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

Iteration 1 - ISAS
IBM Software Analytics Solution

Contents

1. Business understanding

1.1 Identify the objectives of the business

Agrifood systems encompass various stages of the agricultural value chain, including the production of both food and non-food agricultural products. These stages involve food storage, aggregation, post-harvest handling, transportation, processing, distribution, marketing, disposal, and consumption. Food systems within agrifood systems encompass a wide range of food products derived from various sources, including crop and livestock production, forestry, fisheries, aquaculture, and synthetic biology, with the primary purpose of being consumed by humans (‘Agrifood Systems’, 2023).

Agrifood system has three elements:

- Primary production, which encompasses both agricultural and non-agricultural food sources and non-food agricultural products that function as inputs for other industries.
- Food distribution, which connects production with consumption through supply chains and domestic transport networks. Food supply chains encompass a comprehensive range of participants and processes engaged in the post-harvest management, storage, consolidation, transportation, transformation, dissemination, and commercialization of food products.
- Household consumption, as a consequence of operational agrifood systems, which is susceptible to different levels of demand shocks, such as a decrease in income, contingent upon the prevalence of vulnerable segments within the population. As the proportion increases, safeguarding food security and nutrition from shocks becomes increasingly challenging.

Agrifood systems substantially impact anthropogenic greenhouse gas (GHG) emissions, accounting for approximately one-third of the overall emission (LavagnedOrtigue, n.d.). The emissions in question are derived from many sources, encompassing on-farm activities that pertain to the cultivation of crops and the rearing of livestock. Moreover, alterations in land use, such as deforestation and the drainage of peatlands to facilitate agricultural expansion, are significant contributors to greenhouse gas (GHG) emissions. In addition, emissions are also produced throughout the pre-and post-production phases, which include activities such as food manufacturing, retail operations, household consumption, and food disposal procedures (LavagnedOrtigue, n.d.).

This study is with the following objectives:

- Deeply understand the environmental impact, focusing on climate change and

global warming, from the agri-food industry.

- Provide evidence of policy setting to reduce the CO2 emissions from the agri-food sector.

1.2 Assess the situation

1.2.1 Resource inventory

The software, IBM SPSS Modeler, used for this project is from the https://www.ibm.com/academic/home . The datasets used for this project are from www.kaggle.com/datasets. All references are from the websites: www.nzagrc.org.nz, www.fao.org, www.iaea.org, and www.beehive.govt.nz.

1.2.2 Requirements, assumptions, and constraints

Agricultural departments or organisations responsible for policymaking may benefit from establishing a dedicated data science team to undertake data mining and analysis tasks. Alternatively, they could consider engaging the services of a data mining consulting company to provide the necessary technical expertise.

From a database security standpoint, when data mining tasks are outsourced to a consulting firm, it becomes necessary for the consulting company to gain access to the backend database system. Ensuring the database system's security is paramount for agricultural organisations.

Economic factors significantly influence the outcome of the data mining project. The consideration of consulting fees and the comparative costs of competing products may play a significant role in determining whether to establish an internal team or seek the services of a consulting firm. Budgetary limitations may influence the decision-making process.

Assumptions regarding the quality of data play a pivotal role. The availability, accuracy, and integration of emissions, temperature, and agricultural data influence the reliability of the analysis. The resolution of data gaps and inconsistencies is of utmost importance. A specific assumption is that all agri-food factors are independent of the average temperature rise for implementing a linear regression model.

Gaining insight into the perspective of the project sponsor or management team is crucial. Are they interested in a comprehensive understanding of the data mining model, or are they primarily focused on obtaining practical and implementable outcomes?

Adapting communication strategies to align with individuals' areas of expertise is crucial for facilitating optimal decision-making processes. Achieving a successful project is contingent upon the careful consideration and management of various factors, including the harmonisation

of economic constraints, the dependability of data, and the fulfilment of stakeholder expectations.

In data access, it is imperative to acquire passwords for essential data sources to facilitate uninterrupted analysis. It is imperative to adhere to data security protocols. In the context of legal limitations, it is imperative to ascertain data usage rights and adhere to regulatory frameworks to mitigate potential legal complications and safeguard against privacy breaches. Concerning financial limitations, it is imperative to develop a comprehensive project budget that encompasses all expenditures, such as consulting fees, tool expenses, and any unforeseen costs that may arise. By considering these factors, data access protection, adherence to legal requirements, and preservation of budgetary integrity are ensured, facilitating a seamless and compliant project implementation.

1.2.3 Risks and contingencies

Regarding risk management, exercising control over consulting fees within the project budget is imperative. In addition to this, it is crucial to consider the cost of time, as policy formulation is frequently intertwined with strategic planning and the annual report. Data risks, such as inadequate data quality or coverage, can compromise the accuracy of analysis. Implementing rigorous data validation and preparation protocols is imperative to address this concern effectively. The management of potential risks associated with the outcomes, such as the possibility of less influential preliminary findings, can be effectively addressed by implementing transparent communication strategies. Effectively managing stakeholder expectations can be achieved by contextually presenting findings and emphasising the potential for further insights as the analysis progresses. It is imperative to ensure meticulous and comprehensive scheduling of the project.

1.3 Determine data mining objectives

With the help of a particular data mining team or a consulting company, the business objectives can be transferred to data mining objectives. The data mining goals of this project to be completed are the following:

- Examine the correlation between carbon dioxide (CO2) emissions within the agri-food sector and the subsequent temperature rise.
- Analyse the influence of various countries based on aggregated data on emissions and temperature change.
- Identify the countries with the highest average temperature increase by year and

analyse their contributions to the overall environmental impact.

1.4 Produce a project plan

Table 1. Project plan

Phase	Time	Resources	Risks
Business understanding – Identify the objectives of the business	20th July	All analysts	Data problems
Business understanding – Assess the situation	21st July	All analysts	Data problems
Business understanding – Determine data mining objectives	22nd July	All analysts	Data problems
Business understanding – Produce a project plan	23rd July	All analysts	Data problems
Data understanding – Collect initial data	24th July	All analysts	Data problems, technology problems
Data understanding – Describe the data	25th July	All analysts	Data problems, technology problems
Data understanding – Explore the data	26th July	All analysts	Data problems, technology problems
Data understanding – Verify the data quality	27th July	All analysts	Data problems, technology problems
Data preparation – Select the data	29th July	Data mining consultant, database analyst	Data problems, technology problems
Data preparation – Clean the data	30th July	Data mining consultant, database analyst	Data problems, technology problems
Data preparation – Construct the data	31st July	Data mining consultant, database analyst	Data problems, technology problems

Data preparation – Integrate various data sources	1st August	Data mining consultant, database analyst	Data problems, technology problems
Data preparation – Format the data as required	2nd August	Data mining consultant, database analyst	Data problems, technology problems
Data transformation – Reduce the data	3rd August	Data mining consultant, database analyst	Data problems, technology problems
Data transformation – Project the data	4th August	Data mining consultant, database analyst	Data problems, technology problems
Data-mining methods selection – Match and discuss the objectives of data-mining to data mining methods	5th August	Data mining consultant, database analyst	Technology problems, inability to find an adequate model
Data-mining methods selection – Select the appropriate data-mining method based on discussion	6th August	Data mining consultant, database analyst	Technology problems, inability to find an adequate model
Data-mining algorithms selection – Conduct exploratory analysis and discuss	7th August	Data mining consultant, database analyst	Technology problems, inability to find an adequate model
Data-mining algorithms selection – Select data-mining algorithms based on discussion	8th August	Data mining consultant, database analyst	Technology problems, inability to find an adequate model
Data-mining algorithms selection – Build/Select appropriate models and choose relevant parameters	9th August	Data mining consultant, database analyst	Technology problems, inability to find an adequate model
Data mining – Create and justify test designs	10th August	Data mining consultant,	Technology problems

		database analyst	
Data mining – Conduct data mining: classify, regress, cluster, etc. (models must execute)	11th August	Data mining consultant, database analyst	Technology problems
Data mining – Search for patterns	12th August	Data mining consultant, database analyst	Technology problems
Interpretation – Study and discuss the mined patterns	13th August	All analysts	Inability to implement results
Interpretation – Visualize the data, results, models, and patterns	14th August	All analysts	Inability to implement results
Interpretation – Interpret the results, models, and patterns	15th August	All analysts	Inability to implement results
Interpretation – Assess and evaluate results, models, and patterns	16th August	All analysts	Inability to implement results
Interpretation – Iterate prior steps (1-7) as required	17th August	All analysts	Inability to implement results

2. Data understanding

2.1 Collect initial data

The compilation of the agricultural carbon dioxide (CO2) emission dataset involved the integration and refinement of around twelve distinct datasets sourced from the Food and Agriculture Organisation (FAO) as well as data obtained from the Intergovernmental Panel on Climate Change (IPCC). The dataset is from the website https://www.kaggle.com/datasets/alessandrolobello/agri-food-co2-emission-dataset-forecasting-ml.

2.2 Describe the data

All features show the corresponding CO2 emissions. CO2 is recorded in kilotons (kt); 1 kt represents 1,000,000 kg of CO2. The "Average Temperature C°" feature serves as the

machine learning model's target variable and signifies the mean annual temperature rise. For instance, when the value is 0.12, the temperature experienced at a specific location has risen by 0.12 degrees Celsius.

All the dataset features are the following:

Table 2. Dataset features

Features	Explanation
Savanna fires	Emissions from fires in savanna ecosystems
Forest fires	Emissions from fires in forested areas.
Crop residues	Emissions from burning or decomposing leftover plant material after crop harvesting.
Rice cultivation	Emissions from methane released during rice cultivation.
Drained organic soils (CO2)	Emissions from carbon dioxide released when draining organic soils.
Pesticides manufacturing	Emissions from the production of pesticides.
Food transport	Emissions from transporting food products.
Forestland	Land covered by forests.
Net forest conversion	Change in forest area due to deforestation and afforestation.
Food household consumption	Emissions from food consumption at the household level.
Food retail	Emissions from the operation of retail establishments selling food.
On-farm electricity use	Electricity consumption on farms.
Food packaging	Emissions from the production and disposal of food packaging materials.
Agrifood system waste disposal	Emissions from waste disposal in the agrifood system.
Food processing	Emissions from processing food products.
Fertilizers manufacturing	Emissions from the production of fertilizers.
IPPU	Emissions from industrial processes and product use.
Manure applied to soils	Emissions from applying animal manure to agricultural soils.
Manure left on pasture	Emissions from animal manure on pasture or grazing land.
Measure management	Emissions from managing and treating animal manure.

Fires in organic soils	Emissions from fires in organic soils.
Fires in humid tropical forests	Emissions from fires in humid tropical forests.
On-farm energy use	Energy consumption on farms.
Rural population	Number of people living in rural areas.
Urban population	Number of people living in urban areas.
Total population – Male	The total number of male individuals in the population.
Total population – Female	The total number of female individuals in the population.
Total emission	Total greenhouse gas emissions from various sources.
Average temperature ℃	The average increase of temperature (by year) in degrees Celsius,

2.3 Explore the data

Figure 1 shows the partial content of the whole dataset, generated from the IBM SPSS Modeler.

Figure 1. Partial content of the dataset

	Area	Year	Savanna fires	Forest fires	Crop Residues	Rice Cultivation	Drained organic soils (CO2)	Pesticides Manufacturing	Food Transport	Forestland	Net Forest conversion	Food Household Consumption	Food Retail	On-farm Electricit
1	Afghanistan	1990	14.724	0.056	205.608	686.000	0.000	11.807	63.115	-2388.803	0	79.085	109.645	
2	Afghanistan	1991	14.724	0.056	209.497	678.160	0.000	11.712	61.212	-2388.803	0	80.489	116.679	
3	Afghanistan	1992	14.724	0.056	196.534	686.000	0.000	11.712	53.317	-2388.803	0	80.769	126.172	
4	Afghanistan	1993	14.724	0.056	230.817	686.000	0.000	11.712	54.362	-2388.803	0	85.068	81.461	
5	Afghanistan	1994	14.724	0.056	242.049	705.600	0.000	11.712	53.987	-2388.803	0	88.806	90.401	
6	Afghanistan	1995	14.724	0.056	243.815	666.400	0.000	11.712	54.645	-2388.803	0	90.163	98.868	
7	Afghanistan	1996	38.930	0.201	249.036	686.000	0.000	11.712	53.164	-2388.803	0	93.790	21.646	
8	Afghanistan	1997	30.938	0.119	276.294	705.600	0.000	11.712	52.039	-2388.803	0	93.970	28.213	
9	Afghanistan	1998	64.141	0.326	287.435	705.600	0.000	11.712	52.705	-2388.803	0	95.260	30.887	
10	Afghanistan	1999	46.168	0.089	247.498	548.800	0.000	11.712	35.763	-2388.803	0	98.988	39.432	
11	Afghanistan	2000	22.781	0.711	168.807	509.600	0.000	11.712	38.556	-2388.803	0	103.419	73.740	
12	Afghanistan	2001	0.222	0.000	170.988	474.320	0.000	11.712	39.194	121.902	0	105.341	102.012	
13	Afghanistan	2002	9.056	0.000	266.197	529.200	0.000	11.712	37.525	121.902	0	108.796	128.807	
14	Afghanistan	2003	55.805	0.000	324.219	568.400	0.000	11.712	60.701	121.902	0	110.807	157.558	
15	Afghanistan	2004	11.976	0.000	267.000	764.400	0.000	11.712	48.759	121.902	0	105.601	190.242	
16	Afghanistan	2005	5.326	0.000	383.750	627.200	0.000	11.983	73.181	121.902	0	115.725	230.999	
17	Afghanistan	2006	4.408	0.000	333.609	627.200	0.000	12.931	103.285	121.902	0	107.851	241.914	
18	Afghanistan	2007	2.824	0.000	403.375	666.400	0.000	13.429	114.756	121.902	0	113.617	246.270	
19	Afghanistan	2008	27.762	0.000	287.910	744.800	0.000	29.920	230.595	121.902	0	130.390	254.078	
20	Afghanistan	2009	2.618	0.000	451.865	784.000	0.000	75.016	385.583	121.902	0	188.672	261.103	
21	Afghanistan	2010	24.811	0.000	413.647	815.360	0.000	81.611	468.253	121.902	0	286.095	255.116	
22	Afghanistan	2011	1.841	0.000	335.038	823.200	0.000	81.611	478.814	-246.219	0	522.628	270.607	
23	Afghanistan	2012	2.896	0.000	445.596	803.600	0.000	107.386	530.821	-246.219	0	534.407	271.240	
24	Afghanistan	2013	3.159	0.000	455.073	803.600	0.000	76.062	391.078	-246.219	0	833.232	276.240	
25	Afghanistan	2014	2.680	0.000	473.417	862.400	0.000	49.783	304.180	-246.219	0	1094.134	333.425	
26	Afghanistan	2015	0.845	0.000	403.318	642.880	0.000	81.853	440.031	-246.219	0	1570.339	370.604	
27	Afghanistan	2016	1.656	0.000	387.613	466.480	0.000	54.910	340.893	154.657	0	1649.838	425.935	
28	Afghanistan	2017	0.402	0.000	344.645	429.052	0.000	55.148	345.761	154.657	0	1431.530	477.331	
29	Afghanistan	2018	0.201	0.000	291.784	460.753	0.000	72.743	407.631	154.657	0	1392.548	534.826	
30	Afghanistan	2019	7.105	0.000	395.269	499.918	0.000	80.807	489.725	154.657	0	1520.258	590.611	
31	Afghanistan	2020	10.843	0.000	427.528	578.416	0.000	107.628	545.324	154.657	0	1548.140	630.758	1
32	Albania	1990	5.556	7.025	59.239	23.520	110.570	2.000	46.965	72.858	0	16.012	8.301	
33	Albania	1991	5.556	7.025	31.462	6.272	110.570	2.000	47.952	72.858	0	11.466	2.745	
34	Albania	1992	5.556	7.025	29.937	1.882	110.570	2.000	40.527	72.858	0	7.908	3.428	
35	Albania	1993	5.556	7.025	44.055	1.098	110.570	2.000	57.659	72.858	0	16.522	7.010	
36	Albania	1994	5.556	7.025	42.425	0.000	110.570	3.000	72.424	72.858	0	17.563	5.780	
37	Albania	1995	5.556	7.025	41.833	0.000	110.570	3.000	72.692	72.858	0	11.555	4.276	
38	Albania	1996	1.347	13.328	33.198	283.181	110.570	1.000	71.226	72.858	0	9.323	4.671	
39	Albania	1997	2.123	33.092	39.309	265.098	110.570	3.000	56.804	72.858	0	11.233	4.587	
40	Albania	1998	2.062	16.283	39.639	265.098	110.570	3.000	85.656	72.858	0	15.107	6.311	
41	Albania	1999	4.300	13.160	32.508	265.098	110.534	3.000	149.611	72.858	0	49.681	8.290	
42	Albania	2000	3.651	77.533	36.513	265.098	110.534	4.000	177.899	72.858	0	64.999	11.193	

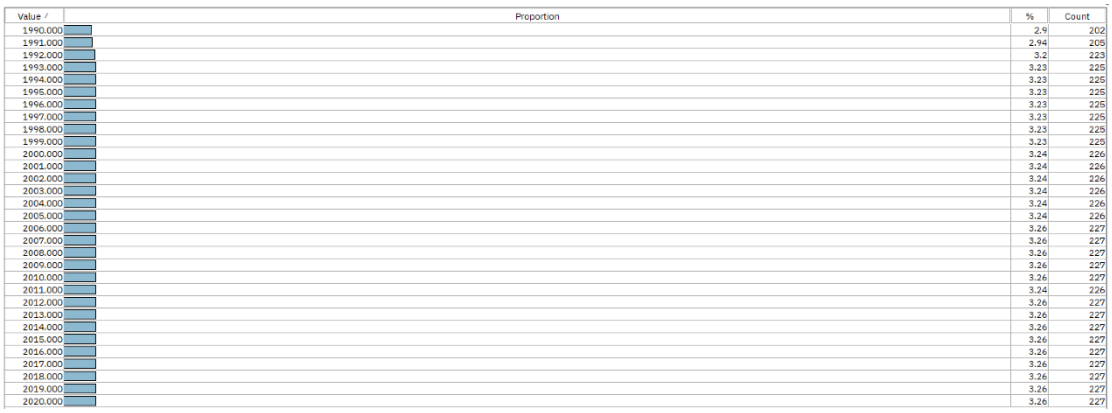
Figure 2 shows the partial content of the data distribution based on countries.

Figure 2. Partial content of the data distribution based on countries

Value	Proportion	%	Count
Afghanistan		0.45	31
Albania		0.45	31
Algeria		0.45	31
American Samoa		0.45	31
Andorra		0.45	31
Angola		0.45	31
Anguilla		0.45	31
Antigua and Barbuda		0.45	31
Argentina		0.45	31
Armenia		0.42	29
Aruba		0.45	31
Australia		0.45	31
Austria		0.45	31
Azerbaijan		0.42	29
Bahamas		0.45	31
Bahrain		0.45	31
Bangladesh		0.45	31
Barbados		0.45	31
Belarus		0.42	29
Belgium		0.3	21
Belgium-Luxembourg		0.14	10
Belize		0.45	31
Benin		0.45	31
Bermuda		0.45	31
Bhutan		0.45	31
Bolivia (Plurinational State of)		0.45	31
Bosnia and Herzegovina		0.42	29
Botswana		0.45	31
Brazil		0.45	31
British Virgin Islands		0.45	31
Brunei Darussalam		0.45	31
Bulgaria		0.45	31
Burkina Faso		0.45	31
Burundi		0.45	31
Cabo Verde		0.45	31
Cambodia		0.45	31
Cameroon		0.45	31
Canada		0.45	31
Cayman Islands		0.45	31
Central African Republic		0.45	31
Chad		0.45	31
Channel Islands		0.45	31
Chile		0.45	31

Figure 3 shows the content of the data distribution based on years.

Figure 3. Data distribution based on years



Value	Proportion	%	Count
1990.000		2.9	202
1991.000		2.94	205
1992.000		3.2	223
1993.000		3.23	225
1994.000		3.23	225
1995.000		3.23	225
1996.000		3.23	225
1997.000		3.23	225
1998.000		3.23	225
1999.000		3.23	225
2000.000		3.24	226
2001.000		3.24	226
2002.000		3.24	226
2003.000		3.24	226
2004.000		3.24	226
2005.000		3.24	226
2006.000		3.26	227
2007.000		3.26	227
2008.000		3.26	227
2009.000		3.26	227
2010.000		3.26	227
2011.000		3.24	226
2012.000		3.26	227
2013.000		3.26	227
2014.000		3.26	227
2015.000		3.26	227
2016.000		3.26	227
2017.000		3.26	227
2018.000		3.26	227
2019.000		3.26	227
2020.000		3.26	227

Figure 4 shows the total global carbon dioxide emission each year from 1990 to 2020.

Figure 4. Total global emission of carbon dioxide each year

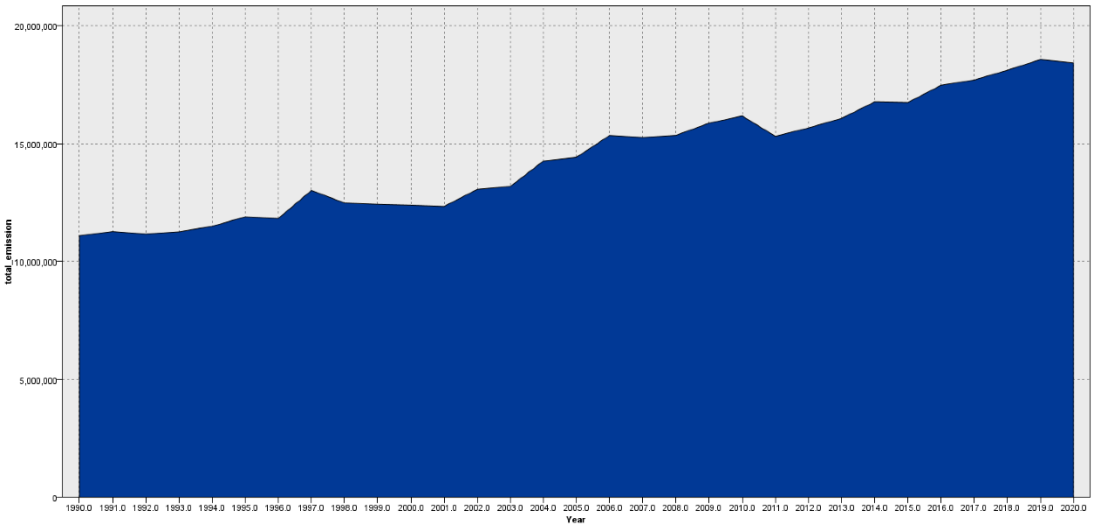


Figure 5 shows each country's total carbon dioxide emission from 1990 to 2020.

Figure 5. Each country's total carbon dioxide emission

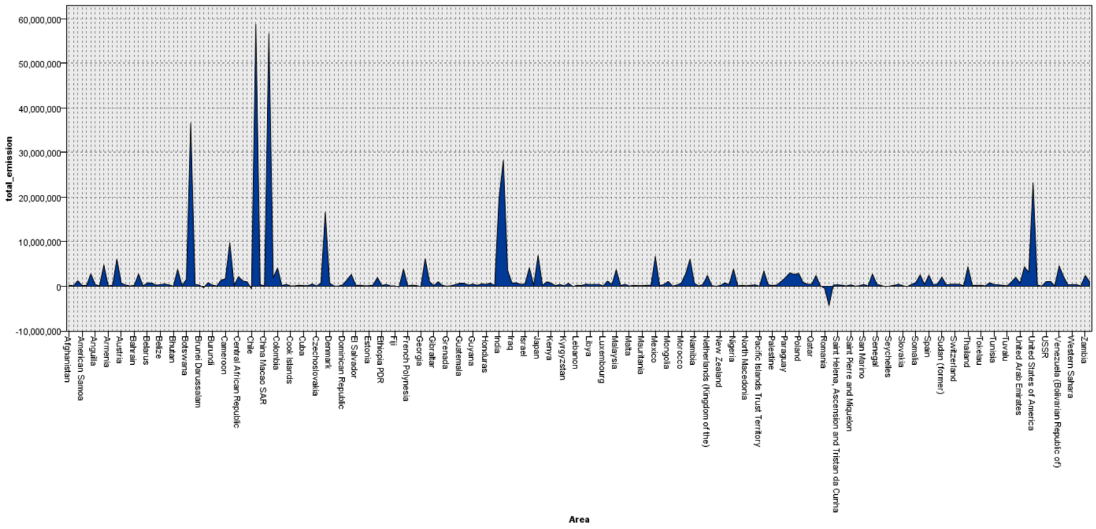


Figure 6 shows the CO2 emission and temperature change by year.

Figure 6. CO2 emission and temperature

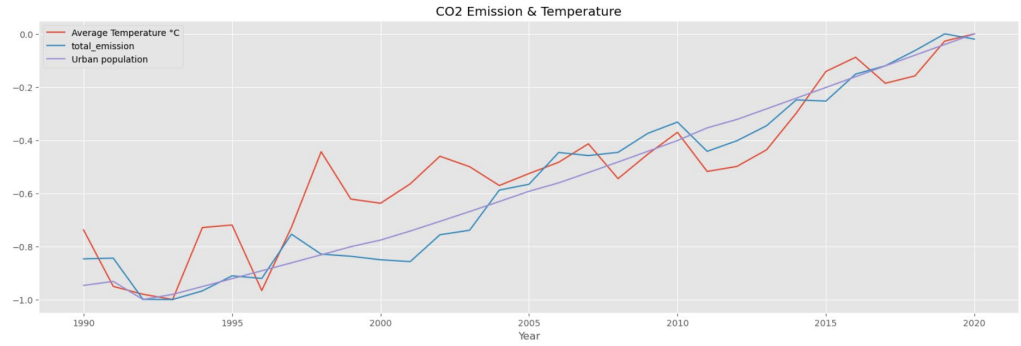
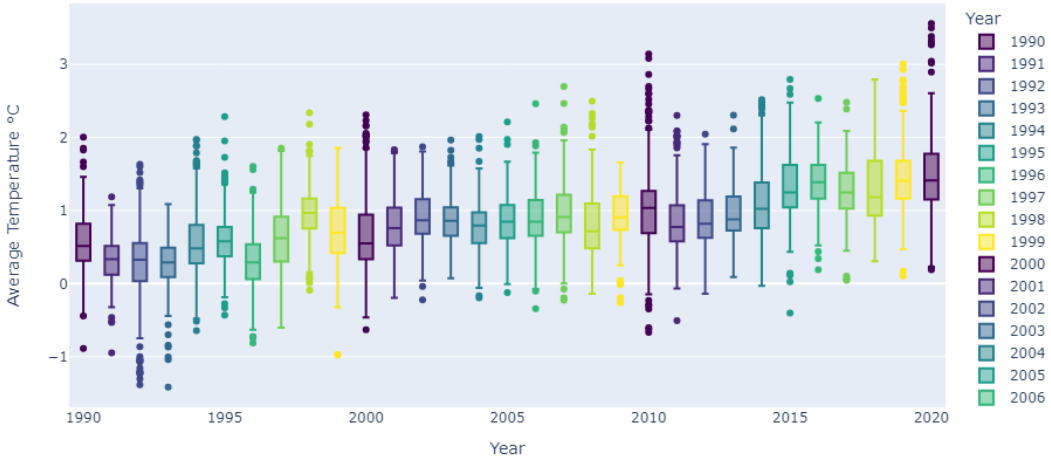


Figure 7 shows the average temperature distribution by year.

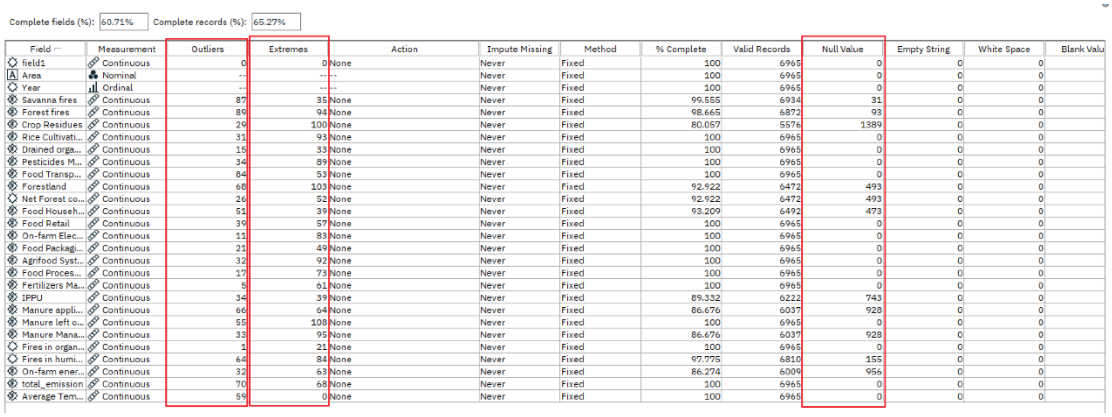
Figure 7. Average temperature distribution by years



2.4 Verify the data quality

Data need to be cleaned and prepared for machine learning models. Missing values, outliers, and feature engineering should be handled with advanced regression techniques. Data quality assessment is frequently conducted throughout description and exploration stages. Figure 8 shows each feature's missing values, outliers, and extremes.

Figure 8. Missing values, outliers, and extremes of each feature



Complete fields (%): 60.71% Complete records (%): 65.27%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Valu
field1	Continuous	0	0	None	Never	Fixed	100	6965	0	0	0	
Area	Nominal	--	--	--	Never	Fixed	100	6965	0	0	0	
Year	Ordinal	--	--	--	Never	Fixed	100	6965	0	0	0	
Savanna fires	Continuous	87	35	None	Never	Fixed	99.555	6934	31	0	0	
Forest fires	Continuous	89	94	None	Never	Fixed	98.665	6872	93	0	0	
Crop Residues	Continuous	29	100	None	Never	Fixed	80.057	5576	1389	0	0	
Rice Cultivati...	Continuous	31	93	None	Never	Fixed	100	6965	0	0	0	
Drained orga...	Continuous	15	33	None	Never	Fixed	100	6965	0	0	0	
Pesticides M...	Continuous	34	89	None	Never	Fixed	100	6965	0	0	0	
Food Transp...	Continuous	84	53	None	Never	Fixed	100	6965	0	0	0	
Forestland	Continuous	68	103	None	Never	Fixed	92.922	6472	493	0	0	
Net Forest co...	Continuous	26	52	None	Never	Fixed	92.922	6472	493	0	0	
Food Househ...	Continuous	51	39	None	Never	Fixed	93.209	6492	473	0	0	
Food Retail	Continuous	39	57	None	Never	Fixed	100	6965	0	0	0	
On-farm Elec...	Continuous	11	83	None	Never	Fixed	100	6965	0	0	0	
Food Packagi...	Continuous	21	49	None	Never	Fixed	100	6965	0	0	0	
Agrifood Syst...	Continuous	32	92	None	Never	Fixed	100	6965	0	0	0	
Food Proces...	Continuous	17	73	None	Never	Fixed	100	6965	0	0	0	
Fertilizers Ma...	Continuous	5	61	None	Never	Fixed	100	6965	0	0	0	
IPPU	Continuous	34	39	None	Never	Fixed	89.332	6222	743	0	0	
Manure appli...	Continuous	66	64	None	Never	Fixed	86.676	6037	928	0	0	
Manure left o...	Continuous	55	108	None	Never	Fixed	100	6965	0	0	0	
Manure Mana...	Continuous	33	95	None	Never	Fixed	86.676	6037	928	0	0	
Fires in organ...	Continuous	1	21	None	Never	Fixed	100	6965	0	0	0	
Fires in humi...	Continuous	64	84	None	Never	Fixed	97.775	6810	155	0	0	
On-farm ener...	Continuous	32	63	None	Never	Fixed	86.274	6009	956	0	0	
total_emission	Continuous	70	68	None	Never	Fixed	100	6965	0	0	0	
Average Tem...	Continuous	59	0	None	Never	Fixed	100	6965	0	0	0	

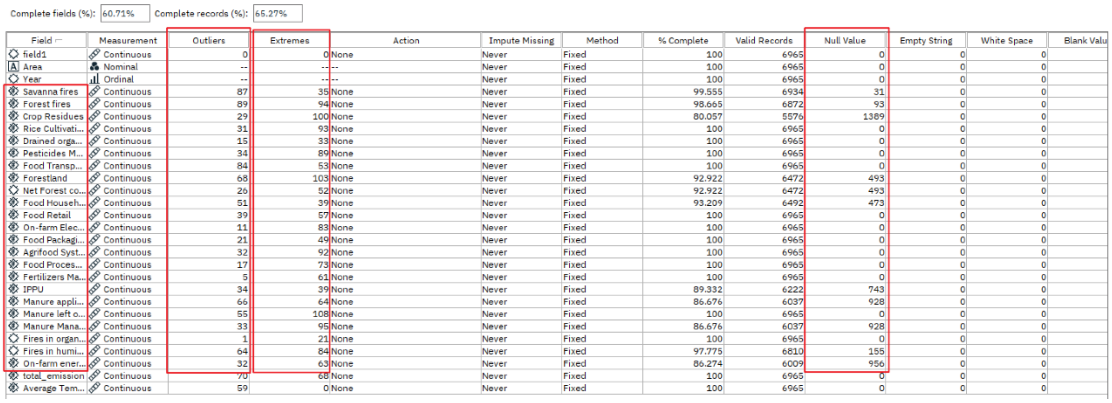
3. Data preparation

3.1 Select the data

After a profound understanding, according to the data collected during the initial phase of the CRISP-DM methodology, the data relevant to the data mining goals is selected. This part should contain selecting items and selecting attributes. In this project, the crucial data mining objectives are to analyse the influence of various countries based on aggregated data on emissions and temperature change, identify the countries with the highest average temperature increase by year, and analyse their contributions to the overall environmental impact. Thus, all countries are considered, which means all items should be considered, so all items are selected. For selecting attributes, one of the data mining goals is to examine the correlation between carbon dioxide (CO2) emissions within the agri-food sector and the subsequent temperature rise, the attribute ‘total_emission’ is the summation of all types of carbon dioxide emissions from the agri-food system. Therefore, all features relevant to carbon dioxide emissions from the agrifood system are selected and considered. Only ‘Rural population’, ‘Urban population’, ‘Total Population – Male’, and ‘Total population – Female’ are excluded.

