

Predictive Analytics of CO2 Emission from Agri-Food Activities Using Machine Learning



SANDESH RAGASHETTI - 20014943

**Applied Research Project submitted in partial fulfillment of the requirements
for the degree of M.Sc. in Information Systems with Computing
at Dublin Business School**

Supervisor: Professor Swati Dongre

August 2024

DECLARATION

I declare that this Applied Research Project that I have submitted to Dublin Business School for the award of MSc. Information Systems with Computing is the result of my own investigations, except where otherwise stated, where it is clearly acknowledged by references. Furthermore, this work has not been submitted for any other degree.

Signed: Sandesh Ragashetti

Student Number: 20014943

Date: 26/08/2024

ACKNOWLEDGMENT

I take this opportunity to express my sincere gratitude to my supervisor **Prof. Swati Dongre** for her great support and guidance in completing my Applied Research Project. Her valuable input, instructions, and encouragement towards this project have been significant, which helped me to finish the project smoothly and properly within the time frame.

Furthermore, I would like to stretch my hand and express gratitude and appreciation to the Data Analytics and Visualisation lecturer **Prof. Shazia Afzal**, who taught me the module with all concepts from the beginning with all guidance. Also, my thanks to the **Dublin Business School Library** for providing the necessary resources for my applied research project which helps me to refer to required content.

Lastly, I would like to share my love and gratitude to my family and friends for their unconditional support in this project which helped to keep my confidence and hopes always high.

Table Of Contents

DECLARATION	2
ACKNOWLEDGMENT	3
ABSTRACT	8
1. INTRODUCTION	9
1.1. Background	9
1.2. Motivation	10
1.3. Problem Statement	11
1.4. Research Questions	11
1.5. Research Objectives	12
1.6. Research Roadmap	13
2. LITERATURE REVIEW	14
2.1. Introduction	14
2.2. CO2 Emission	14
2.3. Emissions from the agri-food sector	14
2.4. Impact of CO2 emissions in the world	15
2.5. Evolution of CO2	15
2.6. Predictive analysis for agri-food CO2 emissions	16
2.7. Previous approaches for CO2 emission prediction	17
2.8. Conclusion	21
2.9. Research Gap	21
2.10. Hypothesis	22
3. RESEARCH METHODOLOGY	23
3.1. Research Approach and Design	23
3.2. Tools Used	26
3.3. Data Collection And Preparation	26
3.3.1. Data Collection	26
3.3.2. Data Preparation	29
3.4. Exploratory Data Analysis	34
3.5. Modeling	35
3.5.1. Model selection	35
3.6. Ethical Consideration	41
3.7. Conclusion	42

4. RESULTS	43
4.1. Exploratory Data Analysis	43
4.1.1. Descriptive Statistics	43
4.1.2. Data Visualizations	45
4.2. Model Evaluation	51
4.2.1. Comparative Analysis	54
4.2.2. Hypothesis Testing	56
4.2.3. Regression Coefficients Significance	57
4.2.4. Actual Vs Predict Values for Linear Regression Model	58
5. CONCLUSION, LIMITATIONS AND FUTURE WORK	60
5.1. Conclusion	60
5.2. Limitations	61
5.3. Future Work	62
6. REFERENCES	64

List of Figures

Figure 1.1: Research Roadmap.....	13
Figure 3.1: Workflow of CRISP-DM Design.....	24
Figure 3.2: Code for Selection of Relevant Columns.....	30
Figure 3.3: Code for Data Aggregation.....	31
Figure 3.4: Code for Filling Missing Values.....	32
Figure 3.5: Code for Data Splitting for Training and Testing.....	32
Figure 3.6: Code for Data Transformation.....	33
Figure 3.7: Linear Regression Model Implementation.....	37
Figure 3.8: Decision Tree Implementation.....	38
Figure 3.9: Random Forest Implementation.....	39
Figure 3.10: Neural Network model function.....	40
Figure 3.11: Neural Network model Implementation.....	40
Figure 4.1: Data Describe.....	43
Figure 4.2: Displaying First 5 Rows.....	44
Figure 4.3: Data Information about Null Values and Data Type.....	44
Figure 4.4: Graph Plot for Total Emission and Average Temperature Rise Over Time.....	45
Figure 4.5: Graph Plot for Emission Breakdown by Category Over Time.....	46
Figure 4.6: Correlation Matrix.....	47
Figure 4.7: Graph Plot for Top 20 Countries of CO2 Emission.....	48
Figure 4.8: Graph Plot for Mean CO2 emission of the agri-food activities.....	49
Figure 4.9: Bar Plot for Total Emission for Each Year.....	50
Figure 4.10: CO2 Emission Comparative Analysis between models metrics.....	54
Figure 4.11: Temperature Rise Comparative Analysis Between Models Metrics.....	55
Figure 4.12: Code for CO2 Emission t-test.....	56
Figure 4.13: Code for Temperature t-test.....	56
Figure 4.14: Code to Check Each Feature's Coefficient In Linear Regression.....	57
Figure 4.15: Actual vs Predict value line of CO2 emission for Linear Regression.....	58
Figure 4.16: Actual vs Predict value line of temperature rise for Linear Regression.....	58

List of Tables

Table 4.1: CO2 Emission: Model Performance Metrics..... 52

Table 4.2: Temperature Rise: Model Performance Metrics..... 53

ABSTRACT

Global warming, Climate change, and Human health are getting impacted due to excessive agri-food emissions. Hence, the predictive analysis of CO₂ emissions from agri-food activities is important for policymakers and researchers to develop strategies for sustainable agricultural practices. This study collected and explored secondary historical data on agri-food CO₂ emissions in various countries around the world for a time span of 30 years (1990–2020) with machine learning techniques. Since previous research studies left a gap in predicting emissions from the agri-food sector and corresponding temperature rise, this project explores this area by implementing the four predictive models Linear Regression, Decision Trees, Random Forests, and Neural Networks. As a result, exploratory data analysis helps to understand the descriptive statistics, and data visualizations on agri-food activities, emissions, temperature rise, and their relationships. The four predictive models are trained and measured with metrics like MSE, RMSE, MAE, and R-squared. The Linear Regression model emerged as the best model with the highest predictive accuracy, with the lowest RMSE ($1.55e-11$), MAE ($8.37e-12$), and highest R²-score (1.00) for CO₂ emissions. The study concludes that Linear Regression can serve as a robust tool in predicting CO₂ emissions from agri-food activities and helps the policymakers, government bodies, and sustainable environment by providing useful insights and strategies to reduce the environmental impact of agriculture.

Keywords: CO₂ emissions, Predictive models, Agri-Food Activities, Linear Regression, Performance Metrics

1. INTRODUCTION

1.1. Background

Agri-Food is the sector that involves all activities from agriculture to delivery of foods, which covers food production, processing, transforming, distributing and consuming food. This sector also involves cultivation, animal breeding, grain and livestock farming processes, and distribution such as packaging and transportation until marketing at retail. It essentially revolves around the food supply chain which addresses both agricultural practices and the food industry's role in feeding the population (El Bilali et al., 2021).

The agri-food field is one of the main contributors to the global CO₂ emission which is responsible for releasing large amounts of green gasses into the atmosphere. There are many sources which are causing these emissions such as rice cultivation, production of pesticides and fertilizers, waste disposal and manure (Mangla et al., 2018). Due to globalization, the need for food is also increasing and environmental impact is getting worse which indirectly impacts on the farming methods (Zheng et al., 2019).

Clean and healthier environment is important for human life and growth. Pollution adversely affects the people, animals and plants which also destroys the natural cycle. The fossil fuels which are used in farming releases CO₂ emissions which causes climate change and global warming. According to the World Resources Institute, 6 billion tons of greenhouse gas emissions were produced in 2014. It is reported that 12.5% of total CO₂ emissions are produced by agriculture (Serafeim and Caicedo, 2022). According to the Food and Agriculture Organisation (FAO) of the United Nations, 16% emissions were increased from the agriculture and food

systems globally in the year 1990-2020. In 2020, these emissions increased to 17 billion tones and accounted for 30% of the global emission (Chowdhury et al.,2021).

1.2. Motivation

Nowadays, Global CO₂ emission has become a big environmental issue especially in the field of agriculture which contributes to greenhouse gas emissions significantly. Strict policies have to be introduced by the policy makers and government bodies and also they need to conduct research in the agriculture sector through their climate policy efforts which is aimed at reducing methane as well as CO₂ emissions(Singh, P.K et al.,2021).

Historically, prediction and reduction of CO₂ emission have depended on the traditional statistical models and methods, but these models are not consistent to capture the complicated, non-linear relationships in agricultural areas. For example, models like ARIMA and linear regression cannot adapt to changing conditions(Fatima et al., 2019). This problem has led to inventing more advanced models and tools that can handle more complex data with more accurate results. Machine learning is coming with new opportunities for predicting CO₂ emissions accurately due to recent advancements. Machine Learning models can analyze large datasets, identify patterns and predict future emissions with greater accuracy (Serafeim and Caicedo, 2022). By understanding the use of data and advanced algorithms, policymakers, researchers, and industry stakeholders can implement more effective strategies to prevent CO₂ emissions.

1.3. Problem Statement

Despite the advancement of machine learning in the modern era, there is a significant gap in comprehensive prediction of CO₂ emissions particularly from the agri-food field. Existing research and studies focused only on general CO₂ emission prediction from all the area or particular field, with minimum attention to the unique features and challenges arriving from the agriculture sector. Furthermore, predictive analysis of global average temperature rise due to CO₂ emission released from the agricultural field is unexplored, leaving a crucial gap in identifying the correlation between the agricultural activities, temperature and CO₂ emissions.

1.4. Research Questions

1. How machine learning models are effective in predicting CO₂ emissions and temperature rise from the different countries of the world based on the historical data of agri-food emissions?
2. How does the performance of the selected machine learning models are measured and compared through the metrics like RMSE, MSE, MAE and R²? And Which model performs best in predicting CO₂ emission and temperature rise?
3. Which are the top countries contributing more to the global emission and temperature rise from the agri-food area? And Which agri-food activities emit these gasses the most?
4. What is the trend observed between CO₂ emission and temperature rise in agri-food areas over the 30 years?

1.5. Research Objectives

- To collect and analyze the historical data of agri-food CO₂ emissions that includes various agri-food activities like crop residues, rice cultivation, agrifood systems waste disposal, organic soil emission, food supply chain, agrochemical manufacturing and emission from Manure.
- To conduct Exploratory data analysis (EDA) on collected data and show descriptive statistics, different visualization graph plots to understand the structure, trends, and CO₂ emission rate of agri-food activities.
- To measure the effectiveness and accuracy of machine learning models in forecasting CO₂ emissions and temperature rise based on historical data from different countries over the years.
- To conclude the best performing model among four different models (Linear regression, Random Forest, Decision Tree and Neural Networks) by calculating and comparing the metrics like RMSE, MSE, MAE and R².

1.6. Research Roadmap

Below attached picture shows the roadmap for this research project:

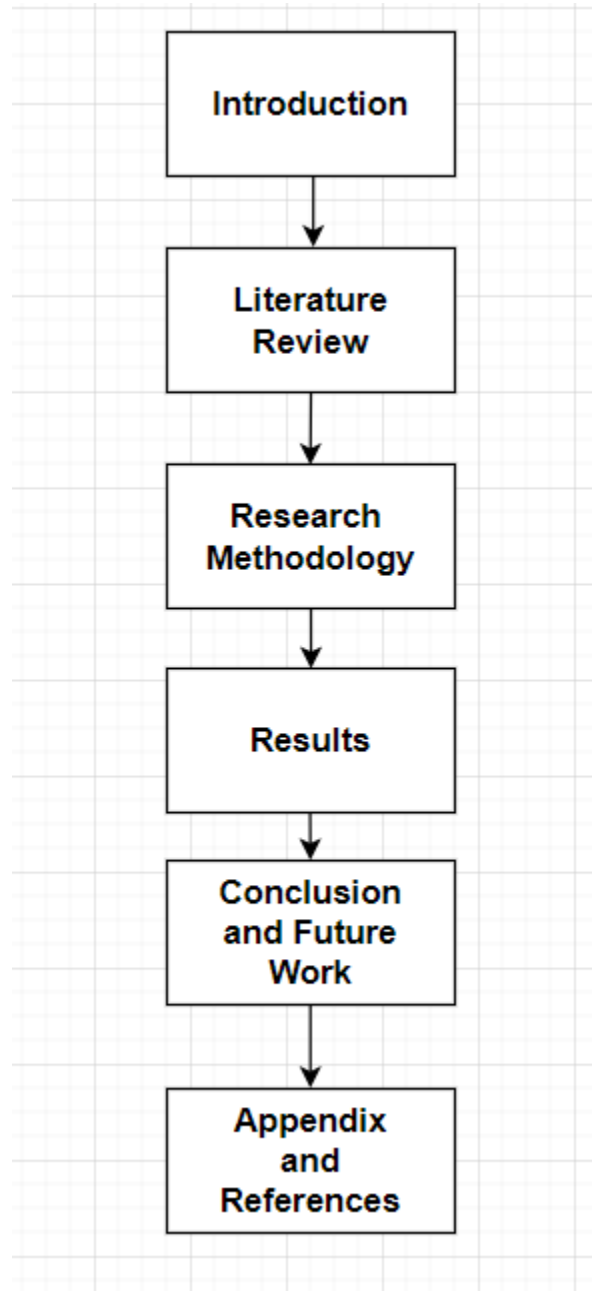


Figure 1.1: Research Roadmap

2. LITERATURE REVIEW

2.1. Introduction

The literature review explains about the CO₂ emission, emission of CO₂ from the agri-food, impact of CO₂ emission in the world, evolution of CO₂ emission and predictive analysis for agri-food CO₂ emissions. After that, several researches conducted by the people in the past are explained. Approach to the different emission problems, key findings from their researches, machine learning models they used are clearly detailed in this section. Finally, the literature study is concluded in identifying the research gap and hypothesis.

