Analysis of Food Production-Related Emissions in European Countries: Trends and Impacts on Climate Change

15.02.2024

—

Group Members:
Muhammad Haseeb Abbasi
Hamza
Manahil

Overview

Report provides a comprehensive analysis of greenhouse gas emissions from both pre- and post-production activities across various sectors in Europe, covering the years 1990 to 2021. Through detailed examination and visualization using interactive plots, This report explore the temporal trends and geographical distributions of emissions for N_2O, CH_4, and CO_2. The report highlights key contributors to emissions within different categories, such as pesticides and fertilizer manufacturing, as well as post-production activities like food processing and industrial wastewater. Mitigation strategies are proposed based on these insights, emphasizing the importance of policy interventions, technological innovation, and behavioural changes. Additionally, the report underscores the pivotal role of machine learning models in supporting these mitigation efforts by providing predictive insights and facilitating decision-making processes to reduce emissions effectively.

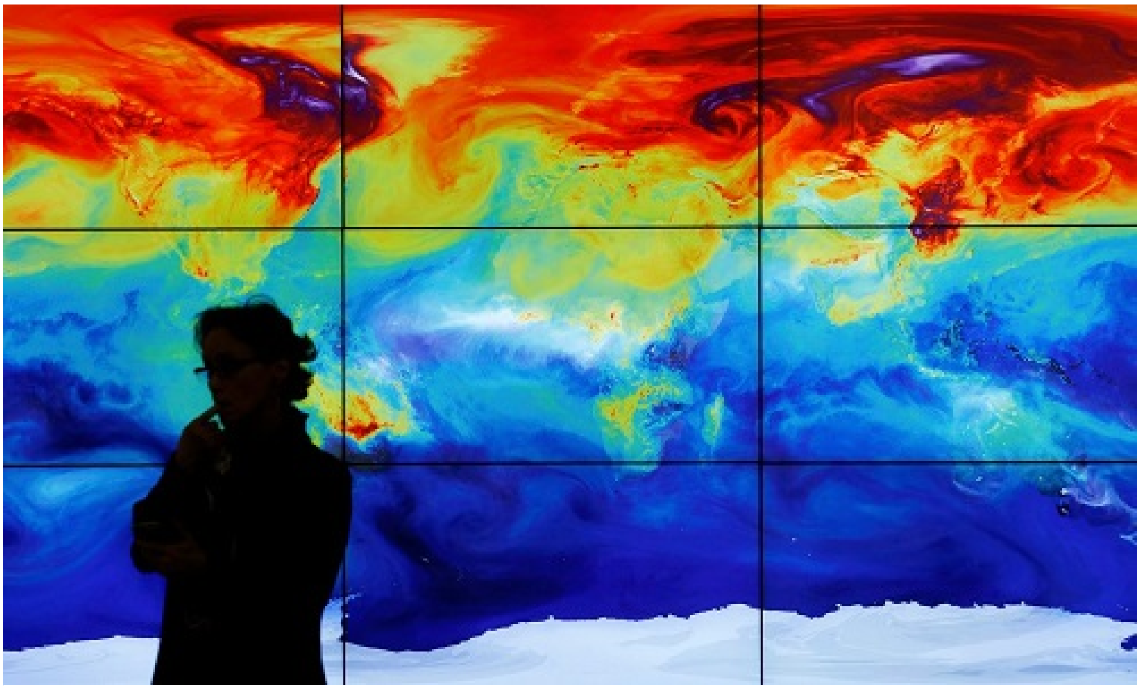
Goals

1. The primary goal of this study is to analyse and understand the patterns and trends of greenhouse gas emissions from both pre- and post-production activities across various sectors in Europe.
2. By leveraging data analysis techniques and visualization tools, the study aims to uncover insights into the temporal and spatial distribution of emissions for N_2O, CH_4, and CO_2.
3. The study seeks to identify key contributors to emissions within different categories and propose mitigation strategies to address these sources effectively.

Special Thanks:

Special thanks are extended to Dr. Ammar Tufail for his invaluable guidance, expertise, and mentorship throughout this project. His insightful advice and unwavering support have been instrumental in shaping the direction and success of this study. We are deeply grateful for his dedication and contributions.

Table of Contents



Chapter 01:

Introduction to Climate Change: Agrifood system chain focusing on GHG's emission from pre and post production systems.

Introduction

The intersection of climate change and agrifood production emissions is a critical area of study that has significant implications for global sustainability and food security. Climate change, characterized by long-term shifts in temperature, precipitation, and other atmospheric conditions, is intricately linked to the agrifood system, which encompasses the entire food production and consumption chain. This link is manifested through the emission of greenhouse gases (GHGs) from various stages of agrifood production, including pre and post agricultural activities.

Climate Change and Agrifood Production Emissions:

Climate change poses a substantial threat to agrifood production due to its potential to disrupt agricultural systems, alter growing conditions, and increase the frequency of extreme weather events. These changes can have far-reaching consequences for food availability, access, and utilization, making it imperative to understand and mitigate the associated emissions from the agrifood system.

Pre-Agricultural Production:

The pre-agricultural production stage encompasses activities that occur before the cultivation of crops or the raising of livestock. Within this stage, two significant sources of emissions are 'Pesticides Manufacturing' and 'Fertilizers Manufacturing'. Pesticides are chemical substances used to control pests and diseases in crops, while fertilizers are applied to enhance soil fertility and crop yields. The manufacturing processes of these inputs can result in the release of GHGs, such as carbon dioxide (CO2) and nitrous oxide (N2O), contributing to the overall carbon footprint of agrifood production.

Post-Agricultural Production:

Following agricultural activities, the post-agricultural production stage involves a series of interconnected processes that transform raw agricultural commodities into consumable food products and manage associated waste streams. This stage encompasses various sources of emissions, including 'Food Processing', 'Food Transport', 'Food Packaging', 'Food Retail', 'Food Household Consumption', 'Solid Food Waste', 'Domestic Wastewater', 'Industrial Wastewater', and 'Incineration'.

Each of these sources plays a distinct role in the emission profile of the agrifood system. For instance, food processing involves energy-intensive operations that can lead to the release of CO2 and other pollutants. Similarly, food transport contributes to emissions through the combustion of fossil fuels in vehicles, while food packaging and retail activities generate waste and energy-related emissions. Furthermore, food consumption, waste management, and wastewater treatment processes are associated with additional emissions throughout the post-agricultural production stage.

Hypothesis:

The hypothesis for this study, "Comparative Emissions Analysis," aims to investigate and compare the emissions associated with pre-agricultural production, including 'Pesticides Manufacturing' and 'Fertilizers Manufacturing,' with those from post-agricultural activities such as 'Food Processing,' 'Food Transport,' 'Food Packaging,' 'Food Retail,' 'Food Household Consumption,' 'Solid Food Waste,' 'Domestic Wastewater,' 'Industrial Wastewater,' and 'Incineration.' The primary objective is to discern which stages of the agrifood production system have the most significant impact in terms of emissions. By conducting this comparative analysis, the study seeks to provide valuable insights into the relative environmental footprint of different stages within the agrifood system, thereby informing potential mitigation strategies and sustainable practices to address the most impactful areas of emissions.

Aim and Objectives:

The aim and objectives of the study align with the hypothesis and have been supported through key analytical steps in the Jupyter Notebook within Visual Studio Code (VSCode). The study's primary aim is to conduct a comprehensive analysis of emissions from pre and post-agricultural production activities, with the following specific objectives:

- Feature Engineering: The study involved the identification and transformation of relevant variables and features to enhance the predictive power of the models and provide a more nuanced understanding of emissions across different stages of agrifood production.
- Geographical Analysis using Interactive Plots: By leveraging interactive plots, the study aimed to visualize and explore the spatial distribution of emissions, allowing for a geospatial understanding of the environmental impact of pre and post-agricultural activities.
- Temporal Trends in Greenhouse Gases Emissions: The study analysed temporal trends in greenhouse gases emissions from pre and post-production activities over a specific period (1990-2021), providing insights into how emissions have evolved over time.
- Analysis of Emissions in Pre and Post Production Items: The study conducted a detailed analysis of emissions associated with specific pre and post-production items, enabling a granular assessment of their environmental impact.
- Machine Learning Models: The study employed various machine learning models, including Linear Regressor, Gradient Boosting Algorithm, Random Forest, and Catboost Algorithms, to predict and understand emissions patterns, thereby facilitating a

quantitative comparison of the environmental impact of different stages of agrifood production.

- Cross-Validation and Model Evaluation: To ensure the robustness of the machine learning models, the study performed cross-validation and model evaluation, validating the predictive performance of the models and their suitability for addressing the research questions.
- Feature Importance: The study assessed the importance of features within the machine learning models, shedding light on the key factors driving emissions in pre and post-agricultural production activities.

Introduction to the Dataset:

The dataset is downloaded from FAOSTAT website and it provides comprehensive insights into emissions stemming from pre and post-food production activities, shedding light on their contributions to climate change. With a total of 20,760 entries, the dataset offers a robust foundation for understanding the environmental impact of various stages within the food production cycle.

Dataset Details

Total Rows: 20,760

Total Columns: 14

Column Descriptions:

Area: This column categorizes different European countries where emissions data was collected. The dataset includes emissions data from countries such as Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, and Sweden.

Element: This column delineates the type of emissions recorded. The dataset captures emissions of carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O).

Item: This column categorizes various activities or processes associated with food production and consumption, such as pesticides manufacturing, food processing, food transport, food packaging, food retail, food household consumption, solid food waste, domestic wastewater, and industrial wastewater.

Chapter 02:

Methodology and Analytical Framework

(This chapter would encompass the detailed description of the methods, tools, and analytical approaches used to address the research questions and achieve the stated objectives)

In this chapter we embark on a comprehensive exploration of the methods and analytical approaches employed to address the research hypothesis and achieve the stated aims and objectives. This chapter serves as a critical foundation for the subsequent analysis and findings, providing a detailed insight into the systematic process undertaken to investigate and compare emissions from pre and post-agricultural production activities. The methodology encompasses a range of analytical techniques, including feature engineering, geospatial analysis, temporal trend assessment, and the application of machine learning models, all of which are strategically aligned with the study's overarching goal of understanding the relative environmental impact of different stages within the agrifood system. By delving into the intricacies of the analytical framework, this chapter sets the stage for a rigorous and data-driven examination of agrifood production emissions, ultimately contributing to a more nuanced understanding of the factors driving environmental impact within the food supply chain.

