Data-Driven Insights into the Environmental Impact of the Agri-Food System: An Analysis Using SPSS, Python, and PySpark

Mengzhe Zhao
Faculty of Science
The University of Auckland
Auckland, New Zealand
mzha910@aucklanduni.ac.nz

***Abstract* — This research delves into the environmental impact of the agrifood system, with a specific focus on greenhouse gas emissions. Utilising a dataset and applying data mining techniques through tools like SPSS, Python, and PySpark, the study identifies key relationships and patterns within the agrifood system. Linear and random forest regression models highlight the linear relationships between certain agrifood variables and greenhouse gas emissions, emphasizing the significant influence of the food retail sector. Clustering models, such as K-Means and Agglomerative clustering, reveal inherent groupings within the system, potentially indicating different farming practices or regions with similar environmental impacts. The research underscores the need for sustainable practices, especially within the food retail industry, and suggests targeted interventions based on data clustering. The study also advocates for the transition to eco-friendly transportation methods and the promotion of carbon sequestration in agriculture. The proposed actions, grounded in data-driven insights, aim to provide stakeholders with actionable steps to mitigate the environmental impact of the agrifood system.**

Keywords — Agrifood System, Greenhouse Gas Emissions, Data Mining, Linear Regression, Random Forest Regression, K-Means Clustering, Agglomerative Clustering, Sustainable Practices

I. Introduction

Data mining plays a crucial role in data science, offering the ability to extract meaningful patterns from extensive datasets. A pressing global concern is the impact of CO2 emissions on global temperatures, particularly from the agri-food sector. This relationship, deeply connected to the broader issue of climate change, requires thorough analysis.

This research focuses on a dataset detailing CO2 emissions and their associated temperature changes. The primary goal is to understand the correlation between these emissions and temperature fluctuations. Additionally, the study aims to evaluate the contributions of different countries to these changes. A combination of tools was employed to achieve these objectives: SPSS for its statistical capabilities, Python for detailed data analysis, and PySpark for managing large-scale data.

By leveraging these tools, this research seeks to provide a comprehensive understanding of the interplay between CO2 emissions and global temperature variations.

II. Practical Problem

The world is witnessing unprecedented climate change, with rising global temperatures being a primary concern. This temperature rise is closely linked to human activities, especially the emissions of greenhouse gases like carbon dioxide (CO2). The agri-food sector, a vital component of global economies, plays a significant role in these emissions (LavagnedOrtigue, n.d.). Understanding the extent of these emissions and their direct impact on temperature is essential for policymakers, businesses, and communities.

From a business perspective, the agri-food sector faces dual challenges. On one hand, there is a growing demand for food due to the increasing global population. On the other, there is an urgent need to reduce CO2 emissions to combat climate change.

Balancing these demands requires a deep understanding of the relationship between agri-food activities, CO2 emissions, and the resulting temperature changes (LavagnedOrtigue, n.d.).

Furthermore, the influence of various countries in this scenario must be considered. Different nations have varied agricultural practices, levels of industrialisation, and consumption patterns. This diversity leads to different levels of emissions and, consequently, varied impacts on global temperatures.

In essence, the practical problems this research addresses are:

- Quantifying the correlation between CO2 emissions from the agri-food sector and the resulting rise in global temperatures.
- Analysing the role of different countries in this context based on their aggregated data on emissions and temperature change.

By addressing these issues, the research aims to provide actionable insights that can guide sustainable practices in the agri-food sector and inform international climate policies.

III. Research Problem

While the practical challenges posed by CO2 emissions and their impact on global temperatures are evident, a gap exists in understanding the intricate relationships within the data. The agri-food sector's contribution to CO2 emissions and the subsequent temperature changes is a complex interplay of various factors (LavagnedOrtigue, n.d.). These factors range from agricultural practices and industrial activities to consumption patterns across different countries.

The primary research problem is to study the following:

- The exact relationship between CO2 emissions from the agri-food sector and the observed changes in global temperatures.
- The varied contributions of different countries to this phenomenon. Given the diverse agricultural and industrial practices across nations, there is a need to quantify and compare the influence of each country based on their aggregated data on emissions and temperature change.

Furthermore, while the dataset provides a wealth of information, extracting meaningful insights requires applying appropriate data mining techniques. The choice of tools and methodologies and their effectiveness in addressing the research problem is another area of exploration.

In essence, this research seeks to transform raw data into actionable knowledge, providing a clearer understanding of the agri-food sector's role in global temperature changes and the varying influences of different countries.

IV. Research Objectives

This research aims to delve into the intricate relationships between CO2 emissions, especially from the agri-food sector, and global temperature changes. To achieve a comprehensive understanding, the research is guided by the following specific objectives:

- Correlation Analysis: To examine the correlation between carbon dioxide (CO2) emissions within the agri-food sector and the subsequent temperature rise. This objective seeks to understand the direct impact of emissions from agricultural and food-related activities on global temperatures.
- Country-specific Analysis: To analyse the influence of various countries based on aggregated data on emissions and temperature change. Recognising that different countries have diverse agricultural practices and industrial activities; this objective aims to quantify and compare the contributions of individual nations to global CO2 emissions and temperature variations.

By addressing these objectives, the research intends to provide a clear and detailed understanding of the role of the agri-food sector in global temperature changes and the varying influences of different countries.

V. Literature Review

The role of agriculture in climate change mitigation has been a subject of significant interest in recent literature.

A. Agriculture's Role in Climate Change Mitigation

Horowitz and Gottlieb (2010) emphasize that agriculture could be pivotal in U.S. efforts to address climate change. They suggest that farms and ranches can undertake activities that either reduce greenhouse gas (GHG) emissions or sequester greenhouse gases. Such activities may include conservation tillage, reducing nitrogen fertiliser usage, changing livestock and manure management practices, and planting trees or grass (Horowitz & Gottlieb, n.d.).

B. Policy Implications

The authors discuss the potential of the Federal Government to offer carbon offsets and incentive payments to encourage rural landowners to adopt climate-friendly activities (Horowitz & Gottlieb, n.d.). The extent of adoption would depend on costs, potential revenues, and other economic incentives created by climate policy.

C. *Carbon Offsets in a Cap-and-Trade System*

The literature explores the possibility of a nationwide cap-and-trade system on GHG emissions, where fossil fuel sources must have a permit for every ton of emissions (Horowitz & Gottlieb, n.d.). Under such a system, GHG permits would not be required for emissions from agricultural practices or land use changes. This system could encourage agricultural mitigation by making agricultural activities eligible for offset credits.

D. *Conservation Reserve Program (CRP)*

As highlighted by Horowitz and Gottlieb, the CRP is the most extensive environmental program on private lands in the U.S. The program has been instrumental in sequestering carbon, with estimates suggesting that CRP acres sequestered 48 million more metric tons of CO2 in 2008 than if the land had remained in previous uses (Horowitz & Gottlieb, n.d.).

E. *Potential Solutions*

The literature suggests that shifting to conservation tillage, nutrient management, and tree planting can contribute significantly to GHG reduction efforts (Horowitz & Gottlieb, n.d.). These practices not only help in reducing emissions but also have implications for farm profits.

VI. Research Methodology

The methodology for this research is based on an enhanced version of the CRISP-DM model, which offers a more granular approach to data mining projects (*IBM SPSS Modeler CRISP-DM Guide*, n.d.). This improved framework consists of eight distinct phases, each tailored to this study's specific objectives and context.

A. *Business Understanding*

The primary focus was to grasp the intricate relationship between CO2 emissions from the agri-food sector and global temperature changes. This chapter provides a foundational understanding of the significance and implications of this relationship.

B. *Data Understanding*

The dataset was initially explored to discern its structure, quality, and potential intricacies. This phase ensured familiarity with the data's attributes, potential outliers, and missing values.

C. *Data Preparation*

Preliminary steps, such as data cleaning and handling missing values, were executed to enhance the dataset's quality and consistency.

D. *Data Transformation*

The data was further processed to create derived attributes, normalize scales, and ensure it was in the optimal format for subsequent analysis.

E. *Data Mining Methods Selection*

Appropriate data mining techniques were identified based on the research objectives and the nature of the data. This phase involved choosing regression and clustering data mining methods.

F. *Data Mining Algorithms Selection*

Once the methods were chosen, specific algorithms suitable for the dataset and objectives were selected. K-means clustering, and agglomerative clustering are considered for the clustering. Linear regression, and random forest regression are considered for the regression.