The next section will clean all data qualities including outliers and missing values. As Figure 9 shows, for the missing values, the attributes ‘Savanna fires’, ‘Forest fires’, ‘Crop Residues’, ‘Forestland’, ‘Net Forest conversion’, ‘Food Household Consumption’, ‘IPPU’, ‘Manure applied to Soils’, ‘Manure Management’, ‘Fires in humid tropical forests’, and ‘On-farm energy use’ are cleaned. For the outliers and extremes, the attributes from ‘Savanna fires’ to ‘On-farm energy use’ are cleaned.

Figure 9. Data selection



Complete fields (%): 60.71% Complete records (%): 65.27%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Valu
field1	Continuous	0	0	None	Never	Fixed	100	6965	0	0	0	
Area	Nominal	--	---		Never	Fixed	100	6965	0	0	0	
Year	Ordinal	--	---		Never	Fixed	100	6965	0	0	0	
Savanna fires	Continuous	87	35	None	Never	Fixed	99.555	6934	31	0	0	
Forest fires	Continuous	89	94	None	Never	Fixed	98.665	6872	93	0	0	
Crop Residues	Continuous	29	100	None	Never	Fixed	80.057	5576	1389	0	0	
Rice Cultivati...	Continuous	31	93	None	Never	Fixed	100	6965	0	0	0	
Drained orga...	Continuous	15	33	None	Never	Fixed	100	6965	0	0	0	
Pesticides M...	Continuous	34	89	None	Never	Fixed	100	6965	0	0	0	
Food Transp...	Continuous	84	53	None	Never	Fixed	100	6965	0	0	0	
Forestland	Continuous	68	103	None	Never	Fixed	92.922	6472	493	0	0	
Net Forest co...	Continuous	26	52	None	Never	Fixed	92.922	6472	493	0	0	
Food Househ...	Continuous	51	39	None	Never	Fixed	93.209	6492	473	0	0	
Food Retail	Continuous	39	57	None	Never	Fixed	100	6965	0	0	0	
On-farm Elec...	Continuous	11	83	None	Never	Fixed	100	6965	0	0	0	
Food Packagi...	Continuous	21	49	None	Never	Fixed	100	6965	0	0	0	
Agrifood Syst...	Continuous	32	92	None	Never	Fixed	100	6965	0	0	0	
Food Proces...	Continuous	17	73	None	Never	Fixed	100	6965	0	0	0	
Fertilizers Ma...	Continuous	5	61	None	Never	Fixed	100	6965	0	0	0	
IPPU	Continuous	34	39	None	Never	Fixed	89.332	6222	743	0	0	
Manure appli...	Continuous	66	64	None	Never	Fixed	86.676	6037	928	0	0	
Manure left o...	Continuous	55	108	None	Never	Fixed	100	6965	0	0	0	
Manure Mana...	Continuous	33	95	None	Never	Fixed	86.676	6037	928	0	0	
Fires in organ...	Continuous	1	21	None	Never	Fixed	100	6965	0	0	0	
Fires in humi...	Continuous	64	84	None	Never	Fixed	97.775	6810	155	0	0	
On-farm ener...	Continuous	32	63	None	Never	Fixed	86.274	6009	956	0	0	
total_emission	Continuous	70	68	None	Never	Fixed	100	6965	0	0	0	
Average Tem...	Continuous	59	0	None	Never	Fixed	100	6965	0	0	0	

3.2 Clean the data

3.2.1 Missing values cleaning

For the missing values of each attribute, typically, one country has no value of the one whole attribute from 1990 to 2020 (Figure 10).

Figure 10. Missing value of one country

	Area	Year	Savanna fires	Forest fires	Crop Residues	Rice Cultivation	Drained organic soils (CO2)	Pesticides Manufacturing	Food Transport	Forestland	Net Forest conversion	Food Household Consumption	Food Retail	On-farm
2000	El Salvador	2015	9.623	33.092	50.714	4.752	0.000	35.000	382.912	-61.501	1142	550.782	266.436	
2001	El Salvador	2016	0.548	0.786	60.226	3.124	0.000	28.000	399.999	-61.501	1142	443.108	266.341	
2002	El Salvador	2017	9.366	31.543	55.657	3.512	0.000	65.000	392.020	-61.501	1142	338.111	245.784	
2003	El Salvador	2018	7.123	9.788	47.307	2.789	0.000	66.000	414.654	-61.501	1142	500.674	247.473	
2004	El Salvador	2019	7.556	18.299	51.025	2.709	0.000	59.000	447.277	-61.501	1142	684.772	264.343	
2005	El Salvador	2020	4.599	7.158	56.138	2.869	0.000	75.000	390.822	-61.501	1142	569.932	267.792	
2006	Equatorial Guinea	1990	0.001	0.223	$null$	1011.444	0.000	11.481	1.415	-5730.083	9560	4.485	10.049	
2007	Equatorial Guinea	1991	0.001	0.223	$null$	1011.444	0.000	11.481	1.415	-5730.083	9560	4.331	10.901	
2008	Equatorial Guinea	1992	0.001	0.223	$null$	921.664	0.000	11.481	1.415	-5730.083	9560	4.571	11.284	
2009	Equatorial Guinea	1993	0.001	0.223	$null$	921.664	0.000	11.481	1.710	-5730.083	9560	4.375	7.206	
2010	Equatorial Guinea	1994	0.001	0.223	$null$	921.664	0.000	11.481	1.530	-5730.083	9560	4.306	8.213	
2011	Equatorial Guinea	1995	0.001	0.223	$null$	921.664	0.000	11.481	2.022	-5730.083	9560	4.241	9.852	
2012	Equatorial Guinea	1996	0.000	0.372	$null$	921.664	0.000	11.481	4.320	-5730.083	9560	4.331	11.442	
2013	Equatorial Guinea	1997	0.000	0.372	$null$	921.664	0.000	11.481	10.260	-5730.083	9560	4.234	13.101	
2014	Equatorial Guinea	1998	0.000	0.372	$null$	921.664	0.000	11.481	8.843	-5730.083	9560	4.026	9.146	
2015	Equatorial Guinea	1999	0.000	0.372	$null$	921.664	0.000	11.481	12.821	-5730.083	9560	6.419	12.175	
2016	Equatorial Guinea	2000	0.000	0.000	$null$	921.664	0.000	11.481	20.181	-5730.083	9560	6.569	21.330	
2017	Equatorial Guinea	2001	0.000	0.000	$null$	764.893	0.000	11.481	23.799	0.000	3830	6.345	28.895	
2018	Equatorial Guinea	2002	0.000	0.000	$null$	764.893	0.000	11.481	32.711	0.000	3830	8.755	36.030	
2019	Equatorial Guinea	2003	0.000	0.000	$null$	764.893	0.000	11.481	37.390	0.000	3830	9.076	43.760	
2020	Equatorial Guinea	2004	0.000	0.000	$null$	764.893	0.000	11.481	39.469	0.000	3830	10.150	52.668	
2021	Equatorial Guinea	2005	0.000	0.000	$null$	764.893	0.000	11.481	41.804	0.000	3830	10.629	65.648	
2022	Equatorial Guinea	2006	0.011	0.416	$null$	764.893	0.000	11.481	46.307	0.000	3830	11.642	70.552	
2023	Equatorial Guinea	2007	0.000	0.139	$null$	764.893	0.000	11.481	48.910	0.000	3830	12.538	72.097	
2024	Equatorial Guinea	2008	0.000	0.278	$null$	764.893	0.000	11.481	50.395	0.000	3830	14.517	73.107	
2025	Equatorial Guinea	2009	0.000	0.000	$null$	764.893	0.000	11.481	52.512	0.000	3830	16.050	76.372	
2026	Equatorial Guinea	2010	0.000	0.139	$null$	764.893	0.000	11.481	54.630	0.000	3830	18.067	76.112	
2027	Equatorial Guinea	2011	0.000	0.278	$null$	764.893	0.000	11.481	57.111	0.000	3831	25.782	76.092	
2028	Equatorial Guinea	2012	0.000	0.139	$null$	764.893	0.000	11.481	59.237	0.000	3831	26.698	88.900	
2029	Equatorial Guinea	2013	0.015	0.000	$null$	764.893	0.000	11.481	61.718	0.000	3831	30.008	89.698	
2030	Equatorial Guinea	2014	0.000	0.416	$null$	764.893	0.000	11.576	63.843	0.000	3831	30.966	97.159	
2031	Equatorial Guinea	2015	0.000	0.000	$null$	764.893	0.000	11.576	65.617	0.000	3831	48.658	104.637	
2032	Equatorial Guinea	2016	0.000	0.000	$null$	764.893	0.000	11.576	67.035	0.000	3832	51.508	117.585	
2033	Equatorial Guinea	2017	0.000	0.278	$null$	764.893	0.000	11.576	70.729	0.000	3832	52.255	129.384	
2034	Equatorial Guinea	2018	0.000	0.416	$null$	764.893	0.000	11.576	65.605	0.000	3832	50.714	146.290	
2035	Equatorial Guinea	2019	0.000	0.000	$null$	764.893	0.658	11.576	72.542	0.000	3832	45.281	159.857	
2036	Equatorial Guinea	2020	0.000	1.388	$null$	764.893	0.658	11.481	51.609	0.000	3832	55.320	174.515	
2037	Eritrea	1993	22.927	3.462	14.749	248.061	8.340	0.000	23.460	-243.151	671	6.371	15.484	
2038	Eritrea	1994	22.927	3.462	20.939	248.061	8.340	0.000	31.993	-243.151	671	6.561	17.342	
2039	Eritrea	1995	22.927	3.462	15.245	248.061	8.340	0.000	36.615	-243.151	671	6.753	24.200	
2040	Eritrea	1996	58.686	0.000	14.936	248.061	8.340	0.000	42.661	-243.151	671	7.322	25.179	
2041	Eritrea	1997	96.251	0.000	17.778	248.061	8.340	1.000	40.166	-243.151	671	7.808	26.445	

Thus, considering that one of the data mining goals is to identify the countries with the highest average temperature increase by year, the cleaning method, fixed as mean, is used to impute the missing values (Figure 11).

Figure 11. Missing values imputing method

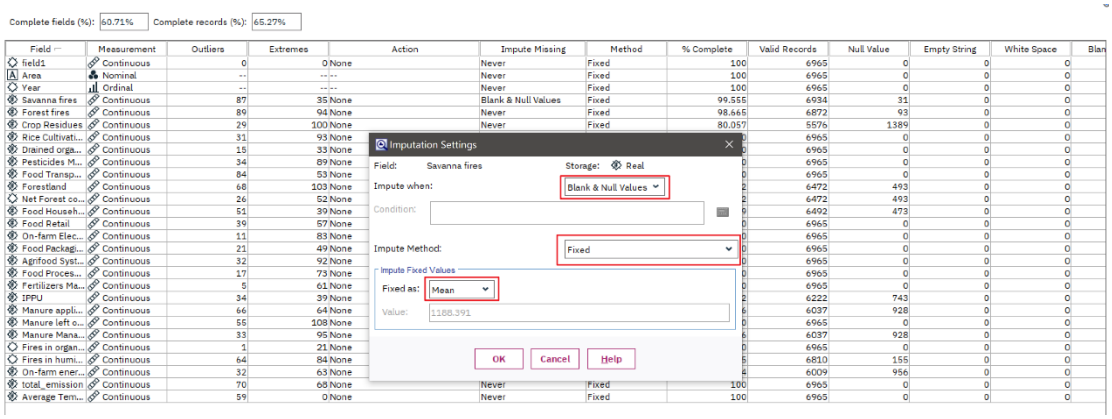
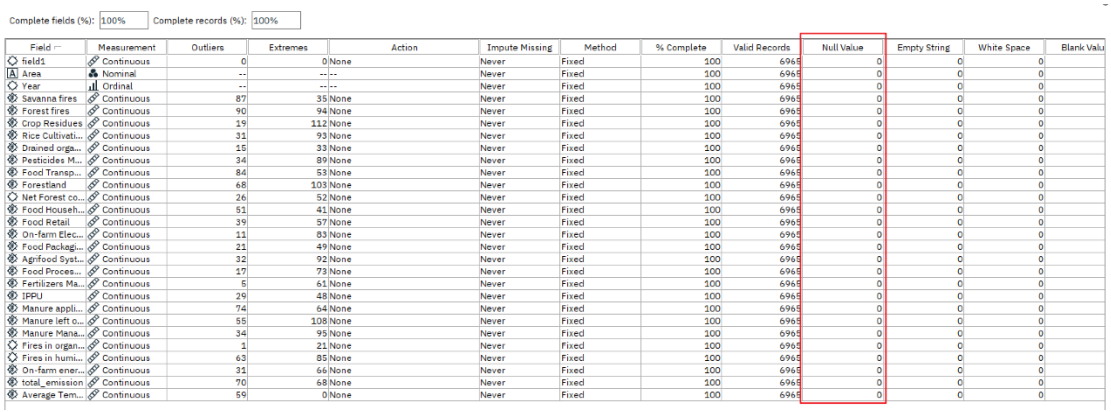


Figure 12 shows the result after cleaning the missing values.

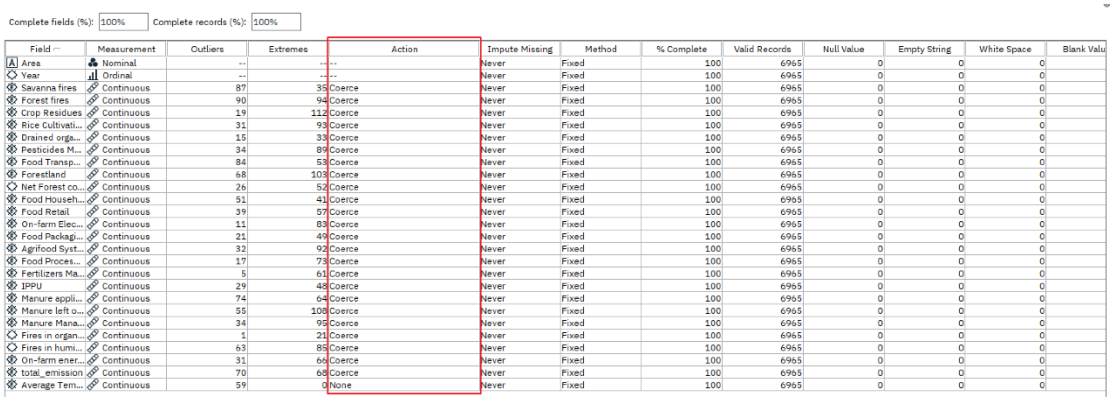
Figure 12. Missing values cleaned



3.2.2 Outliers and extremes cleaning

For outliers and extremes, a standard method of dealing with this is to coerce outliers and extremes. Thus, this project uses this method to clean the outliers and extremes of each attribute (Figure 13).

Figure 13. Outliers and extremes cleaning method

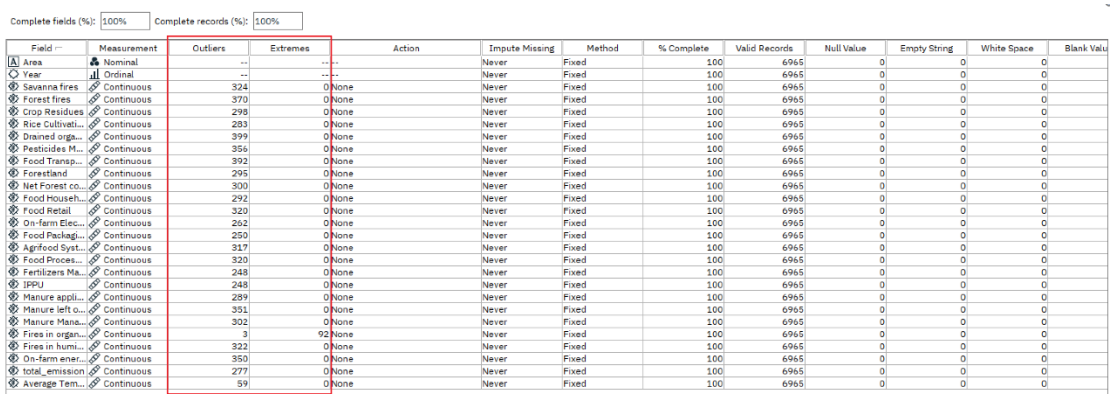


Complete fields (%): 100% Complete records (%): 100%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Valu
Area	Nominal	--	--	--	Never	Fixed	100	6965	0	0	0	
Year	Ordinal	--	--	--	Never	Fixed	100	6965	0	0	0	
Savanna fires	Continuous	87	35	Coerce	Never	Fixed	100	6965	0	0	0	
Forest fires	Continuous	90	94	Coerce	Never	Fixed	100	6965	0	0	0	
Crop Residues	Continuous	19	112	Coerce	Never	Fixed	100	6965	0	0	0	
Rice Cultivati...	Continuous	31	93	Coerce	Never	Fixed	100	6965	0	0	0	
Drained orga...	Continuous	15	33	Coerce	Never	Fixed	100	6965	0	0	0	
Pesticides M...	Continuous	34	89	Coerce	Never	Fixed	100	6965	0	0	0	
Food Transp...	Continuous	84	53	Coerce	Never	Fixed	100	6965	0	0	0	
Forestland	Continuous	68	103	Coerce	Never	Fixed	100	6965	0	0	0	
Net Forest co...	Continuous	26	52	Coerce	Never	Fixed	100	6965	0	0	0	
Food Househ...	Continuous	51	41	Coerce	Never	Fixed	100	6965	0	0	0	
Food Retail	Continuous	39	57	Coerce	Never	Fixed	100	6965	0	0	0	
On-farm Elec...	Continuous	11	83	Coerce	Never	Fixed	100	6965	0	0	0	
Food Packagi...	Continuous	21	49	Coerce	Never	Fixed	100	6965	0	0	0	
Agrifood Syst...	Continuous	32	92	Coerce	Never	Fixed	100	6965	0	0	0	
Food Proces...	Continuous	17	73	Coerce	Never	Fixed	100	6965	0	0	0	
Fertilizers Ma...	Continuous	5	61	Coerce	Never	Fixed	100	6965	0	0	0	
IPPU	Continuous	29	48	Coerce	Never	Fixed	100	6965	0	0	0	
Manure appli...	Continuous	74	64	Coerce	Never	Fixed	100	6965	0	0	0	
Manure left o...	Continuous	55	108	Coerce	Never	Fixed	100	6965	0	0	0	
Manure Mana...	Continuous	34	99	Coerce	Never	Fixed	100	6965	0	0	0	
Fires in organ...	Continuous	1	21	Coerce	Never	Fixed	100	6965	0	0	0	
Fires in humi...	Continuous	63	85	Coerce	Never	Fixed	100	6965	0	0	0	
On-farm ener...	Continuous	31	66	Coerce	Never	Fixed	100	6965	0	0	0	
total_emission	Continuous	70	68	Coerce	Never	Fixed	100	6965	0	0	0	
Average Tem...	Continuous	59	0	None	Never	Fixed	100	6965	0	0	0	

Figure 14 shows the result after coercing the outliers and the extremes by three times iterations.

Figure 14. Outliers and extremes coerced



Complete fields (%): 100% Complete records (%): 100%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Valu
Area	Nominal	--	--	--	Never	Fixed	100	6965	0	0	0	
Year	Ordinal	--	--	--	Never	Fixed	100	6965	0	0	0	
Savanna fires	Continuous	324	0	None	Never	Fixed	100	6965	0	0	0	
Forest fires	Continuous	370	0	None	Never	Fixed	100	6965	0	0	0	
Crop Residues	Continuous	298	0	None	Never	Fixed	100	6965	0	0	0	
Rice Cultivati...	Continuous	283	0	None	Never	Fixed	100	6965	0	0	0	
Drained orga...	Continuous	399	0	None	Never	Fixed	100	6965	0	0	0	
Pesticides M...	Continuous	356	0	None	Never	Fixed	100	6965	0	0	0	
Food Transp...	Continuous	392	0	None	Never	Fixed	100	6965	0	0	0	
Forestland	Continuous	295	0	None	Never	Fixed	100	6965	0	0	0	
Net Forest co...	Continuous	300	0	None	Never	Fixed	100	6965	0	0	0	
Food Househ...	Continuous	292	0	None	Never	Fixed	100	6965	0	0	0	
Food Retail	Continuous	320	0	None	Never	Fixed	100	6965	0	0	0	
On-farm Elec...	Continuous	292	0	None	Never	Fixed	100	6965	0	0	0	
Food Packagi...	Continuous	250	0	None	Never	Fixed	100	6965	0	0	0	
Agrifood Syst...	Continuous	317	0	None	Never	Fixed	100	6965	0	0	0	
Food Proces...	Continuous	320	0	None	Never	Fixed	100	6965	0	0	0	
Fertilizers Ma...	Continuous	248	0	None	Never	Fixed	100	6965	0	0	0	
IPPU	Continuous	248	0	None	Never	Fixed	100	6965	0	0	0	
Manure appli...	Continuous	289	0	None	Never	Fixed	100	6965	0	0	0	
Manure left o...	Continuous	351	0	None	Never	Fixed	100	6965	0	0	0	
Manure Mana...	Continuous	302	0	None	Never	Fixed	100	6965	0	0	0	
Fires in organ...	Continuous	3	92	None	Never	Fixed	100	6965	0	0	0	
Fires in humi...	Continuous	322	0	None	Never	Fixed	100	6965	0	0	0	
On-farm ener...	Continuous	350	0	None	Never	Fixed	100	6965	0	0	0	
total_emission	Continuous	277	0	None	Never	Fixed	100	6965	0	0	0	
Average Tem...	Continuous	59	0	None	Never	Fixed	100	6965	0	0	0	

3.3 Construct the data

All missing values, outliers, and extremes are processed; in the previous table, the feature 'total_emission' is not the summation of all types of carbon dioxide emissions from the agri-food system. A new attribute called 'Updated_total_emission' is constructed, which calculates the summation of the cleaned data of all types of carbon dioxide emissions from the agrifood system (Figure 15). The values of 'Updated_total_emission' is different from the values of 'total_emission' (Figure 16).

Figure 15. New attribute – 'Updated_total_emission'

Complete fields (%): 100% Complete records (%): 100%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Sp
Area	Nominal	--	--	--	Never	Fixed	100	5318	0	0	
Year	Ordinal	--	--	--	Never	Fixed	100	5318	0	0	
Sevanna fires	Continuous	234	0	None	Never	Fixed	100	5318	0	0	
Forest fires	Continuous	155	62	None	Never	Fixed	100	5318	0	0	
Crop Residues	Continuous	2	0	None	Never	Fixed	100	5318	0	0	
Rice Cultivation	Continuous	195	0	None	Never	Fixed	100	5318	0	0	
Drained organic soils (CO2)	Continuous	52	73	None	Never	Fixed	100	5318	0	0	
Pesticides Manufacturing	Continuous	74	41	None	Never	Fixed	100	5318	0	0	
Food Transport	Continuous	60	25	None	Never	Fixed	100	5318	0	0	
Forestland	Continuous	127	0	None	Never	Fixed	100	5318	0	0	
Net Forest conversion	Continuous	132	0	None	Never	Fixed	100	5318	0	0	
Food Household Consumption	Continuous	53	0	None	Never	Fixed	100	5318	0	0	
Food Retail	Continuous	120	25	None	Never	Fixed	100	5318	0	0	
On-farm Electricity Use	Continuous	99	48	None	Never	Fixed	100	5318	0	0	
Food Packaging	Continuous	129	22	None	Never	Fixed	100	5318	0	0	
Agrifood Systems Waste Disposal	Continuous	106	32	None	Never	Fixed	100	5318	0	0	
Food Processing	Continuous	79	18	None	Never	Fixed	100	5318	0	0	
Fertilizers Manufacturing	Continuous	131	0	None	Never	Fixed	100	5318	0	0	
IPPU	Continuous	22	0	None	Never	Fixed	100	5318	0	0	
Manure applied to Soils	Continuous	48	0	None	Never	Fixed	100	5318	0	0	
Manure left on Pasture	Continuous	113	0	None	Never	Fixed	100	5318	0	0	
Manure Management	Continuous	31	0	None	Never	Fixed	100	5318	0	0	
Fires in organic soils	Continuous	1	11	None	Never	Fixed	100	5318	0	0	
Fires in humid tropical forests	Continuous	219	53	None	Never	Fixed	100	5318	0	0	
On-farm energy use	Continuous	17	0	None	Never	Fixed	100	5318	0	0	
total_emission	Continuous	110	15	None	Never	Fixed	100	5318	0	0	
Average Temperature	Continuous	51	0	None	Never	Fixed	100	5318	0	0	
Updated_total_emission	Continuous	55	6	None	Never	Fixed	100	5318	0	0	

Figure 16. Difference between 'Updated_total_emission' and 'total_emission'

	i	Fertilizers Manufacturing	IPPU	Manure applied to Soils	Manure left on Pasture	Manure Management	Fires in organic soils	Fires in humid tropical forests	On-farm energy use	total_emission	Average Temperature	Updated_total_emission
1	4	11.997	209.978	260.143	1590.532	319.176	0.000	0.000	3008.982	2198.964	0.536	5207.946
2	4	12.854	217.039	268.629	1657.236	342.308	0.000	0.000	3008.982	2323.877	0.022	5332.859
3	4	13.493	222.116	264.790	1633.507	349.122	0.000	0.000	3008.982	2356.304	-0.260	5365.286
4	4	14.056	201.206	261.722	1642.962	352.295	0.000	0.000	3008.982	2368.471	0.102	5377.453
5	4	15.127	182.291	267.622	1689.359	367.678	0.000	0.000	3008.982	2500.769	0.372	5509.751
6	4	15.912	174.365	275.236	1779.314	397.550	0.000	0.000	3008.982	2624.613	0.286	5633.595
7	4	16.953	165.423	310.131	1900.587	465.205	0.000	0.000	3008.982	2838.921	0.037	5847.903
8	5	18.108	164.468	338.933	2110.637	511.593	0.000	0.000	3008.982	3204.180	0.415	6213.162
9	3	19.181	163.505	362.568	2305.394	541.660	0.000	0.000	3008.982	3560.717	0.893	6569.699
10	0	20.421	163.550	400.556	2554.690	611.061	0.000	0.000	3008.982	3694.807	1.058	6703.789
11	5	21.278	164.603	343.392	2247.032	517.493	0.000	0.000	3008.982	3113.528	0.976	6122.510
12	3	22.310	151.675	297.226	1873.850	426.206	0.000	0.000	3008.982	5038.534	1.409	8047.516
13	3	44.588	157.754	313.813	2151.399	592.561	0.000	0.000	3008.982	6035.816	1.084	9044.798
14	4	31.323	158.842	320.903	2192.624	603.102	0.000	0.000	3008.982	6449.089	0.679	9458.071
15	4	36.438	160.927	324.195	2222.926	576.037	0.000	0.000	3008.982	6734.998	1.399	9743.980
16	0	29.126	161.001	339.439	2229.918	604.767	0.000	0.000	3008.982	7001.298	0.457	10010.280
17	4	33.989	168.067	325.850	2149.666	626.243	0.000	0.000	3008.982	7076.182	1.477	10085.164
18	3	31.476	170.125	316.075	2050.130	647.468	0.000	0.000	3008.982	7281.053	0.786	10290.035
19	1	28.599	177.178	367.378	2343.206	715.934	0.000	0.000	3008.982	8069.086	0.836	11078.068
20	1	26.495	179.237	383.555	2403.608	725.441	0.000	0.000	3008.982	8735.042	0.897	11744.024
21	5	22.380	192.300	438.123	2737.493	849.335	0.000	0.000	3008.982	9871.639	1.529	12880.621
22	4	32.493	192.372	449.471	2840.266	843.171	0.000	0.000	3008.982	10925.534	1.258	13934.516
23	3	29.443	266.451	441.855	2798.736	836.909	0.000	0.000	3008.982	11280.277	0.344	14289.259
24	0	29.981	333.533	432.003	2741.099	832.896	0.000	0.000	3008.982	11694.004	1.291	14702.986
25	7	28.639	282.615	437.620	2770.467	842.504	0.000	0.000	3008.982	12128.698	0.469	15137.680
26	0	27.235	256.703	423.504	2719.153	797.901	0.000	0.000	3008.982	12652.876	1.104	15661.858
27	4	27.847	302.756	421.927	2692.957	793.938	0.000	0.000	3008.982	12988.384	1.839	15997.366
28	5	35.309	273.836	422.402	2680.838	768.680	0.000	0.000	3008.982	12786.219	1.290	15795.201
29	3	35.501	298.915	427.507	2716.908	789.336	0.000	0.000	3008.982	13054.983	1.612	16063.965
30	5	35.555	308.968	387.950	2557.433	766.501	0.000	0.000	3008.982	13354.360	0.927	16363.342
31	3	38.218	341.048	426.436	2743.595	785.332	0.000	0.000	3008.982	14032.421	0.209	17041.403
32	9	356.491	767.146	196.644	383.307	474.695	0.000	0.000	320.880	3475.291	0.736	3475.291
33	9	2903.861	628.660	194.185	394.069	469.142	0.000	0.000	195.188	5680.136	-0.462	5680.136
34	9	2903.861	247.204	187.375	400.054	453.218	0.000	0.000	163.133	5221.132	0.318	5221.132
35	9	2903.861	263.811	198.809	426.417	481.384	0.000	0.000	126.849	5316.973	0.271	5316.973
36	9	2903.861	212.374	260.301	546.010	620.299	0.000	0.000	94.954	5564.288	1.203	5564.288
37	3	2903.861	278.264	267.215	546.743	639.653	0.000	0.000	159.286	5582.371	-0.039	5582.371
38	0	377.408	266.198	258.445	463.034	634.935	0.000	0.000	152.943	3231.957	-0.146	3231.957
39	3	377.408	199.124	242.095	433.554	586.431	0.000	0.000	114.769	3035.164	0.087	3035.164
40	4	377.408	193.148	230.314	411.251	558.918	0.000	0.000	142.219	3033.239	0.582	3033.239
41	3	377.408	223.296	235.616	427.838	570.972	0.000	0.000	255.629	3336.639	0.942	3336.639
42	5	377.408	245.508	242.061	428.117	589.822	0.000	0.000	248.065	3516.551	1.119	3516.551

3.4 Integrate various data resources

A new dataset is merged into the previous dataset. The new dataset has the following features: 'Area', 'Year', 'Rural population', 'Urban population', 'Total Population – Male', 'Total Population – Female', 'total_emission', and 'Average Temperature'. The features 'Area', 'Year', 'total_emission', and 'Average Temperature' are duplicated, the same as the corresponding features of the previous dataset. Thus, the 'Keys' merge method is used, combining the duplicated key fields (Figure 17). Figure 18 and Figure 19 show the merged table.

