and practical application utilising the Python programming language and relevant ML libraries like scikit-learn. Furthermore, this chapter discussed the layered approach of forecasting CO2E, focusing on predicting data for the input features, such as electricity production and economic growth, rather than utilising random input values for these features. In Chapter Five, the ML results were evaluated using the relevant performance metrics, highlighting the superior performance of AdaBoost models in forecasting CO2E, in comparison to the traditional ML models that were studied. Notably, the AdaBoost models exhibited robust evaluation scores, including an average RMSE score of 10,143.17 kt, a MAE score of 9,642.64 kt, an R^2 score of

0.90 and an accuracy of 96.74%. Subsequently, the forecasted CO2E trends were validated against past trends identified by the CPA model, using line graphs, revealing consistency between the two trends. The validation of electricity production and economic growth, as primary contributors to carbon emissions in SA, underscores the urgent need for targeted interventions to address these factors. Policymakers must prioritise investments in renewable energy sources and implement sustainable economic policies, to curb emissions effectively and mitigate the adverse environmental impacts of industrial growth.

6.2.5 To evaluate the effectiveness of the ML models by using the relevant performance metrics [RO5]

This objective was accomplished across Chapters Four and Five of this study. In Chapter Four, the rationale behind selecting RMSE, MAE, and R^2 performance metrics for model evaluation was justified, as they are commonly used in regression scenarios. The mathematical formulas for these metrics were discussed, along with the introduction of an accuracy score, as this metric is typically utilised for classification scenarios. In Chapter Five, these performance metrics were applied to assess ML model performance, with the results presented in various tables throughout the chapter. To quantify the results, the statistical properties of these metrics were also calculated and presented in this chapter. Statistical visualisations such as box and whisker plots were employed, to further quantify and illustrate the results of the performance metric scores, whilst statistical analyses, including scatterplots, aided in evaluating the impact of the repeated holdout method on cross-validation results.

6.3 Implications of the study

The findings of this study have significant implications for South Africa's environmental and economic trajectory. Firstly, the identification of economic growth and electricity production as primary drivers of CO2E emphasises the urgent need for strategic interventions to balance economic development with environmental sustainability. As SA strives to meet its emissions targets outlined in the 2030 National Development Plan and align with global climate goals, as indicated in the Paris Agreement, policymakers must prioritise initiatives that promote cleaner energy production and energy efficiency measures, whilst fostering sustainable economic growth.

This is because economic expansion drives energy demand and carbon emissions; hence, the transition to cleaner energy sources can decouple GDP growth from emissions growth.

Moreover, the observed fluctuations in CO2E trends, influenced by factors such as load-shedding and economic downturns, highlight the vulnerability of South Africa's energy sector to external shocks. This underscores the importance of investing in diversified energy sources and enhancing infrastructure resilience, to mitigate the adverse impacts of disruptions in energy supply on both the economy and the environment. Additionally, the study highlights the importance of proactive government intervention in addressing energy security, climate change, and infrastructure challenges. Policies to mitigate the impact of load-shedding on GDP are essential, as interruptions in electricity supply disrupt economic activity. This necessitates robust energy management strategies and investments in renewable energy to ensure long-term economic stability.

From a political perspective, the response of the government to the challenges posed by load-shedding and other economic downturns reflects the broader political landscape in SA. It underscores the importance of political leadership, policy coherence and accountability in addressing economic and environmental challenges. Political decisions regarding infrastructure investment, energy policy and regulatory frameworks plays a crucial role in shaping the country's economic trajectory and environmental sustainability. The government's handling of crises such as load-shedding can have significant implications on the public’s perception on their political legitimacy. Persistent challenges in the energy sector and perceived failures in governance can erode public trust in political parties and leaders.

The findings from the forecasted CO2E trends reveal a concerning discrepancy between South Africa's emissions trajectory and its commitments under the Paris Agreement. With the forecast indicating that SA is unlikely to meet the targeted CO2E level of 334 Mt by 2030, the nation risks falling short of its pledge to limit the global temperature to 1.5 degrees Celsius. This underscores the urgency of implementing more determined emission reduction measures, to align with the climate objectives set forth in the Paris Agreement.

Despite the challenges highlighted in meeting the CO2E commitments outlined in the Paris Agreement, there have been notable achievements in terms of progress towards the 2023 National Development Plan targets. The forecasted CO2E trends suggest that SA is on track to meet its emissions levels, ranging from 350 to 420 Mt of CO2E by 2030, as stipulated in the National

Development Plan. However, it is crucial to emphasise the importance of aiming for the lower end of these targets, to ensure meaningful contributions towards global climate mitigation efforts.

Looking ahead to the long-term objectives of achieving the 1.5 degrees Celsius temperature threshold by 2030, as mandated by the Paris Agreement, or a zero net emissions rate by 2050, SA faces significant challenges in transitioning its economy towards sustainability. Whilst progress has been made in reducing emissions and promoting renewable energy adoption, the pathway to achieving zero net emissions requires a comprehensive and concerted effort across all sectors of society. This includes sustained political commitment, substantial investment in clean energy technologies, and transformative policy measures to drive systemic change and ensure a just transition to a low-carbon economy by 2050.

6.3.1 Implications on data mining and ML research

The successful integration of CPA with ML in this study opens doors for a wider adoption of this analytical approach across various research domains. By combining traditional statistical methods with advanced ML algorithms, this approach offers a versatile framework for gaining deeper insights into complex research beyond carbon emissions. This promotes interdisciplinary collaboration and newer methodological approaches.

Moreover, the utilisation of ensemble learning techniques and repeated holdout methods demonstrates significant potential for their use in other studies with similar constraints. These methods not only enhance model performance and generalisation, but also optimise resource utilisation and promote reproducibility, offering valuable tools for a wide range of research endeavors across various disciplines.

6.4 Future works

Future research should focus on incorporating load-shedding dynamics into the analysis by assessing the correlation between load-shedding cycles or stages, economic growth, electricity production, and CO2E in SA. This could offer crucial insights into South Africa's energy landscape. Understanding and confirming how load-shedding impacts carbon emissions, economic activity, and energy production is vital for policymakers, to develop effective strategies to mitigate environmental impact, whilst ensuring economic stability. This analysis can shed light on the

interplay between energy supply and demand, and environmental sustainability in a South African context, informing more targeted policy interventions and infrastructure investments.

Additionally, future research endeavors should focus on forecasting when the next change-points may occur for the key socio-economic indicators employed in this study. Anticipating these shifts is essential for proactive decision-making and risk management, allowing stakeholders to adopt policies and investments accordingly. By identifying potential change-points in these indicators, policymakers can better prepare for economic fluctuations, energy demand shifts, and environmental challenges. This ultimately aids in adopting more resilient and sustainable development pathways for SA.

One notable limitation of this study is the lack of up-to-date data on renewable energy production, which restricts the ability to assess the contribution of renewable energy sources to overall electricity generation. This limitation emphasises the need for periodic monitoring of the ratio of coal-generated electricity to net electricity generation, particularly given the surge in renewable energy projects in SA, which are aimed at meeting global climate objectives. As the country increasingly embraces renewable energy initiatives, the ratio of coal-generated electricity to overall net generation is expected to fluctuate. This could potentially impact the predictive accuracy of the models trained in this study in the future.

Furthermore, with the rise in renewable energy projects in the country, there is a logical expectation of an increase in net electricity generation without a corresponding increase in CO2E. This underscores the importance of incorporating net electricity generated from renewable energy sources into model training, thereby ensuring the relevance and accuracy of the models over time. Hence, future research should also focus on incorporating renewable energy production into the analysis.

6.5 Chapter summary

This chapter outlined how each research objective was met and further provided detailed explanations of the chapters dedicated to addressing them. Thereafter, the study's implications for South Africa's economic, energy and environmental landscape were discussed. This chapter concluded by providing recommendations for future research, including the incorporation of load-

shedding into the analysis, to assess and confirm its impact on economic growth, energy production, and CO2E in the country. Additionally, it was suggested that future research should focus on predicting future change-points of the variables examined in this study, to anticipate fluctuations in energy demands, economic growth, and CO2 levels in SA. Lastly, it was suggested that future research should include renewable energy production into the analysis, as it plays a crucial role in shaping South Africa's future energy landscape and mitigating CO2E in the country.

REFERENCES

Abbas, S. and Fried, R. 2020. Robust control charts for the mean of a locally linear time series. *Journal of Statistical Computation and Simulation*, 90: 2741 - 2765.

Abelha, M., Fernandes, S., Mesquita, D., Seabra, F. and Ferreira-Oliveira, A. T. 2020. Graduate Employability and Competence Development in Higher Education—A Systematic Literature Review Using PRISMA. *Sustainability*, 12 (15): 5900.

Acartürk, C. 2014. Towards a systematic understanding of graphical cues in communication through statistical graphs. *Journal of Visual Languages & Computing*, 25 (2): 76-88.

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.

Agrawal, R. 2024. Polynomial Regression for Beginners. Available: https://www.analyticsvidhya.com/blog/2021/07/all-you-need-to-know-about-polynomial-regression/ (Accessed 05 July 2024).

Akpeji, K. O., Olasoji, A. O., Gaunt, C., Oyedokun, D. T., Awodele, K. O. and Folly, K. A. 2020. Economic impact of electricity supply interruptions in South Africa. *SAIEE Africa Research Journal*, 111 (2): 73-87.

Albert, M. S. L., Arroyo Marioli, F., Baffes, J., Hill, S. C., Inami, O., Kamin, S. B., Kasyanenko, S., Kenworthy, P. G., Khadan, J. and Kirby, P. A. 2023. Global Economic Prospects, June 2023.

Albrecher, H., Bladt, M., Kortschak, D., Prettenthaler, F. and Swierczynski, T. 2019. Flood occurrence change-point analysis in the paleoflood record from Lake Mondsee (NE Alps). *Global and Planetary Change*, 178: 65-76.

Al-Shehri, H., Al-Qarni, A., Al-Saati, L., Batoaq, A., Badukhen, H., Alrashed, S., Alhiyafi, J. and Olatunji, S. O. 2017. Student performance prediction using Support Vector Machine and K-Nearest Neighbor. In: Proceedings of 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE). 30 April-3 May 2017. 1-4.

Aminikhanghahi, S. and Cook, D. J. 2017. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51 (2): 339-367.

