ESCOLA
SUPERIOR
DE SAÚDE
POLITÉCNICO
DO PORTO

P.PORTO

M

MESTRADO
BIOESTATÍSTICA E BIOINFORMÁTICA APLICADAS À SAÚDE

Life Cycle Assessment using Machine Learning

Sofia Carolina Moura Gomes

10/**2024**

Life Cycle Assessment using Machine Learning

Autor

Sofia Carolina Moura Gomes

Orientadores

Professora Doutora Brígida Mónica Faria / E2S – P.PORTO; LIACC – Artificial Intelligence and Computer Science Lab (member of LASI).

Professora Doutora Alexandra Alves Oliveira / E2S – P.PORTO; LIACC – Artificial Intelligence and Computer Science Lab (member of LASI).

Professor Doutor Edgar Pinto / E2S – P.PORTO; REQUIMTE/LAQV - E2S - P.PORTO.

Dissertação apresentada para cumprimento dos requisitos necessários à obtenção do grau de Mestre em **Bioestatística e Bioinformática aplicadas à Saúde** pela Escola Superior de Saúde do Instituto Politécnico do Porto.

Acknowledgement of financial support

This research work was conducted as part of the project RETAIL – REtail using Technology based on Artificial InteLLigence, funded by COMPETE2023 (Ref. COMPETE2030-FEDER-00386800) and supported by ITEA4.

Acknowledgments

I want to thank Paulo, my boyfriend, for the way he always helped me throughout these two years in this master's degree. He has been my support since 2015.

I want to express my gratitude to my family and friends for their support and encouragement. Special thanks to my mother, who always encouraged me to study, and to my father for always being there.

I want to give a special thanks to my advisors who helped me with their knowledge and motivated me.

I also want to thank Dr. Filipa Heitor who made me feel more competent along this journey.

To my director, Elsa Castro, I am grateful for her understanding and encouragement.

To everyone who made this somehow possible, I express my gratitude.

Resumo

A Avaliação do Ciclo de Vida (ACV) é uma metodologia científica que permite avaliar o impacto de um produto ou serviço no meio ambiente, ao longo de todo o ciclo de vida. Engloba as fases de definição de objetivos e contexto, inventário, avaliação do impacto e interpretação.

A Inteligência Artificial (IA) refere-se a sistemas computacionais capazes de realizar tarefas que normalmente requerem inteligência humana. Machine Learning (ML) é uma área da IA que envolve o desenvolvimento de algoritmos capazes de aprender com os dados e fazer previsões ou tomar decisões com base neles. A ACV e ML têm sido combinados para ultrapassar a complexidade das várias fases da ACV e para diferentes objetivos, nomeadamente para desenvolver ferramentas de ACV substitutas.

O presente estudo centra-se na fase de Inventário do Ciclo de Vida (ICV), procurando encontrar as emissões de poluentes geradas para o ambiente para completar a fase de ICV da ACV. Esta dissertação procura responder à seguinte questão de investigação: "Podem ser aplicadas técnicas de ML para prever variáveis de resultado da fase de ICV da ACV?". Estas variáveis incluem todos os inputs e outputs ao longo do ciclo de vida de um produto.

O conjunto de dados utilizado neste trabalho é composto por 865 observações e refere-se a variáveis de *input* agrícola (e.g. fertilizantes químicos, pesticidas, mão de obra, gasóleo) e o *output* da produção e emissões de poluentes. Os dados foram retirados da literatura, referem-se produções agrícolas de kiwi, melancia, citrinos, chá e avelã, na província de Guilan, no norte do Irão. O cálculo das emissões de poluentes foi validado por um especialista na área, utilizando o Agri-footprint 4.0 e a versão atualizada como o Agri-footprint 6. Foram também consultadas outras metodologias, normas e relatórios importantes para esta investigação. Os modelos de Árvore de Decisão e Redes Neuronais desenvolvidos foram capazes de estimar as emissões de poluentes geradas para o ambiente ao longo do processo produtivo. Os resultados médios do Erro Normalizado Absoluto para a Árvore de Decisão, Rede Neuronal 1 e Rede Neuronal 2 foram, respetivamente, 1124.79, 0.07, e 0.14. O teste de Friedman, com p-*value* <0.001, menor que α=0.05, revela diferenças estatisticamente significativas nos valores de Erro Absoluto Normalizado em pelo menos um dos modelos. O teste de Wilcoxon (p-*value* <0.001) indica diferenças significativas entre todos os modelos.

Palavras-chave: Avaliação do Ciclo de Vida, Sustentabilidade, Machine Learning, Árvores de Decisão, Análise de Dados

Abstract

Life Cycle Assessment (LCA) is a scientific methodology that allows for assessing the impact of a product or service on the environment, throughout its life cycle. It includes defining objectives and context, inventory, impact assessment, and interpretation phases.

Artificial Intelligence (AI) refers to computer systems capable of performing tasks that typically require human intelligence. Machine Learning (ML) is an area of AI that involves the development of algorithms capable of learning from data and making predictions or decisions based on data. LCA and ML have been combined to overcome LCA's complexity at various stages and for different purposes, namely, to develop surrogate LCA tools.

This study focuses on the application of ML in the Life Cycle Inventory (LCI) phase to find the pollutant emissions generated into the environment to complete the LCI phase of the LCA. The present work seeks to answer the following question: "Can Machine Learning techniques be applied to predict outcome variables of the LCI phase of LCA?". These variables include all the inputs and outputs throughout the life cycle of a product.

The database used in this work comprises 865 observations containing agricultural input variables (e.g. chemical fertilizer, pesticides, human labor, diesel fuel) and production output (yield and environmental emissions). The data was collected from literature and refers to kiwi, watermelon, citrus, tea, and hazelnut crops in Guilan province in northern Iran. An expert in the field validated the estimation of pollutant emissions, calculated using Agri-footprint 4.0 and the updated version Agri-footprint 6. Additional key methodologies, standards and reports were also consulted for this research. The Decision Tree and Neural Network models developed were able to estimate the pollutant emissions generated into the environment throughout the production process. The results of the Absolute Normalized Error for the Decision Tree, Neural Network 1 and Neural Network 2 were 1124.79, 0.07 and 0.14 respectively. The Friedman test, with p-value <0.001, less than α=0.05, reveals statistically significant differences in the Absolute Normalized Error values in at least one of the models. The Wilcoxon test (p-value <0.001) indicates significant differences between all the models.

Keywords: Life Cycle Assessment, Sustainability, Machine Learning, Decision Trees, Data analysis

Abbreviations and Acronyms

ACV – Avaliação do Ciclo de Vida

ADP – Abiotic Depletion Potential

AI – Artificial Intelligence

ANN – Artificial Neural Network

AP – Acidification Potential

BUIDT – Bottom-up Induction Decision Trees

CART – Classification and Regression Trees

CV – Cross-validation

DT – Decision Tree

EEA – European Environment Agency

EU – European Union

EUP – Eutrophication Potential

FAEP – Freshwater Aquatic Ecotoxicity Potential

FAO – Food and Agriculture Organization

FYM – Farmyard manure

GHG – Greenhouse Gas

GW – Global warming

GWP – Global Warming Potential

HESTIA – Harmonized Environmental Storage and Tracking of the Impacts of Agriculture

HTP – Human Toxicity Potential

IA – Inteligência Artificial

ICV – Invetário do Ciclo de Vida

ID3 – Iterative Dichotomiser 3

INE – National Statistics Institute

IPCC – Intergovernmental Panel on Climate Change

IQR – Interquartile range

ISO – International Organization for Standardization

K-fertilizers – Potassium fertilizers

LCA – Life Cycle Assessment

LCI – Life Cycle Inventory

MAE - Mean Absolute Error

MAEP – Marine Aquatic Ecotoxicity Potential

ML – Machine Learning

MLP – Mullti-layer perceptron

MSE – Mean Squared Error

NAE – Normalized Absolute Error

N-fertilizers – Nitrogen fertilizers

NGOs – Non-Governmental Organizations

OLDP – Ozone Layer Depletion Potential

PEF – Product Environmental Footprint

P-fertilizers – Phosphorous fertilizers

PHOP – Photochemical Oxidation Potential

ReLU – Rectified Liner Unit

RETAILL – Retail using Technology based on Artificial Intelligence

RMSE – Root Mean Squared Error

R^2 – R-squared

SDGs – Sustainable Development Goals

TDIDT – Top Down Induction Decision Trees

TEP – Terrestrial Ecotoxicity Potential

UN – United Nations

Table of Contents

Figures

Tables

1. Introduction

Climate change, urbanization, rapid population growth, and recent changes in consumption patterns are all negatively contributing to achieving the United Nations (UN) Sustainable Development Goals (SDGs) (1). One of the primary objectives of Agenda 2030 is to ensure that production and consumption systems are sustainable, safeguarding the food security and nutrition of future generations (2,3). According to the SDGs, by 2030, food losses along production and supply chains should be halved (2,4). The Portuguese National Statistics Institute (INE) reported that, in 2021, more than 1.87 million tons of food were wasted in Portugal (5). The causes of food waste may be related to overproduction, characteristics of fresh products, transportation, storage, and retail processes (6).

Agriculture's contribution to the European Union (EU) Greenhouse Gas (GHG) emissions total has been estimated to account for as much as 11%, while global emissions represent 12% (7,8). The food supply chain may negatively impact the life-support system: it remains as a main contributor to the emissions of harmful air pollutants and it primarily contributes to deforestation, biodiversity loss, and unsustainable water extraction (8). As a result, the agri-food sector faces some challenges related to sustainability. These include efficient water resource management, reducing greenhouse gas and pollutants emissions, preserving biodiversity, and generally mitigating the negative consequences to human health. It should be highlighted that the food system, particularly agriculture, influences various aspects of human health. Agrochemicals (pesticides and fertilizers) and other pollutants in groundwater (such as manure) are often related to health problems, such as respiratory illnesses, waterborne diseases, and even mental health issues (8–10).

On the other hand, there are opportunities for the sector, such as technological innovation and AI to increase efficiency and sustainability, alongside the growing interest in learning more about product origin and the environmental impacts associated with food production. Many authors recognize that addressing food system sustainability requires a multifaceted approach and focusing on a single field or direction would be simplistic (10–12).

Sustainability is the practice of meeting present needs without compromising the ability of future generations to meet theirs. Sustainability also plays a crucial role, with increasing emphasis on minimizing environmental impact in supply chains.

Ensuring sustainability alongside profitability is essential for long-term success across different industries. This balance supports environmental health, economic viability, and social

equity, promoting a sustainable future for all. To assure a sustainable supply chain, various areas of action are needed, such as optimizing inventory management and routes, reducing ecological footprint in transportation, assessing the life cycle of food, monitoring environmental indicators, improving work methods, and increasing process efficiency (13,14). A key tool in achieving these goals is Life Cycle Assessment (LCA).

LCA evaluates the environmental impacts of all stages of a product's life cycle, from raw material extraction to production, use, and disposal. By integrating LCA into sustainability strategies, companies can identify opportunities to reduce environmental footprints, enhance resource efficiency, and improve overall performance, as sustainability and profitability are not just compatible but mutually reinforcing.

LCA requires extensive data collection of inputs and outputs such as raw materials, energy, water, used chemicals, and pollutants emissions at each life cycle stage. A common concern in many products LCA is the data collecting stage or Life Cycle Inventory (LCI), which may impact the study's overall quality level as it requires extensive datasets (15). Data is usually obtained from different sources like producers or farmers (primary data), literature reviews, government reports, and scientific publications (secondary data) or from associations, non-governmental organizations (NGOs), and international organizations (16).

Manufacturing processes often involve linked steps where the output of one process is the input of another. Changes in upstream processes can ripple through the supply chain, affecting downstream sustainability metrics. Often, not all necessary data is available or it may be incomplete, leading to gaps in the analysis. Furthermore, inaccurate data collection methods or reporting as well as different metrics and standards, can mislead the LCA process into a difficult aggregation and comparison. Countries and regions have contrasting characteristics and regulations, making it difficult to have a unified approach (17).

Artificial intelligence (AI) refers to computer systems capable of performing tasks that typically require human intelligence, such as learning from large datasets, pattern recognition, and decision-making. AI has a lot of potential, offering innovative solutions in many fields, namely agriculture (18). Applying AI to LCA represents an opportunity to automate data collection and analysis, improve the accuracy of environmental impact models, and facilitate the integration of large volumes of data from different sources. AI can also help identify patterns and trends that may not be easily perceptible through traditional methods.