Figure 17. Merge method

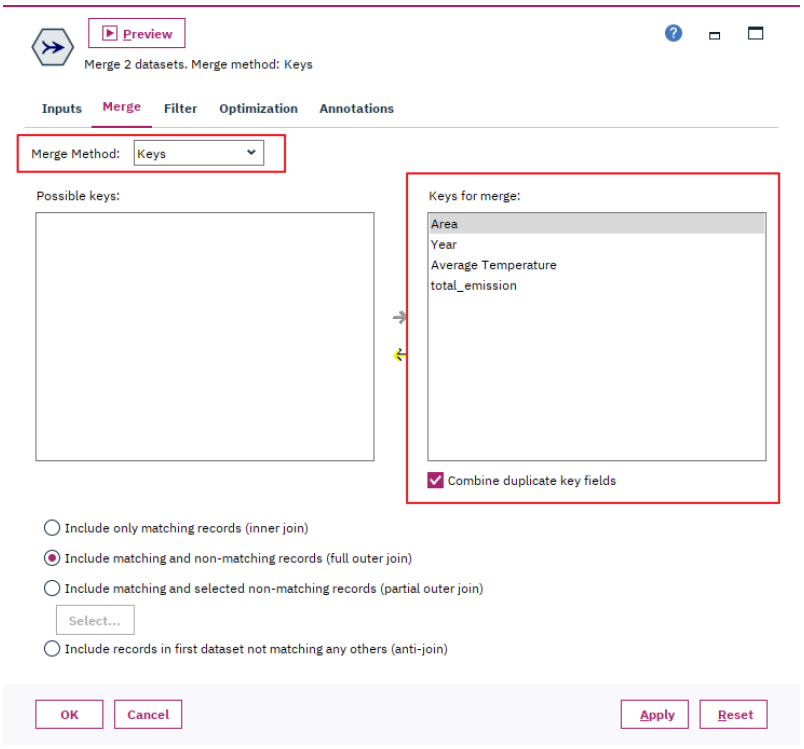


Figure 18. Merged table

	Area	Year	Average Temperature	total_emission	Savanna fires	Forest fires	Crop Residues	Rice Cultivation	Drained organic soils (CO2)	Pesticides Manufacturing	Food Transport	Forestland	Net Forest conversion	Food Household Consu
1	Afghanistan	1990	0.536	2198.964	14.724	0.056	205.608	686.000	0.000	11.807	63.115	-2388.803	0.000	
2	Afghanistan	1991	0.021	2323.877	14.724	0.056	209.497	678.160	0.000	11.712	61.212	-2388.803	0.000	
3	Afghanistan	1992	-0.260	2356.304	14.724	0.056	196.534	686.000	0.000	11.712	53.317	-2388.803	0.000	
4	Afghanistan	1993	0.102	2368.471	14.724	0.056	230.817	686.000	0.000	11.712	54.362	-2388.803	0.000	
5	Afghanistan	1994	0.372	2500.769	14.724	0.056	242.049	705.600	0.000	11.712	53.987	-2388.803	0.000	
6	Afghanistan	1995	0.286	2624.613	14.724	0.056	243.815	666.400	0.000	11.712	54.645	-2388.803	0.000	
7	Afghanistan	1996	0.037	2838.921	38.930	0.201	249.036	686.000	0.000	11.712	53.164	-2388.803	0.000	
8	Afghanistan	1997	0.415	3204.180	30.938	0.119	276.294	705.600	0.000	11.712	52.039	-2388.803	0.000	
9	Afghanistan	1998	0.891	3560.717	64.141	0.326	287.435	705.600	0.000	11.712	52.705	-2388.803	0.000	
10	Afghanistan	1999	1.058	3694.807	46.168	0.089	247.498	548.800	0.000	11.712	35.763	-2388.803	0.000	
11	Afghanistan	2000	0.976	3113.528	22.781	0.711	168.807	509.600	0.000	11.712	38.556	-2388.803	0.000	
12	Afghanistan	2001	1.409	5038.534	0.222	0.000	170.988	474.320	0.000	11.712	39.194	121.902	0.000	
13	Afghanistan	2002	1.084	6035.816	9.056	0.000	266.197	529.200	0.000	11.712	37.525	121.902	0.000	
14	Afghanistan	2003	0.679	6449.089	55.805	0.000	324.219	568.400	0.000	11.712	60.701	121.902	0.000	
15	Afghanistan	2004	1.399	6734.998	11.976	0.000	267.000	764.400	0.000	11.712	48.759	121.902	0.000	
16	Afghanistan	2005	0.457	7001.298	5.326	0.000	383.750	627.200	0.000	11.983	73.181	121.902	0.000	
17	Afghanistan	2006	1.477	7076.182	4.408	0.000	333.609	627.200	0.000	12.931	103.285	121.902	0.000	
18	Afghanistan	2007	0.786	7281.053	2.824	0.000	403.375	666.400	0.000	13.429	114.756	121.902	0.000	
19	Afghanistan	2008	0.836	8069.086	27.762	0.000	287.910	744.800	0.000	29.920	230.595	121.902	0.000	
20	Afghanistan	2009	0.897	8735.042	2.618	0.000	451.865	784.000	0.000	75.016	385.583	121.902	0.000	
21	Afghanistan	2010	1.529	9871.639	24.811	0.000	413.647	815.360	0.000	81.611	468.253	121.902	0.000	
22	Afghanistan	2011	1.255	10925.534	1.841	0.000	335.038	823.200	0.000	81.611	478.814	-246.219	0.000	
23	Afghanistan	2012	0.344	11280.277	2.896	0.000	445.596	803.600	0.000	107.386	530.821	-246.219	0.000	
24	Afghanistan	2013	1.291	11694.004	3.159	0.000	455.073	803.600	0.000	76.062	391.078	-246.219	0.000	
25	Afghanistan	2014	0.469	12128.698	2.680	0.000	473.417	862.400	0.000	49.783	304.180	-246.219	0.000	1
26	Afghanistan	2015	1.104	12652.876	0.845	0.000	403.318	642.880	0.000	81.853	440.031	-246.219	0.000	1
27	Afghanistan	2016	1.839	12988.384	1.656	0.000	387.613	466.480	0.000	54.910	340.893	154.657	0.000	1
28	Afghanistan	2017	1.290	12786.219	0.402	0.000	344.645	429.052	0.000	55.148	345.761	154.657	0.000	1
29	Afghanistan	2018	1.612	13054.983	0.201	0.000	291.784	460.753	0.000	72.743	407.631	154.657	0.000	1
30	Afghanistan	2019	0.927	13354.360	7.105	0.000	395.269	499.918	0.000	80.807	489.725	154.657	0.000	1
31	Afghanistan	2020	0.209	14032.421	10.843	0.000	427.528	578.416	0.000	107.628	545.324	154.657	0.000	1
32	Albania	1990	0.736	3475.291	5.556	7.025	59.239	23.520	110.570	2.000	46.965	72.858	0.000	
33	Albania	1991	-0.462	5680.136	5.556	7.025	31.462	6.272	110.570	2.000	47.952	72.858	0.000	
34	Albania	1992	0.318	5221.132	5.556	7.025	29.937	1.882	110.570	2.000	40.527	72.858	0.000	
35	Albania	1993	0.271	5316.973	5.556	7.025	44.055	1.098	110.570	2.000	57.659	72.858	0.000	
36	Albania	1994	1.203	5564.288	5.556	7.025	42.425	0.000	110.570	3.000	72.424	72.858	0.000	
37	Albania	1995	-0.039	5582.371	5.556	7.025	41.833	0.000	110.570	3.000	72.692	72.858	0.000	
38	Albania	1996	-0.146	3231.957	1.347	13.328	33.198	283.181	110.570	1.000	71.226	72.858	0.000	
39	Albania	1997	0.087	3035.164	2.123	33.092	39.309	265.098	110.570	3.000	56.804	72.858	0.000	
40	Albania	1998	0.582	3033.239	2.062	16.283	39.639	265.098	110.570	3.000	85.656	72.858	0.000	
41	Albania	1999	0.941	3336.639	4.300	13.160	32.508	265.098	110.534	3.000	149.611	72.858	0.000	
42	Albania	2000	1.119	3516.551	3.651	77.533	36.513	265.098	110.534	4.000	177.899	72.858	0.000	

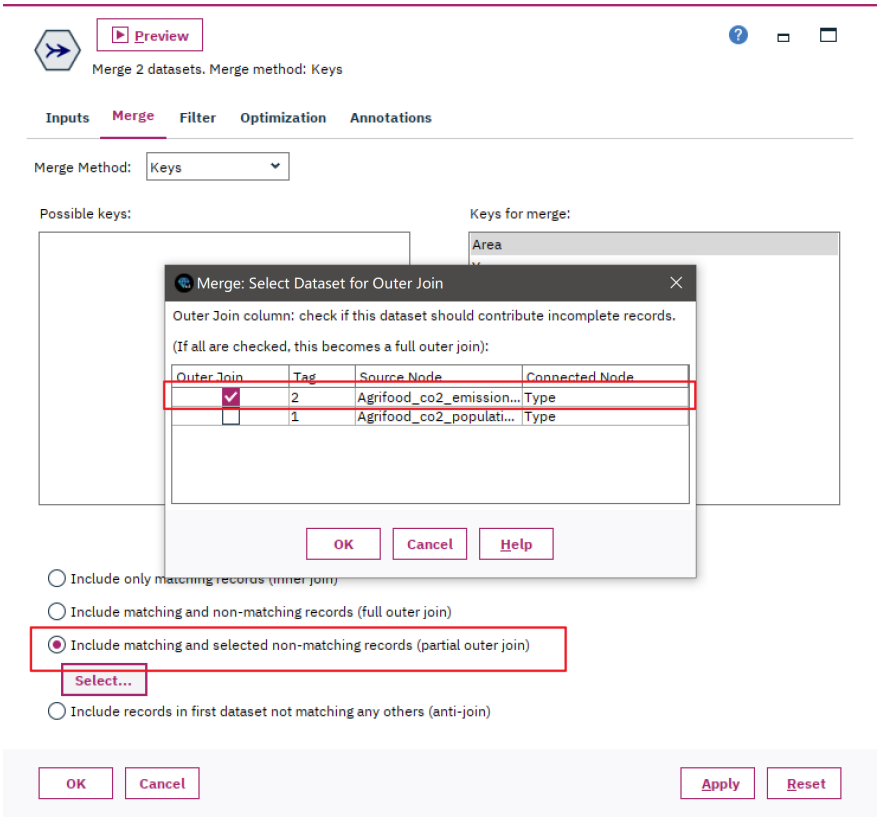
Figure 19. Merged table

	pplied to Soils	Manure left on Pasture	Manure Management	Fires in organic soils	Fires in humid tropical forests	On-farm energy use	Updated_total_emission	Rural population	Urban population	Total Population - Male	Total Population - Female
1	260.143	1590.532	319.176	0.000	0.000	3008.982	5207.940	9699107	2993947	5340507.000	5342409.000
2	268.629	1657.236	342.308	0.000	0.000	3008.982	5332.859	10230490	2763167	5372959.000	5372208.000
3	264.790	1653.507	349.122	0.000	0.000	3008.982	5365.286	10995568	2985663	6028494.000	6028939.000
4	261.722	1642.962	352.295	0.000	0.000	3008.982	5377.453	11858090	3237009	7003641.000	7000119.000
5	267.622	1689.359	367.678	0.000	0.000	3008.982	5509.751	12690115	3482604	7733458.000	7722096.000
6	275.236	1779.314	397.550	0.000	0.000	3008.982	5633.595	13401971	3697570	8219467.000	8199445.000
7	310.131	1900.587	465.205	0.000	0.000	3008.982	5847.903	13952791	3870093	8569175.000	8537421.000
8	338.933	2110.637	511.593	0.000	0.000	3008.982	6213.162	14373573	4008032	8916862.000	8871958.000
9	362.568	2305.394	541.660	0.000	0.000	3008.982	6569.699	14733655	4130344	9275541.000	9217591.000
10	400.556	2554.690	611.061	0.000	0.000	3008.982	6703.789	15137497	4266179	9667811.000	9595036.000
11	343.392	2247.032	517.493	0.000	0.000	3008.982	6122.510	15667474	4436282	9815442.000	9727541.000
12	297.226	1873.850	426.206	0.000	0.000	3008.982	8047.516	16318324	4648139	9895467.000	9793166.000
13	313.813	2151.399	592.561	0.000	0.000	3008.982	9044.798	17086910	4893013	10562202.000	10438055.000
14	320.903	2192.624	603.102	0.000	0.000	3008.982	9458.071	17909063	5155788	11397483.000	11247647.000
15	324.195	2222.926	576.037	0.000	0.000	3008.982	9743.980	18692107	5426872	11862726.000	11690825.000
16	339.439	2229.918	604.767	0.000	0.000	3008.982	10010.280	19378962	5691836	12302104.000	12109086.000
17	325.850	2149.666	626.243	0.000	0.000	3008.982	10085.164	19961972	5931478	12828447.000	12614497.000
18	315.075	2050.130	647.468	0.000	0.000	3008.982	10290.035	20464923	6151869	13067961.000	12835340.000
19	367.378	2343.206	715.934	0.000	0.000	3008.982	11078.068	20929119	6364912	13339006.000	13088192.000
20	383.555	2403.608	725.441	0.000	0.000	3008.982	11744.024	21415593	6588738	13827977.000	13557331.000
21	438.123	2737.493	849.335	0.000	0.000	3008.982	12880.621	21966187	6836980	14240377.000	13949295.000
22	449.471	2840.266	843.171	0.000	0.000	3008.982	13934.516	22594126	7114473	14780282.000	14468875.000
23	441.855	2798.736	836.909	0.000	0.000	3008.982	14289.259	23280663	7416295	15399105.000	15067373.000
24	432.003	2741.099	832.896	0.000	0.000	3008.982	14702.986	23997856	7733832	15946572.000	15594637.000
25	437.620	2770.467	842.504	0.000	0.000	3008.982	15137.680	24703798	8054222	16543889.000	16172321.000
26	423.504	2719.153	797.901	0.000	0.000	3008.982	15661.858	25368923	8367571	17071446.000	16682054.000
27	421.927	2692.957	793.938	0.000	0.000	3008.982	15997.366	25985093	8670939	17520861.000	17115346.000
28	422.402	2680.838	768.680	0.000	0.000	3008.982	15795.201	26558609	8971472	18028696.000	17614722.000
29	427.507	2716.908	789.336	0.000	0.000	3008.982	16063.965	27099874	9273302	18549862.000	18136922.000
30	387.950	2557.433	766.501	0.000	0.000	3008.982	16363.342	27626382	9582625	19090409.000	18679089.000
31	426.436	2743.595	785.332	0.000	0.000	3008.982	17041.403	28150604	9904337	19692301.000	19279930.000
32	196.644	383.307	474.695	0.000	0.000	320.880	3475.291	2086075	1195379	1676902.000	1618163.000
33	194.185	394.069	469.142	0.000	0.000	195.188	5680.136	2073348	1202083	1675168.000	1626914.000
34	187.375	400.054	453.218	0.000	0.000	163.133	5221.132	2033508	1207079	1670550.000	1633189.000
35	198.809	426.417	481.384	0.000	0.000	126.849	5316.973	1983942	1205641	1663515.000	1637195.000
36	260.301	546.010	620.299	0.000	0.000	94.954	5564.288	1936056	1204539	1654796.000	1639203.000
37	267.215	546.743	639.653	0.000	0.000	159.286	5582.371	1897862	1208874	1644902.000	1639462.000
38	258.445	463.034	634.935	0.000	0.000	152.943	3231.957	1871645	1220583	1633759.000	1637572.000
39	242.095	433.554	586.431	0.000	0.000	114.769	3035.164	1854740	1238301	1620795.000	1632924.000
40	230.314	411.251	558.918	0.000	0.000	142.219	3033.239	1843604	1260155	1606618.000	1625558.000
41	235.616	427.838	570.972	0.000	0.000	255.629	3336.639	1832914	1282662	1592223.000	1616037.000
42	242.061	428.117	589.822	0.000	0.000	248.065	3516.551	1818833	1303137	1577591.000	1604429.000

3.5 Format the data as required

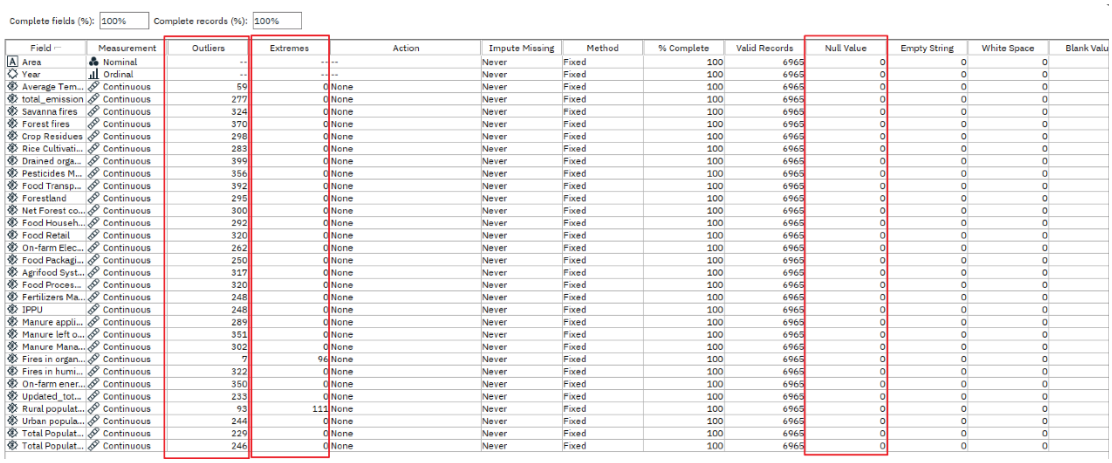
Because the original dataset is selected for partial outer join, the cleaned missing values, outliers, and extremes of the original attributes are not changed after merging a new dataset (Figure 20).

Figure 20. Selected original dataset for partial outer join



For the newly added attributes, ‘Rural population’, ‘Urban population’, ‘Total Population – Male’, and “Total Population – Female’, the data values are cleaned with the same methods as the previous steps. The mean imputes the missing values, and all outliers and extremes are coerced. Figure 21 shows the result table after the data cleaning.

Figure 21. Result table after data cleaning

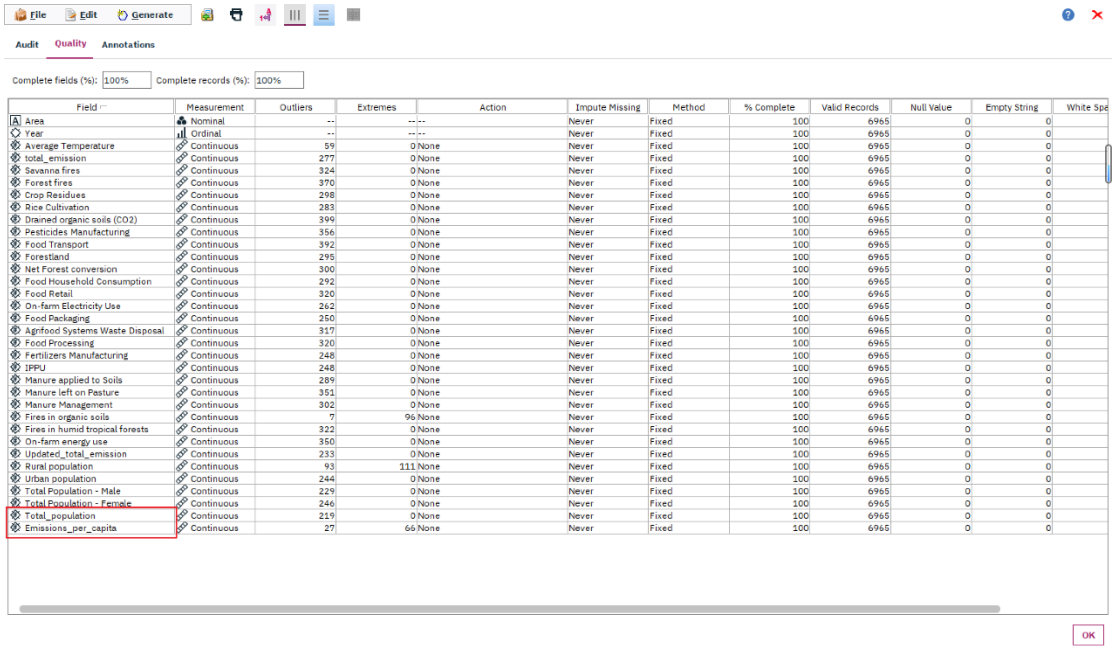


Complete fields (%): 100% Complete records (%): 100%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Valu
Area	Nominal	--	--	--	Never	Fixed	100	6965	0	0	0	
Year	Ordinal	--	--	--	Never	Fixed	100	6965	0	0	0	
Average Tem...	Continuous	59	0	None	Never	Fixed	100	6965	0	0	0	
total_emission	Continuous	277	0	None	Never	Fixed	100	6965	0	0	0	
Savanna fires	Continuous	324	0	None	Never	Fixed	100	6965	0	0	0	
Forest fires	Continuous	370	0	None	Never	Fixed	100	6965	0	0	0	
Crop Residues	Continuous	298	0	None	Never	Fixed	100	6965	0	0	0	
Rice Cultivati...	Continuous	283	0	None	Never	Fixed	100	6965	0	0	0	
Drained orga...	Continuous	399	0	None	Never	Fixed	100	6965	0	0	0	
Pesticides M...	Continuous	356	0	None	Never	Fixed	100	6965	0	0	0	
Food Transp...	Continuous	392	0	None	Never	Fixed	100	6965	0	0	0	
Forestland	Continuous	295	0	None	Never	Fixed	100	6965	0	0	0	
Net Forest co...	Continuous	300	0	None	Never	Fixed	100	6965	0	0	0	
Food Househ...	Continuous	292	0	None	Never	Fixed	100	6965	0	0	0	
Food Retail	Continuous	320	0	None	Never	Fixed	100	6965	0	0	0	
On-farm Elec...	Continuous	262	0	None	Never	Fixed	100	6965	0	0	0	
Food Packagi...	Continuous	250	0	None	Never	Fixed	100	6965	0	0	0	
Agrifood Syst...	Continuous	317	0	None	Never	Fixed	100	6965	0	0	0	
Food Proces...	Continuous	320	0	None	Never	Fixed	100	6965	0	0	0	
Fertilizers Ma...	Continuous	248	0	None	Never	Fixed	100	6965	0	0	0	
IPPU	Continuous	248	0	None	Never	Fixed	100	6965	0	0	0	
Manure appli...	Continuous	289	0	None	Never	Fixed	100	6965	0	0	0	
Manure left o...	Continuous	351	0	None	Never	Fixed	100	6965	0	0	0	
Manure Mana...	Continuous	302	0	None	Never	Fixed	100	6965	0	0	0	
Fires in organ...	Continuous	7	96	None	Never	Fixed	100	6965	0	0	0	
Fires in humi...	Continuous	322	0	None	Never	Fixed	100	6965	0	0	0	
On-farm ener...	Continuous	350	0	None	Never	Fixed	100	6965	0	0	0	
Updated_tot...	Continuous	233	0	None	Never	Fixed	100	6965	0	0	0	
Rural populat...	Continuous	93	111	None	Never	Fixed	100	6965	0	0	0	
Urban popula...	Continuous	244	0	None	Never	Fixed	100	6965	0	0	0	
Total Populat...	Continuous	229	0	None	Never	Fixed	100	6965	0	0	0	
Total Populat...	Continuous	246	0	None	Never	Fixed	100	6965	0	0	0	

Other than that, considering the third data mining objective, two new features are constructed to identify the countries with the highest average temperature increase by year and analyse their contributions to the overall environmental impact (Figure 22). The first is ‘Total_population’, which is the summation of ‘Rural population’ and ‘Urban population’. The second is ‘Emissions_per_capita’, whose value is ‘Updated_total_emission’ divided by ‘Total_population.

Figure 22. New features – 'Total_population' and 'Updated_total_populati



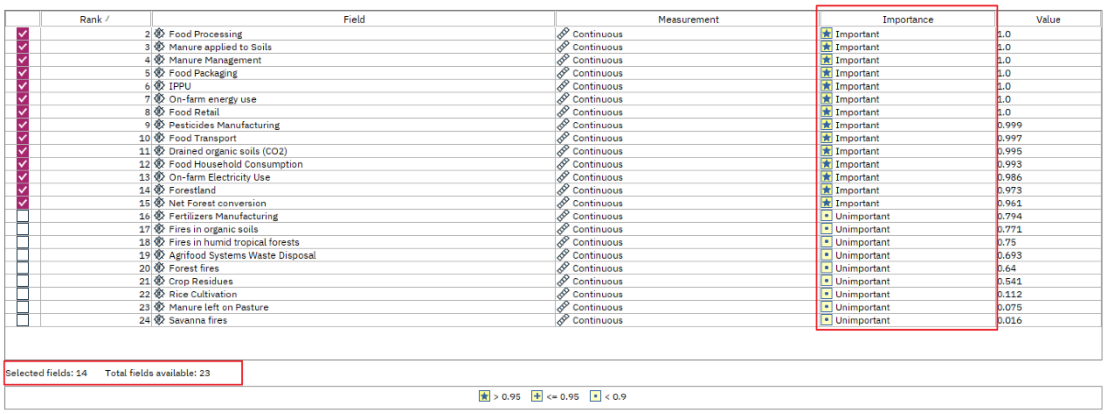
Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Spe
Area	Nominal	--	--	--	Never	Fixed	100	6965	0	0	
Year	Ordinal	--	--	--	Never	Fixed	100	6965	0	0	
Average Temperature	Continuous	59	0	None	Never	Fixed	100	6965	0	0	
total_emission	Continuous	277	0	None	Never	Fixed	100	6965	0	0	
Savanna fires	Continuous	324	0	None	Never	Fixed	100	6965	0	0	
Forest fires	Continuous	370	0	None	Never	Fixed	100	6965	0	0	
Crop Residues	Continuous	298	0	None	Never	Fixed	100	6965	0	0	
Rice Cultivation	Continuous	283	0	None	Never	Fixed	100	6965	0	0	
Drained organic soils (CO2)	Continuous	399	0	None	Never	Fixed	100	6965	0	0	
Pesticides Manufacturing	Continuous	356	0	None	Never	Fixed	100	6965	0	0	
Food Transport	Continuous	392	0	None	Never	Fixed	100	6965	0	0	
Forestland	Continuous	295	0	None	Never	Fixed	100	6965	0	0	
Net Forest conversion	Continuous	300	0	None	Never	Fixed	100	6965	0	0	
Food Household Consumption	Continuous	292	0	None	Never	Fixed	100	6965	0	0	
Food Retail	Continuous	320	0	None	Never	Fixed	100	6965	0	0	
On-farm Electricity Use	Continuous	262	0	None	Never	Fixed	100	6965	0	0	
Food Packaging	Continuous	250	0	None	Never	Fixed	100	6965	0	0	
Agrifood Systems Waste Disposal	Continuous	317	0	None	Never	Fixed	100	6965	0	0	
Food Processing	Continuous	320	0	None	Never	Fixed	100	6965	0	0	
Fertilizers Manufacturing	Continuous	248	0	None	Never	Fixed	100	6965	0	0	
IPPU	Continuous	248	0	None	Never	Fixed	100	6965	0	0	
Manure applied to Soils	Continuous	289	0	None	Never	Fixed	100	6965	0	0	
Manure left on Pasture	Continuous	351	0	None	Never	Fixed	100	6965	0	0	
Manure Management	Continuous	302	0	None	Never	Fixed	100	6965	0	0	
Fires in organic soils	Continuous	7	96	None	Never	Fixed	100	6965	0	0	
Fires in humid tropical forests	Continuous	322	0	None	Never	Fixed	100	6965	0	0	
On-farm energy use	Continuous	350	0	None	Never	Fixed	100	6965	0	0	
Updated_total_emission	Continuous	233	0	None	Never	Fixed	100	6965	0	0	
Rural population	Continuous	93	111	None	Never	Fixed	100	6965	0	0	
Urban population	Continuous	244	0	None	Never	Fixed	100	6965	0	0	
Total Population - Male	Continuous	229	0	None	Never	Fixed	100	6965	0	0	
Total Population - Female	Continuous	246	0	None	Never	Fixed	100	6965	0	0	
Total_population	Continuous	219	0	None	Never	Fixed	100	6965	0	0	
Emissions_per_capita	Continuous	27	66	None	Never	Fixed	100	6965	0	0	

4. Data transformation

4.1 Reduce the data

The data is reduced by selecting relevant features. 'Average Temperature' is the target feature, and 'Savanna fires', 'Forest fires', 'Crop Residues', 'Rice Cultivation', 'Drained organic soils (CO2)', 'Pesticides Manufacturing', 'Food Transport', 'Forestland', 'Net Forest conversion', 'Food Household Consumption', 'Food Retail', 'On-farm Electricity Use', 'Food Packaging', 'Agrifood Systems Waste Disposal', 'Food Processing', 'Fertilizers Manufacturing', 'IPPU', 'Manure applied to Soils', 'Manure left on Pasture', 'Manure Management', 'Fires in organic soils', 'Fires in humid tropical forests', and 'On-farm energy use' are the inputted features. Before the data transformation, 14 out of 23 features are selected as 'Important' (Figure 23).

Figure 23. 'Important' features before data transformation



Rank	Field	Measurement	Importance	Value
2	Food Processing	Continuous	Important	1.0
3	Manure applied to Soils	Continuous	Important	1.0
4	Manure Management	Continuous	Important	1.0
5	Food Packaging	Continuous	Important	1.0
6	IPPU	Continuous	Important	1.0
7	On-farm energy use	Continuous	Important	1.0
8	Food Retail	Continuous	Important	1.0
9	Pesticides Manufacturing	Continuous	Important	0.999
10	Food Transport	Continuous	Important	0.997
11	Drained organic soils (CO2)	Continuous	Important	0.995
12	Food Household Consumption	Continuous	Important	0.993
13	On-farm Electricity Use	Continuous	Important	0.986
14	Forestland	Continuous	Important	0.973
15	Net Forest conversion	Continuous	Important	0.961
16	Fertilizers Manufacturing	Continuous	Unimportant	0.794
17	Fires in organic soils	Continuous	Unimportant	0.771
18	Fires in humid tropical forests	Continuous	Unimportant	0.75
19	Agrifood Systems Waste Disposal	Continuous	Unimportant	0.693
20	Forest fires	Continuous	Unimportant	0.64
21	Crop Residues	Continuous	Unimportant	0.541
22	Rice Cultivation	Continuous	Unimportant	0.112
23	Manure left on Pasture	Continuous	Unimportant	0.075
24	Savanna fires	Continuous	Unimportant	0.016

Selected fields: 14 Total fields available: 23

> 0.95 <= 0.95 < 0.9

The skewness and kurtosis of the target feature, 'Average Temperature', are 0.329 and 1.032, respectively. The distribution of it approaches normal distribution (Figure 24). If the data is balanced by boosting and reducing, the distribution will not be close to normal distribution (Figure 25 & Figure 26). Thus, in this project, the target feature data keeps the original data value, not using boosting or reducing.