2.2. CO₂ Emission

CO₂ emission is the type of an emission which releases carbon dioxide gas into the atmosphere. This gas results from different actions like fossil fuel combustion, burning coal, oil, industrial processes, agricultural activities, cutting trees and many more. CO₂ is a greenhouse gas that heats the earth's atmosphere, which plays a key role in global warming and climate change (Yaacob et al., 2020). Moreover, the excessive release of CO₂ gas leads to an increase in global temperature and results in variations of climate conditions along with natural habitats (Jayakrishnan et al., 2022).

2.3. Emissions from the agri-food sector

Various agriculture activities such as soil management, pesticides and fertilizer production, livestock farming, crop cultivation, manure management etc are causing the release

of CO₂ emission. These factors are not only the reason for CO₂ emission, but also releases gasses like nitrous oxide (N₂O) and methane (CH₄). For example, crop cultivation in 'stand water' produces a substantial amount of CO₂ and methane due to anaerobic decomposition of organic matter (Hosseini et al., 2019). Nitrogen based fertilizers on the soils example where nitrogen results in release of CO₂ emission through nitrification and denitrification from soil (Mangla et al., 2018).

2.4. Impact of CO₂ emissions in the world

CO₂ emissions have shown its effect on the environment through driving climate change, global warming, and environmental damage. Higher emission of these hazardous gasses makes an impact on global temperatures, melting icebergs and causes to the rise of hurricanes, heatwaves and floods that pose environmental risks and danger for human lives (Sarfraz et al., 2023; Serafeim and Velez Caicedo, 2022). CO₂ emission has some serious effects on both the environment and human health. A rise of CO₂ in the oceans is responsible for acidification, which damages marine life and ecosystems leading to decrease biodiversity (Talukder et al., 2022). Increased levels of CO₂ also result in poor air, and create health risks linked to respiratory and heart diseases (Arora et al., 2018). Changes in temperature and precipitation patterns impact farming which leads to food shortages and economic losses (Serafeim and Caicedo, 2022).

2.5. Evolution of CO₂

The toxic emissions have been closely linked to industrialization and economic development. The combustion of fossil fuels such as coal, oil, and natural gas has led to an increase in emissions of gasses, since the 18th century Industrial Revolution (Millot and Maizi,

2021). The post World War II economic change accelerated the Industrialization, which caused the developed countries to contribute more towards the CO₂ level. The 20th century saw a rapid rise in atmospheric pollution and emissions due to the expansion of industrial production, urbanization, and the increase of motor vehicles (Zheng et al., 2019). Despite environmental impact awareness, global CO₂ emissions have continued to rise, motivated by economic growth of countries and energy demands (Srivastav, 2019). The evolution of CO₂ emissions is closely linked to human industrial and economic activities (Zheng et al., 2019). Despite international efforts to reduce emissions through projects like the Kyoto Protocol and the Paris Agreement, global CO₂ emission is not decreasing effectively to meet climate targets (Fatima et al., 2019). This ongoing rise emphasizes the challenges on economic development balance along with environmental sustainability.

2.6. Predictive analysis for agri-food CO₂ emissions

Predictive analysis for CO₂ emissions using machine learning has emerged as a powerful technique for forecasting future conditions and developing strategies. Models like ARIMA have been used extensively, however researchers were not able to capture the complex, non-linear relationships in data of emissions (Fatima et al., 2019). Further, Models like Support Vector Machines (SVMs), Random Forests (RFs), and Artificial Neural Networks (ANNs) have shown improved accuracy in results by preprocessing large datasets and identifying patterns that other traditional models might miss (Serafeim and Caicedo, 2022). One study has shown that the ANN model has achieved higher accuracy in predicting agricultural methane and CO₂ emissions (Hosseini et al., 2019). By combining traditional algorithms and advanced ML methods, GA-ELM (hybrid model) enhanced the prediction capabilities by optimizing methodologies

(Shabani et al., 2021). By integrating predictive models with real time data and global climate, models can enhance their effectiveness in managing and reducing CO₂ emissions from the agri-food sector more effectively (Chowdhury et al., 2021).

2.7. Previous approaches for CO₂ emission prediction

Ahmed S., Ahmed K., and Ismail M. (2020) explored environmental technology, energy use, and economic activity which affected CO₂ emissions in emerging economies. Environmental technology involved new methods to reduce environmental harm, and total energy consumed. Their study found that both environmental technology and energy use have a big impact on CO₂ emissions by using statistical methods to analyze the data. Finally, authors concluded that improving these areas could help lower emissions. This study showed other researchers that advancement in technology and optimizing energy use were main factors to reduce emissions. It helped to understand sustainable development which offered practical solutions for growing economies.

Bussaban K., Kularbphetong K., and Boonseng C. (2023) addressed CO₂ emissions through machine learning models, including regression and neural networks, for predicting future emissions. Model results showed high accuracy with an indication of hope that machine learning technology is optimistic when it comes to mitigating environmental issues. Their research presented the work in environmental science and practical ways of predicting emissions by using machine learning. It succeeded in using advanced predictive models and clear data analysis. However, the inherent weakness is within dependencies with past data, which may affect the accuracy of models under changed conditions.

Chowdhury S., Rubi M. A., and Bijoy Md. H. I. (2021) team conducted research on the model Artificial Neural Networks (ANN) in estimating methane and CO₂ emissions from agriculture in Bangladesh. This elaborated that ANNs are computational models and the gasses being released methane and CO₂ resulted from the greenhouse effect, which trapped heat in the atmosphere. ANN models were trained to predict emissions using historical data, and the model could predict the emissions that were efficient in environmental forecasting. While this method is reliable for the prediction of emissions and solving environmental problems, its complexity and the need for a big set of data during training make the ANN model weak.

Guo X., Yang J., Shen Y., and Zhang X. (2023) investigated the agricultural carbon emissions in China by implementing a hybrid GA-ELM model, which combines Genetic Algorithm (GA) for optimization and Extreme Learning Machine (ELM) for learning. Their study used the GA-ELM model on past emissions data and identified that it to be robust and more effective than other models. Though these researchers contributed by presenting an advanced hybrid model for emissions prediction, offering more accurate and efficient forecasting tools, its weaknesses are computational complexity and the resources needed to implement the hybrid model.

Hamrani A., Akbarzadeh A., and Madramootoo C. A. (2020) discovered the use of machine learning models to predict greenhouse gas emissions from agricultural soils. Key points included in this study were greenhouse gas emissions, gasses that trap heat in the atmosphere, and the use of algorithms to analyze these data and make predictions. The comparisons made by the authors were on the grounds of the models like regression and neural networks, with the grounds of predicting accuracy. It was observed that these models had enhanced prediction with efficiency. In this study, the authors derived comprehensive model comparison and practical

implications of different data. However, they found a requirement for large data and overfitting of the models to be the pain points.

Hosseini S. M., Saifoddin A., Shirmohammadi R., and Aslani A. (2019) used time series and regression models for analysis of CO2 emissions and addressed the challenge of forecasting it in Iran. Their study mainly includes time series analysis, the study of time ordered data points, and a statistical method for examining relationships between variables. Their study explained that models could effectively forecast CO2 emissions, providing prediction accuracy. This study offered specific insights into Iran's context and contributed by evaluating the use of these statistical methods.

Ma N., Shum W. Y., Han T., and Lai F. (2021) team studied the application of Gaussian Process Regression model to analyze and predict CO2 emissions. This study applied the GPR models to old emissions data and found that the accuracy of GPR models outperformed some traditional regression models. Their studies improved the understanding by presenting a more sophisticated statistical process and method for emission's prediction. The authors implication helped in the model's accuracy and robustness, but it could fail on computational complexity and requirement for enough resources.

Serafeim G., and Velez Caicedo G. (2022) team focused on prediction of Scope 3 carbon emissions using machine learning models. They developed regression and classification algorithms for the prediction of Scope 3 emissions. Study discovered that these models well predict Scope 3 emissions, hence an essential tool for organizations in the evaluation of their carbon footprints. This publication helped to align with other researches on AI applications in sustainability, contributing to the emissions management understanding. Though their study

includes innovative use of machine learning and practical relevance, it struggled to adapt to the complexity of the models and the need for large data.

With the use of the Inclusive Multiple Model (IMM) method, Shabani E., Hayati B., Pishbahar E., Ghorbani M. A., and Ghahremanzadeh M. (2021) team attempted to estimate CO₂ emissions in Iran's agriculture field. This technique was better than other individual models such as Multiple Regression (MLR), Gaussian Process Regression (GPR), and Artificial Neural Network (ANN). Their results explained the better accuracy of the IMM model in predicting CO₂ emissions, supporting the model's purpose in guiding air pollution reduction plans. Their research contributes to the literature by inventing a more robust and accurate method for emissions estimation. Study showed by validating the IMM model as a superior tool for environmental forecasting with the model's high accuracy and innovative approach. On the other hand, it had some complexity in implementation and arranging required resources.

Singh P. K., Pandey A. K., Ahuja S., and Kiran R. (2021) team deployed multiple forecasting approaches, to predict CO₂ emissions from paddy crops in India. They used ARIMA and regression models to predict future values, and analyzed the release of carbon dioxide from agricultural activities. Their study found that combining multiple models provided more accurate predictions than individual models and supported literature on emphasizing the benefits of model integration. Their research contributed by demonstrating the efficacy of combined prediction methods in agriculture and practical implications for policy-making, but they had some limitations with the process complexity and combining multiple models.

Zhu and Huo (2022) team analyzed the impact of agricultural production efficiency on the release of carbon dioxide during agricultural activities in China. Their study showed that

higher the production efficiency with lower carbon emissions. Their research aligned with other studies on sustainable agricultural practices highlighting the importance of reducing emissions. It showed the link between production efficiency and emissions reduction factors which guides for the clear implications of policy and practice, with specific limitations to data.

2.8. Conclusion

The major findings from the research papers showed that there was a continuous growth in the predictive analysis of CO₂ emissions through various models and methodologies from the different sectors and different regions. Many models like Artificial neural networks, GA-ELM, Gaussian Process Regression, ARIMA and regression models have shown high accuracy in forecasting CO₂ emissions. Also studies demonstrated that CO₂ emissions can be reduced globally with improvements in environmental technology and optimization of energy use.

2.9. Research Gap

Despite the above mentioned developments, several research gaps remain with respect to CO₂ emission prediction in the agriculture field. Most studies mainly focussed on the specific regions such as China, Bangladesh, Iran etc. And also there is a lack of comprehensive models that integrate the effects of CO₂ emissions on global temperature rise due to excessive emissions. Most of the existing studies have concentrated on emission prediction without addressing the subsequent impact on global climate change and temperature rise. These studies left a gap to develop integrated models that not only predict CO₂ emissions but also the global temperature rises from the different countries of the world.

2.10. Hypothesis

Based on the above studies and research gap, this project aims to carry out predictive analysis of CO₂ emission from the agri-food sector where various activities of agriculture are selected to understand the factors that affect most of the emission and cause temperature rise in the global. This study will identify the best performing model among four defined models, which includes the calculation of metrics and comparative analysis of these models. Exploratory data analysis will show clear visualization of affecting factors to the total emission and temperature rise around the globe from the past 30 years. Finally, predictive analyses of CO₂ emission and temperature rise will be done through the best performing model and conclude the outcomes of results which helps the policy makers, researchers, and industry stakeholders to take the actions in coming years against the CO₂ emission reduction and global climate change.

3. RESEARCH METHODOLOGY

3.1. Research Approach and Design

This Applied research project follows a quantitative approach where it involves the use of statistical, mathematical and machine learning techniques to analyze the numerical data and uncover patterns, relationships, and trends. For the Agri-food CO₂ emission, quantitative approach entails collecting and processing data on different agricultural activities. Then applying predictive modeling techniques such as Linear regression, Decision Tree, Random Forest and Neural Networks to the CO₂ emission and temperature rise prediction. Along with this some of the metrics are evaluated like MSE, RMSE, MAE and R². This approach allows for rigorous analysis, providing a data driven basis for decision making and policy development.

This Applied Research Project utilized the CRISP-DM (Cross Industry Standard Process for Data Mining) design or framework for the successive prediction of the CO₂ emission model which was originally proposed by the Shearer (Shearer, 2000). This design provides a structured and easy path for conducting CO₂ emission research and its prediction by using data mining steps. It is a structured design which ensures thoroughness, repeatability, and alignment with research objectives. This structured approach helps to meet all the objectives where its focus on thorough data analysis, effective modeling and evaluating metrics makes it an invaluable tool for this project.

The below diagram follows the different stages of the CRISP-DM framework, where CO₂ emission data undergoes all the phases, each with specific tasks and deliverables.

