MODELING AGRICULTURAL CONTRIBUTIONS TO CO_2 EMISSIONS: A COMPARATIVE STUDY OF MACHINE LEARNING APPROACHES

Project Submitted to the University of Kerala

in partial fulfilment of the requirements

of the Degree of Master of Science in Statistics

Submitted By

SANJAY RAJ.K.S

(Reg no : 98723615021)

Department of Statistics

University of Kerala

Thiruvananthapuram

2023-2025

DEPARTMENT OF STATISTICS

UNIVERSITY OF KERALA

Dr. Manoj Chacko
Professor and Head
Department of Statistics
University of Kerala

P. O. Kariavattom
Trivandrum-695 581
Phone: 0471-2308905

CERTIFICATE

This is to certify that the project entitled **"MODELING AGRICULTURAL CONTRIBUTIONS TO CO_2 EMISSIONS: A COMPARATIVE STUDY OF MACHINE LEARNING APPROACHES"** is a bonafide record of original work done by **SANJAY RAJ.K.S,** Master of Science in Statistics programme of Department of Statistics, University of Kerala, during 2023-2025 under my supervision and guidance and it is hereby approved for submission.

Dr. Manoj Chacko

July 2025

Thiruvananthapuram

DECLARATION

I hereby declare that the project report entitled **"IMPACT OF AGRICULTURAL ACTIVITIES ON CO_2 EMISSIONS: A STATISTICAL AND MACHINE LEARNING ANALYSIS"** is a bonafide record of original work done by me under the supervision and guidance of **DR. MANOJ CHACKO**, Professor, Department of Statistics, University of Kerala in partial fulfilment of the requirements of the Degree of **Master of Science in Statistics** of University of Kerala, Karyavattom and this work has not been formed the basis for the award of any other academic qualification, fellowship or any other similar title of any other University or Board.

SANJAY RAJ.K.S

July 2025

Thiruvananthapuram

ACKNOWLEDGEMENT

The successful completion of this project owes to the inspiration and constant support that I received from various sources. I avail this opportunity to express my sincere gratitude to all those who helped me directly or indirectly for the completion of work.

I wish to express my deep gratitude to my project guide DR. MANOJ CHACKO, Professor, Department of Statistics, University of Kerala for his valuable guidance throughout the period of this project work. His valuable and inspiring suggestions, constructive criticism, and moral support helped me in the successful completion of this work. I would like to thank all the teachers, the administration officer, the librarian, and research scholars of our department, my parents, and all my friends for their help and encouragement that they have rendered to me during the course of this project work.

Above all, I thank the Almighty for His blessings showered on us throughout the course of this project work.

SANJAY RAJ.K.S

Table of Contents

List of Tables

List of Figures

Chapter 1
INTRODUCTION

1.1 Introduction

Carbon dioxide (CO_2) emissions are one of the leading causes of global warming. They cause higher global temperatures, altered weather patterns, increased sea levels, and damage to ecosystems. Everyone is aware that cars, factories, and power plants are large emitters of these emissions, but agriculture is also a significant factor, even if one that usually does not receive the same level of attention.

Agriculture contributes to CO_2 and other greenhouse gases in several different ways. Deforestation to provide land for farms, burning crop residues, applying huge amounts of chemical pesticides and fertilizers, and raising livestock all emit noxious gases. And it's not just what occurs on the farm packaging, refrigeration, shipping, and retail sale, for example, also contribute emissions. The extent to which farming emits can vary greatly between nations. This is based on how agriculture is practiced, the economy of the country, land regulations, and even weather conditions.

This project examines how various aspects of agriculture contribute to CO_2 emissions in various nations over time. Drawing on data from across the globe, it seeks out patterns and correlations between agricultural practices and emission rates. It also examines population size, land area, temperature variation, and economic conditions to determine how these influence emissions.

The core focus of this study is to build and compare various machine learning models to analyse agricultural CO_2 emissions. Regression models are used to predict total CO_2 emissions, while classification models are applied to categorize average temperature

levels. By evaluating model performance based on appropriate metrics, the study aims to identify the most effective algorithms for each task, helping to improve predictive accuracy and support future environmental decision-making.

1.2 Objective of the study

- Analyse Agricultural Contributions to CO_2 Emissions: Investigate the role of various agricultural activities, including rice cultivation, manure management, and deforestation, in contributing to CO_2 emissions.
- Predict CO_2 Emissions Using Machine Learning: Apply and compare machine learning models (e.g., regression and classification) to accurately predict total agricultural CO_2 emissions and classify temperature levels.
- Examine Temporal Trends in Emissions: Study how agricultural emissions have changed over time, particularly from 1990 to 2020, and identify key shifts in emission patterns.
- Identify Key Emission Drivers: Determine the most significant agricultural activities contributing to emissions and understand how factors like population and temperature influence emission levels.
- Evaluate and Compare Model Performance: Assess the effectiveness of various machine learning models in predicting CO_2 emissions, using metrics like RMSE, MSE, and R-squared.
- Provide Policy Insights: Offer insights into sustainable agricultural practices and strategies for reducing emissions based on the findings.

1.3 Data and variables

This study uses a panel dataset that includes information on CO_2 emissions from different agricultural activities across several countries and years. The data covers a variety of sources such as savanna and forest fires, rice cultivation, manure use, food processing, and transport. It also includes important details like rural and urban population and average temperature increase. In this dataset, the agricultural sector is responsible for around 62% of the total CO_2 emissions, showing its major role in climate change. This makes the dataset very useful for understanding how farming and related activities affect the environment. By studying these variables, we can explore

patterns, identify the main sources of emissions, and build models to better explain and predict the impact of agriculture on climate change.

Source: The dataset is publicly available on the Kaggle website and was uploaded by Alessandro Lo Bello. It was created by combining and cleaning about a dozen different datasets from the Food and Agriculture Organization (FAO) and the Intergovernmental Panel on Climate Change (IPCC). The data shows agricultural CO_2 emissions from many countries over several years and includes different sources like land use, livestock, and food production. This dataset helps us better understand how farming activities contribute to climate change and can guide efforts to reduce emissions.

Dataset Features

- Area: Country or region name.
- Year: The calendar year of observation.
- Savanna fires: Emissions from fires in savanna ecosystems.
- Forest fires: Emissions from fires in forested areas.
- Crop Residues: Emissions from burning or decomposing leftover plant material after crop harvesting.
- Rice Cultivation: Emissions from methane released during rice cultivation.
- Drained organic soils (CO2): Emissions from carbon dioxide released when draining organic soils.
- Pesticides Manufacturing: Emissions from the production of pesticides.
- Food Transport: Emissions from transporting food products.
- Forestland: Land covered by forests.
- Net Forest conversion: Change in forest area due to deforestation and afforestation.
- Food Household Consumption: Emissions from food consumption at the household level.
- Food Retail: Emissions from the operation of retail establishments selling food.

- On-farm Electricity Use: Electricity consumption on farms.
- Food Packaging: Emissions from the production and disposal of food packaging materials.
- Agrifood Systems Waste Disposal: Emissions from waste disposal in the agrifood system.
- Food Processing: Emissions from processing food products.
- Fertilizers Manufacturing: Emissions from the production of fertilizers.
- IPPU: Emissions from industrial processes and product use.
- Manure applied to Soils: Emissions from applying animal manure to agricultural soils.
- Manure left on Pasture: Emissions from animal manure on pasture or grazing land.
- Manure Management: Emissions from managing and treating animal manure.
- Fires in organic soils: Emissions from fires in organic soils.
- Fires in humid tropical forests: Emissions from fires in humid tropical forests.
- On-farm energy use: Energy consumption on farms.
- Rural population: Number of people living in rural areas.
- Urban population: Number of people living in urban areas.
- Total Population - Male: Total number of male individuals in the population.
- Total Population - Female: Total number of female individuals in the population.
- Total emission: Total greenhouse gas emissions from various sources.
- Average Temperature °C: The average increasing of temperature (by year) in degrees Celsius,

1.4 Summary of the project

The objective of this study is to examine the role of agriculture in contributing to CO_2 emissions, with a particular focus on identifying the major sources of emissions, such as rice cultivation, manure management, and deforestation, and understanding their impact on climate change. The study aims to explore the relationship between agricultural emissions and key factors such as population density (urban vs. rural), land area, average temperature, and socio-economic indicators across different countries and regions. By applying and comparing various machine learning models, including regression and classification algorithms, the study seeks to predict total CO_2 emissions from agricultural activities and classify temperature changes, while evaluating model performance using metrics like R-squared, RMSE, and accuracy. The research also aims to analyse temporal trends and regional variations in agricultural emissions from 1990 to 2020, with a focus on countries with significant agricultural emissions and land use changes. Furthermore, the study will assess the effectiveness of different machine learning models, such as Random Forest, XGBoost, and Multiple Linear Regression, to estimate agricultural CO_2 emissions, comparing their accuracy, precision, and robustness. Ultimately, the study aims to provide actionable insights and recommendations for reducing agricultural CO_2 emissions through sustainable farming practices and better resource management, offering data driven support for policymakers to develop strategies for climate change mitigation.

Chapter 2
LITERATURE REVIEW

2.1 Background and Context

Carbon dioxide (CO_2) emissions are a primary driver of global climate change and have been a subject of extensive academic and policy-oriented research. With industrialization and urbanization accelerating in developing economies particularly in China the relationship between economic growth and environmental degradation has become increasingly complex. Traditional econometric models have offered partial insights, often constrained by assumptions of linearity and multicollinearity. In contrast, recent advancements in machine learning have enabled the modelling of high-dimensional, non-linear systems, offering more nuanced analyses of CO_2 emission dynamics. This chapter explores past studies on the determinants of CO_2 emissions, the environmental impact of industrial and urban growth, and the emerging use of artificial intelligence in predictive environmental modelling.

2.2 Economic, Industrial, and Urban Drivers of Emissions

Past studies have widely examined how economic growth, industrial structure, and urbanization contribute to CO_2 emissions. The Environmental Kuznets Curve (EKC), proposed by Grossman and Krueger (1995), suggested an inverted U-shaped relationship between economic development and environmental degradation. However, empirical findings have varied. For example, Holtz-Eakin and Selden (1995) and Friedl and Getzner (2003) supported the inverted U-shape, while others such as

Shafik (1994) and Murshed and Dao (2020) observed a direct positive correlation between income and emissions. On industrial structure, Bernardini and Galli (1993) found that as economies moved from agriculture to heavy industries and finally to service-oriented sectors, energy intensity initially rose and later declined. Similarly, urbanization has been shown to influence CO_2 emissions by increasing energy demand, transportation needs, and land-use changes. York et al. (2003) and Wang et al. (2012) demonstrated that urban growth directly correlates with higher emissions, although the magnitude varies based on regional development patterns.

2.3 Technology, Investment, and Environmental Innovation

Technological progress has consistently been cited as a mitigating factor for CO_2 emissions. Lin and Du (2014) argued that R&D investment improved energy efficiency and reduced emissions intensity. Ang (2009) highlighted the role of endogenous technological change through R&D and trade openness. Other studies, such as Wei et al. (2010), found that foreign direct investment (FDI) and technology imports played key roles in transferring cleaner technologies to developing countries. However, the pollution haven hypothesis complicates this narrative, as suggested by Candau and Dienesch (2017), who noted that foreign investments may be attracted to regions with lax environmental regulations. Thus, while technological advancement can be a protective factor, its effects are often mediated by governance quality and regulatory frameworks.

2.4 Machine Learning Applications in Emissions Forecasting

The limitations of linear econometric models have driven a growing interest in machine learning for forecasting CO_2 emissions. Li et al. (2021) applied multiple machine learning algorithms including linear regression, artificial neural networks (ANN), ensemble models, and k-nearest neighbors (KNN)—to assess CO_2 emission trends in China between 2000 and 2018. The KNN model outperformed others in predictive accuracy, based on root mean square error (RMSE). Additionally, sensitivity analyses using this model revealed optimal industrialization and urbanization levels for minimizing emissions in specific provinces. These findings underscored the value of machine learning in uncovering localized, non-linear relationships. Moreover,

interpretability tools like SHAP were recommended to enhance model transparency for policymaking, although they were not explicitly used in this study.

2.5 Research Gaps and Study Justification

Despite significant advancements in modelling CO_2 emissions, key gaps remain in the existing literature:

- Overemphasis on limited variables: Most prior studies focused on single-factor or bivariate relationships (e.g., GDP vs. CO_2 emissions) and often neglected multi-factorial frameworks that integrated variables such as urbanization, technological development, foreign investment, and energy consumption.
- Limited use of ensemble machine learning models: Although some studies applied individual machine learning techniques, few utilized ensemble models such as XGBoost or LightGBM to assess CO_2 emissions across provinces or countries comprehensively.
- Lack of model interpretability: Many predictive models prioritized accuracy but failed to address transparency. Even though models like K-nearest neighbors (KNN) achieved strong predictive performance, their lack of interpretability limited their relevance for policy-making.
- Absence of policy-oriented indices: There was limited effort to develop integrated forecasting frameworks that connected model outputs with actionable indices accounting for regional disparities and threshold effects, which could better inform targeted environmental policies.