Machine Learning (ML) is a subset of AI that involves the development of algorithms capable of learning from data and making predictions or decisions based on data (19). Unlike traditional programming, where explicit instructions are given, ML algorithms improve their performance as they are exposed to more data.

As the need for environmental protection becomes increasingly recognized worldwide, the application of ML in LCA has gained attention. Integration of AI in LCA has the potential to improve the accuracy and efficiency of these assessments. ML can help process large datasets, identify patterns, and make predictive analyses that traditional LCA methods might struggle with. LCA allows the understanding and mitigating of the environmental impacts of products, but it faces challenges due to its resource-intensive nature. In fact, despite its benefits, carrying out an LCA is often costly, time-consuming, and requires a lot of data. The adoption of AI and ML techniques and modern data technologies offers promising solutions to these challenges, enabling more efficient and accurate environmental assessments. As climate regulations push for greater environmental accountability, the integration of ML in LCA becomes increasingly important for driving innovation and sustainability (14,20,21).

Some examples of how ML can be used in LCA are the capacity of optimization, impact parameter forecasting, or missing data prediction. It can also be applied to address data incompleteness and uncertainty (21). The ability of ML to lower data-collecting costs is also one of the possible benefits of integrating ML in LCA. It provides a quantitative view of the environmental impacts of products throughout their entire life cycle. ML serves as a powerful tool to support and improve decision-making towards reducing the overall environmental footprint of products.

To enhance prediction models, some authors have studied the application of ML in LCA and showed that life cycle inventory, life cycle impact assessment, and interpretation were improved due to ML's ability to predict values and uncover hidden patterns accurately. However, the review also identified several challenges such as the size of training datasets. Additionally, there is a lack of detailed descriptions and established metrics and standards guidelines to evaluate these models' performance. Based on these findings one of the recommendations is to explore ML models in LCA studies and deeply integrate ML into different LCA stages to address complex environmental sustainability challenges (14,20,22).

This dissertation is conducted as part of the Retail using Technology based on Artificial Intelligence (RETAILL) project, which aims to make great strides in the development of a

software system that combines environmental and territorial data, as well as relevant data to support the decision-making of the different agents throughout the supply chain (23). Overall, the dissertation contributes to the RETAILL project by providing a framework that combines LCA and ML to promote sustainable agricultural practices and continuously improve production processes.

The present work represents a contribution to the field of sustainability in the supply chain, especially in the agricultural production sector, by providing relevant information to guide LCI and promote sustainability in the contemporary retail context. This dissertation seeks to answer the following research question: "Can Machine Learning techniques be applied to predict outcome variables of the LCI in LCA?".

The result of applying the models consists of, based on the inputs of the agricultural production, finding the emissions generated to the environment to complete the LCI. Selected environmental indicators include the amount of water and energy used, quantity of materials or food wasted, carbon dioxide (CO_2) emissions, and chemical usage (pesticides, fertilizers, and other elements).

1.1. Objectives and goals

The dissertation, integrated into the RETAILL project within the context of the Eureka Clusters – Sustainability Call (23,24) aims to contribute to the development of a decision support tool for choosing products with lower environmental impact. One of the RETAILL project goals is to develop the application of AI in LCA as a value-added solution. Based on these proposals, specific objectives have been established for this dissertation:

1. Provide an overview of LCA, LCA software and ML, detailing their principles, methodologies, and applications. This aims to contextualize the relationship between LCA and ML, explaining their significance in the dissertation's broader scope.
2. Develop ML models for emission estimation. This involves developing a Decision Tree (DT) model capable of estimating pollutant emissions from the production of kiwi, watermelon, citrus, tea, and hazelnut. It is intended for the model to support the Life Cycle Inventory (LCI) phase. Furthermore, building two Artificial Neural Network (ANN) models for the same purpose, providing alternative methods for emission estimation.

3. Compare model performance. Evaluate and compare the performance of the DT and ANN models to identify the optimal method for emission prediction in the context of the LCI phase.

As mentioned, specific objectives include the development of a Machine Learning model capable of translating the inputs of the agricultural production process into emissions to the environment. Essentially, the proposed framework will allow the compilation and processing of information during the LCI phase, the second phase in LCA. The framework aims to address data incompleteness, enhance the accuracy of emission estimations, and provide a more efficient and cost-effective approach to data
collection and analysis. Furthermore, the proposal model could be an alternative to commonly used software in LCI that lacks transparency and is not freely available.

To guide the work a hypothesis was defined: "DT and ANN models can estimate emissions, learn from data, and effectively predict data in the LCI phase?"

By addressing this hypothesis, the dissertation aims to demonstrate the feasibility and benefits of integrating ML into the LCI phase of LCA, ultimately contributing to more sustainable production practices.

The application of ML techniques to develop complementary and favorable models for LCA aims to enhance accuracy and efficiency in assessing environmental impacts. This facilitates informed decision-making on sustainable practices. Translating environmental indicators into measurable impacts on both the environment and human health reinforces the holistic approach of the circular economy, which considers not only economic aspects but also social and environmental impacts. The circular economy aims to minimize waste and maximize resource reuse. LCA for calculating the environmental impact of products aligns with the principles of circular economy. It allows the identification of stages where efficiency can be improved, waste reduced, and resources reused or recycled.

This dissertation is aligned with the principles of the circular economy by providing a framework that combines LCA and ML to promote sustainable agricultural practices, reduce waste, optimize resource use, and continuously improve production processes, integrating environmental and human health considerations into the product lifecycle.

Overall, the completion of this work can leverage future work within the RETAILL project and disseminate strategies to combat pollution and waste generated by agriculture, consequently

reducing the ecological footprint associated with the activity. Any action taken in this direction could improve the food quality of the produced goods, resulting in benefits for human health.
This dissertation contributes relevant information to the construction of a tool that allows combining environmental, territorial, and other relevant data in supporting decision-making by different agents throughout the supply chain.

1.2. Dissertation structure

This dissertation is organized into six main chapters, from fundamental concepts to the results analysis and conclusions. The work begins with an introduction that outlines the objectives and goals of the research, setting the stage by introducing the research question and providing an overview of the key issues being addressed.
Following the introduction that outlines the objectives and research questions, the background and state of the art provide a review of the relevant literature and the main subjects for this study. Essential concepts are described such as the product life cycle and life cycle assessment, focusing on the life cycle inventory. The second chapter will focus on sustainability concepts in the supply chain, and ML, providing context for the research.
Important ML techniques to the research include decision trees and neural networks. The section on decision trees is especially detailed, covering their advantages, disadvantages, and construction.
This chapter also covers data sources and specific approaches used for calculating agricultural emissions. This includes calculations for carbon dioxide emissions from various sources, water use, and heavy metal emissions.
The application of ML techniques is then described, beginning with descriptive statistics and data pre-processing, followed by model training focusing on decision trees and model performance evaluation.
The dissertation also addresses limitations encountered during the research process. Then the key findings are presented. These results are discussed in the context, exploring the implications of the findings. Finally, recommendations for future research, suggestions for further investigation, and habitual references.

2. Background and State of the Art

The following chapter includes the background context and the current state of the art in the fields relevant to this research. It begins with the concept of the product life cycle being examined to understand its stages. Following this, an exploration of LCA. In this chapter there is also a section about LCA software to give an overview of the existing tools and how they look or work. The chapter then covers a range of ML techniques, including decision trees, and neural networks, highlighting their potential applications in LCA.

2.1. Product life cycle

The product life cycle is illustrated in Figure 1. It consists of the following phases: raw material extraction, manufacturing and processing, packaging and distribution (transportation), retail and usage, and end-of-life that can result in disposal or recycling. The raw materials are not the only input in this process. In the remaining phases, other resources are needed, including human labor, energy, and water. During all the product life cycle phases pollutant emissions are generated to soil, air, and water (25).

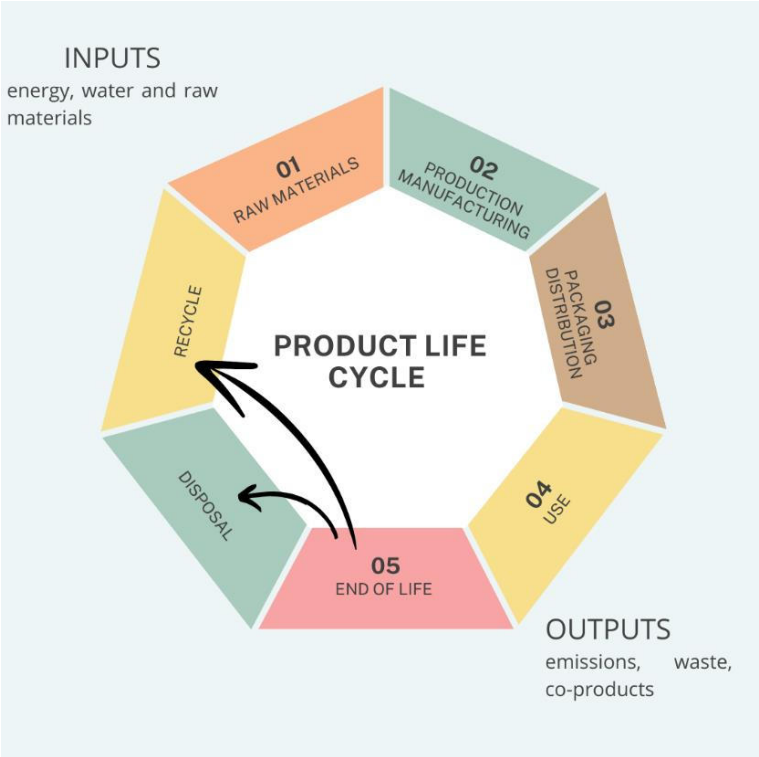

Figure 1. Product life cycle.

The European Climate and Health Observatory is a joint initiative of, among others, the European Commission and the European Environment Agency (EEA). Such organizations provide access to many data and resources related to climate change and health (26,27).

Different activity sectors, such as agriculture, industry, energy supply, transportation, and residential/commercial activities, have environmental impacts. Figure 2 illustrates how the main activity sectors have a negative impact thus affecting human health. The use of pesticides and fertilizers leads to contaminating water bodies. Overuse of chemicals degrades soil quality. Contaminated water sources can cause diseases. Air pollutants and chemical exposure can lead to respiratory issues or other health problems. Greenhouse gas emissions drive global warming causing extreme heat waves that increase morbidity. In addition, intensive use of natural resources (minerals, water, fossil fuels) leads to environmental degradation (28,29).

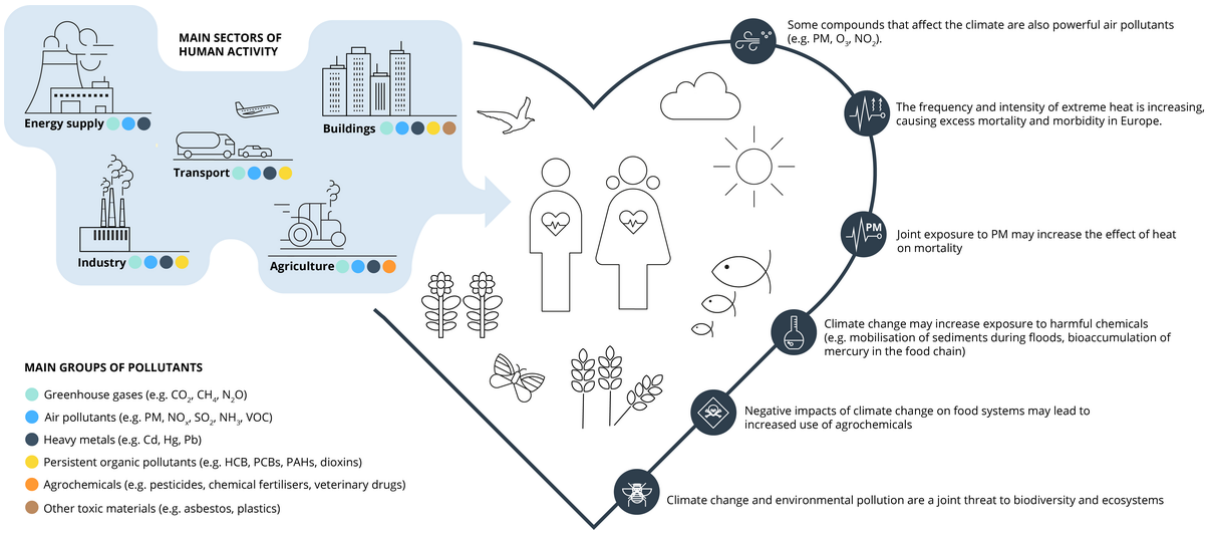

Figure 2. Intersection between pollution and environmental and human health impacts. Adapted from EEA (26).