G. *Data Mining*

The selected algorithms were applied to the dataset. This phase involved training models, testing them, and refining them based on performance.

H. *Interpretation*

The results obtained from the data mining phase were analysed and interpreted in the context of the research objectives. This phase ensured the findings were statistically significant, meaningful, and actionable regarding CO2 emissions and global temperature changes.

By adopting this enhanced CRISP-DM framework, the research ensures a thorough, structured, and rigorous approach to understanding the dynamics between CO2 emissions from the agri-food sector and global temperature variations.

VII. Processes Design

The design process for this research is characterized by a unique iterative approach, where the dataset undergoes the complete CRISP-DM process three times, each time utilising a different tool: SPSS, Python, and PySpark. This iterative methodology ensures a comprehensive exploration of the data, leveraging the strengths of each tool.

A. *Iteration with SPSS*

1) *Business Understanding*

Defined the research objectives and identified the key questions to be addressed.

2) *Data Understanding*

Imported the dataset into SPSS and conducted preliminary statistical analyses to understand its properties.

3) Data Preparation

Used SPSS's data cleaning functionalities to handle missing values, outliers, and duplicates.

4) Data Transformation

Applied transformations, such as normalization and encoding, to prepare the data for mining.

5) Data Mining Methods and Algorithms Selection

Appropriate data mining methods were chosen based on the research objectives, and specific SPSS-supported algorithms were selected.

6) Model Training and Testing

Models were trained and tested within the SPSS environment, evaluating their performance against predefined criteria.

7) Interpretation

Results were interpreted using SPSS's statistical and visualisation tools, translating the findings into preliminary insights.

B. Iteration with Python

1) Business Understanding

Revisited the research objectives and refined the key questions based on insights from the SPSS iteration.

2) Data Understaning:

Using Python's Pandas library, further explored the dataset to identify patterns and anomalies.

3) Data Preparation

Leveraged Python libraries, such as Pandas and NumPy, for data cleaning and preprocessing.

4) Data Transformation

Applied advanced transformations using Python to enhance the dataset's analytical value.

5) Data Mining Methods and Algorithms Selection:

Utilised Python's scikit-learn library to select and implement suitable data mining algorithms.

6) Model Training and Testing

Models were trained and tested using Python, with performance metrics evaluated to ensure robustness.

7) Interpretation

Leveraged Python's visualisation libraries, like Matplotlib and Seaborn, to interpret and visualise the results.

C. Iteration with PySpark

1) Business Understanding

With insights from previous iterations, the research objectives and questions were refined for a more focused exploration.

2) Data Understanding

Used PySpark to handle large-scale data operations and better understand the dataset's structure.

3) Data Preparation

Leveraged PySpark's data processing capabilities for cleaning and preprocessing.

4) Data Transformation

Applied large-scale data transformations using PySpark to optimize the dataset for mining.

5) Data Mining Methods and Algorithms Selection

Utilised PySpark's MLlib to select and implement data mining algorithms suitable for big data.

6) Model Training and Testing

Trained and tested models within the PySpark environment, ensuring scalability and efficiency.

7) Interpretation

Interpreted the results within the PySpark framework, generating insights that build upon the findings from the SPSS and Python iterations.

By iterating over the dataset three times and implementing the entire CRISP-DM process with a different tool, this research ensures a multi-faceted, in-depth exploration of the data, maximizing the potential for comprehensive insights.

VIII. Implementation Description

A. Python Iteration

1) Business Understanding

The objectives of the business were identified, focusing on the agrifood systems and their impact on greenhouse gas emissions.

A comprehensive project plan was developed, outlining the steps to achieve the data mining objectives.

2) Data Understanding

Initial data was collected and described, focusing on attributes relevant to agrifood systems.

Data was explored using visualisations and statistical methods, leveraging libraries like Matplotlib and Pandas.

Data quality was verified, ensuring its reliability for further analysis.

3) Data Preparation

Relevant data attributes were selected for the analysis.

The Multiple Imputation by Chained Equations (MICE) method addressed missing values (*Sklearn.Feature_selection.Mutual_info_regression*, n.d.).

Outliers were processed to ensure data consistency.

The RobustScaler from the sklearn.preprocessing (*Sklearn.Preprocessing.RobustScaler*, n.d.) module was used for data projection, ensuring resilience to outliers.

4) Data Mining Methods and Algorithms Selection

a) Regression

- Linear Regression: Relationships between variables were modeled using sklearn.linear_model.LinearRegression (*Sklearn.Linear_model.LinearRegression*, n.d.).
- Random Forest Regression: Complex relationships in the data were captured using sklearn.ensemble.RandomForestRegressor (*Sklearn.Ensemble.RandomForestRegressor*, n.d.).

b) Clustering

- K-Means Clustering: Data was segmented using sklearn.cluster.KMeans (*Sklearn.Cluster.KMeans*, n.d.).
- Agglomerative Clustering: A hierarchical approach was employed using sklearn.cluster.AgglomerativeClustering (*Sklearn.Cluster.AgglomerativeClustering*, n.d.).

5) Model Training and Testing

Using the train_test_split function from sklearn.model_selection, the dataset was divided into training and testing subsets.

Models were trained on the training subset and evaluated on the testing subset to assess their performance.

Metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were used to quantify the models' accuracy.

6) Interpretation

Results were visualised using Matplotlib and Seaborn to represent the findings.

Insights derived from the models were discussed regarding the agrifood system's impact on greenhouse gas emissions.

B. SPSS Iteration:

1) Business Understanding

Similar to the Python iteration, the business objectives and project plan were outlined, considering the statistical strengths of SPSS.

2) Data Understanding and Preparation

Data exploration and preprocessing were conducted using SPSS's graphical tools and built-in statistical functions.

3) Data Mining Methods and Algorithms Selection

a) Regression

- Linear Regression: Modeled relationships between variables.
- Random Forest Regression: Captured complex relationships, albeit with limited options compared to Python.

b) Clustering

- K-Means Clustering: Segmented data.
- TwoStep Clustering: A unique SPSS method combining hierarchical and density-based clustering.

4) Model Training and Testing

The dataset was partitioned into training and testing sets using SPSS's built-in functionalities.

Models were evaluated using SPSS's statistical tools, focusing on metrics relevant to regression and clustering.

5) Interpretation

SPSS's graphical tools were used to visualise the results.

Findings were contextualised, emphasizing their relevance to the research objectives.

C. PySpark Iteration

1) Business Understanding

The business objectives and project plan were outlined, emphasizing the scalability and distributed processing capabilities of PySpark.

2) Data Understanding and Preparation

Large-scale data operations were optimized using PySpark's distributed computing capabilities, leveraging the pyspark.sql and pyspark.ml modules .

3) Data Mining Methods and Algorithms Selection

a) Regression

- Random Forest Regression: Captured intricate relationships in big data using pyspark.ml.regression.RandomForestRegressor .

b) Clustering

- K-Means Clustering: Used for clustering large datasets with pyspark.ml.clustering.KMeans.

c) Model Training and Testing

The dataset was split into training and testing subsets using PySpark's data-handling capabilities.

Models were trained and evaluated, leveraging the distributed processing power of PySpark.

4) Interpretation:

Results were interpreted within the PySpark framework, generating insights that build upon the findings from the Python and SPSS iterations.

Visualisations, where applicable, were generated using PySpark's visualisation tools.

IX. Patterns and Results

A. *Correlation between CO2 Emissions and Temperature Rise*

The linear regression model demonstrated that certain variables within the agri-food system, such as food production or transportation methods, have a linear relationship with greenhouse gas emissions. As these factors change, there is a predictable change in emissions.

Understanding these linear relationships is pivotal for stakeholders in the agri-food system. If transportation becomes a significant factor, transitioning to eco-friendly methods becomes a viable solution. This insight provides a data-driven foundation for stakeholders to prioritize emission reduction strategies, directly addressing the practical problem of the agri-food system's environmental impact.

B. *Influence of Various Countries on Emissions and Temperature*

The random forest regression model underscored food retail as the most critical agri-food factor influencing the average temperature rise. This suggests that practices within the food retail sector significantly impact the environment.

This finding emphasizes the importance of sustainable practices within the food retail industry. By focusing on this sector with sustainable initiatives, there is potential for a substantial reduction in greenhouse gas emissions, aligning with the research objective of understanding the agri-food system's environmental impact.