Chapter 3
METHODOLOGY

3.1 Introduction

This chapter outlines the materials and techniques employed in the study. It is divided into three sections for better clarity and organization. The first section introduces the key terms, variables, and data handling methods used in the study. It describes the dataset structure, including how the data was collected, the nature of each variable, and the initial steps taken to prepare the data for analysis such as cleaning and transformation. The second section focuses on the statistical techniques and tools applied to analyse the data. This includes descriptive statistics to summarize the data, correlation analysis to explore relationships between variables, and machine learning methods for predictive modelling and pattern identification. The final section highlights the statistical software and programming tools used in the study. It discusses the platforms and libraries chosen such as R or Python and explains how they were used to implement the analysis and interpret the results efficiently.

.

3.2 Software Tools Used

In this study, the following software tools and libraries were used to process data, build models, and visualize results:

- Python: The primary programming language used for data analysis and machine learning.
    - Pandas: Used for data manipulation and cleaning.
    - NumPy: For numerical computations and handling arrays.

- Matplotlib and Seaborn: For data visualization (e.g., line charts, bar charts, heatmaps).
- Scikit-learn: For machine learning models (e.g., Multiple Linear Regression, Decision Trees, Random Forest, etc.).
- XGBoost: For implementing the XGBoost model.
- SciPy: Used for statistical analysis and hypothesis testing.

- R Language: Used for Descriptive Statistics and additional statistical analysis and visualizations.
- Microsoft Excel: Used for initial data cleaning, preparation, and summary statistics.

3.3 Descriptive Statistics

Descriptive statistics plays a crucial role in summarizing and organizing information to reveal key patterns and insights. This technique focuses on providing a clear overview of the characteristics of a dataset by calculating measures such as mean, median, standard deviation, and range. By using these measures, we can determine central tendencies, assess the variability of the data, and identify any outliers or anomalies. The goal of descriptive statistics is to provide a comprehensive understanding of the distribution and spread of values, which forms the foundation for further analysis. In addition to numerical summaries, graphical representations like histograms, bar charts, and box plots are often employed to visually interpret these characteristics, offering a more intuitive grasp of the data's structure.

3.4 Data Visualization

Data visualization is an essential technique used to present complex information in a clear and accessible manner. It involves the use of graphical tools to represent data visually, allowing for a more intuitive understanding of patterns, trends, and relationships. By transforming raw data into visual formats such as line charts, bar graphs, pie charts, and heatmaps, data visualization helps in uncovering insights that

might not be immediately apparent from raw numbers alone. This approach not only enhances data comprehension but also aids in identifying key features, comparing variables, and spotting anomalies or outliers. Effective data visualization plays a critical role in communicating findings, enabling stakeholders to interpret the data quickly and make informed decisions based on visual representations.

3.4.1 Line Chart

A line chart is a graphical representation used to display information that changes over time. It is one of the most common types of charts used in data visualization to show trends, patterns, and relationships between variables. A line chart typically consists of points connected by straight lines, with each point representing a data value at a specific time or interval. The x-axis usually represents time, while the y-axis shows the variable being measured. Line charts are particularly effective for illustrating continuous data and understanding how a variable evolves or fluctuates over a period. They are widely used in fields such as economics, science, and business to visualize trends and forecast future patterns.

3.4.2 Bar Chart

A bar chart is a visual representation used to compare quantities or frequencies of different categories. It consists of rectangular bars, where the length or height of each bar is proportional to the value or frequency it represents. The categories are typically displayed on the x-axis, while the y-axis represents the numerical values. Bar charts are particularly useful for comparing discrete data, making them ideal for showing comparisons between different groups, such as population size, sales figures, or emission levels. They can be presented in either horizontal or vertical orientation, depending on the data's nature. Bar charts are widely used in various fields, including business, economics, and research, to easily highlight differences across categories and identify trends.

A grouped bar chart is a variation where multiple bars are placed side by side for each category, allowing for comparison of different sub-groups or variables within that category. It’s useful for comparing multiple sets of data, such as sales figures across different years or regions.

3.4.3 Heatmap

A heatmap is a data visualization technique that uses varying colours to represent data values, ranging from low to high. This graphical tool is especially useful when analysing complex datasets, as it highlights detailed information through colour gradients. Typically, heatmaps employ a range of colours, with warmer colours like red indicating higher values, while cooler colours like blue represent lower values. This method of visualization is widely used in fields such as statistics, data analysis, and biology, as it allows for a more intuitive understanding of relationships and trends within the data compared to traditional numerical tables. Heatmaps simplify the interpretation of large volumes of data by visually pointing out areas of significance or concern, making it easier to detect patterns or anomalies at a glance.

3.4.4 Stacked Area Chart

A stacked area chart is a data visualization tool that displays the cumulative total of multiple data series over time or categories, emphasizing the contribution of each individual category to the overall total. It is similar to an area chart but with the added feature of stacking the areas on top of one another. Each area represents a different category, and the height of each stacked segment reflects its value at a given point. This type of chart is useful for visualizing how different categories contribute to a total over time, showing both the overall trend and the individual trends of each category. Stacked area charts are commonly used in fields like business, economics, and environmental studies, where it's important to visualize the breakdown of a total value and its changes over time or across different groups.

3.4.5 Box Plot

A box plot, also known as a box-and-whisker plot, is a graphical representation that displays the distribution of a dataset based on five key summary statistics: the minimum, first quartile (Q1), median, third quartile (Q3), and maximum. The plot consists of a rectangular "box" that spans from the first to the third quartile, with a line (or "whisker") extending from each end of the box to show the range of the data. The

median is typically marked inside the box. Box plots are useful for identifying the central tendency, spread, and variability of data, as well as for detecting outliers. They provide a clear visual summary of the distribution, making it easy to compare multiple datasets or understand the overall range and distribution of values. Box plots are widely used in statistical analysis and are particularly helpful for visualizing large datasets with multiple variables.

3.4.6 Pie Chart

A pie chart is a circular data visualization tool used to represent the proportions of a whole, with each segment or "slice" corresponding to a category's relative percentage of the total. The entire circle represents 100% of the data, and the slices are scaled according to their contribution to that total. Pie charts are effective for illustrating simple parts-to-whole relationships and are commonly used when there are a limited number of categories to compare. They are particularly useful for showing how different components contribute to the total, making it easier to understand the relative size of each category. Pie charts are widely used in business, marketing, and social sciences to present categorical data in a visually appealing and easy-to-understand manner. However, they may not be suitable for displaying complex datasets with too many categories, as the slices can become hard to differentiate.

3.4.7 Subplot

A subplot is a smaller plot within a larger figure that allows multiple visualizations to be displayed together in a single layout. Each subplot can represent a different data set or variable, with its own axes, titles, and labels. This arrangement makes it easier to compare and analyse multiple plots simultaneously, especially when dealing with related data or trends across different categories. Subplots are often used to organize complex visualizations in a compact, clear, and easy-to-read format.

3.5 Machine Learning Models

Machine learning (ML) is a subset of artificial intelligence (AI) focused on developing algorithms that enable systems to learn from data and make decisions based on this learning. Unlike traditional programming, where rules are explicitly written by

the programmer, machine learning algorithms allow computers to detect patterns and insights from data, improving over time without human intervention. For instance, in natural language processing, ML models can understand and respond to previously unheard combinations of words, accurately interpreting the intent behind them.

Key Components of Machine Learning:

1. **Data**: The foundation of machine learning, data can be structured (e.g., databases) or unstructured (e.g., text, images). The quality and quantity of the data significantly influence the model's performance.
2. **Features**: Input variables or attributes used by the model to make predictions. Feature engineering involves selecting, modifying, or creating features to enhance model performance.
3. **Labels**: The target variables the model aims to predict. In supervised learning, the labels are provided during the training phase.
4. **Algorithms**: Mathematical models or procedures that help the system learn from data. Different algorithms are suited for different types of problems.

Steps in the Machine Learning Life Cycle:

1. **Gathering Data**: The first step involves identifying and obtaining relevant data. This stage focuses on addressing data-related challenges.
2. **Data Preparation**: This step involves organizing and preparing the data for machine learning. It includes gathering data, randomizing the dataset, and performing data exploration to understand its nature.
3. **Data Wrangling**: The process of cleaning the data, selecting relevant variables, and transforming the dataset into a suitable format for analysis. Issues such as missing values, duplicate data, and noise are handled during this stage to ensure accurate outcomes.
4. **Data Analysis**: This step involves determining the type of problem (e.g., classification, regression, clustering) and selecting the appropriate machine learning technique. The model is then built using the prepared data and evaluated.

5. **Train Model**: In this step, the model is trained using the dataset. Training is essential for the model to learn various patterns, rules, and features in the data, improving its performance.

6. **Test Model**: After training, the model is tested using a test dataset to evaluate its accuracy. Testing helps determine how well the model performs and whether it meets the project requirements.

7. **Deployment**: The final step involves deploying the model into a real-world system. If the model provides accurate results with acceptable speed, it is integrated into the operational environment. This phase is similar to preparing a final report for the project, where the model is implemented and monitored for its ongoing performance.

The ML life cycle ensures that the model is robust, accurate, and able to perform well in real-world applications

Supervised Machine Learning

Supervised machine learning is a type of machine learning where the model is trained on a labelled dataset. In this approach, the algorithm is provided with input-output pairs, where the output (label) is already known. The model learns to map inputs to the correct outputs by identifying patterns in the data. Once trained, the model can predict the output for new, unseen data. Supervised learning is typically used for tasks like classification and regression.

Unsupervised Machine Learning

Unsupervised machine learning is a type of machine learning where the model is trained on an unlabelled dataset. In this case, the algorithm is provided with data without corresponding output labels. The goal is to discover hidden patterns, relationships, or structures within the data. The model learns from the input data without any guidance on what the expected output should be. Unsupervised learning is commonly used for tasks like clustering and dimensionality reduction.

Semi-Supervised Machine Learning

Semi-supervised machine learning lies between supervised and unsupervised learning. It involves training the model on a dataset that is partially labelled, with a

small portion of the data containing labels and the majority remaining unlabelled. This approach helps leverage the large amount of unlabelled data while still using the labelled data to guide the learning process. Semi-supervised learning is often used when labelling data is costly or time-consuming.

Regression Model

A regression model in machine learning is a supervised learning algorithm used to predict a continuous target variable based on one or more input features. The model learns the relationship between the features and the target variable from the training data, and it makes predictions on new, unseen data by applying this learned relationship.

3.5.1 Multiple Linear Regression

Multiple Linear Regression is a supervised learning algorithm for predicting a continuous target variable based on multiple independent features. It assumes a linear relationship between the target and the predictors, fitting a linear equation (hyperplane) to the data. The model aims to minimize the sum of squared residuals between the actual and predicted values. It's simple to interpret but may not capture complex relationships between features.

3.5.2 Decision Tree Regression

Decision Tree Regression is a type of regression model that splits the data into subsets based on feature values. Each internal node represents a test on a feature, and each leaf node gives a prediction for the target variable. The tree is constructed by recursively splitting the data to reduce the variance in the target variable. It's simple to understand but can overfit the data if not properly pruned or regularized.

3.5.3 Random Forest Regression

Random Forest Regression is an ensemble technique that creates multiple decision trees using random sampling of the data and features. The final prediction is made by averaging the predictions from each tree. It helps reduce the overfitting risk commonly associated with individual decision trees, making it more robust and

accurate for regression tasks. However, it can be computationally intensive and may be less interpretable than a single decision tree.

3.5.4 Gradient Boosting Regression

Gradient Boosting Regression is an ensemble method where trees are built sequentially to correct the errors made by previous trees. Each subsequent tree focuses on reducing the residual errors from earlier trees. This boosting approach often yields high predictive accuracy, but it can be prone to overfitting, especially with many trees, and is computationally expensive.

3.5.5 XGBoost Regression

XGBoost (Extreme Gradient Boosting) is an optimized version of gradient boosting that improves upon traditional methods with regularization techniques to prevent overfitting and a more efficient algorithm for faster computation. It's highly popular for its performance on large datasets and is often the go-to algorithm for many machine learning competitions.

3.5.6 KNN Regression

K-Nearest Neighbors (KNN) Regression is a non-parametric method where the prediction for a data point is based on the average (or weighted average) of the target values of its k-nearest neighbors in the feature space. The choice of k and the distance metric significantly affect performance. KNN is simple and intuitive, but it may struggle with high-dimensional data and large datasets, requiring significant computation at prediction time.

Classification Model

A classification model in machine learning is a supervised learning algorithm used to predict a categorical target variable based on one or more input features. The model learns the relationship between the features and the target variable from the training data and assigns new data points to one of the predefined classes or categories.

3.5.7 Decision Tree

Decision Tree is a supervised learning algorithm used for classification tasks. It recursively splits the data into subsets based on the feature that provides the best separation between classes (usually based on Gini impurity or entropy). Each internal node represents a feature test, and each leaf node represents a class label. Decision Trees are easy to interpret but are prone to overfitting, especially with deep trees.

3.5.8 Random Forest

Random Forest is an ensemble method that builds multiple decision trees and combines their predictions. Each tree is trained on a random subset of data, and a random subset of features is used at each split. The final prediction is made by taking a majority vote from all trees. Random Forest improves the stability and accuracy of decision trees and reduces the risk of overfitting, but it can be computationally intensive.

3.5.9 KNN (K-Nearest Neighbors)

K-Nearest Neighbors (KNN) is a simple classification algorithm where the class of a data point is determined by the majority class among its k-nearest neighbors. The algorithm is non-parametric and does not require training. However, KNN can be computationally expensive at prediction time, especially with large datasets, and is sensitive to irrelevant features and the choice of k.