Angeyo, K. H., Muthama, MakokhaJ., W. and MuthamaJ., N. 2016. Long Term Change Point Detections in Total Ozone Column over East Africa via Maximal Overlap Discrete Wavelet Transform.

Anwar, A., Sinha, A., Sharif, A., Siddique, M., Irshad, S., Anwar, W. and Malik, S. 2022. The nexus between urbanization, renewable energy consumption, financial development, and CO2 emissions: evidence from selected Asian countries. *Environment, Development and Sustainability*, 24 (5): 6556-6576.

Aquilani, B., Piccarozzi, M., Abbate, T. and Codini, A. 2020. The role of open innovation and value co-creation in the challenging transition from industry 4.0 to society 5.0: Toward a theoretical framework. *Sustainability*, 12 (21): 8943.

Archer, E., Landman, W., Malherbe, J., Tadross, M. and Pretorius, S. 2019. *South Africa's winter rainfall region drought: A region in transition? Climate Risk Manage., 25, 100188*.

Arif, S., Mohamad Mohsin, M. F., Abu Bakar, A., Hamdan, A. and Syed Abdullah, S. 2017. Change point analysis: A statistical approach to detect potential abrupt change. *Jurnal Teknologi*, 79 (5): 147-159.

Aron, J. and Kingdon, G. 2007. *South African economic policy under democracy*: Oxford University Press.

Ateba, B. B., Prinsloo, J. J. and Gawlik, R. 2019. The significance of electricity supply sustainability to industrial growth in South Africa. *Energy Reports*, 5: 1324-1338.

Aurelio, Y. S., De Almeida, G. M., de Castro, C. L. and Braga, A. P. 2019. Learning from imbalanced data sets with weighted cross-entropy function. *Neural processing letters*, 50: 1937-1949.

Awan, T. M., Aziz, M., Sharif, A., Ch, T. R., Jasam, T. and Alvi, Y. 2022. Fake news during the pandemic times: A Systematic Literature Review using PRISMA. 6 (1): 49-60.

Awe, O. O. and Adepoju, A. A. 2020. Change-point detection in CO2 emission-energy consumption nexus using a recursive Bayesian estimation approach. *Statistics in Transition New Series*, 21 (1): 123-136.

Ba, A. and McKenna, S. A. 2014. Water quality monitoring with online change-point detection methods. *Journal of Hydroinformatics*, 17 (1): 7-19.

Bakay, M. S. and Ağbulut, Ü. 2021. Electricity production based forecasting of greenhouse gas emissions in Turkey with deep learning, support vector machine and artificial neural network algorithms. *Journal of Cleaner production*, 285: 125324.

Baker, L. 2017. Post-apartheid electricity policy and the emergence of South Africa's renewable energy sector. *The Political Economy of Clean Energy Transitions, Oxford University Press, Oxford*: 371-390.

Baldwin, S. A. and Larson, M. J. 2017. An introduction to using Bayesian linear regression with clinical data. Behaviour Research and Therapy, 98: 58-75.

Baloch, M. A., Ozturk, I., Bekun, F. V. and Khan, D. 2021. Modeling the dynamic linkage between financial development, energy innovation, and environmental quality: does globalization matter? *Business Strategy and the Environment*, 30 (1): 176-184.

Baltagi, B. H., Kao, C. and Liu, L. 2017. Estimation and identification of change points in panel models with nonstationary or stationary regressors and error term. *Econometric Reviews*, 36 (1-3): 85-102.

Banks, S. 2019. *Greenhouse Effects on Animals*. Available: https://sciencing.com/greenhouse-effects-animals-8050452.html (Accessed 15 December 2022).

Bardenet, R., Brendel, M., Kégl, B. and Sebag, M. 2013. Collaborative hyperparameter tuning. In: Proceedings of *International conference on machine learning*. PMLR, 199-207.

Bergstra, J. and Bengio, Y. 2012. *Random search for hyper-parameter optimization*.

Bergstra, J., Yamins, D. and Cox, D. 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In: Proceedings of *International conference on machine learning*. PMLR, 115-123.

Beskopylny, A. N., Stel'makh, S. A., Shcherban', E. M., Mailyan, L. R., Meskhi, B., Razveeva, I., Chernil'nik, A. and Beskopylny, N. 2022. Concrete strength prediction using machine learning methods CatBoost, k-nearest neighbors, support vector regression. Applied Sciences, 12 (21): 10864.

Bilina, R. and Lawford, S. 2012. Python for unified research in econometrics and statistics. *Econometric Reviews*, 31 (5): 558-591.

Binder, M., Moosbauer, J., Thomas, J. and Bischl, B. 2020. Multi-objective hyperparameter tuning and feature selection using filter ensembles. In: Proceedings of *Proceedings of the 2020 genetic and evolutionary computation conference*. 471-479.

Blokhin, A., Smith, A. and Perez, Y. 2021. *The 5 Countries That Produce the Most Carbon Dioxide (CO2)*. Available: https://www.investopedia.com/articles/investing/092915/5-countries-produce-most-carbon-dioxide-co2.asp (Accessed 20 January 2023).

Boamah, K. B., Du, J., Xu, L., Nyarko Mensah, C., Khan, M. A. S. and Allotey, D. K. 2020. A study on the responsiveness of the environment to international trade, energy consumption, and economic growth. The case of Ghana. *Energy Science & Engineering*, 8 (5): 1729-1745.

Borofsky, Y. 2021. *The Love-Hate Relationship with Self-Generation*. Available: https://energyforgrowth.org/article/the-love-hate-relationship-with-self-generation/ (Accessed 15 December 2023).

Borovicka, T., Jirina Jr, M., Kordik, P. and Jirina, M. 2012. Selecting representative data sets. *Advances in data mining knowledge discovery and applications*, 12: 43-70.

Brown, K. E., Hottle, T. A., Bandyopadhyay, R., Babaee, S., Dodder, R. S., Kaplan, P. O., Lenox, C. S. and Loughlin, D. H. 2018. Evolution of the United States energy system and related emissions under varying social and technological development paradigms: plausible scenarios for use in robust decision making. *Environmental Science & Technology*, 52 (14): 8027-8038.

Brownstone, D. and Valletta, R. G. 2001. The Bootstrap and Multiple Imputations: Harnessing Increased Computing Power for Improved Statistical Tests. *Journal of Economic Perspectives*, 15: 129-141.

Burton, J., Caetano, T. and McCall, B. 2018. *Coal transition in South Africa - Understanding the implications of a 2oC-compatible coal phase-out plan for South Africa*. DDRI & Climate Strategies.

Burton, J., Marquard, A. and McCall, B. 2019. Socio-economic considerations for a Paris Agreement-compatible coal transition in South Africa.

Buzun, N. and Avanesov, V. 2017. *Bootstrap for change point detection*: arXiv preprint arXiv:1710.07285.

Cai, Y., Sam, C. Y. and Chang, T. 2018. Nexus between clean energy consumption, economic growth and CO2 emissions. *Journal of Cleaner production*, 182: 1001-1011.

Cairoli, S. n.d. *Consequences of Carbon Emissions for Humans*. Available: https://sciencing.com/consequences-of-carbon-emissions-for-humans-12730960.html (Accessed 05 June 2022).

Camci, F. 2010. Change Point Detection in Time Series Data Using Support Vectors. *Int. J. Pattern Recognit. Artif. Intell.*, 24: 73-95.

Cerqueira, V., Torgo, L. and Mozetič, I. 2020. Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, 109: 1997-2028.

Chelani, A. B. 2011. Change detection using CUSUM and modified CUSUM method in air pollutant concentrations at traffic site in Delhi. *Stochastic Environmental Research and Risk Assessment*, 25: 827-834.

Chen, H., Tackie, E. A., Ahakwa, I., Musah, M., Salakpi, A., Alfred, M. and Atingabili, S. 2022. Does energy consumption, economic growth, urbanization, and population growth influence carbon emissions in the BRICS? Evidence from panel models robust to cross-sectional dependence and slope heterogeneity. *Environmental Science and Pollution Research*, 29 (25): 37598-37616.

Chen, S., Mihara, K. and Wen, J. 2018. Time series prediction of CO2, TVOC and HCHO based on machine learning at different sampling points. Building and Environment, 146: 238-246.

Chicco, D., Warrens, M. J. and Jurman, G. 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.

Chih-Pei, H. and Chang, Y.-Y. 2017. John W. Creswell, research design: Qualitative, quantitative, and mixed methods approaches. *Journal of Social and Administrative Sciences*, 4 (2): 205-207.

Chimphlee, S. and Chimphlee, W. 2023. Forecasting Carbon Dioxide Emission in Thailand Using Machine Learning Techniques. Indonesian Journal of Electrical Engineering and Informatics (IJEEI), 11 (3): 896-910.

Cho, H. 2016. Change-point detection in panel data via double CUSUM statistic. *Electronic Journal of Statistics*, 10: 2000-2038.

Clifford, C. 2022. *Weather disasters caused $145 billion in damage last year in the U.S., says NOAA*. Available: https://www.cnbc.com/2022/01/11/weather-disasters-cost-145-billion-last-year-in-the-us-says-noaa.html (Accessed 15 February 2022).

Climate Action Tracker. 2020. *Global update: Paris Agreement Turning Point*. Available: https://climateactiontracker.org/press/global-update-paris-agreement-turning-point/ (Accessed 27 December 2023).

Climate Analytics. 2022. *What is South Africa's pathway to limit global warming to 1.5°C?* Available: http://1p5ndc-pathways.climateanalytics.org/countries/south-africa/ (Accessed 26 December 2023).

Daumé, H. 2017. *A course in machine learning*. Hal Daumé III.

Dawari, A. 2022. *All About Adaboost*. Available: https://pub.towardsai.net/all-about-adaboost-ba232b5521e9 (Accessed 17 May 2023).

De Marchi, D. and Soille, P. 2018. *Advances in Interactive Processing and Visualisation with Jupyter on the JRC Big Data Platform (JEODPP)*. EasyChair.

Debone, D., Leite, V. P. and Miraglia, S. G. E. K. 2021. Modelling approach for carbon emissions, energy consumption and economic growth: A systematic review. *Urban Climate*, 37: 100849.

Delgado-Rodríguez, M. and Sillero-Arenas, M. 2018. Systematic review and meta-analysis. *Medicina Intensiva*, 42 (7): 444-453.