2.1 Feature Engineering:

In the context of feature engineering, the categorization of items into 'Pre-Production Items' and 'Post Production Items' directly correlates with the emissions of greenhouse gases, including N_2O, CH_4, and CO_2. 'Pre-Production Items' primarily involve the manufacturing of inputs such as pesticides and fertilizers, which are associated with the emission of these gases during their production processes. On the other hand, 'Post Production Items' encompass a diverse range of activities within the food production chain, such as 'Food Processing', 'Food Transport', 'Food Packaging', 'Food Retail', 'Food Household Consumption', 'Solid Food Waste', 'Domestic Wastewater', and 'Industrial Wastewater', all of which can lead to the emission of N_2O, CH_4, and CO_2 at various stages. This categorization lays the groundwork for a detailed analysis of the emissions associated with each stage of agrifood production, providing a structured approach to understanding and comparing the environmental impact of pre and post-production activities.

2.2 Geographical Analysis using Interactive Plots:

The second aim of the study involves conducting a Geographical Analysis using Interactive Plots to visualize the emissions of N_2O, CH_4, and CO_2 from both pre- and post-production items. By leveraging interactive plots, the study aims to provide a spatial understanding of the distribution of these greenhouse gas emissions, allowing for a comprehensive assessment of their geographical impact.

2.2.1 Emission of Nitrogen di Oxide (NO2)

- **Exploring N2O Emissions from Pre-Production Items: An Interactive Plot Analysis (1990-2021)**

In the report, while the interactive plots cannot be directly manipulated, the provided numerical details and insights from the interactive plot analysis in VSCode offer a comprehensive understanding of the geographical distribution of N2O, CH4, and CO2 emissions across the included European countries.

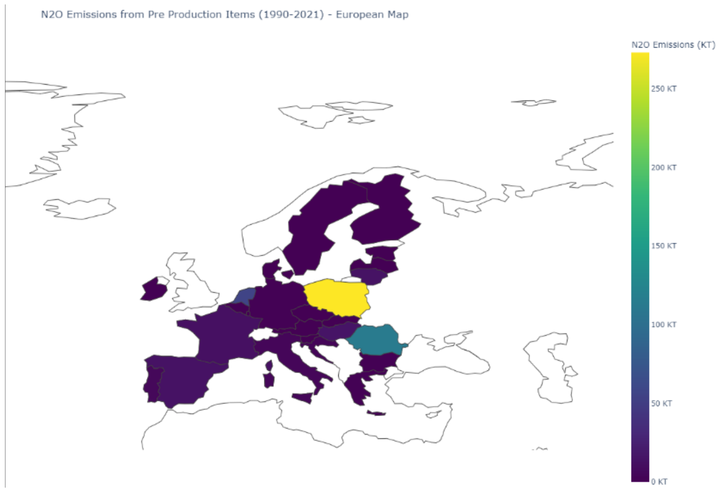


Figure showing NO2 emission in European countries from 1991 to 2021

For instance, the plot "Exploring N2O Emissions from Pre-Production Items: An Interactive Plot Analysis (1990-2021)" reveals that Poland has the highest N2O emissions at 273 KT, followed by Romania at 113.38 KT and the Netherlands at 56.10 KT.

- **<u>Exploring N2O Emissions from Post-Production Items: An Interactive Plot Analysis (1990-2021)</u>**

The N2O emissions from post-production items in the European countries exhibit notable variations. Germany emerges as the leading emitter, with a substantial 1945.13 Kt, signifying a significant environmental impact. Following Germany, France and Italy rank as the second and third highest emitters, with 147.52 Kt and 134.83 Kt, respectively. These insights underscore the importance of targeted emission reduction strategies in these countries to mitigate the environmental impact of post-production activities. Additionally, Spain and Poland register 97.2 Kt and 80.93 Kt, respectively.

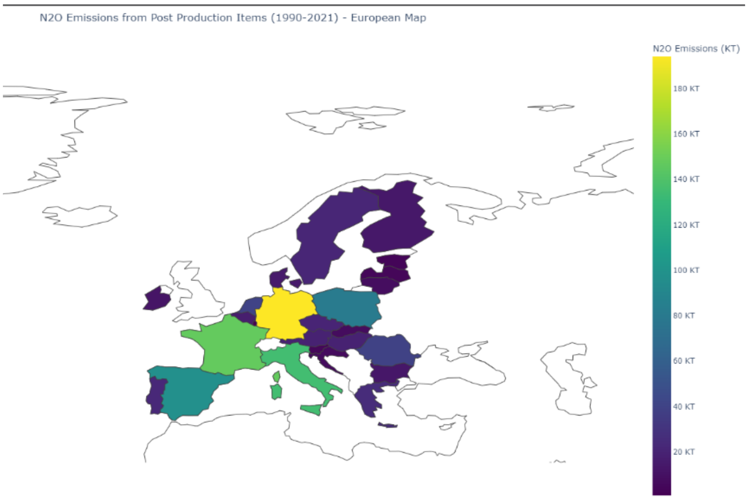


Figure shows N2O Emissions from Post-Production Items from 1990 to 2021

2.2.2 Emission of Methane (CH4)

- **Objecting CH4 Emissions from Pre-Production Items (1990-2021)**

The CH4 emissions choropleth map across European countries highlights France as the top emitter with 1.5 kT, followed closely by Germany and Italy at 1.31 kT and 1.34 kT, respectively. Spain ranks third with 0.868 kT, while the Netherlands and Romania follow at the fourth and fifth positions with emissions of 0.456 kT and 0.221 kT, respectively.

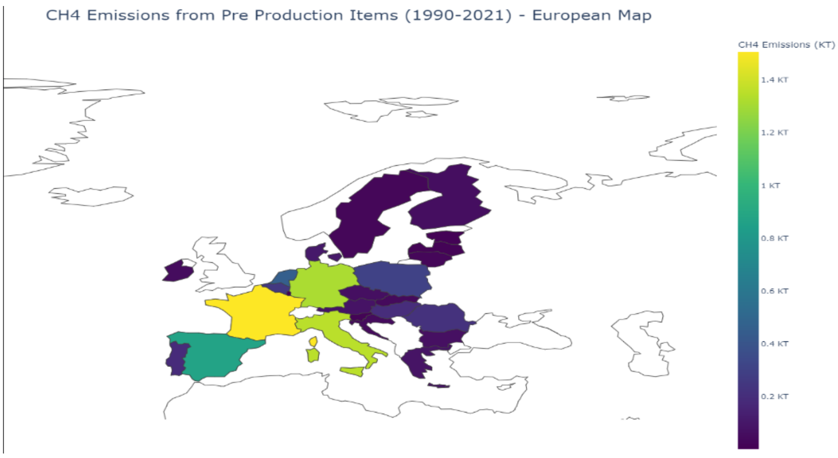


Figure shows CH4 Emissions from Pre-Production Items from 1990 to 2021

- **Objecting CH4 Emissions from Post-Production Items (1990-2021)**

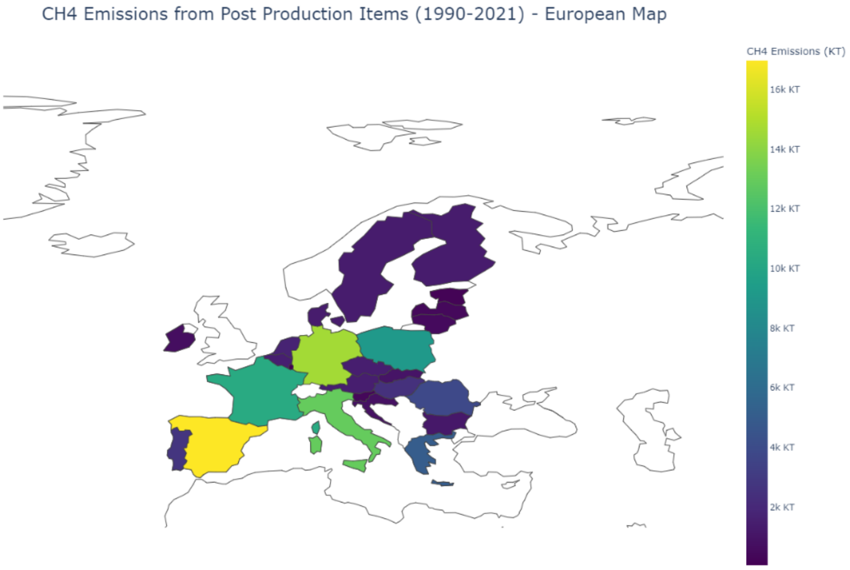


Figure shows CH4 Emissions from Post-Production Items from 1990 to 2021

The interactive map illustrating CH4 emissions from post-production items across European countries from 1990 to 2021 reveals notable trends. Spain emerges as the leading contributor with 16.97 KT of emissions, followed closely by Germany at 14.16 KT. Italy and France also feature prominently, with 12.92 KT and 10.33 KT, respectively. Despite ranking fifth, Poland still demonstrates a significant impact with 9.14 KT of emissions.

2.2.3 Emission of Carbon dioxide (CO2)

- **Envisioning CO2 Emissions from Pre-Production Items (1990-2021)**

The choropleth map visualizes CO2 emissions from pre-production items across European countries from 1990 to 2021. Poland emerges as the highest emitter, with CO2 emissions totaling 158 KT, followed closely by the Netherlands with 151.7 KT, and Germany with 89.81 KT. Spain and Italy also exhibit significant emissions, with values of 87.1 KT and 76.03 KT, respectively.

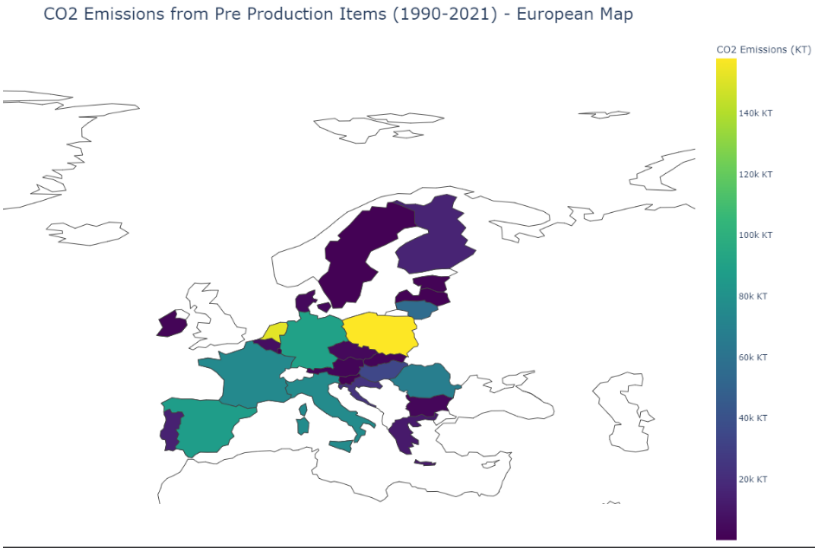


Figure shows CO2 Emissions from Pre-Production Items from 1990 to 2021

- **Envisioning CO2 Emissions from Post-Production Items (1990-2021)**

The choropleth map illustrates CO2 emissions from post-production items across European countries from 1990 to 2021. Germany leads as the highest emitter with 2.5 million KT, followed closely by Italy at 2.3 million KT. France also demonstrates significant emissions at 2.01 million KT, while Poland follows with 959.24 KT. Spain and the Netherlands exhibit emissions of 742.19 KT and 613.07 KT, respectively. Greece shows emissions totalling 285.32 KT.