3.5.10 SVM (Support Vector Machine)

Support Vector Machine (SVM) is a supervised learning algorithm used for both classification and regression tasks. For classification, it finds the hyperplane that best separates data points of different classes in a high-dimensional feature space. SVMs can handle non-linear decision boundaries using kernel functions. The algorithm is effective in high-dimensional spaces and for cases where there is a clear margin of separation, but it can be memory-intensive and may require careful tuning of parameters.

Chapter 4
DESCRIPTIVE STATISTICS

4.1 Introduction

Descriptive statistics refers to a set of methods used to summarize, organize, and present data in a clear and concise way. It focuses on describing the main features and characteristics of a dataset without making conclusions beyond the data itself. The main goal is to provide a straightforward summary, which typically includes measures of central tendency such as the mean, median, and mode; measures of spread like range, variance, and standard deviation; and the shape of the data distribution through skewness and kurtosis. Descriptive statistics also use visual tools such as charts, graphs, and tables to help understand and interpret the data more easily. Common graphical techniques include histograms, bar charts, pie charts, scatter plots, and box plots.

MAIN PURPOSE OF DESCRIPTIVE STATISTICS

- To help better understand the agricultural emissions data and build a foundation for further analysis.
- To provide basic information about the different variables in the dataset, such as emission sources and climate indicators.
- To identify possible patterns or relationships between agricultural activities and CO_2 emissions.
- To assess the scope and scale of the emissions data across different variables.

4.2 Summary Statistics for Agricultural Emissions Activities and Total Emissions

Table 2.1 : Summary Statistics for Agricultural Emissions

Variable	Mean	Median	Standard Deviation	Minimum	Maximum
Crop Residues	998.71	103.70	3700.35	0.0002	33,490.07
Rice Cultivation	4259.67	534.82	17613.83	0.0000	164,915.30
Manure applied to Soils	923.23	120.44	3226.99	0.0490	34,677.36
Manure left on Pasture	3518.03	972.57	9103.56	0.0007	92,630.76
Manure Management	2263.34	269.86	7980.54	0.4329	70,592.65
Total Emission	64091.24	12147.65	228312.96	-391884.06	3115114.00

The table above shows the range and average levels of emissions from different farming activities, along with the total emissions. In almost every case, the average value is much higher than the middle value (median), which means that most countries or regions have low emissions, but a few have very high levels that increase the overall average. For example, emissions from crop residues have an average of about 999, but the median is only around 104. This pattern is even more noticeable in rice cultivation, where the average is over 4,200, but the median is just about 535. Such differences suggest that emissions vary a lot between places or over time.

Similar trends can be seen in emissions from manure-related activities. For example, manure left on pasture has a high average of 3,518, but the median is much lower at around 973, and the values range up to more than 92,000. This shows that while most data points are relatively low, a few regions have very large emissions. The total emissions column also shows a big gap between the average and the median, with an extremely wide range overall. Interestingly, there is even a negative value, which may be due to an error in the data or could reflect a situation where more carbon was

absorbed than released. These differences and extreme values suggest that the data may need to be adjusted or transformed before using it in further analysis, so that the results are more accurate and reliable.

4.3 Distribution of Observations by Emission Range and Average Temperature

Table 2.2 : Emission Range and Temperature Distribution

Emission Range	Average Temperature (°C)	Observation Count	Percentage (%)
<1000	0.973517	53	0.79
1000 - 2000	0.818604	249	3.70
2000 - 5000	0.796036	1105	16.43
>5000	0.890160	5320	79.08

This table groups emission levels and looks at how they relate to average temperature increases. Surprisingly, the group with the lowest emissions (less than 1000) shows the highest average temperature, around 0.97°C, even though it includes only a small number of cases (just 0.79% of the data). In contrast, the largest group, which includes areas with very high emissions (above 5000), makes up over 79% of the observations and shows a slightly lower average temperature of about 0.89°C.

The middle two groups those with emissions between 1000 and 5000 have the lowest average temperatures, ranging from 0.80°C to 0.82°C. This pattern suggests that the connection between emissions and temperature is not always direct or simple. While we might expect higher emissions to always lead to higher temperatures, other factors like local climate conditions, natural carbon absorption, or differences in how emissions are managed could also play a role. This reminds us that climate patterns

are complex, and looking at data in groups like this helps uncover some of that complexity.

4.4 Population Distribution by Gender

Table 2.3 : Population Distribution by Gender

Gender	Total Count	Percentage (%)
Total Population - Male	122,720,720,341	50.42
Total Population - Female	120,664,928,634	49.58

This table shows the total population divided by gender. The overall distribution is nearly balanced, with males making up 50.42% and females accounting for 49.58% of the total population. The male population count is approximately 122.7 billion, while the female population is slightly lower at around 120.7 billion. Though the difference in percentage is small, this slight imbalance may reflect demographic trends commonly observed across countries, such as higher male birth rates or gender-based differences in population reporting. The near-equal distribution also suggests that both male and female population dynamics are likely to influence patterns in agricultural activity, resource use, and emission contributions, making gender an important demographic factor to consider in broader environmental analyses.

4.5 Gender-wise Total Population and Percentage Distribution

Table 2.4 : Gender-wise Population Distribution

Population Type	Total Count	Percentage (%)
Rural Population	124,379,127,014	51.33
Urban Population	117,933,448,426	48.67

This table presents the total population divided by location type: rural and urban. The rural population accounts for 51.33% of the total, with a count of approximately 124.4 billion people, while the urban population makes up 48.67%, totalling around 117.9 billion people. This slight majority of rural residents reflects the continued importance of rural areas, particularly in the context of agriculture and land-based livelihoods. Given that agriculture is a major source of emissions in this study, the higher rural population may also point to a greater potential impact of rural communities on land use, resource consumption, and emission patterns. The close split between rural and urban populations also suggests that both types of areas should be considered when evaluating policy or sustainability efforts related to emissions and climate change.

Chapter 5
DATA VISUALISATION

5.1 Introduction

This section presents a visual exploration of agricultural CO_2 emissions using Python-based tools, aimed at uncovering key trends, regional differences, and contributing factors in a clear and accessible way. The analysis begins with a global overview of total emissions from 1990 to 2020, illustrating how emissions have generally increased over time, with specific years showing fluctuations tied to shifts in agricultural practices and economic factors.

A focused analysis on India follows, offering insights into the country's emissions trends during the same period, highlighting national patterns and factors driving changes in emissions. The study also compares countries with the largest and smallest agricultural emission-related land areas, revealing regional differences in land use and the varying growth of emissions across different regions.

To further explore the relationships between emissions and influencing factors, a correlation heatmap is presented, showing how agricultural, demographic, and environmental variables (such as temperature, crop residues, and population density) interact with total emissions.

All visualizations were created using Python libraries such as Matplotlib and Seaborn, which helped present the data clearly and effectively. These visuals not only support the statistical analysis but also offer actionable insights that can inform sustainable agriculture practices and guide policy development for reducing agricultural CO_2 emissions in the future.

5.2 Total Emission All Over Years

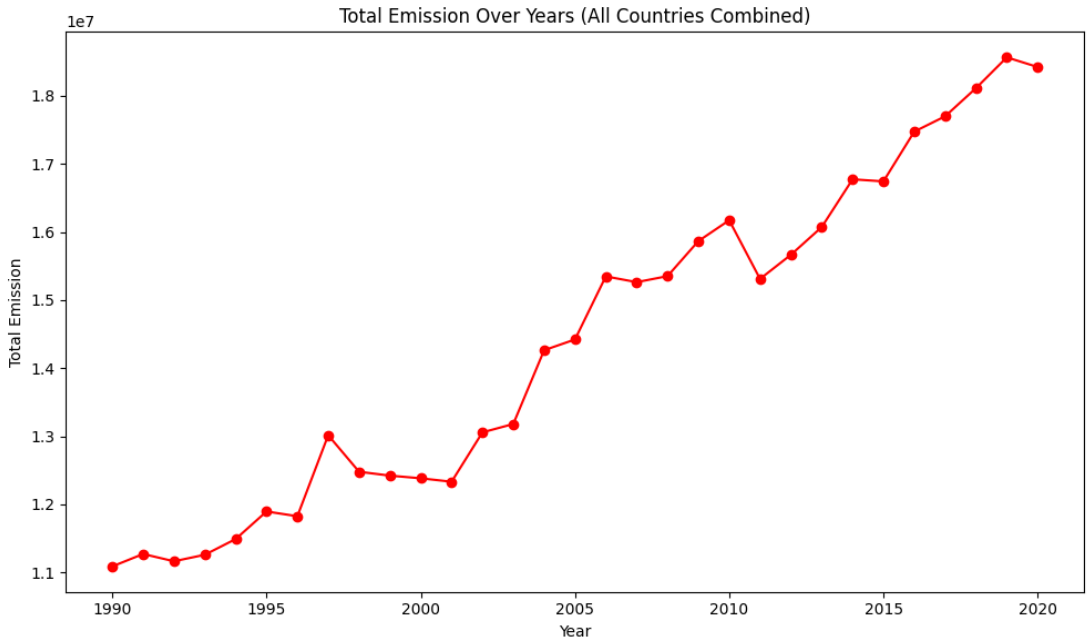

Figure 2.1: Total Emissions Over the Years

Overall agricultural CO_2 emissions exhibited a steady increasing trend between 1990 and 2020 for all countries combined. Beginning around 11 million units in 1990, emissions have continued to rise year after year, reaching over 18.5 million units in 2019. While the general trend is upward, some years showed minor decreases or levelling off, for example, around 1998, 2011, and 2014. These variations can be indicative of shifts in farming practices, climate fluctuation, financial conditions, or policy initiatives. Yet the overall trend shows constant growth in emissions, which points towards increasing environmental pressure through agriculture. The trend buttresses the necessity of sustainable farm practices and gives a sound basis for more advanced statistical modelling and impact assessment with respect to climate.

5.3 Total Emission Trend Over Years For India

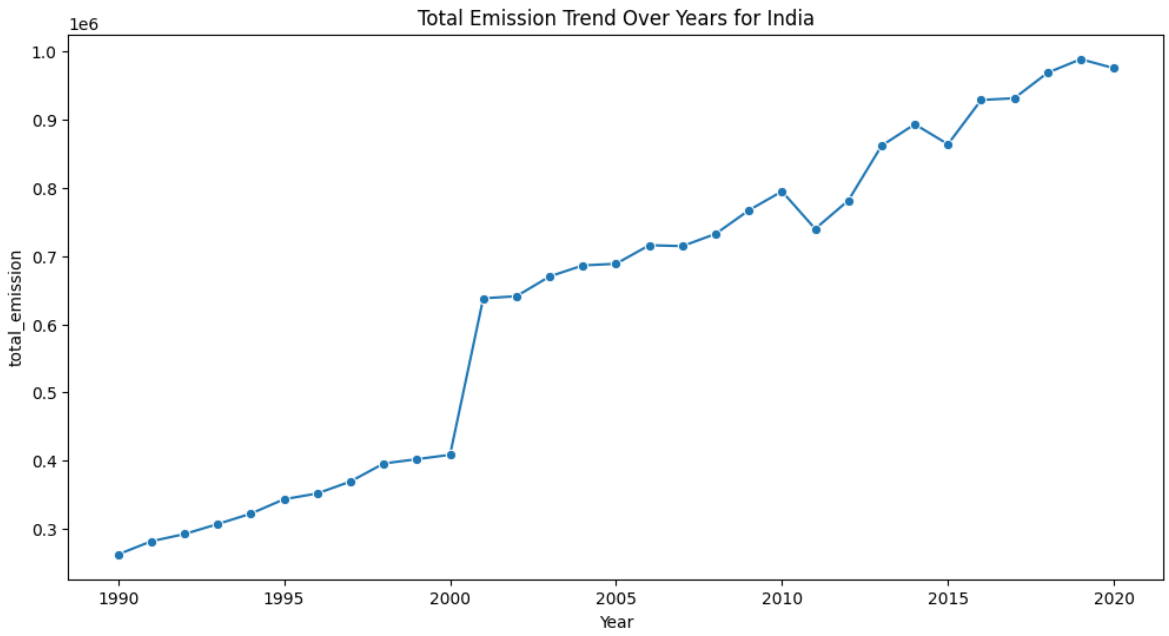

Figure 2.2: Emission Trend in India (1990-2020)

The graph above shows how total agricultural emissions in India have changed from 1990 to 2020. In the early 1990s, emissions were low, around 260,000 units. They increased slowly through the 1990s. Around 2001, there was a sudden jump, with emissions rising sharply to over 640,000 units. After that, emissions kept increasing, although there were some ups and downs between 2010 and 2015. By 2019, emissions reached their highest point, close to 1,000,000 units. There was a small drop in 2020, but overall, the trend shows a strong and steady rise over time. This increase suggests that agricultural activities in India have grown over the years, possibly due to more use of fertilizers, larger crop production, and a rise in livestock numbers.