C. *Countries with Highest Emissions and Temperature Increase in 2020*

The K-Means clustering model segmented the data into distinct clusters, revealing inherent groupings within the agri-food system. These clusters represent different farming practices, types of crops, or regions with similar environmental impacts.

These clusters offer valuable insights for policymakers and industry leaders. By pinpointing high-impact clusters, targeted interventions can be designed, promoting sustainable practices within specific segments of the agri-food system, and addressing the practical problem of mitigating greenhouse gas emissions.

X. Proposed Actions

A. *Promotion of Sustainable Practices in the Agri-food System*

1) Action

Encourage stakeholders within the agri-food system, especially within the food retail sector, to adopt sustainable practices. This can be achieved through awareness campaigns, training programs, and incentives.

2) Rationale

The random forest regression model highlighted the significant impact of the food retail sector on the environment. By promoting sustainable practices in this sector, we can directly address the practical problem of the agri-food system's environmental impact.

B. *Adoption of Eco-friendly Transportation Methods*

1) Action

Advocate for transitioning to eco-friendly transportation methods within the agri-food system. This can include electric vehicles, biofuel-powered vehicles, and efficient transportation routes.

2) Rationale

The linear regression model identified transportation as a potentially significant factor affecting greenhouse gas emissions. Addressing this can help in reducing the environmental footprint of the agri-food system.

C. *Targeted Interventions Based on Data Clustering:*

1) Action

Design and implement targeted interventions based on the clusters identified by the K-Means and Agglomerative clustering models. This can involve promoting specific farming practices or introducing sustainable initiatives in identified high-impact regions.

2) Rationale

The clustering models revealed inherent groupings within the agri-food system. Focusing on these specific clusters can create tailored solutions that address the practical problem more effectively.

D. *Incentivize Carbon Sequestration in Agriculture:*

1) Action

Offer carbon offsets and incentive payments to encourage rural landowners to pursue climate-friendly activities, such as conservation tillage or planting trees. This can be part of a broader effort to combat climate change.

2) Rationale

The literature suggests that agriculture could play a prominent role in U.S. efforts to address climate change if farms and ranches undertake activities that reduce greenhouse gas emissions or sequester carbon (Horowitz & Gottlieb, n.d.).

E. Engage in Continuous Data Analysis and Feedback:

1) Action

Regularly analyse the data from the agri-food system using the tools and methodologies identified in this research. Use the insights gained to refine and adjust the proposed actions continuously.

2) Rationale

Continuous analysis ensures that the actions remain relevant and effective in addressing the practical problems and meeting the research objectives.

XI. Reference

Horowitz, J., & Gottlieb, J. (n.d.). *The Role of Agriculture in Reducing Greenhouse Gas Emissions*.

IBM SPSS Modeler CRISP-DM Guide. (n.d.).

LavagnedOrtigue, O. (ESS). (n.d.). *Greenhouse gas emissions from agrifood systems*.

Sklearn.cluster.AgglomerativeClustering. (n.d.). Scikit-Learn. Retrieved 21 September 2023, from https://scikit-learn/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

Sklearn.cluster.KMeans. (n.d.). Scikit-Learn. Retrieved 21 September 2023, from https://scikit-learn/stable/modules/generated/sklearn.cluster.KMeans.html

Sklearn.ensemble.RandomForestRegressor. (n.d.). Scikit-Learn. Retrieved 21 September 2023, from https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

Sklearn.feature_selection.mutual_info_regression. (n.d.). Scikit-Learn. Retrieved 21 September 2023, from https://scikit-learn/stable/modules/generated/sklearn.feature_selection.mutual_info_regression.html

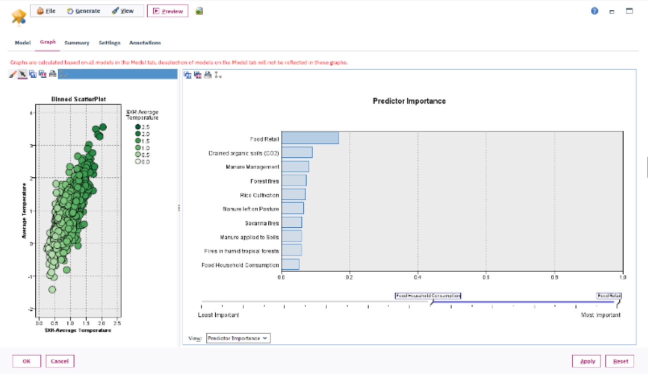
Sklearn.linear_model.LinearRegression. (n.d.). Scikit-Learn. Retrieved 21 September 2023, from https://scikit-learn/stable/modules/generated/sklearn.linear_model.LinearRegression.html

Sklearn.preprocessing.RobustScaler. (n.d.). Scikit-Learn. Retrieved 21 September 2023, from https://scikit-learn/stable/modules/generated/sklearn.preprocessing.RobustScaler.html

XII. Appendix

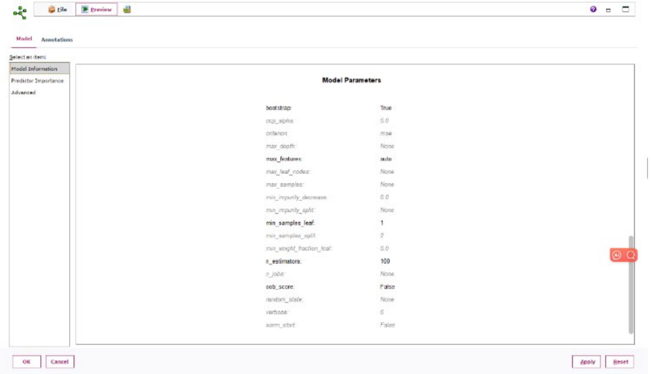
A. IBM Software Analytics Solution Visualisation

Figure 1. Predictor Importance based on all auto numeric regression models



Graphs are calculated based on all models in the Model tab, deselection of models on the Model tab will not be reflected in these graphs.

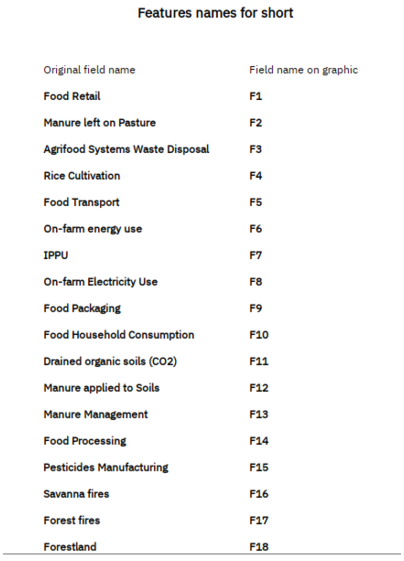
Figure 2. Model parameters of random forest



Model Parameters

bootstrap: True
ccp_alpha: 0.0
criterion: mse
max_depth: None
max_features: auto
max_leaf_nodes: None
max_samples: None
min_impurity_decrease: 0.0
min_impurity_split: None
min_samples_leaf: 1
min_samples_split: 2
min_weight_fraction_leaf: 0.0
n_estimators: 100
n_jobs: None
oob_score: False
random_state: None
verbose: 0
warm_start: False

Figure 3. Predictor Importance based on random forest model



Features names for short

Original field name	Field name on graphic
Food Retail	F1
Manure left on Pasture	F2
Agrifood Systems Waste Disposal	F3
Rice Cultivation	F4
Food Transport	F5
On-farm energy use	F6
IPPU	F7
On-farm Electricity Use	F8
Food Packaging	F9
Food Household Consumption	F10
Drained organic soils (CO2)	F11
Manure applied to Soils	F12
Manure Management	F13
Food Processing	F14
Pesticides Manufacturing	F15
Savanna fires	F16
Forest fires	F17
Forestland	F18