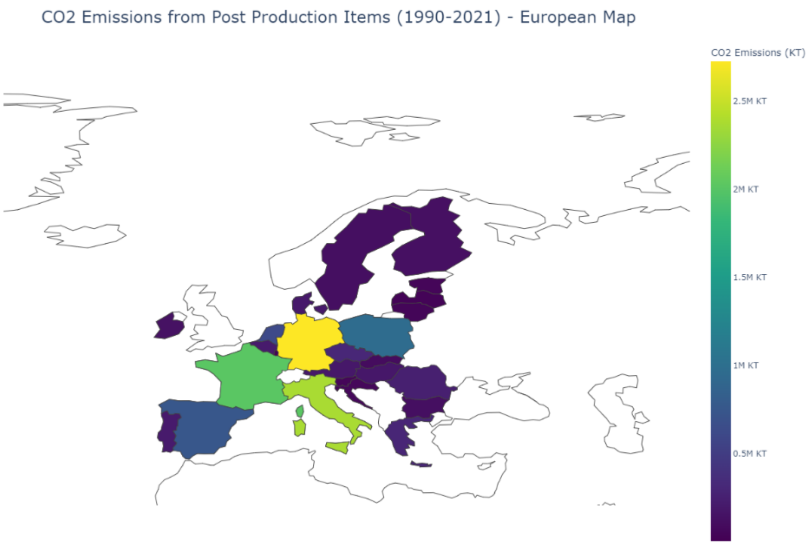


Figure shows CO2 Emissions from Post-Production Items from 1990 to 2021

2.3 Temporal Trends in Greenhouse Gas Emissions from Pre-Post Production Activities (1990-2021)

In recent decades, the discourse surrounding climate change and its impact on the environment has intensified, necessitating a critical examination of human activities contributing to greenhouse gas emissions. One significant area of focus within this discourse pertains to the temporal trends in greenhouse gas emissions stemming from pre-post production activities across various industrial sectors in Europe.

2.3.1 Overall N2O Pre and Post Production Emission Items in Euorpe (1990 - 2021)

- **Pre-Production Emission**

The analysis of N2O emissions from pre-production items spanning 1990 to 2021 indicates notable trends. Initially, in 1991, emissions were recorded at 0.27 KT, followed by a slight decline in the subsequent two years. However, by 1995, there was a significant increase, marking the beginning of an upward trend. This trend continued until around 2004, with a minor decrease thereafter. From 2005 onwards, emissions stabilized, fluctuating within the range of 0.4 to 0.61 KT.

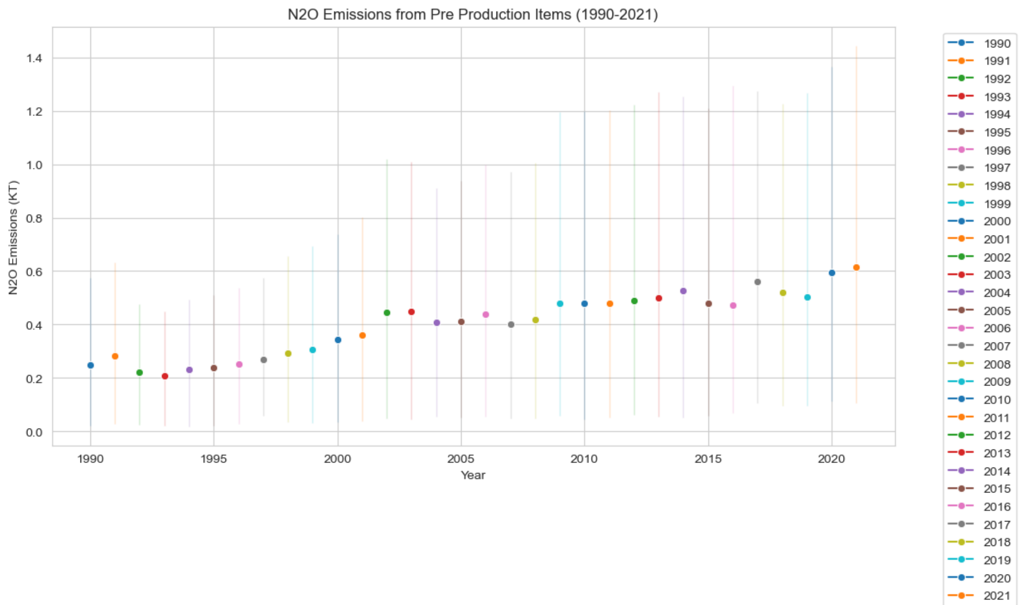


Figure showing NO2 emission from Pre-Production Items1991 to 2021

- **Post Production Emission**

The examination of N2O emissions stemming from post-production items over the period from 1990 to 2021 unveils discernible trends. Initially, in 1990 and 1991, N2O emissions reached their peak, registering at 0.23 KT. However, starting from 1992, there was a gradual decline, leading to emissions fluctuating within the range of 0.15 to 0.20 KT from 1993 onwards until the end of the observed period. This pattern suggests a stabilization in emissions following the initial surge, indicating potential adjustments in post-production practices or regulatory measures aimed at curbing N2O release from these activities over the analyzed timeframe.

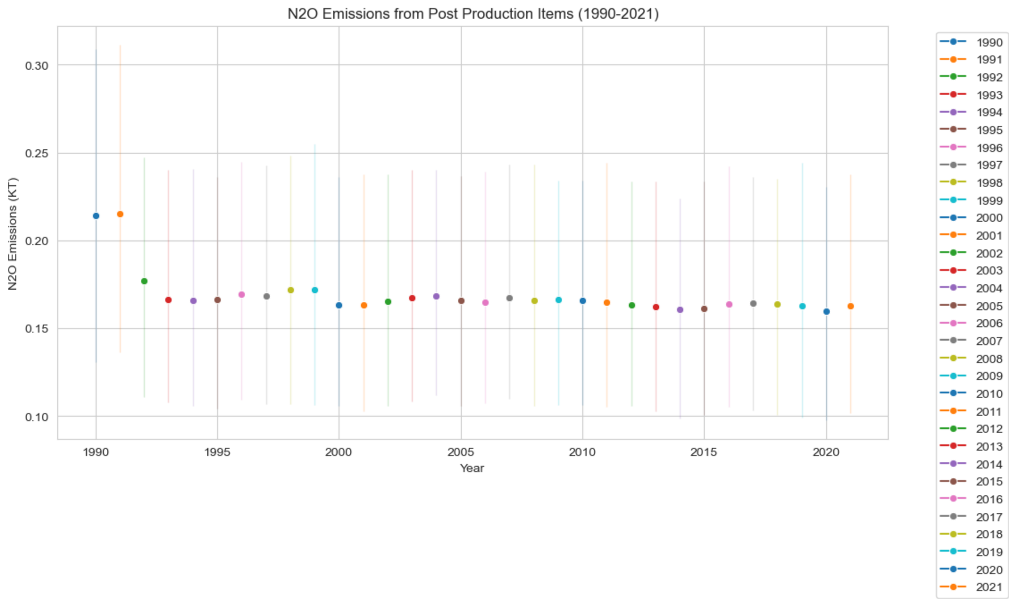


Figure showing NO2 emission from Post-Production Items1991 to 2021

The examination of N2O emissions stemming from post-production items over the period from 1990 to 2021 unveils discernible trends. Initially, in 1990 and 1991, N2O emissions reached their peak, registering at 0.23 KT. However, starting from 1992, there was a gradual decline, leading to emissions fluctuating within the range of 0.15 to 0.20 KT from 1993 onwards until the end of the observed period. This pattern suggests a stabilization in emissions following the initial surge, indicating potential adjustments in post-production practices or regulatory measures aimed at curbing N2O release from these activities over the analysed timeframe.

2.3.2 Comprehensive CH4 Emission from Pre and Post Production Items Categories in Europe from 1990 to 2021

- **Pre-Production Emission**

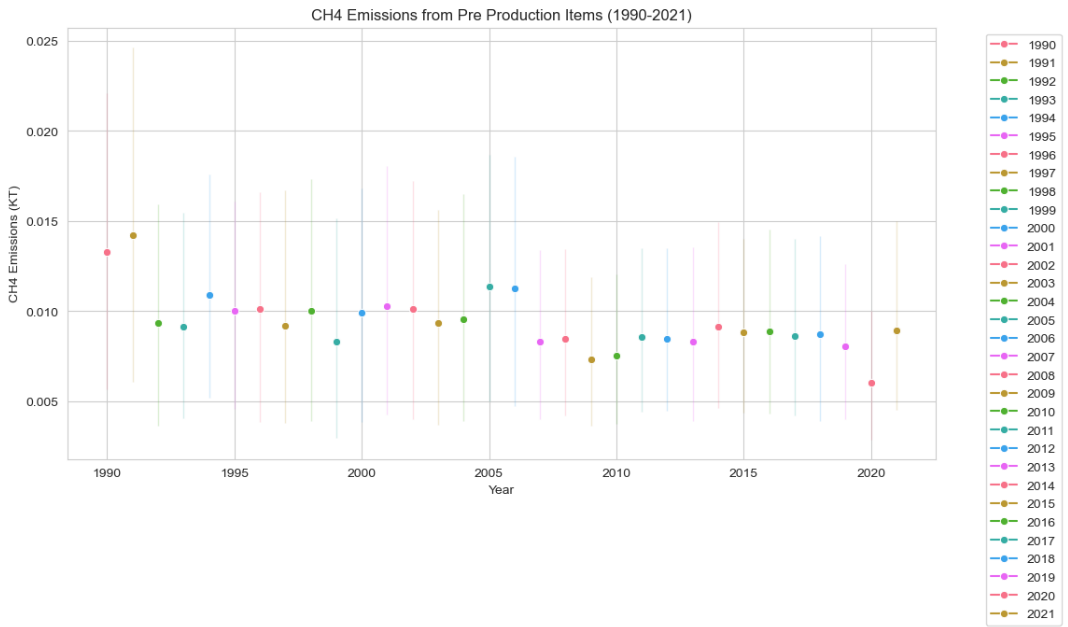


Figure showing CH4 emission from Pre-Production Items1991 to 2021

The analysis of CH4 emissions originating from pre-production items spanning the years 1990 to 2021 reveals distinctive patterns. Initially, in 1990, emissions commenced at 0.014 KT, experiencing a marginal increase to 0.0148 KT in 1991, followed by a sudden spike to 0.095 KT. Subsequently, from 1993 to 2004, emissions fluctuated within the range of 0.08 to 0.011 KT. However, in 2005, there was a notable uptick, reaching 0.12 KT, which was followed by a sharp decline to 0.075 KT. From then onwards until 2021, the emission levels stabilized, oscillating between 0.06 to 0.010 KT.