5.4 Countries with the Highest Agricultural Emission Associated Land Area

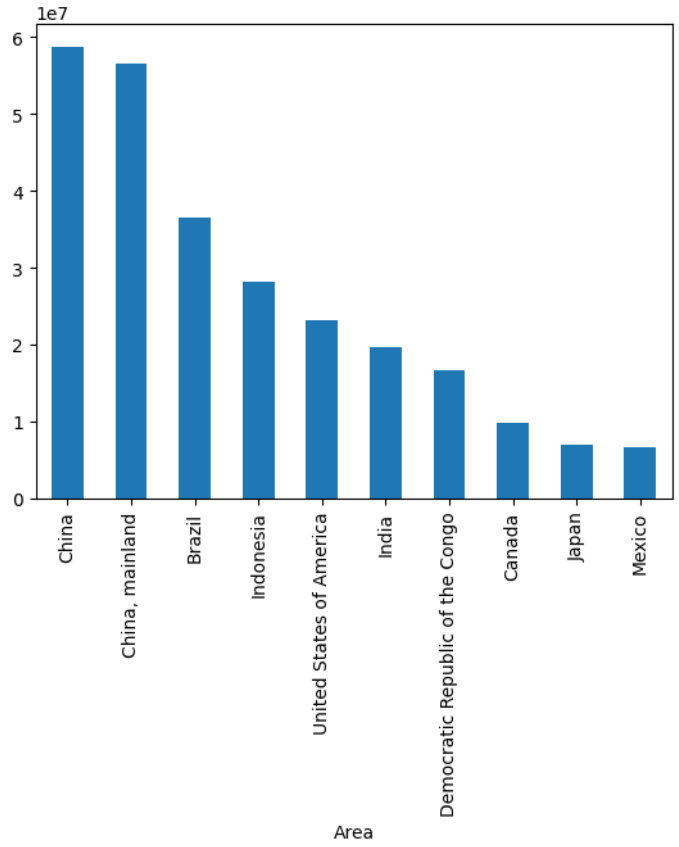

Figure 2.3: Countries with Highest Agricultural Emission Land Area

The chart above shows the top countries with the largest areas related to agricultural emissions. China has the highest value, and it appears twice in the data once as “China” and again as “China, mainland” both with very large areas. Brazil comes next, followed by Indonesia and the United States. India is in the middle of the list, showing a significant amount of agricultural emission area, but lower than countries like Brazil and Indonesia. Other countries like the Democratic Republic of the Congo, Canada, Japan, and Mexico have smaller values in comparison. This chart helps show how much land in each country is linked to agricultural practices that may contribute to emissions. Larger countries or those with more farming land tend to have higher values.

5.5 Countries with the Greatest Decline in Agricultural Emission-Associated Land Area

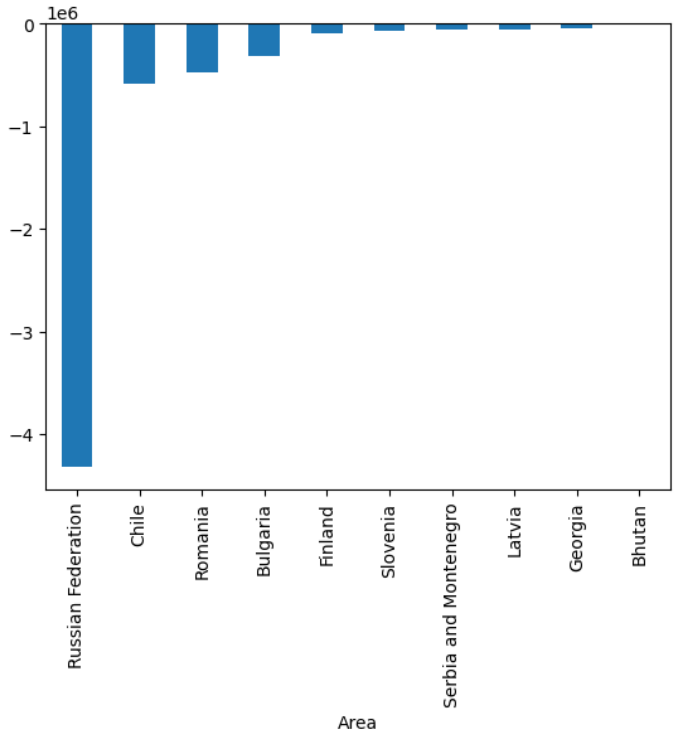

Figure 2.4: Countries with Decline in Agricultural Emission Land Area

The bar chart above shows countries where agricultural emission areas have gone down. The Russian Federation has the biggest decrease by a large margin, with values far lower than the rest. Other countries like Chile, Romania, and Bulgaria also show clear drops. The remaining countries, such as Finland, Slovenia, and Bhutan, have smaller decreases. These negative values may suggest that these countries reduced the size of land used in farming that causes emissions, possibly due to changes in farming methods or better environmental rules. This chart helps highlight where emission-related land use is going down, which could be a positive sign for the environment.

5.6 Heatmap Analysis of Key Contributors to Total Agricultural Emissions

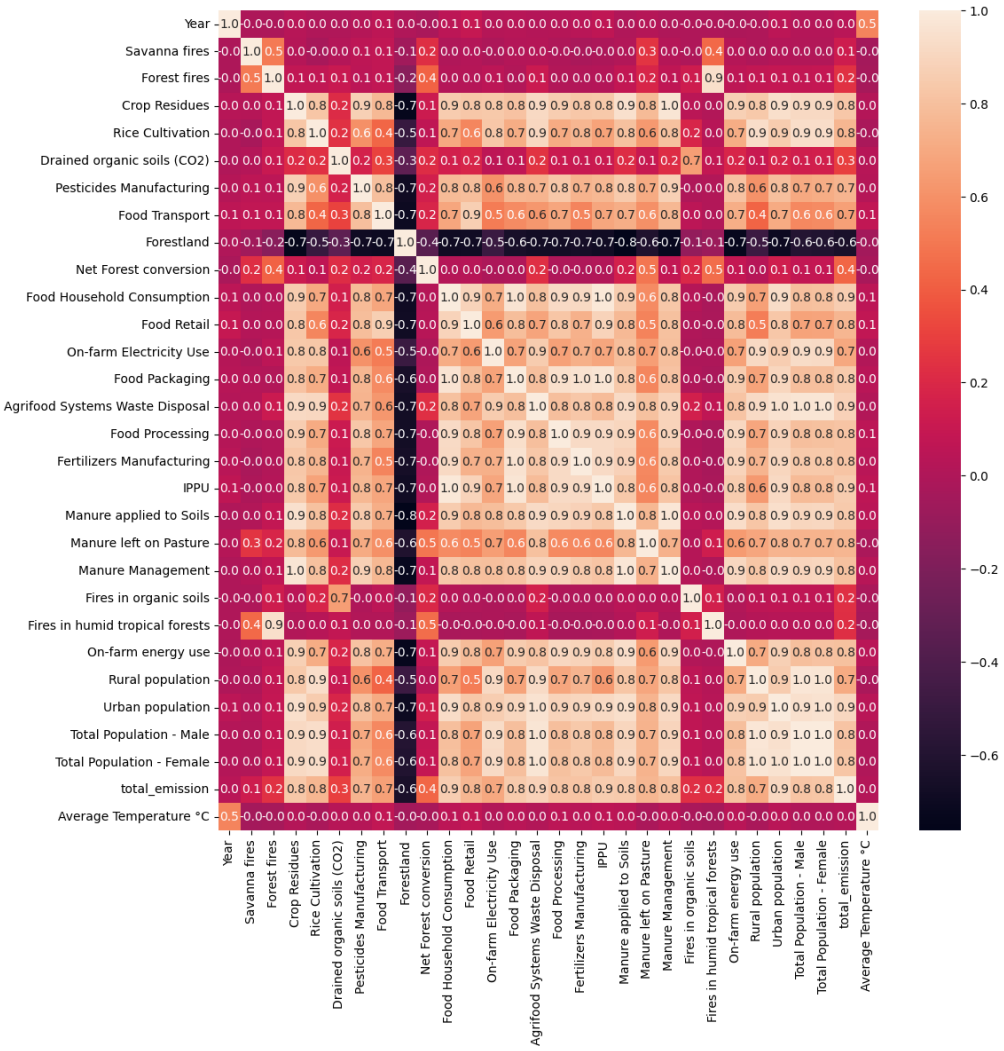

Figure 2.5: Heatmap of Key Contributors to Emissions

The heatmap shows that total emissions have a strong positive relationship with rice cultivation, food transport, food processing, fertilizer manufacturing, manure management, and on-farm energy use. These activities are closely linked to the rise in emissions. Urban and rural population growth also show a strong connection with total emissions. As population increases, more food and energy are needed, leading to higher emissions. Average temperature has a moderate positive correlation with total emissions, suggesting a possible link between emissions and temperature changes. Factors like forest fires, savanna fires, and net forest conversion have weak or negative

relationships with total emissions, showing they may not be major contributors in this dataset.

5.7 Breakdown of Gas Emissions by Activity (1990-2020)

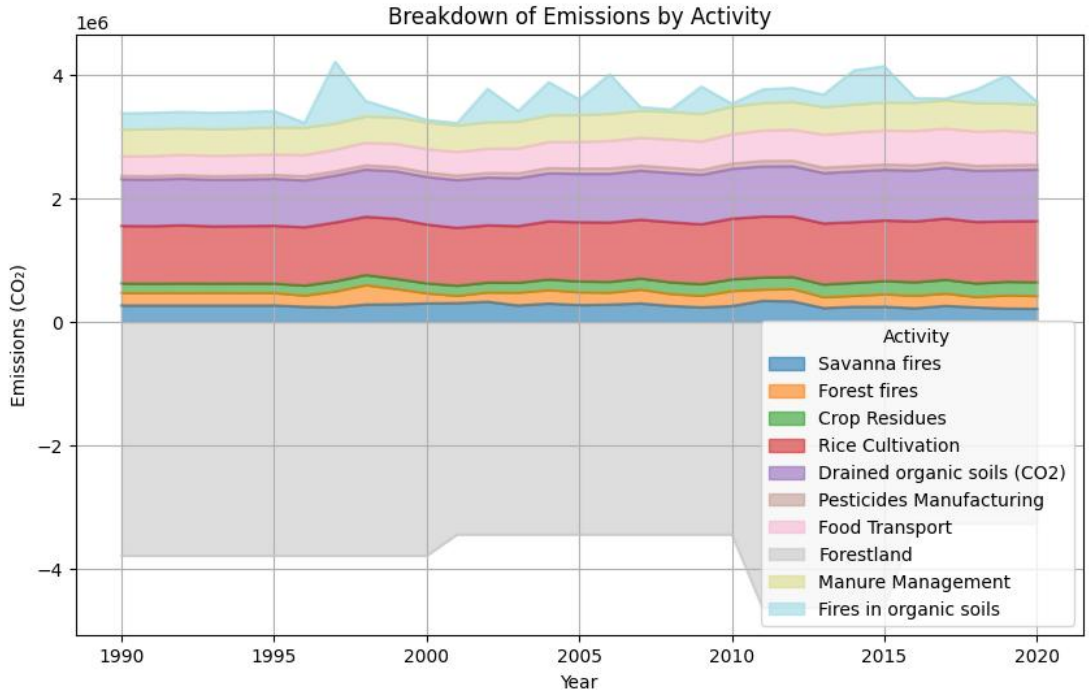

Figure 2.6: Breakdown of Gas Emissions by Activity (1990-2020)

This stacked area chart visualizes the breakdown of CO_2 emissions from various agricultural activities over time, spanning from 1990 to 2020. Each colour band in the graph represents a different agricultural activity contributing to CO_2 emissions. Over the years, emissions from savanna fires, forest fires, and crop residues show noticeable fluctuations, with forest fires and savanna fires showing occasional peaks during specific years, indicating higher levels of fire activity in those periods. Activities like rice cultivation and drained organic soils remain significant and stable contributors to emissions throughout the years, with drained organic soils (CO2) contributing heavily, particularly in the earlier years. Pesticides manufacturing, food transport, manure management, and fires in organic soils contribute relatively smaller amounts, yet still play a role in overall emissions. This chart helps to clearly observe how agricultural activities have influenced CO_2 emissions, with certain activities, such

as rice cultivation and drained organic soils, maintaining a steady and substantial impact, while others show more variability based on specific conditions or events. The data suggests that, while some emissions sources remain consistent, others, like fires, may be driven by external factors such as climate events or policy changes, which are worth investigating further for climate action and policy development.