Figure 4. Predictor Importance based on linear regression model

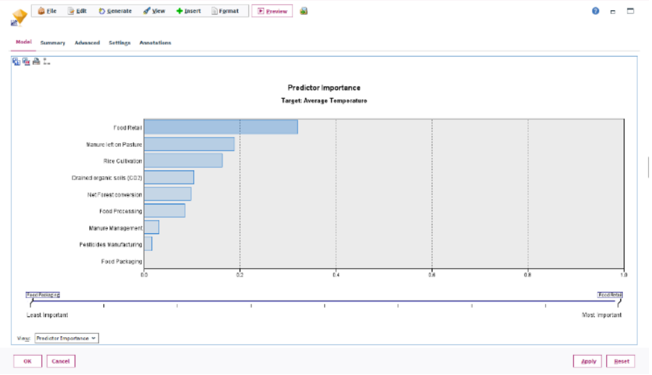
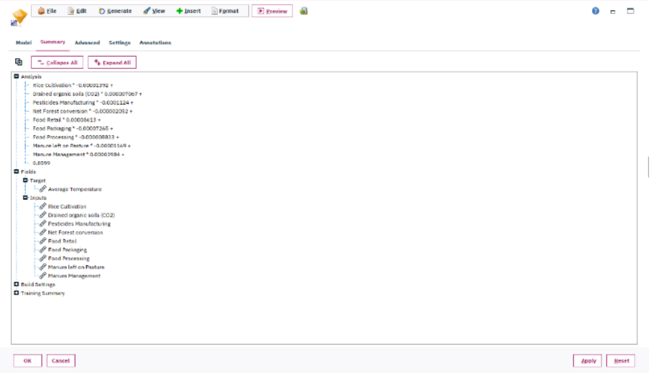
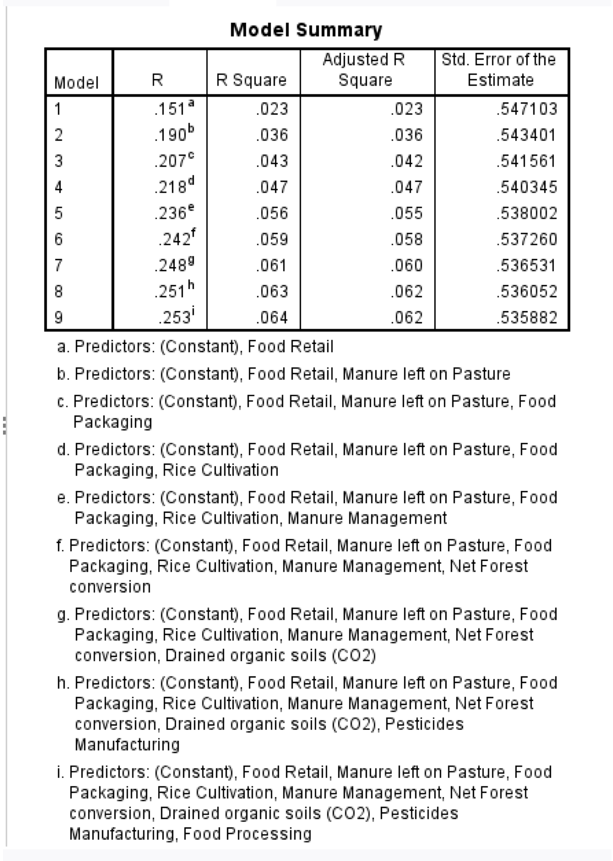


Figure 5. Linear relationship analysis



```
Analysis
    Rice Cultivation * -0.00001392 +
    Drained organic soils (CO2) * 0.000007067 +
    Pesticides Manufacturing * -0.0001124 +
    Net Forest conversion * -0.000002052 +
    Food Retail * 0.00008613 +
    Food Packaging * -0.00007265 +
    Food Processing * -0.000008833 +
    Manure left on Pasture * -0.00001169 +
    Manure Management * 0.00003984 +
    0.8599
Fields
    Target
        Average Temperature
    Inputs
        Rice Cultivation
        Drained organic soils (CO2)
        Pesticides Manufacturing
        Net Forest conversion
        Food Retail
        Food Packaging
        Food Processing
        Manure left on Pasture
        Manure Management
Build Settings
Training Summary
```

Figure 6. Linear regression model summary



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.151[a]	.023	.023	.547103
2	.190[b]	.036	.036	.543401
3	.207[c]	.043	.042	.541561
4	.218[d]	.047	.047	.540345
5	.236[e]	.056	.055	.538002
6	.242[f]	.059	.058	.537260
7	.248[g]	.061	.060	.536531
8	.251[h]	.063	.062	.536052
9	.253[i]	.064	.062	.535882

a. Predictors: (Constant), Food Retail

b. Predictors: (Constant), Food Retail, Manure left on Pasture

c. Predictors: (Constant), Food Retail, Manure left on Pasture, Food Packaging

d. Predictors: (Constant), Food Retail, Manure left on Pasture, Food Packaging, Rice Cultivation

e. Predictors: (Constant), Food Retail, Manure left on Pasture, Food Packaging, Rice Cultivation, Manure Management

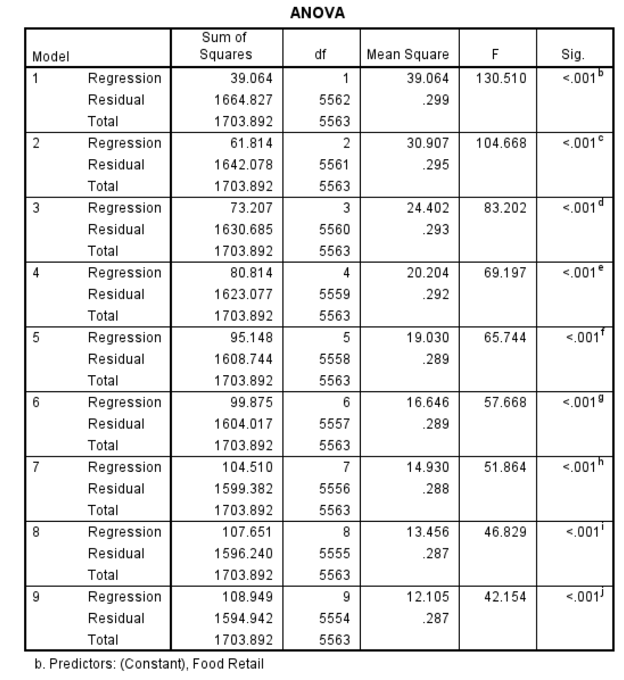
f. Predictors: (Constant), Food Retail, Manure left on Pasture, Food Packaging, Rice Cultivation, Manure Management, Net Forest conversion

g. Predictors: (Constant), Food Retail, Manure left on Pasture, Food Packaging, Rice Cultivation, Manure Management, Net Forest conversion, Drained organic soils (CO2)

h. Predictors: (Constant), Food Retail, Manure left on Pasture, Food Packaging, Rice Cultivation, Manure Management, Net Forest conversion, Drained organic soils (CO2), Pesticides Manufacturing

i. Predictors: (Constant), Food Retail, Manure left on Pasture, Food Packaging, Rice Cultivation, Manure Management, Net Forest conversion, Drained organic soils (CO2), Pesticides Manufacturing, Food Processing

Figure 7. Linear regression model ANOVA

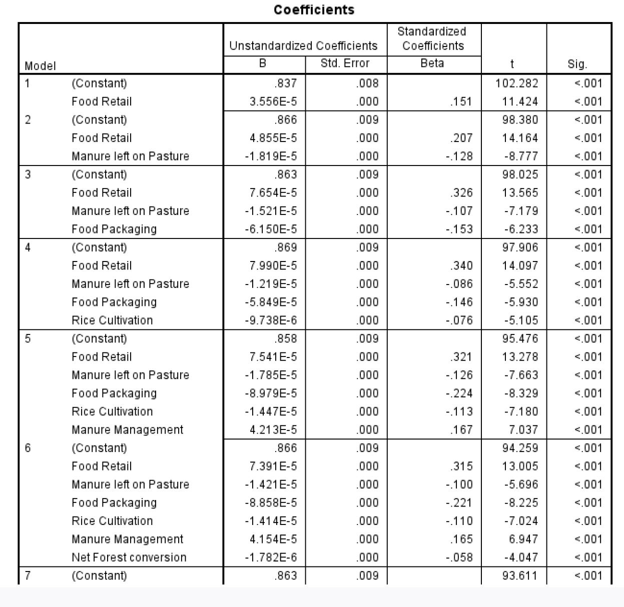


ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	39.064	1	39.064	130.510	<.001[b]
	Residual	1664.827	5562	.299		
	Total	1703.892	5563			
2	Regression	61.814	2	30.907	104.668	<.001[c]
	Residual	1642.078	5561	.295		
	Total	1703.892	5563			
3	Regression	73.207	3	24.402	83.202	<.001[d]
	Residual	1630.685	5560	.293		
	Total	1703.892	5563			
4	Regression	80.814	4	20.204	69.197	<.001[e]
	Residual	1623.077	5559	.292		
	Total	1703.892	5563			
5	Regression	95.148	5	19.030	65.744	<.001[f]
	Residual	1608.744	5558	.289		
	Total	1703.892	5563			
6	Regression	99.875	6	16.646	57.668	<.001[g]
	Residual	1604.017	5557	.289		
	Total	1703.892	5563			
7	Regression	104.510	7	14.930	51.864	<.001[h]
	Residual	1599.382	5556	.288		
	Total	1703.892	5563			
8	Regression	107.651	8	13.456	46.829	<.001[i]
	Residual	1596.240	5555	.287		
	Total	1703.892	5563			
9	Regression	108.949	9	12.105	42.154	<.001[j]
	Residual	1594.942	5554	.287		
	Total	1703.892	5563			