As mentioned before, agriculture is one of the activities that has a negative impact on the environment and human health. Figure 3 illustrates how the pressures of each stage have consequences:

- Nitrogen and phosphorus fertilizers cause eutrophication of aquatic and terrestrial ecosystems.

- Ammonia emissions contribute to air pollution and are harmful to sensitive ecosystems.
- Climate change is influenced by greenhouse gas emissions, as mentioned above.
- Pesticide use has been associated with biodiversity and human health problems.
- The improper or excessive use of antibiotics in veterinary medicine is harmful.
- Overexploitation of freshwater resources is frequently associated with agriculture.

Figure 3 facilitates understanding LCA in agriculture because it visually illustrates the environmental pressures at various stages of farming processes. It highlights key sources of pollution and impact, such as water pollution, biodiversity loss, and climate change. By breaking down these processes it is possible to better identify the phases that require more attention when performing the study (29).

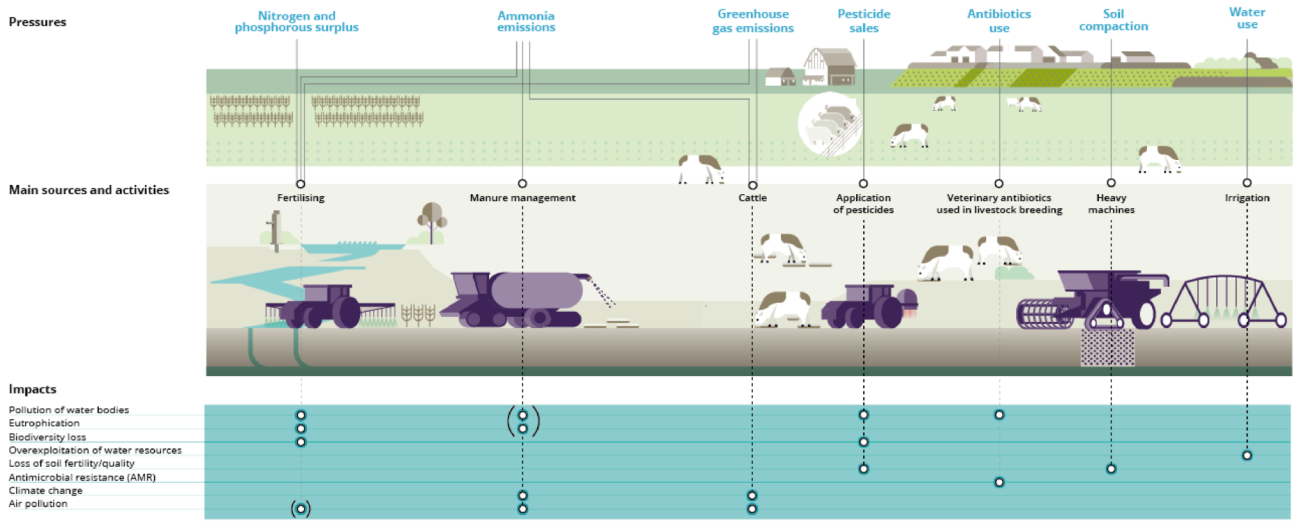

Figure 3. Agriculture impacts on the environment. Adapted from EEA (29).

Depending on the stages willing to study there are four common models: cradle-to-grave, cradle-to-gate, cradle-to-cradle, and gate-to-gate. Cradle-to-grave is when the full product life cycle is analyzed. Cradle-to-gate assesses the impact until the product leaves the industry and before it is transported to the retail or consumer. Cradle-to-cradle is often associated with the circular economy. It consists of replacing the waste stage with a recycling process (closed-loop recycling). Gate-to-gate is appropriate for product life cycles with many value-added processes. To reduce the complexity of the assessment, only one value-added process in the production chain is assessed. These assessments can later be linked, completing a larger-scale LCA (25).

2.2. Life Cycle Assessment

Life Cycle Assessment (LCA) is a quantitative tool used to measure the potential environmental impact of a product or service throughout its life cycle (25). Food systems include multiple stages, including activities before farming, the actual production of crops and feed, harvesting, transport, distribution, consumption and waste or disposal, following the same logic of the product life cycle illustrated in Figure 1 of the previous topic. LCA is used to assess the environmental impacts by considering all inputs and outputs in the entire process (25,30). When comparing LCA for different products, it is crucial to ensure that the same stages of the supply chain are included in all analyses.

Goal and scope definition, life cycle inventory, life cycle impact assessment, and interpretation are the four phases of LCA as shown in Figure 4. The purpose of the study, the expected product of the research, and the environmental impact categories are, among others, defined in the goal and scope phase (31). Due to the iterative nature of the process, the preliminary results are interpreted to help clarify the first three phases of the process that lead to the results (30).

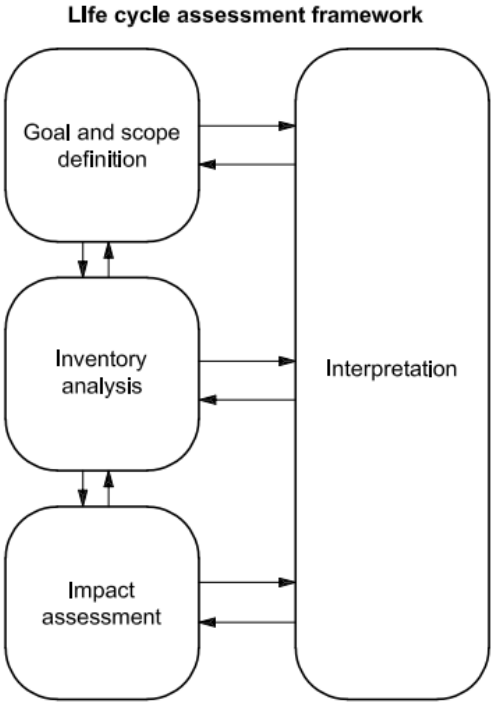

Figure 4. The four LCA stages. Adapted from ISO 14040 (31).

International Organization for Standardization (ISO) operates as an independent, non-governmental international organization that gathers experts worldwide to establish consensus on optimal practices, ranging from product manufacturing to process

management. The organization has different committees to develop standards and address the needs of different fields. ISO is dedicated to simplifying, enhancing safety, and improving the quality of life for everyone across the globe (32).

In the domain of environmental sustainability, ISO has published many standards namely about circular economy, water quality, and environmental management systems. The principles associated with LCA are found in ISO 14040 and ISO 14044. ISO 14040 defines LCA as the compilation of input and output flows and assessment of environmental impacts associated with a product throughout its life cycle (25). The ISO 14040:2006, *Environmental Management – Life Cycle Assessment – Principles and Framework* – contributes to the thirteenth Sustainable Development Goal: Climate action (33). This is explained by the common objective of integrating climate change measures and policies focusing on understanding the root of pollution and waste by assessing the life cycle of products, services, and processes. In 2020 a corrigenda was published: ISO 14040:2006/Amd 1:2020 (31).

The ISO 14044:2006, *Environmental Management – Life Cycle assessment – Requirements and guidelines* – also contributes to the thirteenth SDG, as well as to the twelfth (3). The norm provides essential tools and methodologies to guide LCA thus contributing to responsible consumption and production (SDG 12) and climate action (SDG 13) by promoting efficient resource use, reducing waste, identifying and mitigating GHG emissions, and fostering sustainable innovation. The most up-to-date amendment is ISO 14044:2006/Amd 2:2020 (34).

There are many knowledge gaps in LCA and ML can be used to fill in those gaps (35). Some of these gaps are a lack of data for specific temporal gaps or sectors, not standardized characterization factors, and computational power demand. In the impact assessment phase, supervised learning, such as regression or classification, can be used to estimate missing characterization factors or life cycle impacts. Another example is in the interpretation phase, unsupervised learning (as clustering algorithms) can be used to identify groups based on their sustainability performance (14).

2.2.1. LCA software tools

LCA software tools are widely used to facilitate the calculations and analyses involved in evaluating the environmental impacts of products, processes, or services. These tools may provide an advantage in terms of ease of use but often have built-in equations and databases

that are not free or disclosed to the user. Some common and recognized LCA software and databases are ecoinvent, SimaPro, and OpenLCA (36–38). The currently available environmental data is gathered, reviewed, and published regularly in LCI database – ecoinvent, that supports best practices in LCA as well as in database management.

Figure 5 illustrates the process of manufacturing an aluminum road bike. The first box indicates the manufacturing of the aluminum frame, where one unit of the frame is produced. This process flows into transport to the next production phase. Both the manufacturing and transportation processes feed into the production of the aluminum bike frame, where one unit of the frame is produced by combining the two inputs. After all the intermediate processes, the output moves into the final stage "My Road Bike" where the complete road bike is assembled, producing one unit of the bike. The "1 p" notation refers to one functional unit. The next step after outlining the model is to go to the process list to add library processes. Library processes can be for example adding tap water to the manufacturing and a container ship to the transport (39). In the context of agriculture, an example can be given with some fruit. Considering apple production, the first stage represents the cultivation of apple trees, where the trees are nurtured with water and nutrients, producing apples. This process flows into manual harvesting and the post-harvest phase. Harvesting processes feed into the grading and packaging of apples, where the yield of apples is prepared for sale. The final product moves to the consumption phase, where the apples are available to the consumer. This is a simplified example of what a production process in the agricultural sector might look like.

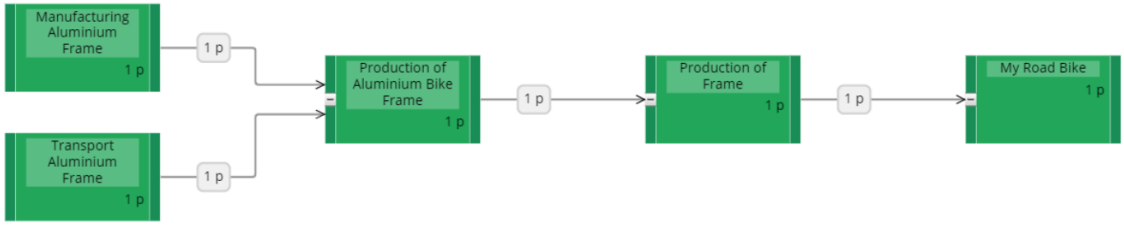

Figure 5. Preliminary structure of a model in SimaPro. Obtained from SimaPro (39).

Figure 6 is from the openLCA software, and it displays the inventory results for grape production under organic farming practices in the Languedoc-Roussillon region. Inputs section lists the resources consumed during the process. The outputs section shows the emissions and waste associated. Some output examples are ammonia emissions to air and phosphorus released to water. The "Total Requirements" section summarizes the resources required to produce one unit of the product, giving insights into the environmental impact of the production process (37). The inventory phase represented in Figure 6 is just one part of the overall LCA.

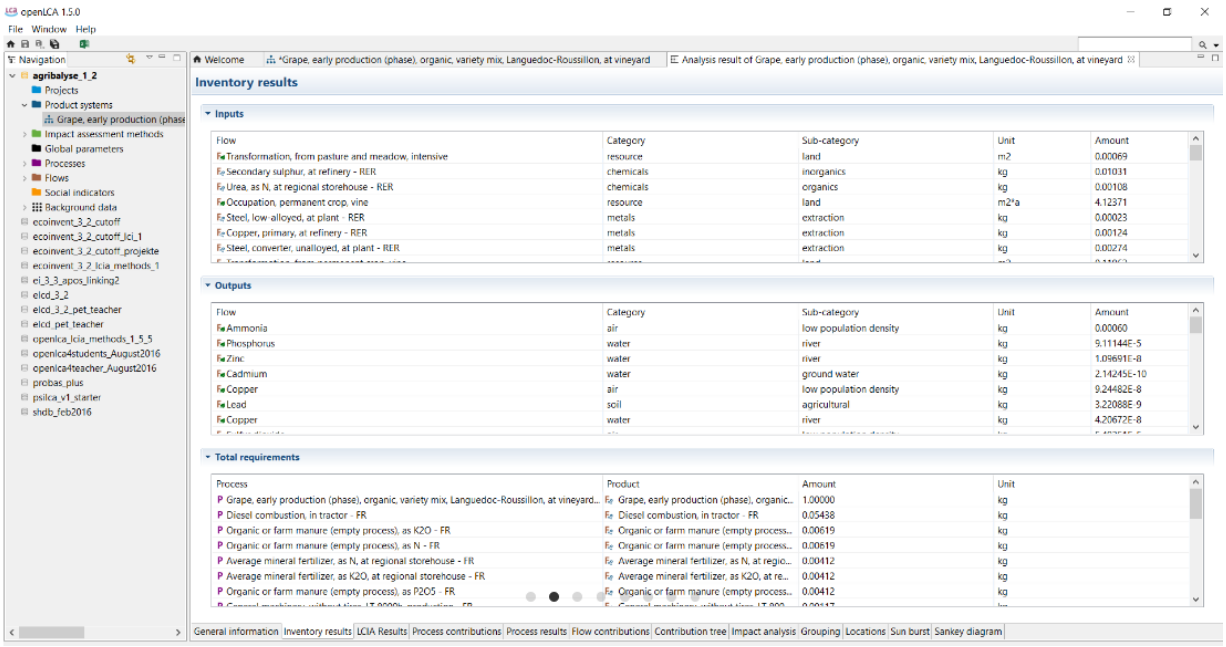

Figure 6. Inventory results in OpenLCA. Obtained from OpenLCA (37).