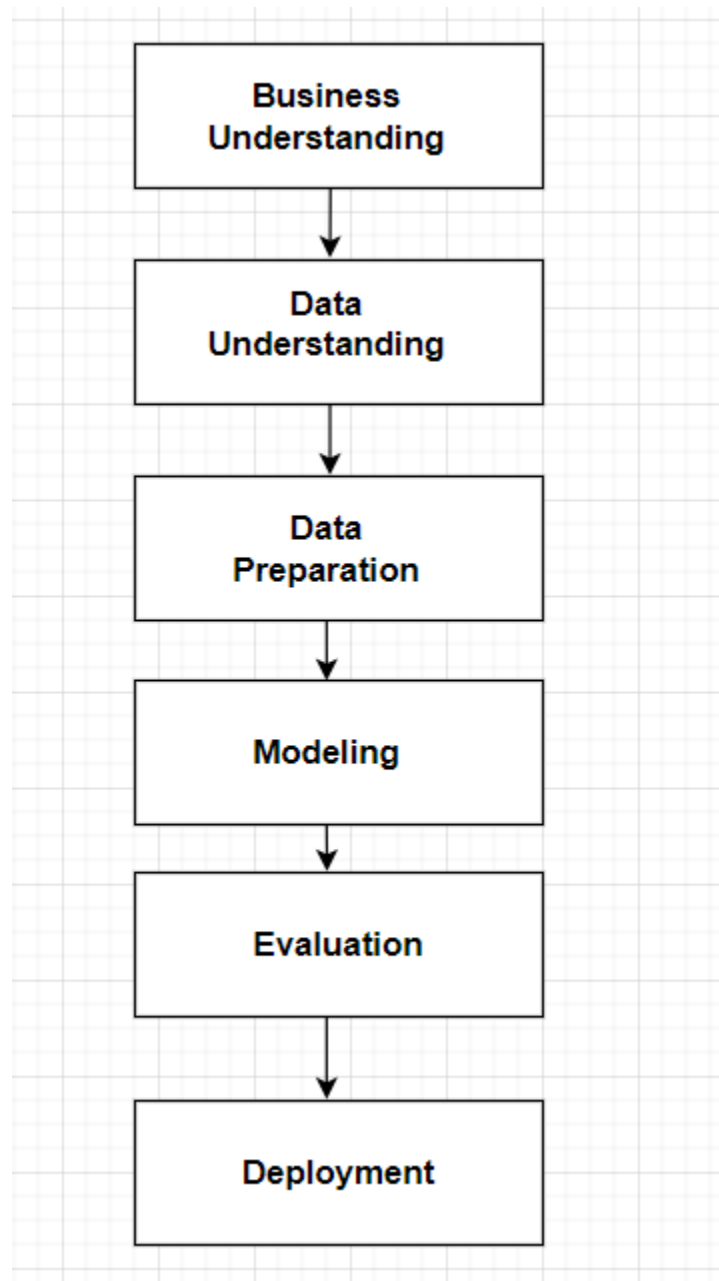


Figure 3.1: Workflow of CRISP-DM Design

1. Business Understanding: The first phase focuses on understanding the project problem statement, objectives and requirements of resources. Then converting this knowledge into a machine learning problem definition and a proper plan. Identifying the business objectives

defined for this project, assessing the data from the agri-food sector, defining analytics goals, and developing a well structured project plan for CO₂ emission prediction are the major steps of this stage.

2. Data Understanding: The second phase of CRISP-DM design involves collecting the existing CO₂ emission data required for analysis, understanding the data structure and type, identifying any potential problem with data quality, descriptive statistics and exploratory data analysis through visualization graphs. For this project, data from various agricultural activities contributing to CO₂ emissions are collected from sources such as the Food and Agriculture organization(FAO) and Intergovernmental Panel on Climate Change(IPCC).

3. Data Preparation: In this phase, conversion of raw data into useful form of data required for data analysis and modeling will be handled. This process includes data preprocessing, feature engineering, data aggregation, data transformation etc.

4. Modeling: The modeling phase involves selecting models and applying defined modeling techniques to them based on the problem and objectives. The processed data is splitted here into training and testing sets where models get trained with training sets. For this project, models such as Linear Regression, Random Forest, Decision Tree, and Neural network are used.

5. Evaluation: This phase involves thoroughly evaluating the selected models with the help of evaluation metrics. Finally, identifying the best model based on the performance metrics and reviewing those steps executed to ensure they properly achieve the project's objectives. For this project, the evaluation measures are MSE, RMSE, MAE, and R-squared to assess the accuracy of the models. Comparative analysis will be carried out in this step to check the evaluation results, and identify the best model among used models.

6. Deployment: Deployment stage includes demonstrating the output result of the evaluation and giving recommendations to policymakers or industry stakeholders. However, in real time this stage includes integration of the best-performing model into practical use where it can generate insights, make predictions and decisions.

3.2. Tools Used

Python: As part of this project, Python programming language is used to build the artifact. Python is used with google collab notebook which supports writing and running the code. Pandas, Numpy, Matplotlib, Seaborn, Scikit-learn and Statsmodels libraries are used in this project to access the modules required for the predictive analysis.

Machine Learning: Machine learning algorithms are used in this project to build and train predictive models for CO2 emissions and temperature.

Google Colab Notebook: This is the development environment for writing and running the python code. It has most of the pre-installed python libraries to support the development.

Draw.io: This tool is used for creating flowchart and diagrams required for this project.

3.3. Data Collection And Preparation

3.3.1. Data Collection

The dataset used in this project is collected from Kaggle, named as "Agri-food CO2 emission dataset - Forecasting ML" by author Alessandro Lobello. This secondary dataset provides detailed information on CO2 emissions from different agri-food activities across 236

different regions or countries around the world over 30 years from the year 1990 to 2020. Author constructed this data set by merging and reprocessing approximately a dozen individual datasets from the Food and Agriculture Organization (FAO) and data from IPCC. The data includes variables such as the country, year, different agri-food activities, and the respective CO2 emissions measured in kilotons and temperature measured in celsius. This dataset is curated by the author, making it easy for analyzing and predicting CO2 emissions in the agriculture sector.

Original data columns: Area, Year, Savanna fires, Forest fires, Crop Residues, Rice Cultivation, Drained Organic Soils (CO2), Pesticides Manufacturing, Food Transport, Forestland, Net Forest Conversion, Food Household Consumption, Food Retail, On-farm Electricity Use, Food Packaging, Agrifood Systems Waste Disposal, Food Processing, Fertilizers Manufacturing, IPPU, Manure Applied to Soils, Manure Left on Pasture, Manure Management, Fires in Organic Soils, Fires in Humid Tropical Forests, On-farm Energy Use, Rural Population, Urban Population, Total Population - Male, Total Population - Female, Total Emission, Average Temperature C.

Size of the data

Rows: The dataset consists of 6965 rows. Each row represents a unique yearly record of CO2 emission value for different agri-food activities of 236 different countries in the world from the year 1990 to 2020.

Columns: There are 31 columns present in the collected dataset. These columns include data such as the area, year, CO2 emission produced by agricultural activities, total emission and the average temperature rise. Here each CO2 factor is measured in kilotons (kt). 1kt is equal to 1000

kg of CO₂. Average temperature column represents the average rise in temperature for the particular year which is measured in celsius.

Thus, the actual size of the data is 6965 rows X 31 columns, which covers the emission from all over the world for 30 years ranging from 1990 to 2020.

Independent Variables

The independent variables in this dataset are columns that influence CO₂ emissions in the agricultural sector. Below are the details of the only selected features.

Area: Country or Region name.

Year: The year of data.

Below are the selected activities which are contributing to the CO₂ emission:

Crop Residues, Rice Cultivation, Drained organic soils (CO₂), Pesticides Manufacturing, Food Transport, Forestland, Net Forest conversion, Food Household Consumption, Food Retail, Food Packaging, Agrifood Systems Waste Disposal, Food Processing, Fertilizers Manufacturing, IPPU, Manure applied to Soils, Manure left on Pasture, Manure Management, Fires in organic soils, Fires in humid tropical forests.

On-farm Electricity Use: Total electricity consumed on farms.

On-farm energy use: Total energy consumed on farms.

Rural population: people living in rural areas.

Urban population: People living in urban areas.

Total Population-Male: Total male population of the country.

Total Population-Female: Total Female population of the country.

Average Temperature: Average temperature rise of the country in degree celsius.

Dependent Variable: Total emission is the single dependent variable which is the total summation of all the agriculture activities from the sheet for the particular year.

total_emission: total CO2 emission recorded for a country in a particular year.

3.3.2. Data Preparation

3.3.2.1. Feature Selection

In the feature selection phase, the project aimed to streamline the dataset to focus on the most relevant variables and derive a consolidated measure of total CO2 emissions from the selected features within the agri-food sector. Hence below steps are essential to carry out the selection and transformation.

Selecting Relevant Columns: In this step, below code snippet extract the dataset to include only the required columns that are useful for analysis of CO2 emissions. The selected columns can be seen in the figure. Change the below figure as per updated code.

```
[ ] # Considering only relevant columns required for the analysis and modeling
  relevant_columns = [
    'Area', 'Year', 'Crop Residues', 'Rice Cultivation', 'Drained organic soils (CO2)',
    'Pesticides Manufacturing', 'Agrifood Systems Waste Disposal', 'Food Processing',
    'Fertilizers Manufacturing', 'Manure applied to Soils', 'Fires in organic soils',
    'Food Transport', 'Food Packaging', 'Manure left on Pasture', 'Manure Management',
    'Average Temperature °C'
  ]
  data_f = data_f[relevant_columns]
```

Figure 3.2: Code for Selection of Relevant Columns

3.3.2.2. Data Cleaning

Missing values are handled in this step by filling them with the median of each column, to ensure the integrity and completeness of the dataset. By this approach, the central tendency of

the data is preserved without any gaps and then data is ready for further steps providing a solid foundation for exploratory data analysis (EDA) and modeling.

```
# Filling missing values with the median value of each column for numeric columns only
filled_data = data_f.fillna(data_f.median(numeric_only=True))

#Verifying if there are any missing values left after filling missing values
filled_data.isnull().sum()
```

	0
Area	0
Year	0
Crop Residues	0
Rice Cultivation	0
Agrifood Systems Waste Disposal	0
Organic Soil Emission	0
Food Supply Chain	0
Agrochemical Manufacturing	0
Emission from Manure	0
Total Emission (kt)	0
Average Temperature °C	0

dtype: int64

Figure 3.4: Code for Filling Missing Values

3.3.2.3. Data Wrangling

Data wrangling is a process for manipulating raw data into an appropriate format. This involves the data aggregation, splitting and transformation of collected data which helps for further analyses and modeling.

Data Aggregation: In the below code snippet, new features or columns are generated by merging related columns to capture broader categories of the emission. Then by using new features, the dataset becomes easier to work for analyses and modeling.

From the picture below, Organic Soil Emission, Food Supply Chain, Agrochemical Manufacturing, Emission from Manure are the new features obtained from merging the columns.

▼ Data Aggregation

Merging existing columns to create new features

```
data_f['Organic Soil Emission'] = data_f['Drained organic soils (CO2)'] + data_f['Fires in organic soils']
data_f['Food Supply Chain'] = data_f['Food Processing'] + data_f['Food Packaging'] + data_f['Food Transport']
data_f['Agrochemical Manufacturing'] = data_f['Fertilizers Manufacturing'] + data_f['Pesticides Manufacturing']
data_f['Emission from Manure'] = data_f['Manure applied to Soils'] + data_f['Manure left on Pasture'] + data_f['Manure Management']
```

Dropping the columns columns

Dropping the original columns after merging them to create new features to simplify the dataset.

```
[ ] # Dropping the original columns after merging the columns to new variable
columns_to_drop = [
    'Drained organic soils (CO2)', 'Fires in organic soils',
    'Food Processing', 'Food Packaging', 'Food Transport',
    'Fertilizers Manufacturing', 'Pesticides Manufacturing',
    'Manure applied to Soils', 'Manure left on Pasture', 'Manure Management'
]
data_f = data_f.drop(columns=columns_to_drop)
```

Calculating Total Emission

```
[ ] # Calculating the total emission and updating the 'Total Emission (kt)' column, based on the selected relevant columns.
data_f['Total Emission (kt)'] = data_f[['Crop Residues', 'Rice Cultivation', 'Agrifood Systems Waste Disposal', 'Organic Soil Emission',
    'Food Supply Chain', 'Agrochemical Manufacturing', 'Emission from Manure']].sum(axis=1)
```

Figure 3.3: Code for Data Aggregation

After creating new features, existing columns which are used to generate these features have to be dropped to simplify the dataset and recalculate the total emission by adding new features and other existing columns to match the value.

Data Splitting: Data splitting is one of the important steps in the data preparation process which involves the splitting of the collected and transformed data into training and testing purposes. As part of this project approach, the data is splitted in the ratio of 70:30 as into training and testing sets. 70% accounts for training data and 30% accounts for the testing data. The training set is used to feed the data for models training and the testing data set is used to evaluate models performance on unseen data. This helps us to determine how well the model generalizes new data and performs on an unseen dataset.

```

[12] # Splitting the data into training and testing sets
#training and testing set for emission
X_train_emissions, X_test_emissions, Y_train_emissions, Y_test_emissions = train_test_split(x_features, y_emissions, test_size=0.3, random_state=10)

#training and testing set for temperature
X_train_temp, X_test_temp, Y_train_temp, Y_test_temp = train_test_split(x_features, y_temperature, test_size=0.3, random_state=10)

```

Figure 3.5: Code for Data Splitting for Training and Testing

As per above picture, data is splitted for CO2 emission and temperature. `train_test_split()` is the function from the `sklearn.model_selection` module that helps to split the data into training and testing sets. Parameter '`test_size=0.3`' specifies that 30% of the data is allocated for testing, and the remaining 70% for training data. Parameter `random_state=10` ensures reproducibility of the split by setting a random seed. Here, variables '`X_train_emissions`' and '`X_test_emissions`' are splitted for training and testing the selected predictive models. And '`Y_train_emissions`' and '`Y_test_emissions`' are the target emissions values for training and testing. Similarly, for temperature '`X_train_temp`' and '`X_test_temp`' are splitted sets for training and testing the temperature prediction model. And, '`Y_train_temp`' and '`Y_test_temp`' are the target temperature values for training and testing.

Data Transformation: The Data Transformation step helps in improving the quality and consistency of the data which is later utilized into selected models. This step standardizes and organizes the collected data by using available methods. As part of this project CO2 emission data, standardization and Imputation data transformation methods are implemented.

Standardization: Standardization transforms the data so that it has an average of 0 and a standard deviation of 1, which ensures that each variable has an equal effect on the model by eliminating biases from different scales. It also helps models to work much faster and effectively

which will be creating feature space. Additionally, it helps to stabilize the model and its performance by making variables consistent.

As per the code below, a standard scaler object is created for the ‘StandardScaler()’ module, which is imported from the ‘scikit-learn’ library. The ‘fit_transform’ method is used here, which takes a training set as a parameter for emission (X_train_emissions) and temperature(X_train_temp) to calculate the mean and standard deviation and then transform the data accordingly using ‘transform()’ function for test data.