- **Post Production Emission**

The examination of CH4 emissions stemming from post-production items over the period from 1990 to 2021 reveals fluctuating trends. Initially, in 1990, CH4 emissions commenced at 20.75 KT, experiencing marginal variations hovering around 21 KT. However, from 1991 to 1995, there was a notable decline, with emissions reaching a low of 15 KT. Subsequently, there were fluctuations, with emissions fluctuating between approximately 18 KT in 2000 and declining to 11 KT by 2021.

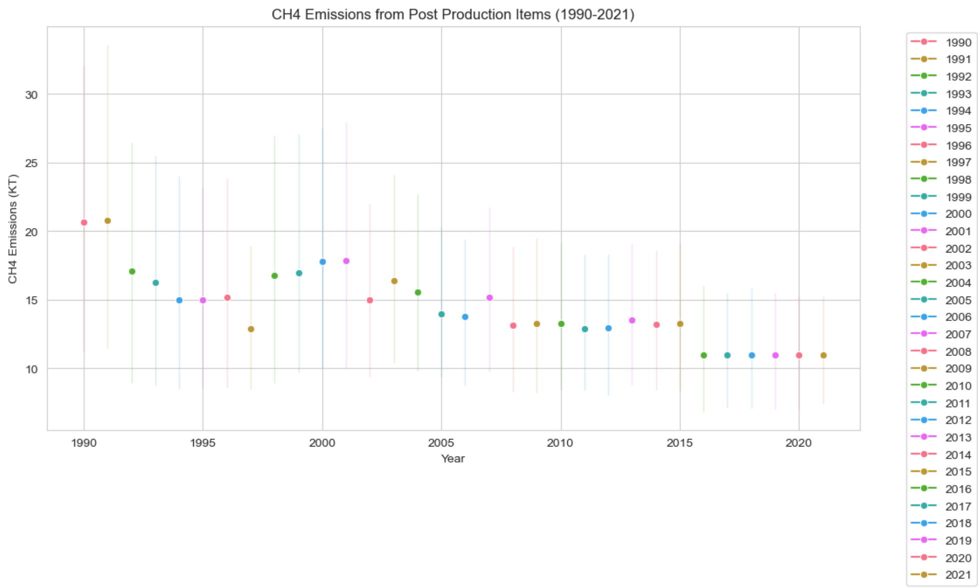


Figure showing CH4 emission from Post-Production Items1991 to 2021

2.3.3 Inclusive CO2 Emission from Pre and Post Production Items in Europe from 1990 to 2021

- **Pre-Production Emission**

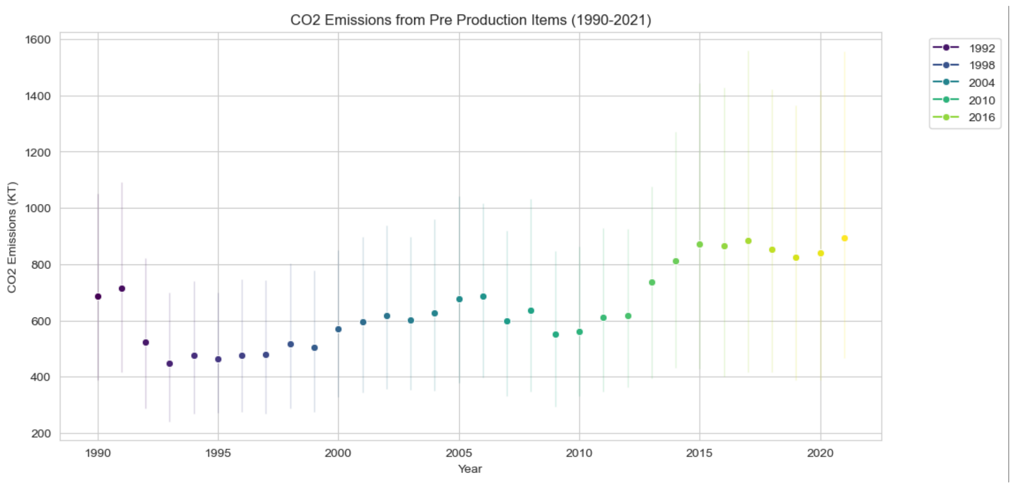
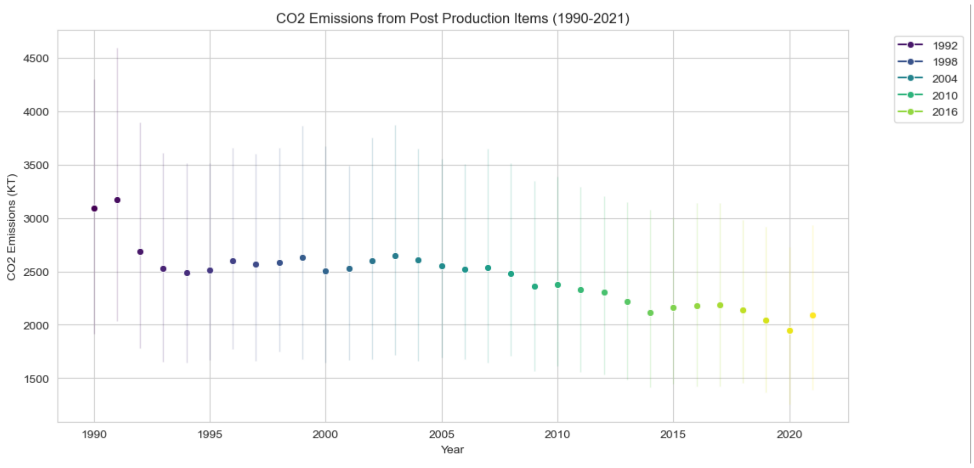


Figure showing CO2 emission from Post-Production Items1991 to 2021

The analysis of CO2 emissions stemming from pre-production items between 1990 and 2021 reveals dynamic fluctuations. Initially, in 1990, CO2 emissions were recorded at 700 KT, experiencing a slight increase the following year before declining sharply to 440 KT in 1993. Subsequently, after a slight rise in 1994, emissions fluctuated between 430 KT and 610 KT up to 2002. From then until 2012, there were fluctuations characterized by modest increases and decreases. However, starting in 2012, emissions began to rise significantly, culminating in a sharp increase around 2015, with values reaching approximately 900 KT. The pattern continued with fluctuations until 2021, where emissions were recorded at approximately 910 KT.

- **Post Production Emission**



The examination of CO_2 emissions originating from post-production items between 1990 and 2021 unveils fluctuating trends. In 1990, emissions commenced at 3100 KT, with a notable increase to 3200 KT in 1991. However, a descending pattern ensued from 1993 to 1995. Subsequently, from 1995 to 2009, emissions continued to decline with slight fluctuations. There were slight increases observed in 2015 and 2016, reaching 2200 KT and 2250 KT, respectively. However, emissions reached a record low in 2020, plummeting to approximately 1950 KT, before rebounding to 2150 KT in 2021.

2.4 Analysis of Emissions in Pre-Production Items

There are following two Items fall in pre-production Emission category from my data

- Pesticides Manufacturing
- Fertilizer Manufacturing

Pesticides manufacturing includes the industrial process of producing chemical substances or mixtures used to kill, repel, or control pests such as insects, weeds, fungi, and rodents to protect crops, livestock, and structures.

Emission of N_2O, CH_4 and CO_2 from Pesticides Manufacturing:

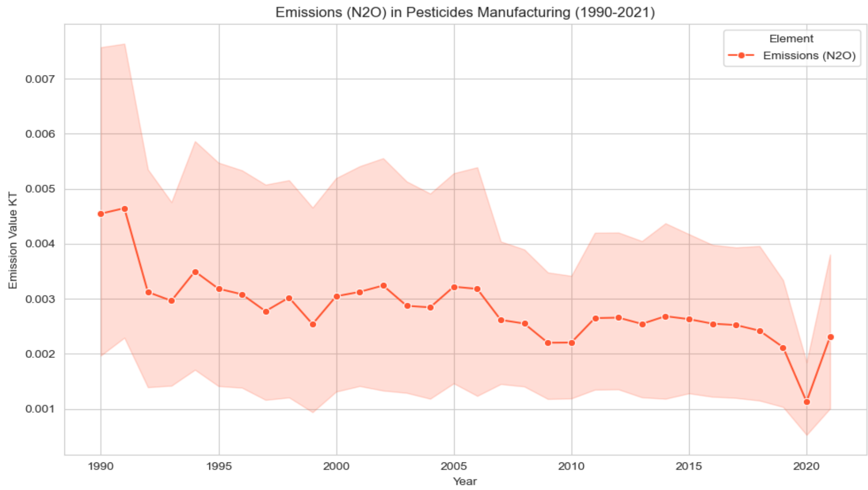


Figure shows Emission of N2O in Pesticides Manufacturing (1990 - 2021)