5.8 Emissions Distribution by Year (1990-2020)

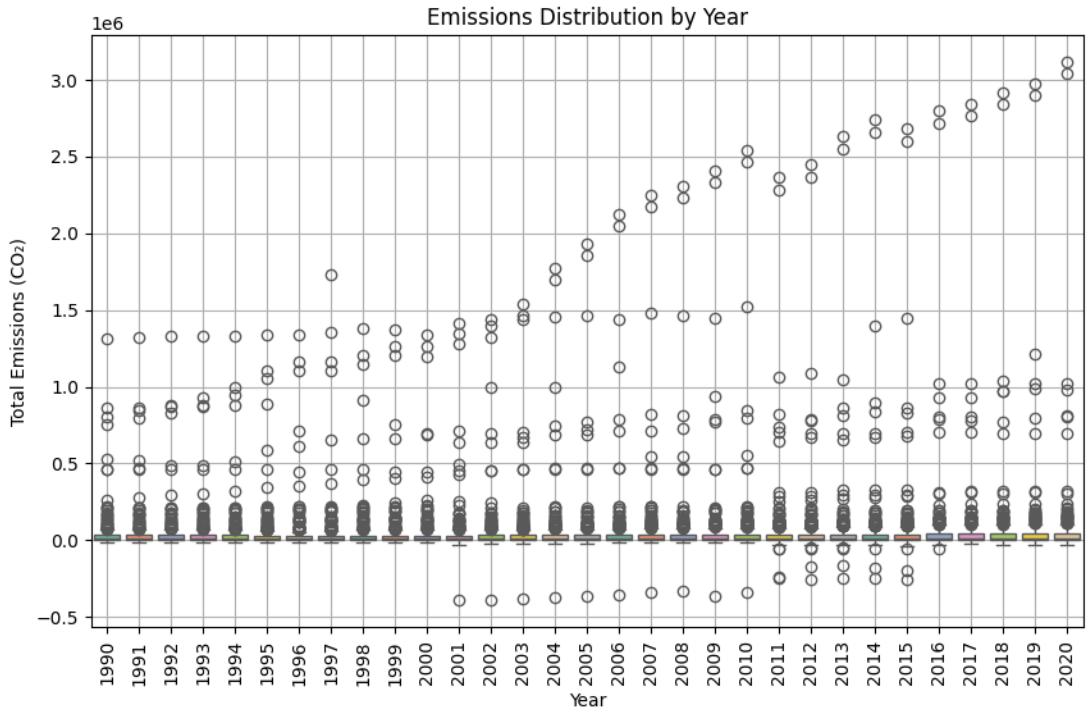

Figure 2.7: Emissions Distribution by Year (1990-2020)

This box plot illustrates the distribution of total CO_2 emissions by year from 1990 to 2020, showing how emissions have varied over time. From 1990 to around 2005, the emissions remained relatively stable, with a narrow interquartile range (IQR) and no significant outliers, indicating that the emissions were fairly consistent across these years. However, starting in 2005, there is a noticeable increase in both the median emissions and the spread of the data, with the whiskers extending higher, suggesting greater variability in emissions. The years after 2010 show several outliers, indicating periods where emissions spiked significantly, likely due to large-scale agricultural events, changes in farming practices, or external environmental factors. Overall, the

chart highlights a general upward trend in CO_2 emissions over the years, with occasional extreme fluctuations, particularly in the later years, pointing to growing agricultural activities or shifts in emission sources.

5.9 Average Emissions by Activity

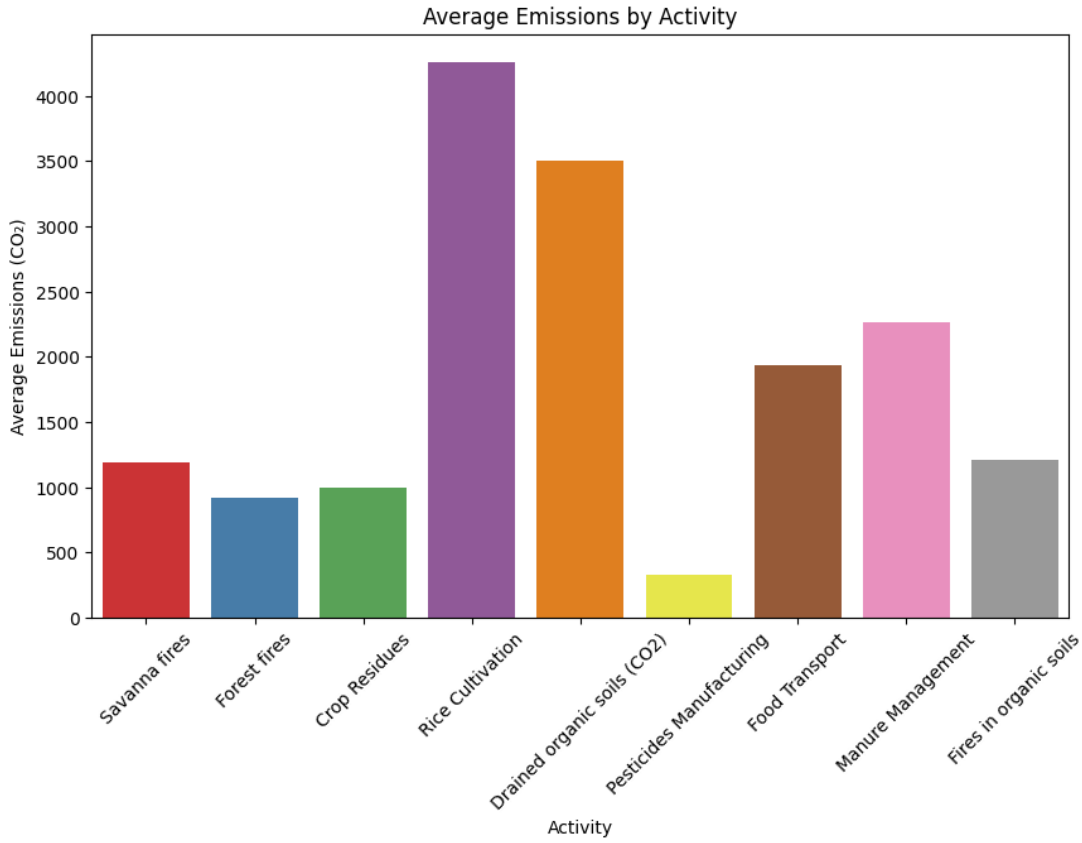

Figure 2.8: Average Emissions by Activity

This bar chart visualizes the average CO_2 emissions by activity, showcasing the relative contribution of different agricultural activities to total emissions. Rice cultivation stands out as the most significant contributor, with emissions reaching over 4000 units, far surpassing the other activities. Drained organic soils (CO_2) is the second-largest source, with emissions just below rice cultivation, indicating a substantial impact on overall CO_2 emissions. Other activities, such as manure management and food transport, contribute notably, with emissions between 1000 and

2000 units on average. Savanna fires, forest fires, and crop residues are among the lower contributors, with emissions ranging from 500 to 1000 units. Pesticides manufacturing and fires in organic soils have the least impact, contributing just a fraction of the emissions compared to the larger sources like rice cultivation and drained organic soils. This graph highlights which activities are most responsible for CO_2 emissions in the dataset, offering valuable insights for targeting mitigation strategies in agricultural practices.

5.10 Pie Chart of Emissions Proportion from Different Activities in 2020

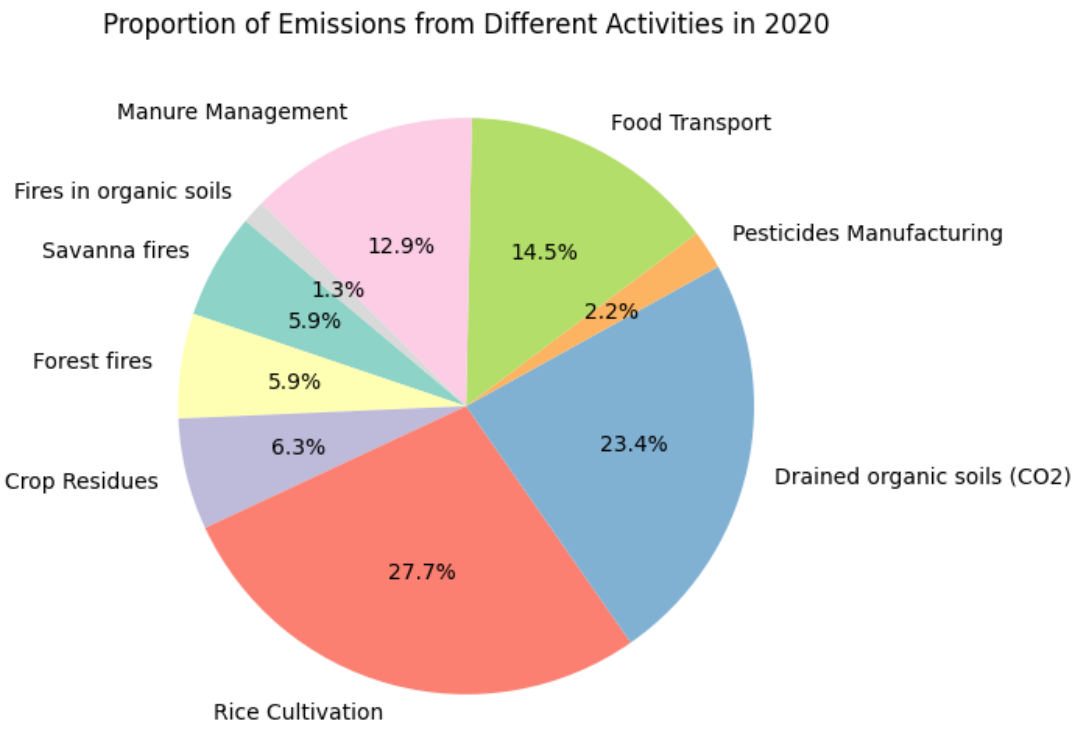

Figure 2.9: Pie Chart of Emission Proportions in 2020

This pie chart illustrates the proportion of emissions from different agricultural activities in 2020, highlighting the significant contributors to CO_2 emissions that year. Rice cultivation stands out as the largest source, responsible for 27.7% of total

emissions, indicating its substantial impact, likely due to methane and other gases released during cultivation. Drained organic soils (CO_2) follow closely with 23.4% of the emissions, emphasizing the importance of soil management in reducing emissions from agricultural lands, particularly those with drained peatlands or wetlands. Activities such as manure management and food transport contribute 12.9% and 14.5%, respectively, showing the significant emissions associated with livestock management and the transportation of agricultural goods. Savanna fires and forest fires contribute 5.9% each, while crop residues account for 6.3%. Lastly, pesticides manufacturing makes up the smallest portion at 2.2%. Overall, the chart highlights that rice cultivation and drained organic soils are the dominant sources of emissions in 2020, while other activities like manure management and food transport also play key roles in contributing to agricultural emissions.

Chapter 6
MODEL COMPARISON

6.1 Introduction

In this section, various machine learning models were evaluated for their ability to predict CO2 emissions based on a set of environmental and agricultural features, including forest fires, crop residues, food transport, temperature, population density, and industrial activities. The models selected for this comparison include Multiple Linear Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, XGBoost Regression, and K-Nearest Neighbors (KNN). Each model was trained using the pre-processed dataset and evaluated on a separate test set.

The models were assessed using key performance metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2), which provided insights into their accuracy and predictive power. The results of these evaluations are presented and compared to identify the most effective model for estimating CO2 emissions. This comparison highlights the strengths and weaknesses of each model, providing a clear understanding of which approach offers the most reliable predictions for this specific task.

6.2 Steps Involved in Applying Machine Learning Methods

1.Data Collection:

The dataset includes a variety of environmental and agricultural features, such as forest fires, crop residues, food transport, temperature, population density, and

industrial activities. These factors serve as the input variables for predicting CO2 emissions.

2.Data Preprocessing:

- Handling Missing Values: Missing values in the dataset were identified and handled using imputation techniques (e.g., filling missing values with the mean, median, or mode).
- Feature Engineering: Derived new features like Total Population by summing Total Population - Male and Total Population - Female to represent the population more accurately.
- Feature Scaling: Numerical features were scaled using StandardScaler to standardize their values, ensuring that all features had a mean of 0 and a standard deviation of 1. This step was particularly important for models like KNN and Linear Regression, which are sensitive to the scale of the data.

3.Splitting the Dataset:

The dataset was split into training and testing sets, typically using an 80/20 or 70/30 ratio. The training set was used to train the models, while the testing set was used to evaluate their performance.

4.Model Selection:

Several regression models were selected for comparison, including:

- Multiple Linear Regression
- Decision Tree Regression
- Random Forest Regression
- Gradient Boosting Regression
- XGBoost Regression
- K-Nearest Neighbors (KNN) Regression

5.Model Training:

Each model was trained using the training dataset (x_train, y_train). The models were fit to the data, learning the underlying relationships between the input features and the target variable (CO2 emissions).

6.Model Evaluation:

The trained models were evaluated using the test dataset (X_test, y_test) to assess their predictive performance. Key performance metrics used for evaluation include:

- Mean Absolute Error (MAE): Measures the average magnitude of errors in predictions.
- Mean Squared Error (MSE): Represents the average squared difference between predicted and actual values.
- Root Mean Squared Error (RMSE): The square root of MSE, providing an error metric in the same units as the target variable.
- R-squared (R^2): Indicates how well the model explains the variance in the target variable.

7.Model Comparison:

The performance metrics for each model were compared to determine the most effective regression approach for predicting CO2 emissions. The models were analysed in terms of their accuracy, ability to handle non-linear relationships, and generalization to unseen data.

By following these steps, the project successfully applied machine learning methods to predict CO2 emissions, offering insights into environmental factors that influence emissions and contributing to future sustainability efforts.

6.3 Model Comparison

In this section, multiple machine learning models are compared to evaluate their effectiveness in predicting total CO_2 emissions (regression task) and classifying average temperature levels (classification task) based on agricultural and environmental features. For the regression task, models such as Multiple Linear Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regression (SVR), and Gradient Boosting Regression are employed. For classification, models including Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Support Vector Machine (SVM), and Gradient Boosting Classifier are used. Each model's performance is assessed using appropriate metrics

Root Mean Squared Error (RMSE) and R-squared for regression, and accuracy, confusion matrix, and classification report for classification. The goal of this comparison is to identify the most accurate and reliable model for each task, thereby enhancing the overall quality of prediction and decision-making based on the dataset.

6.3.1 Model Comparison for Total Emission (Regression)

In this section, various regression models are evaluated to predict total CO_2 emissions. The models are compared based on RMSE, MAE, MSE, and R^2 scores.