Desai, A., Gandhi, S., Gupta, S., Shah, M. and Patel, S. 2022. *Carbon Emission Prediction on the World Bank Dataset for Canada*: arXiv preprint arXiv:2211.17010.

Dethier, J.-J. 2007. Producing knowledge for development: Research at the World Bank. *Global Governance*, 13: 469.

Dhansay, T., Maupa, T., Twala, M., Sibewu, Z., Nengovhela, V., Mudau, P. T., Schalenkamp, M., Mashale, N., Muedi, T., Ndou, C., Zilibokwe, N. J., Mothupi, T., Safi, M. and Hicks, N. 2022. CO2 storage potential of basaltic rocks, Mpumalanga: Implications for the Just Transition. *South African Journal of Science*, 117 (7-8): 1-7.

Dikgwatlhe, P. 2018. Coal as a strategic resource in South Africa.

Dong, K., Dong, X. and Dong, C. 2019. Determinants of the global and regional CO2 emissions: What causes what and where? *Applied Economics*, 51 (46): 5031-5044.

Doyle-Kent, M. and Kopacek, P. 2020. Industry 5.0: Is the manufacturing industry on the cusp of a new revolution? In: Proceedings of *Proceedings of the International Symposium for Production Research 2019*. Springer, 432-441.

Efron, B. 1992. Bootstrap methods: another look at the jackknife. In: *Breakthroughs in statistics: Methodology and distribution*. Springer, 569-593.

El Bouchefry, K. and de Souza, R. S. 2020. Learning in big data: Introduction to machine learning. In: *Knowledge discovery in big data from astronomy and earth observation*. Elsevier, 225-249.

Elton Bryson Stephens Company. n.d. *GreenFILE*. Available: https://www.ebsco.com/products/research-databases/greenfile (Accessed 17 January 2023).

Embarak, D. O. 2018. *Data analysis and visualization using python*. Springer.

Emmert-Streib, F. and Dehmer, M. 2022. Taxonomy of machine learning paradigms: A data-centric perspective. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12 (5): e1470.

Erschens, R., Keifenheim, K. E., Herrmann-Werner, A., Loda, T., Schwille-Kiuntke, J., Bugaj, T. J., Nikendei, C., Huhn, D., Zipfel, S. and Junne, F. 2019. Professional burnout among medical students: Systematic literature review and meta-analysis. *Medical Teacher*, 41 (2): 172-183.

Eskom. n.d. *Our Recent Past*. Available: https://www.eskom.co.za/heritage/history-in-decades/eskom-2003-2012/ (Accessed 30 November 2023).

Espoir, D. K., Sunge, R. and Bannor, F. 2023. Exploring the dynamic effect of economic growth on carbon dioxide emissions in Africa: evidence from panel PMG estimator. *Environmental Science and Pollution Research*: 1-18.

European Union Joint Research Centre. 2021. *Carbon Footprint by Country 2023*. Available: https://worldpopulationreview.com/country-rankings/carbon-footprint-by-country (Accessed 20 January 2022).

Fantom, N. J. and Serajuddin, U. 2016. *The World Bank's classification of countries by income*: World Bank Policy Research Working Paper.

Faulkner, D. and Loewald, C. 2008. *Policy change and economic growth: A case study of South Africa*. International Bank for Reconstruction and Development/The World Bank.

Fernandez-Basso, C., Ruiz, M. D. and Martin-Bautista, M. J. 2020. A fuzzy mining approach for energy efficiency in a Big Data framework. *IEEE Transactions on Fuzzy Systems*, 28 (11): 2747-2758.

Fernando, J. 2023. *R-Squared: Definition, Calculation Formula, Uses, and Limitations*. Available: https://www.investopedia.com/terms/r/r-squared.asp (Accessed 22 December 2023).

Formplus. 2021. *Experimental Research Designs: Types, Examples & Methods*. Available: https://www.formpl.us/blog/experimental-research (Accessed 12 November 2022).

Fortea-Sanchis, C. and Escrig-Sos, J. 2019. Quality control techniques in surgery: application of cumulative sum (CUSUM) charts. *Cirugía Española (English Edition)*, 97 (2): 65-70.

Frost, J. n.d. *How To Interpret R-squared in Regression Analysis*. Available: https://statisticsbyjim.com/regression/interpret-r-squared-regression/ (Accessed 22 December 2023).

Galama, J. T. and Scholtens, B. 2021. A meta-analysis of the relationship between companies' greenhouse gas emissions and financial performance. *Environmental Research Letters*, 16 (4): 043006.

Gan, W., Lin, J. C. W., Chao, H. C. and Zhan, J. 2017. Data mining in distributed environment: a survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7 (6): e1216.

Gao, L., Peng, K., Gao, F. and Guan, X. 2010. AdaBoost regression algorithm based on classification-type loss. In: Proceedings of *2010 8th World Congress on Intelligent Control and Automation*. 7-9 July 2010. 682-687.

Georgescu, V. 2012. Online change-point detection in financial time series: challenges and experimental evidence with frequentist and Bayesian setups. In: *Methods for Decision Making in an Uncertain Environment*. World Scientific, 131-145. Available: https://doi.org/10.1142/9789814415774_0009 (Accessed 15 May 2023).

Gevorkyan, M. N., Demidova, A. V. and Kulyabov, D. S. 2020. Comparative analysis of machine learning methods by the example of the problem of determining muon decay. *Discrete and Continuous Models and Applied Computational Science*, 28 (2): 105-119.

Ghawi, R. and Pfeffer, J. 2019. Efficient hyperparameter tuning with grid search for text categorization using kNN approach with BM25 similarity. *Open Computer Science*, 9 (1): 160-180.

Glavas, D. 2024. Determinants of CO2 Emissions: A Machine Learning Approach. International Advances in Economic Research, 30 (2): 215-217.

Goldberg, A. 2015. The economic impact of load shedding: The case of South African retailers. University of Pretoria.

Grosse, K., Lee, T., Park, Y., Backes, M. and Molloy, I. 2020. *A new measure for overfitting and its implications for backdooring of deep learning*: CoRR.

Gu, S., Fu, B., Thriveni, T., Fujita, T. and Ahn, J. W. 2019. Coupled LMDI and system dynamics model for estimating urban CO2 emission mitigation potential in Shanghai, China. *Journal of Cleaner production*, 240: 118034.

Gu, X. and Angelov, P. P. 2021. Multiclass fuzzily weighted adaptive-boosting-based self-organizing fuzzy inference ensemble systems for classification. *IEEE Transactions on Fuzzy Systems*, 30 (9): 3722-3735.

Guan, H. 2021. Application of improved adaboost iteration in small sample for prediction of oral performance. In: Proceedings of *Journal of Physics: Conference Series*. IOP Publishing, 012049.

Gupta, M., Wadhvani, R. and Rasool, A. 2022. Real-time Change-Point Detection: A deep neural network-based adaptive approach for detecting changes in multivariate time series data. Expert Systems with Applications, 209: 118260.

Hambali, M., Saheed, Y., Oladele, T. and Gbolagade, M. 2019. ADABOOST ensemble algorithms for breast cancer classification. *Journal of Advances in Computer Research*, 10 (2): 31-52.

Han, C. H. and Sim, K.-B. 2008. Real-time face detection using AdaBoot algorithm. In: Proceedings of *2008 International Conference on Control, Automation and Systems*. IEEE, 1892-1895.

Han, P., Zeng, N., Oda, T., Zhang, W., Lin, X., Liu, D., Cai, Q., Ma, X., Meng, W. and Wang, G. 2020. A city-level comparison of fossil-fuel and industry processes-induced CO2 emissions over the Beijing-Tianjin-Hebei region from eight emission inventories. *Carbon Balance and Management*, 15 (1): 1-16.

Hansen, J., Sato, M. and Ruedy, R. 2012. Perception of climate change. *Proceedings of the National Academy of Sciences*, 109 (37): E2415-E2423.

Hanto, J., Schroth, A., Krawielicki, L., Oei, P.-Y. and Burton, J. 2022. South Africa's energy transition – Unraveling its political economy. *Energy for sustainable development*, 69: 164-178.

Hao, J. and Ho, T. K. 2019. Machine learning made easy: a review of scikit-learn package in python programming language. *Journal of Educational and Behavioral Statistics*, 44 (3): 348-361.

Hasani, H., Mardi, S., Shakerian, S., Taherzadeh-Ghahfarokhi, N. and Mardi, P. 2020. The Novel Coronavirus Disease (COVID-19): A PRISMA Systematic Review and Meta-Analysis of Clinical and Paraclinical Characteristics. *BioMed Research International*, 2020: 3149020.

Haseeb, M., Wattanapongphasuk, S. and Jermsittiparsert, K. 2019. Financial Development, Market Freedom, Political Stability, Economic Growth and CO 2 Emissions: An Unexplored Nexus in ASEAN Countries. *Contemporary Economics*, 13 (3): 363-374.

Hastie, T., Rosset, S., Zhu, J. and Zou, H. 2009. Multi-class adaboost. *Statistics and its Interface*, 2 (3): 349-360.

Hellerstein, J. M. 2008. Quantitative data cleaning for large databases. *United Nations Economic Commission for Europe (UNECE)*, 25: 1-42.

Henneman, L. R. F., Rafaj, P., Annegarn, H. J. and Klausbruckner, C. 2016. Assessing emissions levels and costs associated with climate and air pollution policies in South Africa. *Energy Policy*, 89: 160-170.

Herbst, L. and Lalk, J. 2015. A review of the policy documents behind South Africa's Renewable Energy Independent Power Producer Procurement Programme: How its hits and misses impact society. In: Proceedings of *2015 IEEE International Symposium on Technology and Society (ISTAS)*. IEEE, 1-6.

Hirsch, A., Levy, B. and Nxele, M. 2021. Politics and Economic Policymaking in South Africa since 1994. *The Oxford Handbook of the South African Economy*: 66.

Hlongwane, N. W., Mahapa, R. and Nthebe, T. C. 2023. The nexus between foreign direct investment and electricity consumption in South Africa. *International Journal of Energy Economics and Policy*, 13 (5): 213-220.

Hofer, S. M. and Piccinin, A. M. 2009. Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. *Psychological methods*, 14 (2): 150.

Horváth, L. and Rice, G. 2014. Extensions of some classical methods in change point analysis. Test, 23 (2): 219-255.

Hosseini, S. M., Saifoddin, A., Shirmohammadi, R. and Aslani, A. 2019. Forecasting of CO2 emissions in Iran based on time series and regression analysis. *Energy Reports*, 5: 619-631.