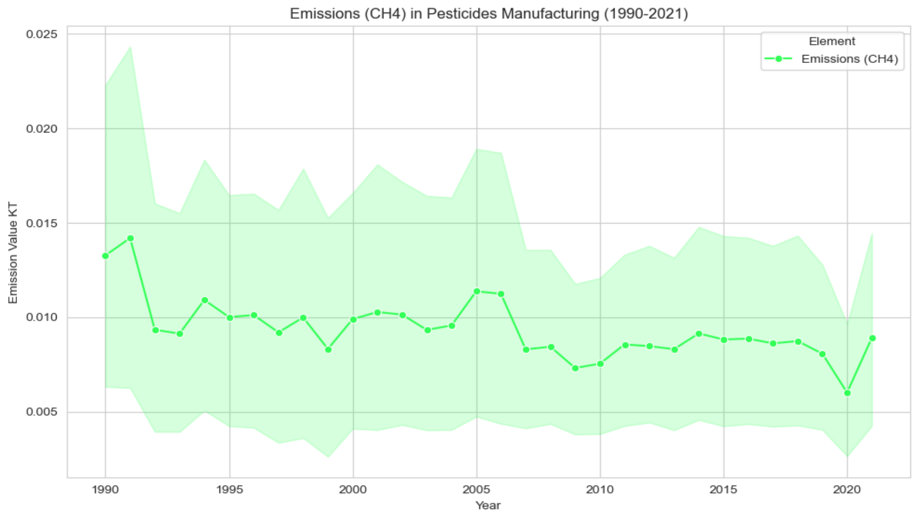


Figure shows Emission of CH4 in Pesticides Manufacturing (1990 - 2021)

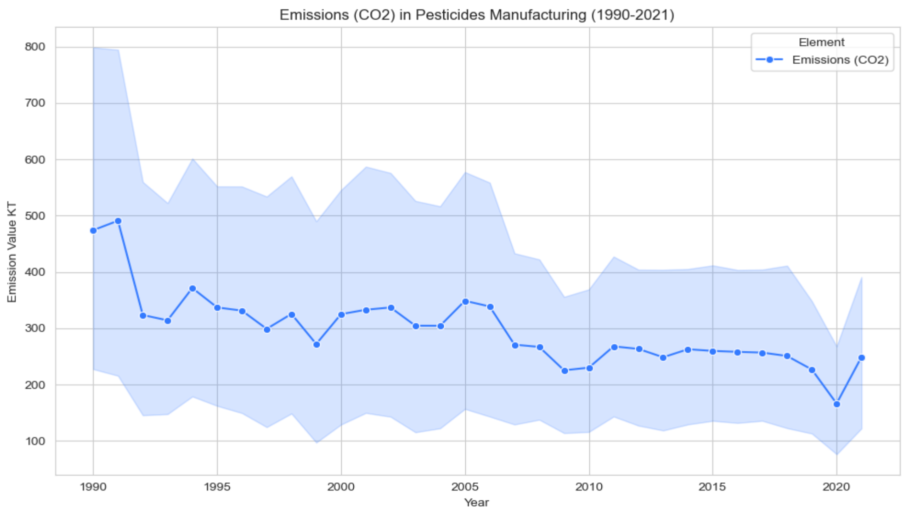


Figure shows Emission of CO2 in Pesticides Manufacturing (1990 - 2021)

The analysis of emissions from Pesticides Manufacturing between 1990 and 2021 unravels distinctive trends across N2O, CH4, and CO2. Beginning with N2O emissions, in 1990, the level was observed at 0.0045 KT, with a slight increase in 1991 followed by a rapid decline to 0.0031 KT by 1993. Subsequently, a fluctuating pattern emerged from 1995 to 2010, with values oscillating between 0.002 and 0.0034 KT. A notable rise occurred in 2011, maintaining a relatively smooth trajectory until 2019, where a significant drop to 0.0011 KT was recorded, before rising again to 0.00235 KT in 2021. For CH4 emissions, the trend mirrored N2O initially, with a starting value of 0.0135 KT in 1990, followed by a slight increase in 1991 and a rapid decline to 0.09 KT by 1993. Fluctuations persisted from 1995 to 2010, ranging between 0.010 and 0.075 KT. A notable increase was observed in 2005 to 0.012 KT, maintaining a relatively stable trajectory until 2019, with a significant drop to 0.06 KT, before rising to 0.085 KT in 2021. Concerning CO2 emissions, a similar pattern emerged with a starting value of 475 KT in 1990, rising to a record high of nearly 497 KT in 1991, before declining sharply to 315 KT in 1993. Fluctuations persisted from 1994 to 2010, with values ranging between 480 KT and 220 KT. A slight rise occurred in 2011 to 280 KT, maintaining relatively stable values until 2019, with a record low of 175 KT in 2020, followed by a rise to 250 KT in 2021.

Fertilizer Manufacturing The process of producing chemical or organic substances containing essential nutrients, such as nitrogen, phosphorus, and potassium, utilized to improve soil fertility and promote plant growth in agriculture and horticulture.

Emission of N2O, CH4 and CO2 from Fertilizer Manufacturing:

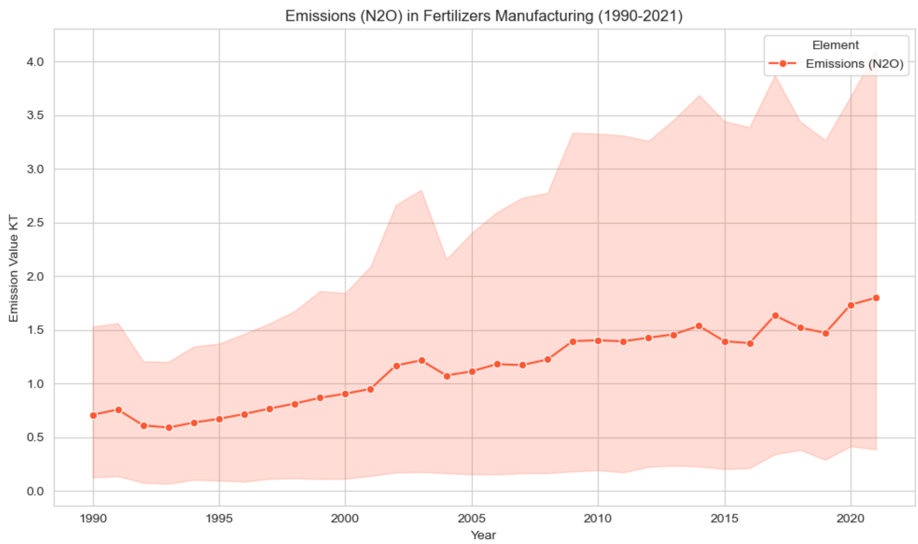


Figure shows Emission of N2O in Fertilizer Manufacturing (1990 - 2021)

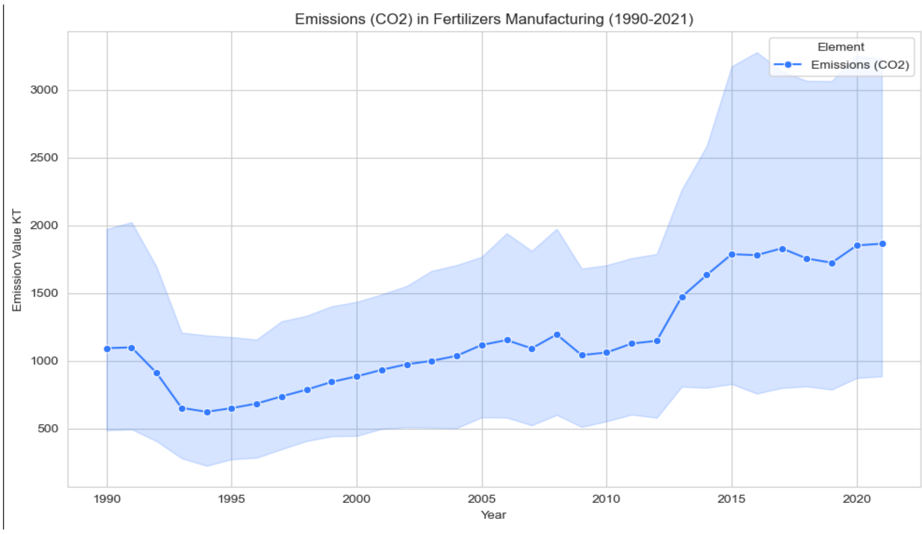


Figure shows Emission of CO2 in Fertilizer Manufacturing (1990 - 2021)

The analysis of emissions from Fertilizers Manufacturing between 1990 and 2021 reveals intriguing patterns in N2O and CO2 emissions. Notably, CH4 emissions were not observed in this sector, which could be attributed to the manufacturing processes not involving methane. Beginning with N2O emissions, in 1990, levels started at 0.75 KT, rising to 0.8 KT in 1991, followed by a slight decline to 0.6 KT in 1992, reaching a record low of 0.58 KT in 1993. However, a gradual increase ensued, with incremental rises leading to a peak of 1.8 KT in 2021, with fluctuations observed over the years, particularly notable in the early 2000s and between 2014 and 2020. Regarding CO2 emissions, a similar pattern emerged, with levels starting at 1100 KT in 1990, rising to 1150 KT in 1991, then sharply declining to 650 KT in 1993, reaching a record low of 600 KT. Subsequently, a steady increase was observed, with peaks of 1850 KT in 2021, with fluctuations evident, particularly notable in the early 2000s and between 2013 and 2015.

2.5 Green House Gases Emissions in Post Production Items

N2O:

The analysis of N2O emissions from post-production items reveals varied emission levels across different categories. Among the categories examined, domestic wastewater exhibits the highest emission level, reaching approximately 0.75 KT, indicating a significant contribution to N2O emissions from post-production activities. Following closely behind is food processing, with emissions recorded at 0.35 KT, highlighting its substantial impact on N2O release. Industrial wastewater also contributes to emissions, albeit at a lower level, with emissions recorded at 0.1 KT. Food transport and food household consumption show moderate emissions, at 0.195 KT and 0.025 KT, respectively, indicating their contribution to post-production N2O emissions. Food retail and food packaging exhibit relatively lower emissions, recorded at 0.01 KT and 0.005 KT, respectively.