b. Predictors: (Constant), Food Retail

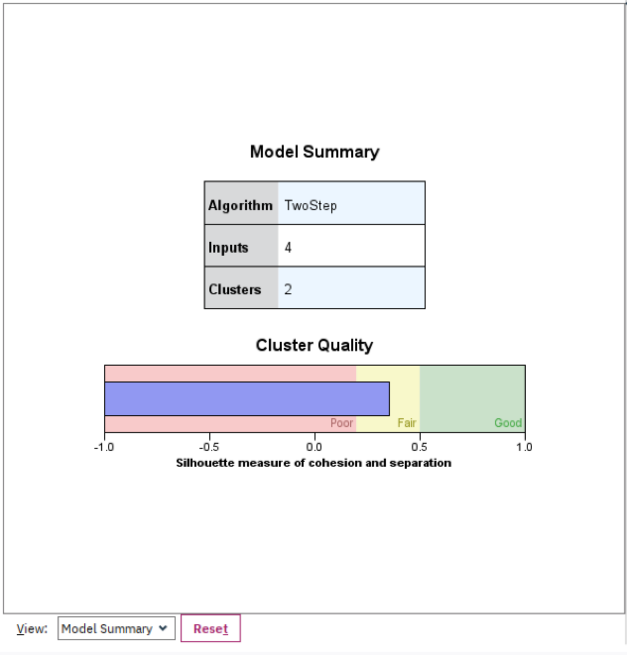
Figure 8. Linear regression model coefficients



Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.837	.008		102.282	<.001
	Food Retail	3.556E-5	.000	.151	11.424	<.001
2	(Constant)	.866	.009		98.380	<.001
	Food Retail	4.855E-5	.000	.207	14.164	<.001
	Manure left on Pasture	-1.819E-5	.000	-.128	-8.777	<.001
3	(Constant)	.863	.009		98.025	<.001
	Food Retail	7.654E-5	.000	.326	13.565	<.001
	Manure left on Pasture	-1.521E-5	.000	-.107	-7.179	<.001
	Food Packaging	-6.150E-5	.000	-.153	-6.233	<.001
4	(Constant)	.869	.009		97.906	<.001
	Food Retail	7.990E-5	.000	.340	14.097	<.001
	Manure left on Pasture	-1.219E-5	.000	-.086	-5.552	<.001
	Food Packaging	-5.849E-5	.000	-.146	-5.930	<.001
	Rice Cultivation	-9.738E-6	.000	-.076	-5.105	<.001
5	(Constant)	.858	.009		95.476	<.001
	Food Retail	7.541E-5	.000	.321	13.278	<.001
	Manure left on Pasture	-1.785E-5	.000	-.126	-7.663	<.001
	Food Packaging	-8.979E-5	.000	-.224	-8.329	<.001
	Rice Cultivation	-1.447E-5	.000	-.113	-7.180	<.001
	Manure Management	4.213E-5	.000	.167	7.037	<.001
6	(Constant)	.866	.009		94.259	<.001
	Food Retail	7.391E-5	.000	.315	13.005	<.001
	Manure left on Pasture	-1.421E-5	.000	-.100	-5.696	<.001
	Food Packaging	-8.858E-5	.000	-.221	-8.225	<.001
	Rice Cultivation	-1.414E-5	.000	-.110	-7.024	<.001
	Manure Management	4.154E-5	.000	.165	6.947	<.001
	Net Forest conversion	-1.782E-6	.000	-.058	-4.047	<.001
7	(Constant)	.863	.009		93.611	<.001

Figure 9. Clustering model summary

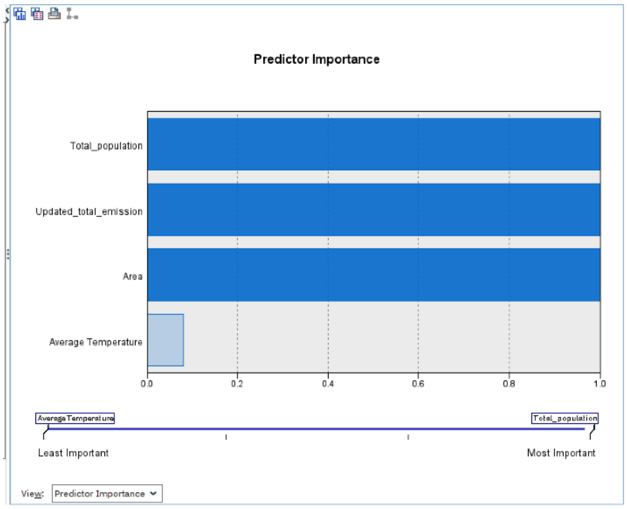


Model Summary

Algorithm	TwoStep
Inputs	4
Clusters	2

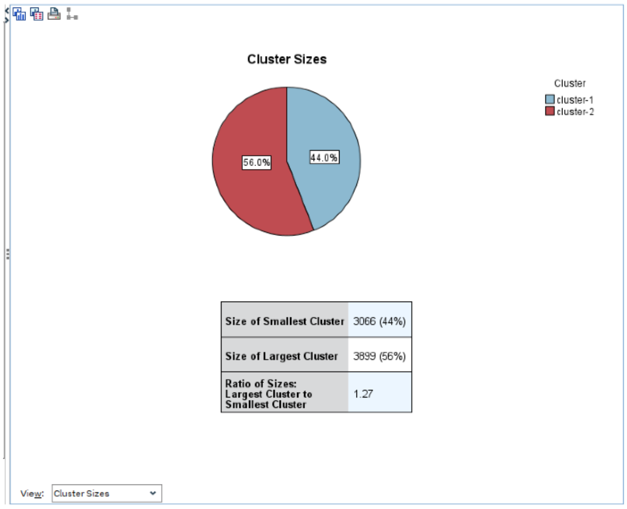
View: Model Summary

Figure 10. Clustering model predictor importance



View: Predictor Importance

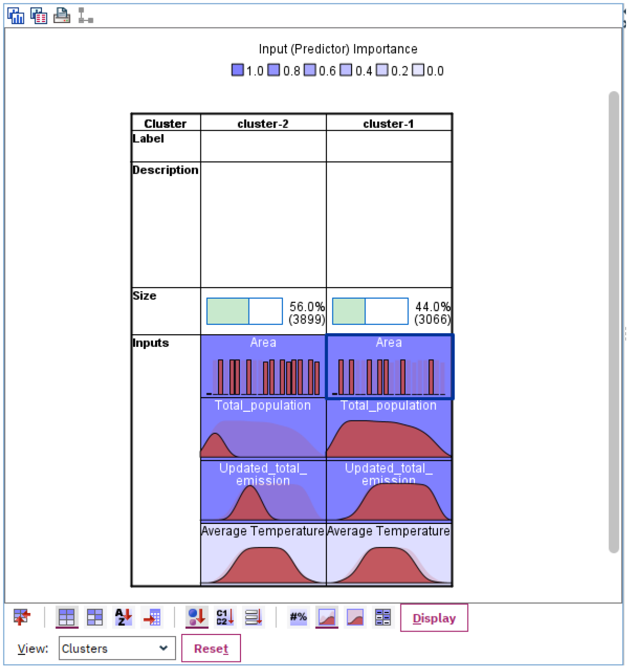
Figure 11. Clustering sizes



Size of Smallest Cluster	3066 (44%)
Size of Largest Cluster	3899 (56%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	1.27