Huang, M. and Wang, B. 2016. Factors influencing CO2 emissions in China based on grey relational analysis. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, 38 (4): 555-561.

Hussain, S., Song, L.-x., Ahmad, S. and Riaz, M. 2019. A new auxiliary information based cumulative sum median control chart for location monitoring. *Frontiers of Information Technology & Electronic Engineering*, 20: 554-570.

Ibrahim, M. and Alagidede, P. 2018. Nonlinearities in financial development–economic growth nexus: Evidence from sub-Saharan Africa. *Research in International Business and Finance*, 46: 95-104.

International Research. n.d. *What is Quantitative Research?* Available: https://www.sisinternational.com/what-is-quantitative-research/ (Accessed 30 May 2022).

Iorember, P. T., Jelilov, G., Usman, O., Işık, A. and Celik, B. 2021. The influence of renewable energy use, human capital, and trade on environmental quality in South Africa: multiple structural breaks cointegration approach. *Environmental Science and Pollution Research*, 28 (11): 13162-13174.

Jakhrani, A. Q., Rigit, A. R. H., Othman, A.-K., Samo, S. R. and Kamboh, S. A. 2012. Estimation of carbon footprints from diesel generator emissions. In: Proceedings of *2012 International Conference on Green and Ubiquitous Technology*. IEEE, 78-81.

James, N. and Menzies, M. 2022. Global and regional changes in carbon dioxide emissions: 1970–2019. *Physica A: Statistical Mechanics and its Applications*, 608: 128302.

James, N., Menzies, M., Azizi, L. and Chan, J. 2020. Novel semi-metrics for multivariate change point analysis and anomaly detection. *Physica D: Nonlinear Phenomena*, 412: 132636.

Javaid, M., Haleem, A., Singh, R. P., Suman, R. and Gonzalez, E. S. 2022. Understanding the adoption of Industry 4.0 technologies in improving environmental sustainability. *Sustainable Operations and Computers*, 3: 203-217.

Jayakumar, N., Brunckhorst, O., Dasgupta, P., Khan, M. S. and Ahmed, K. 2015. e-Learning in Surgical Education: A Systematic Review. *Journal of Surgical Education*, 72 (6): 1145-1157.

Jiang, F., Zhu, Q. and Tian, T. 2023. An ensemble interval prediction model with change point detection and interval perturbation-based adjustment strategy: A case study of air quality. Expert Systems with Applications, 222: 119823.

Jin, X., Li, S., Zhang, W., Zhu, J. and Sun, J. 2020. Prediction of soil-available potassium content with visible near-infrared ray spectroscopy of different pretreatment transformations by the boosting algorithms. *Applied Sciences*, 10 (4): 1520.

Johnsson, F., Kjärstad, J. and Rootzén, J. 2019. The threat to climate change mitigation posed by the abundance of fossil fuels. *Climate Policy*, 19 (2): 258-274.

Joseph, V. R. and Mustaffa, N. K. 2023. Carbon emissions management in construction operations: a systematic review. *Engineering, Construction and Architectural Management*, 30 (3): 1271-1299.

Joshua, U. and Alola, A. A. 2020. Accounting for environmental sustainability from coal-led growth in South Africa: the role of employment and FDI. *Environmental Science and Pollution Research*, 27: 17706-17716.

Josselin, J.-M. and Le Maux, B. 2017. Sampling and Construction of Variables. In: *Statistical Tools for Program Evaluation : Methods and Applications to Economic Policy, Public Health, and Education*. Cham: Springer International Publishing, 15-43. Available: https://doi.org/10.1007/978-3-319-52827-4_2 (Accessed 05 August 2023).

Kanade, V. 2023. What Is Linear Regression? Types, Equation, Examples, and Best Practices for 2022. Available: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-linear-regression/ (Accessed 03 July 2024).

Kapoor, N. R., Kumar, A., Kumar, A., Kumar, A., Mohammed, M. A., Kumar, K., Kadry, S. and Lim, S. 2022. Machine Learning-Based CO2 Prediction for Office Room: A Pilot Study. Wireless Communications and Mobile Computing, 2022 (1): 9404807.

Kebede, S. 2017. Modeling energy consumption, CO2 emissions and economic growth nexus in Ethiopia: evidence from ARDL approach to cointegration and causality analysis.

Kebonye, N. M. 2021. Exploring the novel support points-based split method on a soil dataset. *Measurement*, 186: 110131.

Khashaba, A. S. 2020. Evaluation of the Effectiveness of Online Peer-Based Formative Assessments (PeerWise) to Enhance Student Learning in Physiology: A Systematic Review Using PRISMA Guidelines. *International Journal of Research in Education and Science*, 6 (4): 613-628.

Killick, R. and Eckley, I. 2014. changepoint: An R package for changepoint analysis. *Journal of statistical software*, 58 (3): 1-19.

Kim, K. and Kim, Y. 2012. International comparison of industrial CO2 emission trends and the energy efficiency paradox utilizing production-based decomposition. *Energy Economics*, 34 (5): 1724-1741.

Koca Akkaya, E. and Akkaya, A. V. 2023. Development and performance comparison of optimized machine learning-based regression models for predicting energy-related carbon dioxide emissions. Environmental Science and Pollution Research, 30 (58): 122381-122392.

Koehrsen, W. 2018. Introduction to Bayesian Linear Regression. Available: https://towardsdatascience.com/introduction-to-bayesian-linear-regression-e66e60791ea7 (Accessed 06 July 2024).

Kohler, M. 2014. Differential electricity pricing and energy efficiency in South Africa. *Energy*, 64: 524-532.

Köne, A. Ç. and Büke, T. 2010. Forecasting of CO2 emissions from fuel combustion using trend analysis. *Renewable and Sustainable Energy Reviews*, 14 (9): 2906-2915.

Kong, D., Zhu, J., Duan, C., Lu, L. and Chen, D. 2020. Bayesian linear regression for surface roughness prediction. Mechanical Systems and Signal Processing, 142: 106770.

Kordos, J. 2018. Small area statistics and quality management–the Polish perspective. *Śląski Przegląd Statystyczny*, (16 (22)): 37-54.

Kotsiantis, S. B., Zaharakis, I. and Pintelas, P. 2007. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160 (1): 3-24.

Kraus, S., Breier, M. and Dasí-Rodríguez, S. 2020. The art of crafting a systematic literature review in entrepreneurship research. *International Entrepreneurship and Management Journal*, 16 (3): 1023-1042.

Krikler, B. E., Davignon, O., Kreczko, L., Linacre, J., Olaiya, E. O. and Sakuma, T. 2019. Pandas DataFrames for a FAST binned analysis at CMS. In: Proceedings of *EPJ Web of Conferences*. EDP Sciences, 06035.

Kumar, B. V., Mahalanobis, A. and Juday, R. D. 2005. *Correlation pattern recognition*. Cambridge university press.

Kwakwa, P. A. and Adusah-Poku, F. 2020. The carbon dioxide emission effects of domestic credit and manufacturing indicators in South Africa. *Management of Environmental Quality: An International Journal*, 31 (6): 1531-1548.

Law, P.-M., Endert, A. and Stasko, J. T. 2020. Characterizing Automated Data Insights. *2020 IEEE Visualization Conference (VIS)*: 171-175.

Lawlor, P. 2023. *SA's load shedding constraint and its impact on different economic sectors* Available: https://www.investec.com/en_za/focus/economy/sa-s-load-shedding-how-the-sectors-are-being-affected.html (Accessed 20 December 2023).

Le Quéré, C., Jackson, R. B., Jones, M. W., Smith, A. J., Abernethy, S., Andrew, R. M., De-Gol, A. J., Willis, D. R., Shan, Y. and Canadell, J. G. 2020. Temporary reduction in daily global CO2 emissions during the COVID-19 forced confinement. *Nature climate change*, 10 (7): 647-653.

Le Thien, M.-A., Cordier, Q., Lifante, J.-c., Carty, M. J. and Duclos, A. 2023. Control Charts Usage for Monitoring Performance in Surgery: A Systematic Review. *Journal of Patient Safety*, 19: 110 - 116.

Lee, L. C., Liong, C.-Y. and Jemain, A. A. 2018. Validity of the best practice in splitting data for hold-out validation strategy as performed on the ink strokes in the context of forensic science. *Microchemical Journal*, 139: 125-133.

Lee, S. H. and Lim, S. 2017. Clinical effectiveness of acupuncture on Parkinson disease: A PRISMA-compliant systematic review and meta-analysis. *Medicine (Baltimore)*, 96 (3): e5836.

Letete, T., Guma, M. and Marquard, A. 2010. *Information on climate change in South Africa: greenhouse gas emissions and mitigation options*.

Levendis, J. D. 2018. Stationarity and Invertibility. In: Levendis, J. D. ed. *Time Series Econometrics: Learning Through Replication*. Cham: Springer International Publishing, 81-99. Available: https://doi.org/10.1007/978-3-319-98282-3_4 (Accessed 20 November 2023).

Li, M., Wang, W., De, G., Ji, X. and Tan, Z. 2018. Forecasting Carbon Emissions Related to Energy Consumption in Beijing-Tianjin-Hebei Region Based on Grey Prediction Theory and Extreme Learning Machine Optimized by Support Vector Machine Algorithm. *Energies*, 11 (9): 2475.

Li, Y. 2020. Forecasting Chinese carbon emissions based on a novel time series prediction method. *Energy Science & Engineering*, 8 (7): 2274-2285.

Li, Y. and Guo, J. 2022. The asymmetric impacts of oil price and shocks on inflation in BRICS: a multiple threshold nonlinear ARDL model. *Applied Economics*, 54 (12): 1377-1395.

Li, Y. and Sun, Y. 2021. Modeling and predicting city-level CO2 emissions using open access data and machine learning. Environmental Science and Pollution Research, 28 (15): 19260-19271.

Li, Y., Li, T. and Lu, S. 2021. Forecast of urban traffic carbon emission and analysis of influencing factors. *Energy Efficiency*, 14 (8): 84.

Liao, C., Wang, S., Fang, J., Zheng, H., Liu, J. and Zhang, Y. 2019. Driving forces of provincial-level CO2 emissions in China's power sector based on LMDI method. *Energy Procedia*, 158: 3859-3864.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., Clarke, M., Devereaux, P. J., Kleijnen, J. and Moher, D. 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Annals of internal medicine*, 151 (4): W-65-W-94.