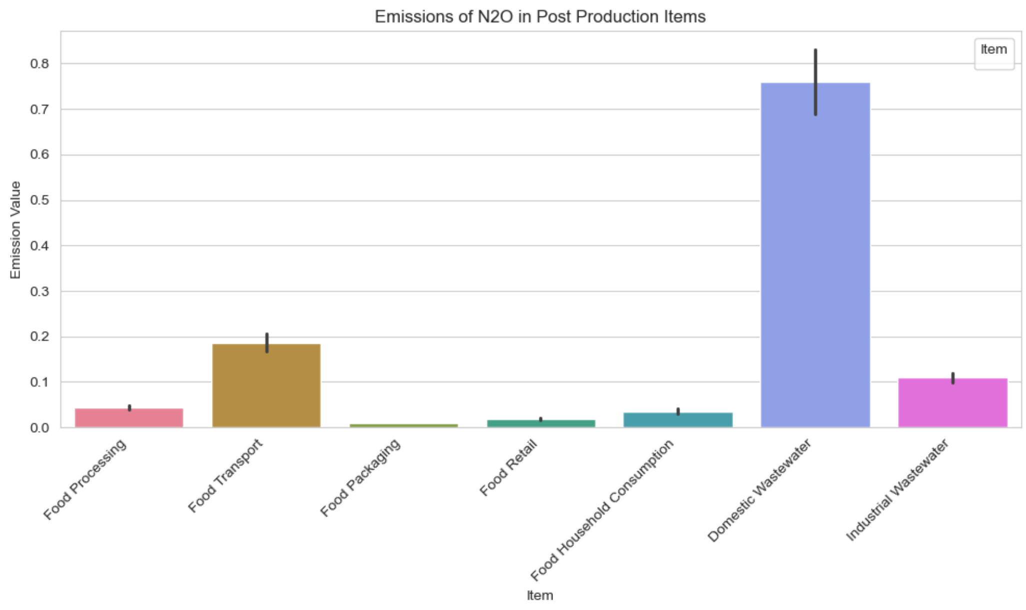


Figure shows Emission of N2O from different source types (1990 - 2021)

CH4:

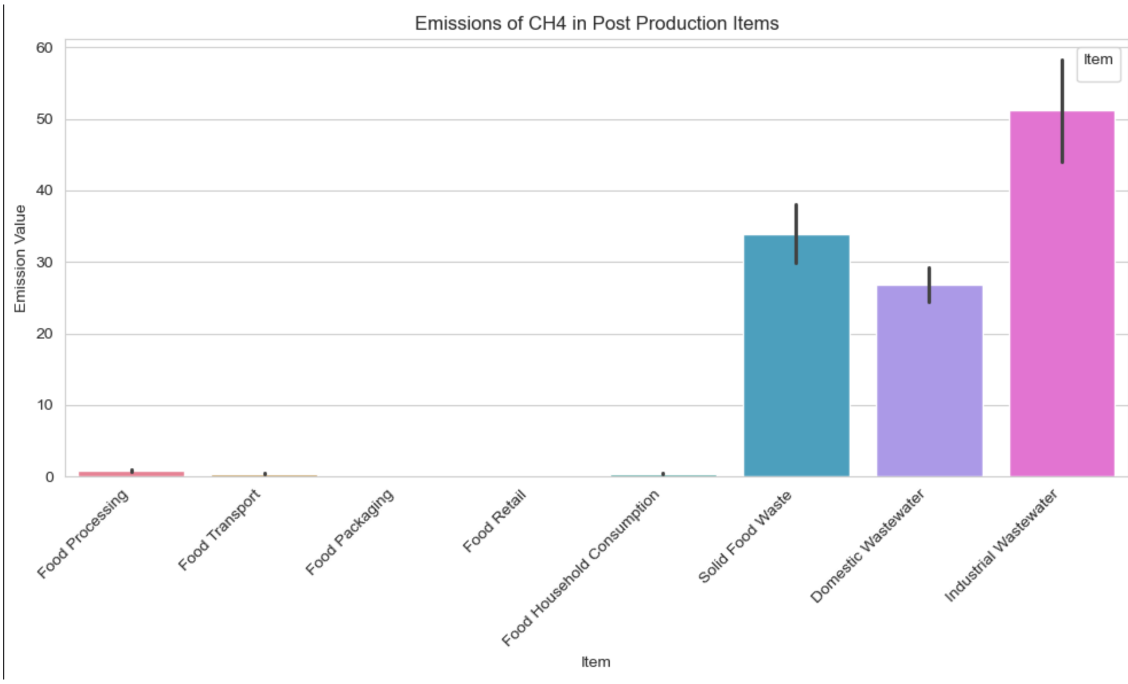


Figure shows Emission of CH4 from different source types (1990 - 2021)

The analysis of CH4 emissions from post-production items reveals significant variation in emission levels across different categories. Notably, industrial wastewater emerges as the primary contributor to CH4 emissions, recording the highest level at 51 KT. Solid food waste also exhibits substantial emissions, with levels reaching 33.5 KT, indicating its significant contribution to CH4 release from post-production activities. Domestic wastewater follows closely behind, with emissions recorded at 27 KT, further underscoring its role in CH4 emissions. In contrast, emissions from other categories, including food processing, food transport, food packaging, food retail, and food household consumption, are relatively low, ranging between 0 and 0.005 KT.

CO2:

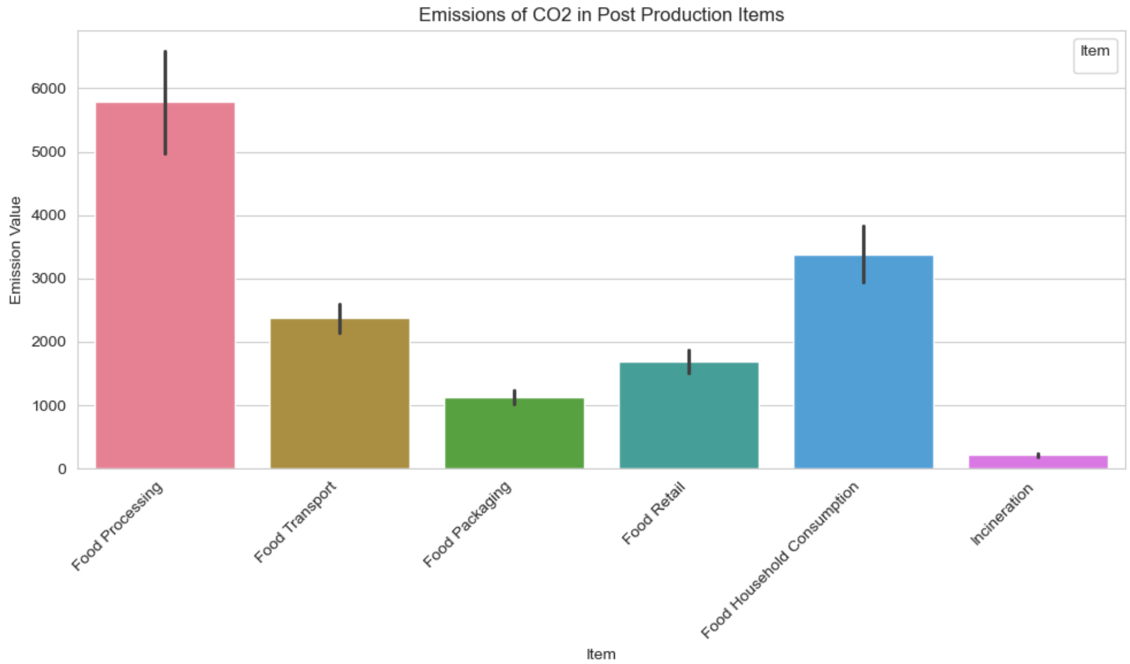
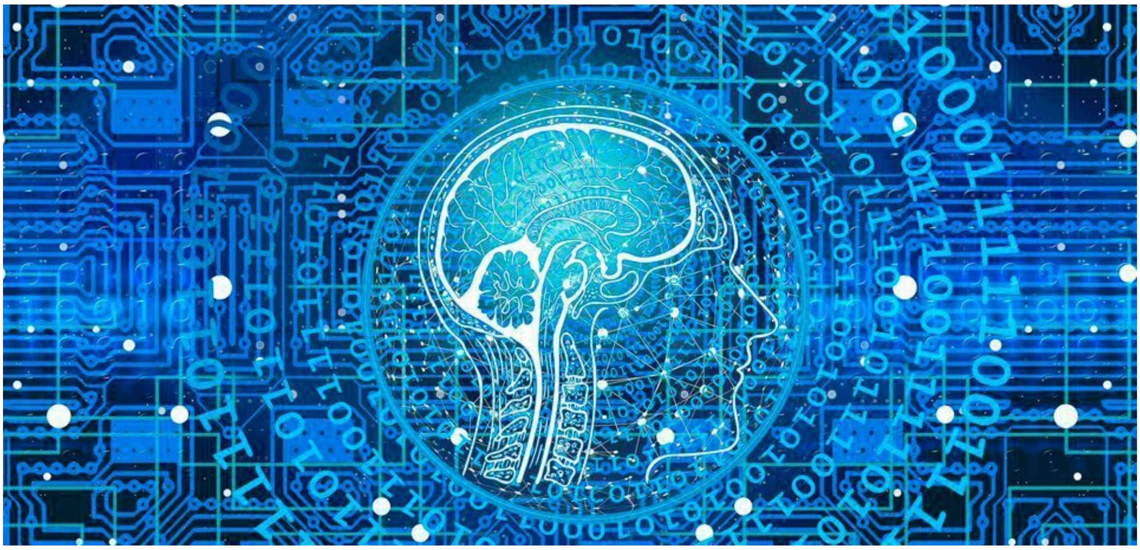


Figure shows Emission of CO2 from different source types (1990 - 2021)

The analysis of CO2 emissions from post-production items unveils significant variations in emission levels across different categories. Among the categories examined, food processing emerges as the primary contributor to CO2 emissions, recording the highest level at 5900 KT. This indicates the substantial impact of food processing activities on CO2 release within the post-production phase. Following closely behind are food household consumption and food transport, with emissions recorded at 3300 KT and 2200 KT, respectively, highlighting their significant contributions to CO2 emissions. Food retail also exhibits substantial emissions, with levels reaching 1700 KT, indicating its role in contributing to CO2 emissions from post-production activities. Additionally, food packaging contributes to CO2 emissions, with levels recorded at 1050 KT, underscoring the importance of considering packaging materials and waste management practices in mitigating CO2 emissions. Incineration, while comparatively lower in emissions at 200 KT, still contributes to CO2 release within the post-production phase.



Chapter 03:

Harnessing Intelligence: Building and Optimizing Machine Learning Models

(In this chapter, we explore the implementation of various machine learning models, delve into the significance of cross-validation and model evaluation techniques for robust performance assessment, and investigate feature importance methodologies to unveil key factors driving predictive outcomes)

Machine Learning:

Machine learning is a branch of artificial intelligence focused on the development of algorithms and statistical models that enable computers to learn from and make predictions or decisions based on data, without being explicitly programmed. It encompasses a range of techniques that allow machines to identify patterns, extract insights, and make intelligent decisions, ultimately enabling automation, prediction, and optimization across diverse domains.

3.1 Mechanism of applying Machine Learning Models in my notebook:

The Machine learning models are applied in following steps in one snippet code.

In this process, the data is first prepared by separating the features from the target variable, followed by splitting the dataset into training and testing sets for model evaluation.