View: Cluster Sizes

Figure 12. Cluster 1 VS Cluster 2



Cluster	cluster-2	cluster-1
Label		
Description		
Size	56.0% (3899)	44.0% (3066)
Inputs	Area	Area
	Total_population	Total_population
	Updated_total_emission	Updated_total_emission
	Average Temperature	Average Temperature

View: Clusters

Figure 13. Total emissions distribution of two clusters

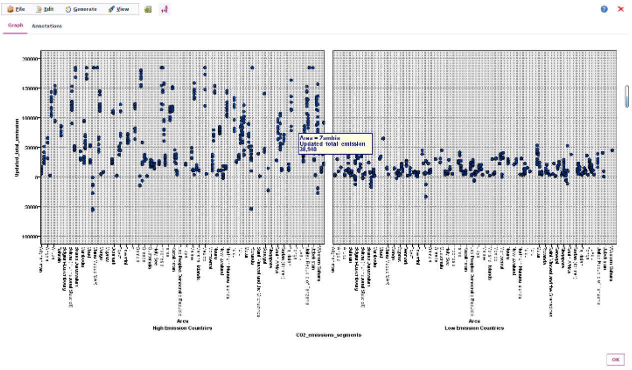


Figure 14. Total population distribution of two clusters

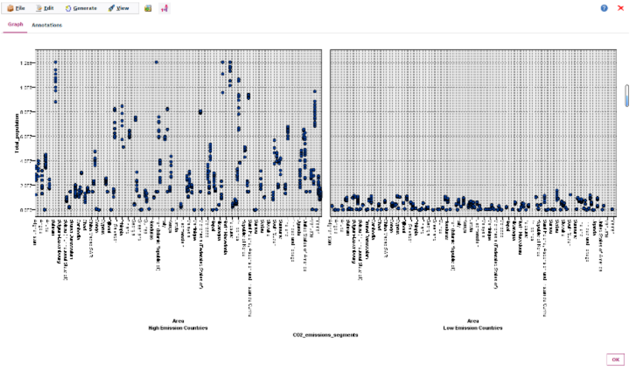


Figure 15. Average temperature rise distribution of two clusters

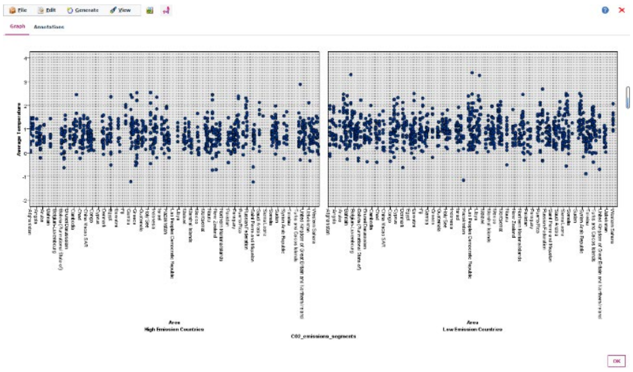


Figure 16. Emissions per capita distribution of two clusters

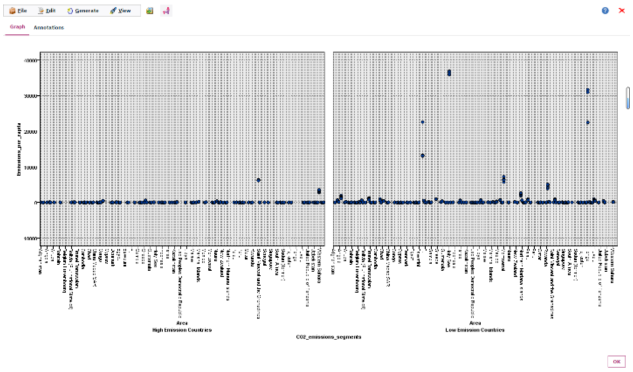
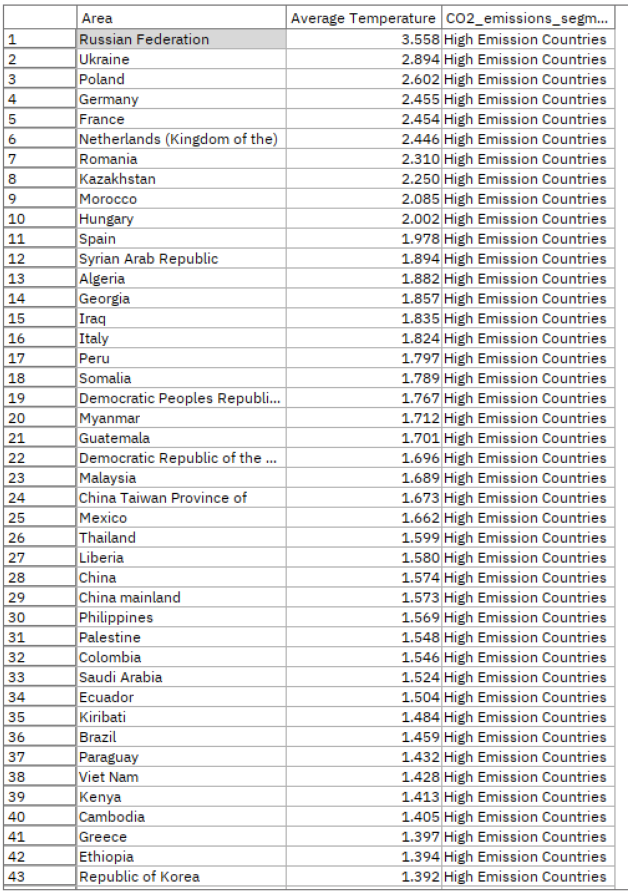


Figure 17. Country with the highest average temperature increase by year



	Area	Average Temperature	CO2_emissions_segm...
1	Russian Federation	3.558	High Emission Countries
2	Ukraine	2.894	High Emission Countries
3	Poland	2.602	High Emission Countries
4	Germany	2.455	High Emission Countries
5	France	2.454	High Emission Countries
6	Netherlands (Kingdom of the)	2.446	High Emission Countries
7	Romania	2.310	High Emission Countries
8	Kazakhstan	2.250	High Emission Countries
9	Morocco	2.085	High Emission Countries
10	Hungary	2.002	High Emission Countries
11	Spain	1.978	High Emission Countries
12	Syrian Arab Republic	1.894	High Emission Countries
13	Algeria	1.882	High Emission Countries
14	Georgia	1.857	High Emission Countries
15	Iraq	1.835	High Emission Countries
16	Italy	1.824	High Emission Countries
17	Peru	1.797	High Emission Countries
18	Somalia	1.789	High Emission Countries
19	Democratic Peoples Republi...	1.767	High Emission Countries
20	Myanmar	1.712	High Emission Countries
21	Guatemala	1.701	High Emission Countries
22	Democratic Republic of the ...	1.696	High Emission Countries
23	Malaysia	1.689	High Emission Countries
24	China Taiwan Province of	1.673	High Emission Countries
25	Mexico	1.662	High Emission Countries
26	Thailand	1.599	High Emission Countries
27	Liberia	1.580	High Emission Countries
28	China	1.574	High Emission Countries
29	China mainland	1.573	High Emission Countries
30	Philippines	1.569	High Emission Countries
31	Palestine	1.548	High Emission Countries
32	Colombia	1.546	High Emission Countries
33	Saudi Arabia	1.524	High Emission Countries
34	Ecuador	1.504	High Emission Countries
35	Kiribati	1.484	High Emission Countries
36	Brazil	1.459	High Emission Countries
37	Paraguay	1.432	High Emission Countries
38	Viet Nam	1.428	High Emission Countries
39	Kenya	1.413	High Emission Countries
40	Cambodia	1.405	High Emission Countries
41	Greece	1.397	High Emission Countries
42	Ethiopia	1.394	High Emission Countries
43	Republic of Korea	1.392	High Emission Countries