Lin, S., Wang, S., Marinova, D., Zhao, D. and Hong, J. 2017. Impacts of urbanization and real economic development on CO2 emissions in non-high income countries: Empirical research based on the extended STIRPAT model. *Journal of Cleaner production*, 166: 952-966.

Ling, W., Li, Y., Jiang, W., Sui, Y. and Zhao, H. L. 2015. Common Mechanism of Pathogenesis in Gastrointestinal Diseases Implied by Consistent Efficacy of Single Chinese Medicine Formula: A PRISMA-Compliant Systematic Review and Meta-Analysis. *Medicine (Baltimore)*, 94 (27): e1111.

Liu, D., Guo, X. and Xiao, B. 2019. What causes growth of global greenhouse gas emissions? Evidence from 40 countries. *Science of The Total Environment*, 661: 750-766.

Liu, H., Wong, W.-K., The Cong, P., Nassani, A. A., Haffar, M. and Abu-Rumman, A. 2023. Linkage among Urbanization, energy Consumption, economic growth and carbon Emissions. Panel data analysis for China using ARDL model. *Fuel*, 332: 126122.

Liu, X., Lu, D., Zhang, A., Liu, Q. and Jiang, G. 2022. Data-driven machine learning in environmental pollution: gains and problems. *Environmental Science & Technology*, 56 (4): 2124-2133.

Lohwasser, J., Schaffer, A. and Brieden, A. 2020. The role of demographic and economic drivers on the environment in traditional and standardized STIRPAT analysis. *Ecological Economics*, 178: 106811.

Ma, L., Grant, A. J. and Sofronov, G. 2020. Multiple change point detection and validation in autoregressive time series data. *Statistical Papers*, 61: 1507-1528.

Ma, X., Wang, C., Dong, B., Gu, G., Chen, R., Li, Y., Zou, H., Zhang, W. and Li, Q. 2019. Carbon emissions from energy consumption in China: Its measurement and driving factors. *Science of The Total Environment*, 648: 1411-1420.

Mabuni, D. and Babu, S. 2021. High accurate and a variant of k-fold cross validation technique for predicting the decision tree classifier accuracy. *Int J Innov Tech Explor Eng*, 10: 105-110.

Madubeko, V. 2010. The global financial crisis and its impact on the South African economy. University of Fort Hare Alice.

Mafukata, M. A. 2017. The Impact of the 2008–2009 Global Financial Crisis on Employment Creation and Retention in the Platinum Group Metals (PGMs) Mining Sub-sector in South Africa. *Global Financial Crisis and Its Ramifications on Capital Markets: Opportunities and Threats in Volatile Economic Conditions*: 569-585.

Mahata, A. 2024. Polynomial Regression. Available: https://akashdeepmahata.medium.com/polynomial-regression-ebf3f959a14a (Accessed 06 July 2024).

Maisonnave, H., Mabugu, R., Chitiga, M., Robichaud, V. and Decaluwé, B. 2009. The Impact of the international economic crisis in South Africa. *CIRPéE. Cahier de recherché/Working Paper*: 09-52.

Mali, K. 2024. What is Linear Regression? Available: https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/ (Accessed 02 July 2024).

Mandeya, S. M. T. and Ho, S.-Y. 2021. Inflation, inflation uncertainty and the economic growth nexus: An impact study of South Africa. *MethodsX*, 8: 101501.

Männistö, M., Mikkonen, K., Kuivila, H.-M., Virtanen, M., Kyngäs, H. and Kääriäinen, M. 2020. Digital collaborative learning in nursing education: a systematic review. *Scandinavian Journal of Caring Sciences*, 34 (2): 280-292.

Manzano-León, A., Camacho-Lazarraga, P., Guerrero, M. A., Guerrero-Puerta, L., Aguilar-Parra, J. M., Trigueros, R. and Alias, A. 2021. Between Level Up and Game Over: A Systematic Literature Review of Gamification in Education. *Sustainability*, 13 (4): 2247.

Mardani, A., Liao, H., Nilashi, M., Alrasheedi, M. and Cavallaro, F. 2020. A multi-stage method to predict carbon dioxide emissions using dimensionality reduction, clustering, and machine learning techniques. *Journal of Cleaner production*, 275: 122942.

Mardani, A., Streimikiene, D., Cavallaro, F., Loganathan, N. and Khoshnoudi, M. 2019. Carbon dioxide (CO2) emissions and economic growth: A systematic review of two decades of research from 1995 to 2017. *Science of The Total Environment*, 649: 31-49.

Maroney, D. B. 2019. *Forecasting Electricity Generation: An AR(1) Approach*: Econometric Modeling: Agriculture.

Martinez, Q., Chen, C., Xia, J. and Bahai, H. 2023. Sequence-to-sequence change-point detection in single-particle trajectories via recurrent neural network for measuring self-diffusion. Transport in Porous Media, 147 (3): 679-701.

Martinez, W. and Gray, J. B. 2014. The role of margins in boosting and ensemble performance. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6 (2): 124-131.

Mashishi, N. 2021. *South Africa the 12th biggest source of greenhouse gases? Yes, but that's not the only measure that matters*. Available: https://www.polity.org.za/article/south-africa-the-12th-biggest-source-of-greenhouse-gases-yes-but-thats-not-the-only-measure-that-matters-2021-04-19 (Accessed 30 January 2022).

Maulud, D. and Abdulazeez, A. M. 2020. A review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends, 1 (2): 140-147.

Mays, K. L., Shepson, P. B., Stirm, B. H., Karion, A., Sweeney, C. and Gurney, K. R. 2009. Aircraft-based measurements of the carbon footprint of Indianapolis. *Environmental Science & Technology*, 43 (20): 7816-7823.

McKinney, W. 2022. *Python for data analysis*. " O'Reilly Media, Inc.".

Medlock, K. B. and Soligo, R. 2001. Economic Development and End-Use Energy Demand. *The Energy Journal*, 22: 77 - 105.

Meng, L., Guo, J. e., Chai, J. and Zhang, Z. 2011. China's regional CO2 emissions: characteristics, inter-regional transfer and emission reduction policies. *Energy Policy*, 39 (10): 6136-6144.

Mensah, I. A., Sun, M., Gao, C., Omari-Sasu, A. Y., Zhu, D., Ampimah, B. C. and Quarcoo, A. 2019. Analysis on the nexus of economic growth, fossil fuel energy consumption, CO2 emissions and oil price in Africa based on a PMG panel ARDL approach. *Journal of Cleaner production*, 228: 161-174.

Menyah, K. and Wolde-Rufael, Y. 2010. Energy consumption, pollutant emissions and economic growth in South Africa. *Energy Economics*, 32 (6): 1374-1382.

Miller, D., Saunders, R. and Oloyede, O. 2008. South African Corporations and post-Apartheid Expansion in Africa–creating a new regional space. *African Sociological Review/Revue Africaine de Sociologie*, 12 (1): 1-19.

Miller, T. H., Gallidabino, M. D., MacRae, J. I., Hogstrand, C., Bury, N. R., Barron, L. P., Snape, J. R. and Owen, S. F. 2018. Machine Learning for Environmental Toxicology: A Call for Integration and Innovation. *Environmental Science & Technology*, 52 22: 12953-12955.

Millman, K. J. and Aivazis, M. 2011. Python for scientists and engineers. *Computing in science & engineering*, 13 (2): 9-12.

Min, H. and Luo, X. 2016. Calibration of soft sensor by using Just-in-time modeling and AdaBoost learning method. *Chinese journal of chemical engineering*, 24 (8): 1038-1046.

Mirzania, P., Gordon, J. A., Balta-Ozkan, N., Sayan, R. C. and Marais, L. 2023. Barriers to powering past coal: Implications for a just energy transition in South Africa. *Energy Research & Social Science*, 101: 103122.

Mishra, A. 2011. Accessing the World Bank open data programmatically. *XRDS: Crossroads, The ACM Magazine for Students*, 18 (2): 44-45.

Mısır, O. and Akar, M. 2022. Efficiency and core loss map estimation with machine learning based multivariate polynomial regression model. Mathematics, 10 (19): 3691.

Mngomezulu, B. R. 2016. Endogenous and Exogenous Factors Affecting Energy Reforms in Africa: A Critical Analysis. *Journal of Social Sciences*, 49 (3-1): 195-204.

Mohamed, G. M., Patel, S. S. and Naicker, N. 2023. Data augmentation for deep learning algorithms that perform driver drowsiness detection. International Journal of Advanced Computer Science and Applications, 14 (1): 233-248.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P. and Stewart, L. A. 2015. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic reviews*, 4 (1): 1-9.

Mohri, M., Rostamizadeh, A. and Talwalkar, A. 2018. *Foundations of machine learning*. MIT press.

Monyeki, P. 2021. Data mining to analyse recurrent crime in South Africa.

Monyeki, P., Naicker, N. and Obagbuwa, I. C. 2020. Change-point analysis: An effective technique for detecting abrupt change in the homicide trends in a democratic South Africa. *The Scientific World Journal*, 2020: 1-10.

Morales-Hernández, A., Van Nieuwenhuyse, I. and Rojas Gonzalez, S. 2023. A survey on multi-objective hyperparameter optimization algorithms for machine learning. *Artificial Intelligence Review*, 56 (8): 8043-8093.

Morse, A. 2018. *The Importance of Reducing a Carbon Footprint*. Available: https://sciencing.com/the-importance-of-reducing-a-carbon-footprint-5229039.html (Accessed 17 December 2022).

Musa, K. S., Maijama'a, R. and Yakubu, M. 2021. The causality between urbanization, industrialization and CO2 emissions in Nigeria: Evidence from Toda and Yamamoto Approach. *Energy Economics Letters*, 8 (1): 1-14.

Naidoo, C. 2023. The Impact of Load Shedding on the South Africa Economy. *Journal of Public Administration*, 58 (1): 7-16.

Naidoo, P. 2021. South Africa's unemployment rate is now highest in the world. *The Capital News*, 44 (36): 15.

Nathaniel, S., Barua, S., Hussain, H. and Adeleye, N. 2021. The determinants and interrelationship of carbon emissions and economic growth in African economies: Fresh insights from static and dynamic models. *Journal of Public Affairs*, 21 (1): e2141.