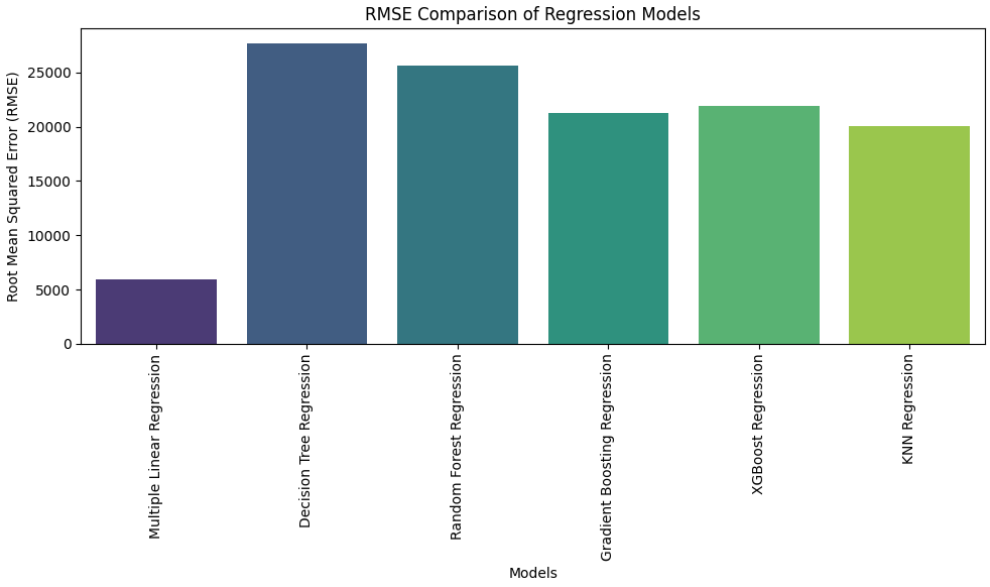

Figure 3.1 : RMSE Comparison of Regression Models

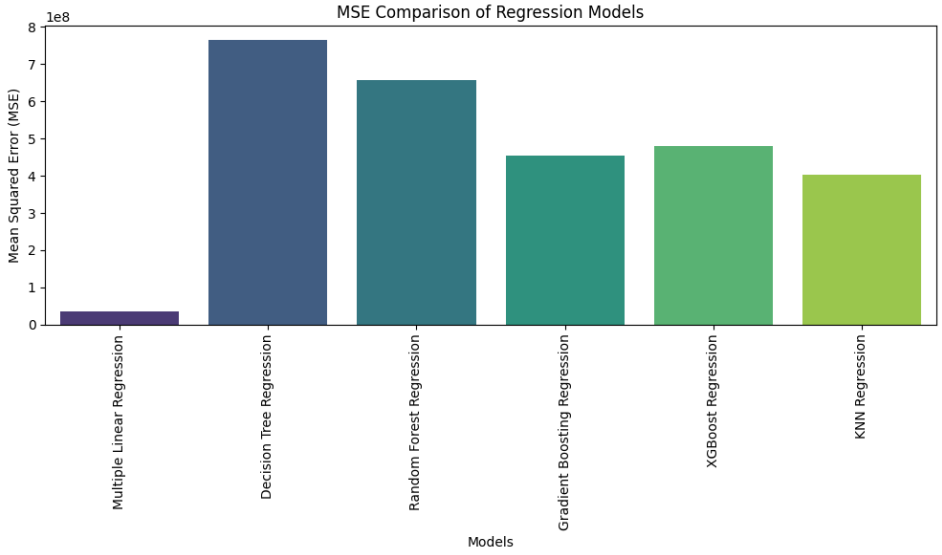

Figure 3.2 : MSE Comparison of Regression Models

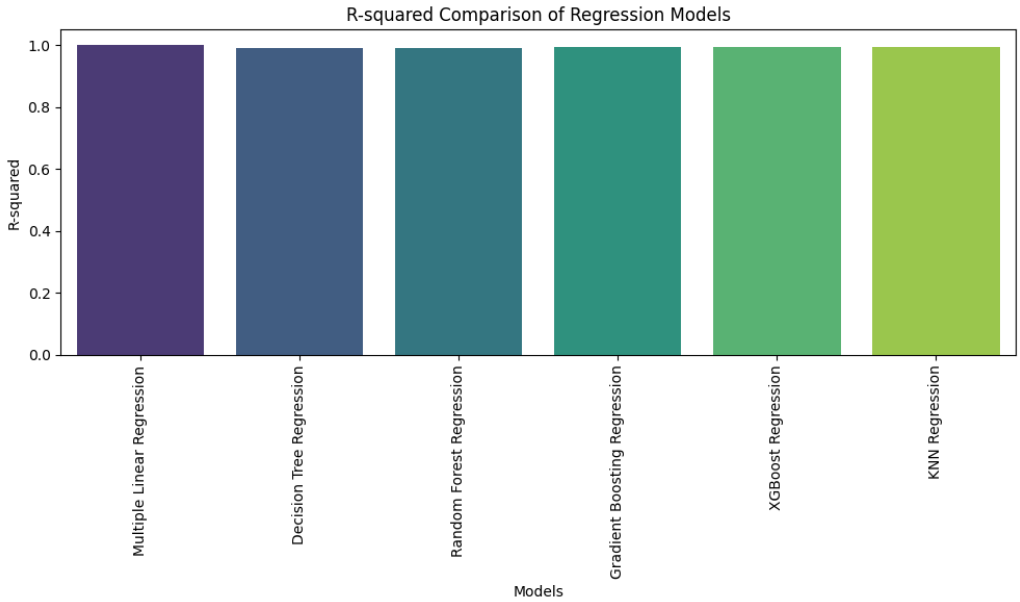

Figure 3.3 : R-squared Comparison of Regression Models

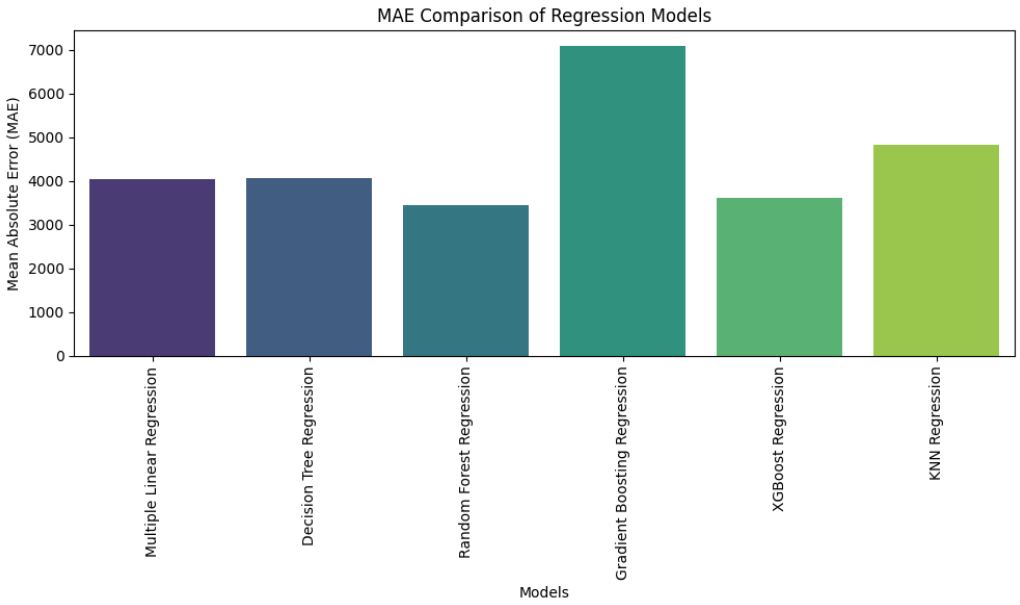

Figure 3.4 : MAE Comparison of Regression Models

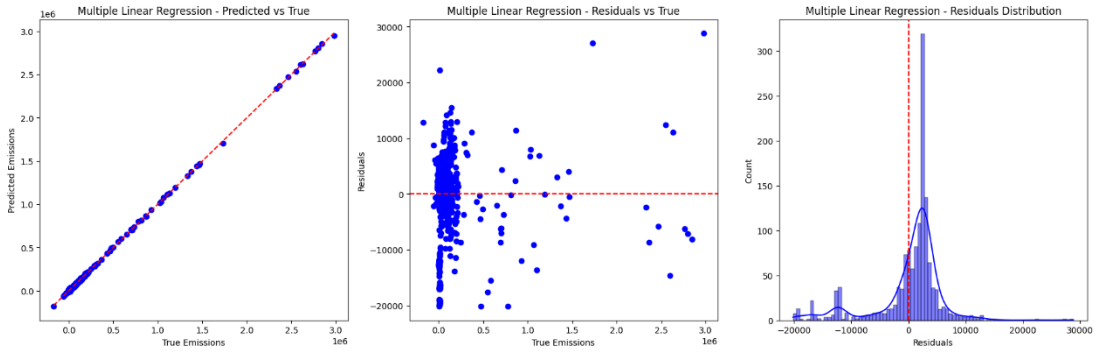

Figure 3.5 : Multiple Linear Regression Diagnostic

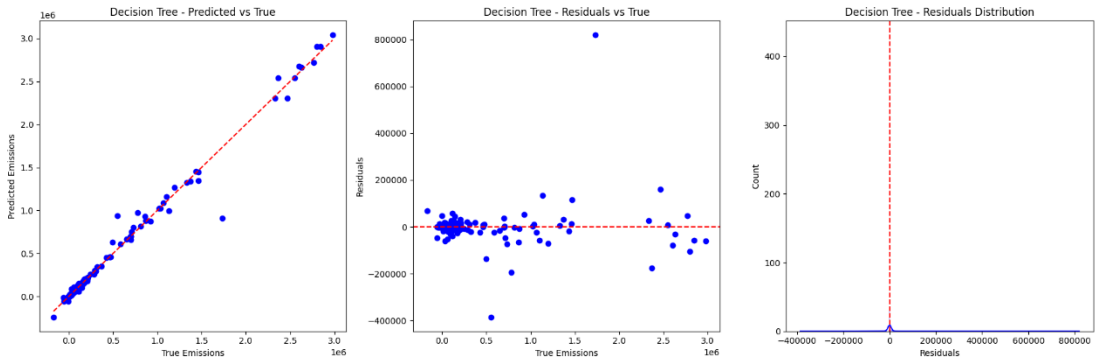

Figure 3.6 : Decision Tree Regression Diagnostics

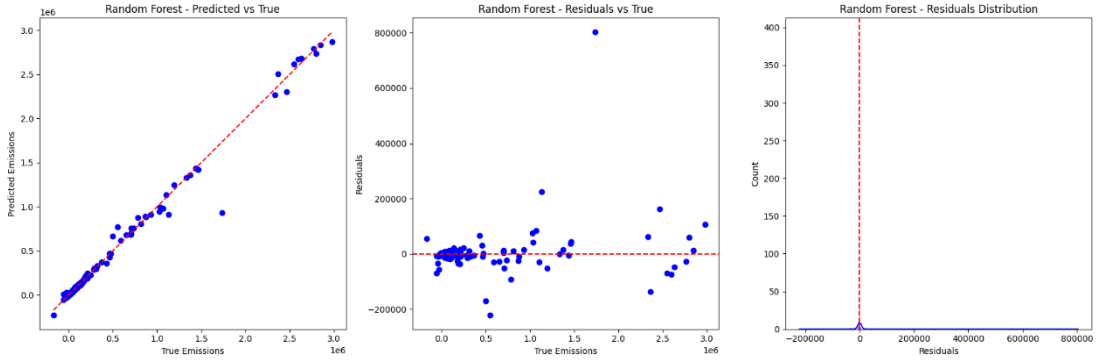

Figure 3.7 : Random Forest Regression Diagnostics

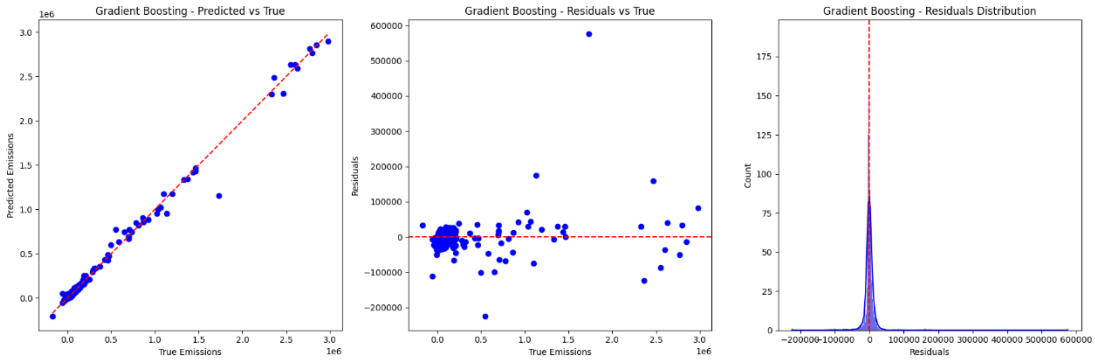

Figure 3.8 : Gradient Boosting Regression Diagnostics

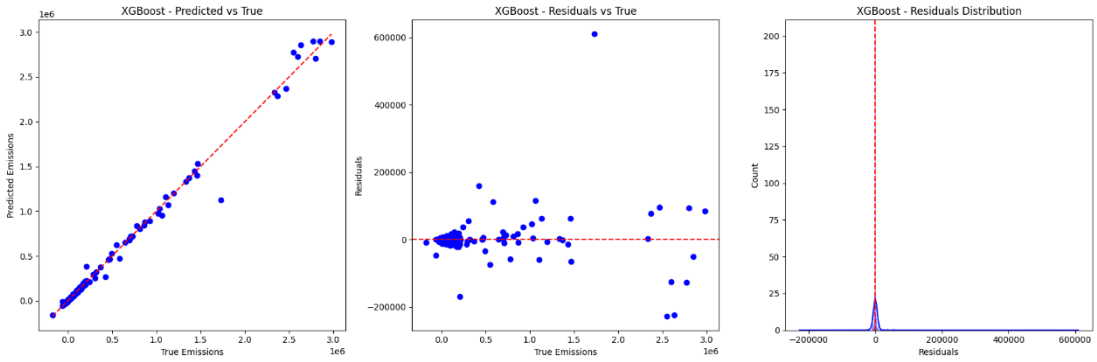

Figure 3.9 : XGBoost Regression Diagnostics

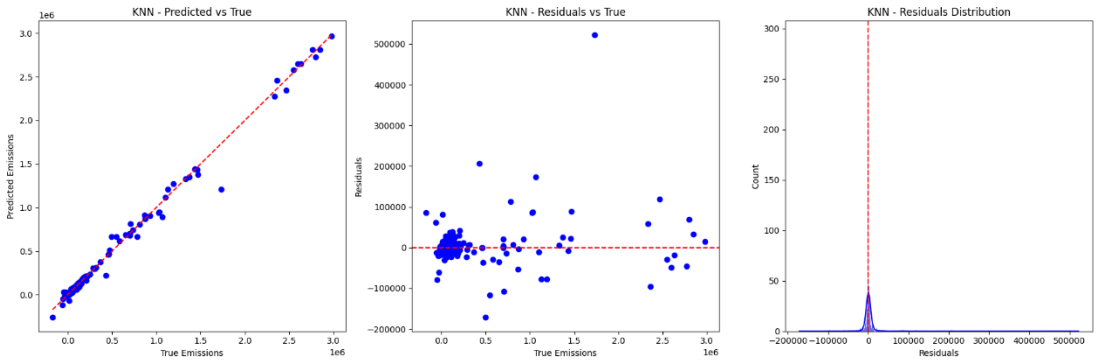

Figure 3.10 : KNN Regression Diagnostics

- Multiple Linear Regression shows strong predictive performance, with the Predicted vs True plot indicating a good fit as the predicted and actual values are closely aligned. However, the Residuals vs True plot reveals some larger residuals, suggesting a few outliers or extreme errors. The Residuals Distribution shows a skewed pattern, with a peak near zero and a tail extending to higher residuals, indicating potential over-prediction for lower emission values. Overall, the model is effective, but there is room for improvement in handling outliers and capturing non-linear relationships.
- Decision Tree Regression shows a good fit in the Predicted vs True plot, but the Residuals vs True plot indicates overfitting, with large residuals and significant variation in errors. The Residuals Distribution highlights a large spike near zero and spreads in the errors, suggesting the model struggles with extreme predictions. While effective in capturing patterns, the model's overfitting and large residuals point to areas for improvement.
- Random Forest Regression shows a strong fit in the Predicted vs True plot, indicating accurate predictions. The Residuals vs True plot suggests slight overfitting, and the Residuals Distribution shows a peak near zero with a few extreme errors. While the model performs well, it may benefit from adjustments to reduce large residuals.
- Gradient Boosting Regression shows a good fit in the Predicted vs True plot, with predictions closely aligned to actual values. The Residuals vs True plot suggests some variance, indicating slight overfitting, and the Residuals Distribution shows a sharp peak with some extreme outliers. While the model performs well overall, it may benefit from handling large residuals better.
- XGBoost Regression shows a strong fit in the Predicted vs True plot, with predicted values closely aligning with actual values. The Residuals vs True plot indicates some variation in residuals, suggesting a slight presence of overfitting. The Residuals Distribution plot shows a sharp peak near zero, indicating most residuals are small, but there are a few extreme errors. Overall, XGBoost performs well, though it may benefit from adjustments to handle large residuals or outliers.
- KNN Regression shows a good fit in the Predicted vs True plot, with predicted values closely following the red dashed line. The Residuals vs True plot

indicates some variance in residuals, but it appears less pronounced compared to other models, suggesting minimal overfitting. The Residuals Distribution shows a sharp peak near zero, indicating that most residuals are small, though a few extreme values remain. Overall, KNN performs well but may benefit from adjustments to handle the larger residuals and outliers more effectively.

Table 2.1 : Model Performance Metrics for Regression Models