B. Open Source Analytics Solution Visualisation

Figure 18. Correlative importance of different features

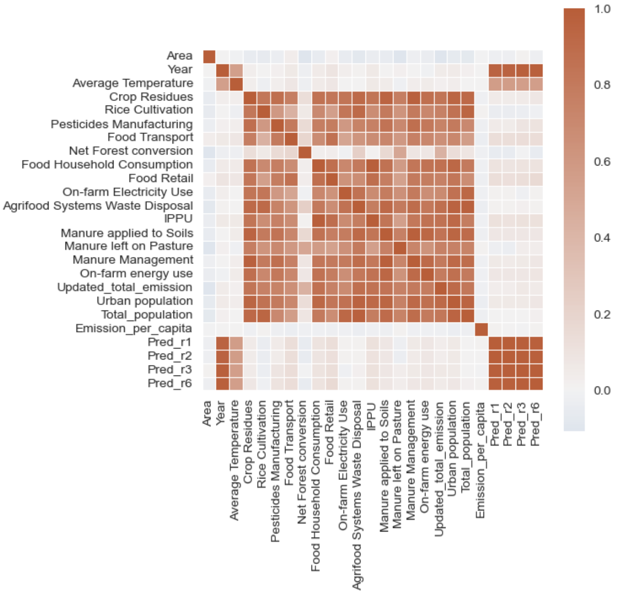


Figure 19. Predictor importance based on random forest regression model

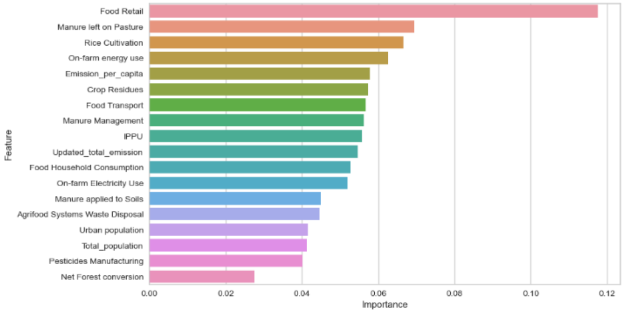


Figure 20. Accuracy of different linear regression model

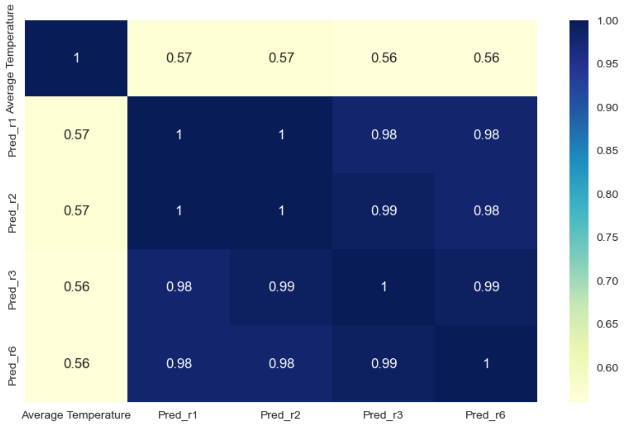


Figure 21. K-means clustering results

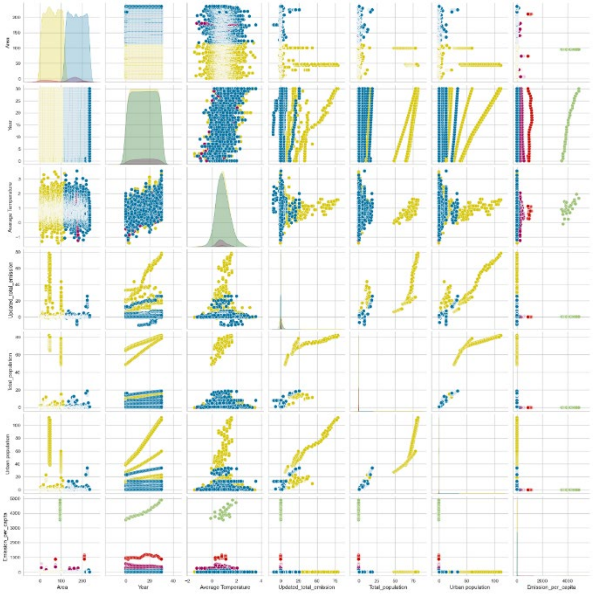


Figure 22. Agglomerative clustering results

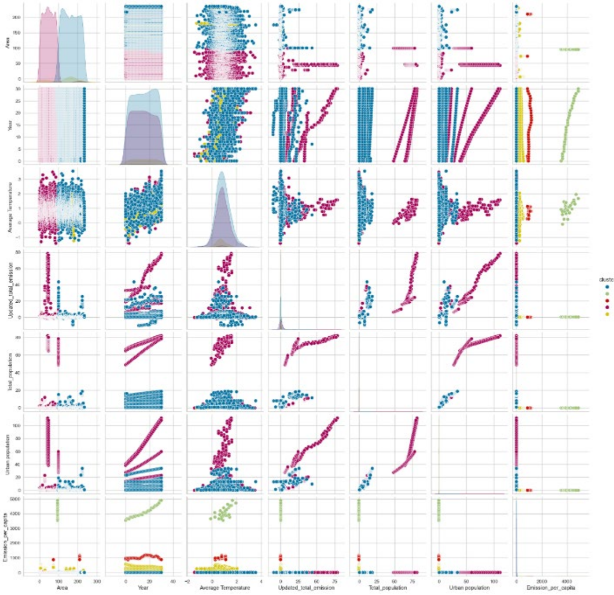


Figure 23. Average temperature change by top 30 countries in 2020

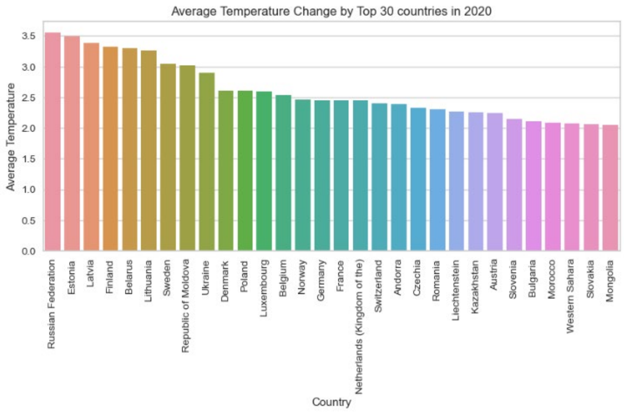


Figure 24. Total agri-food CO2 emissions by top 30 countries in 2020

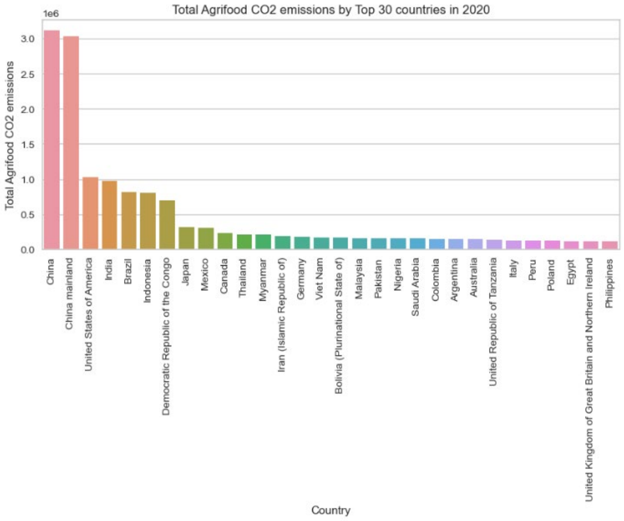
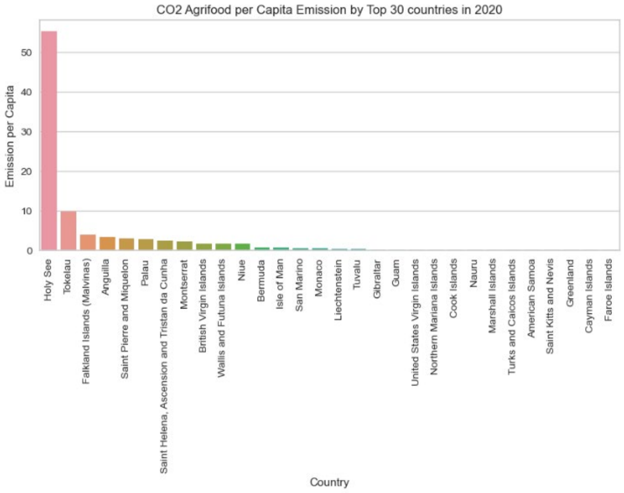