Figure 24. Distribution of Average Temperature

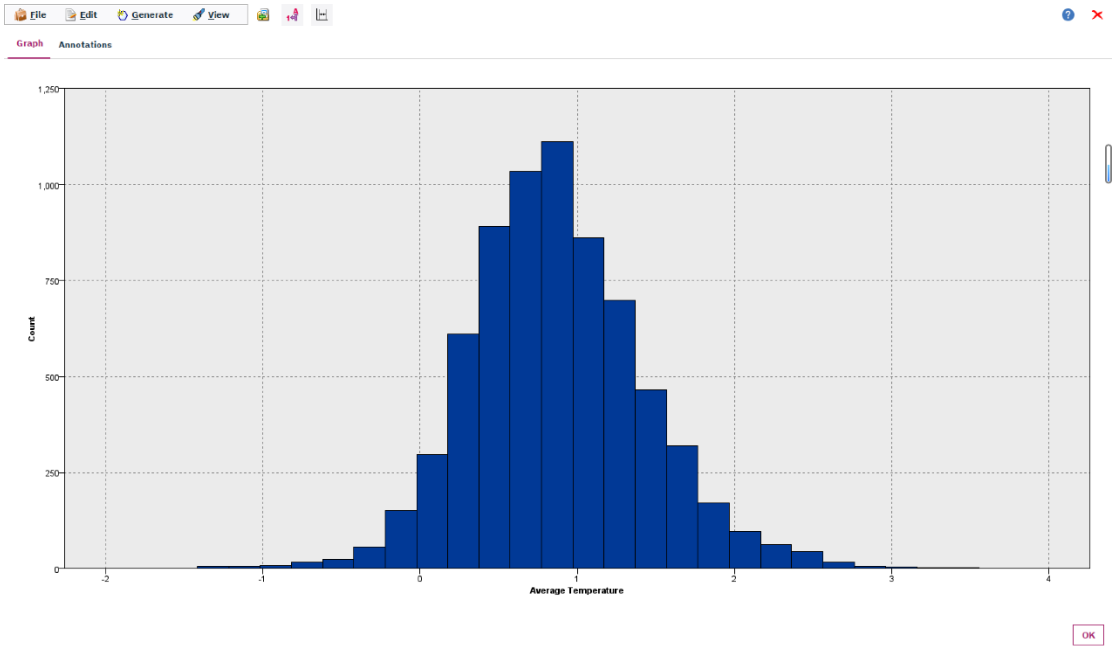


Figure 25. Distribution of Average Temperature after boosting

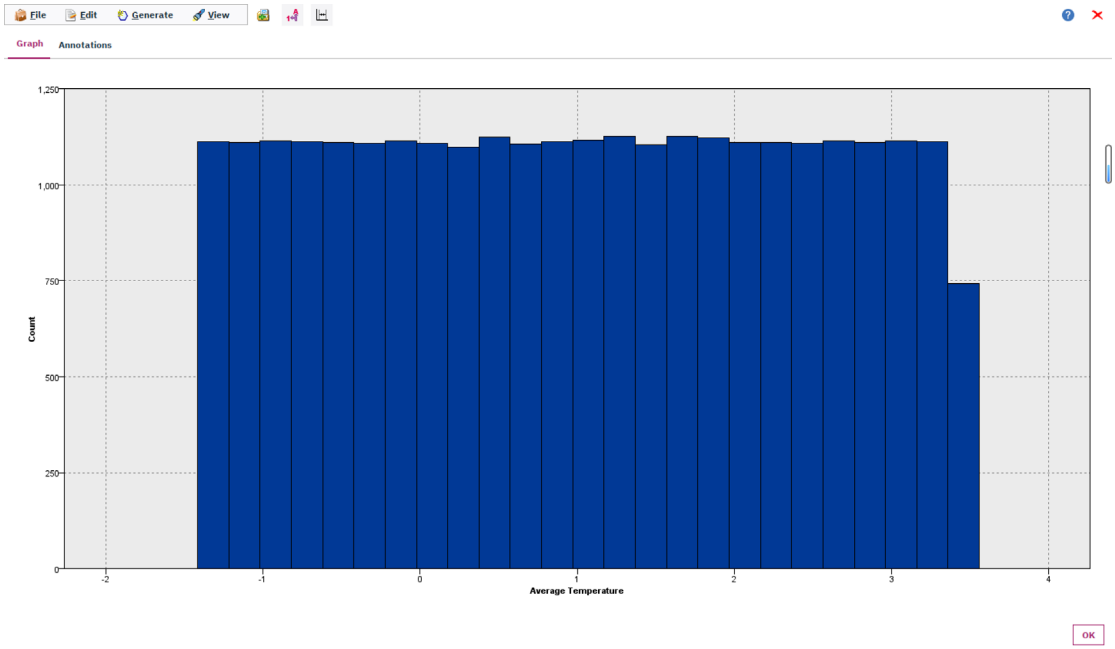
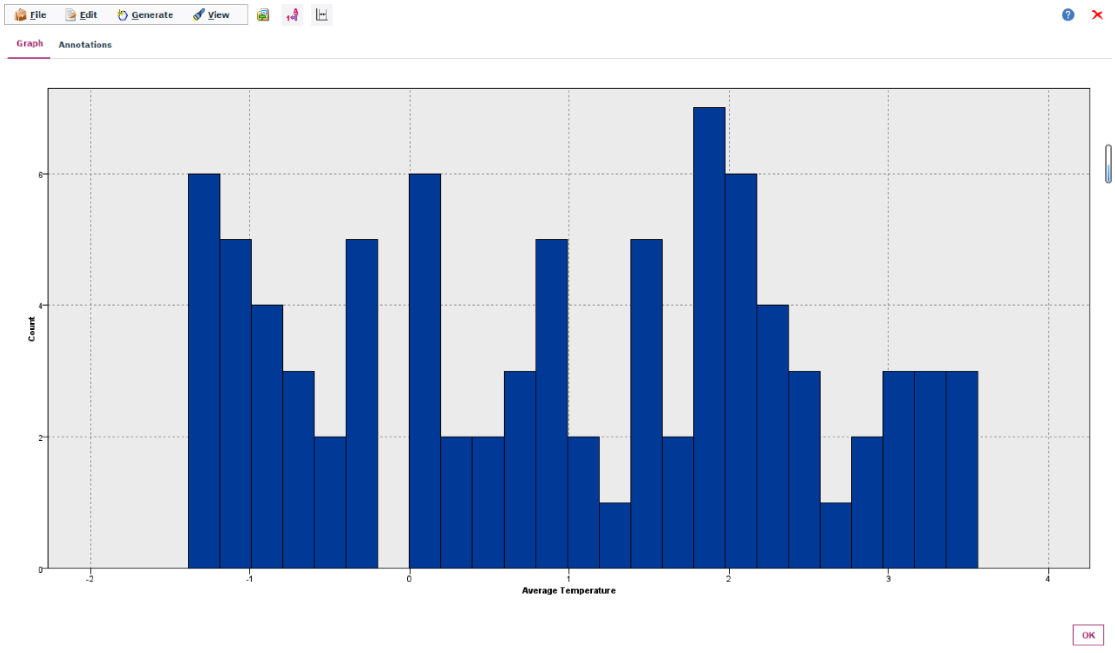


Figure 26. Distribution of Average Temperature after reducing



Other than that, 'Feature Selection' modelling is used to evaluate the importance of all 23 agri-food features to the target feature, 'Average Temperature'. Figure 27 shows the model parameters, and Figure 28 shows the selection result. The importance of 'Fertilizers Manufacturing' and 'Crop Residues' is marginal, so these features are filtered (Figure 29).

Figure 27. Feature selection model parameters

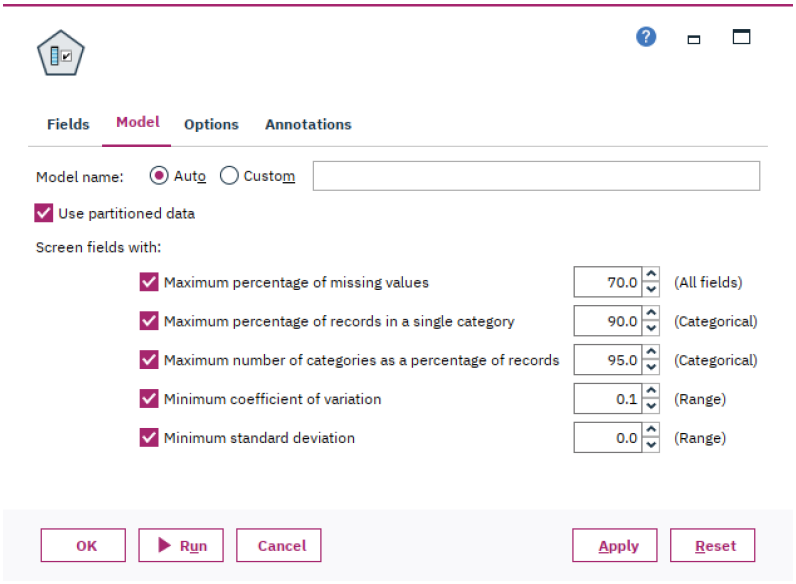


Figure 28. Feature selection result

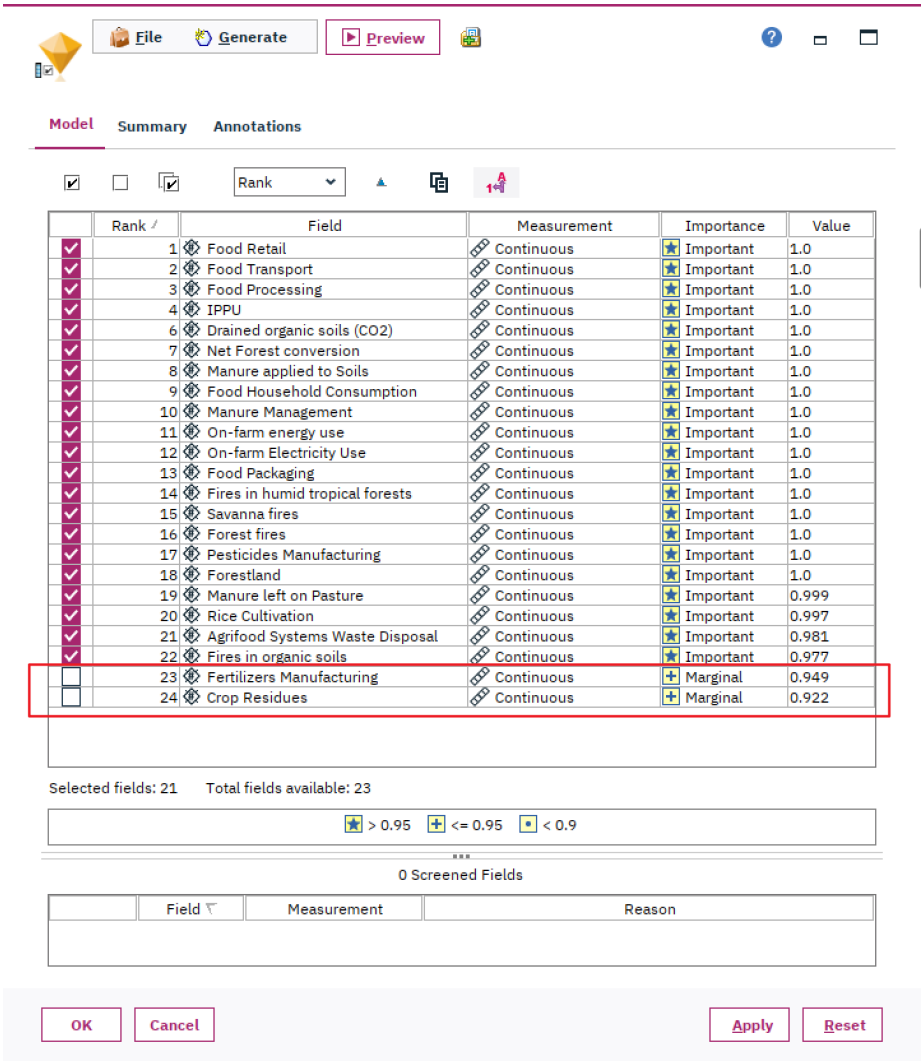
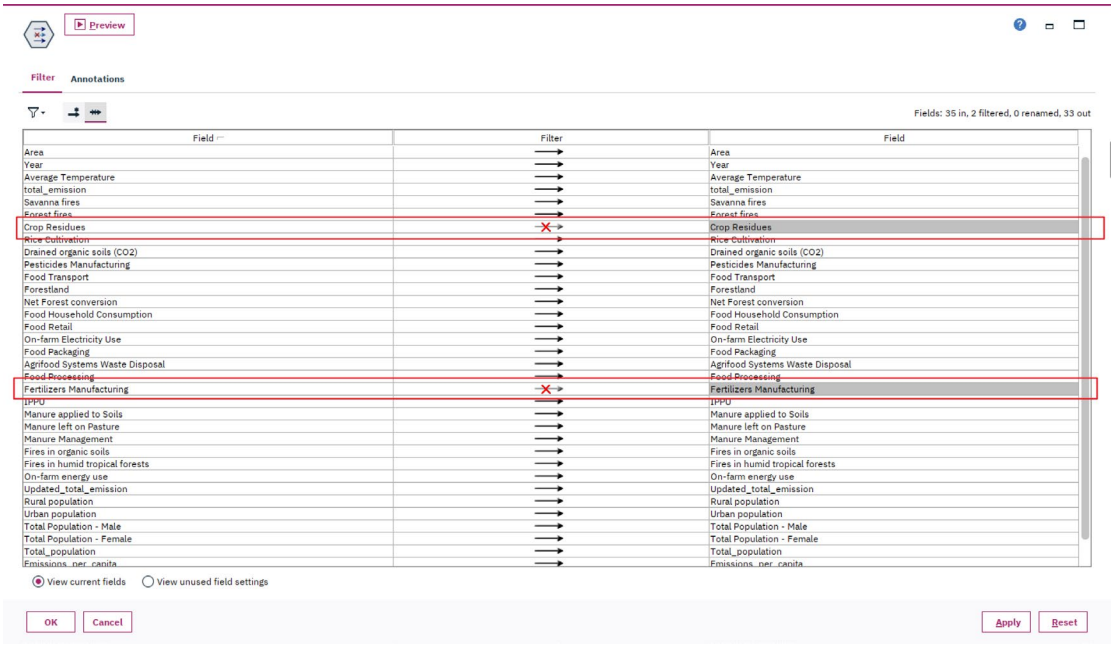


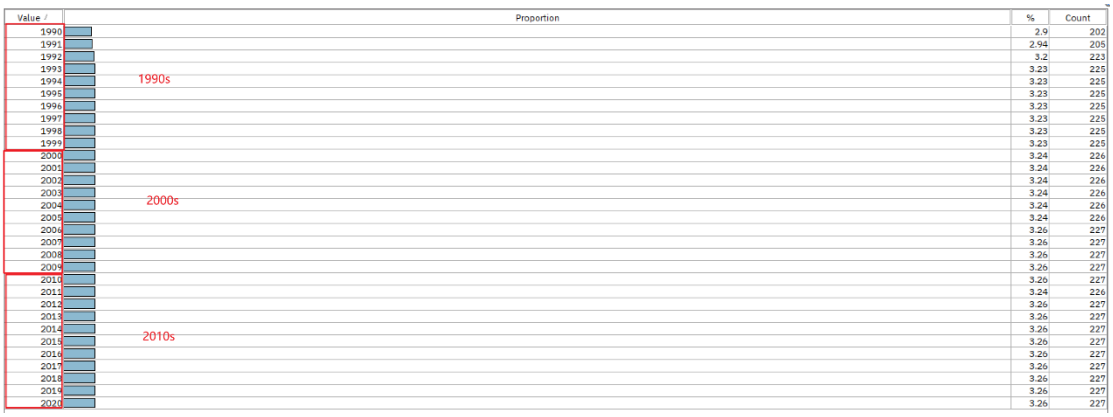
Figure 29. Feature filter



4.2 Project the data

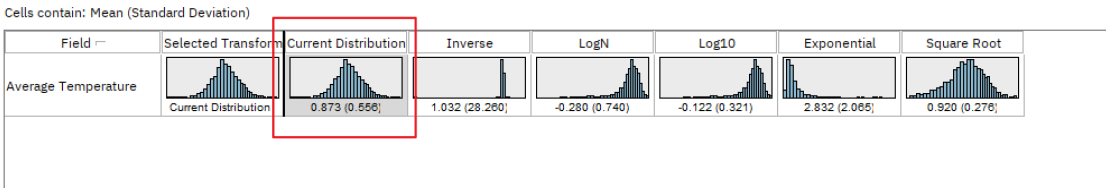
The period of the whole dataset is from 1990 to 2020. The years from 1990 to 1999 are grouped as '1990s', the years from 2000 to 2009 are grouped as '2000s', and the years from 2010 to 2020 are grouped as '2010s' (Figure 30).

Figure 30. Grouped years



After trying to transform the data value of 'Average Temperature' in several ways, including inverse, logN, log10, exponential, and square root, the distribution of the original data value is the most closed to normal distribution (Figure 31). Therefore, transforming the target feature data value is not executed, and the original data value is used to process the subsequent steps.

Figure 31. Transforming 'Average Temperature' data value



5. Data mining methods selection

5.1 Match and discuss the objectives of data mining to data mining methods

Three different data mining methods are discussed and matched to three data mining objectives, respectively.

The first data mining objective, to examine the correlation between carbon dioxide (CO2) emissions within the agri-food sector and the subsequent temperature rise, involves studying the correlation between two variables. Correlation analysis or regression analysis can be used to examine the relationship between carbon dioxide emissions and temperature rise. Correlation analysis is a statistical technique employed to assess the magnitude and direction of the association between two or more variables (Dean, n.d.). The process entails the computation of a correlation coefficient, a quantitative measure that assesses the extent of the relationship between the variables. The correlation coefficient is a statistical measure that varies between -1 and 1. When the coefficient is close to -1, it suggests a robust negative relationship. Conversely, a coefficient near 1 indicates a strong positive relationship. On the other hand, a coefficient close to 0 signifies the absence of any relationship. Regression analysis is a statistical technique employed to establish a mathematical model that describes the association between a dependent variable and one or more independent variables (Dean, n.d.). The process entails using a line or curve to establish a relationship with the data, thereby enabling the generation of predictions regarding the dependent variable by considering the values of the independent variables. Regression analysis encompasses various techniques, such as linear regression, multiple linear regression, and nonlinear regression, employed to model relationships between variables. Therefore, these methods can determine the strength and direction of the relationship between these two variables.

The second data mining objective, to analyse the influence of various countries based on aggregated data on emissions and temperature change, involves assessing how different countries' emissions impact temperature change. Clustering or segmentation techniques can group countries based on their emissions and temperature change data. Clustering is a method

used to locate different subgroups within a more enormous collection (Dean, n.d.). When analysts divide the data into subgroups, often referred to as clusters, their goal is to distribute the data so that the cases within a group are pretty like one another, while the cases in other clusters are incredibly distinct from one another. On the other hand, segmentation refers to categorising consumers or other things into different groups based on the commonalities they share. When it comes to grouping and segmentation, there are a wide variety of algorithmic and methodological options. Examples of popular approaches are clustering techniques such as k-means, hierarchical clustering, and decision trees (Dean, n.d.). Using these methods, the data can be automatically segmented based on criteria, such as the degree of similarity or distance between two points in the data. These methods can identify patterns and trends in the data and understand how different countries contribute to emissions and temperature change.

The third data mining objective, to identify the countries with the highest average temperature increase by year and analyse their contributions to the overall environmental impact, involves finding countries with the highest average temperature increase by year and understanding how their emissions contribute to the overall environmental impact. Descriptive statistics or ranking methods can identify the countries with the highest average temperature increase by year (Marr, n.d.). Once these countries are identified, regression or decision tree analysis can be used to understand their contribution to the overall environmental impact. The process of clustering is one method that can be used to determine the existence of subgroups within a more extensive set. In dividing the data into subgroups, often referred to as clusters, analysts intend to distribute the data so that the cases within a group are incredibly like one another. However, the cases in other clusters are incredibly dissimilar to one another. The process of classifying consumers or other things into subcategories according to the shared characteristics of those subcategories is known as segmentation. When it comes to clustering and segmentation, there are a wide variety of options in terms of algorithms and methods. Common approaches include clustering techniques such as k-means, hierarchical clustering, and decision trees (Marr, n.d.). Using these methods, the data can be automatically segmented depending on criteria, such as similarities between the segments or distances. A decision tree is a type of decision support tool that uses a tree-like model of decisions and the probable repercussions of those actions. These potential implications include the outcomes of random events, the costs of resources, and the utility of those resources. Displaying an algorithm that consists solely of conditional control statements can be done in this manner. Decision trees are a prominent tool in machine learning, in addition to their widespread application in operations research, specifically in decision analysis. These trees are used to determine which approach is

most likely to achieve a given objective.

5.2 Select the appropriate data mining methods based on discussion

- Examine the correlation between carbon dioxide (CO2) emissions within the agri-food sector and the subsequent temperature rise: regression analysis is utilised to investigate the association between carbon dioxide (CO2) emissions and the increase in temperature. This methodology facilitates the assessment of the magnitude and orientation of the association between the variables mentioned above.
- Analyse the influence of various countries based on aggregated data on emissions and temperature change: clustering techniques are utilised to categorise countries according to their emissions and temperature change data. This approach enables the identification of patterns and trends within the dataset, facilitating a comprehensive comprehension of the various countries' contributions to emissions and temperature fluctuations.
- Identify the countries with the highest average temperature increase by year and analyse their contributions to the overall environmental impact: descriptive statistics is utilised to ascertain the nations exhibiting the most significant average temperature increase by year. After identifying these countries, decision tree analysis is used to gain insights into their contributions to the overall environmental impact.

6. Data mining algorithms selection

6.1 Conduct exploratory analysis and discuss

The first data mining objective is to examine the correlation between carbon dioxide (CO2) emissions within the agri-food sector and the subsequent temperature rise, for which regression analysis is used. The second data mining goal is to analyse the influence of various countries based on aggregated data on emissions and temperature change, for which clustering is utilised. The third data mining objective is to identify the countries with the highest average temperature increase by year and analyse their contributions to the overall environmental impact, for which descriptive statistics is used. Descriptive statistics is not a typical data mining method, and the results of the third goal are presented directly in the eighth step. This step discusses the regression analysis for the first objective and the clustering for the second goal.

6.1.1 Regression

Based on the first objective, the specific regression algorithm is selected using 'Auto Numeric' modelling. 'Average Temperature' is the target value. The inputs contain 'Savanna fires', 'Forest fires', 'Rice Cultivation', 'Drained organic soils (CO2), 'Pesticides Manufacturing', 'Food Transport', 'Forestland', 'Net Forest conversion', 'Food Household Consumption', 'Food Retail', 'On-farm Electricity Use', 'Food Packaging', 'Agrifood Systems Waste Disposal', 'Food Processing', 'IPPU', 'Manure applied to Soils', Manure left on Pasture', 'Manure Management', 'Fires in organic soils', 'Fires in humid tropical forests', and 'On-farm energy (Figure 32). Based on the dataset, there are 17 possible regression algorithms, including general regression, generalized linear, generalized linear engine, KNN algorithm, linear-AS, random trees, SVM, tree-AS, XGBoost linear, XGBoost tree, linear regression, CHAID, C&R tree, XGBoost-AS, random forest, and neural net (Figure 33).

Figure 32. Target and inputs

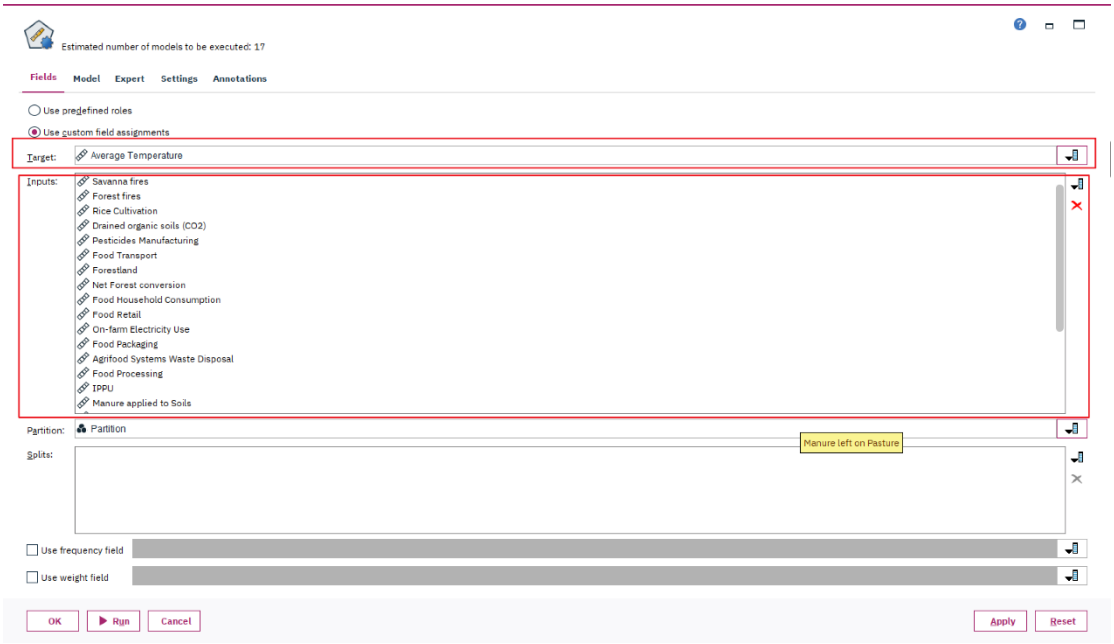
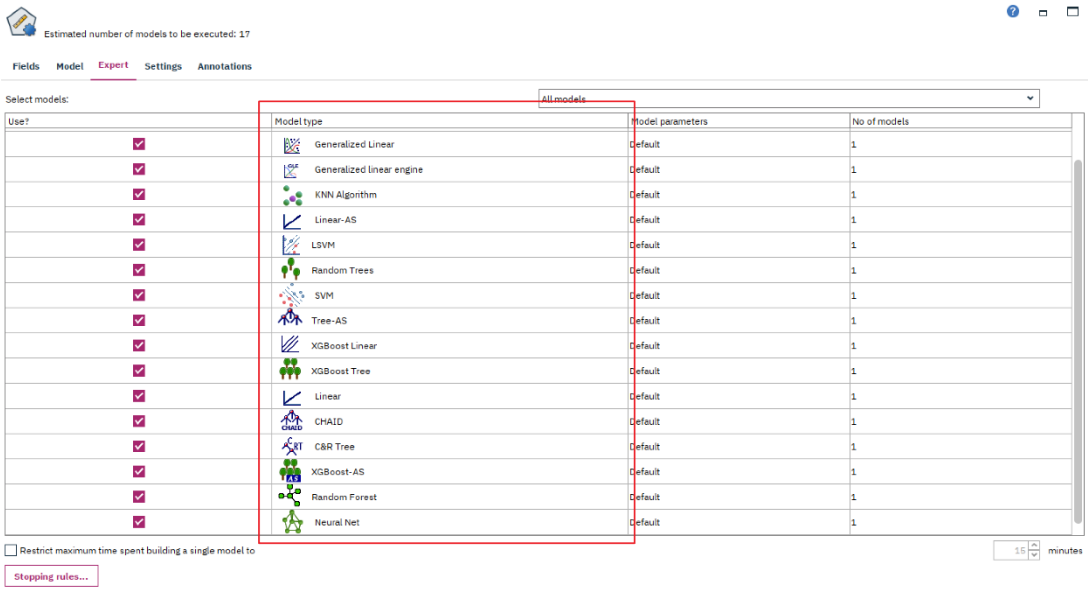


Figure 33. Possible regression algorithms



After running the auto numeric modelling, the result shows that random forest is the optimal algorithm. For the training set, the correlation value is the highest, 0.952, and the relative error value is the lowest, 0.105, compared to other algorithms (Figure 34). For the testing set, the correlation value is 0.667, and the relative error value is 0.557 (Figure 35).

Figure 34. Training set with random forest

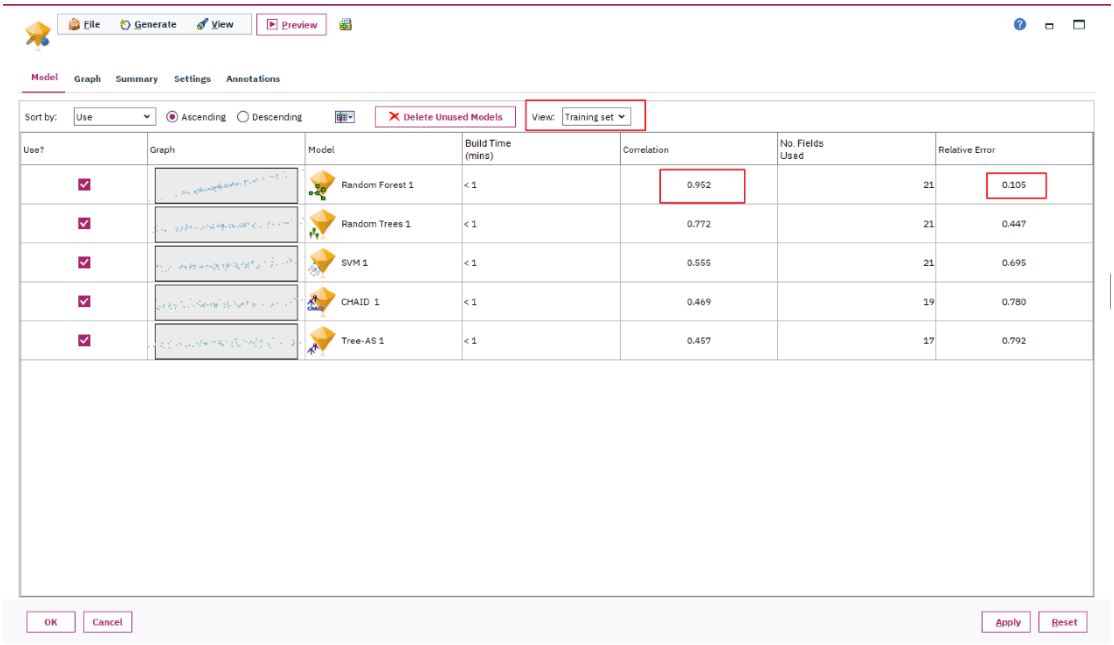
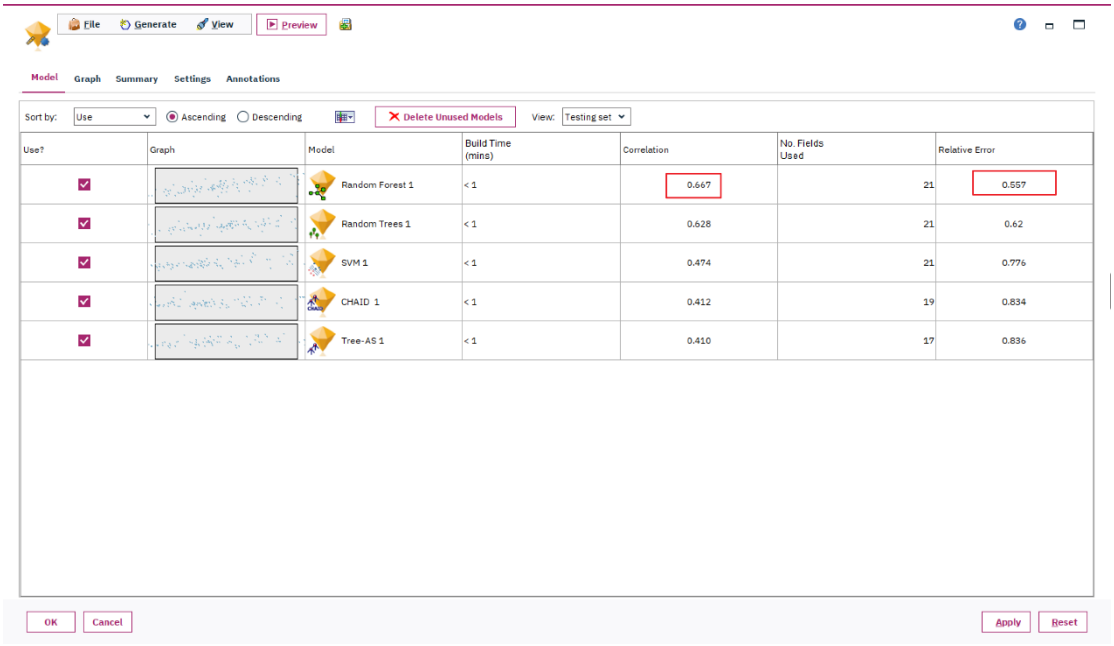


Figure 35. Testing set with random forest



6.1.2 Clustering

Based on the second goal, the specific clustering algorithm is selected using 'Auto Cluster'. 'Average Temperature' is the evaluation value, and the inputs contain 'Area', 'Updated_total_emission', 'Total_population' and 'Average Temperature' (Figure 36). Based on the dataset, there are three possible clustering algorithms, including Kohonen, K-means, and TwoStep (Figure 37).