National Oceanic and Atmospheric Administration. 2021. *Climate change impacts* Available: https://www.noaa.gov/education/resource-collections/climate/climate-change-impacts (Accessed 25 January 2022).

Nguyen, D. K., Huynh, T. L. D. and Nasir, M. A. 2021. Carbon emissions determinants and forecasting: Evidence from G6 countries. *Journal of Environmental Management*, 285: 111988.

Nowak, M. 2005. The first ten years after apartheid: an overview of the South African economy. *Post-Apartheid South Africa: The First Ten Years. Washington, DC: International Monetary Fund*: 1-10.

Nunez, C. 2019. *Carbon dioxide levels are at a record high. Here's what you need to know.* Available: https://www.nationalgeographic.com/environment/article/greenhouse-gases (Accessed 02 February 2022).

Nyeadi, J. D. 2023. The impact of financial development and foreign direct investment on environmental sustainability in Sub-Saharan Africa: using PMG-ARDL approach. *Economic Research-Ekonomska Istraživanja*, 36 (2): 2106270.

Omarjee, L. 2022. *Table Mountain wildfire, floods and 'severe weather' cost SA billions in 2021*. Available: https://www.news24.com/fin24/economy/table-mountain-wildfire-floods-and-severe-weather-cost-sa-billions-in-2021-20220126#:~:text=%22Approximately%2010%20500%20people%20lost,or%20about%20R5.2%20trillion (Accessed 01 March 2022).

Osmanski, S. 2020. *How Do Carbon Emissions Affect the Environment?* Available: https://www.greenmatters.com/p/how-do-carbon-emissions-affect-environment (Accessed 17 December 2022).

Özöğür-Akyüz, S., Windeatt, T. and Smith, R. 2015. Pruning of Error Correcting Output Codes by optimization of accuracy–diversity trade off. *Machine Learning*, 101: 253-269.

Padhma, M. 2023. *A Comprehensive Introduction to Evaluating Regression Models*. Available: https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/ (Accessed 22 November 2023).

Page, M. J. and Moher, D. 2017. Evaluations of the uptake and impact of the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) Statement and extensions: a scoping review. Systematic reviews, 6 (1): 263.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P. and Moher, D. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery*, 88: 105906.

Parola, A., Di Fuccio, R., Marcionetti, J. and Limone, P. 2023. Digital games for career guidance: a systematic review using PRISMA guidelines. *Behaviour & Information Technology*: 1-11.

Parpoula, C. and Karagrigoriou, A. 2022. On optimal segmentation and parameter tuning for multiple change-point detection and inference. *Journal of Statistical Computation and Simulation*, 92: 3789 - 3816.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. and Dubourg, V. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12: 2825-2830.

Pérez, J. Q., Daradoumis, T. and Puig, J. M. M. 2020. Rediscovering the use of chatbots in education: A systematic literature review. *Computer Applications in Engineering Education*, 28 (6): 1549-1565.

Plevris, V., Solorzano, G., Bakas, N. P. and Ben Seghier, M. E. A. 2022. Investigation of performance metrics in regression analysis and machine learning-based prediction models. In: Proceedings of *8th European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS Congress 2022)*.

Pretorius, I., Piketh, S. and Burger, R. 2015. The impact of the South African energy crisis on emissions. *WIT Trans. Ecol. Environ*, 198: 255-264.

Qiao, W., Lu, H., Zhou, G., Azimi, M., Yang, Q. and Tian, W. 2020. A hybrid algorithm for carbon dioxide emissions forecasting based on improved lion swarm optimizer. *Journal of Cleaner production*, 244: 118612.

Qu, H., Suphachalasai, S., Thube, S. D. and Walker, S. 2023. South Africa Carbon Pricing and Climate Mitigation Policy. *Selected Issues Papers*, 2023 (040).

Quarato, M., De Maria, L., Gatti, M. F., Caputi, A., Mansi, F., Lorusso, P., Birtolo, F. and Vimercati, L. 2017. Air Pollution and Public Health: A PRISMA-Compliant Systematic Review. *Atmosphere*, 8 (10): 183.

Rafindadi, A. A. and Ozturk, I. 2017. Dynamic effects of financial development, trade openness and economic growth on energy consumption: evidence from South Africa. *International Journal of Energy Economics and Policy*, 7 (3): 74-85.

Rahman, M. M. 2021. The dynamic nexus of energy consumption, international trade and economic growth in BRICS and ASEAN countries: A panel causality test. *Energy*, 229: 120679.

Rajagopalan, G. 2021. Prepping your data with pandas. *A Python Data Analyst's Toolkit: Learn Python and Python-based Libraries with Applications in Data Analysis and Statistics*: 147-241.

Rajkoomar, M., Marimuthu, F., Naicker, N. and Damascene Mvunabandi, J. 2022. A meta-analysis of the economic impact of carbon emissions in Africa. *Environmental Economics*, 13 (1): 89-100.

Raschka, S. 2018. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*.

Reich, P. B. 2010. The carbon dioxide exchange. *Science*, 329 (5993): 774-775.

Rena, R. and Msoni, M. 2014. Global financial crises and its impact on the South African economy: A further update. *Journal of Economics*, 5 (1): 17-25.

Rengachary, G. S., Chandran, H., Vijayan, N., Yadav, V. and Mishra, S. 2023. A machine learning assisted optical multistage interconnection network: Performance analysis and hardware demonstration. *ETRI Journal*, 45 (1): 60-74.

Ritchie, H., Roser, M. and Rosado, P. 2022. South Africa: Energy Country Profile. *Our World in Data*.

Rodriguez, J. D., Perez, A. and Lozano, J. A. 2009. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*, 32 (3): 569-575.

Rogelj, J., Shindell, D., Jiang, K., Fifita, S., Forster, P., Ginzburg, V., Handa, C., Kheshgi, H., Kobayashi, S. and Kriegler, E. 2018. Mitigation pathways compatible with 1.5 C in the context of sustainable development. In: *Global warming of 1.5 C*. Intergovernmental Panel on Climate Change, 93-174.

Rogerson, P. A. 2003. A cumulative sum approach to syndromic surveillance in geographic regions. *Journal of Urban Health*, 80 (1): 130.

Rong, S. and Bao-Wen, Z. 2018. The research of regression model in machine learning field. In: Proceedings of MATEC Web of Conferences. EDP Sciences, 01033.

Ross, A. and Willson, V. L. 2017. Multiple Regression with Two Continuous Predictors and the Interactions between Them: Data Centered. In: *Basic and Advanced Statistical Tests*. Brill, 75-86.

Ryu, J.-h., Wan, H. and Kim, S. 2010. Optimal Design of a CUSUM Chart for a Mean Shift of Unknown Size. *Journal of Quality Technology*, 42: 311 - 326.

Sadorsky, P. 2020. Energy related CO2 emissions before and after the financial crisis. *Sustainability*, 12 (9): 3867.

Saha, A. 2015. *Doing Math with Python: Use Programming to Explore Algebra, Statistics, Calculus, and More!* No Starch Press.

Sahoo, M. and Sahoo, J. 2022. Effects of renewable and non-renewable energy consumption on CO2 emissions in India: Empirical evidence from disaggregated data analysis. *Journal of Public Affairs*, 22 (1): e2307.

Saib, M. O., Rajkoomar, M., Naicker, N. and Olugbara, C. T. 2022. Digital pedagogies for librarians in higher education: a systematic review of the literature. *Information Discovery and Delivery*, 51 (1): 13-25.

Salahuddin, M., Gow, J., Ali, M. I., Hossain, M. R., Al-Azami, K. S., Akbar, D. and Gedikli, A. 2019. Urbanization-globalization-CO2 emissions nexus revisited: empirical evidence from South Africa. *Heliyon*, 5 (6): e01974.

Şalvarlı, Ş. İ. and Griffiths, M. D. 2021. Internet Gaming Disorder and Its Associated Personality Traits: A Systematic Review Using PRISMA Guidelines. *International Journal of Mental Health and Addiction*, 19 (5): 1420-1442.

Sampene, A., Li, C., Agyeman, F. and Brenya, R. 2021. Analysis of the BRICS countries' pathways towards a low-carbon environment. *BRICS Journal of Economics*, 2 (4): 77-102.

Sarker, I. H. 2021. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2 (3): 1-21.

Sarkis-Onofre, R., Catalá-López, F., Aromataris, E. and Lockwood, C. 2021. How to properly use the PRISMA Statement. *Systematic reviews*, 10 (1): 117.

Sayeed, S., Ahmad, A. F. and Peng, T. C. 2022. Smartic: a smart tool for big data analytics and iot. *F1000Research*, 11: 17.

Schleussner, C.-F., Ganti, G., Rogelj, J. and Gidden, M. J. 2022. An emission pathway classification reflecting the Paris Agreement climate objectives. *Communications Earth & Environment*, 3 (1): 135.

Schorfheide, F. and Wolpin, K. I. 2012. On the use of holdout samples for model selection. *American Economic Review*, 102 (3): 477-481.

Scimago Journal Rank. 2022. Scimago Journal & Country Rank. Available: https://www.scimagojr.com/journalrank.php (Accessed 01 March 2023).

Sen, P. C., Hajra, M. and Ghosh, M. 2020. Supervised classification algorithms in machine learning: A survey and review. In: Proceedings of *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018*. Springer, 99-111.

Seymore, R., Inglesi-Lotz, R. and Blignaut, J. 2014. A greenhouse gas emissions inventory for South Africa: A comparative analysis. *Renewable and Sustainable Energy Reviews*, 34: 371-379.

Sguazzin, A. 2023. *SA beats climate goal as load shedding slashes emissions* Available: https://www.moneyweb.co.za/news-fast-news/sa-beats-climate-goal-as-load-shedding-slashes-emissions/ (Accessed 17 December 2023).

Shahbaz, M., Bhattacharya, M. and Ahmed, K. 2017. CO2 emissions in Australia: economic and non-economic drivers in the long-run. *Applied Economics*, 49 (13): 1273-1286.

Shahhosseini, M., Hu, G. and Pham, H. 2022. Optimizing ensemble weights and hyperparameters of machine learning models for regression problems. *Machine Learning with Applications*, 7: 100251.

Shahraki, A., Abbasi, M. and Haugen, Ø. 2020. Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost. *Engineering Applications of Artificial Intelligence*, 94: 103770.

Shan, Y., Huang, Q., Guan, D. and Hubacek, K. 2020. China CO2 emission accounts 2016–2017. *Scientific Data*, 7 (1): 54.