1.2.3 Data Transformation

```
[ ] # Standardizing the selected features
obj_scaler = StandardScaler()

#for emission
X_train_emissions = obj_scaler.fit_transform(X_train_emissions)
X_test_emissions = obj_scaler.transform(X_test_emissions)

#for temperature
X_train_temp = obj_scaler.fit_transform(X_train_temp)
X_test_temp = obj_scaler.transform(X_test_temp)

[ ] from sklearn.impute import SimpleImputer
#Imputing missing values into the column, using mean imputation as an example
obj_imputer = SimpleImputer(strategy='mean')

#for emission
X_train_emissions_imp = obj_imputer.fit_transform(X_train_emissions)
X_test_emissions_imp = obj_imputer.transform(X_test_emissions)

#for temperature
X_train_temp_imp = obj_imputer.fit_transform(X_train_temp)
X_test_temp_imp = obj_imputer.transform(X_test_temp)
```

Figure 3.6: Code for Data Transformation

Imputation: Imputation involves filling the missing values in the data which ensures that the dataset is complete and ready for modeling. These missing values are due to many reasons, but making sure those spaces are handled before modeling is the crucial step.

As per the above code, ‘SimpleImputer’ class from ‘sklearn.impute’ library is used to fill missing values where ‘strategy=‘mean’ specifies that empty values should be filled with the mean value of the column. ‘fit_transform’ method from the ‘SimpleImputer’ class helps to

impute the missing values for both emission and temperature training data sets. 'transform' on the testing data applies the imputation using the mean calculated from the training data. These steps confirm that there are no missing values left on training and testing data, making the dataset smooth for modeling.

3.4. Exploratory Data Analysis

Exploratory data analysis is the important part of the project where it involves visualization of data based on the relationship between the variables, patterns and characteristics of the data. The primary goal of EDA is to summarize and display the main properties of the collected data by using available methods. This data exploration helps in further analyses and is useful for modeling.

There are several techniques used in this project to explore and understand data which are listed below and their results are explained in the 'Results' section of this report.

Understanding the data: To understand the dataset and its characteristics initially, some steps are necessary which are listed below.

1. Displaying first five rows
2. Reviewing data structures and types
3. Identifying Numeric and Non-Numeric Columns

Descriptive Statistics: This step provides descriptive statistics for all the numerical columns in the dataset which includes metrics such as count, mean, standard deviation, minimum, and maximum values, 25th, 50th, and 75th percentiles. This information helps to understand the

distribution, central tendency, and variability of the data which eventually helps in further analyses and modeling.

Data Visualization: Data Visualization step provides a graphical representation of the data which is useful for the exploration of data pattern, trends and relationship between the variables. There are several types of data visualization techniques available where this project used and showed certain types such as dual axis plot, box plot, bar plot, scatter plot and histograms.

Correlation Analyses: A Correlation analysis identifies relationships between each numerical variable with all required features in the data. The 'corr()' function is used here for correlation matrix and 'heatmap' function is used to visualize the correlation matrix with annotations and a cool warm color. The heatmap shows the correlation coefficients between pairs of numerical variables. Positive values indicate positive correlations and Negative values (down to -1) indicate negative correlations. Positive correlations indicate that an increase in one variable is associated with an increase in another where negative correlations indicate an increase in one variable is associated with a decrease in another.

3.5. Modeling

3.5.1. Model selection

For predictive analyses on CO2 emission, four models have been selected where these models help for feature extraction and prediction of CO2 emissions. They are Linear Regression, Decision Tree, Random Forest, and Neural Networks. Each model has its own unique strengths

and capabilities to predict CO2 emissions. Data from different angles are captured by these models and also ensures thorough analysis of using the best features of each model. The selection of these four models provides a balanced methodology for predicting CO2 emissions. Using the unique strengths and capabilities of each model, the project aims to estimate performance metrics of each model.

Linear Regression: Linear Regression is a phenomenal and most used statistical model for predicting continuous target variables. To measure CO2 emissions and average temperature metrics, Linear Regression model is trained and an instance of 'LinearRegression()' is created and fits it on the trained emission and temperature data by using 'fit()' method . After training the models, 'predict()' function is used to predict CO2 emissions and average temperature on the test sets. To evaluate the model's performance, 'evaluate_model()' function has been created, which calculates key metrics such as Mean Squared Error (MSE), RMSE(Root Mean Squared Error), MAE(Mean Absolute Error) and R -squared(Coefficient of Determination).

Linear Regression

```
[ ] # Linear Regression
    lin_reg_emission = LinearRegression()

    #For emission
    lin_reg_emission.fit(X_train_emissions_imp, Y_train_emissions)
    y_pred_lin_reg_emission = lin_reg_emission.predict(X_test_emissions_imp)
    lin_reg_metric_emission = model_evaluation(Y_test_emissions, y_pred_lin_reg_emission)

    #for temperature
    lin_reg_temp = LinearRegression()
    lin_reg_temp.fit(X_train_temp_imp, Y_train_temp)
    y_pred_lin_reg_temp = lin_reg_temp.predict(X_test_temp_imp)
    lin_reg_metrics_temp = model_evaluation(Y_test_temp, y_pred_lin_reg_temp)
```

Figure 3.7: Linear Regression Model Implementation

Decision Tree: Due to its fine tuning and capable of making accurate predictions the decision tree model ensures that it is flexible to capture complex relationships in the data.

Below code displays how a Decision Tree model is optimized to calculate emissions and temperature. The steps involve the use of the ‘DecisionTreeRegressor()’ class with separate training models for emissions and temperature. The training model uses the ‘GridSearchCV()’ method which tests various combinations of hyperparameters defined in ‘dt_param_grid_ht’. After this the Decision Tree model is fitted to the trained data for emission and temperature. Then the best estimator is selected and predictions are made on the test set by fitting the training set for CO2 emissions and temperature rise. The model's performance is evaluated using the ‘evaluate_model()’ function, which calculates metrics like MSE, RMSE, MAE and R-squared (R2) for both temperature and emission.

```

# Decision Tree
dec_tre_emission = DecisionTreeRegressor(random_state=10)
dt_emission_grid_search = GridSearchCV(estimator=dec_tre_emission, param_grid=dt_param_grid_ht, cv=5, n_jobs=-1, scoring='r2')

#for emission
dt_emission_grid_search.fit(X_train_emissions_imp, Y_train_emissions)
dt_best_emission = dt_emission_grid_search.best_estimator_
y_pred_dt_emission = dt_best_emission.predict(X_test_emissions_imp)
dec_tre_metric_emission = model_evaluation(Y_test_emissions, y_pred_dt_emission)

#for temperature
dec_tre_temp = DecisionTreeRegressor(random_state=10)
grid_search_dt_temp = GridSearchCV(estimator=dec_tre_temp, param_grid=dt_param_grid_ht, cv=5, n_jobs=-1, scoring='r2')
grid_search_dt_temp.fit(X_train_temp_imp, Y_train_temp)
best_dt_temp = grid_search_dt_temp.best_estimator_
y_pred_temp_dt = best_dt_temp.predict(X_test_temp_imp)
dec_tre_metric_temp = model_evaluation(Y_test_temp, y_pred_temp_dt)

```

Figure 3.8: Decision Tree Implementation

Random Forest: To measure CO2 emissions and average temperature using Random Forest, firstly separate models are trained for each prediction task. For CO2 emissions, ‘RandomForestRegressor()’ variable is initialized with a random state 10 for reproducibility.

‘RandomizedSearchCV()’ is used to optimize the hyperparameters of the model($n_iter=10$) from a predefined distribution (`rf_param_grid_ht`) and measured each combination using 3 fold cross-validation. The R-squared metric is used to calculate the model performance and the best model is selected based on the highest R-squared value. The best estimator is selected and predictions are made on the test set by fitting the training set for CO2 emissions. The model's performance is evaluated using the ‘`evaluate_model()`’ function, which calculates metrics like MSE, RMSE, MAE and R-squared (R2). Similarly, temperature performance is evaluated using the Random Forest model.

```
[38] from sklearn.model_selection import RandomizedSearchCV
# Random Forest
#for emission
ran_for_emission = RandomForestRegressor(random_state=10)
rf_random_search_emission = RandomizedSearchCV(estimator=ran_for_emission, param_distributions=rf_param_grid_ht,
                                              n_iter=10, cv=3, n_jobs=-1, scoring='r2', random_state=10)
rf_random_search_emission.fit(X_train_emissions_imp, Y_train_emissions)
rf_best_emission = rf_random_search_emission.best_estimator_
y_pred_rf_emission = rf_best_emission.predict(X_test_emissions_imp)
rf_metric_emission = model_evaluation(Y_test_emissions, y_pred_rf_emission)

#for temperature
ran_for_temp = RandomForestRegressor(random_state=10)
rf_random_search_temp = RandomizedSearchCV(estimator=ran_for_temp, param_distributions=rf_param_grid_ht,
                                           n_iter=10, cv=3, n_jobs=-1, scoring='r2', random_state=10)
rf_random_search_temp.fit(X_train_temp_imp, Y_train_temp)
rf_best_temp = rf_random_search_temp.best_estimator_
y_pred_rf_temp = rf_best_temp.predict(X_test_temp_imp)
rf_metric_temp = model_evaluation(Y_test_temp, y_pred_rf_temp)
```

Figure 3.9: Random Forest Implementation

Neural Networks: The Neural Network used in this project is specifically designed for regression tasks and the type of this neural network known as ‘Feedforward neural network(FNN)’. The below code snippet involves the initiation, training and evaluation of a neural network model designed to predict CO2 emission and temperature rise.

```
[ ] # Function to create and compile the neural network model
def neu_net_model_creation(input_dim):
    model = Sequential()
    model.add(Dense(64, input_dim=input_dim, activation='relu'))
    model.add(Dense(32, activation='relu'))
    model.add(Dense(1)) # Output layer for regression
    model.compile(optimizer='adam', loss='mse')
    return model
```

Figure 3.10: Neural Network model function

The function 'neu_net_model_creation()' is declared initially which defines the architecture of the neural network. The input dimension (input_dim) is determined by the number of features in the training data and the network includes two hidden layers where the first hidden layer has 64 neurons and the second hidden layer has 32 neurons with the activation parameter as 'ReLU' (Rectified Linear Unit) which imparts non linearity to the model and helps to absorb the complex relationships between the data.

The output layer is assigned with a 1 neuron with a linear activation function where continuous values are predicted. Last step of this function compiles where it takes 'adam' as optimizer which is a popular optimization algorithm in deep learning, 'mse' (Mean squared error) assigned to the loss function which is the standard for regression problems.

```

▶ # Neural Network for CO2 Emissions
nn_emission = neu_net_model_creation(input_dim=X_train_emissions_imp.shape[1])
early_stopping = EarlyStopping(monitor='val_loss', patience=10, restore_best_weights=True)
hist_emission = nn_emission.fit(X_train_emissions_imp, Y_train_emissions,
                                validation_split=0.3,
                                epochs=100,
                                batch_size=32,
                                callbacks=[early_stopping],
                                verbose=1)

⚙ Show hidden output

[ ] # Predictions and evaluation for CO2 Emissions
y_pred_nn_emission = nn_emission.predict(X_test_emissions_imp)
nn_metric_emission = model_evaluation(Y_test_emissions, y_pred_nn_emission)

⚙ 66/66 ————— 0s 2ms/step

[ ] # Neural Network for Temperature
nn_temp = neu_net_model_creation(input_dim=X_train_temp_imp.shape[1])
history_temp = nn_temp.fit(X_train_temp_imp, Y_train_temp,
                            validation_split=0.3,
                            epochs=100,
                            batch_size=32,
                            callbacks=[early_stopping],
                            verbose=1)

⚙ Show hidden output

[ ] # Predictions and evaluation for Temperature
y_pred_temp_nn = nn_temp.predict(X_test_temp_imp)
nn_metric_temp = model_evaluation(Y_test_temp, y_pred_temp_nn)

```

Figure 3.11: Neural Network model implementation

As shown in the above picture, the neural network model is trained for both CO2 emission and temperature rise. ‘EarlyStopping’ function is used here to monitor the validation loss during training and ‘patience’ is 10. This means the model weights from the best epoch are restored if the validation loss does not improve for 10 consecutive epochs, training is halted early. This method ensures the model reacts well to unseen data and prevents overfitting.

Here the model is trained on 70% of the data and the 30% used for testing with 100 epochs and a batch size of 32 which means the model updates its weights every 32 samples.

After training, the model is used to make predictions on the test data. The predicted values are then evaluated using the 'model_evaluation()' function. This neural network feedforward approach and early stopping provides a robust method for predicting CO₂ emissions and temperature variables.

3.6. Ethical Consideration

This applied research project follows ethical guidelines to ensure the responsible use, analysis, and implementation of CO₂ emissions data with models in the agri-food field. There is no need for informed consent or anonymity since this secondary data is available publicly and has given rights to use it. However, Project is adhering to data providers guidelines and usage terms.

Machine learning techniques and models like Linear Regression, Decision Trees, Random Forest, and Neural Networks are used in this project transparently. The methods and processes are chosen carefully to match the research objectives and results are reported honestly with clear explanations of each model's performance avoiding any bias and manipulation. The research aims to analyze and predict CO₂ emissions in agriculture by supporting policymaking and sustainable practices. The results are shared openly with clear documentation of methodologies by ensuring the reproducibility and reviewed by others. The data and analytical methods are used appropriately with these ethical practices, the study contributes to sustainable environment and social responsibility.

3.7. Conclusion

To conclude about the methodology of this applied research project, study followed CRISP-DM methodology for data mining process which includes all the steps in making the right predictive analyses of the data. In this section only the implementation part has shown with clear details about the data, code and different ML steps. The evaluation and findings from this implementation will be shown clearly in the 'Results' section with all details.