Model	MAE	MSE	RMSE	R-squared
Multiple Linear Regression	4049.42	34,864,832.70	5904.65	0.9995
Decision Tree Regression	4077.09	764,586,304.61	27,651.15	0.9889
Random Forest Regression	3444.26	657,166,379.47	25,635.26	0.9905
Gradient Boosting Regression	7084.46	453,906,572.83	21,305.08	0.9934
XGBoost Regression	3625.44	479,715,852.89	21,902.42	0.9931
KNN Regression	4839.58	402,671,843.91	20,066.68	0.9942

The performance of several machine learning models Multiple Linear Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, XGBoost Regression, and KNN Regression was assessed based on key metrics: MAE, MSE, RMSE, and R-squared (R^2). Multiple Linear Regression performed the best with an R^2 of 0.9995, indicating it could explain nearly all the variance in the target variable and showed low prediction errors. Random Forest and Gradient Boosting followed closely with R^2 values of 0.9905 and 0.9934, respectively, demonstrating strong performance, though with slightly higher error values compared to the linear model. XGBoost had similar performance to Gradient Boosting but with slightly higher errors. Decision Tree exhibited lower performance with an R^2 of 0.9889, suggesting overfitting, and KNN showed the lowest R^2 of 0.9942, with higher RMSE compared to the other models. Overall, Multiple Linear Regression proved to be the most reliable model for predicting CO2 emissions, with Random Forest and Gradient Boosting offering strong alternatives for handling more complex relationships.

6.3.2 Model Comparison for Average Temperature (Classification)

To assess the classification performance in predicting average temperature, a range of machine learning models was employed, including Decision Tree, Random Forest, K-Nearest Neighbors, and Support Vector Machine. The target variable average temperature was converted into categorical labels representing different temperature levels such as High, Medium, and Low. These models were trained and tested using a stratified split of the dataset. Evaluation was conducted based on metrics such as accuracy, precision, recall, and the confusion matrix to compare how well each model classified the temperature categories.

Decision Tree

The table below shows the performance metrics of the Decision Tree model used to classify average temperature levels. It includes precision, recall, and F1-score for each temperature category, along with overall accuracy, macro average, and weighted average scores.

Table 2.2 : Classification Report per Class for Decision Tree Model

	precision	recall	f1-score	support
High	0.25	0.29	0.27	56
Low	0.38	0.38	0.38	145
Medium	0.89	0.89	0.89	1192
Accuracy			0.81	1393
Macro avg	0.51	0.52	0.51	1393
Weighted avg	0.81	0.81	0.81	1393

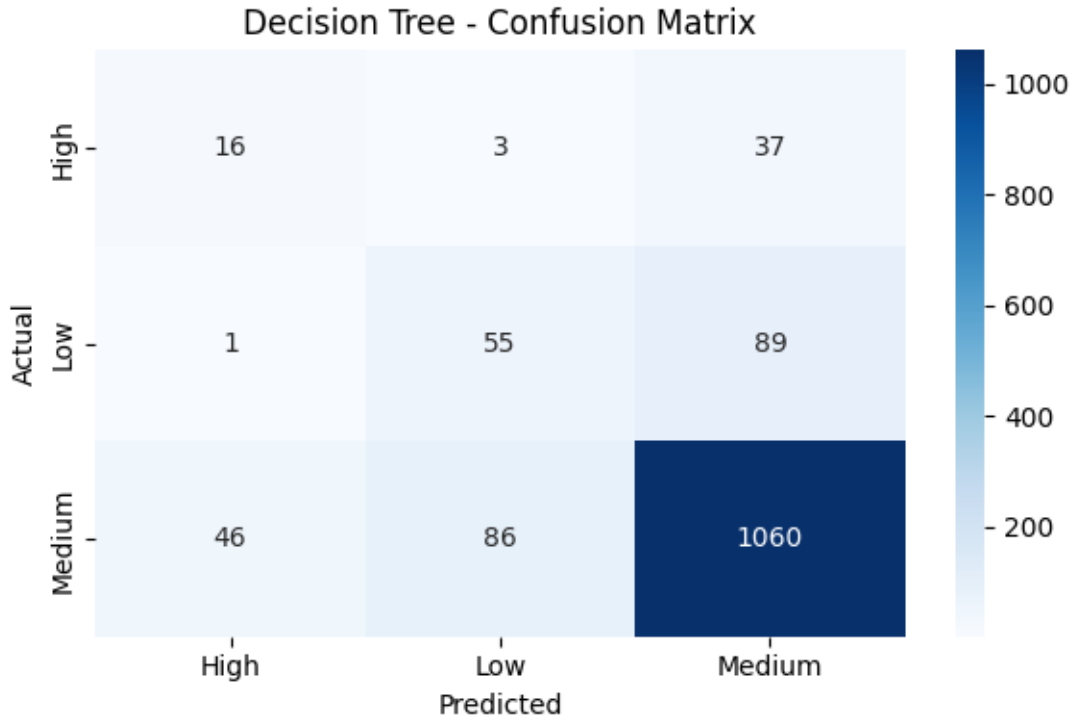

Figure 3.11 : Decision Tree - Confusion Matrix

Random Forest

The table below shows the performance metrics of the Random Forest model used to predict temperature categories. It includes precision, recall, and F1-score for each class, along with overall accuracy, macro average, and weighted average scores.

Table 2.3 : Classification Report per Class for Random Forest Model

	precision	recall	f1-score	support
High	0.28	0.12	0.17	56
Low	0.54	0.31	0.39	145
Medium	0.88	0.95	0.92	1192
Accuracy			0.85	1393
Macro avg	0.57	0.46	0.50	1393
Weighted avg	0.82	0.85	0.83	1393

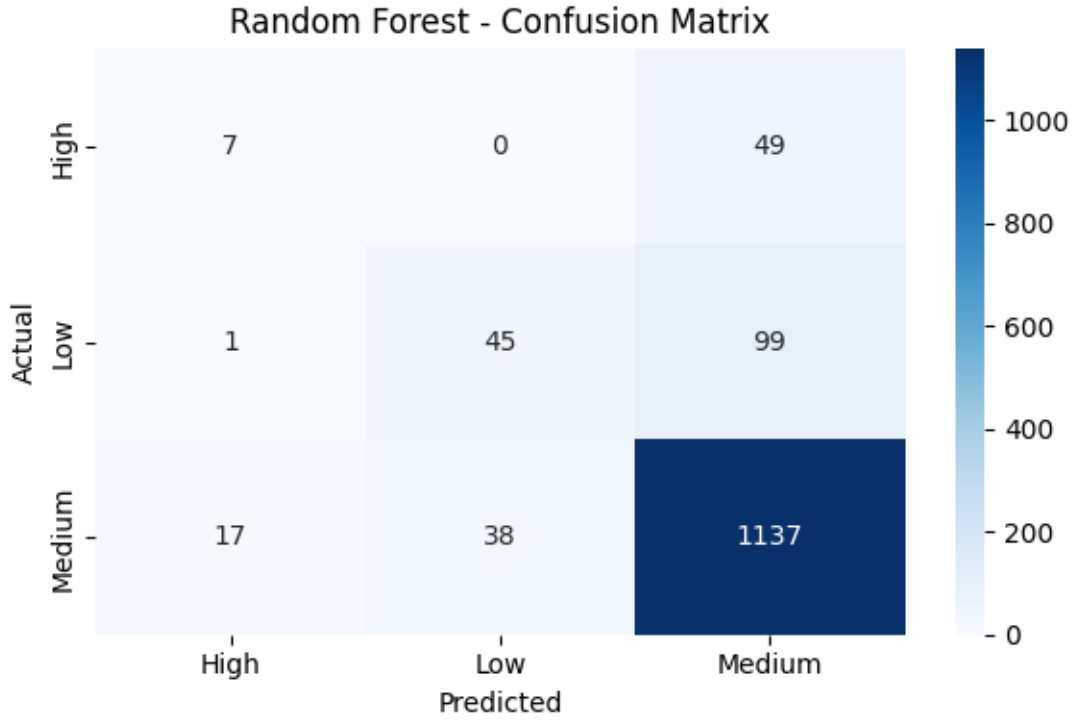

Figure 3.12 : Random Forest - Confusion Matrix

K-Nearest Neighbors

The table below presents the performance metrics of the K-Nearest Neighbors (KNN) model used for classifying temperature categories. It outlines precision, recall, and F1-score for each class, as well as overall accuracy, macro average, and weighted average scores.