Sharif, A., Raza, S. A., Ozturk, I. and Afshan, S. 2019. The dynamic relationship of renewable and nonrenewable energy consumption with carbon emission: A global study with the application of heterogeneous panel estimations. *Renewable Energy*, 133: 685-691.

Sharma, S. S. 2011. Determinants of carbon dioxide emissions: Empirical evidence from 69 countries. *Applied energy*, 88: 376-382.

Sharma, S., Swayne, D. A. and Obimbo, C. 2016. Trend analysis and change point techniques: a survey. *Energy, Ecology and Environment*, 1: 123-130.

Shatnawi, A., Alkassar, H. M., Al-Abdaly, N. M., Al-Hamdany, E. A., Bernardo, L. F. A. and Imran, H. 2022. Shear strength prediction of slender steel fiber reinforced concrete beams using a gradient boosting regression tree method. *Buildings*, 12 (5): 550.

Shen, L., Wu, Y., Lou, Y., Zeng, D., Shuai, C. and Song, X. 2018. What drives the carbon emission in the Chinese cities?—A case of pilot low carbon city of Beijing. *Journal of Cleaner production*, 174: 343-354.

Shuai, C., Chen, X., Wu, Y., Tan, Y., Zhang, Y. and Shen, L. 2018. Identifying the key impact factors of carbon emission in China: Results from a largely expanded pool of potential impact factors. *Journal of Cleaner production*, 175: 612-623.

Shuai, C., Shen, L., Jiao, L., Wu, Y. and Tan, Y. 2017. Identifying key impact factors on carbon emission: Evidences from panel and time-series data of 125 countries from 1990 to 2011. *Applied energy*, 187: 310-325.

Siegel, A. F. 2012. Chapter 5 - Variability: Dealing with Diversity. In: Siegel, A. F. ed. *Practical Business Statistics (Sixth Edition)*. Boston: Academic Press, 95-121. Available: https://www.sciencedirect.com/science/article/pii/B9780123852083000055 (Accessed 15 November 2023).

Simon, D. 2001. Trading spaces: imagining and positioning the 'new'South Africa within the regional and global economies. *International Affairs*, 77 (2): 377-406.

Simplilearn. 2023. What Is Python Polynomial Regression In Machine Learning? Available: https://www.simplilearn.com/what-is-python-polynomial-regression-in-machine-learning-article (Accessed 05 July 2024).

Singh, A. 2024. KNN algorithm: Introduction to K-Nearest Neighbors Algorithm for Regression. Available: https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/ (Accessed 07 July 2024).

Singh, R., Chauhan, P., Sawan, N. and Smarandache, F. 2009. Ratio Estimators in Simple Random Sampling Using Information on Auxiliary Attribute. *Pakistan Journal of Statistics and Operation Research*, 4: 47-53.

Siriwardana, M. and Dollery, B. 2002. The impact of the Asian economic crisis on Southern African economies. *African Development Review*, 14 (2): 276-297.

Smith, D. M. 2023. *Living under apartheid: aspects of urbanization and social change in South Africa*. Taylor & Francis.

Song, Y., Liang, J., Lu, J. and Zhao, X. 2017. An efficient instance selection algorithm for k nearest neighbor regression. Neurocomputing, 251: 26-34.

Sorensen, P. 2011. Mining in South Africa: a mature industry? *International journal of Environmental studies*, 68 (5): 625-649.

South African Government. 2002. *ENERGY OUTLOOK FOR SOUTH AFRICA: 2002*. Available: https://www.gov.za/sites/default/files/gcis_document/201409/dmeenergyoutlook230120040.pdf (Accessed 30 November 2023).

Spalding-Fecher, R., Williams, A. and van Horen, C. 2000. Energy and environment in South Africa: charting a course to sustainability. *Energy for sustainable development*, 4 (4): 8-17.

Sperber, A. D. 2004. Translation and validation of study instruments for cross-cultural research. *Gastroenterology*, 126: S124-S128.

Sravani, B. and Bala, M. M. 2020. Prediction of Student Performance Using Linear Regression. In: Proceedings of 2020 International Conference for Emerging Technology (INCET). 5-7 June 2020. 1-5.

Stančin, I. and Jović, A. 2019. An overview and comparison of free Python libraries for data mining and big data analysis. In: Proceedings of *2019 42nd International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 977-982.

Statistics South Africa. 2023. Economic growth muted as 2023 draws to a close. Available: https://www.statssa.gov.za/?p=17053 (Accessed 26 December 2023).

Steenkamp, H. 2016. The influence of load shedding on the productivity of hotel staff in Cape Town, South Africa. In: *The influence of load shedding on the productivity of hotel staff in Cape Town, South Africa: Steenkamp, Henriette*.

Stoddard, I., Anderson, K., Capstick, S., Carton, W., Depledge, J., Facer, K., Gough, C., Hache, F., Hoolohan, C. and Hultman, M. 2021. Three decades of climate mitigation: why haven't we bent the global emissions curve? *Annual Review of Environment and Resources*, 46: 653-689.

Strahler Rivero, T., Herrera Nuñez, L. M., Uehara Pires, E. and Amodeo Bueno, O. F. 2015. ADHD Rehabilitation through Video Gaming: A Systematic Review Using PRISMA Guidelines of the Current Findings and the Associated Risk of Bias. *Frontiers in Psychiatry*, 6: 151.

Swaminathan, B. and Muraligopal, S. 2017. Climate Change and Emerging Agriculture Complexities. *Advances in Plants and Agriculture Research*, 6 (3): 59-60.

Tang, K. S., Cheng, D. L., Mi, E. and Greenberg, P. B. 2020. Augmented reality in medical education: a systematic review. *Can Med Educ J*, 11 (1): e81-e96.

Tartakovsky, A. G. and Moustakides, G. V. 2010. State-of-the-Art in Bayesian Changepoint Detection. *Sequential Analysis*, 29 (2): 125-145.

Tartakovsky, A. G., Rozovskii, B. L., Blazek, R. B. and Kim, H. 2006. A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods. *IEEE transactions on signal processing*, 54 (9): 3372-3382.

Tartakovsky, A., Nikiforov, I. and Basseville, M. 2014. *Sequential analysis: Hypothesis testing and changepoint detection*. CRC press.

Tatachar, A. V. 2021. Comparative Assessment of Regression Models Based On Model Evaluation Metrics. *International Journal of Innovative Technology and Exploring Engineering*, 8 (9): 853-860.

Taylor Enterprises. n.d. *Change-Point Analyzer*. Available: https://variation.com/product/change-point-analyzer/#tab-description (Accessed 15 March 2023).

Taylor, W. A. 2000. *Change-point analysis: a powerful new tool for detecting changes*.

Taylor, W. A. 2018. Change-point analysis: a powerful new tool for detecting changes. *Taylor Enterprises*: 1-19.

The World Bank. 2012. *World development indicators 2012*. The World Bank.

The World Bank. 2018. *Sustainable Energy for all*. Available: https://databank.worldbank.org/source/sustainable-energy-for-all (Accessed 02 August 2023).

The World Bank. 2022. *Country Climate and Development Report*. Available: https://databank.worldbank.org/source/country-climate-and-development-report-(ccdr) (Accessed 03 August 2023).

The World Bank. 2023. *World Development Indicators*. Available: https://databank.worldbank.org/source/world-development-indicators (Accessed 06 August 2023).

The World Bank. n.d. *About the World Bank*. Available: https://www.worldbank.org/en/about (Accessed 03 July 2023).

Thompson, A. 2023. *Load Shedding: What Is It and Why Is It Affecting South Africa?* Available: https://theculturetrip.com/africa/south-africa/articles/load-shedding-what-it-is-and-why-is-it-affecting-south-africa (Accessed 02 December 2023).

Tian, W., Song, J., Li, Z. and de Wilde, P. 2014. Bootstrap techniques for sensitivity analysis and model selection in building thermal performance analysis. *Applied energy*, 135: 320-328.

Torchio, M. F., Lucia, U. and Grisolia, G. 2020. Economic and Human Features for Energy and Environmental Indicators: A Tool to Assess Countries' Progress towards Sustainability. *Sustainability*, 12 (22): 9716.

Tsirikoglou, P., Abraham, S., Contino, F., Lacor, C. and Ghorbaniasl, G. 2017. A hyperparameters selection technique for support vector regression models. *Applied Soft Computing*, 61: 139-148.

Tunde, O. L., Adewole, O. O., Alobid, M., Szűcs, I. and Kassouri, Y. 2022. Sources and Sectoral Trend Analysis of CO2 Emissions Data in Nigeria Using a Modified Mann-Kendall and Change Point Detection Approaches. *Energies*, 15 (3): 766.

Turok, I. 2012. *Urbanisation and development in South Africa: Economic imperatives, spatial distortions and strategic responses*. Human Settlements Group, International Institute for Environment and Development.

Tutmez, B. 2006. Trend analysis for the projection of energy-related carbon dioxide emissions. *Energy exploration & exploitation*, 24 (1-2): 139-149.

U.S. Energy Information Administration. 2021. *Greenhouse gases and the climate*. Available: https://www.eia.gov/energyexplained/energy-and-the-environment/greenhouse-gases-and-the-climate.php (Accessed 15 February 2022).

United Nations Development Programme. 2021. *South Africa - UNDP Climate Promise*. Available: https://climatepromise.undp.org/what-we-do/where-we-work/south-africa (Accessed 25 December 2023).

Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F. and Mueller, A. 2015. Scikit-learn: Machine learning without learning the machinery. *GetMobile: Mobile Computing and Communications*, 19 (1): 29-33.

Vélez-Henao, J.-A., Vivanco, D. F. and Hernández-Riveros, J.-A. 2019. Technological change and the rebound effect in the STIRPAT model: A critical view. *Energy Policy*, 129: 1372-1381.

Vidyullatha, P., Rao, D. R., Prasanth, Y., Changala, R. and Narayana, L. 2016. Integrating Different Machine Learning Techniques for Assessment and Forecasting of Data. In: Proceedings of *Emerging Research in Computing, Information, Communication and Applications: ERCICA 2015, Volume 2*. Springer, 123-130.

Wachs, S. 2010. *What is a CUSUM chart and when should i use one*.

Wachs, S. 2022. *What is a CUSUM Chart and When Should I Use One?* Available: https://accendoreliability.com/cusum-chart-use-one/ (Accessed 15 June 2023).