4. RESULTS

4.1. Exploratory Data Analysis

4.1.1. Descriptive Statistics

This step gives descriptive statistics of the numeric variables or columns in the dataset. The picture below depicts the descriptive statistics of each numerical variable in agri-food emission, which includes metrics such as variable counts along with their averages and standard deviations for minimums and maximum values, and 25th, 50th, and 75th percentiles. This result helps to understand the range (std, min and max), central tendency (mean and median) as well as overall distribution of particular numerical variables in the agri-food data. Finally, this overall detail helps in capturing the scale of data, potential outliers and variability in the dataset.

```
[18] data_f.describe().drop(columns=['Year'])
```

	Crop Residues	Rice Cultivation	Agrifood Systems	Waste Disposal	Organic Soil Emission	Food Supply Chain	Agrochemical Manufacturing	Emission from Manure	Total Emission (kt)	Average Temperature °C
count	5576.000000	6965.000000	6965.000000	6965.000000	6.965000e+03	6965.000000	6965.000000	6037.000000	6.965000e+03	6965.000000
mean	998.706309	4259.666673	6018.444633	4.713544e+03	7470.936029	3369.141749	6772.713670	3.250161e+04	0.872989	
std	3700.345330	17613.825187	22156.742542	3.515691e+04	34842.031265	12786.505252	19537.979375	1.119052e+05	0.555930	
min	0.000200	0.000000	0.340000	0.000000e+00	23.186100	0.473200	0.482600	7.262568e+02	-1.415833	
25%	11.006525	181.260800	86.680500	0.000000e+00	346.608417	404.705622	195.438400	3.667300e+03	0.511333	
50%	103.698200	534.817400	901.275700	0.000000e+00	620.441780	1440.678229	1383.779500	6.953818e+03	0.834300	
75%	377.640975	1536.640000	3006.442100	6.916873e+02	2803.326650	2338.417463	4800.693500	1.580989e+04	1.206750	
max	33490.074100	164915.255600	213289.701600	1.181812e+06	478554.086400	182269.423300	159007.844900	1.347692e+06	3.558083	

Figure 4.1: Data Describe

- **Displaying First 5 rows**

```
[ ] data_f.head()
```

	Area	Year	Crop Residues	Rice Cultivation	Agri-food Systems Waste Disposal	Organic Soil Emission	Food Supply Chain	Agrochemical Manufacturing	Emission from Manure	Total Emission (kt)	Average Temperature °C
0	Afghanistan	1990	205.6077	686.00	691.7888	0.0	382.960756	23.804483	2169.8513	4160.013039	0.536167
1	Afghanistan	1991	209.4971	678.16	710.8212	0.0	381.058056	24.565973	2268.1735	4272.275829	0.020667
2	Afghanistan	1992	196.5341	686.00	743.6751	0.0	373.162556	25.204973	2267.4190	4291.995729	-0.259583
3	Afghanistan	1993	230.8175	686.00	791.9246	0.0	374.207256	25.767973	2256.9791	4365.696429	0.101917
4	Afghanistan	1994	242.0494	705.60	831.9181	0.0	373.832956	26.838973	2324.6596	4504.899029	0.372250

Figure 4.2: Displaying First 5 Rows

Above table gives an initial glimpse into the data structure and the types of values by displaying first rows from the data set. ‘head()’ method is used here to display the first five rows of the data.

- **Data type and Non-null count**

```
[ ] data_f.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6965 entries, 0 to 6964
Data columns (total 11 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   Area                                       6965 non-null   object
1   Year                                       6965 non-null   int64
2   Crop Residues                             5576 non-null   float64
3   Rice Cultivation                         6965 non-null   float64
4   Agri-food Systems Waste Disposal         6965 non-null   float64
5   Organic Soil Emission                    6965 non-null   float64
6   Food Supply Chain                        6965 non-null   float64
7   Agrochemical Manufacturing               6965 non-null   float64
8   Emission from Manure                     6037 non-null   float64
9   Total Emission (kt)                     6965 non-null   float64
10  Average Temperature °C                   6965 non-null   float64
dtypes: float64(9), int64(1), object(1)
memory usage: 598.7+ KB
```

Figure 4.3: Data Information about Null Values and Data Type

The above information about data includes non-null values count and data types of the selected features from the data. ‘info()’ method is used here to display this data. A high degree of

completeness achieved by most of the columns with 6965 non-null values. However, "Crop Residues" and "Emission from Manure" have some null values in the columns, which are handled in data cleaning. All columns present in the data are numerical values, except the "Area" column which is object (categorical) type.

4.1.2. Data Visualizations

- **Total Emission and Average temperature rise over the time**

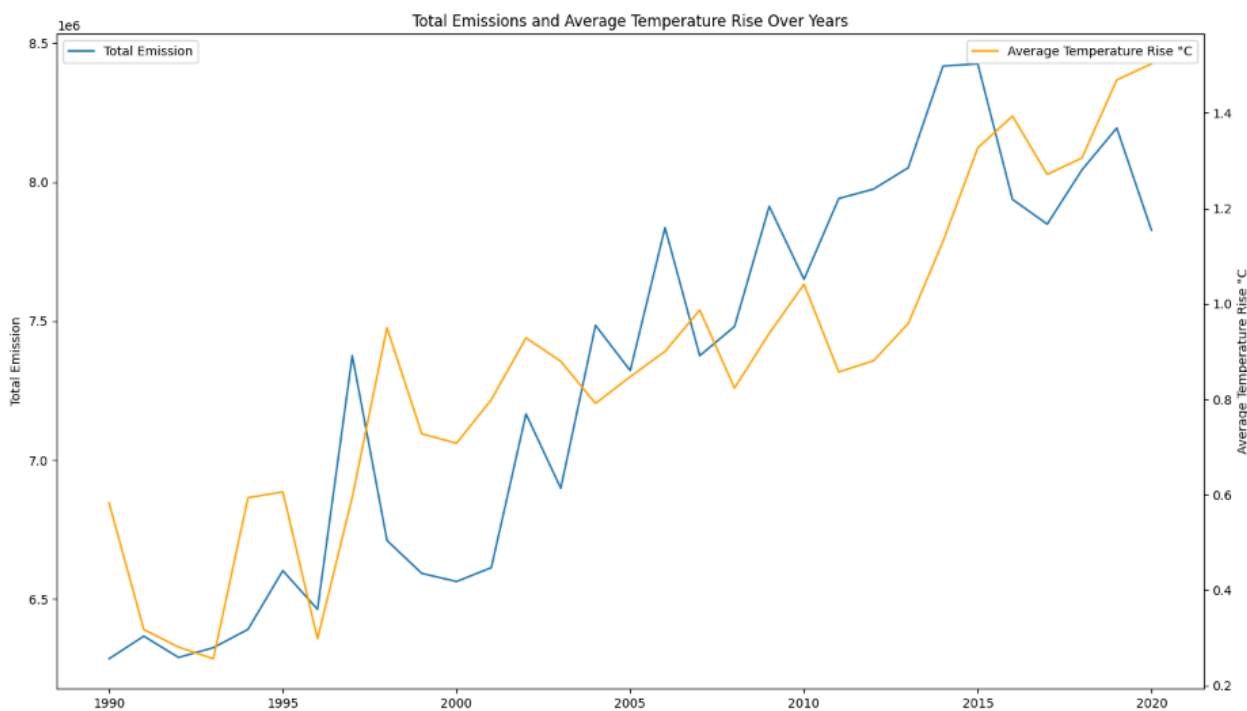


Figure 4.4: Graph Plot for Total Emission and Average Temperature Rise Over Time

The above line chart graph compares Total Emissions (kilotons) with Average Temperature Rise (degrees Celsius) over a span of 30 years. Years are denoted by X axis ranging from 1990 to 2020, 'Total Emission' is denoted by the left vertical Y axis with the scale in the

range of millions(1e6) and 'Average Temperature Rise' is denoted by the right vertical Y axis with scale ranges from 0.2 C to 1.5 C. 'Total Emission' is indicated by the blue line and 'Average Temperature' is indicated by Orange line.

Total emissions and temperature are gradually increasing over time which indicates a potential correlation between them. Both CO2 emissions and temperature continue to rise sharply in recent years, indicating ongoing environmental challenges with respect to global warming and greenhouse gas emissions.

- **Emission breakdown by category over time**

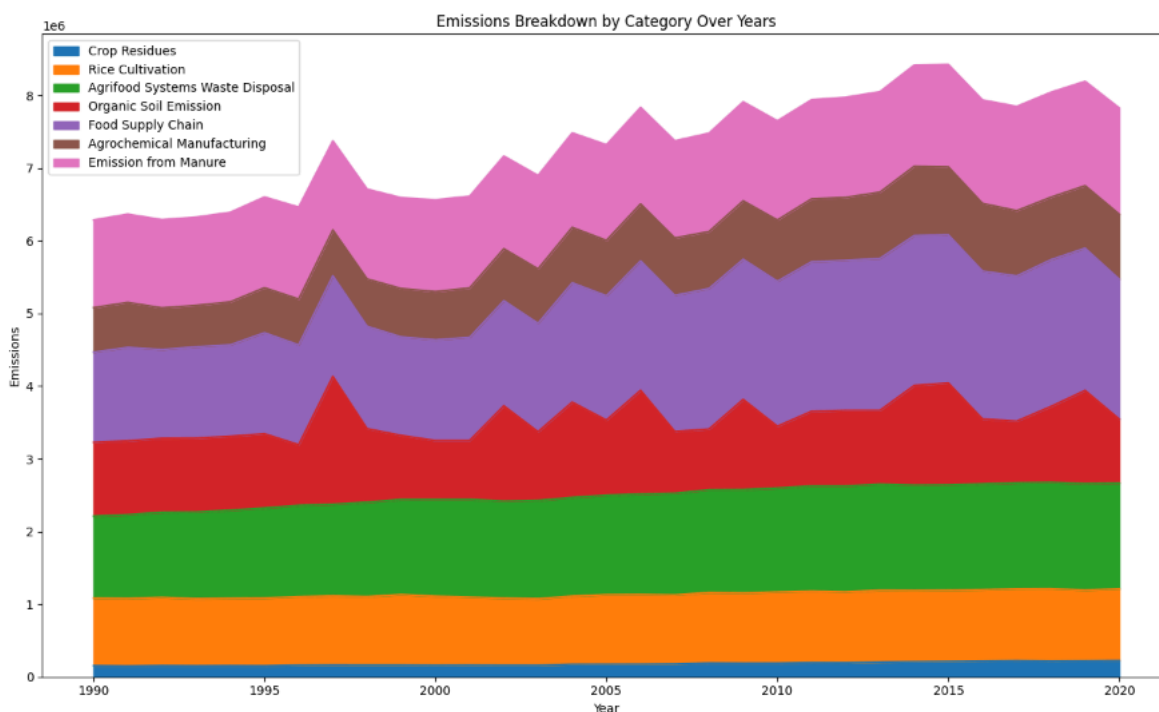


Figure 4.5: Graph Plot for Emission Breakdown by Category Over Time

The above stacked area chart shows the emissions by category over the years ranging from 1990 to 2020. Each color in the chart denotes different agri-food activities by their

emissions, Y axis represents the magnitude of the emissions (kilo tonnes) and X axis represents the years, ranging from 1990 to 2020.

Each category of agri-food emission is represented in different colors with a stack on top of each other. Lowest emission category ‘Crop Residues’ is at the bottom where most emission category ‘Emission from Manure’ is at the top. This stacking of each category shows the contribution to the total emissions and how each category's emissions have changed from 1990 to 2020. Below graph clearly tells that there is a noticeable increase in total emission with an upward trend. Graph provides insights of agri-food categories which need more focus in terms of emission reduction strategies.

- **Correlation Matrix**

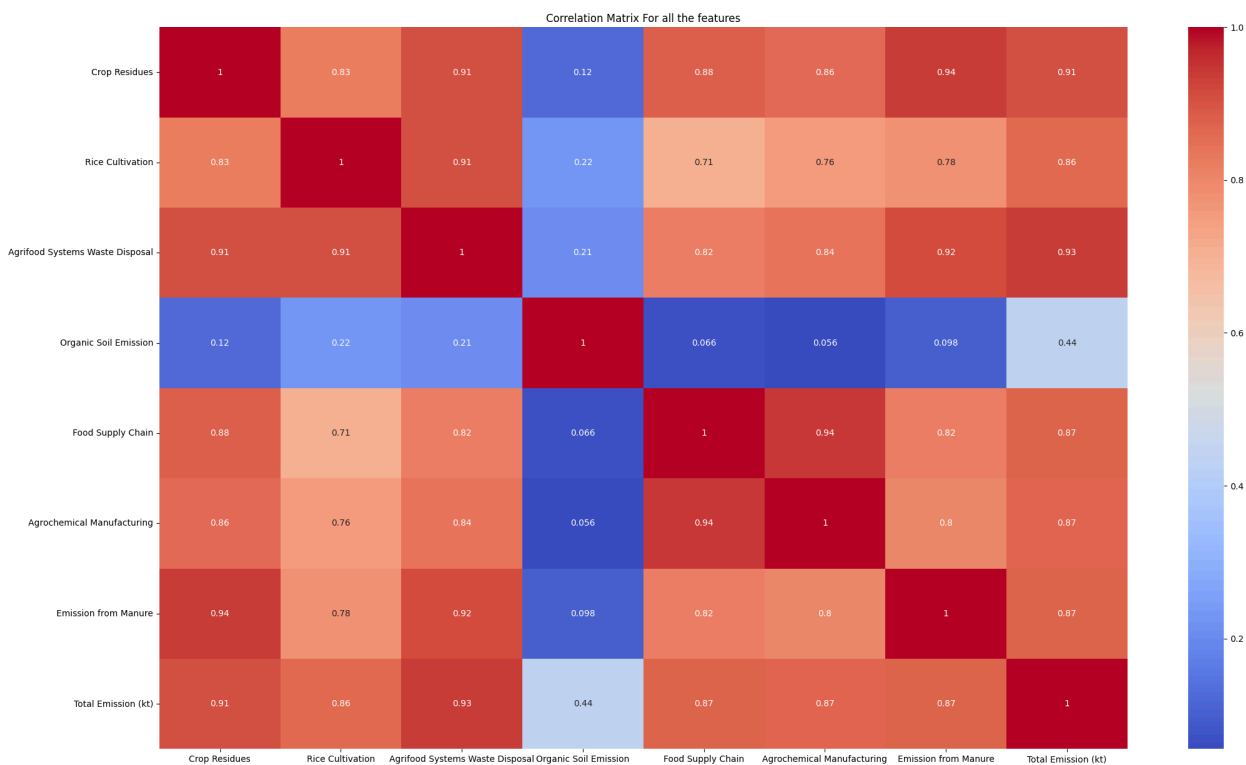


Figure 4.6: Correlation Matrix

Correlation coefficients values range from -1 to 1. The positive value of coefficient (+1) represents a positive correlation, negative value(-1) coefficient represents a negative correlation and zero (0) coefficient indicates no correlation. On the color side, Red color denotes a positive correlation, Blue color denotes a negative correlation. The intensity of the color reflects the strength of the coefficient value of the variable. For example, darker colors represent stronger positive correlation between the two variables. Each cell in the matrix shows the correlation coefficient between two variables. The diagonal cells showed a stronger correlation value of 1 because here the variable value is compared itself.