Figure 36. Evaluation and inputs

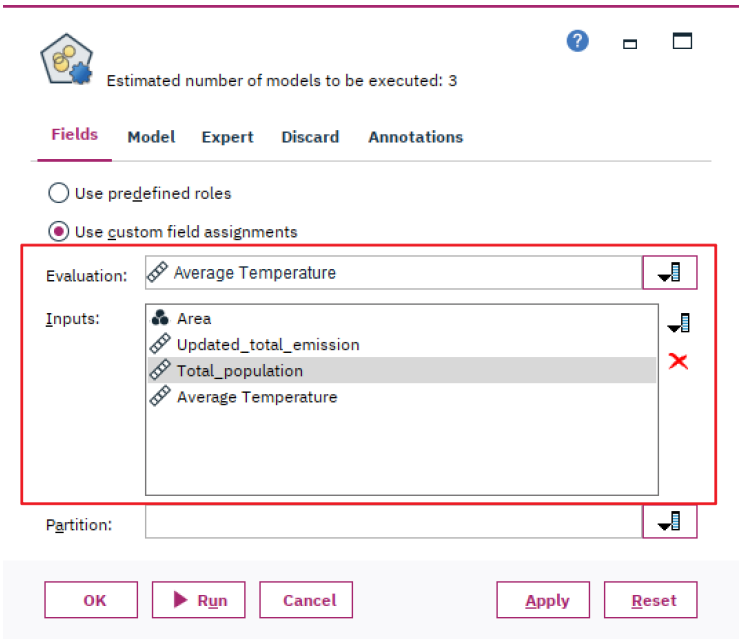
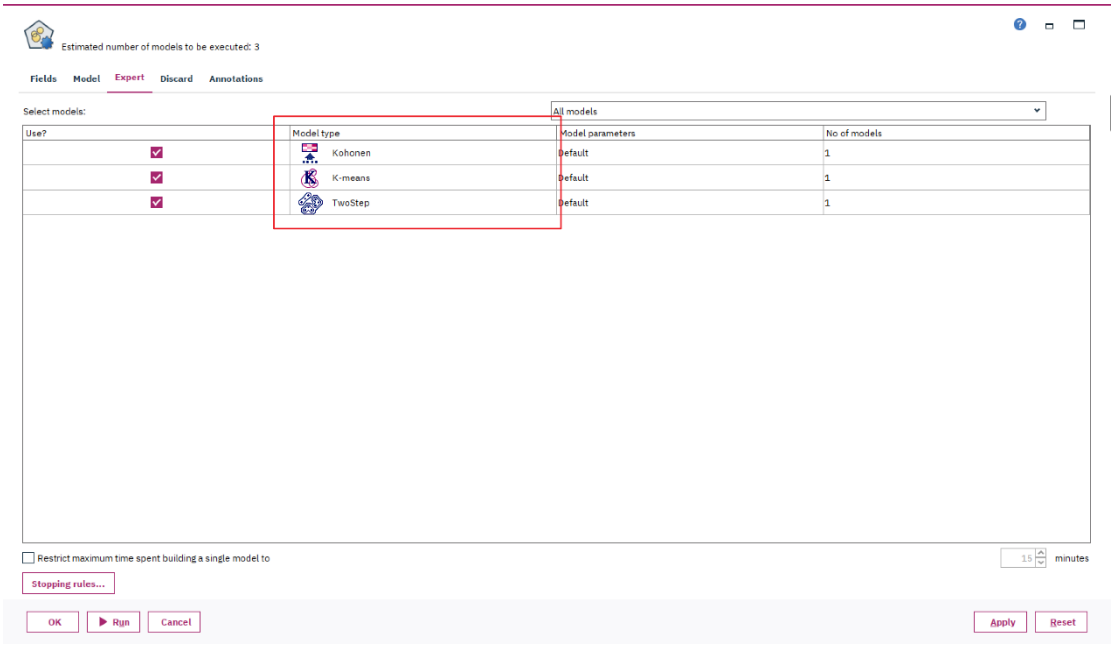
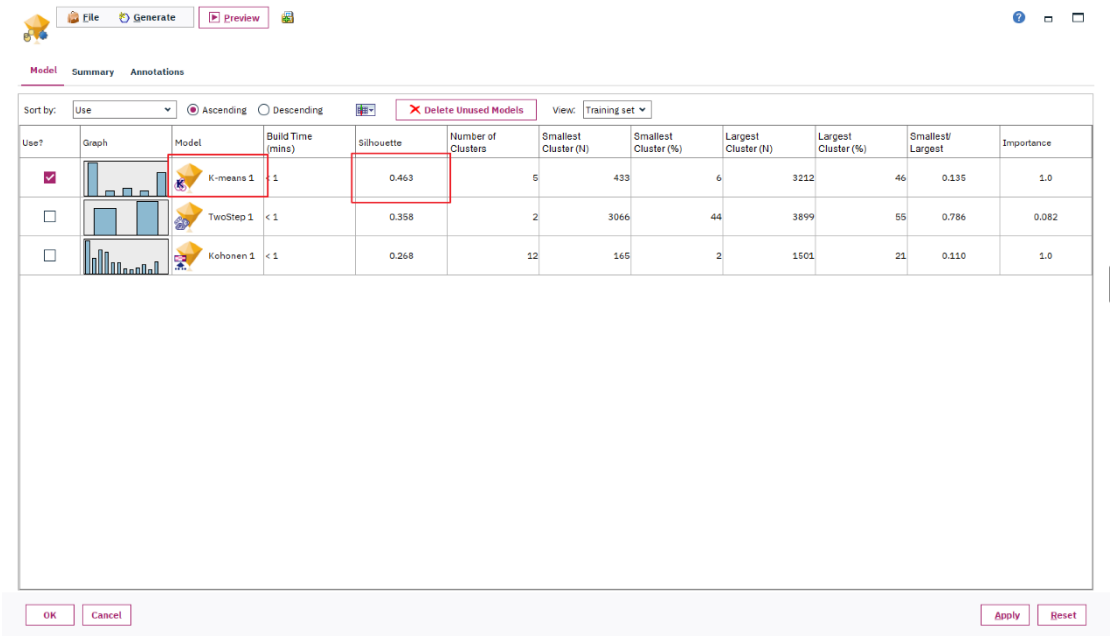


Figure 37. Possible clustering algorithms



After running the auto cluster modelling, the generated result shows that K-means is the optimal algorithm. The silhouette value is the highest, 0.463 (Figure 38).

Figure 38. Optimal algorithm



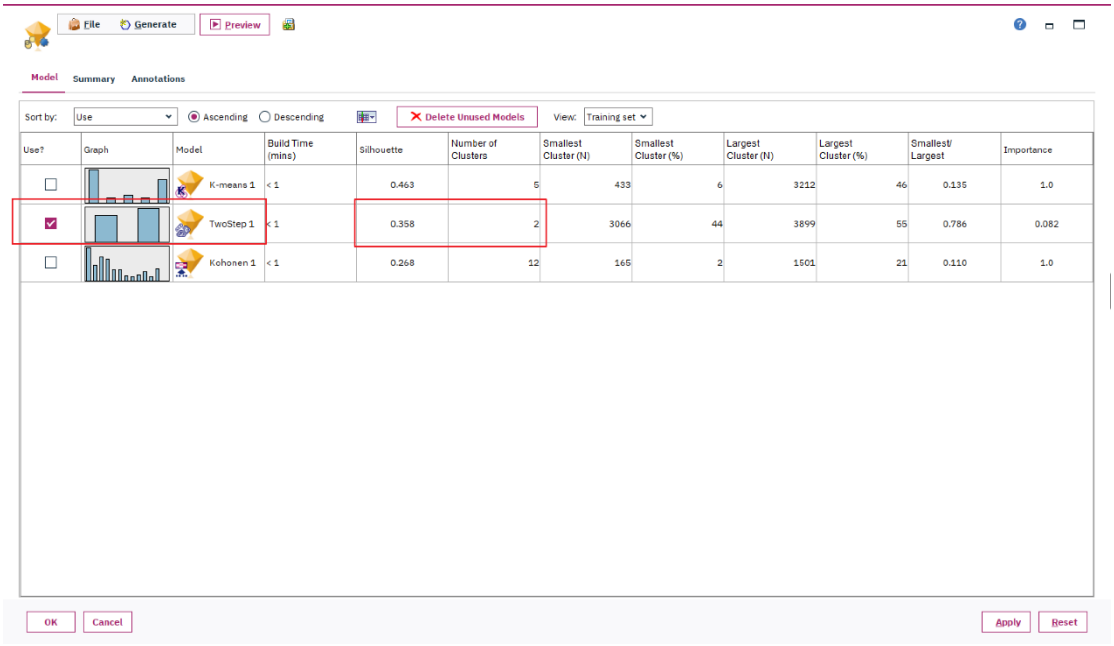
6.2 Select data mining algorithms based on discussion

The random forest is the optimal regression algorithm and selected for the first data mining objective to examine the correlation between carbon dioxide (CO2) emissions within the agri-

food sector and the subsequent temperature rise,. The random forest algorithm is classified as an ensemble learning technique that combines multiple decision trees to enhance the precision and resilience of predictive models (Dean, n.d.). The algorithm in question is widely recognised in machine learning and can perform classification and regression tasks. The algorithm generates numerous decision trees during the training phase. It establishes the class that manifests itself most frequently among the trees (for classification) or the average prediction made by the various trees (for regression). Random forests are renowned for their adeptness in managing extensive datasets, feature spaces with numerous dimensions, and intricate interdependencies among features. Moreover, these tools are known for their user-friendly nature and straightforward interpretation, rendering them highly favoured across various domains. Additionally, this project also uses a linear algorithm for the first data mining objective. Linear regression is a supervised machine learning algorithm to determine the linear association between a dependent variable and one or more independent features. The dependent variable in this study is the average temperature rise, while the independent variables consist of the different factors associated with CO2 emissions within the agri-food sector. Linear regression is a statistical technique that interprets how alterations in the independent variables correspond to variations in the dependent variable. The linear regression algorithm can determine the optimal linear equation for predicting the value of the dependent variable, taking into account the independent variables. The coefficients associated with the independent variables in the equation provide insights into the extent to which each variable contributes to the predictive model of the dependent variable.

The K-means is the optimal clustering algorithm for the second data mining goal, to analyse the influence of various countries based on aggregated data on emissions and temperature change. However, if the K-means algorithm is used, the number of clusters is 5, and the smallest cluster is only 5% of the whole dataset, which is meaningless. In comparison, when TwoStep is utilized, the number of clusters is 2, and the smallest cluster is 37%, which is more rational (Figure 39).

Figure 39. K-means VS TwoStep



Hence, the TwoStep clustering algorithm is selected. The TwoStep clustering algorithm is an unsupervised learning technique employed to identify inherent groupings within a given dataset, also known as clusters (*IBM Documentation*, 2021). The algorithm possesses several advantages in comparison to conventional clustering techniques. Firstly, it can generate clusters by considering both categorical and continuous variables. Secondly, it autonomously determines the appropriate number of clusters, thereby eliminating the need for manual intervention. Lastly, it demonstrates efficiency in analysing extensive data files. The TwoStep clustering technique is a method that involves two distinct steps for clustering. The initial step entails conducting a singular scan of the data, wherein the raw input data is compressed into a set of subclusters that can be more easily managed. The subsequent step employs a hierarchical clustering technique to progressively combine the subclusters into increasingly larger clusters without re-examining the data. One advantage of hierarchical clustering is its ability to eliminate the need for preselecting the number of clusters. Numerous hierarchical clustering techniques commence by considering individual records as initial clusters, subsequently merging them iteratively to generate progressively larger clusters. While the effectiveness of this approach diminishes when confronted with substantial volumes of data, the preliminary clustering performed by TwoStep enables hierarchical clustering to execute swiftly, even when handling extensive datasets (*IBM Documentation*, 2021).

6.3 Build/Select appropriate models and choose relevant parameters

6.3.1 Regression

For the first data mining objective, the random forest algorithm model and the linear regression model are built, 'Average Temperature' is the target, and 'Savanna fires', 'Forest fires', 'Rice Cultivation', 'Drained organic soils (CO2), 'Pesticides Manufacturing', 'Food Transport', 'Forestland', 'Net Forest conversion', 'Food Household Consumption', 'Food Retail', 'On-farm Electricity Use', 'Food Packaging', 'Agrifood Systems Waste Disposal', 'Food Processing', 'IPPU', 'Manure applied to Soils', Manure left on Pasture', 'Manure Management', 'Fires in organic soils', 'Fires in humid tropical forests', and 'On-farm energy use' are the predictors (Figure 40 & Figure 41).

Figure 40. Building random forest model

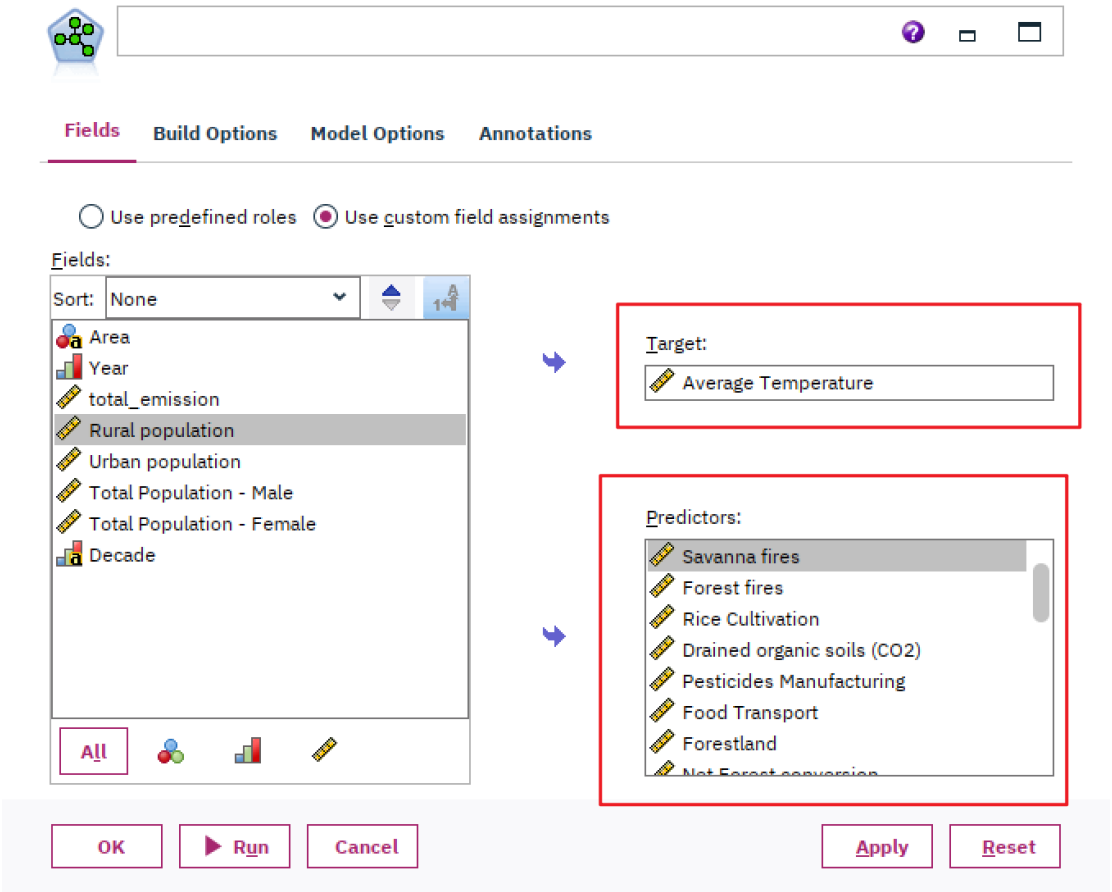


Figure 41. Building linear regression model

The relevant parameters of the random forest model contain the number of trees, max depth, minimum leaf node size, and the number of features to use for splitting (Figure 42). The relevant parameters of the linear regression model contain the singularity tolerance (Figure 43).

Figure 42. Parameters of random forest model

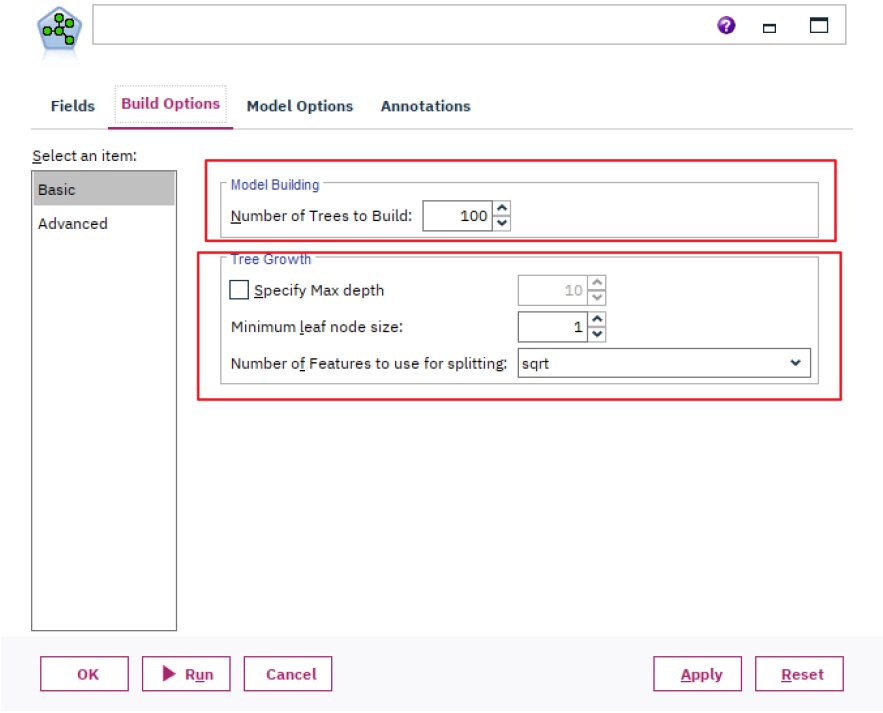
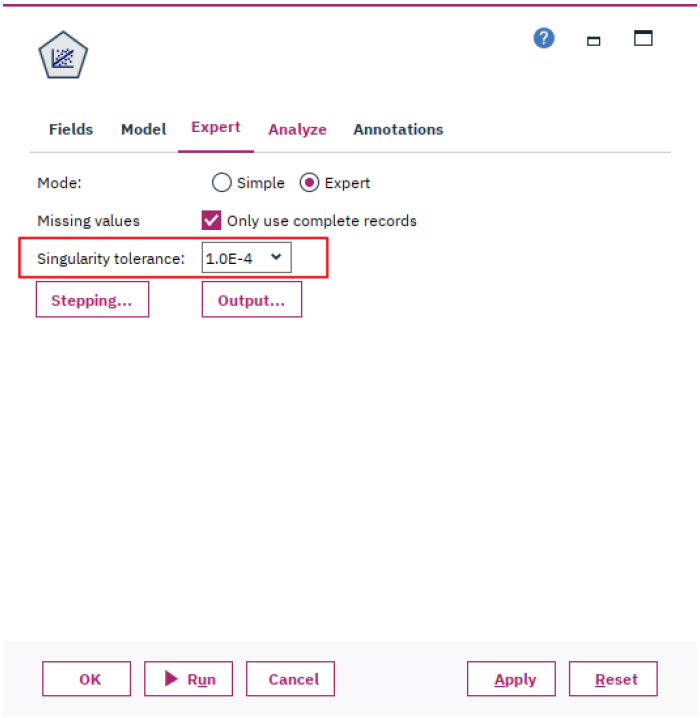


Figure 43. Parameters of linear regression model



6.3.2 Clustering

For the second goal, the TwoStep algorithm model is built; the inputs include 'Area', 'Total_population', 'Average Temperature', and 'Updated_total_emission' (Figure 44). The relevant parameters of this model contain the number of clusters, distance measures, and clustering criterion (Figure 45).

Figure 44. Building TwoStep model

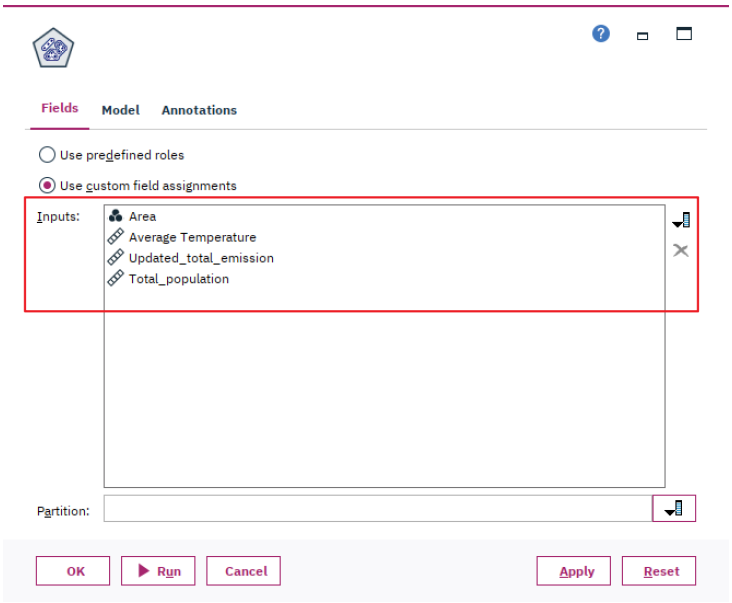
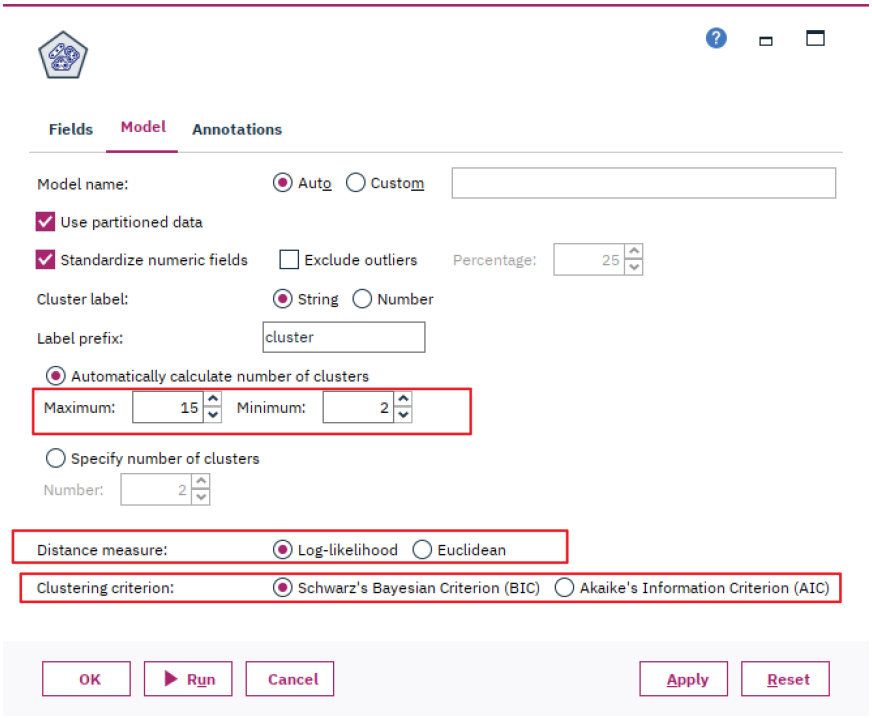


Figure 45. Parameters of TwoStep model

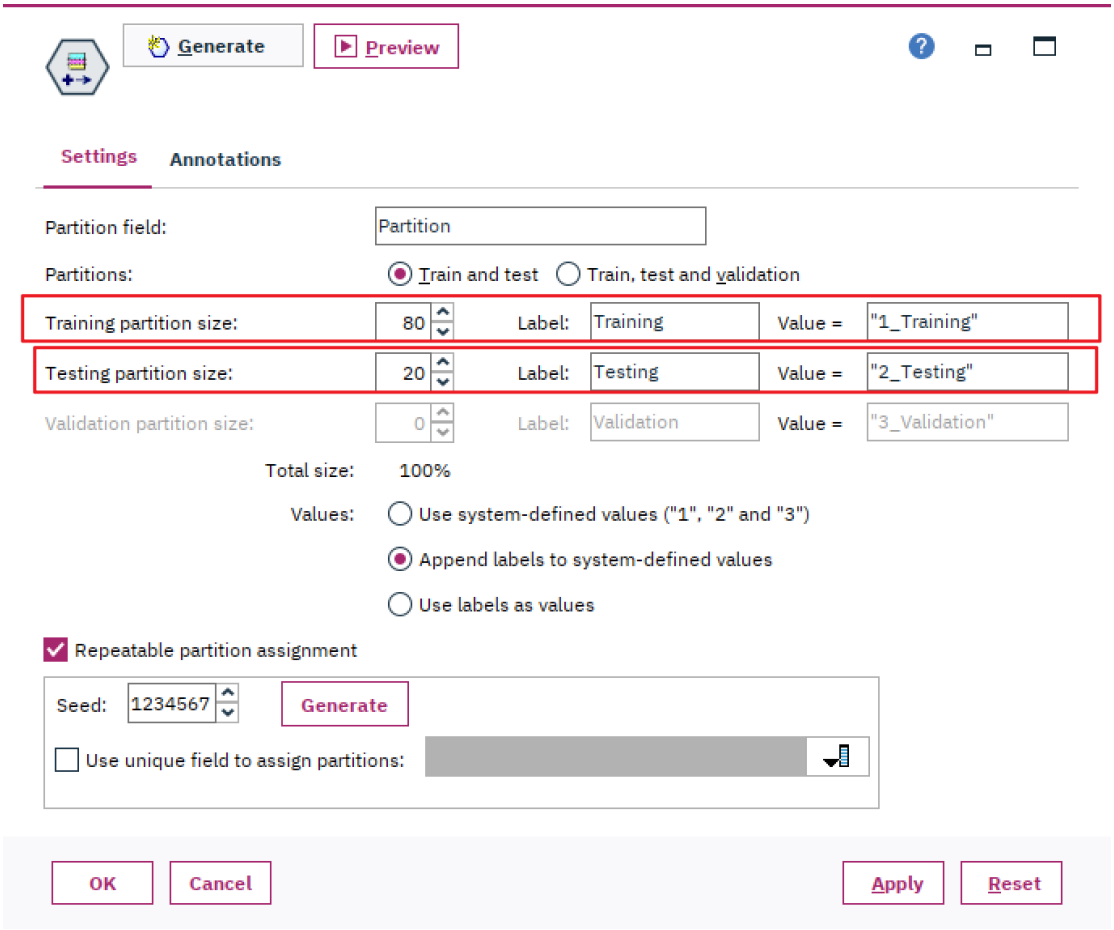


7. Data mining

7.1 Create and justify test designs

The random forest is a supervised learning algorithm. Thus, training and testing sets are separated from the whole dataset. The training set is 80% of the whole dataset, and the testing set is 20% (Figure 46). The utilisation of an 80/20 split for training and testing sets is widely acknowledged as a prevalent guideline within machine learning. The guideline above possesses broad applicability across various models and problem domains. The rationale behind this division is allocating a subset of the data to evaluate the model's performance while utilising the more significant data to train the model. The exact training-to-testing data ratio may vary depending on the analysis's requirements and the dataset's inherent attributes. The crucial aspect is to guarantee sufficient data in the training set to effectively train the model while simultaneously setting aside an adequate amount of data in the testing set to yield a dependable evaluation of the model's performance on unfamiliar data.

Figure 46. Training set and testing set



The TwoStep is an unsupervised learning algorithm. Unsupervised learning algorithms generally do not necessitate dividing data into separate training and testing sets. Unsupervised learning algorithms are advantageous due to their ability to train models without the need for labelled data. Instead, these algorithms rely on the calculation of relationships between data points to uncover the underlying structure of the data. Consequently, the entirety of the dataset is utilised to train an unsupervised learning model.

7.2 Conduct data mining – regression and clustering

7.2.1 Regression

The random forest model runs successfully (Figure 47). The data size of the training set is 5564. Figure 48 shows all model parameters of the random forest.

Figure 47. Random forest stream output

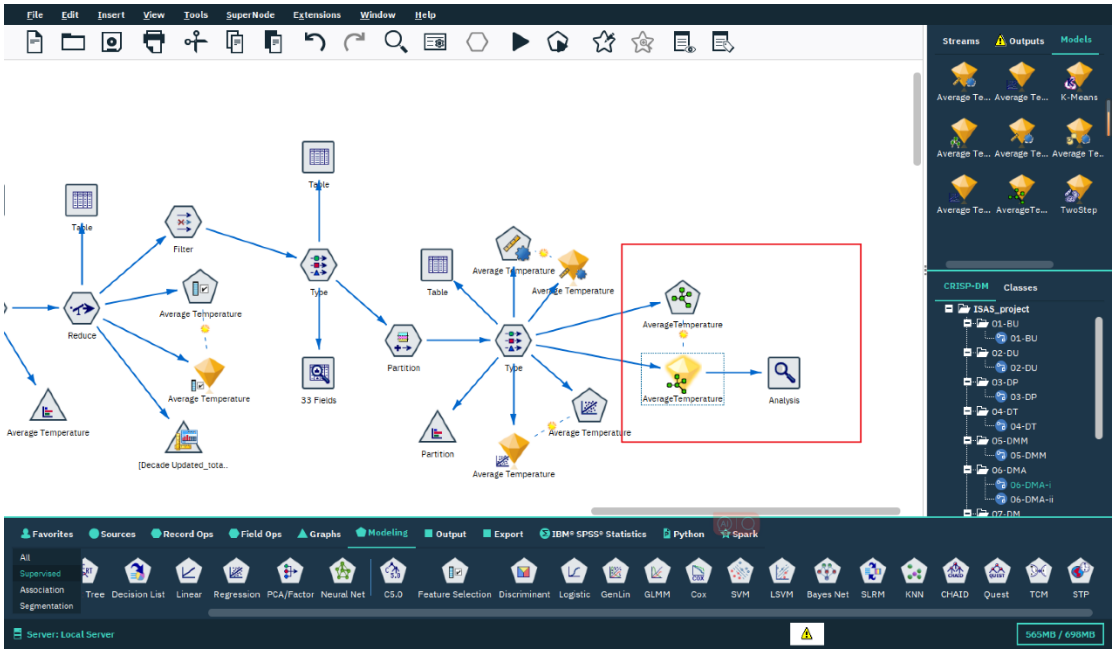
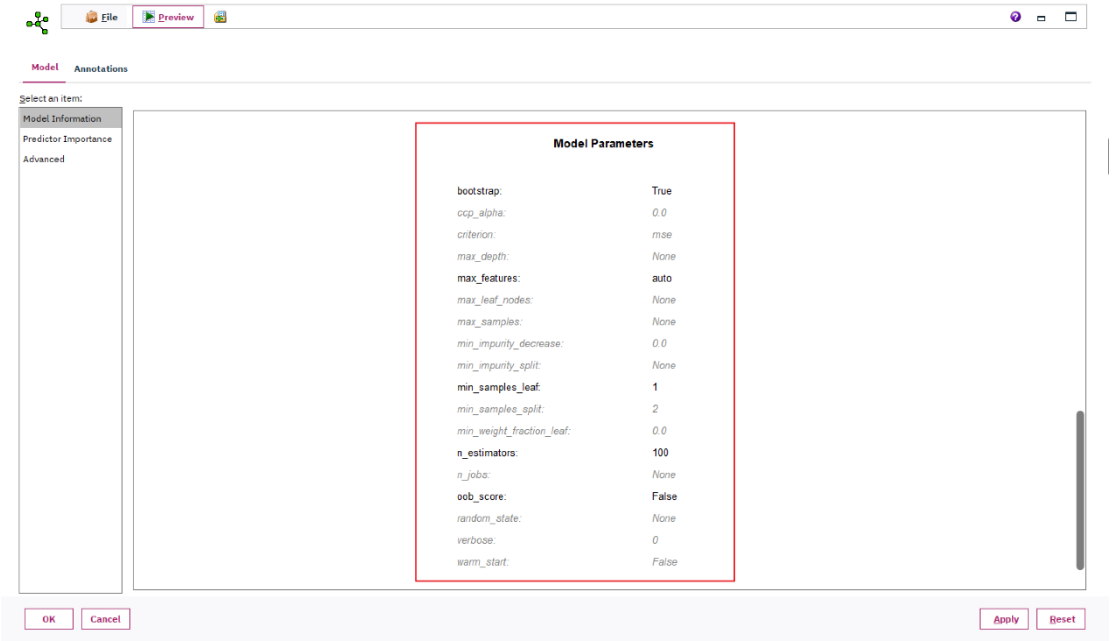


Figure 48. All model parameters



The linear regression model runs successfully (Figure 49). Figure 50 shows the linear relationship between the input and the target features, ‘Average Temperature.