Pipelines are then created for each machine learning model, incorporating preprocessing steps such as one-hot encoding categorical features and standardizing numeric features. These pipelines are designed to seamlessly integrate data preprocessing with model training. The algorithms used include Linear Regression, Gradient Boosting, Random Forest, and CatBoost.

Hyperparameters, which govern the learning process of the models, are defined for each model, specifying different values for tuning.

Grid search cross-validation is employed to systematically explore the hyperparameter space and find the optimal combination for each model, utilizing 5-fold cross-validation and optimizing for the negative mean squared error metric.

Finally, the performance of each model is evaluated using various metrics including Root Mean Squared Error (RMSE), R-squared (R2) score, Mean Squared Error (MSE), and Mean Absolute Error (MAE).

This comprehensive approach enables the systematic application, comparison, and evaluation of multiple machine learning models to select the most suitable one for the given dataset.

What is achieved in this Machine Learning Model Training:

This code snippet performs a machine learning workflow for regression tasks. Firstly, it prepares the dataset by separating features and the target variable. The features are stored in the variable X, while the target variable is stored in y.

Next, the data is split into training and testing sets using the train_test_split function from sklearn.model_selection, with 80% of the data used for training and 20% for testing.

Subsequently, pipelines are created for each model, incorporating preprocessing steps such as one-hot encoding for categorical features and standardization for numeric features using ColumnTransformer from sklearn.compose.

The models trained include Linear Regression, Gradient Boosting, Random Forest, and CatBoost. Each model's hyperparameters are tuned using grid search cross-validation (GridSearchCV from sklearn.model_selection), optimizing for the negative mean squared error.

Finally, the performance of each model is evaluated using metrics such as Root Mean Squared Error (RMSE), R-squared (R2) score, Mean Squared Error (MSE), and Mean Absolute Error (MAE). These metrics provide insights into how well each model performs in predicting the target variable based on the given features.

Score Comparison:

Linear Regression:

The RMSE (Root Mean Squared Error) is relatively high compared to other models, indicating a higher level of error in predictions.

The R^2 score suggests that the linear regression model explains only about 23% of the variance in the target variable.

Gradient Boosting:

The RMSE is significantly lower than that of linear regression, indicating better predictive performance.

The R^2 score is much higher, around 91%, suggesting that the gradient boosting model explains a substantial portion of the variance in the target variable.

Both the MSE (Mean Squared Error) and MAE (Mean Absolute Error) are lower compared to linear regression, indicating better model performance.

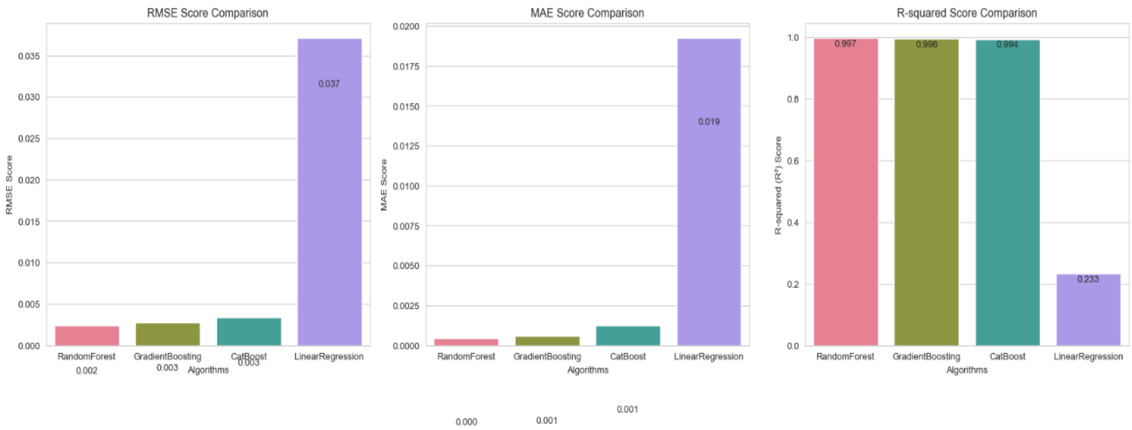


Figure shows comparison score of different machine learning models

Random Forest:

The RMSE is even lower than that of gradient boosting, indicating excellent predictive performance.

The R^2 score is very high, approximately 95%, suggesting that the random forest model explains most of the variance in the target variable.

Both the MSE and MAE are lower than those of both linear regression and gradient boosting, indicating superior model performance.

CatBoost:

The RMSE is the lowest among all models, indicating the best predictive performance.

The R^2 score is extremely high, around 99%, suggesting that the CatBoost model explains almost all of the variance in the target variable.

Both the MSE and MAE are the lowest among all models, indicating the best model performance in terms of accuracy and error.

Overall, it appears that ensemble methods such as Gradient Boosting, Random Forest, and CatBoost outperform linear regression in terms of predictive accuracy. Among them, CatBoost demonstrates the best performance with the lowest RMSE, highest R^2 score, and lowest MSE and MAE values, indicating its superiority in capturing the underlying patterns in the data.

3.2 Cross-Validation and Model Evaluation:

Cross-validation is a technique used to assess the performance of a machine learning model. The basic idea behind cross-validation is to split the dataset into multiple subsets or folds. The model is then trained on a portion of the data (training set) and evaluated on the remaining portion (validation set). This process is repeated multiple times, with each fold serving as both the training and validation set exactly once.

Model evaluation involves assessing the performance of the trained model using various metrics such as accuracy, precision, recall, F1-score, mean squared error (MSE), depending on the type of problem (classification or regression) and the specific requirements of the task. These metrics help to quantify how well the model is performing and guide the selection of the best-performing model for deployment or further optimization.

Let’s move to the Model:

The dataset is divided into training and testing sets using the train_test_split function from sklearn.model_selection. 80% of the data is used for training, while 20% is reserved for testing.

The categorical features ('Area', 'Element', 'Item Categories') are one-hot encoded using OneHotEncoder within a ColumnTransformer. This preprocessing step prepares the data for modeling.

Three pipelines are defined, each comprising a preprocessing step followed by a machine learning model: RandomForestRegressor, GradientBoostingRegressor, and CatBoostRegressor. These pipelines encapsulate the entire modeling process, making it easier to manage and reproduce.

A custom scorer for Root Mean Squared Error (RMSE) is defined using make_scorer. This scorer is used to evaluate model performance during cross-validation.

Cross-validation is performed using cross_val_score with 5 folds. Negative RMSE scores are converted to positive, and the mean RMSE is calculated for each model. This provides an estimate of the model's performance on unseen data.

Each pipeline is fitted to the entire training data, and predictions are made on the test set. Performance metrics such as RMSE, MAE, and R-squared are calculated to assess how well the models generalize to new data.

Overall, this code snippet demonstrates a systematic approach to building, evaluating, and comparing regression models using cross-validation and multiple evaluation metrics.

What is achieved in this Cross Validation and Model Evaluation:

In this code, a machine learning workflow for regression tasks is implemented using three different algorithms: RandomForest, GradientBoosting, and CatBoost. The data is split into training and testing sets using the train_test_split function. Preprocessing steps are defined using ColumnTransformer to one-hot encode categorical features ('Area', 'Element', 'Item Categories') with OneHotEncoder. Three pipelines are created, each comprising a preprocessing step followed by a machine learning model: RandomForestRegressor, GradientBoostingRegressor, and CatBoostRegressor. A custom scorer for Root Mean Squared Error (RMSE) is defined using make_scorer. Cross-validation is performed using cross_val_score with 5 folds. Negative RMSE scores are converted to positive, and the mean RMSE is computed for each model.

Score Comparison

Performance Comparison: The Root Mean Squared Error (RMSE) values from cross-validation for RandomForest, GradientBoosting, and CatBoost are quite similar, indicating that these models perform comparably well on unseen data during cross-validation.

Model Generalization: The RMSE values on the test set for all three models are consistent with the cross-validation results, suggesting that the models generalize well to unseen data.

Mean Absolute Error (MAE): The MAE values on the test set are relatively low for all models, indicating that, on average, the models' predictions are close to the actual values.

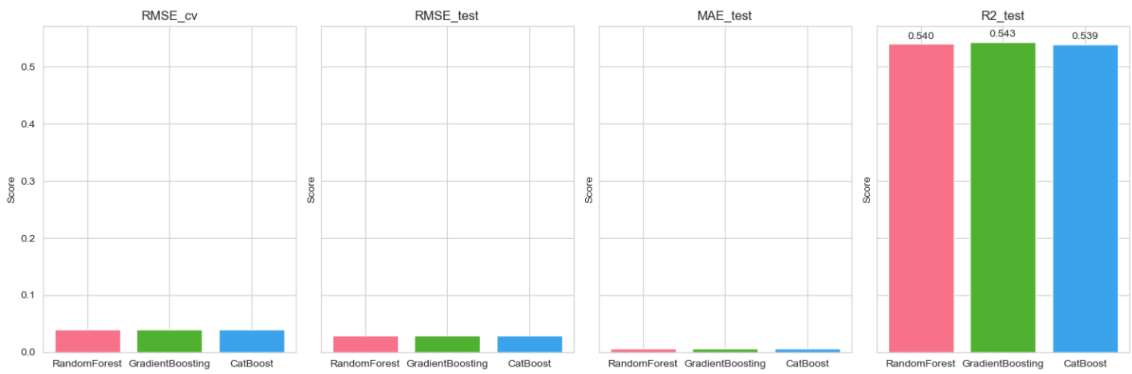


Figure shows Comparison of Model Performance Metrics

R-squared (R2): The R-squared values on the test set for all models are moderate, indicating that the models explain a moderate amount of variance in the target variable.

Comparative Performance: There are no significant differences in performance among RandomForest, GradientBoosting, and CatBoost based on the RMSE, MAE, and R2 values.

However, GradientBoosting shows a slightly lower RMSE and MAE compared to the other models, indicating slightly better predictive performance.

Model Complexity: CatBoost, being a gradient boosting algorithm, might have a more complex internal mechanism compared to RandomForest. However, this does not necessarily translate to better performance in this scenario.

Further Analysis: To gain more insights, it would be beneficial to investigate feature importances and partial dependence plots to understand which features are most influential in predicting emissions. Additionally, considering different hyperparameter settings and model ensembling techniques could potentially improve model performance further.

Overall, all three models appear to provide reasonable performance in predicting emissions, with GradientBoosting showing a slight edge in predictive accuracy. However, further analysis and experimentation may be necessary to optimize the models further.