- **Impact of CO2 emission over top 20 countries**

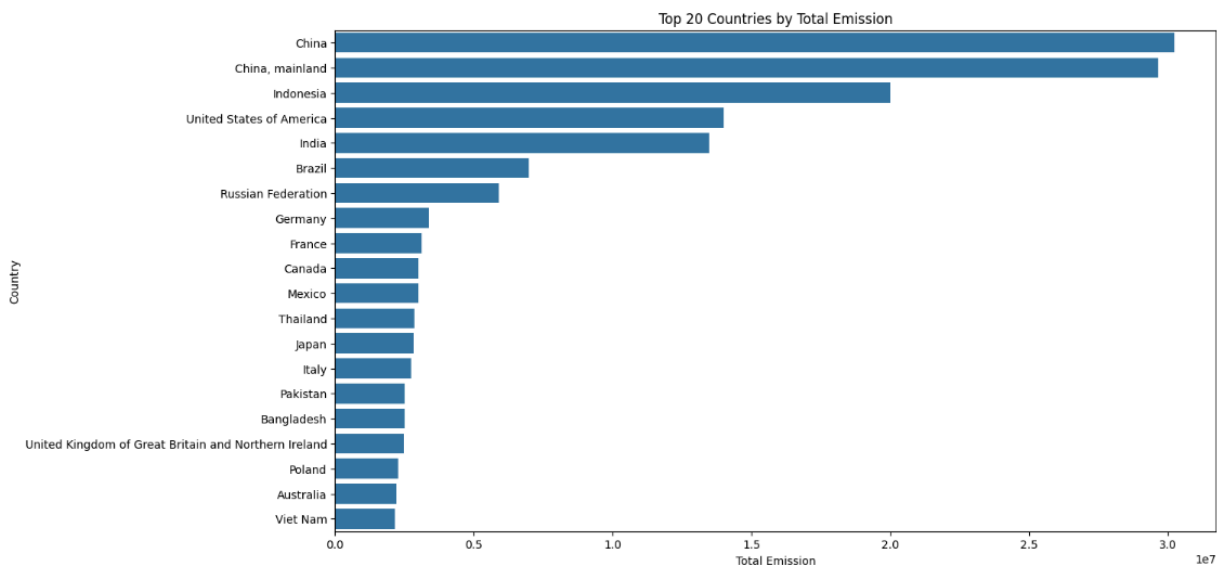


Figure 4.7: Graph Plot for Top 20 Countries of CO2 Emission

The bar chart shows the top 20 countries by total emissions. The X axis represents the total emissions(kilo tonnes) for each country in the range of tens of millions units. And the Y

axis lists the top 20 countries ranked by their total emissions. The length of the blue bar indicates the magnitude of emissions for that particular country.

As per graph, China and mainland China mainland has the highest emissions releasing country from the agri-food sector, followed by Indonesia, USA and India. These countries need targeted efforts to mitigate the impact of climate change. Countries like VietNam, Australia, Poland, UK and Bangladesh are marked as significant emission releasing countries according to data.

- **Mean CO2 emission of the agri-food activities.**

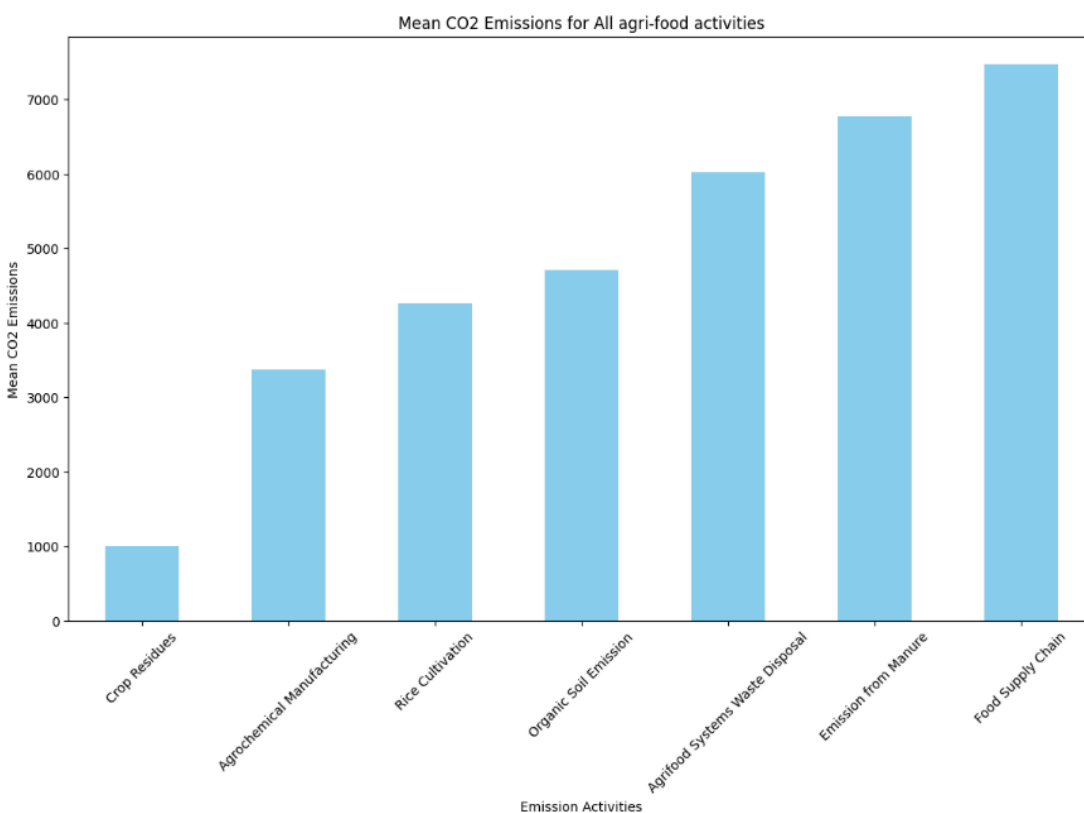


Figure 4.8: Graph Plot for Mean CO2 emission of the agri-food activities

The purpose of calculating and visualizing the mean CO₂ emissions for all agri-food activities is to identify which factors contribute the most to the overall CO₂ emissions which helps in prioritizing the activities for emission reduction efforts. From the above bar chart, x axis represents different agri-food activities where y axis represents mean CO₂ emission in kilo tonnes. The resulting mean emissions are sorted in ascending order from lowest to highest with the activities.

‘Crop Residue’ has the lowest one with release of less than 1000 kt emission and ‘Food supply chain’ factor has the highest emission releasing factor with value more than 7000 kt, followed by ‘Emission from Manure’, ‘Agrifood systems waste disposal’, ‘Organic Soil Emission’ and others.

- **Bar Plot of Total Emissions for Each Year**

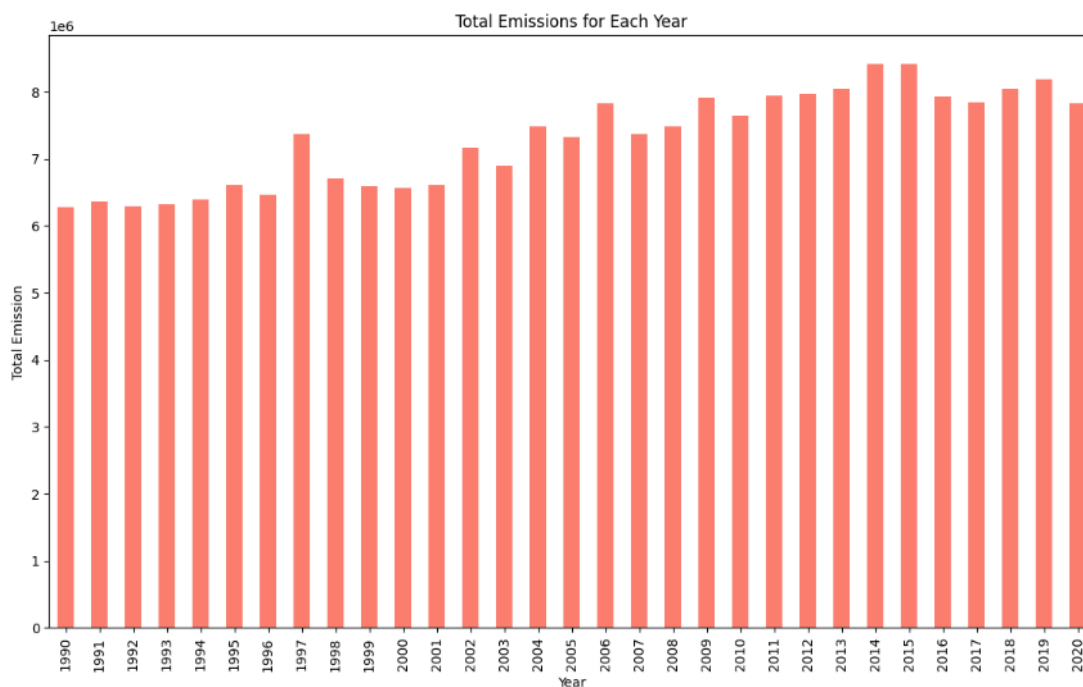


Figure 4.9: Bar Plot for Total Emission for Each Year

The purpose of the above bar plot is to show the total emission released in all countries from the years 1990 to 2020. This plot helps in understanding how emissions have changed annually and identifying trends and patterns over time to implement reduction policies. The x axis represents year values from 1990 to 2020 with the gap of 1 and y axis represents the total emission (kilo tonnes) in the range of one million(1e6).

As per the plotted graph, the emission trend has increased from year to year gradually. The year 2014 and 2015 are the highest CO₂ emission releasing years in the world with a value of more than 8 million kilo tonnes. There is a drastic reduction in the CO₂ emission in the year 2020. This graph helps policy makers and government bodies to identify the possible periods of increased emission activity and mitigation efforts.

4.2. Model Evaluation

Model evaluation metrics such as MSE, RMSE, MAE, and R² are the required measures for the performance of predictive models. These metrics show how well a model predicts CO₂ emissions and average temperature rise which assists in selection of the most accurate and reliable model.

Mean Squared Error (MSE): MSE is defined as the average squared difference between predicted and actual values. The model's predictions are closer to the actual values when the MSE value is less. High value errors will have a more impact on MSE, because the errors are squared. difference between the original and predicted values extracted by averaged the absolute difference over the data set.

Root Mean Squared Error (RMSE): RMSE is the square root of MSE value where a lower RMSE value indicates better model performance. RMSE metric provides an error metric in the same units as the target variable. when comparing RMSE value with actual data, it is more interpretable.

Mean Absolute Error (MAE): MAE is average of the absolute differences between predicted values and actual values. A lower MAE shows better model performance, as it allows one to get an estimate of how close the prediction value is from the actual value. MAE doesn't square the errors unlike MSE, and MAE metrics.

R-squared (R2): R-squared is defined as how well a model's predictions match with the actual data. It tells about the difference in variation of the outcome (dependent variable) can be explained by the input variables (independent variables). If R2 equals to 1, the model perfectly explains all the variance in the data. If R2 is 0, then the model does not show any of the variance in the data, if R2 value is less than zero (Negative), then the model performance is worse to explain the variance in the data.

Table 4.1: CO2 Emission: Model Performance Metrics

Model	MSE	RMSE	MAE	R2
Linear Regression	2.4074e-22	1.5515e-11	8.3798e-12	1.0000
Decision Tree	3.8861e+06	1.9713e+03	2.8318e+02	0.9996
Random Forest	1.9713e+06	1.4040e+03	1.9012e+02	0.9998
Neural Networks	9.1683e+06	3.0279e+03	1.8598e+03	0.9992

Table 4.2: Temperature Rise: Model Performance Metrics

Model	MSE	RMSE	MAE	R2
Linear Regression	1.3389e-31	3.6591e-16	2.7003e-16	1.0000
Decision Tree	3.3540e-05	5.7914e-03	1.2991e-03	0.9998
Random Forest	1.1812e-05	3.4368e-03	6.8835e-04	0.9999
Neural Networks	3.0378e-04	1.7429e-02	9.7416e-03	0.9990

From the above tabular tables of model performance, Linear Regression is the most accurate model with lowest MSE, RMSE and MAE values and R-squared with value 1. This model has perfect accuracy for both CO₂ emissions and temperature rise predictions. The Neural network model has the highest values of MSE, RMSE and MAE with the poorest performance for both CO₂ emissions and temperature predictions. This model shows higher errors and lower R-squared values.