Table 2.4 : Classification Report per Class for KNN Model

	precision	recall	f1-score	support
High	0.25	0.05	0.09	56
Low	0.36	0.11	0.17	145
Medium	0.87	0.97	0.92	1192
Accuracy			0.84	1393
Macro avg	0.49	0.38	0.39	1393
Weighted avg	0.79	0.84	0.80	1393

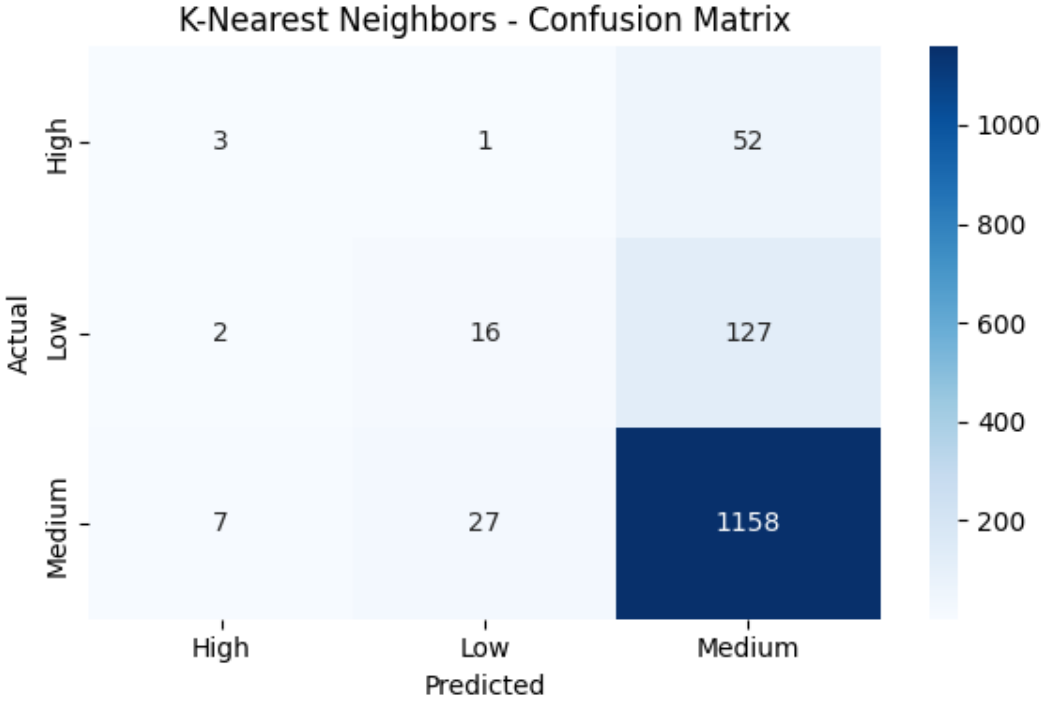

Figure 3.13 : KNN - Confusion Matrix

Support Vector Machine (SVM)

The table below displays the performance metrics of the Support Vector Machine (SVM) model applied to temperature classification. It includes precision, recall, and F1-score for each temperature category, along with the model's overall accuracy, macro average, and weighted average performance.

Table 2.5 : Classification Report per Class for SVM Model

	precision	recall	f1-score	support
High	0.00	0.00	0.00	56
Low	0.00	0.00	0.00	145
Medium	0.86	1.00	0.92	1192
Accuracy			0.86	1393
Macro avg	0.29	0.33	0.31	1393
Weighted avg	0.73	0.86	0.79	1393

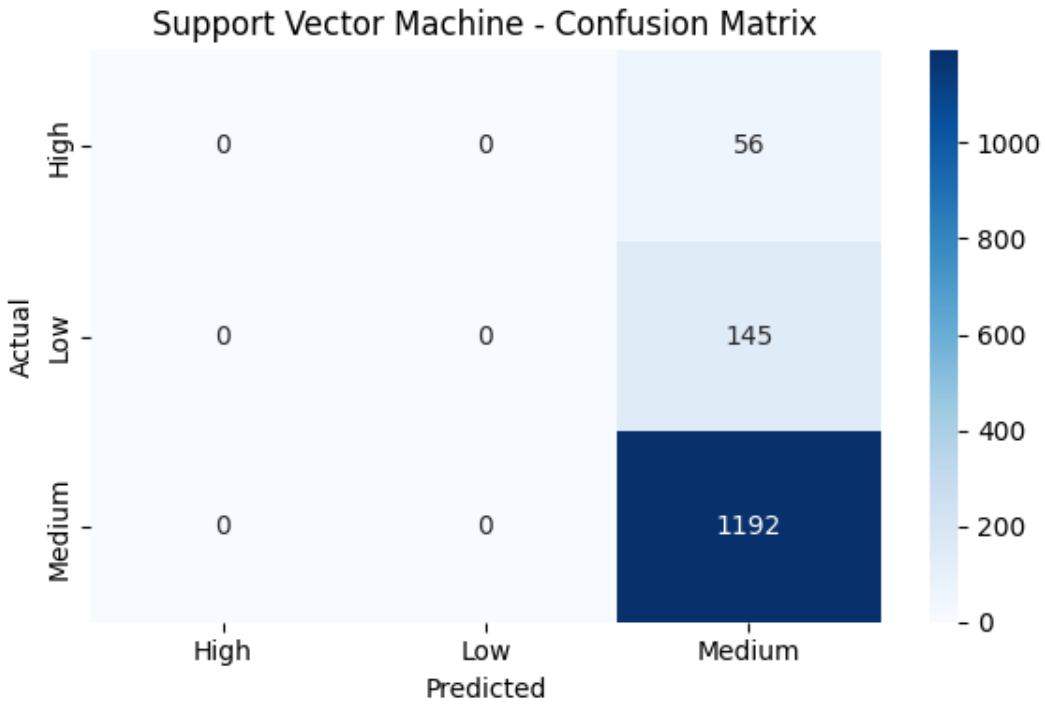

Figure 3.14 : SVM - Confusion Matrix

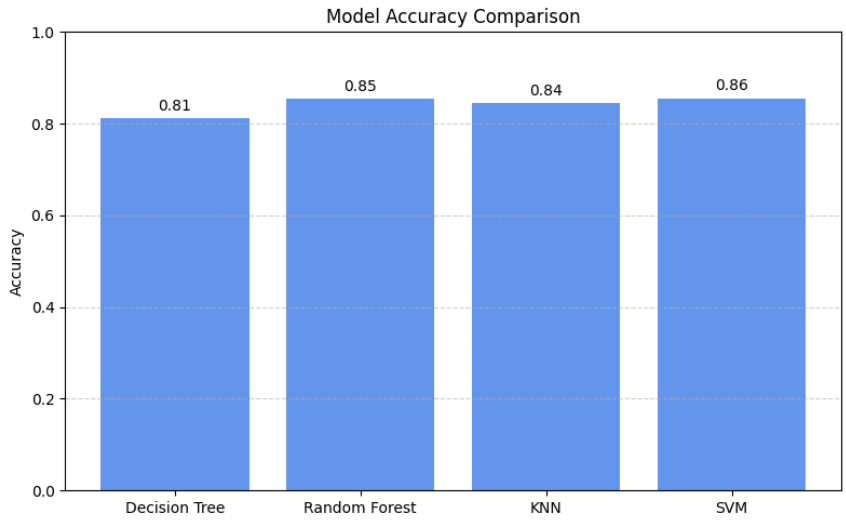

Figure 3.15 : Model Accuracy Comparison

The performance of four classification models Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) was assessed for predicting average temperature categories: Low, Medium, and High. Among them, Random Forest emerged as the best-performing model with an accuracy of 85.36%, demonstrating strong and balanced results across all classes, especially for the dominant 'Medium' category. SVM achieved the highest overall accuracy of 85.57%; however, it completely failed to predict the 'Low' and 'High' classes, making it the least effective in terms of class-wise performance. KNN showed good accuracy (84.49%) but also struggled to classify minority classes accurately. The Decision Tree, with an accuracy of 81.19%, performed moderately well for the 'Medium' category but had limited success in identifying 'Low' and 'High'. Therefore, Random Forest can be considered the most reliable model, while SVM performed poorly in terms of balanced classification.

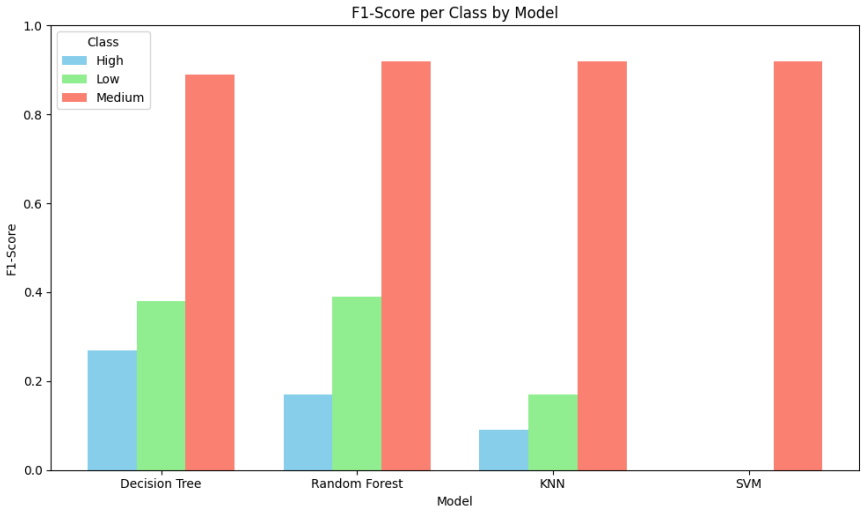

Figure 3.16 : F1-Score per Class by Model

The graph titled "F1-Score per Class by Model" provides a comparative visualization of how four machine learning models Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) perform in predicting the categorical classes: High, Low, and Medium. Across all models, the Medium class shows consistently high F1-scores, nearing or exceeding 0.9, indicating excellent predictive performance. This trend suggests a potential class imbalance where the models are skewed toward the majority class. Among the minority classes, Random Forest demonstrates the strongest ability to identify the Low category, followed by the Decision Tree. On the other hand, KNN shows lower performance on both Low and High classes, while SVM records zero F1-score for these classes, indicating it failed to predict them altogether.

In conclusion, while all models perform well for the Medium class, Random Forest stands out as the most balanced model, achieving better results on the minority classes compared to others. Despite SVM showing high overall accuracy, its inability to identify Low and High categories makes it the least suitable model for this classification task. This highlights the importance of using class-wise metrics like F1-score when evaluating models, especially in imbalanced datasets.

Chapter 6
CONCLUSIONS

This study aimed to investigate the impact of agricultural activities on CO_2 emissions and to apply machine learning models for the prediction and classification of key environmental indicators. By analysing a rich dataset spanning from 1990 to 2020 across multiple countries, the research provided deep insights into the temporal and regional variations in emissions, highlighting the roles of specific agricultural practices such as rice cultivation, manure management, and land use change.

The descriptive and visual analysis revealed a steady increase in agricultural CO_2 emissions over the years, with rice cultivation and drained organic soils identified as major contributors. Countries like China, Brazil, and India were observed to have significant agricultural emission footprints, while regions like the Russian Federation showed notable reductions in emission-associated land use.

The machine learning component of the study involved comparing regression and classification models for predicting total emissions and categorizing average temperatures, respectively. Among the regression models tested, Multiple Linear Regression demonstrated the highest performance with an R^2 value of 0.9995, closely followed by KNN, Random Forest, and XGBoost. These results suggest that even simpler linear models can effectively capture relationships in well-prepared datasets.

For the classification of temperature categories, Random Forest proved to be the most balanced model, achieving high accuracy while maintaining acceptable F1-scores across all classes, especially the minority classes. In contrast, although SVM achieved high overall accuracy, it failed to predict the 'High' and 'Low' temperature categories, illustrating the limitations of relying solely on accuracy in imbalanced classification tasks.

Overall, this study demonstrates the value of machine learning in environmental analysis and policy modelling. It shows that predictive models, when correctly chosen and tuned, can provide highly accurate estimations and classifications for emissions data. The findings highlight the importance of understanding key agricultural emission drivers and leveraging data-driven approaches for environmental decision-making.

REFERENCE

Chen, W., & Lei, Y. (2018). A comparative study on machine learning methods for predicting CO_2 emissions in the agricultural sector. *Computers and Electronics in Agriculture, 150*, 149–157. https://doi.org/10.1016/j.compag.2018.04.003

Feng, Y., Wang, X., & Guo, X. (2021). Machine learning-based evaluation of greenhouse gas emissions from agricultural production in China. *Journal of Cleaner Production, 290*, 125217. https://doi.org/10.1016/j.jclepro.2021.125217

Kumar, R., & Kumar, A. (2022).Comparative analysis of machine learning algorithms for greenhouse gas emission forecasting. *Sustainability, 14*(9), 5071. https://doi.org/10.3390/su14095071

Li, K., Fang, Y., & He, L. (2020). Modeling regional CO_2 emissions using an improved random forest approach: Evidence from Chinese provinces. *Ecological Indicators, 112*, 106149. https://doi.org/10.1016/j.ecolind.2020.106149

Li, X., Zhou, Y., & Ma, Z. (2021).Application of machine learning in predicting carbon emissions: A systematic review. *Environmental Science and Pollution Research, 28*, 54697–54715. https://doi.org/10.1007/s11356-021-14860-6.

Yuan, Y., Zhang, L., & Liu, H. (2020). Comparative analysis of machine learning algorithms for forecasting agricultural carbon emissions. *Sustainability, 12*(14), 5761. https://doi.org/10.3390/su12145761

Zhao, X., Wang, S., & Wang, Y. (2021). Exploring the relationship between agricultural activities and greenhouse gas emissions using machine learning models. *Environmental Science and Pollution Research, 28*(15), 18625–18636. https://doi.org/10.1007/s11356-020-12129-4