Recently some efforts have been made to standardize, harmonize, and aggregate environmental and LCI data. An open-source and open-access platform called the Harmonized Environmental Storage and Tracking of the Impacts of Agriculture (HESTIA) offers a vocabulary of words and a standard style for presenting agri-environmental data (40). Optimeal® is an example of a software tool developed to optimize human nutrition while taking sustainability and nutrient considerations into account (40,41). These examples show that the existing methods and tools have a narrow focus on a specific domain such as agriculture. Some were developed considering restricted geographical contexts.

The lack of transparency regarding the underlying equations and assumptions poses certain challenges. This increases the importance of displaying the known equations. Clear documentation of equations allows for easier identification and correction of any potential errors in the calculations. In section 2.4 the equations considered for this study are presented.

2.2.2. Life Cycle Inventory

The goal of inventory analysis or LCI is to gather all the data of the material and energy flows of the product life cycle. The inventory analysis involves identifying all the relevant inputs and outputs of the system of interest (42). For a corn production unit process, inputs might include seeds, water, fertilizers, machinery, and energy. Outputs include the corn product, and might also include nutrient losses from the fertilizers, and emissions from burning fuel to power

agricultural equipment. Taking inventory of all the inputs and outputs of every unit process can be a huge endeavor. Much work has already been done to build databases of common unit processes and their associated inputs and outputs. However, it is not appropriate to use the data without considering the scope, location, and time it was developed. These are important facts to consider when determining whether it is suitable for a particular LCA (20).

In the Life Cycle Impact Assessment the input and output flows identified in the LCI phase are translated to environmental impacts according to the impact categories defined. Some examples of these categories are eutrophication potential (EUP), abiotic depletion potential (ADP), acidification potential (AP), human toxicity potential (HTP), global warming potential (GWP), freshwater aquatic ecotoxicity potential (FAEP), marine aquatic ecotoxicity potential (MAEP), terrestrial ecotoxicity potential (TEP), photochemical oxidation potential (PHOP) and ozone layer depletion potential (OLDP) (43).

The last step of LCA is the interpretation of the results to find the environmental consequences resultants from production system, draw meaningful conclusions, and identify improvement opportunities (44).

2.3. Machine Learning Techniques

There has been a growing interest in harnessing the power of big datasets and ML techniques associated with LCA due to the rapid advancements in data analytics (14). ML is a branch of AI that comprises the study of computer algorithms that learn automatically through experience and can decipher the complexity of datasets, make predictions, and unveil new information and hidden dataset patterns (14). ML is defined as the ability to learn and improve computationally without the need for specific programming. ML algorithms can find patterns through the training phase on large data sets (19).

ML is a powerful tool for analyzing and interpreting complex data, making predictions, and automating decision-making processes. Building an effective ML model involves several crucial steps as illustrated in Figure 7. The first step is to ensure the raw data cleaning and transformation. The data must be relevant to the problem and contain sufficient information to train the model effectively. The next step is to understand the data distribution, identify patterns, detect outliers, and perform statistical analysis. Then the ML algorithms are selected based on the problem type (e.g., regression, classification, clustering) and the nature of the data. Evaluating multiple algorithms and adjusting parameters to identify the best-performing

model for the given dataset are the following steps. Finally, there is the model training phase and the evaluation of the model. It is important to continuously monitor the model's performance in the real world and make necessary adjustments (45).

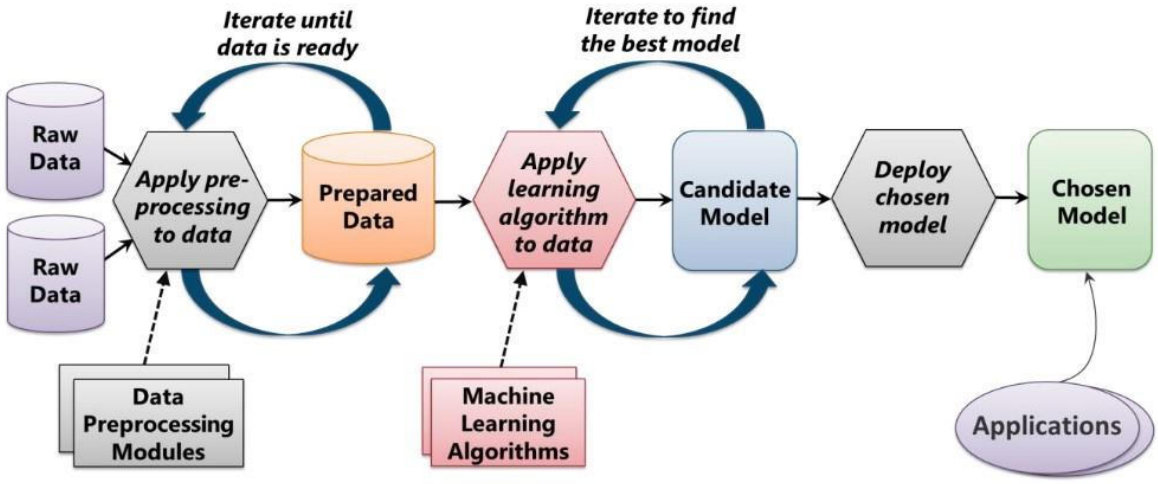

Figure 7. Key steps for ML models building. Obtained from Chappell D. (45).

Supervised learning and unsupervised learning are two categories of ML techniques. Through supervised learning, variables and measurable outcomes are connected in a way that enhances prediction accuracy. Unsupervised learning makes use of intrinsic features of the input datasets to identify patterns and trends without explicitly identifying the desired result (21).

There are several ML algorithms, such as neural networks, decision trees, linear regression, support vector machines, and algorithms for clustering, among others. These methods are extensively and effectively used in many fields, including public health, climate change, and food production.

After the training phase, the algorithms must be evaluated using performance metrics (14). Equations from 1 to 5 are some of the metrics to measure regression performance (46). The predicted value of the $i-th$ sample is $\hat{y}_i$ and y_i is the correspondent true value.

1. Mean Squared Error (MSE):

Measures the average squared difference between the actual and predicted values. When it has a 0 value it means the model has a perfect fit. MSE estimated over $n_{samples}$ is (46,47):

$$MSE\ (y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2 \qquad (1)$$

2. Root Mean Squared Error (RMSE):

RMSE is the square root of the MSE. The lower the RMSE, the better the model is. When it has a 0 value it means the model has a perfect fit. It is most suitable when the aim is to penalize extreme errors more than small ones. However, this makes it sensitive to these extreme values, which can distort the metric (46,47).

$$RMSE\ (\mathrm{y}, \hat{\mathrm{y}}) = \sqrt{\frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i - \hat{\mathrm{y}}_i)^2} \quad (2)$$

3. Mean Absolute Error (MAE):

Measures the average magnitude of the errors in a set of predictions, without considering their direction. When it has a 0 value it means the model has a perfect fit. MAE estimated over $n_{samples}$ is (46,47):

$$MAE\ (\mathrm{y}, \hat{\mathrm{y}}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} |y_i - \hat{\mathrm{y}}_i| \quad (3)$$

4. Normalized Absolute Error (NAE):

It normalizes the absolute error by dividing it by the error that would have occurred if the average of the target variable had been predicted every time. This approach provides a more balanced view of performance because it prevents models from being penalized for large values (46,47).

$$NAE\ (\mathrm{y}, \hat{\mathrm{y}}) = \frac{MAE(\mathrm{y}, \hat{\mathrm{y}})}{\frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} |y_i|} \quad (4)$$

5. R-squared (R^2), the coefficient of determination:

Indicates the proportion of variance in the dependent variable explained by independent variables. R^2 values range from 0 to 1, being 1 the perfect fit and 0 meaning that the model does not explain any variability in the dependent variable (46,47).

$$\mathrm{R}^2(\mathrm{y}, \hat{\mathrm{y}}) = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{\mathrm{y}}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{\mathrm{y}})^2} \quad (5)$$

Where $\bar{\mathrm{y}} = \frac{1}{n} \sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n} (y_i - \hat{\mathrm{y}}_i)^2 = \sum_{i=1}^{n} \epsilon_i{}^2$

These performance metrics help to understand the effectiveness of the ML models.

The ML process involves several stages, starting with raw data that often needs preprocessing before it can be used effectively. This preprocessing aims to clean and prepare the data for analysis. For example, the raw data might contain duplicates or missing

information, necessitating iterative preprocessing steps. This stage typically consumes a significant portion of the project timeline. Once the data is prepared, ML algorithms are applied to find the best solution for the problem at hand. Creating the optimal model involves running multiple experiments, trying various combinations of algorithms, and preparing data until the best-performing model is found. Once an effective model is developed, deployment is crucial as it allows the algorithm implemented by the model to be utilized. Moreover, the ability to deploy new models quickly is essential in adapting to the rapidly changing landscape of Machine Learning applications (45).

To enhance prediction models, LCA researchers have taken advantage of the potential provided by the advancement of ML and the capacity to generate those models and identify a significant pattern from a large amount of data (21).

Some authors have studied the application of ML in LCA. Romeiko et al. (20) tested and compared five ML approaches for computing spatially and temporally explicit life cycle environmental impacts at the county scale, focusing on corn production in the Midwest region as a case study. The aim was to determine the ML model that provides accuracy and efficiency for estimating corn's spatially and temporally life cycle environmental impacts (20). The same author also has a review article that explored the combined use of ML and LCA for quantitative sustainability assessment. The conclusion showed that life cycle inventory, life cycle impact assessment, and interpretation were improved due to ML's ability to predict values and uncover hidden patterns accurately. However, the review also identified several challenges such as the size of training datasets. Additionally, there is a lack of detailed descriptions and established metrics and standards guidelines to evaluate these models' performance. Based on these findings one of the recommendations is to explore ML models in LCA studies and deeply integrate ML into different LCA stages to address complex environmental sustainability challenges (14). Nevertheless, previous studies predominately focused on specific technologies in specific industries (22).

Combining ML techniques with LCA may be a helpful strategy to improve the process and overcome some LCA constraints. ML models can estimate outputs faster and with less storage space than a traditional LCA method. They are also more adaptable when included in simulation platforms and can perform more simulation runs and produce better results for various computationally intensive tasks (21).

2.3.1. Decision Tree

As a non-parametric supervised learning method, decision trees can be used for classification and regression. A classification tree is used to predict the class to which the data belongs. A regression tree is used to predict a real number outcome. Decision trees create a set of if-then-else decision rules by learning from data (48,49).

A Decision Tree (DT) is a group of nodes that resembles a tree. It is a hierarchical structure with nodes and directed edges, used to determine whether values belong in a class or to estimate a target value (50,51). A decision tree has root nodes – with no incoming edges, internal nodes – with exactly one incoming edge and two or more outgoing edges - and leaf or terminal nodes that have precisely one incoming edge and no outgoing edges (52).

Every node represents a different attribute's splitting rule. This rule divides data into classes for classification and divides values into regression groups to minimize the error for the chosen parameter criterion. Until the stopping requirements are satisfied, additional nodes are constructed (51).

Briefly, DT allows for the iterative division of data into subsets based on specific attributes, intending to make predictions or classifications.

Figure 8 shows a representative example of a DT built with the used dataset. Root, node, and leaves are marked and labeled. The product parameter is the root node.