4.2.1. Comparative Analysis

• CO2 Emissions

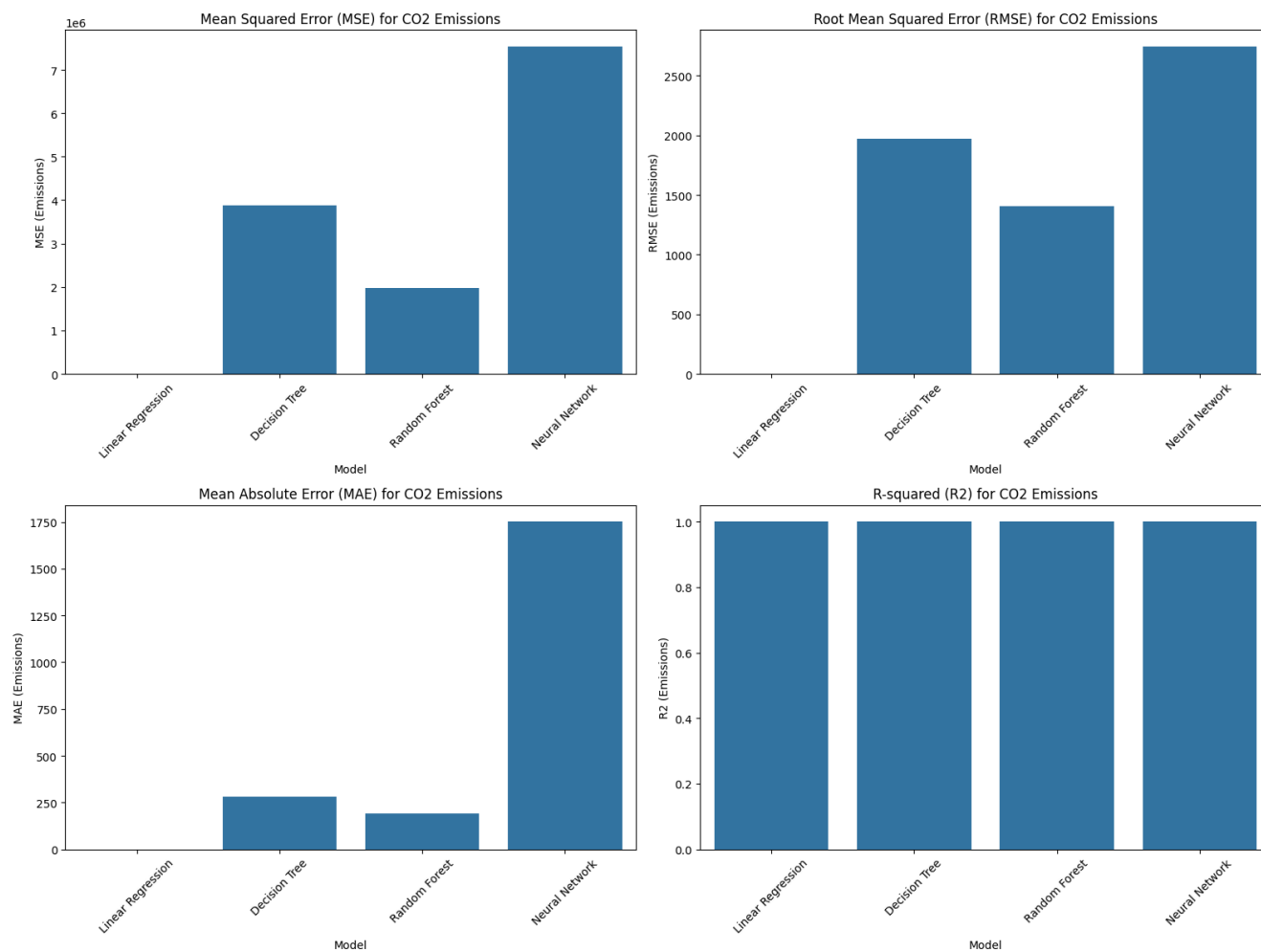


Figure 4.10: CO2 Emission Comparative Analysis between models metrics

From above comparative analysis between models it shows that Linear Regression's MSE, RMSE, MAE are lower than other model values and R-squared value is exactly 1. This indicates that the model predicts CO2 emissions with perfect accuracy and explains all the variance in the data. Random Forest model has performed well but is slightly less accurate than Linear Regression model. And the Decision tree model has moderate accuracy which is

outperformed by Random forest. These two models performed well with relatively low error metrics and R-squared value near to 1. Neural networks model has the highest MSE, RMSE, and MAE values with poorest performance among the other models, despite an R-squared value of 0.9995, it falls short compared to the other models.

- **Temperature Rise**

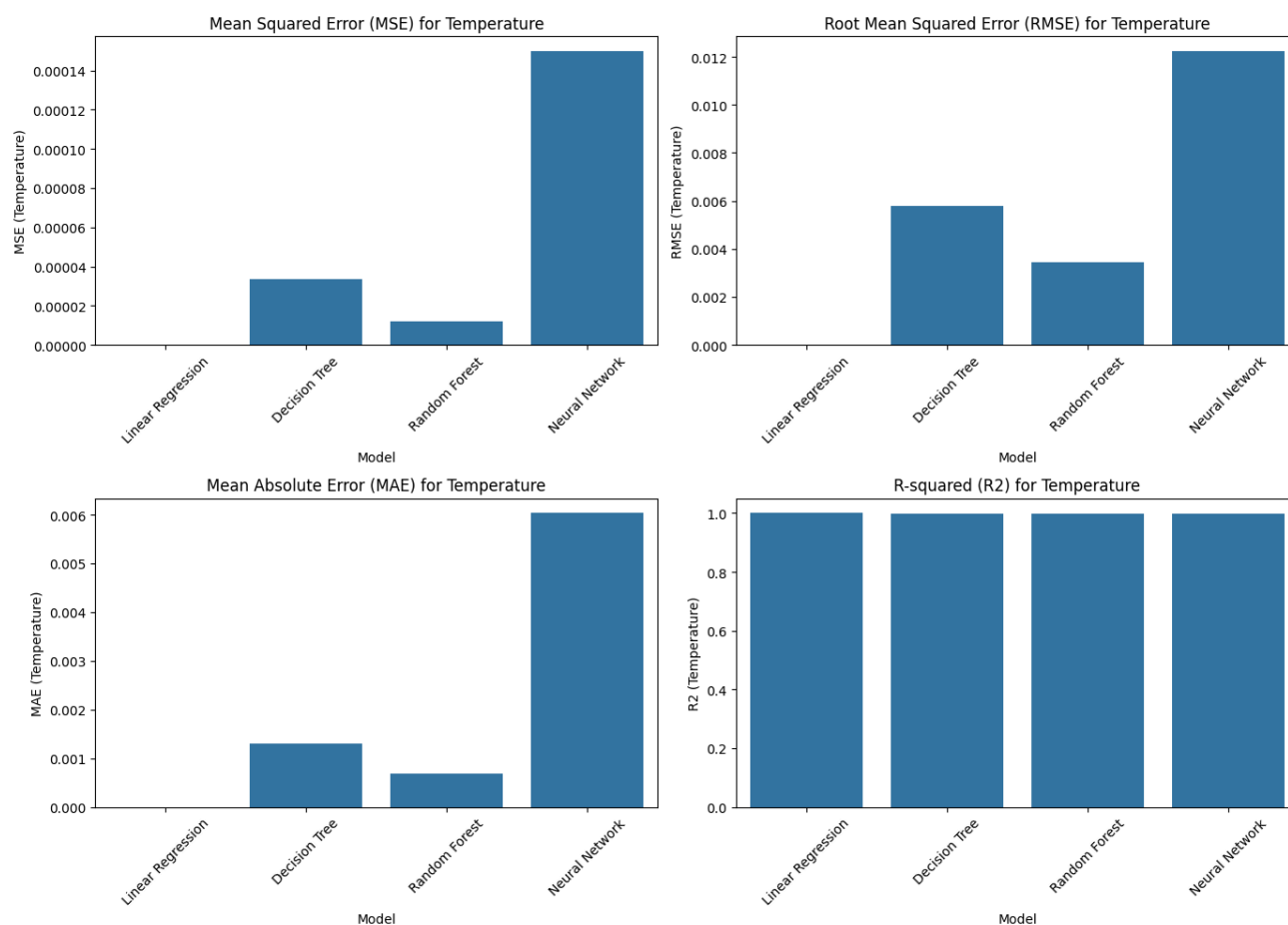


Figure 4.11: Temperature Rise Comparative Analysis Between Models Metrics

Here, Linear Regression proved again to be the accurate model for predicting temperature with negligible error metrics like MSE, RMSE and MAE, achieving an R-squared value of 1. This shows that the model is perfect for predictions, making it the top performer. Decision Tree

and Random Forest models followed closely to the linear regression with low error metrics and an impressive R-squared value of 0.9999. Again, the Neural Network model stood last in predicting temperature compared to other models with the highest MSE, RMSE and MAE values and the lowest R-squared value of 0.9986.

4.2.2. Hypothesis Testing

- **t-Test to compare mean values between Actual vs Predicted Values**

```
[52] from scipy.stats import ttest_rel

# t-test for CO2 Emissions
t_stat_emission, p_value_emission = ttest_rel(Y_test_emissions, y_pred_lin_reg_emission)
print('p-value(CO2 emission):', p_value_emission)

⇒ p-value(CO2 emission): [1.48872768e-28]
```

Figure 4.12: Code for CO2 Emission t-test

Here, t-test helps to check the mean difference between actual and predicted value of the best performing model Linear Regression. 'Y_test_emissions' is the actual value and 'y_pred_lin_reg_emission' is the predicted value. 'ttest_rel()' method from the 'scipy' library is used here to compare the values. The negative p value (1.488e-28) indicates that there is a statistically significant difference between the actual and predicted CO2 emissions. This result shows that the observed difference is not due to random chance.

```
[55] # t-test for Temperature
t_stat_temp, p_value_temp = ttest_rel(Y_test_temp, y_pred_lin_reg_temp)
print('p-value(temperature):', p_value_temp)

⇒ p-value(temperature): [0.01547003]
```

Figure 4.13: Code for Temperature t-test

For temperature, 'Y_test_temp' is the actual value and 'y_pred_lin_reg_emission' is the predicted value. The obtained p-value is below 0.05 which indicates that the difference between the actual and predicted temperature is statistically significant. This shows that the observed difference is not due to random chance.

4.2.3. Regression Coefficients Significance

- Check the significance of each feature's coefficient in Linear Regression

```
import statsmodels.api as sm

# For CO2 emissions
X_train_emissions_sm = sm.add_constant(X_train_emissions_imp)
lin_reg_sm_emission = sm.OLS(Y_train_emissions, X_train_emissions_sm).fit()
print(lin_reg_sm_emission.pvalues)
```

const	0.000000e+00
x1	9.666500e-13
x2	8.421342e-04
x3	6.957660e-05
x4	6.831718e-02
x5	2.479938e-02
x6	3.828358e-04
x7	3.838920e-01
x8	0.000000e+00
x9	1.646838e-02

dtype: float64

Figure 4.14: Code to Check Each Feature's Coefficient In Linear Regression

The p-values for each coefficient from the linear regression model indicates how statistically significant each feature is in evaluating the CO2 emissions. The constant is important in a linear regression model because it allows the line to be positioned anywhere on the graph, not just passing through the origin. The 'add_constant()' function adds a column of ones to the dataset 'X_train_emissions_imp' which represents the intercept term in the model. 'sm.OLS()'

method is used to specify the OLS regression model which takes two main arguments ‘Y_train_emissions’ and ‘X_train_emissions_sm’. ‘Y_train_emissions’ is the target variable for CO2 emission and ‘X_train_emissions_sm’ is the independent variable (features) with the added constant (intercept). The method ‘fit()’ is used to fit the model to the data.

In result, most of the features have p-values less than 0.05 which indicates that these features contribute to CO2 emission predictions meaningfully, also it suggests that the model fits to the data extremely well.

4.2.4. Actual Vs Predict Values for Linear Regression Model

- For CO2 Emission

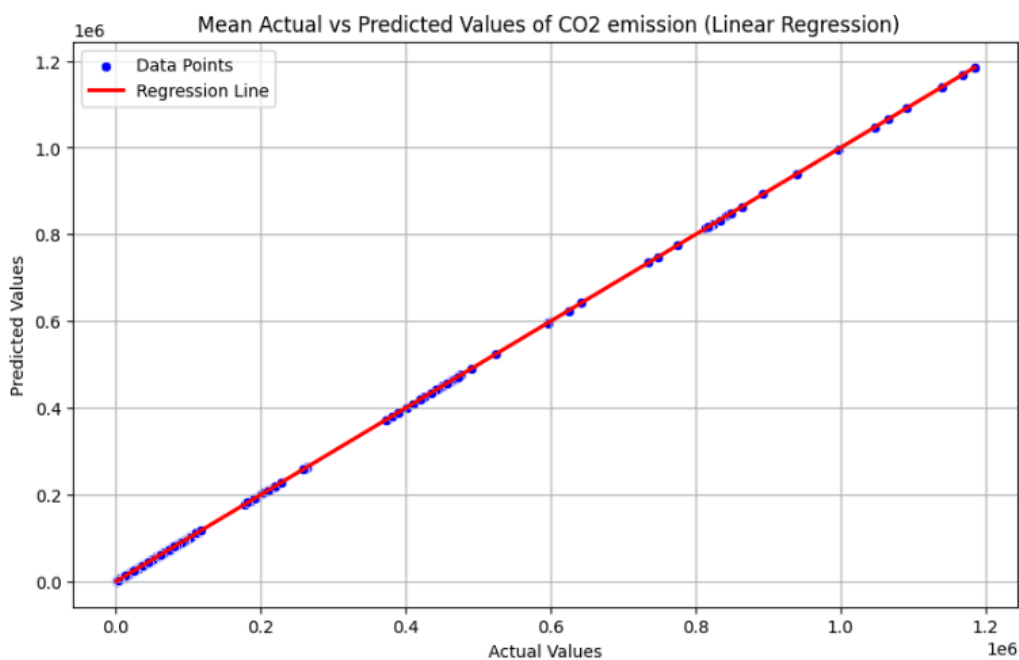


Figure 4.15: Actual vs Predict value line of CO2 emission for Linear Regression

The above graph shows the relationship between actual and predicted values of the Linear Regression model for CO2 emission. The actual values of CO2 emission are plotted

against the predicted values and marked by blue dots. The red line is the best fit perfect line which is derived from the linear regression model. The blue data points align almost over the red regression line, indicating that the model's predictions are very accurate with minimal deviation from the actual values. Hence the prediction errors (difference between actual and predicted values) are small with the reflect of a low residual error and good model performance.