Figure 49. Linear regression stream output

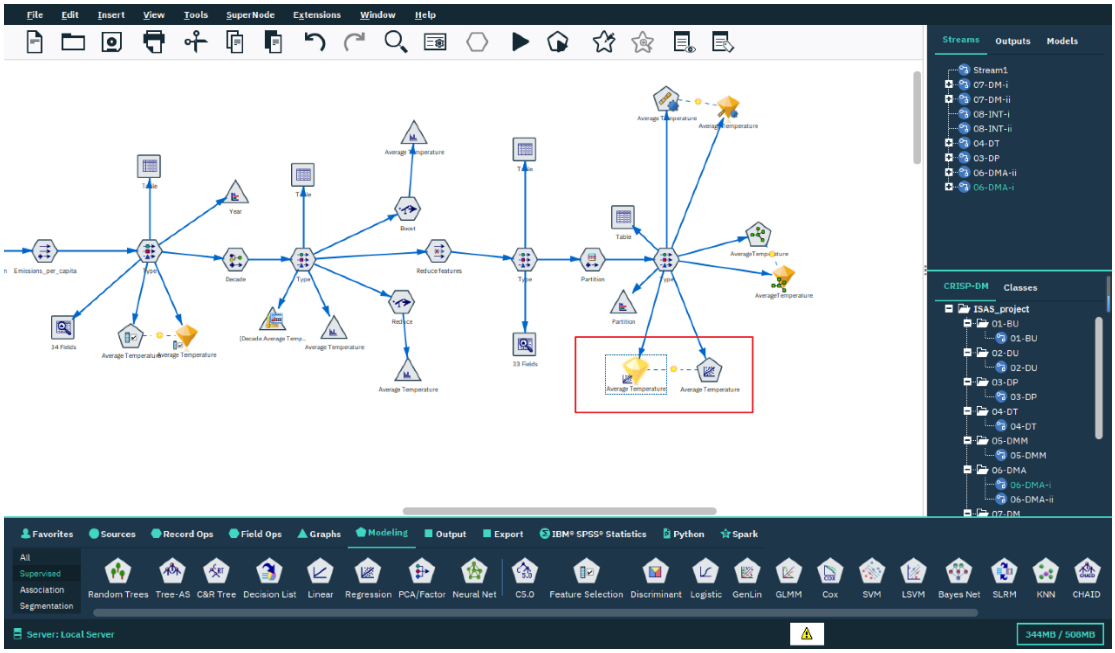
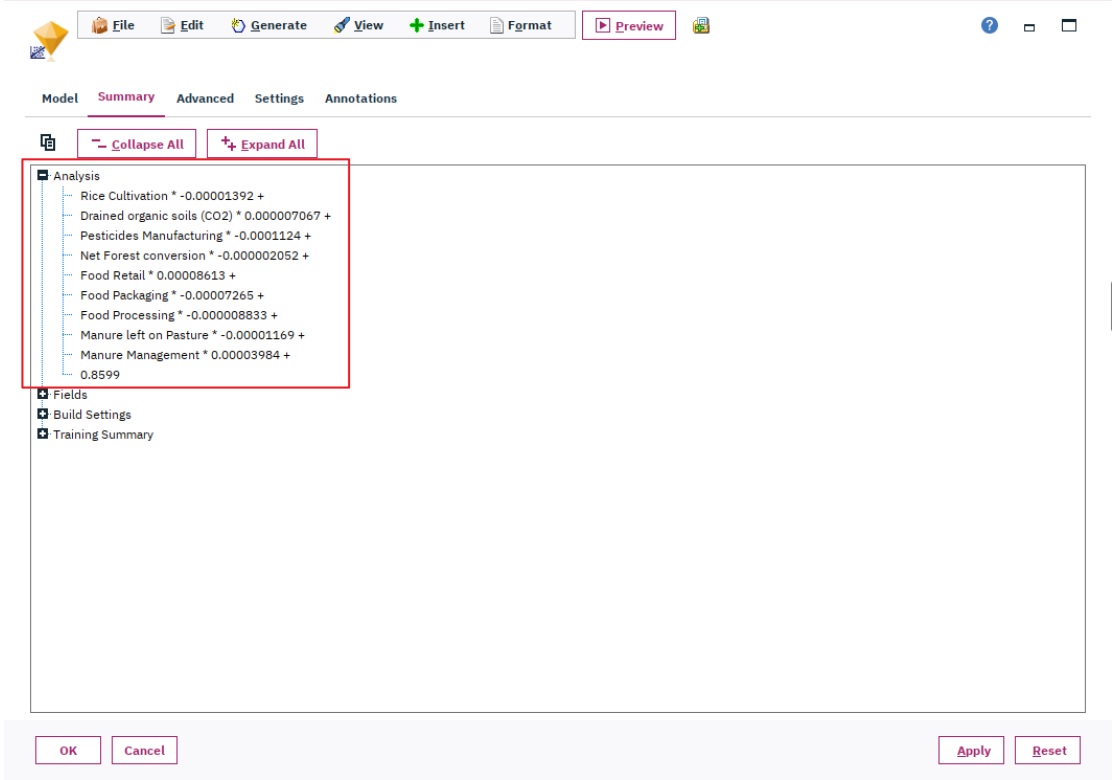


Figure 50. Linear relationship



7.2.2 Clustering

The TwoStep model runs successfully (Figure 51). The silhouette measure of cohesion and separation is close to 0.5, which means that the cluster quality is fair (Figure 52). There are two

clusters, the size of cluster 1 is 44%, and the of cluster 2 is 56% (Figure 53).

Figure 51. TwoStep stream output

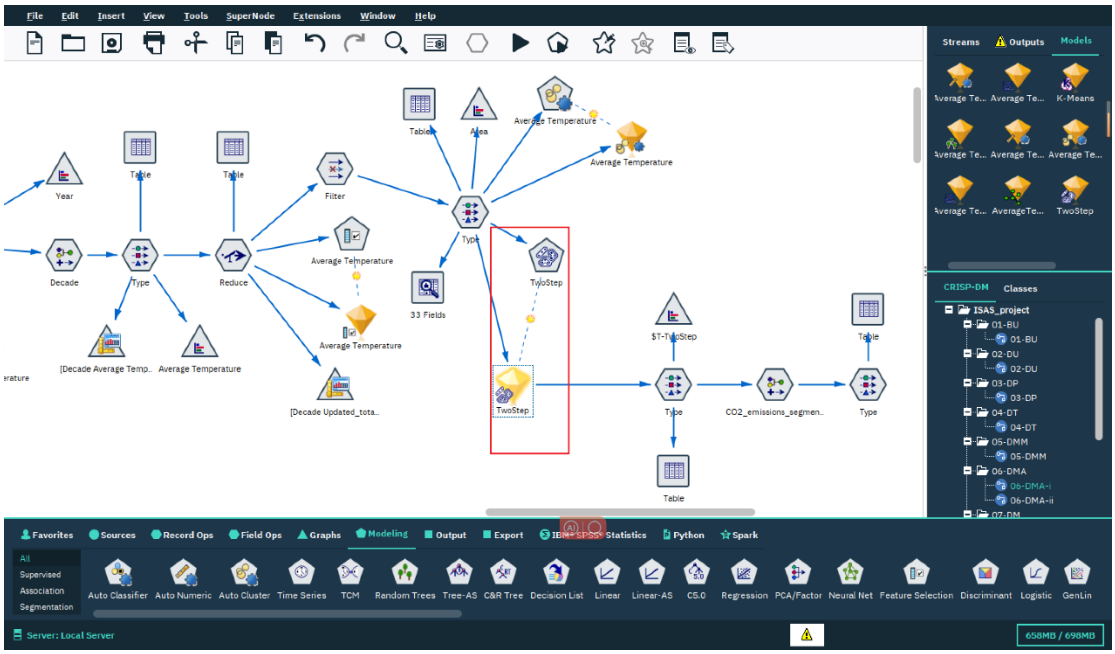


Figure 52. Cluster quality

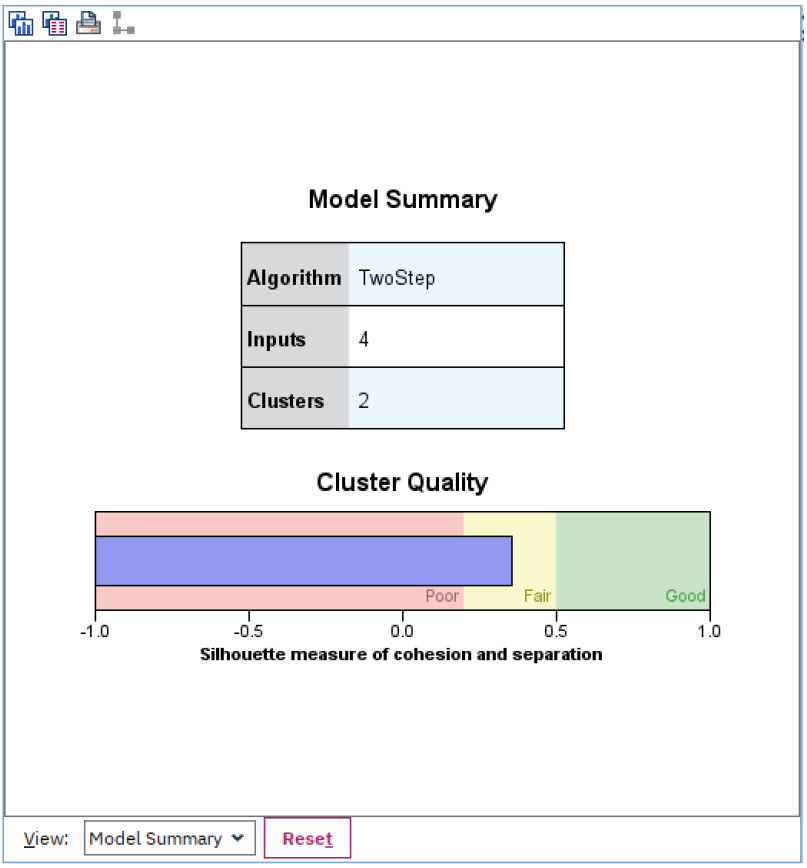
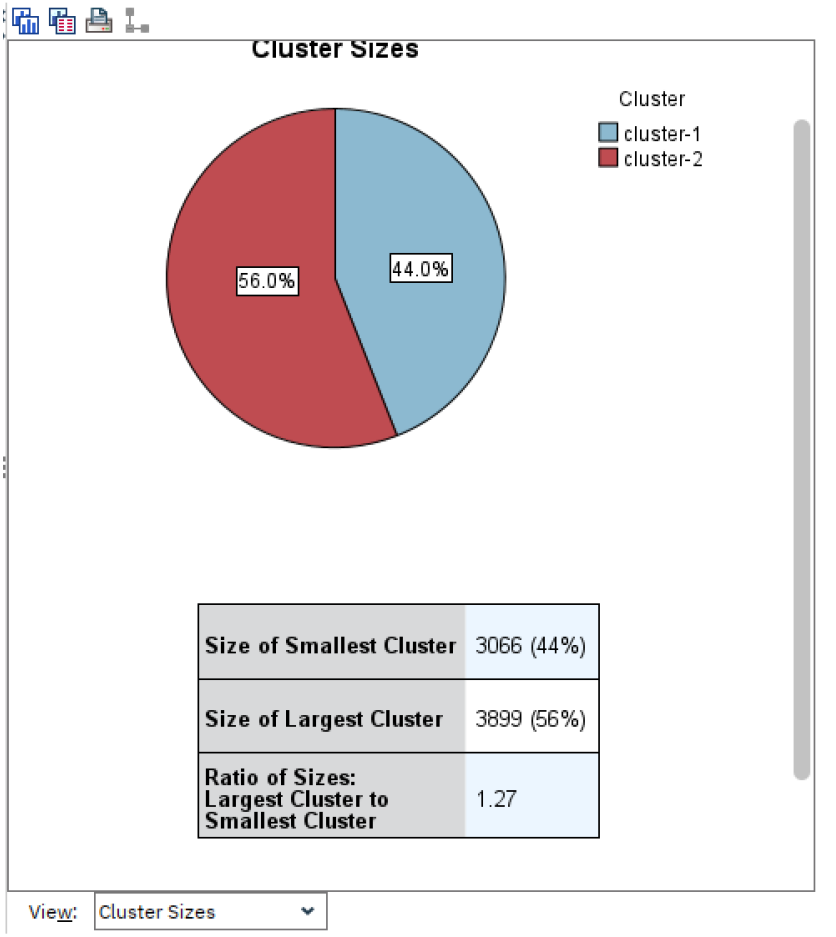


Figure 53. Two clusters



7.3 Search for patterns

7.3.1 Regression

There are three regression models, including all auto numeric algorithms, the random forest, and the linear regression, running successfully, so three patterns are based on different models, respectively (Figure 54, Figure 55, & Figure 56). All patterns present that the most critical agri-food factor relative to the average temperature rise is food retail.

Figure 54. Predictor importance based on all auto numeric algorithms

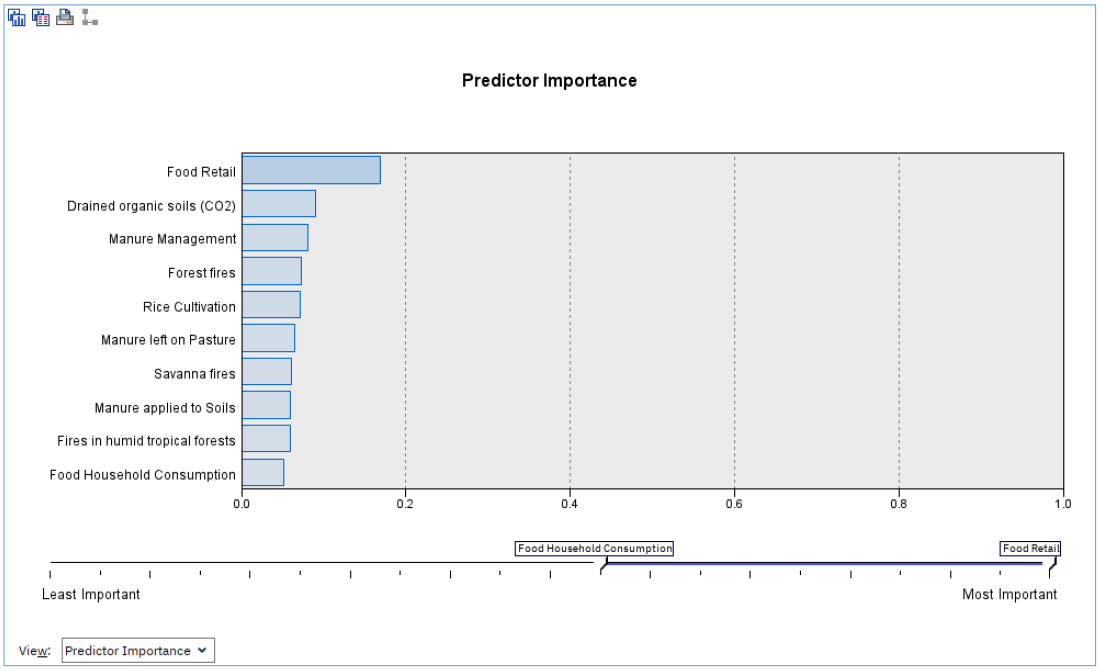


Figure 55. Predictor importance based on random forest

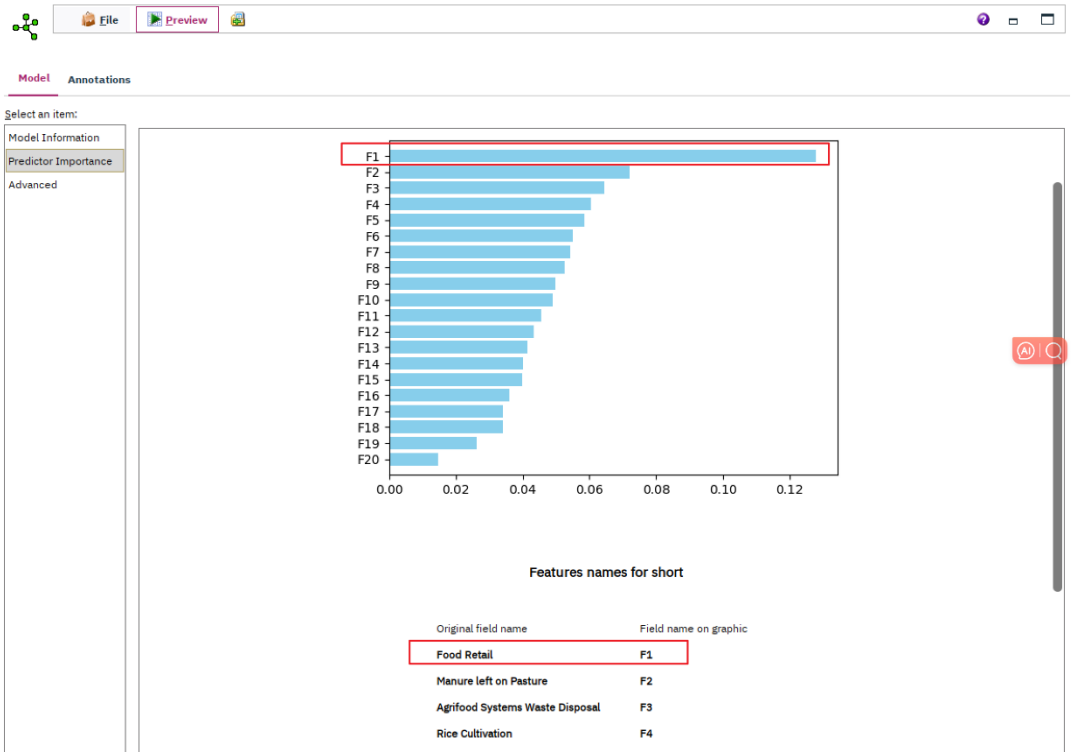
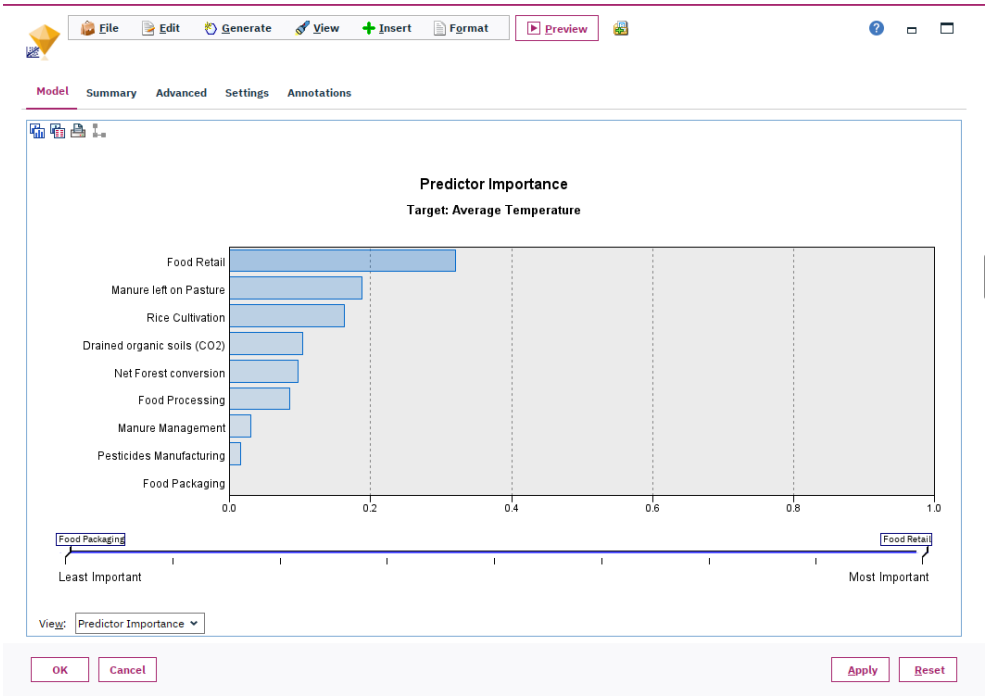


Figure 56. Predictor importance based on linear regression



7.3.2 Clustering

The first cluster is reclassified as ‘High Emission Countries’, and the second is reclassified as 'Low Emission Countries’ (Figure 57 & Figure 58).

Figure 57. Cluster 1 VS Cluster 2

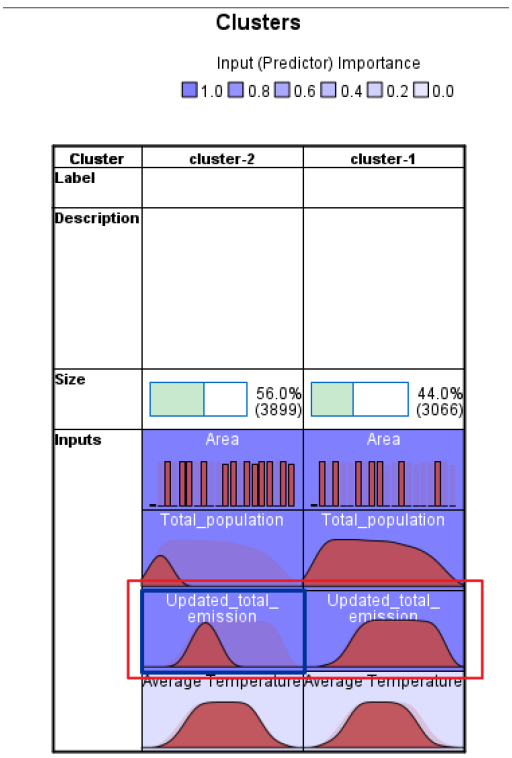
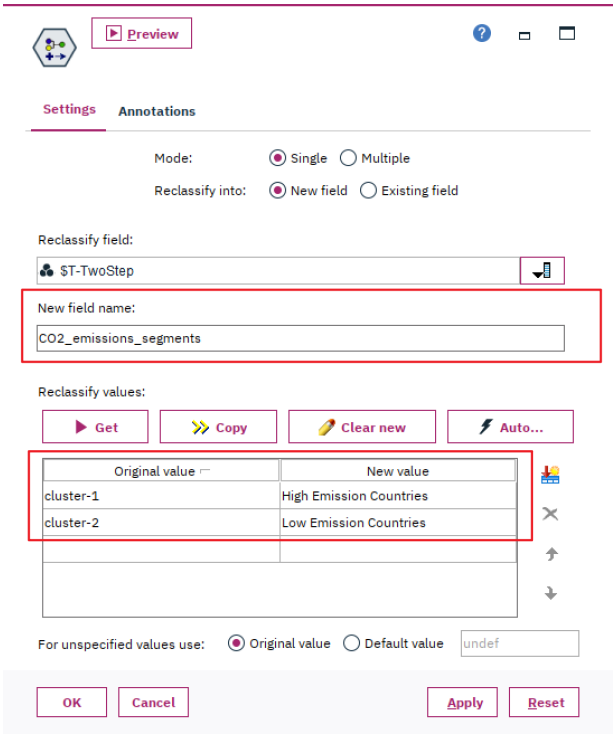
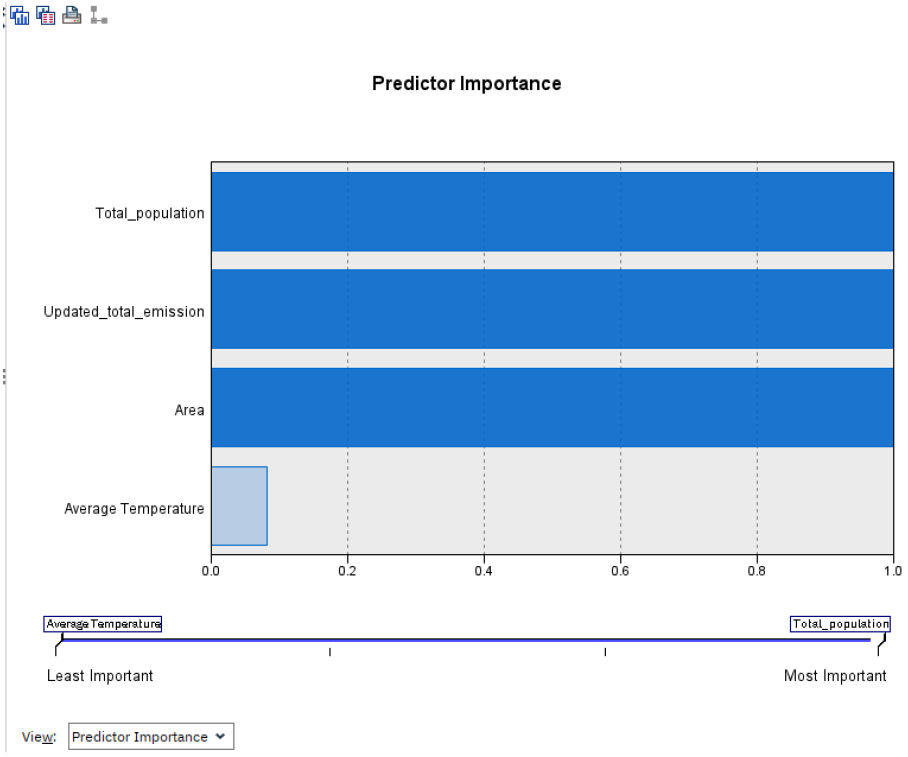


Figure 58. Reclassifying clusters



The most essential features are ‘Total_population’, ‘Updated_total_emission’, and ‘Area’ (Figure 59).

Figure 59. Important predictors



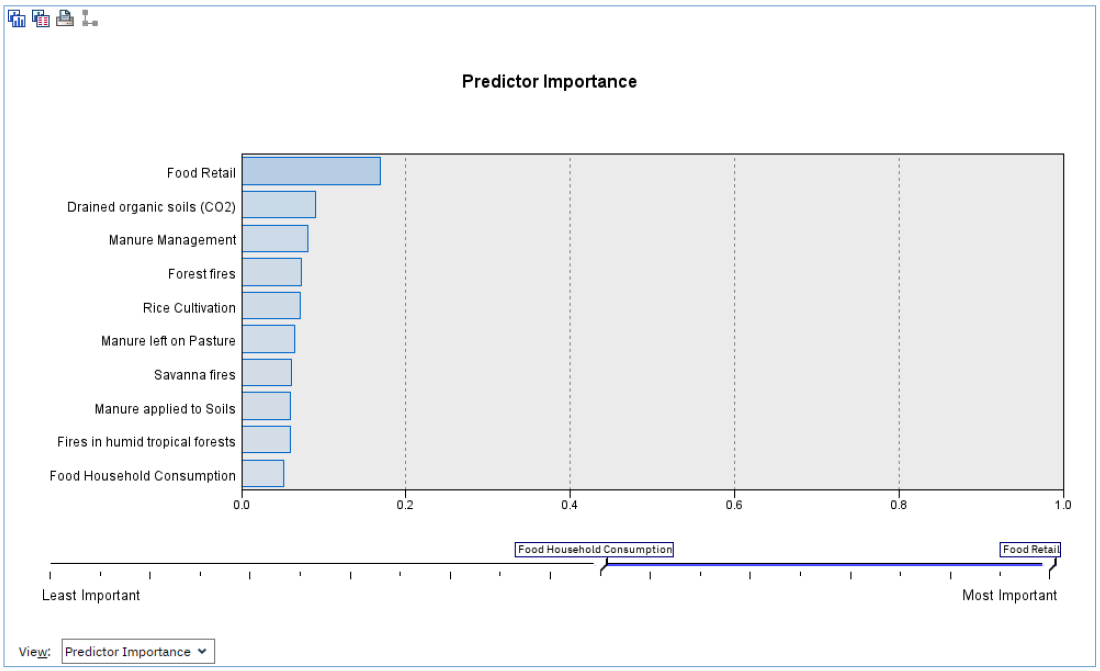
8. Interpretation

8.1 Study and discuss the mined patterns

8.1.1 Regression

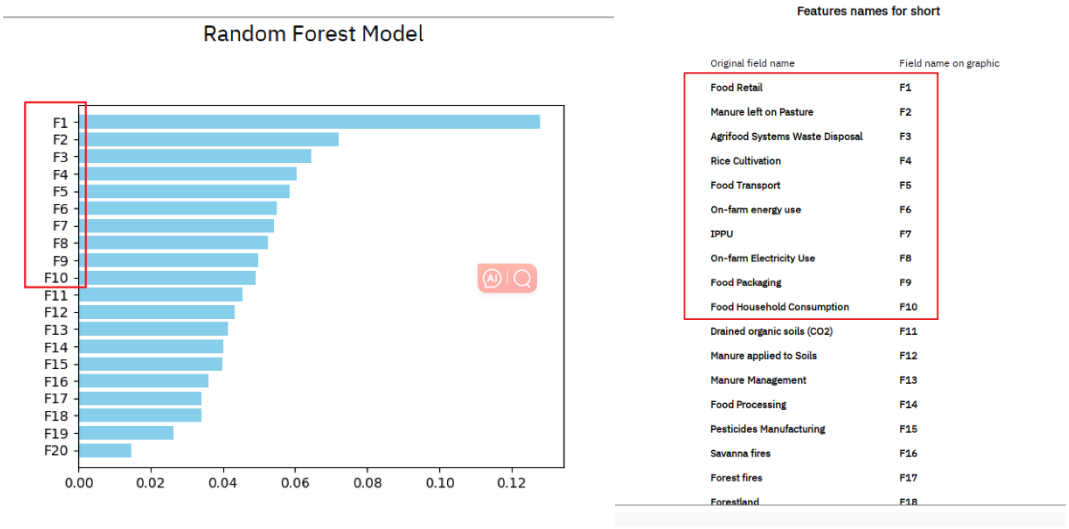
Based on all auto numeric regression algorithms, the top 10 most crucial agri-food features, which affect the subsequent temperature rise, are ‘Food Retail’, ‘Drained organic soils (CO2)’, ‘Manure Management’, ‘Forest fires’, ‘Rice Cultivation’, ‘Manure left on Pasture’, ‘Savanna fires’, ‘Manure applied to Soils’, ‘Fires in humid tropical forests’, and ‘Food Household Consumption’ (Figure 60).