3.3 Feature Selection:

This code demonstrates the process of feature importance calculation and visualization using a RandomForestRegressor model in a machine learning workflow:

One-Hot Encoding: First, categorical features in the training data X_train are encoded using OneHotEncoder from sklearn.preprocessing. This step transforms categorical features into a binary format suitable for machine learning models.

Model Initialization: A RandomForestRegressor model is instantiated.

Model Training: The RandomForestRegressor model is trained using the encoded features X_train_encoded and the target variable y_train. This step involves fitting the model to learn the patterns in the training data.

Feature Importance Calculation: After training, the feature importances are calculated using the feature_importances_ attribute of the trained model. This attribute provides information about the relative importance of each feature in predicting the target variable.

Feature Names Retrieval: The feature names are retrieved using the get_feature_names_out method of the encoder, providing the original feature names from X_train.columns.

Sorting Features: Feature importances and feature names are sorted in descending order based on their importance scores. This step ensures that the most important features appear at the top of the plot.

Plotting Feature Importances: Finally, a bar plot is created to visualize the feature importances. The x-axis represents the features, while the y-axis represents their respective importance scores. The plot provides insights into which features have the most significant impact on the model's predictions.

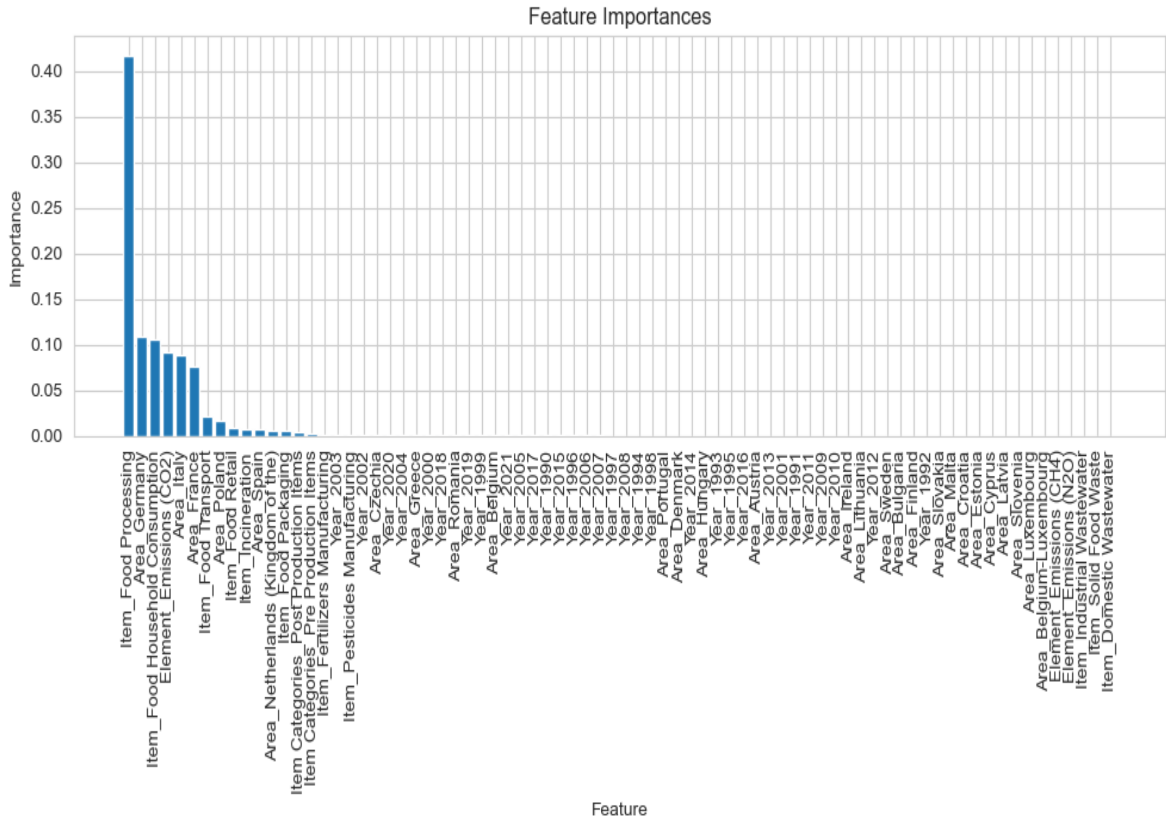


Figure shows The Feature Importances in RandomForestRegressor Model

Insights:

Based on the feature importances provided in the bar plot:

Item Food Processing: This feature has the highest importance, touching a value of 0.425. It indicates that emissions associated with food processing activities have the most significant impact on the model's predictions of overall emissions levels.

Area Germany: The feature representing emissions data from Germany follows closely, with an importance value touching 0.11. This suggests that emissions originating from Germany contribute significantly to the model's predictions, although to a lesser extent compared to food processing activities.

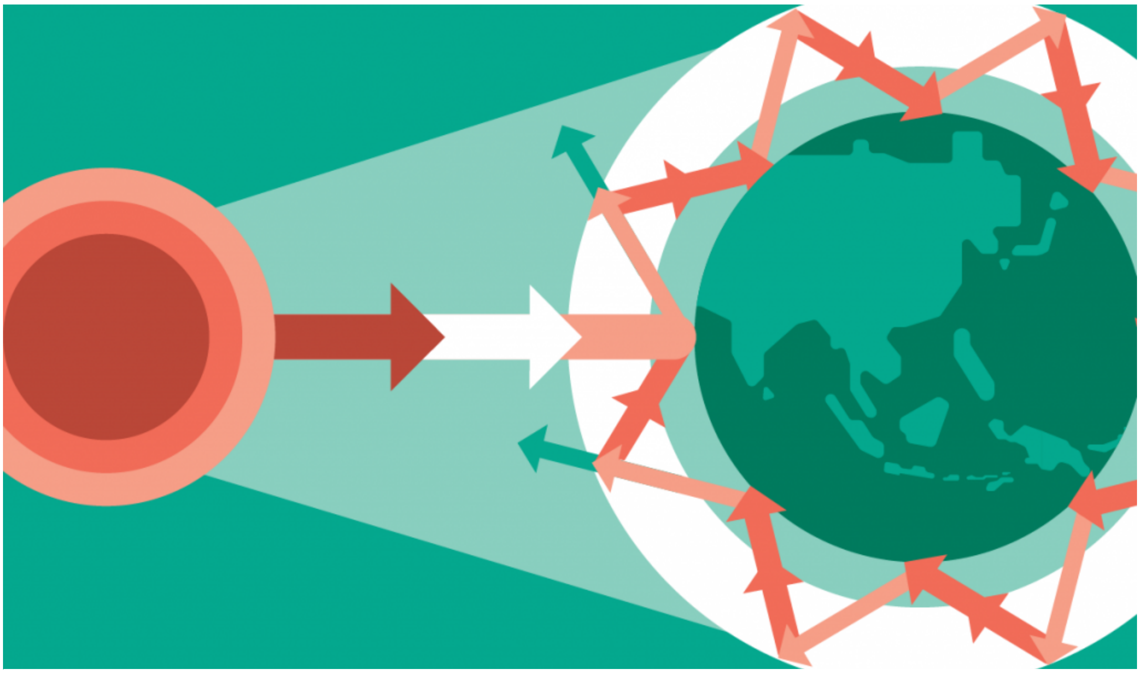
Item Food Household Consumption: The importance of emissions associated with household food consumption, represented by this feature, is notable, touching a value of 0.10. This indicates that household consumption patterns related to food have a moderate impact on overall emissions levels.

Element Emissions (CO_2): This feature, representing emissions of carbon dioxide (CO_2), has a substantial importance value of 0.09. CO_2 emissions are a major contributor to overall greenhouse gas emissions, and their significance in the model underscores their role in predicting emissions levels.

Area Italy: Emissions data from Italy also feature prominently in the model, with an importance value touching 0.85. This suggests that emissions originating from Italy make a significant contribution to the model's predictions, indicating the importance of considering Italy's emissions profile in overall emissions management strategies.

Area France: Similarly, emissions data from France are significant, with an importance value touching 0.75. France's emissions profile plays a notable role in the model's predictions, highlighting the importance of addressing emissions from this geographical region.

These insights provide valuable information about the key factors driving emissions levels, emphasizing the importance of addressing emissions from food processing activities, specific geographical regions like Germany, Italy, and France, as well as focusing on reducing CO_2 emissions to mitigate overall emissions levels effectively.



Chapter 04:

Summary and Mitigation Strategies: Addressing Challenges and Building Resilience

(In this chapter, we delve into a comprehensive analysis aimed at summarizing key insights and proposing effective mitigation strategies. Leveraging interactive plots, we explore Pre and Post Production trends based on sophisticated feature engineering techniques and insightful analysis of the Items column. Utilizing advanced Machine Learning models, coupled with robust cross-validation and thorough model evaluation, we derive actionable insights to inform strategic decision-making and enhance operational efficiency)

Summary and Mitigation:

The analysis encompasses a detailed examination of greenhouse gas emissions from both pre- and post-production activities across various sectors in Europe, spanning the years 1990 to 2021. Through interactive plots and comprehensive data visualization, distinct temporal trends and geographical distributions of emissions for N_2O, CH_4, and CO_2 are elucidated. Notable findings include the identification of key contributors to emissions within different categories, such as pesticides and fertilizer manufacturing, as well as post-production activities like food processing and industrial wastewater.

Mitigation strategies can be formulated based on these insights to address the identified sources of emissions. For instance, targeted regulatory measures and technological advancements can be implemented in the agricultural sector to reduce N_2O emissions from fertilizer and pesticide manufacturing. Similarly, waste management practices and process optimization in food processing and industrial sectors can help mitigate CH_4 emissions from post-production activities. Additionally, investments in renewable energy sources and adoption of sustainable practices in transportation and packaging can contribute to curbing CO_2 emissions. Overall, a multifaceted approach involving policy interventions, technological innovation, and behavioural changes is essential to effectively mitigate greenhouse gas emissions and mitigate their environmental impact in the European context.

Harnessing Machine Learning for Greenhouse Gas Emission Mitigation Strategies:

Machine learning models play a crucial role in supporting these mitigation efforts by providing predictive insights and facilitating decision-making processes. By analyzing historical emissions data and identifying patterns, models can forecast future emissions trends, allowing policymakers and stakeholders to proactively implement measures to reduce emissions. Additionally, models can help optimize resource allocation and prioritize interventions based on their potential impact on emissions reduction. Overall, a combination of policy interventions, technological innovation, and behavioral changes supported by machine learning models is essential for effectively mitigating greenhouse gas emissions and minimizing their environmental impact in the European context and beyond.