Walsh, K., Theron, R. and Reeders, C. 2021. Estimating the economic cost of load shedding in South Africa. In: Proceedings of *Paper submission to Biennial Conference of the Economic Society of South Africa (ESSA)*.

Wang, L., Gong, Z., Gao, G. and Wang, C. 2017. Can energy policies affect the cycle of carbon emissions? Case study on the energy consumption of industrial terminals in Shanghai, Jiangsu and Zhejiang. *Ecological Indicators*, 83: 1-12.

Wang, Q., Alexander, W., Pegg, J., Qu, H. and Chen, M. 2020. HypoML: Visual analysis for hypothesis-based evaluation of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 27 (2): 1417-1426.

Wartner, S., Wiesinger-Widi, M., Girardi, D., Furthner, D. and Schmitt, K. 2019. Semi-automated Quality Assurance for Domain-Expert-Driven Data Exploration–An Application to Principal Component Analysis. In: Proceedings of *Machine Learning and Knowledge Extraction: Third IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2019, Canterbury, UK, August 26–29, 2019, Proceedings 3*. Springer, 128-146.

White, I. R., Royston, P. and Wood, A. M. 2011. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 30 (4): 377-399.

Winkler, H. and Marquand, A. 2009. Changing development paths: From an energy-intensive to low-carbon economy in South Africa. *Climate and development*, 1 (1): 47-65.

Wu, L., Liu, S., Liu, D., Fang, Z. and Xu, H. 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.

Wu, R., Wang, J., Wang, S. and Feng, K. 2021. The drivers of declining CO2 emissions trends in developed nations using an extended STIRPAT model: A historical and prospective analysis. *Renewable and Sustainable Energy Reviews*, 149: 111328.

Wu, W., Ma, X., Zhang, Y., Li, W. and Wang, Y. 2020. A novel conformable fractional non-homogeneous grey model for forecasting carbon dioxide emissions of BRICS countries. *Science of The Total Environment*, 707: 135447.

Xia, M., Dong, L., Jiang, L. and Zeng, M. 2024. Research on time series change point detection and influencing factors under machine learning: based on PM2.5 concentration data in Hefei city. Earth Science Informatics, 17 (1): 351-364.

Xia, T., Krishna, R., Chen, J., Mathew, G., Shen, X. and Menzies, T. 2018. *Hyperparameter Optimization for Effort Estimation. CoRR abs/1805.00336 (2018)*.

Xie, H., Li, D. and Xiong, L. 2014. Exploring the ability of the Pettitt method for detecting change point by Monte Carlo simulation. *Stochastic Environmental Research and Risk Assessment*, 28: 1643-1655.

Xu, Q., Mei, Y. and Moustakides, G. V. 2021. Optimum Multi-Stream Sequential Change-Point Detection With Sampling Control. *IEEE Transactions on Information Theory*, 67 (11): 7627-7636.

Yan, J., Liu, X., Shi, L., Li, C. and Zha, H. 2018. Improving Maximum Likelihood Estimation of Temporal Point Process via Discriminative and Adversarial Learning. In: Proceedings of *International Joint Conference on Artificial Intelligence*.

Yasin, Y. M., Kerr, M. S., Wong, C. A. and Bélanger, C. H. 2020. Factors affecting nurses' job satisfaction in rural and urban acute care settings: A PRISMA systematic review. *Journal of Advanced Nursing*, 76 (4): 963-979.

Yılmaz, E. and Şensoy, F. 2022. Effects of fossil fuel usage in electricity production on CO2Emissions: A STIRPAT model application on 20 selected countries. *International Journal of Energy Economics and Policy*: 224-229.

Yilu, W. 2022. Linear regression in machine learning. In: Proceedings of International Conference on Statistics, Applied Mathematics, and Computing Science. 121634T. Available: https://doi.org/10.1117/12.2628053 (Accessed 03 July 2024).

Yuping, L., Ramzan, M., Xincheng, L., Murshed, M., Awosusi, A. A., Bah, S. I. and Adebayo, T. S. 2021. Determinants of carbon emissions in Argentina: The roles of renewable energy consumption and globalization. *Energy Reports*, 7: 4747-4760.

Zhang, X., Chen, Y., Jiang, P., Liu, L., Xu, X. and Xu, Y. 2020. Sectoral peak CO2 emission measurements and a long-term alternative CO2 mitigation roadmap: A case study of Yunnan, China. *Journal of Cleaner production*, 247: 119171.

Zhu, L., Duan, H.-B. and Fan, Y. 2015. CO2 mitigation potential of CCS in China–an evaluation based on an integrated assessment model. *Journal of Cleaner production*, 103: 934-947.

Ziervogel, G., New, M., Archer van Garderen, E., Midgley, G., Taylor, A., Hamann, R., Stuart-Hill, S., Myers, J. and Warburton, M. 2014. Climate change impacts and adaptation in South Africa. *Wiley Interdisciplinary Reviews: Climate Change*, 5 (5): 605-620.

Annexure A: Language Editing Certificate

Masoodah Mohamed
Published poet in:
Yesterday's and Imagined Realities. A South African Anthology

Certificate of Language Editing

Author: Ghulam Masudh Mohamed
Qualification: Master of Information and Communication Technology
Academic supervisors: Dr S.S. Patel, Prof N. Naicker
Institution: Durban University of Technology

Title of Dissertation: Data Mining and Machine Learning: A Study of the CO2 Emission Trends in South Africa

This serves to certify that the above dissertation has undergone thorough proofreading to ensure correctness in spelling, grammar and punctuation. Additionally, assistance was provided to verify that the citations and reference list adhere to the correct referencing style. The document was sent back to the author to review the comments and address the suggested corrections at their discretion. The feedback process was conducted via email. Please note that the final revised version was not subjected to proofreading.

[redacted]

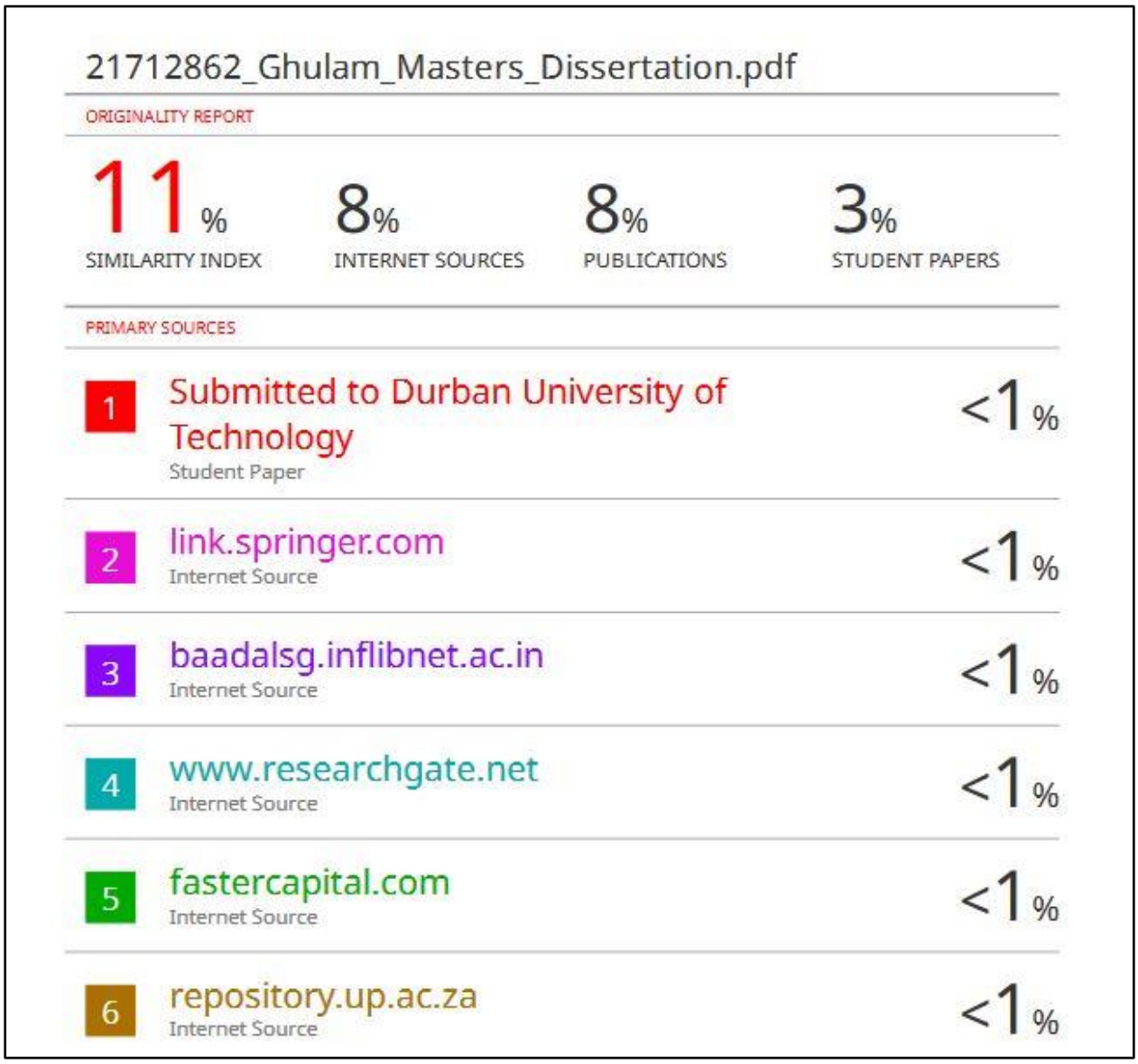
Ms Masoodah Mohamed

Address: Ga-Rankuwa, Pretoria Email: masoodahbegum@gmail.com

Cell: (+27) 81 419 8781

Entry-level member: Psychological Society of South Africa
Qualifications: BA Speech and Hearing Therapy (Wits), BA Honours in Psychology (UNISA)

Annexure B: Cover of Turnitin Report



21712862_Ghulam_Masters_Dissertation.pdf

ORIGINALITY REPORT

11%	8%	8%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Durban University of Technology Student Paper	<1%
2	link.springer.com Internet Source	<1%
3	baadalsg.inflibnet.ac.in Internet Source	<1%
4	www.researchgate.net Internet Source	<1%
5	fastercapital.com Internet Source	<1%
6	repository.up.ac.za Internet Source	<1%