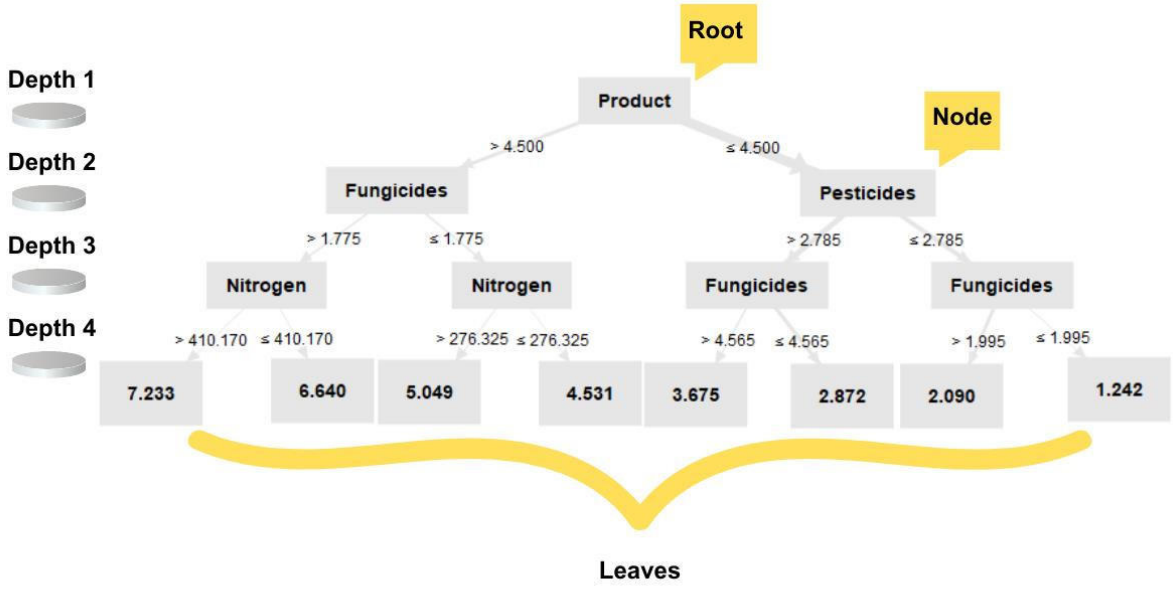

Figure 8. Representative decision tree structure example.

2.3.1.1 Advantages

One of the advantages of decision trees is that they can work with both categorical and numerical data as well as manage problems with more than one output. Furthermore, when a condition is observable in a model, Boolean logic explains the state. Decision trees can be visualized and are simple to interpret. On the other hand, outcomes in a black box model (such as an artificial neural network) can be harder to understand (48,53).

When a DT is used to predict data a logarithmic cost in the number of data points used to train the tree is associated (49).

Statistical tests are often used to validate DT models. This enables the model's dependability to be taken into consideration. Performs well even when the true model used to create the data slightly deviates from its assumptions (48).

2.3.1.2 Disadvantages

One of the common risks when building a DT is having an overfitted model. To prevent it, pruning can be applied. Setting the maximum depth of the tree is also a solution (54).

Training several trees in an ensemble learner, where the features and samples are randomly sampled with replacement, can help to face algorithms that make locally optimum decisions at each node (greedy algorithm). To avoid biased trees, It is advised to balance the dataset before fitting the decision tree (48).

2.3.1.3 Building a Decision Tree

There are several ways to construct decision trees, and each has benefits and drawbacks based on the kind of data and the issue at hand. The efficiency and interpretability of the model can be greatly impacted by selecting the appropriate methodology, which ranges from top-down induction and pruning to methods like random forests and boosting (51,53).

Top Down Induction Decision Trees (TDIDT) are constructed in a descendent and recursive way. This method involves starting from the root of the tree and recursively dividing the data into more homogeneous subsets, based on an attribute selection criterion (55,56).

Bottom-up Induction Decision Trees (BUIDT), a less common approach, where the construction starts from the leaves and gradually merges nodes based on a defined criterion. It creates parent nodes from merged leaf nodes until a single root node is formed (55–57). To

avoid overfitting, decision trees can be pruned. Pruning involves removing parts of the tree that provide little predictive power. There are two types of pruning:

- Pre-Pruning – Halt the tree growth early based on criteria like maximum depth, minimum number of samples per node, or minimum gain in information. When applying pre-pruning, trees do not experience a full set, the model stops splitting nodes early, preventing overgrowth and possibly reducing the risk of overfitting. However, this might also mean that the tree sacrifices some accuracy in exchange for simplicity and faster processing (57).
- Post-Pruning – Fully grow the tree and then remove nodes that do not improve accuracy. Examples include Reduced Error Pruning, where nodes are pruned if removing them does not adversely affect the model's performance on a validation set (55–57).

Building a DT involves several steps from collecting data to evaluating the model.

Selecting a dataset with the relevant information to the study is the first step. Followed by cleaning data and ensuring data quality. Another thing to consider is splitting the dataset into training and testing sets.

Using Cross-validation (CV) to build a decision tree may be a better alternative compared to simply splitting the dataset into training and testing sets. CV provides a more thorough and reliable method for model evaluation. It leverages the entire dataset more effectively, helps detect and prevent overfitting, and offers a more nuanced understanding of model performance across different data partitions (58).

When preprocessing data, generally there is the need to normalize or standardize features, though decision trees are not as sensitive to this step as other models.

Several algorithms can be used. Common algorithms are Iterative Dichotomiser 3 (ID3) and Classification and Regression Trees (CART) (55,59,60). The algorithms will be described in more detail subsequently in this dissertation.

As mentioned above, to avoid some problems that might come up when building a DT, some parameters must be set: maximal depth, and the minimum number of samples required to split an internal node and the leaf node, among others.

After training and testing comes the evaluation of the model using metrics such as accuracy, precision, recall, and F1-score for classification problems. For regression problems metrics

such as MSE, RMSE, MAE, NMAE and R^2 are more suitable. Chapter 2.3 goes into further detail on these metrics.

Finally comes the visualization of the model and the understanding and interpretation.

2.3.1.4 Tree algorithms

Ross Quinlan created the algorithm Iterative Dichotomiser (ID3) to create decision trees from an information gain method. With this algorithm, it is expected to have a tree growing iteratively by selecting features that offer the greatest reduction in entropy or uncertainty (61). The process continues until a prerequisite is met, such as the minimum size of the subset or the maximum tree depth. ID3 is considered a heuristic algorithm, based on Occam's razor [1], where small trees are preferred (59,62).

The advantages of this algorithm are that ID3 uses the entire dataset to build the tree model, creating an understandable set of prediction rules. Furthermore, identifying the leaf nodes makes it possible to prune the test data and lower the number of tests. The linear function of the product of the characteristic number and node number determines the ID3 time. ID3 disadvantages are associated with time consumption and computing demand, due to one attribute being tested individually (49,59).

The C4.5 algorithm was also developed by Ross Quinlan and it comes up as an improvement over the ID3. It can handle numerical (continuous) variables by creating threshold values that split the data into intervals. C4.5 can handle missing attribute values by using fractional counts based on the distribution of observed values. Post-pruning is done by evaluating the rules generated and removing preconditions if their removal improves rule accuracy. Despite the advantages, C4.5 also has disadvantages: it is sensitive to small training sets and small data variations can generate different decision trees (49).

C5.0 improves on C4.5 by using less memory, creating smaller rule sets, and increasing accuracy allowing it to become faster and memory usage and computational power efficient (49).

CART, developed by Leo Breinman, builds binary decision trees, meaning each node splits the data into two branches. Unlike C4.5, CART is suitable for both regression and classification problems (60).

[1] A principle of logic that recommends keeping things as simple as possible.

CART uses cost-complexity pruning to remove branches that provide little additional predictive power (49).

Random Forest, trademarked by Leo Breinman and Adele Cutler, creates a set of decision trees. It is suitable for both classification and regression problems. Every tree considers only a random subset of features to ensure a low correlation among decision trees. Each tree is considered to establish the output. Due to the structure of a Random Forest, the Interpretation is difficulted and it is susceptible to overfitting (49,63,64).

2.3.1.5 Splitting Criterion

In the realm of DT algorithms, the splitting criterion plays a pivotal role in determining how the data is partitioned at each node to achieve the most informative splits. This subsection delves into various splitting criteria, including entropy, information gain, gain ratio, Gini index, and least squares. Each criterion offers an approach to evaluating the quality of a split, influencing the overall performance and accuracy of the DT. By understanding these methods, it is possible to better appreciate their application. The presented equations were all retrieved from the authors Tan et al. (50) and Faria (65).

2.3.1.5.1 Entropy

Entropy is a measure of uncertainty or randomness. It quantifies the impurity of a set of examples. In other words, entropy is a measure of the purity of an interval. In decision trees, entropy is used to determine the best split. Lower entropy indicates a more homogenous subset, so if an interval contains only values of one class it is perfectly pure and entropy is 0. Given a k number of various class labels, m_i the number of values in the i^{th} interval of a partition and m_{ij} being the number of elements of class j in interval i, entropy is defined by (50):

$$Entropy(e_i) = -\sum_{i=1}^{k} p_{ij} log_2 p_{ij} \tag{6}$$

where $p_{ij} = m_{ij}/m_i$ is the probability of class j in the i^{th} interval. The total entropy, e, of the partition, is the weighted average of the individual interval entropies:

$$e = \sum_{i=1}^{n} w_i e_i \tag{7}$$

2.3.1.5.2 Information Gain

Information gain measures the reduction in entropy achieved by partitioning the data according to a given attribute. Decision tree algorithms like ID3 choose the attribute with the highest information gain for splitting.

How to calculate (65):

$$IG(S,A) = Entropy(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} entropy(S_v) \quad (8)$$

Where S_v is the subset of S where attribute A has the v value.

2.3.1.5.3 Gain Ratio

The gain ratio is an extension of information gain that adjusts for the number and size of branches when choosing an attribute. It reduces the bias towards attributes with many values. The C4.5 decision tree algorithm is used to select the best attribute for splitting.

How to calculate (46):

$$Gain\ Ratio\ (S,A) = \frac{IG(S,A)}{Split\ Info(A)} \quad (9)$$

Where,

$$SplitInfo(A)\ = \sum_{v \in values(A)} \frac{|S_v|}{|S|} \log_2 \frac{|S_v|}{|S|} \quad (10)$$

2.3.1.5.4 Gini Index

The Gini index measures impurity or diversity used in classification trees. It represents the probability of a randomly chosen element being misclassified if it was randomly labeled according to the distribution of labels in the subset. Used in the CART algorithm to decide on splits. A lower Gini index indicates a purer node.

Let $p\ (i \mid t)$ denote the fraction of records belonging to class i at a given node t. The reference to node t is sometimes omitted, and the fraction is expressed as pi. How to calculate (50):

$$Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2 \quad (11)$$

2.3.1.5.5 Least squares method

The least squares method is used to select the best attribute for splitting a node in a decision tree, minimizing the sum of squared errors between the predicted values and the real values. During the construction of the tree, the algorithm selects the splitting attribute at each node. This step is crucial and affects the model's accuracy. The goal is to choose the attribute that, when splitting the data, minimizes the sum of squared errors between the predicted values and the real values.

For each possible split based on an attribute, the algorithm calculates the sum of the squared errors between each real value and the average of the values in the node. The attribute with the smallest sum of squared errors is chosen to split the node. This means that the chosen split will, on average, result in more accurate predicted values.

$$SSE = i = \sum_{i=1}^{n} (yi - (mx_i + b))^2 \quad (12)$$

2.3.2. Neural Networks

The human brain and neurons inspire the Artificial Neural Network (ANN) structure as it is a structure of artificial neurons. They are recognized as a reliable decision support tool capable of increasing LCA accuracy (21).

Human brain dendrites (receptors) and axons (effectors) connect a network of neurons. Dendrites collect signals from numerous neurons and combine them to produce a response signal which is then distributed along axons. A neuron is a node in ANN terminology. The biological term synapse is a connection and synaptic efficiency is identified as connection strength or weight. (66). This simple explanation of how the human brain functions helps in better understanding ANNs.

ANNs contain three layers: the input layer representing the information entering the neuron, the hidden layers which are intermediate layers and the output layer, reflective of the outcome information (65).

Several algorithms can be used to build ANNs. The Neural Net Operator in RapidMiner learns a model using a feedforward neural network trained by a backpropagation algorithm (multi-layer perceptron) to train the ANN (67). In a feedforward neural network, connections between neurons do not form cycles. Data flows in one direction, from the input layer through the hidden layers to the output layer. It means that the network processes the input data step by step,

layer by layer until it reaches the output. The backpropagation algorithm works by adjusting the weights of the connections between the neurons to minimize the error between the predicted output and the actual output. It involves calculating the gradient of the loss function and propagating the error backward through the network to update the weights. The two stages of the backpropagation algorithm, a supervised learning technique, are propagation and weight updating. The two stages are iterated until the network reaches an acceptable level of performance. algorithm. The algorithm makes minor adjustments to the weights of each connection based on this information to lower the error function's value (44,66–69).