- **For Temperature rise**

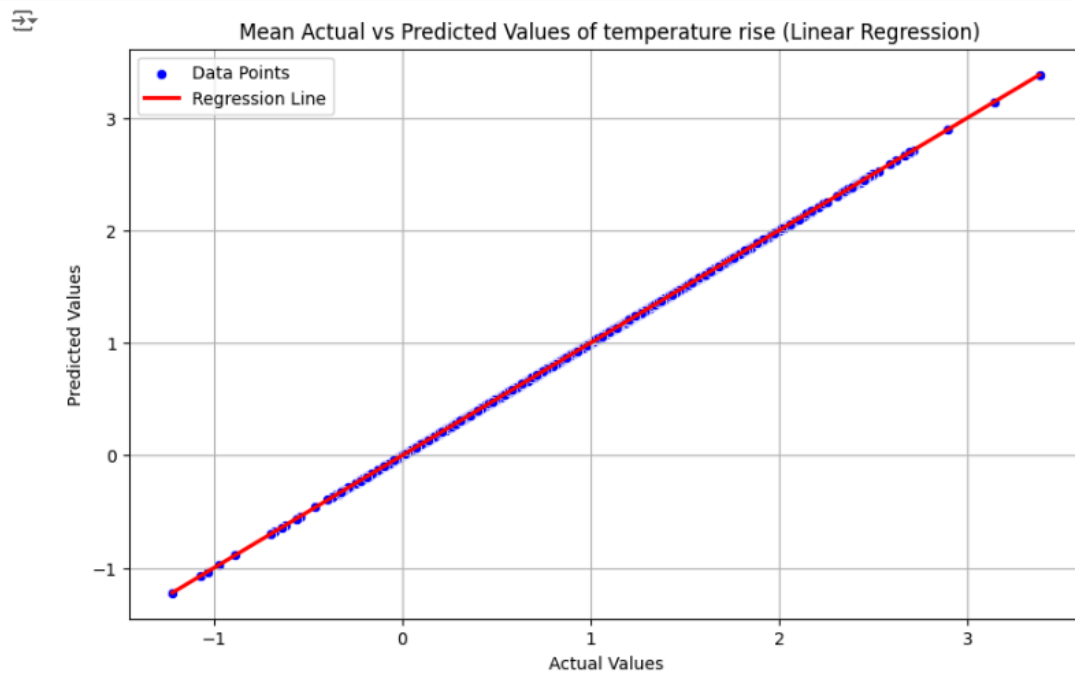


Figure 4.16: Actual vs Predict value line of temperature rise for Linear Regression

Similarly, for temperature rise the above graph shows the perfect accuracy with minimum deviation from the best fit red line. Hence the prediction errors (difference between actual and predicted values) are small with the reflect of a low residual error and provide accurate predictions over the dataset values.

5. CONCLUSION, LIMITATIONS AND FUTURE WORK

5.1. Conclusion

The primary objective of this applied research project was to develop and assess different predictive models for historic data of CO₂ emissions in the agri-food field and to conclude the best performing model among four different models like Linear regression, Random Forest, Decision Tree and Neural Networks by comparing the metrics like MSE, RMSE, MAE and R-squared values. The research revealed that from the calculated metrics and comparative analysis, Linear Regression had the highest accuracy with its lowest value of MSE, RMSE, MAE and highest value of R-squared in predicting CO₂ emissions and the corresponding temperature rise. This model outperformed other models which were less effective in capturing the relationships in the dataset. Global CO₂ emission and temperature rise from data were cleanly analyzed and showed the impact of agri-food activities on global warming and climate change. predictive models for CO₂ emission and temperature rise are built successfully to evaluate the different metrics from them. The evaluation result supports the hypothesis of this project that linear models can effectively help in predicting the CO₂ emissions and temperature rise from different countries.

The study's methodologies and results are on the same line with the previous research studies which highlights the importance of predicting emissions for strategic planning and environmental policy. Some of the research studies showed that the use of advanced machine learning models like neural networks provides best predictive capabilities and their ability to find complex non-linear relationships in data. However, findings from this study showed that neural

networks couldn't help much with its capability in prediction giving its worst results. This difference shows that advanced models may not always be suitable with straightforward variables or features in the dataset. This observation challenges the notion that it is critical to choose between models based on data characteristics and research aims.

There are significant results and benefits from the findings of this study for environmental policy and strategic planning . The Linear regression model stands out as a reliable tool with its accurate prediction of CO₂ emissions and temperature rise globally. This helps policymakers for forecasting and managing future emissions. Furthermore, the identification and implementation of effective strategies by stakeholders that target emissions can reduce global temperature increase.

To conclude, this research successfully implemented data mining techniques for predictive analysis of CO₂ emission from agri-food activities with the historic data from the globe. Further the study showed the relationship between CO₂ emission, agri-food activities and temperature rise and identified Linear Regression as the most robust model in prediction of CO₂ emissions and corresponding temperature rise in the agri-food sector. This implies that reliable predictive models are crucial for making successful policy and climate strategies.

5.2. Limitations

Despite the best performance of the selected models, there are some limitations in prediction of CO₂ emissions. Reliance on historical data stands as a major limitation, which may not capture the live and future changes in agri-food practices, leading to potential biases in the predictions. Also recent data from the year 2020 to 2024 is not readily available and hard to find.

Simple models like Linear regression might not handle the complexity of CO₂ emissions in all the scenarios. Since this project is only considering historical data of internal factors, models ability is limited in prediction of CO₂ emission. External factors like economic trends, climate policies, and advancement in technology will help the prediction of CO₂ emission and temperature rise. Additionally, models' performance depends on the quality and granularity of the available data. There may be a risk of model overfitting in this project due to specific dataset used which means models perform well on the training data, but not on new and unseen data. This may affect the application of models in the real world, where conditions may differ from those in the training data set.

5.3. Future Work

The used and developed predictive models of this study can be further refined and applied in real world policy and decision making strategies. Testing the developed models in practical scenarios by collaborating with government bodies, policymakers and industry stakeholders. This involves evaluating the impact of proposed regulations and the effectiveness of sustainability initiatives.

Future research could explore the use of hybrid models that combine the models with different approaches which will enhance predictive performances. Addition of external factors like economic indicators, climate policies, and advanced technologies in sustainable agriculture will provide a clear understanding of the agri-food activities. Future studies should concentrate more on improving the quality and granularity of the data used for the models. Using big data

sources like satellite imagery and IoT sensors could provide richer datasets that help in improvement of the models ability.

Finally, future researchers can work on validating and testing the selected models of this project with various types of dataset from the different regions, time periods and agricultural practices. This would allow us to measure the performance of models from various dimensions and identify the potential areas for improvement. Also, further studies can concentrate on testing new strategies to reduce CO₂ emissions and temperature rise in the agri-food sector. This involves practices like precision farming, understanding crop changes and adopting sustainable supply chains. These practices allow researchers to develop strategies and design better models to reduce emissions. Lastly, collaborating with the agriculture experts, policy makers and government bodies can help to make improvements in models which better capture the nature of CO₂ emissions. This also supports the development of creative ideas which helps in addressing the environmental, economic and social aspects of sustainability.

6. REFERENCES

1. Ahmed, S., Ahmed, K., & Ismail, M. (2020). Predictive analysis of CO₂ emissions and the role of environmental technology, energy use and economic output: evidence from emerging economies. *Air Quality, Atmosphere & Health*.
<https://doi.org/10.1007/s11869-020-00855-1>
2. Bussaban, K., Kularbphetpong, K., & Boonseng, C. (2023). Prediction of CO₂ Emissions Using Machine Learning. *CONNECT. International Scientific Conference of Environmental and Climate Technologies*, 129. <https://doi.org/10.7250/conect.2023.099>
3. Chowdhury, S., Rubi, M. A., & Bijoy, Md. H. I. (2021, July 1). Application of Artificial Neural Network for Predicting Agricultural Methane and CO₂ Emissions in Bangladesh. *IEEE Xplore*. <https://doi.org/10.1109/ICCCNT51525.2021.9580106>
4. El Bilali, H., Strassner, C., & Ben Hassen, T. (2021). Sustainable Agri-Food Systems: Environment, Economy, Society, and Policy. *Sustainability*, 13(11), 6260.
<https://doi.org/10.3390/su13116260>
5. Food and Agriculture Organization of the United Nations. (n.d.). Food and Agriculture Organization of the United Nations. [online] Available at: <https://www.fao.org/>
6. Guo, X., Yang, J., Shen, Y., & Zhang, X. (2023). Prediction of agricultural carbon emissions in China based on a GA-ELM model. *Frontiers in Energy Research*, 11.
<https://doi.org/10.3389/fenrg.2023.1245820>
7. Hamrani, A., Akbarzadeh, A., & Madramootoo, C. A. (2020). Machine learning for predicting greenhouse gas emissions from agricultural soils. *Science of the Total Environment*, 741, 140338. <https://doi.org/10.1016/j.scitotenv.2020.140338>

8. Hosseini, S. M., Saifoddin, A., Shirmohammadi, R., & Aslani, A. (2019). Forecasting of CO₂ emissions in Iran based on time series and regression analysis. *Energy Reports*, 5, 619–631. <https://doi.org/10.1016/j.egy.2019.05.004>
9. IPCC (2024). IPCC — Intergovernmental Panel on Climate Change. [online] [Ipcc.ch](https://www.ipcc.ch). Available at: <https://www.ipcc.ch/>
10. Ma, N., Shum, W. Y., Han, T., & Lai, F. (2021). Can Machine Learning be Applied to Carbon Emissions Analysis: An Application to the CO₂ Emissions Analysis Using Gaussian Process Regression. *Frontiers in Energy Research*, 9. <https://doi.org/10.3389/fenrg.2021.756311>
11. Maliha Homaira, & Hassan, R. (2021). Prediction of agricultural emissions in Malaysia using the arima, LSTM, and regression models. *International Journal on Perceptive and Cognitive Computing*, 7(1), 33–40.
12. Mangla, S.K., Luthra, S., Rich, N., Kumar, D., Rana, N.P. and Dwivedi, Y.K. (2018). Enablers to implement sustainable initiatives in agri-food supply chains. *International Journal of Production Economics*, [online] 203, pp.379–393. doi:<https://doi.org/10.1016/j.ijpe.2018.07.012>.
13. Mathew, A. (n.d.). Analyzing CO₂ Emission Intensity: A Comprehensive Study of Clean and Unclean Energy Sources using ML Techniques.
14. PREDICTING CO₂ EMISSION FROM POWER INDUSTRY USING MACHINE LEARNING Supervisor: Hamidreza Khaleghzadeh. (2024).
15. Sarfraz, M., Iqbal, K., Wang, Y., Bhutta, M. S., & Jaffri, Z. ul A. (2023). Role of agricultural resource sector in environmental emissions and its explicit relationship with

- sustainable development: Evidence from agri-food system in China. *Resources Policy*, 80, 103191. <https://doi.org/10.1016/j.resourpol.2022.103191>
16. Serafeim, G., & Velez Caicedo, G. (2022). Machine Learning Models for Prediction of Scope 3 Carbon Emissions. *SSRN Electronic Journal*.
<https://doi.org/10.2139/ssrn.4149874>
17. Shabani, E., Hayati, B., Pishbahar, E., Ghorbani, M. A., & Ghahremanzadeh, M. (2021). A novel approach to predict CO₂ emission in the agriculture sector of Iran based on the Inclusive Multiple Model. *Journal of Cleaner Production*, 279, 123708.
<https://doi.org/10.1016/j.jclepro.2020.123708>
18. Shearer, C. (Fall 2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, 5(4), 10.
19. Singh, P. K., Pandey, A. K., Ahuja, S., & Kiran, R. (2021). Multiple forecasting approach: a prediction of CO₂ emission from the paddy crop in India. *Environmental Science and Pollution Research*, 29(17), 25461–25472.
<https://doi.org/10.1007/s11356-021-17487-2>
20. Yeasmin, S., Shmais, L., Noor, S., Syed, J., & Al, R. (n.d.). Artificial Intelligence-based CO₂ Emission Predictive Analysis System.
<https://ebookcentral.proquest.com/lib/dbsie/reader.action?docID=6383434>
21. Zhu, Y., & Huo, C. (2022). The Impact of Agricultural Production Efficiency on Agricultural Carbon Emissions in China. *Energies*, 15(12), 4464.
<https://doi.org/10.3390/en15124464>

22. GeeksforGeeks. (2020). *Step by Step Predictive Analysis - Machine Learning*. [online] Available at:
<https://www.geeksforgeeks.org/step-by-step-predictive-analysis-machine-learning/>.
23. Kelleher, J.D., Mac Namee, B. and D'Arcy, A. (2020). *Fundamentals of machine learning for predictive data analytics : algorithms, worked examples, and case studies*. Cambridge, Massachusetts: The MIT Press.
24. Mckinney, W. (2017). *Python for data analysis : data wrangling with pandas, NumPy, and IPython*. Sebastopol, Ca: O'reilly Media, Inc., October.
25. Tutorialspoint.com. (2024). *Data Mining Tutorial*. [online] Available at:
https://www.tutorialspoint.com/data_mining/.
26. Python, R. (n.d.). *Python Machine Learning – Real Python*. [online] realpython.com. Available at: <https://realpython.com/tutorials/machine-learning/>
27. www.kaggle.com. (n.d.). *Agri-food CO2 emission dataset - Forecasting ML*. [online] Available at:
<https://www.kaggle.com/datasets/alessandrolobello/agri-food-co2-emission-dataset-forecasting-ml/data>.