Figure 60. Important features based on auto numeric regression algorithms



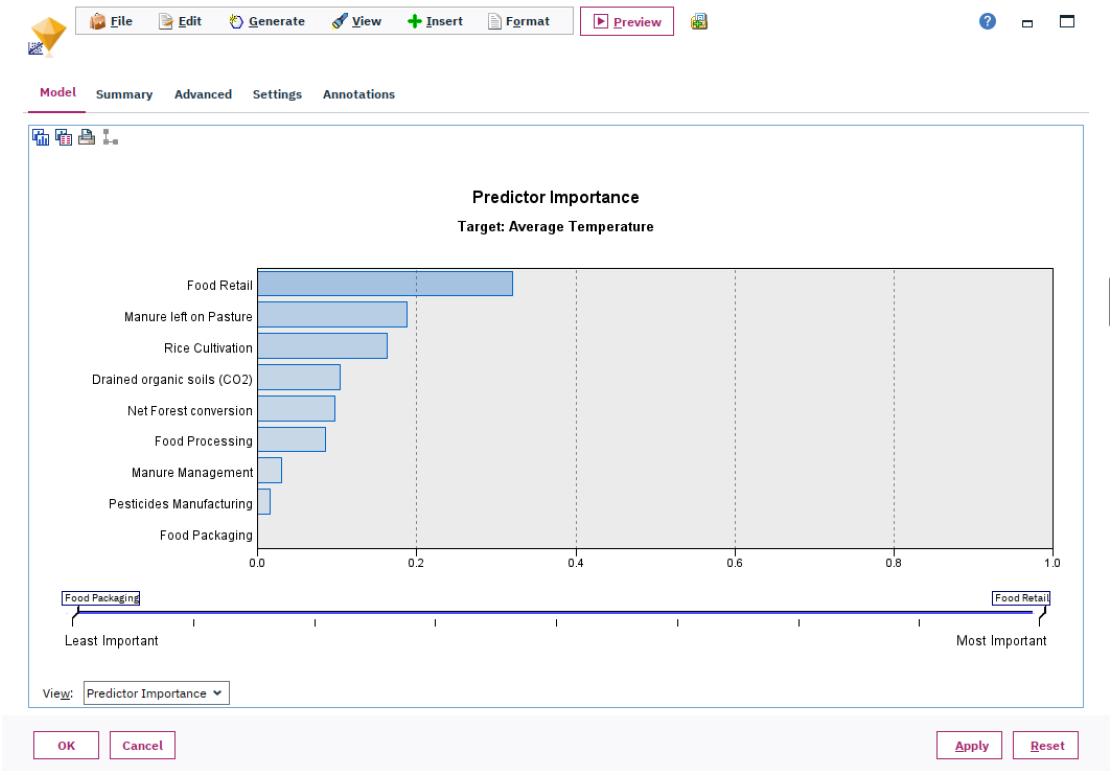
Based on the random forest algorithm model, the top 10 most crucial agri-food features, which influence the temperature rise, are ‘Food Retail’, ‘Manure left on Pasture’, ‘Afrifood system Waste Disposal’, ‘Rice Cultivation’, ‘Food Transport’, ‘On-farm energy use’, ‘IPPU’, ‘On-farm Electricity’, ‘Food Packaging’, and ‘Food Household Consumption’ (Figure 61).

Figure 61. Important features based on random forest model



Based on the linear regression algorithm model, the top 9 most essential agri-food features, which affect the temperature rise, are 'Food Retail', 'Manure left on Pasture', 'Rice Cultivation', 'Drained organic soils (CO2)', 'Net Forest conversion', 'Food Processing', 'Manure Management', 'Pesticide Manufacturing', 'Food Packaging' (Figure 62).

Figure 62. Important features based on linear regression model



8.1.2 Clustering

The high CO2 emission countries have more population than the low CO2 emission countries (Figure 63 & Figure 64). The difference in average temperature rise between the high CO2 emission countries and the low CO2 emission countries is not quite significant (Figure 65).

Figure 63. Total CO2 emissions of two clusters

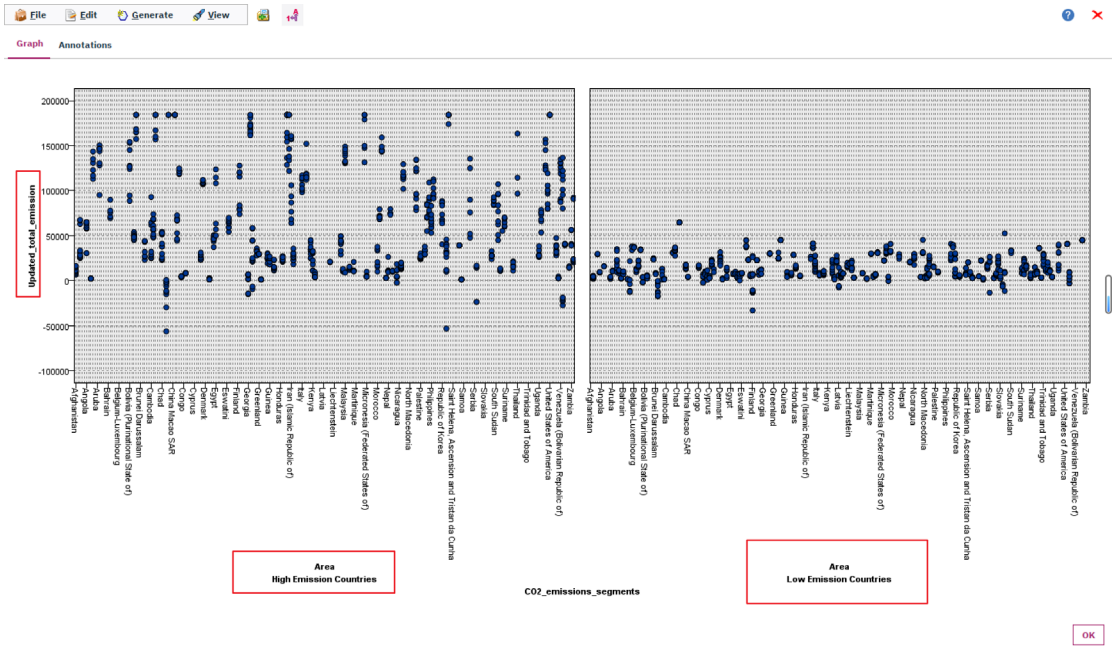


Figure 64. Population of two clusters

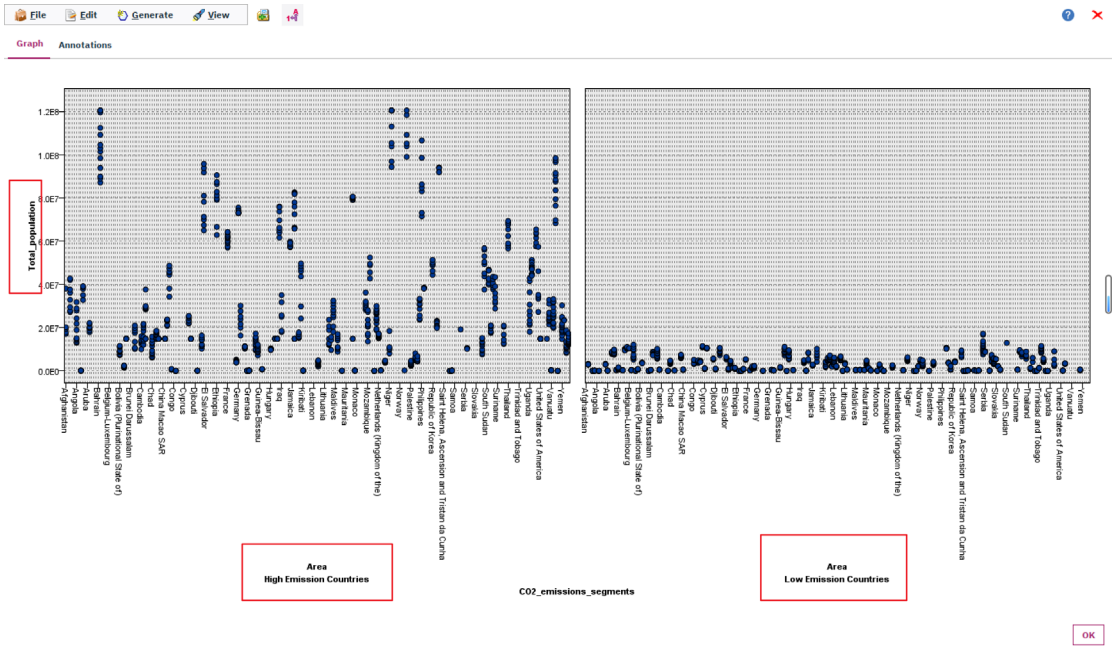
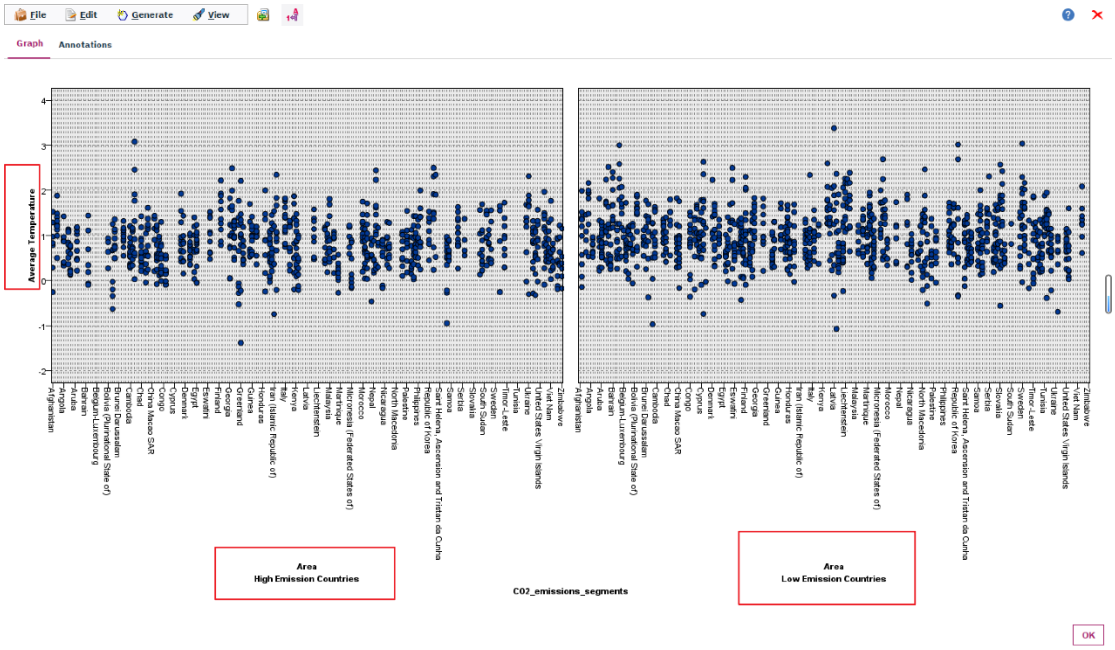


Figure 65. Average temperature rises of two clusters



8.2 Visualize the data, results, models, and patterns

8.2.1 The first data mining objective

The first data mining objective is to examine the correlation between CO2 emissions within the agri-food sector and the subsequent temperature rise.

Figure 66. Predictor Importance based on all auto numeric regression models

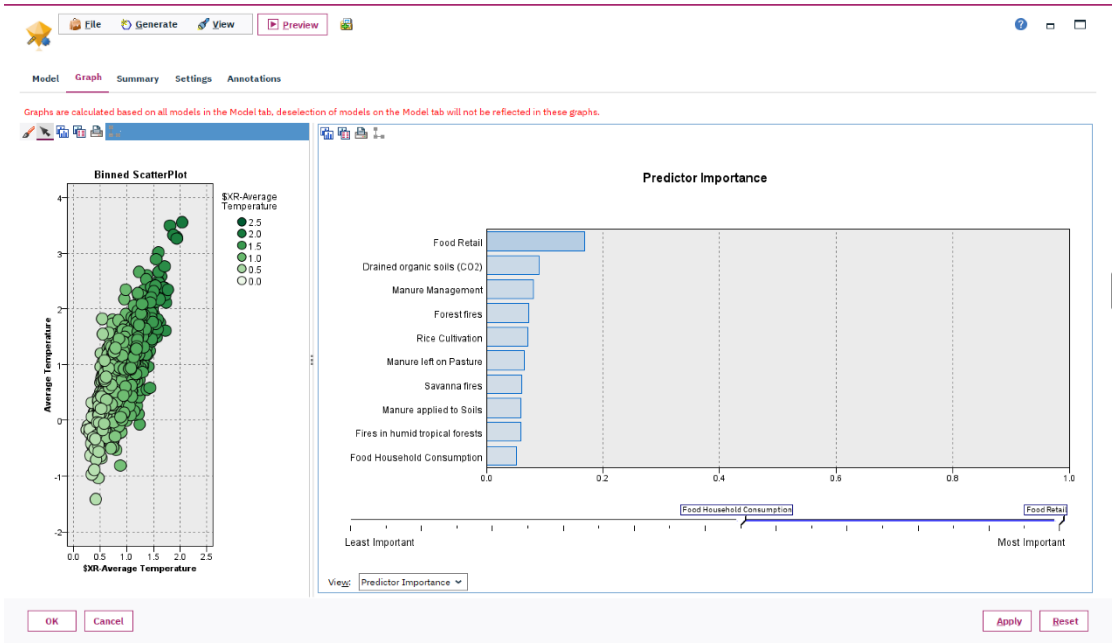


Figure 67. Model parameters of random forest

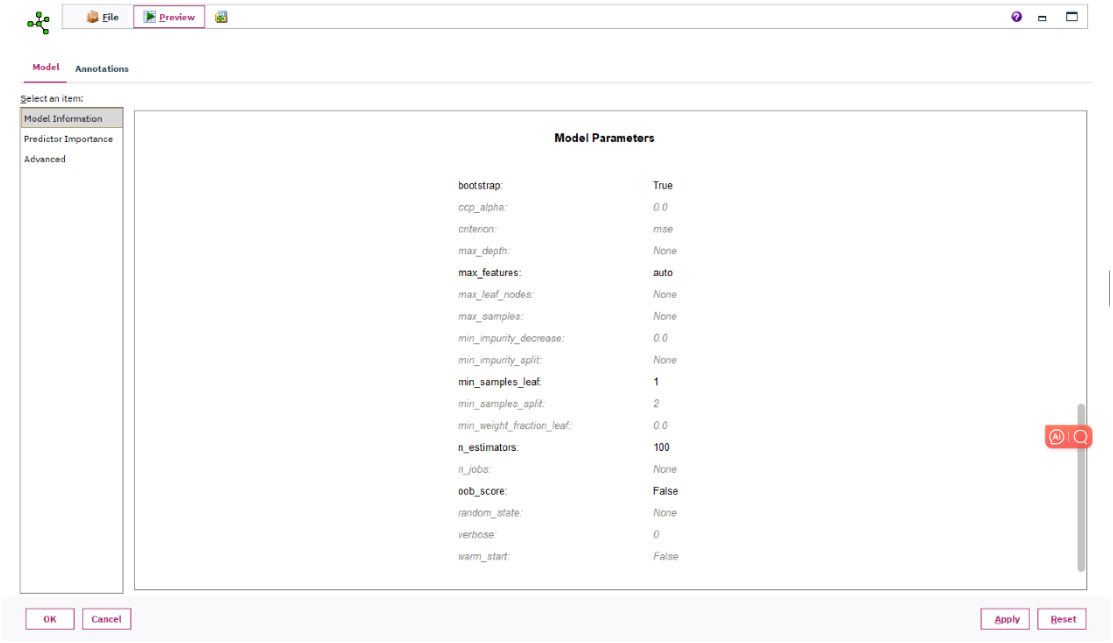


Figure 68. Predictor Importance based on random forest model

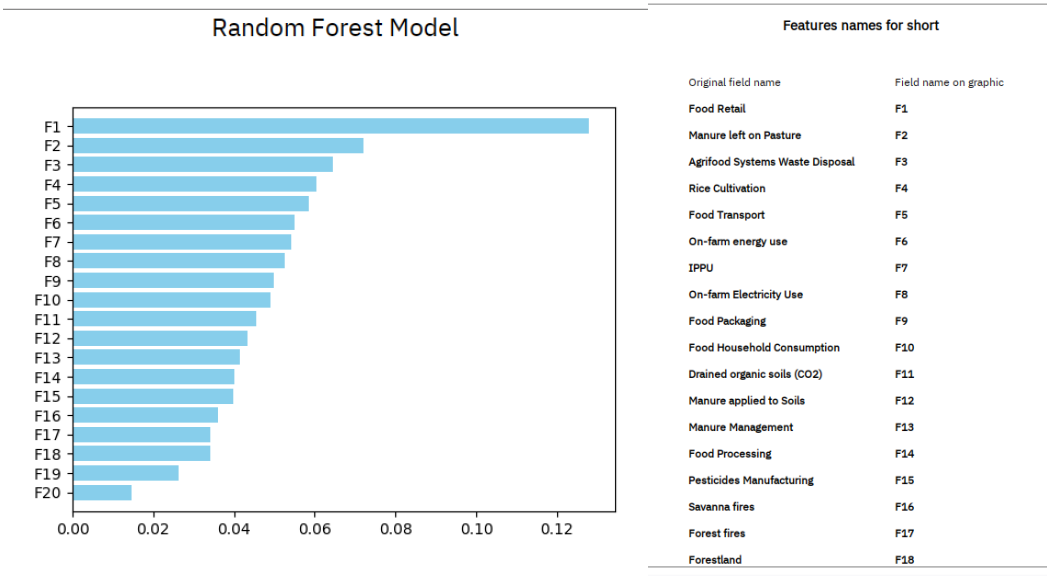


Figure 69. Predictor Importance based on linear regression model

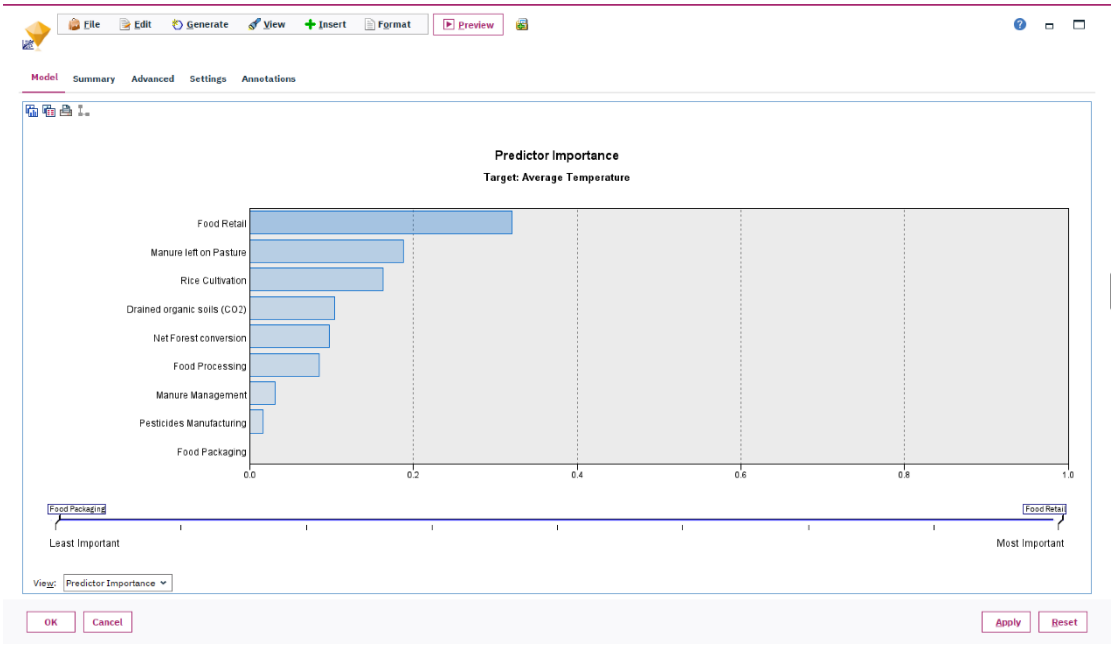


Figure 70. Linear relationship analysis

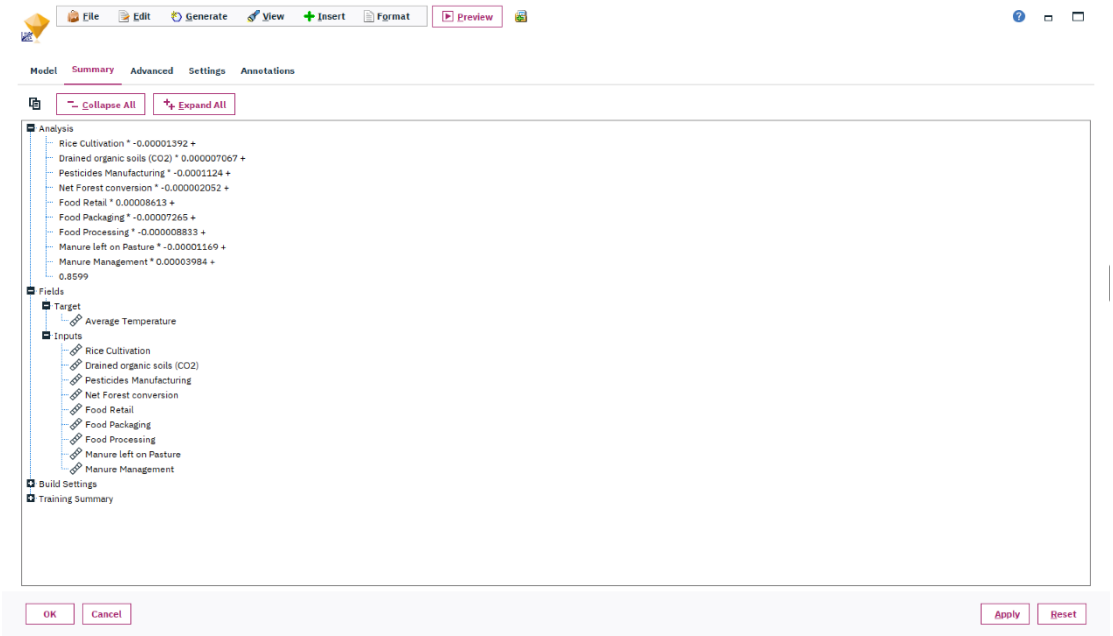


Figure 71. Linear regression model summary

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 72. Linear regression mode 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 73. 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

8.2.2 The second data mining objective

The second data mining objective is to anlalyse the influence of various countries based on aggregated data on emissions and temperature change.