Multi-layer perceptron (MLP) refers to a specific type of feedforward ANN with multiple neurons, allowing the network to perform more complex learning. Every node, except the input nodes, is a neuron, or processing element, with a nonlinear activation function.

The activation function is part of the ANN and transforms an input into a certain output, activating and deactivating neurons. Therefore, the ANN is largely determined by the weight and input-output of the activation function. One of the artificial neural networks that uses the activation function is the backpropagation method (44,67).

ANN's hidden layers represent intermediate neurons between the input and output layers. It allows the ANN to learn complex patterns and relationships between the input data and the output, capturing non-linear interactions. Each neuron in the hidden layer passes a representative weight through an activation function, such as Rectified Liner Unit (ReLU), Sigmoid, or Tanh. ReLU introduces non-linearity to the model. Without non-linear activation functions, neural networks would behave like linear models, limiting their ability to learn complex functions(70). The Sigmoid activation function transforms the input to a value between 0 and 1. It is useful when facing a binary classification model or probabilistic predictions. Hyperbolic Tangent (Tanh) is similar to Sigmoid but the output values range from -1 and 1. Despite having a steeper gradient and due to its bounded output value the gradient tends to vanish with this function (71).

In deep learning, multiple hidden layers can learn hierarchical representations of the data. The first hidden layers may learn simple features, while deeper layers capture more abstract features. Without hidden layers, a neural network could only model linear relationships between inputs and outputs (66,72,73).

Figure 9 represents a feedforward neural network structure. In the input layers, the circles correspond to the dataset features. Three hidden layers are fully connected to every node in

the previous layer. The lines connecting nodes represent the weights that adjust during the learning process. As it represents a regressing problem, predicting a continuous variable, the output layer is a single node.

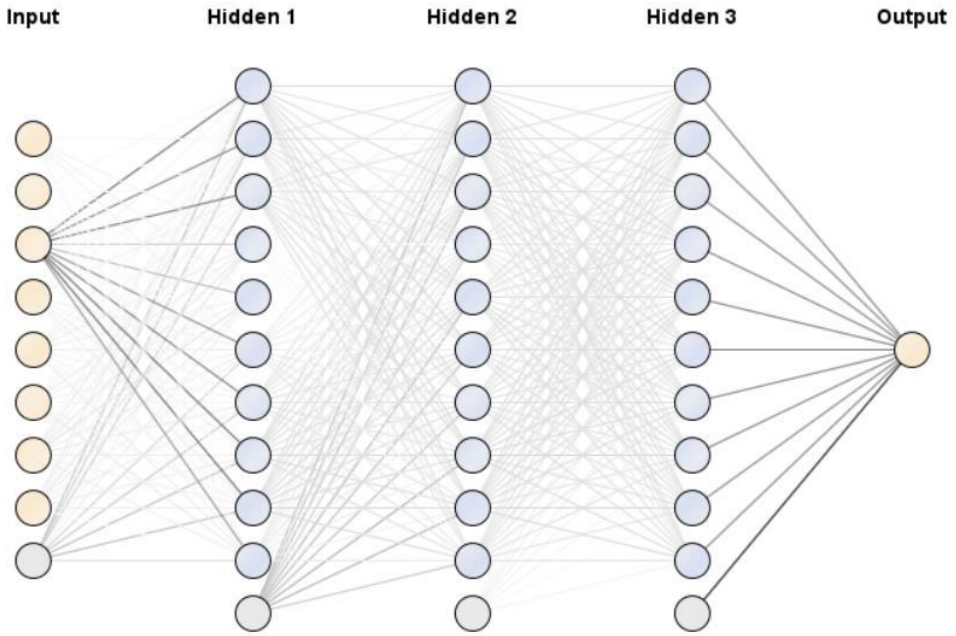

Figure 9. ANN structure example using the study dataset.

While ML is important to process and interpret complex datasets, understanding emissions in agricultural contexts requires a specific approach, grounded in consolidated methodology. Transitioning from ML, the following subchapters address important concepts in Life Cycle Assessment, key emission factors, and methodologies, particularly those associated with the Agri-footprint database.

2.4. Emissions calculation methodologies in agriculture

According to the Portuguese Environment Agency's 2022 National Emissions Inventory, the agriculture sector accounts for 12.2% of Portugal's national GHG emissions. Examples of greenhouse gases are carbon dioxide (CO_2), methane (CH_4), and nitrogen oxide (N_2O) (74).

The EEA stated that, in 2020, agriculture was the main source of ammonia and methane emissions (28). Agriculture activities represent emissions related to chemical fertilizer usage to air and water, and biocide application to air, water, and soil. The computation of each emission involves multiplying the input amounts by their equivalent coefficients or, in some cases, more complex equations. According to Agri-footprint, specific equations are used to

calculate emissions from agricultural activities, which are then translated to potential environmental impacts (75).

2.4.1. Agri-footprint database methodology and basic principles

The support tool is Agri-footprint 6 Methodology Report (75) designed to support LCA practitioners in assessing the environmental impacts of agricultural products. Agri-footprint is a consistent LCI database widely used by the LCA community, the scientific community, and governments worldwide. It comprises data on several industries such as food and biomaterials from more than 63 countries. Agri-footprint shows calculated impact results using the ReCiPe 2016 Midpoint methodology and many processes modeled in alignment with the Product Environmental Footprint (PEF) guidance of the European Commission, which gives it more credibility. There are different paid license options to access Agri-footprint including research, commercial or developer licenses. Agri-footprint is also available in LCA software like SimaPro, which is also a paid software, and in openLCA that, despite being free open-source software for LCA, Agri-footprint is available in openLCA as a paid database (76,77).

Figure 10 illustrates Agri-footprint, which gathers data from different sources, including Food and Agriculture Organization (FAO), scientific publications, industry data, and LCA standards. The background data to build the tool includes information on electricity, waste, heat, and fuel consumption.

Agri-footprint includes over 4800 datasets, mostly country-specific. It provides three allocation options for impact assessment: economic, mass-based, and energy-based. The tool also integrates land use change data to provide a more complete picture. Although the associated software is paid, the background methodology is publicly available, reinforcing transparency. Additionally, the tool is updated every two years to maintain accuracy and relevance.

Figure 10. Scheme of Agri-footprint tool. Obtained from Mérieux NutriSciences | Blonk *(76)*.

According to Agri-footprint 6 Methodology Report the fertilizer inputs, manure inputs, and crop residues generate different emissions to the environment. Table 1 shows which inputs are associated with each emission (75).

Table 1. Chemical and manure inputs and associated emissions. Adapted from Agri-footprint 6 (15).

Emission	Fertilizer	Manure	Crop Residues	Pesticides
Ammonia (NH_3)	Yes	Yes	No	Not directly
Nitrate (NO_3)	Yes	Yes	Yes	Not directly
Carbon dioxide (CO_2)	Yes	-	-	-
Nitrogen monoxide (NO)	Yes	Yes	No	Not directly
Phosphorus	Yes	Yes	No	Not directly
Heavy metals	Yes	Yes	Yes	Not directly
Methane (CH_4)	No	Yes	Yes	No

2.4.2. Carbon dioxide (CO_2) emissions – Lime, dolomite, and urea application

Fertilizers usually enhance agricultural productivity, but their use also contributes to greenhouse gas emissions, particularly carbon dioxide (CO_2). Certain compounds, such as limestone, dolomite, and urea, commonly used as fertilizers or soil amendments, undergo chemical reactions that release CO_2. Accurately estimating these emissions is important for understanding the environmental impact of agricultural practices (75). Carbon dioxide

emissions from lime, dolomite, and urea compounds that are used as fertilizers, can be calculated by the following equation:

$$CO_2 - C_{em} = (M_{Limestone} * EF_{Limestone}) + (M_{Dolomite} * EF_{Dolomite}) + (M_{Urea} * EF_{Urea}) \quad (13)$$

$CO_2 - C_{em}$ is the emissions in kg C. The M factors represent the fertilizers amounts in kg. EF are the emission factors, calculated by the following equation:

$$\frac{kg\ C}{kg\ of\ limestone,dolomite\ and\ urea} \quad (14)$$

2.4.3. Nitrous oxide (N_2O) emissions

Agriculture activities lead to N_2O emissions. The equation to calculate N_2O emissions from nitrogen (N) inputs is:

$$N_2O - N_{Ninputs} = (F_{SN} + F_{ON} + F_{CR} + F_{SOM}) * EF_1 \quad (15)$$

$N_2O - N_{Ninputs}$ = Annual direct N_2O -N emissions from N inputs to managed soils, [Kg N_2O - N]

F_{SN} is the amount of synthetic fertilizer N applied to soils; F_{ON} is the amount of animal manure, compost, sewage sludge and other organic N additions applied to soils; F_{CR} is the amount of N in crop residues (above-ground and below-ground), including N-fixing crops (leguminous), and from forage/pasture renewal, returned to soils; F_{SOM} is the amount of N in mineral soils that is mineralized, in association with loss of soil C from soil organic matter as a result of changes to land use or management. All these are expressed in kg N. EF_1 is the emission factor for N_2O emissions from N inputs, which has a default value of 0.01. The amount of N_2O emitted through atmospheric deposition is calculated with the equation below:

$$N_2O_{(ATD)} - N = [(F_{SN} * Frac_{GASF}) + ((F_{ON} + F_{PRP}) * Frac_{GASM})] * EF_4 \quad (16)$$

EF_4 is the emission factor for N_2O emissions from atmospheric deposition of N on soils and water surfaces.

2.4.4. Ammonia (NH_3) and Nitrate (NO^-_3) emissions

Ammonia (NH_3) emissions from agriculture are an environmental concern due to their contribution to air pollution and eutrophication in water bodies. When nitrogen-based fertilizers, whether synthetic or organic, are applied to fields, a portion of the nitrogen volatilizes into the atmosphere as ammonia. Understanding and quantifying NH_3 emissions

from managed soils is crucial for developing effective strategies to reduce nitrogen losses and mitigate their environmental impact. NH_3 emissions:

$$\mathrm{NH_3}-N = (F_{SN} * Frac_{GASF}) + ((F_{ON} + F_{PRP}) * Frac_{GASM}) \quad (17)$$

$\mathrm{NH_3}-N$ is the ammonia produced from atmospheric deposition of N volatilized from managed soils in KG NH_3-N.

$Frac_{GASF}$ is the fraction of synthetic fertilizer N that volatilizes as NH_3 and NOx.

$Frac_{GASM}$ is the fraction of applied organic N fertilizer (F_{ON}) and urine and dung N deposited by grazing animals (F_{PRP}) that volatilizes as NH_3 and NOx.

Nitrate (NO_3^-) emissions to water:

$$NO_3^- - N = (F_{SN} + F_{ON} + F_{PRP} + F_{CR} + F_{SOM}) * Frac_{LEACH-(H)_{PRP}} * Frac_{WET} \quad (18)$$

$NO_3^- - N$ is the nitrate produced from leaching pf N from managed soils in kg.

$Frac_{LEACH-(H)}$ is the fraction of all N added to / mineralized in managed soils where leaching or runoff occurs that is lost through leaching and runoff.

2.4.5. Water use

During crop cultivation, not all the applied irrigated water is consumed. The unutilized water either evaporates, drains into the soil, or runs off, contributing to inefficiencies in water use. Agri-footprint 6 and ReCiPe Characterization report (77) are useful to estimate the actual water consumption which is calculated in m^3/ha using the following equation:

$$Blue\ water\ footprint\left(\frac{m3}{ton}\right) * Yield\ crop\left(\frac{ton}{ha}\right) * water\ requirement\ ratio \quad (19)$$

Blue water is reported as "Water, unspecified natural origin" refers to the amount of freshwater used for irrigation that originates from surface or groundwater sources.

3. Methodology

This section outlines the methodology and approach taken to achieve the research objectives. It further delves into the methods for calculating emissions, including specific calculations for carbon dioxide (CO_2) emissions, diesel fuel combustion, and chemical fertilizer use. It describes emission sources and their corresponding calculation equations. Additionally, the section explores the environmental indicators and variables considered, providing a framework for assessing the environmental impact within the scope of the research.

The main sources of information and relevant academic research are detailed and explained, as they are the primary sources of information for the background of this study.

3.1. Data source

The publicly available online data was found through Mostashari's (78) study, which focuses on data collection and analysis in Guilan province, northern Iran. This study will be referred as the Iran case ahead in the present dissertation. Primary data, including agricultural inputs and crop yields, were collected for horticultural crops such as hazelnut, citrus, tea, kiwifruit, and watermelon.