Figure 25. CO2 agri-food per capita emissions by top 30 countries in 2020



C. Big Data Analytics Solution Visualisation

Figure 26. Feature importance based on random forest regression model

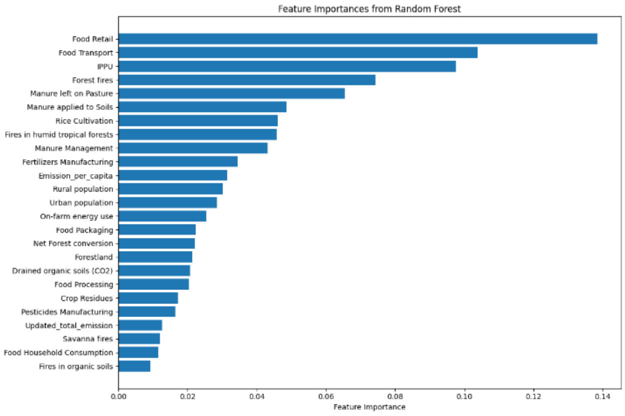


Figure 27. K-Means clustering result – 1

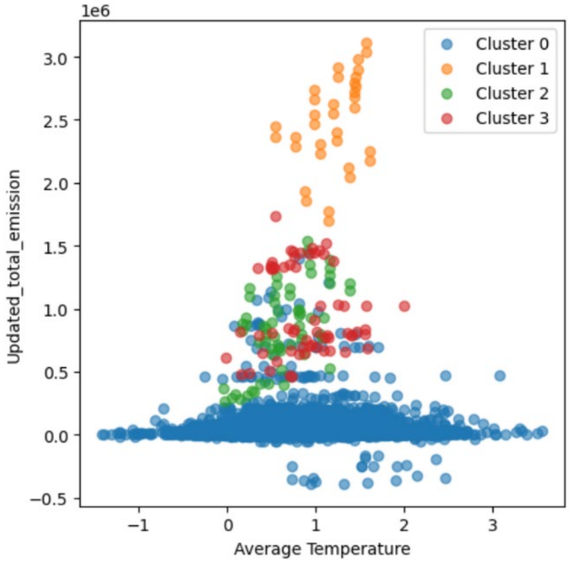


Figure 28. K-Means clustering result – 2

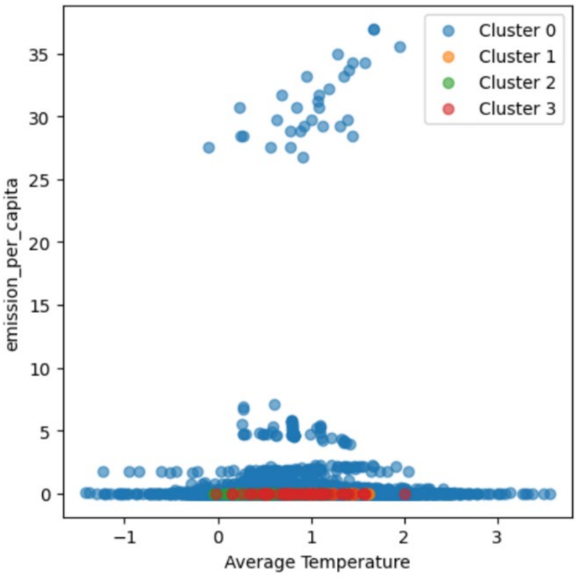


Figure 29. K-Means clustering result – 3

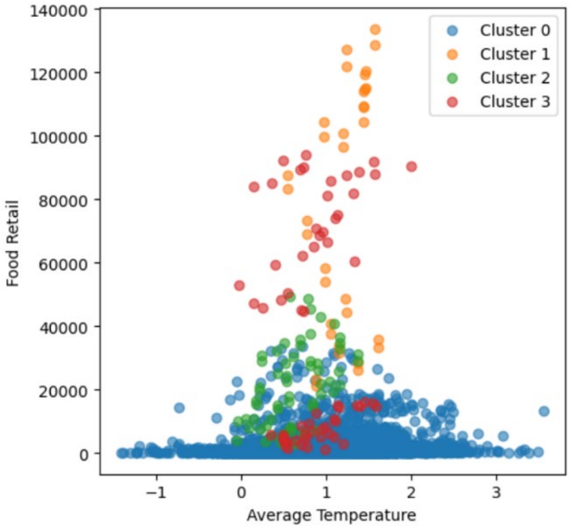


Figure 30. Average temperature change by top 30 countries in 2020

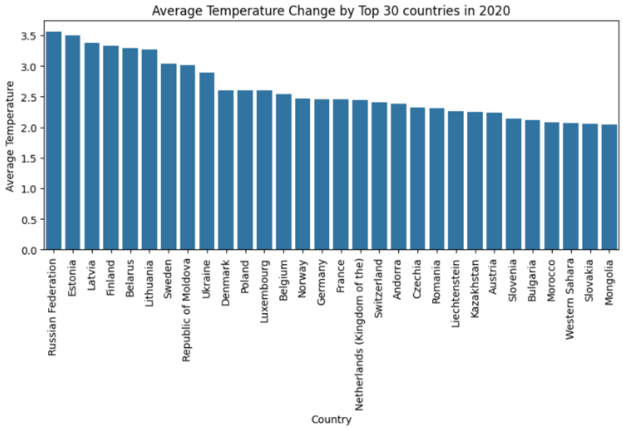


Figure 31. Total agri-food CO2 emissions by top 30 countries in 2020

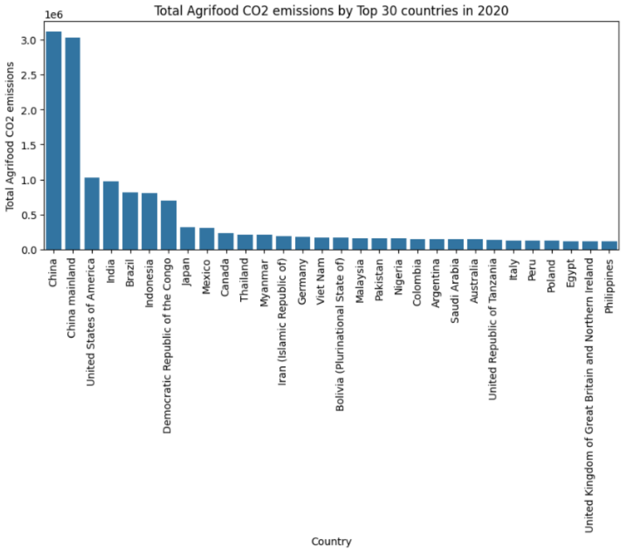
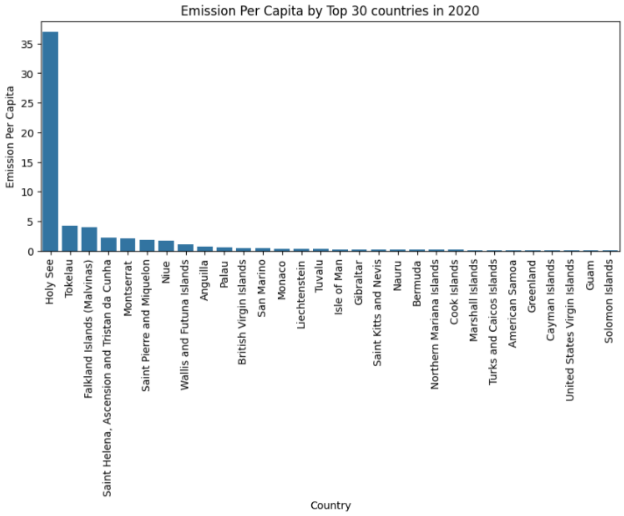


Figure 32. CO2 agri-food per capita emission by top 30 countries in 2020


Emission Per Capita by Top 30 countries in 2020
Emission Per Capita
Holy See
Tokelau
Falkland Islands (Malvinas)
Saint Helena, Ascension and Tristan da Cunha
Montserrat
Saint Pierre and Miquelon
Niue
Wallis and Futuna Islands
Anguilla
Palau
British Virgin Islands
San Marino
Monaco
Liechtenstein
Tuvalu
Isle of Man
Gibraltar
Saint Kitts and Nevis
Nauru
Bermuda
Northern Mariana Islands
Cook Islands
Marshall Islands
Turks and Caicos Islands
American Samoa
Greenland
Cayman Islands
United States Virgin Islands
Guam
Solomon Islands
Country