Figure 74. Clustering model summary

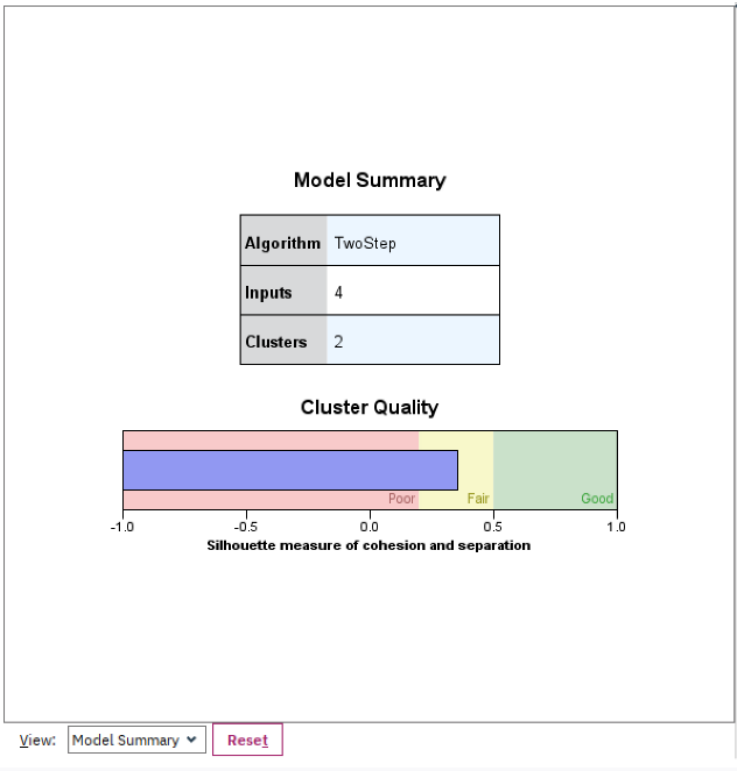


Figure 75. Clustering model predictor importance

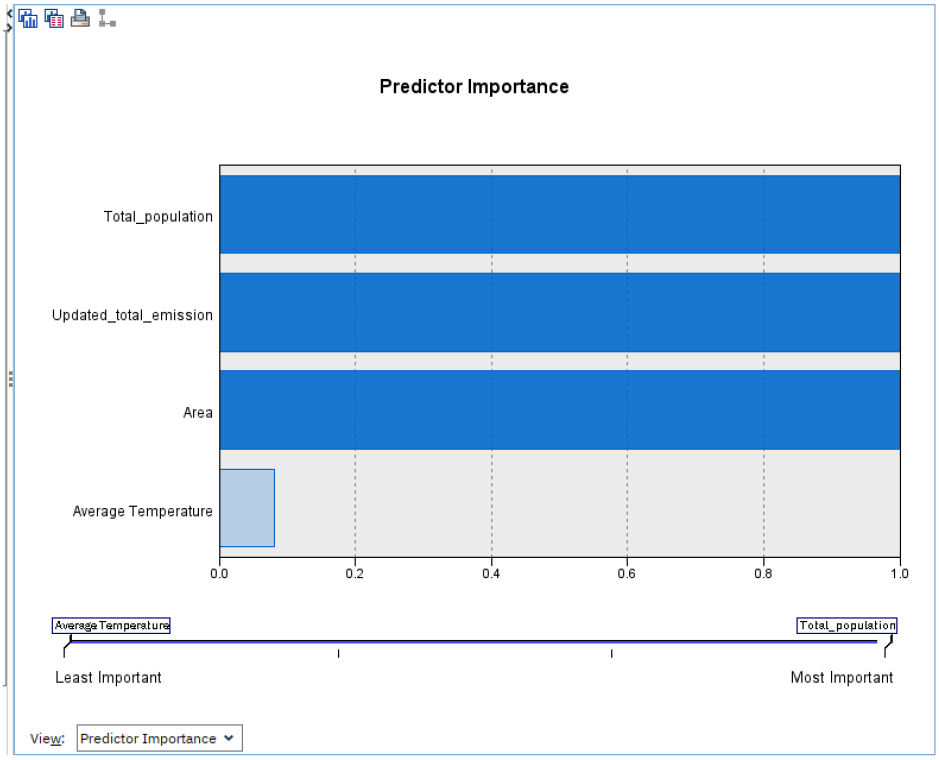


Figure 76. Clusters sizes

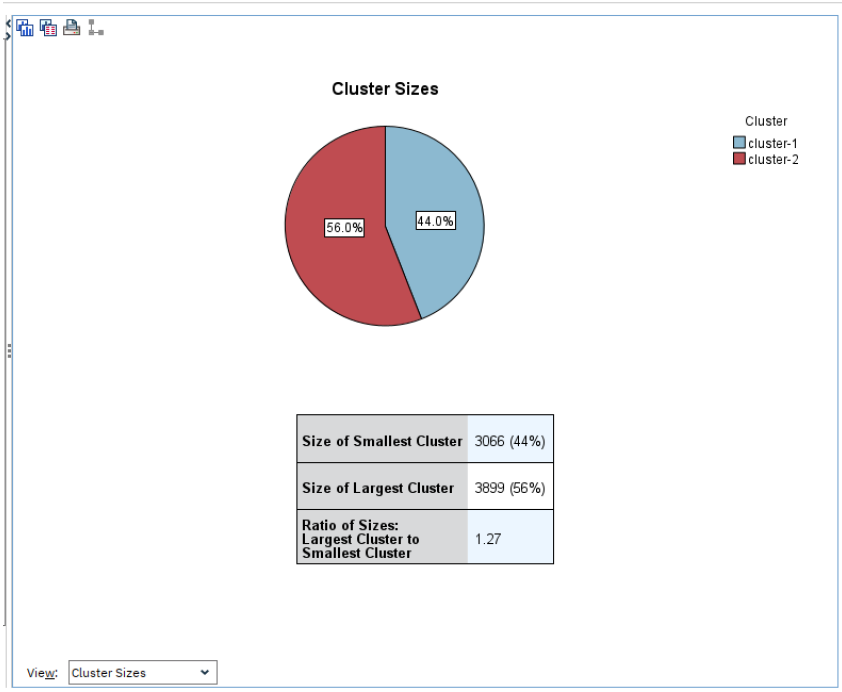


Figure 77. Cluster 1 VS Cluster 2

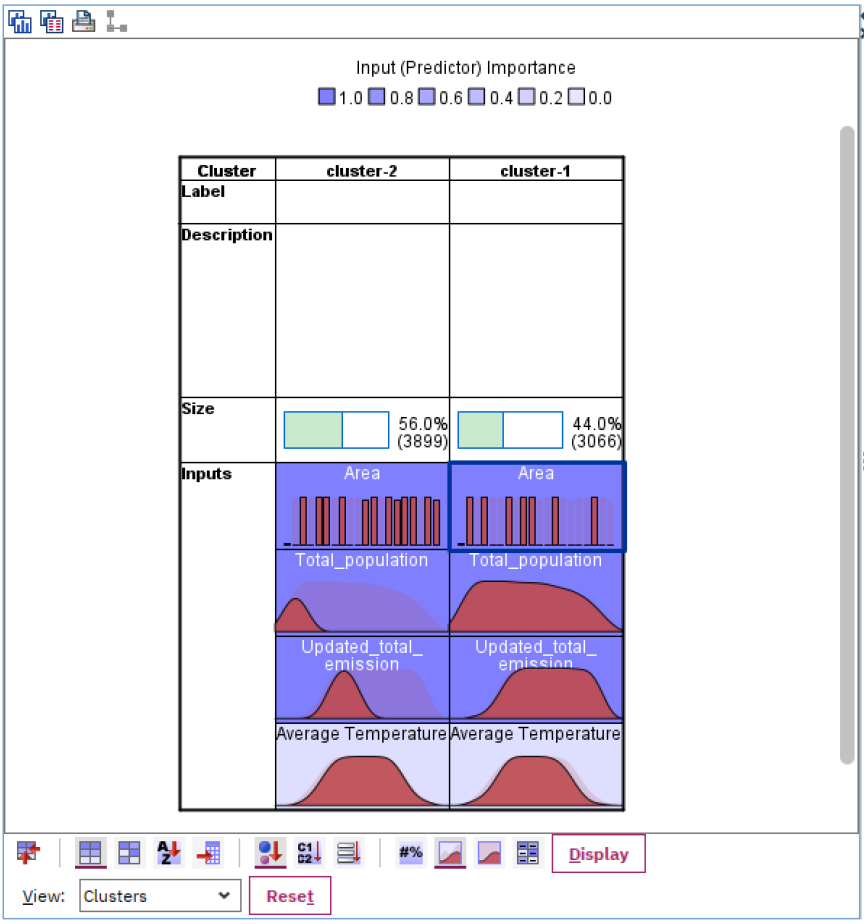


Figure 78. Total emissions distribution of two clusters

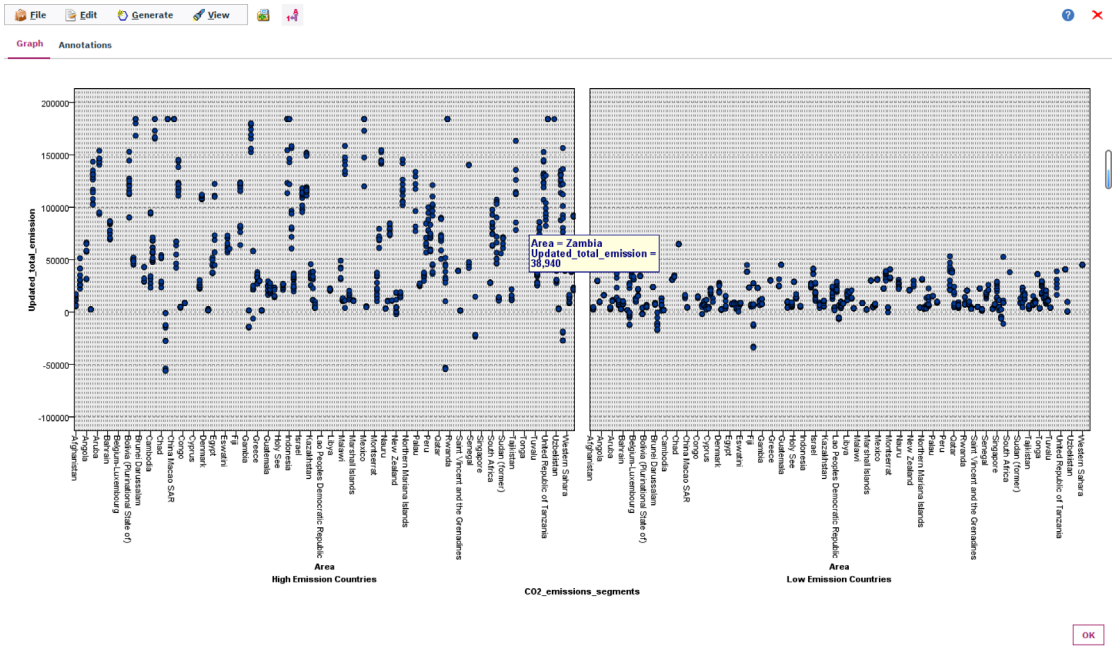


Figure 79. Total population distribution of two clusters

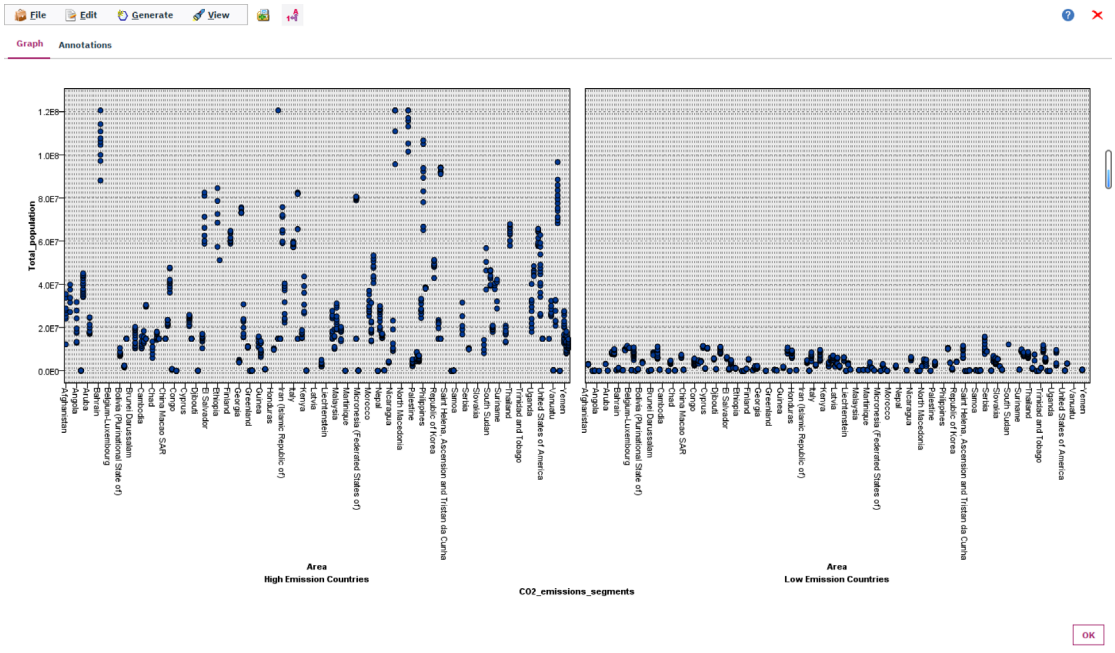


Figure 80. Average temperature rise distribution of two clusters

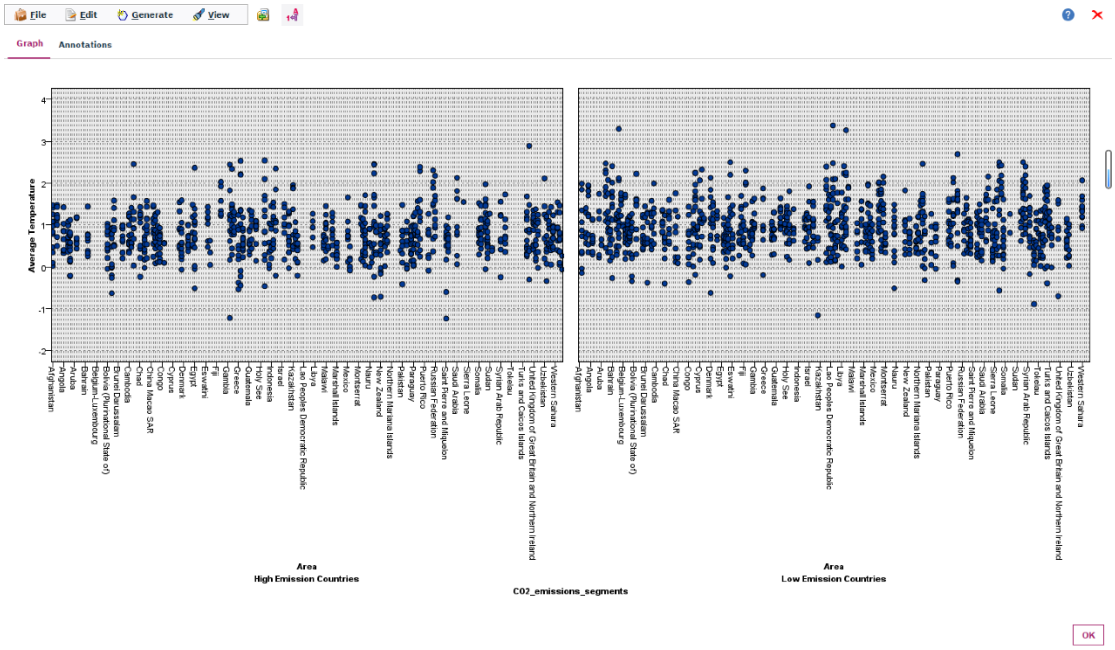
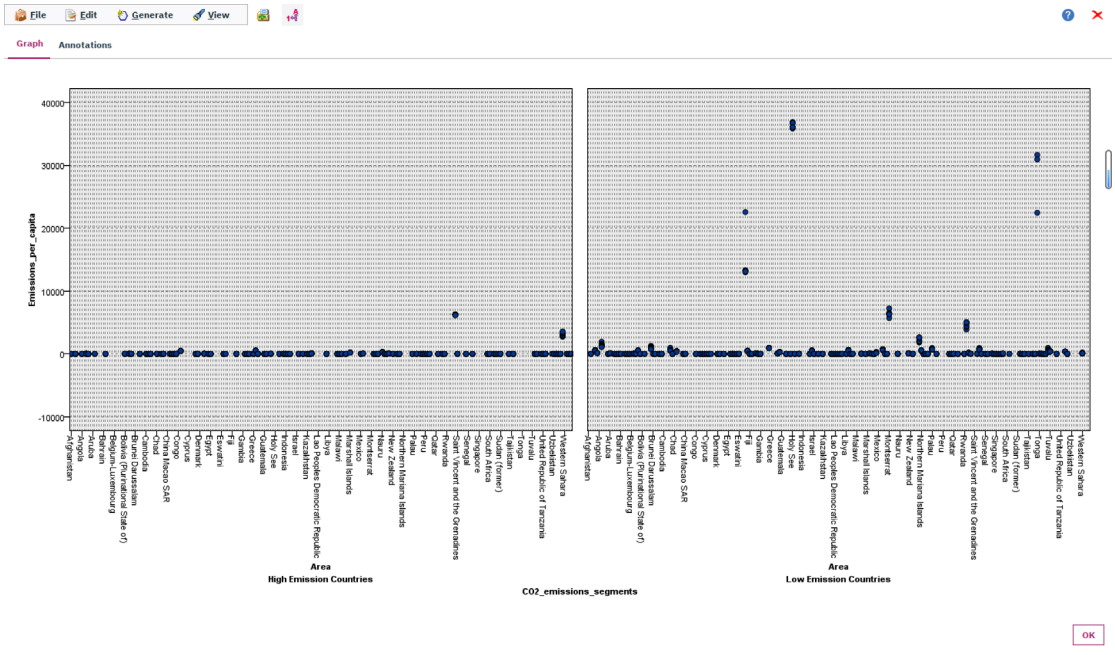


Figure 81. Emissions per capita distribution of two clusters



8.2.3 The third data mining objective

The third data mining objective is to Identify the countries with the highest average temperature increase by year and analyse their contributions to the overall environmental impact.

Figure 82. 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

8.3 Interpret the results, models, and patterns

8.3.1 The first data mining objective

The first data mining objective is to examine the correlation between CO2 emissions within the agri-food sector and the subsequent temperature rise. Through three regression models, ‘Food Retail’ is always the most important feature affecting the temperature increase. Besides that, ‘Rice Cultivation’, ‘Manure left on Pasture’, and ‘Food Household Consumption’ are always the top 10 most essential features in the three regression models. The linear regression analysis presents the linear relationship between the most crucial agri-food features and the temperature increase.

8.3.2 The second data mining objective

The second data mining objective is to anlalyse the influence of various countries based

on aggregated data on emissions and temperature change. Countries with higher CO2 emissions tend to have larger populations than countries with lower CO2 emissions. The disparity in average temperature increases between countries with high CO2 emissions and low CO2 emissions is not particularly substantial.

8.3.3 The third data mining objective

The third data mining objective is to identify the countries with the highest average temperature increase by year and analyse their contributions to the overall environmental impact. In 2020, the country with the highest average yearly temperature increase is Russian Federation. The average temperature increase in this country is 3.558℃, the total CO2 emissions are 184161.771 kilotons, and the CO2 emissions per capita are 2.030 tons (Figure 83).

Figure 83. Russian Federation with the highest average temperature increase

	Area	Year	Average Temperature	Updated_total_emission	Emissions_per_capita
1	Russian Federation	2020	3.558	184161.771	2.030

8.4 Assess and evaluate results, models, and patterns

8.4.1 Regression

For the linear regression model, as Figure 72 shows, when the degree of freedom of the regression values reaches 4, the F value is 69.197, higher than 44.05 (*F-Tables*, n.d.) (Figure 84).

Figure 84. F distribution table with 0.001 Sig.

Table 6: Critical values (percentiles) for the F distribution. Upper one-sided 0.001 significance levels; two-sided 0.002 significance levels; 99.9 percent percentiles.

	Numerator degrees of freedom																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	4053*	5000*	5404*	5625*	5764*	5859*	5929*	5981*	6023*	6056*	6107*	6158*	6209*	6235*	6261*	6287*	6313*	6340*	6366*
2	998.5	999.0	999.2	999.2	999.3	999.3	999.4	999.4	999.4	999.4	999.4	999.4	999.4	999.5	999.5	999.5	999.5	999.5	999.5
3	167.0	148.5	141.1	137.1	134.6	132.8	131.6	130.6	129.9	129.2	128.3	127.4	126.4	125.9	125.0	125.0	124.5	124.0	123.5
4	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47	48.05	47.41	46.76	46.10	45.77	45.43	45.09	44.75	44.40	44.05
5	47.18	37.12	33.20	31.09	29.75	28.84	28.16	27.64	27.24	26.92	26.42	25.91	25.39	25.14	24.87	24.60	24.33	24.06	23.79
6	35.51	27.00	23.70	21.92	20.81	20.03	19.46	19.03	18.69	18.41	17.99	17.56	17.12	16.89	16.67	16.44	16.21	15.99	15.75
7	29.25	21.69	18.77	17.19	16.21	15.52	15.02	14.63	14.33	14.08	13.71	13.32	12.93	12.73	12.53	12.33	12.12	11.91	11.70
8	25.42	18.49	15.83	14.39	13.49	12.86	12.40	12.04	11.77	11.54	11.19	10.84	10.48	10.30	10.11	9.92	9.73	9.53	9.33
9	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11	9.89	9.57	9.24	8.90	8.72	8.55	8.37	8.19	8.00	7.81
10	21.04	14.91	12.55	11.28	10.48	9.92	9.52	9.20	8.96	8.75	8.45	8.13	7.80	7.64	7.47	7.30	7.12	6.94	6.76
11	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12	7.92	7.63	7.32	7.01	6.85	6.68	6.62	6.35	6.17	6.00
12	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48	7.29	7.00	6.71	6.40	6.25	6.09	5.93	5.76	5.59	5.42
13	17.81	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98	6.80	6.52	6.23	5.93	5.78	5.63	5.47	5.30	5.14	4.97
14	17.14	11.78	9.73	8.62	7.92	7.43	7.08	6.80	6.58	6.40	6.13	5.85	5.56	5.41	5.25	5.10	4.94	4.77	4.60
15	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26	6.08	5.81	5.54	5.25	5.10	4.95	4.80	4.64	4.47	4.31
16	16.12	10.97	9.00	7.94	7.27	6.81	6.46	6.19	5.98	5.81	5.55	5.27	4.99	4.85	4.70	4.54	4.39	4.23	4.06

That means that the freedom of the regression values needs to reach 4 at least; the linear regression module is accurate enough. When the degree of freedom of the regression values is 4, the agri-food features, 'Food Retail', 'Manure left on Pasture', 'Food Packaging', and 'Rice Cultivation' are considered. 'Food Retail', 'Manure left on Pasture', and 'Rice Cultivation' are the most important agri-food features in all three regression models.

8.4.2 Clustering

The average silhouette measure of cohesion and separation is 0.4, which is between 0 and 0.5 at the fair level (Figure 74).

8.5 Iterate prior steps 1-7 as required

8.5.1 Regression

Stepwise method is used in the original linear regression model. Now, Backwards and Forwards methods are tried to general a more accurate linear regression algorithm model (Figure 85).

Figure 85. Different methods for linear regression

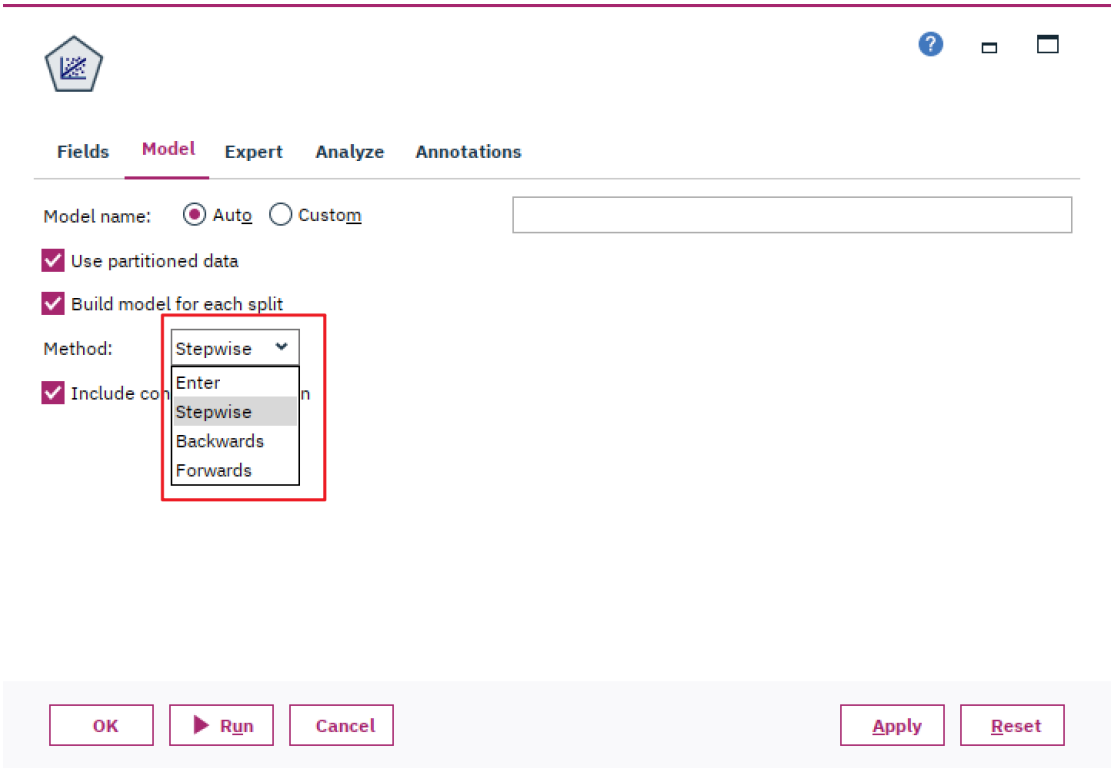


Figure 86 and Figure 87 show the ANOVA tables of Backwards and Forwards methods, respectively.

Figure 86. ANOVA table with Backwards method

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	111.005	21	5.286	18.391	<.001[b]
	Residual	1592.887	5542	.287		
	Total	1703.892	5563			
2	Regression	111.003	20	5.550	19.314	<.001[c]
	Residual	1592.889	5543	.287		
	Total	1703.892	5563			
3	Regression	110.999	19	5.842	20.333	<.001[d]
	Residual	1592.892	5544	.287		
	Total	1703.892	5563			
4	Regression	110.990	18	6.166	21.465	<.001[e]
	Residual	1592.902	5545	.287		
	Total	1703.892	5563			
5	Regression	110.965	17	6.527	22.726	<.001[f]
	Residual	1592.927	5546	.287		
	Total	1703.892	5563			
6	Regression	110.864	16	6.929	24.127	<.001[g]
	Residual	1593.027	5547	.287		
	Total	1703.892	5563			
7	Regression	110.722	15	7.381	25.705	<.001[h]
	Residual	1593.169	5548	.287		
	Total	1703.892	5563			
8	Regression	110.589	14	7.899	27.511	<.001[i]
	Residual	1593.302	5549	.287		
	Total	1703.892	5563			
9	Regression	110.382	13	8.491	29.573	<.001[j]
	Residual	1593.509	5550	.287		
	Total	1703.892	5563			
10	Regression	110.127	12	9.177	31.964	<.001[k]

Figure 87. ANOVA table with Forwards method

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			

For the Backward method, when the degree of freedom of the regression values reaches 6, the F value is 24.127, higher than 15.75 (*F-Tables*, n.d.). For the Forward method, when the degree of freedom of the regression values reaches 4, the F value is 69.197, higher than 44.05 (*F-Tables*, n.d.). The result of the Forward method is the same as the result of Stepwise.

8.5.2 Clustering

The original TwoStep clustering model uses Log-likelihood distance measure and Schwarz's Bayesian Criterion (BIC). Now, Euclidean distance measure and Akaike's Information Criterion (AIC) are tried to generate a more rational clustering algorithm model

(Figure 88).

Figure 88. Different distance measure and clustering criteria

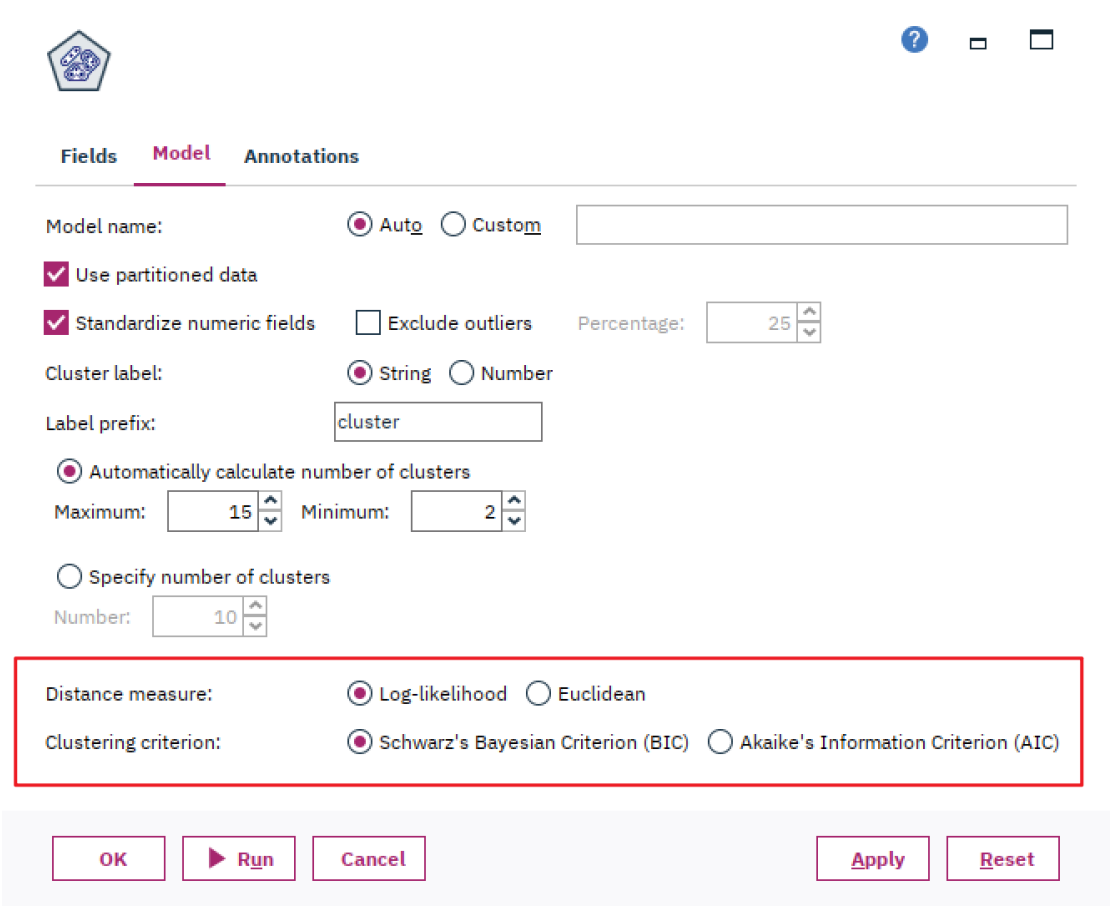


Figure 89 shows the result of combining Log-likelihood distance measure and AIC criterion; the average silhouette measure of cohesion and separation is 0.4, which is between 0 and 0.5 at the fair level. Figure 90 shows the result of combining Euclidean distance measure and AIC criterion; the average silhouette measure of cohesion and separation is 0.4, which is between 0 and 0.5 at the fair level. Figure 91 shows the result of combining Euclidean distance measure and BIC criterion; the average silhouette measure of cohesion and separation is 0.4, which is between 0 and 0.5 at the fair level. All four TwoStep clustering algorithm models have the same result.

Figure 89. Combining Log-likelihood distance measure and AIC criterion

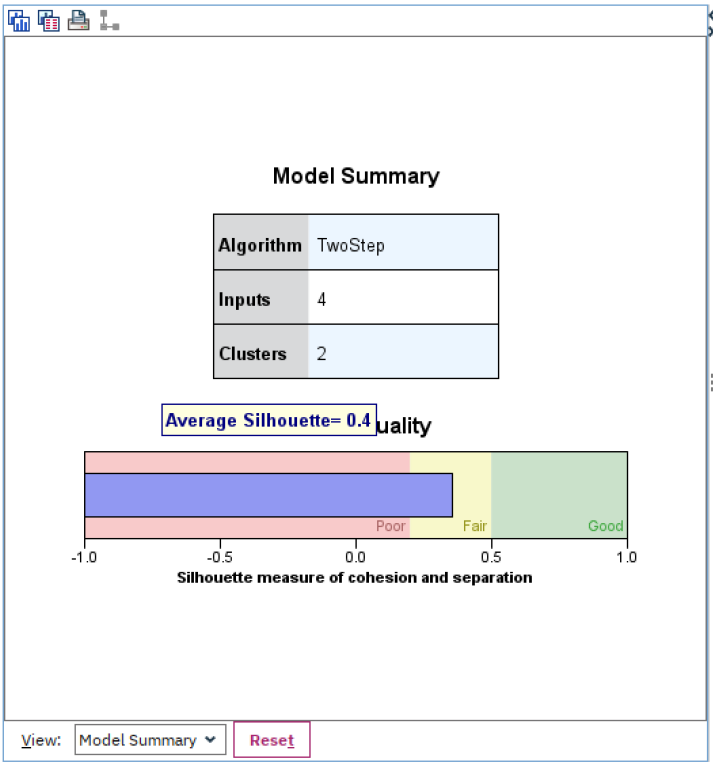


Figure 90. Combining Euclidean distance measure and AIC criterion

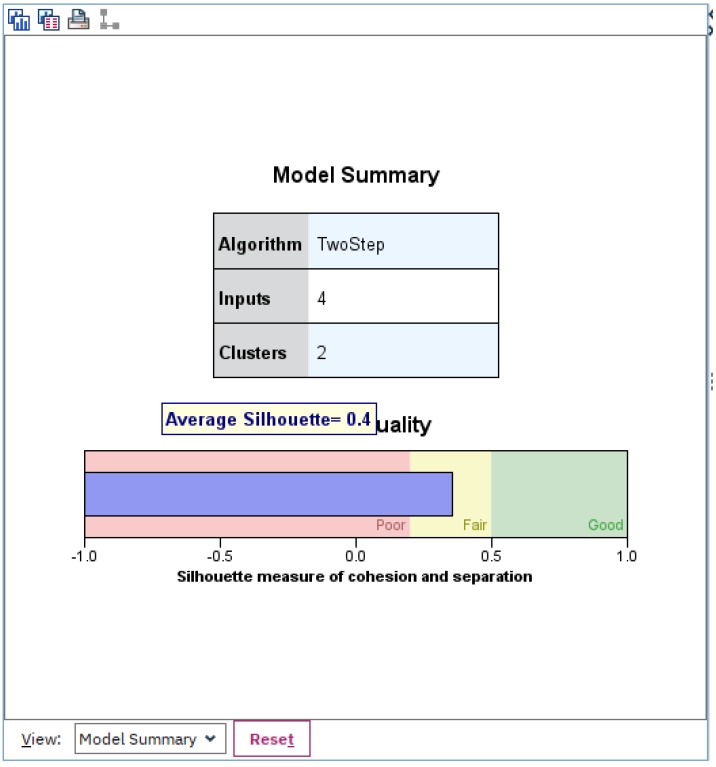
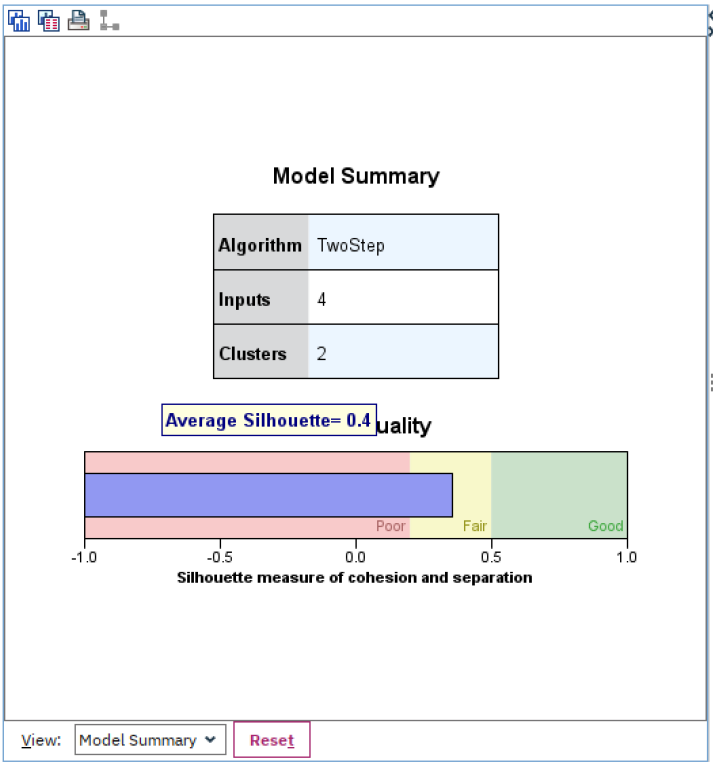


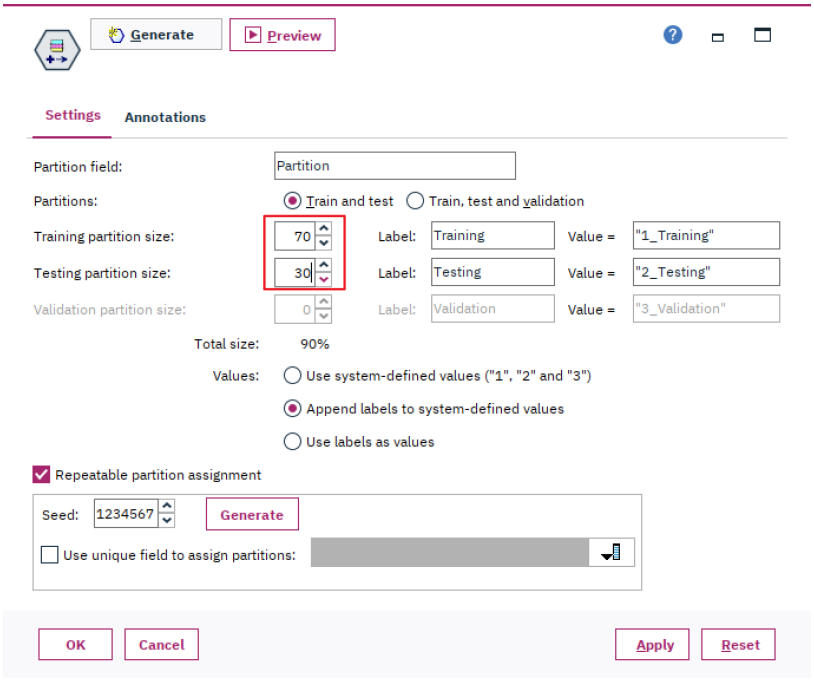
Figure 91. Combining Euclidean distance measure and BIC criterion



8.5.3 Redesign testing set

The original regression model utilizes the 80/20 ratio to split the dataset into the training and the testing sets. Now, 70/20 ratio is tried to generate a more accurate model (Figure 92).

Figure 92. Redesign test



When the degree of freedom of the regression values reaches 4, the F value is 63.362, higher than 44.05 (F-Tables, n.d.) (Figure 93). The result of the 70/30 ratio is the same as the result of the 80/20 ratio.

Figure 93. ANOVA with 70/30 ratio

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	32.997	1	32.997	109.981	<.001[b]
	Residual	1456.041	4853	.300		
	Total	1489.038	4854			
2	Regression	54.549	2	27.275	92.253	<.001[c]
	Residual	1434.489	4852	.296		
	Total	1489.038	4854			
3	Regression	66.393	3	22.131	75.464	<.001[d]
	Residual	1422.645	4851	.293		
	Total	1489.038	4854			
4	Regression	73.949	4	18.487	63.362	<.001[e]
	Residual	1415.089	4850	.292		
	Total	1489.038	4854			
5	Regression	86.309	5	17.262	59.671	<.001[f]
	Residual	1402.729	4849	.289		
	Total	1489.038	4854			
6	Regression	90.443	6	15.074	52.251	<.001[g]
	Residual	1398.595	4848	.288		
	Total	1489.038	4854			
7	Regression	94.216	7	13.459	46.772	<.001[h]
	Residual	1394.822	4847	.288		
	Total	1489.038	4854			
8	Regression	96.974	8	12.122	42.198	<.001[i]
	Residual	1392.064	4846	.287		
	Total	1489.038	4854			
9	Regression	98.493	9	10.944	38.130	<.001[j]
	Residual	1390.545	4845	.287		
	Total	1489.038	4854			

b. Predictors: (Constant), Food Retail

Reference

Agrifood systems. (2023). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Agrifood_systems&oldid=1153743860

Dean, J. (n.d.). *Big Data, Data Mining, and Machine Learning*.

F-Tables. (n.d.). Retrieved 19 August 2023, from http://faculty.washington.edu/heagerty/Books/Biostatistics/TABLES/F-Tables/

IBM Documentation. (2021, March 22). https://www.ibm.com/docs/en/spss-statistics/25.0.0?topic=features-twostep-cluster-analysis

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

Marr, B. (n.d.). *Big Data in Practice*.

Disclaimer

"I acknowledge that the submitted work is my own original work in accordance with the University of Auckland guidelines and policies on academic integrity and copyright. (See: https://www.auckland.ac.nz/en/students/forms-policies-and-guidelines/student-policies-and-guidelines/academic-integrity-copyright.html).

I also acknowledge that I have appropriate permission to use the data that I have utilised in this project. (For example, if the data belongs to an organisation and the data has not been published in the public domain then the data must be approved by the rights holder.) This includes permission to upload the data file to Canvas. The University of Auckland bears no responsibility for the student's misuse of data."