The research employs the LCA approach, which evaluates environmental impacts throughout a product's life cycle. It begins with goal and scope definition, considering factors like functional units and system boundaries. Various environmental impacts, including emissions from agricultural activities and fertilizer usage, are calculated for producing 10 tons of horticultural crops.

The dataset used in this study was obtained from the supplementary data of the Iran case. Inventory data consists of On-orchard and Off-orchard emissions. According to the authors of the referenced article, Off-orchard emissions were obtained from the foreground system, detailing the types and amounts of materials used and the direct energy consumption in horticultural crop production. On-orchard emissions were sourced from the background system, including direct emissions from diesel fuel combustion in agricultural machinery, air emissions from human labor activities, and emissions to soil, water, and air from fertilizers. The authors gathered these emissions from ecoinvent®3.6. Additionally, biocide emissions were derived using the PestLCI 2.0 model in Analytica software (79).

3.2. Emissions calculation methodology

Different emissions may contribute to the same impact but be expressed in different units. Conversion to one unit is very important so that the emissions are translated into one impact category (80). The environmental impact categories are discriminated into the regulations used for this work. In this study, inputs of the system comprise the utilization of water, energy (electricity), agricultural machinery, fertilizers, diesel fuel and biocides. The horticultural crops amount are the outputs: citrus, kiwi, watermelon, hazelnut and tea (78,79).

To enhance transparency, reproducibility and credibility the further topics present the emissions calculation equations that support this study. These methodologies ensure consistency, in line with the latest research and guidelines provided by the referenced databases and institutions. This approach allows for a more precise assessment of environmental impacts, particularly regarding emissions from agricultural practices.

3.2.1. Equations for emissions calculation

To measure environmental impacts it is important to know the relevant equations used in the context of LCA. Based on Equations 15 to 18 in Chapter 2.4, some variables were calculated with variations of those equations according to the available data and a specialist's approval. All the equations were thoroughly referenced and retrieved from Agri-Footprint 4.0, Agri-Footprint 6 and Intergovernmental Panel on Climate Change (IPCC) (75,81,82). All the emissions were computed using the measuring unit from Iran's case which is kilogram per hectare (kg/ha).

For instance, databases like ecoinvent3.6 use normalization factors. The most recent versions of ecoinvent or other databases now include the normalization factors. An illustrative example is the calculation of N_2O emissions from fertilizers to the air (expressed in kg). This is computed using the following formula:

$$Nitrogen + FYM * 0.01 * \left(\frac{44}{28}\right) \quad (20)$$

In this formula, the value 0.01 represents the emission factor. The fraction $44/28$ is the conversion factor, translating the nitrogen-based emission into its equivalent weight of N_2O. Equation 20 illustrates how N_2O emissions from fertilizers to the air were calculated in this work.

3.2.1.1 Carbon dioxide (CO_2) emissions – Human Labor

Human labor activities, particularly those involving physical work, contribute to carbon dioxide (CO_2) emissions through metabolic processes and use energy-consuming tools and machinery. Emissions of carbon dioxide (CO_2) related to human labor are calculated by multiplying the input (hours of human labor) by the coefficient found in the literature, which has a 0,7 value (78,83,84). Equation 21 presents the formula to calculate CO_2 emissions associated with human labor.

$$CO_2\ Human\ Labor = human\ labor\ (hours) * 0{,}7 \quad (21)$$

3.2.1.2 Diesel fuel combustion emissions

The combustion of diesel fuel in agricultural processes results in several emissions to air and soil that can be calculated from standard coefficients of ecoinvent®3.6 (38). These coefficients are also presented in Mostashari-Rad et al. and the emissions are calculated by multiplying the corresponding input and coefficient (78).

3.2.1.3 Chemical fertilizer, farmyard manure (FYM), pesticide, herbicide, and fungicide emissions

This subchapter focuses on the direct N_2O emissions associated with nitrogen (N) from applying synthetic fertilizers, manure and compost, crop residues, and changes in soil organic matter due to land-use or management changes. The calculation of these emissions relies on standardized equations. Ammonia (NH_3) and nitrate (NO_3^-) emissions Equations 17 and 18 are presented in chapter 2.4.4. The components of the fertilizers used on Iran's crops are not known. However, considering the most used components and the possibility of future work, Equations 13 and 14 for calculating the CO_2 emissions associated with these components are presented in Chapter 2 but they will not be used during the study.

3.2.1.4 Nitrous oxide (N_2O) emissions, Ammonia (NH_3) and Nitrate (NO^-_3) emissions

For this work, only direct N_2O emissions were considered because data does not include livestock activities. Changes in mineral soil management and emissions from urine and dung inputs to grazed soils are not considered in this study, as it is assumed that crops are cultivated

on cropland remaining cropland, not grazed, and the organic matter contents of the soils do not substantially change.

Some authors simplify the equation depending on the elements available for the study. In this case, only the available variables will be considered (78).

It was assumed that all nitrogen that volatizes converts to ammonia (NH_3), and that all nitrogen that leaches is emitted as nitrate (NO_3^-) (75).

3.2.1.5 Heavy metal emissions

The heavy metals are released into the atmosphere because of the leaching and removal of biomass, as well as their deposition and output from the application of fertilizer and manure.

Applying fertilizers and manure, as well as deposition, contributes heavy metals to the soil. Based on published research, the heavy metal concentration of manure and fertilizers was determined, as shown in Tables 3-2 and 3-3 of the Agri-footprint 4.0 report. Table 3-4 of the same report lists the deposition of heavy metals (82,85).

The Agri-footprint 4.0 separates mineral fertilizers according to their composition. Table 3-2 of that report includes the heavy metal content of fertilizers in nitrogen fertilizers (N-fertilizers), phosphorous fertilizers (P-fertilizers), and potassium fertilizers (K-fertilizers). This distinction is not verified in the most recent version of Agri-footprint. Furthermore, version 4.0 is followed by the authors of Iran's case, so it is preferably used for the heavy metal emissions calculation.

3.3. Software, Programming Languages and AI-Assisted Technologies

For this work it was used RapidMiner version 9.10.011, Microsoft®Excel® for Microsoft 365 MSO (version 2408 Build 16. 0. 17928. 20114, 64-bit) and IBM SPSS Statistics version 29.0.1.0 -171) (86–88). Preliminary analyses of variables were conducted using RapidMiner and SPSS. Subsequently, the ML models were developed using RapidMiner. Excel was used to organize and process the data and to obtain the results related to emissions.

AI-assisted technologies were used in the writing process to improve readability, grammar, and spelling. Applying the technology is not to replace key authoring tasks and it is done with human oversight and control. All work is afterward reviewed and edited because AI can generate incorrect, incomplete, or biased content.

During the preparation of this work QuillBot, ChatGPT, and Grammarly were used (89–91). ChatGPT was primarily used to guide research on the following topics: product life cycle, LCA, LCI, ML techniques, including decision trees and neural networks. It was also employed to clarify questions on RapidMiner use. ChatGPT was also helpful with generating alternative word synonyms and offering general advice for writing. QuillBot was useful for rewriting sentences and to enhance the flow of text and make it sound more natural. Grammarly, installed in Microsoft Word for Windows shows a list of suggestions by category: grammar and spelling checks, refining written content by identifying errors, and suggesting improvements in tone, punctuation, and clarity. It can also ensure that the work is free of common writing mistakes.

3.4. Data pre-processing

The data targeted for the study by the authors Mostashari-Rad et al. (79) is available online in Excel format. The file features a sheet for each product, and each orchard corresponds to a row, recording values for each of the variables: human labor, agricultural machinery, lubricating oil, nitrogen, phosphate, potassium, farmyard manure (FYM), pesticides, fungicides, diesel fuel, electricity, yield. All variables are expressed in kilograms (kg), except human labor, presented in hours (h), and electricity in kilowatts (kWh), per hectare.

To proceed with the data analysis, original data were grouped, meaning all the products were combined in a unique sheet, and a new variable "Product" was created. It encoded as follows: citrus to 1, kiwi to 2, watermelon to 3, hazelnut to 4, and tea to 5.

Each variable was assigned the letter and number corresponding to the emissions it represents. This was done considering Table 6 of Mostashari-Rad et al. (78). Table 2 shows the descriptions for each emission to which the variables are associated.

Table 2. Abbreviation of variables corresponding to emissions.

Name/abbreviation	Description
B.1.	Emissions by diesel fuel to air
B.2.	Emissions by diesel fuel to soil
B.3.	Emissions by fertilizers to air
B.4.	Emissions by atmospheric deposition of fertilizers to air
B.5.	Emissions by fertilizers to water

B.6.	Emissions by N_2O of fertilizers to soil and air
B.7.	Emissions by heavy metals of fertilizers to soil
B.8.	Emissions by human labor to air
B.9.	Emissions by biocides to air
B.10.	Emissions by biocides to water
B.11.	Emissions by biocides to soil

Biocides emission variables – B.9, B.10 and B.11 – were not considered for this work. On one hand, the information on Pesticide Model in Appendix II of Agri-footprint 6 about the active ingredients of pesticides covers 80% of the substances most used per pesticide supergroup adjusted based on three countries – France, Netherlands and the United States – most used substances (75). On the other hand, this coverage was not considered enough to calculate all the biocide emissions listed in Iran's case because there are substances in the dataset that are not listed in the Pesticide Model, such as benomyl (75). Despite being available on FAO statistics relevant information on the use of pesticides, it was not possible to find details on the amount of active ingredients of each pesticide supergroup used per hectare of the cultivated crop (92).

Table 3 is relevant to describe the LCI emissions variables, associated with various sources such as diesel fuel combustion, fertilizer use, and human labor. In this table, the different pollution sources are organized according to their origin. This table also separates the variables by emissions to air, water, and soil.

Table 3. LCI variables and the corresponding description.

LCI variable	Emissions to air
	From human labor
CO_2 (B.8)	Carbon dioxide emissions
	From diesel fuel combustion/use
NH_3 (B.1)	Ammonia
C_6H_6 (B.1)	Benzene
Benzo(a)pyrene (B.1)	Benzo(a)pyrene, a polycyclic aromatic hydrocarbon (PAH)
Cd (B.1)	Cadmium
CO_2fossil (B.1)	Carbon dioxide
CO fossil (B.1)	Carbon monoxide

Cr (B.1)	Chromium
Cu (B.1)	Copper
N_2O (B.1)	Nitrous oxide
Heat waste (B.1)	Waste heat
CH_4 fossil (B.1)	Methane
Ni (B.1)	Nickel
Nox (B.1)	Nitrogen oxide
NMVOC (B.1)	Non-methane volatile organic compounds
PAH (B.1)	Polycyclic aromatic hydrocarbons
Particulates < 2.5um (B.1)	Particulate matter emissions with a diameter less than 2.5 micrometers
Se (B.1)	Selenium
SO2 (B.1)	Sulfur dioxide
Zn (B.1)	Zinc
	From the use of fertilizers
N_2O (B.3)	Nitrous oxide
NH_3 by chemical fertilizers (B.3)	Ammonia emissions from the use of chemical fertilizers
NH_3 by FYM (B.3)	Ammonia emissions from the use of FYM
N_2O by chemical fertilizers (B.4)	Nitrous oxide emissions from the use of chemical fertilizers
N_2O by FYM (B.4)	Nitrous oxide emissions from the use of FYM
	Emissions to water
Nitrate (B.5)	Nitrate emissions from fertilizer usage
	Emissions to soil
	From diesel fuel combustion/use
Cd (B.2)	Cadmium
Pb (B.2)	Lead
Zn (B.2)	Zinc
	From heavy metals of fertilizers
Cd (B.7)	Cadmium
Cu (B.7)	Copper
Zn (B.7)	Zinc
Pb (B.7)	Lead
Ni (B.7)	Nickel

Cr (B.7)	Chromium
Hg (B.7)	Mercury

3.5. Data analysis and model training

In this subchapter, the building of DT and ANN models using RapidMiner is explained. It begins with a brief introduction of how the work started and then proceeds to describe the application of CV, which provides a more robust understanding of the DT model performance by ensuring that every data point is used both for training and testing. Parameters such as maximum depth, and pruning options are also pointed out in this subsection. Finally, subchapter 3.5.2 describes the process of building two different Artificial Neural Networks.

3.5.1. Decision tree

To develop the DT and guarantee that all the data was used for training and testing CV was applied. In k-fold cross-validation, the dataset is divided into k subsets (folds). The model is trained on k-1 folds and tested on the remaining folds, with this process repeated k times, using each fold as a test set exactly once. This ensures that every data point gets to be in the training and testing set exactly once, giving a greater understanding of how the model performs on unseen data. The average of all testing accuracies is the result. Following this structured approach, a decision tree model was constructed to predict the values of the variables presented in Table 3.

