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

Project Proposal

Contents

1. Business understanding

1.1 Identify the objectives of the business

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.

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

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.

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. It is imperative to ensure meticulous and comprehensive scheduling of the project.

1.3 Determine data mining objectives

With the help of the particular data mining team or the 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 per capita emissions 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	Half a week	All analysts	Data problems
Data understanding	Half a week	All analysts	Data problems, technology problems
Data preparation	Half a week	Data mining consultant, database analyst	Data problems, technology problems
Data transformation	Half a week	Data mining consultant, database analyst	Data problems, technology problems
Data-mining methods selection	Half a week	Data mining consultant, database analyst	Technology problems, inability to find adequate model
Data-mining algorithms selection	Half a week	Data mining consultant, database analyst	Technology problems, inability to find adequate model
Data mining	One week	Data mining consultant, database analyst	Technology problems
Interpretation	Half a week	All analysts	Inability to implement results
Action	Half a week	Data mining consultant, database analyst	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).

2.2 Describe the data

All features show the corresponding CO2 emissions. CO2 is recorded in kilotons (kt); 1 kt represents 1000 kg of CO2. The feature "Average Temperature C°" serves as the target variable for machine learning models and signifies the mean annual temperature rise. For instance, when the value is 0.12, it signifies that 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 CO2 emission and temperature change by years. Figure 2 shows the average temperature distribution by year.

Figure 1. CO2 emission and temperature

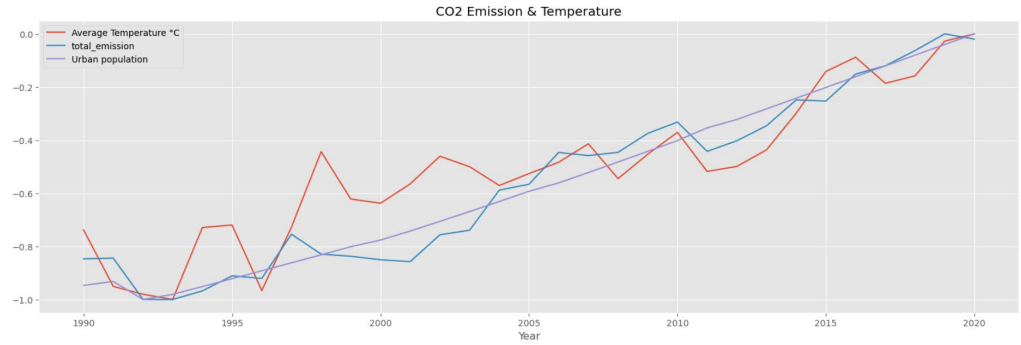
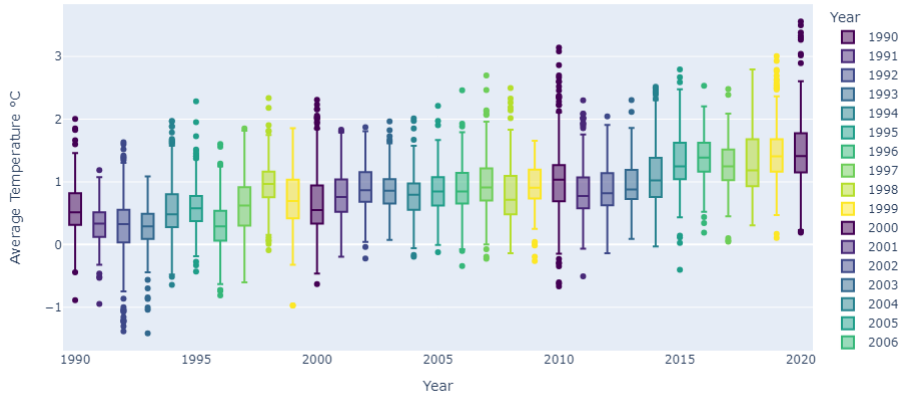


Figure 2. Average temperature distribution by years



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 the stages of data description and exploration.

Reference

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

"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."