Necessary parameters in the Decision Tree operator such as the maximum depth set between 6 and 10, and no pruning options were defined, considering domain knowledge and experimentation for optimization. The following step was to apply the trained decision tree model to the testing set.

A DT was created to predict the values for each variable representing the emissions generated by the inputs in cultivation. The attributes of the DT were the variables human labor, agricultural machinery, lubricating oil, nitrogen, potassium, FYM, pesticides, fungicides, and diesel fuel.

By using the same decision tree model but varying the label, different depths of the tree were tested. It was found that between 6 and 10 the performance values no longer differed greatly, so a smaller depth value was chosen. Less depth was preferable to avoid overfitting the model.

3.5.2. Artificial Neural Network

Two distinct Neural Network models were developed using RapidMiner Studio. Both processes were built using CV. The main difference between the two models is the operator: one uses the Neural Net Operator and the other uses the Deep Learning operator. The model using Neural Net operator will be called Neural Network 1 and the one using Deep Learning operator will be called Neural Network 2.

The Neural Net operator is used to build a feedforward neural network trained by a backpropagation algorithm (multi-layer perceptron) (67). One hidden layer was defined using 865 training cycles. Decay and normalize options were selected. Without a fixed seed, in every run, even with the same dataset and parameters, the results might slightly differ due to the randomness in the training process. The random seed was set to 1000.

When applying the Deep Learning operator, key settings include 10 epochs, adaptive learning rate, and standardization. The model is trained using backpropagation, with hidden layer sizes and Rectifier activation function. A non-default parameter is the hidden layer sizes, where six layers are used, each with 50 neurons. The option of computing variable importances is chosen so that the variables that are most influential in the model's predictions (93). The model performance is then evaluated. The workflow utilizes Cross-validation to ensure robust model assessment.

3.6. Non-parametric tests for paired samples

Statistical tests were applied to compare the results of the errors obtained with the different models and for each variable. The results of the NAE of the two ANN and the DT models across all the variables were analyzed.

The application of parametric tests involves checking the normal distribution and the homogeneity of the variance of the data set. The tests used to verify normality are the Shapiro-Wilk test (for small samples) and the Kolmogorov-Smirnov test. On the other hand, Levene's test is one of the most powerful tests for homogeneity of variance (87).

The Friedman test is a non-parametric test used to compare three or more groups of paired data, for example with measurements from the same sample under different conditions. It is an alternative when the assumptions of ANOVA are not met. The hypotheses of the Friedman test are:

1. The distributions of the values of variable X are identical in the k populations (94):

$$H_0: F(X_1) = F(X_2) = \cdots = F(X_k) \quad (22)$$

2. There is at least one distribution that is different from the rest (94):

$$H_1: \exists\, i, j: F(X_i) \neq F(X_j)\ (i \neq j;\ i, j = 1, \ldots k) \quad (23)$$

To verify if there were differences in the distribution of values, the Friedman test was used in the SPSS software (version indicated in chapter 3.3), and α=0.05.

The Wilcoxon test can be used as a non-parametric alternative to the Student's T-test in cases where the normal distribution is not verified or the robustness of parametric methods cannot be defended (94). The Wilcoxon test makes it possible to formulate hypotheses about the distribution $F(X)$ of variable X in 2 paired samples. The hypothesis for this work can be written as follows:

$$H_0: F(X_1) = F(X_2) \text{ vs } H_1: F(X_1) \neq F(X_2) \quad (24)$$

In this work, the Wilcoxon test was applied to find which pairs of models showed distribution differences, using the SPSS software (version indicated in chapter 3.3), and α=0.05.

4. Results

This chapter presents the findings of the present work, including the comparative analysis between a Decision Tree and two Neural Networks (ANN1 and ANN2).

4.1. Data characterization

Table 4 includes a descriptive statistical analysis. Each variable, ranging from human labor to yield, is characterized by essential statistic metrics including mean, median, standard deviation, quartiles, minimum, and maximum values, derived from a dataset comprising 865 observations (N).

Table 4. Descriptive statistics of the input variables.

Variable	Mean	Standard Deviation	Median	First quartile (25%)	Third quartile (75%)	Minimum	Maximum
Human labor	2.23E+03	2.77E+03	6.90E+02	3.41E+02	4.57E+03	1.15E+02	8.92E+03
Agricultural machinery	2.26E+01	1.55E+01	1.64E+01	1.12E+01	2.89E+01	6.20E+00	7.09E+01
Lubricating oil	3.68E+00	1.86E+00	3.10E+00	2.15E+00	4.99E+00	1.23E+00	8.86E+00
Nitrogen	2.30E+02	1.55E+02	2.40E+02	1.10E+02	3.43E+02	9.80E+00	5.70E+02
Phosphate	8.40E+01	4.62E+01	9.32E+01	2.38E+01	1.20E+02	1.23E+01	1.68E+02
Potassium	1.21E+02	1.01E+02	8.30E+01	5.16E+01	1.39E+02	1.70E+01	3.88E+02
FYM	1.02E+03	8.72E+02	4.54E+02	3.49E+02	1.80E+03	2.28E+02	3.11E+03
Pesticides	2.60E+00	1.06E+00	2.41E+00	1.73E+00	3.36E+00	1.00E+00	5.36E+00
Fungicides	2.99E+00	1.23E+00	2.71E+00	2.01E+00	3.91E+00	1.14E+00	6.15E+00
Diesel fuel	5.23E+01	3.82E+01	4.25E+01	1.87E+01	8.30E+01	5.35E+00	1.32E+02
Electricity	2.31E+02	1.01E+02	2.14E+02	1.42E+02	3.13E+02	8.81E+01	4.53E+02
Yield	1.54E+04	1.03E+04	1.56E+04	8.19E+03	2.43E+04	2.86E+02	3.66E+04

[1] Missing values: 230 in Phosphate and 410 in Electricity

The descriptive statistics presented in Tables 5 to 9 provide an overview of the inputs used to produce five different products:

1. Citrus;
2. Kiwi;
3. Watermelon;
4. Hazelnut
5. Tea.

Statistical analysis reveals differences between products in terms of mean, median, standard deviation, minimum, maximum, and input variables employed in each production process. A complex relationship between input and yield emerges when comparing the five products. Kiwi and watermelon have high yields, but watermelon production reaches an approximate value with fewer inputs, suggesting greater efficiency. Tea is the most labor-intensive but does not produce a proportionally higher yield, which may indicate a less efficient process. Hazelnut production, on the other hand, is the simplest in terms of inputs and yield, reflecting a low-intensity production process comparatively with the other product's production required inputs.

Table 5 shows the descriptive statistics of the input variables and yield for citrus presenting a total of 195 observations (N). The mean value for human labor is approximately 451 hours, with a standard deviation of 92.5 hours, indicating moderate variability in labor usage across orchards. The average yield is 22300 kg/ha, with a standard deviation of 4.64 kg/ha representing the variation in production.

In terms of nitrogen fertilizer, farms use an average of 140 kg per hectare, with a standard deviation of 27.7 kg.

Phosphate application shows a mean of 99.2 kg/ha with a standard deviation of 20.9 kg, and the distribution is centered around a median of 98.7 kg/ha.

For potassium, the average application is 290 kg/ha, with a standard deviation of 61.5 kg/ha, reflecting higher variability when compared to other variables. The minimum application registered is 187 kg/ha and the maximum is 388 kg/ha.

FYM application demonstrated more variability when compared to other products used. The average is 1840 kg/ha with a standard deviation of 393 kg/ha.

The analysis of Table 5 suggests that the yield is influenced by several factors that vary between samples.

Table 5. Descriptive statistics of the input variables and yield for citrus.

Variable	Mean	Standard Deviation	Median	First quartile (25%)	Third quartile (75%)	Minimum	Maximum
Human labor	4.51E+02	9.25E+01	4.44E+02	3.75E+02	5.35E+02	2.90E+02	6.01E+02
Agricultural machinery	2.02E+01	3.73E+00	2.00E+01	1.71E+01	2.32E+01	1.30E+01	2.71E+01
Lubricating oil	4.15E+00	8.14E-01	4.10E+00	3.45E+00	4.79E+00	2.71E+00	5.62E+00
Nitrogen	1.40E+02	2.77E+01	1.39E+02	1.18E+02	1.63E+02	9.04E+01	1.89E+02
Phosphate	9.92E+01	2.09E+01	9.87E+01	7.96E+01	1.18E+02	6.50E+01	1.35E+02
Potassium	2.90E+02	6.15E+01	2.86E+02	2.37E+02	3.48E+02	1.87E+02	3.88E+02
FYM	1.84E+03	3.93E+02	1.82E+03	1.46E+03	2.20E+03	1.18E+03	2.47E+03
Pesticides	2.75E+00	5.43E-01	2.71E+00	2.32E+00	3.22E+00	1.78E+00	3.68E+00
Fungicides	3.14E+00	6.58E-01	3.11E+00	2.59E+00	3.65E+00	2.02E+00	4.22E+00
Diesel fuel	9.08E+01	1.84E+01	8.99E+01	7.41E+01	1.07E+02	5.88E+01	1.22E+02
Electricity	1.36E+02	2.71E+01	1.34E+02	1.15E+02	1.58E+02	8.81E+01	1.82E+02
Yield	2.23E+04	4.64E+03	2.23E+04	1.81E+04	2.66E+04	1.45E+04	3.02E+04

[1] Missing values: 0

Table 6 presents the descriptive statistics of the input variables and yield for kiwi for 140 observations (N). The average yield of 22900 kg/ha is close to the yield of product 1, but the higher number of inputs suggests that the production of this product requires more resources, which may indicate greater complexity or production intensity. The average human labor required in kiwi production is equivalent to 1450 hours, more than three times the needed for citrus production. Table 6 presents higher mean values in the majority of variables (some examples are human labor, agricultural machinery, lubricating oil, diesel fuel, nitrogen and phosphate) when compared to citrus production.

Table 6. Descriptive statistics of the input variables and yield for kiwi.

Variable	Mean	Standard Deviation	Median	First quartile (25%)	Third quartile (75%)	Minimum	Maximum
Human labor	1.45E+03	2.95E+02	1.44E+03	1.19E+03	1.71E+03	9.38E+02	1.94E+03
Agricultural machinery	3.38E+01	6.95E+00	3.42E+01	2.72E+01	3.97E+01	2.22E+01	4.61E+01
Lubricating oil	5.16E+00	1.06E+00	5.22E+00	4.24E+00	6.09E+00	3.36E+00	7.03E+00
Nitrogen	2.59E+02	5.04E+01	2.57E+02	2.19E+02	3.01E+02	1.71E+02	3.49E+02
Phosphate	1.11E+02	2.26E+01	1.10E+02	9.13E+01	1.31E+02	7.33E+01	1.48E+02
Potassium	1.04E+02	2.21E+01	1.04E+02	8.24E+01	1.25E+02	7.02E+01	1.39E+02
FYM	2.33E+03	4.76E+02	2.31E+03	1.89E+03	2.75E+03	1.53E+03	3.11E+03
Pesticides	3.15E+00	6.39E-01	3.14E+00	2.64E+00	3.75E+00	2.03E+00	4.23E+00
Fungicides	3.62E+00	7.28E-01	3.58E+00	3.03E+00	4.28E+00	2.35E+00	4.85E+00
Diesel fuel	9.87E+01	1.98E+01	9.88E+01	8.12E+01	1.16E+02	6.41E+01	1.32E+02
Electricity	3.38E+02	6.95E+01	3.38E+02	2.83E+02	3.96E+02	2.19E+02	4.53E+02
Yield	2.29E+04	4.65E+03	2.28E+04	1.92E+04	2.67E+04	1.48E+04	3.10E+04

[1] Missing values: 0

Table 7 shows the descriptive statistics of the input variables and yield for watermelon, with a total of 120 observations (N).

While pesticides have low variability, evidenced by a relatively small standard deviation (0.36) and an Interquartile Range (IQR) of 0.67, other inputs like human labor, nitrogen, and FYM show higher variability. For example, the standard deviation of nitrogen is 89.4 kg/ha, which implies differences in how farmers manage nitrogen use in watermelon production.

Nitrogen and FYM show comparatively higher means and standard deviations, suggesting that these inputs vary among the dataset. While inputs like pesticides have lower variability and relatively smaller ranges, other factors like human labor, fertilizers, and yield show greater dispersion.

Table 7. Descriptive statistics of the input variables and yield for watermelon.

Variable	Mean	Standard Deviation	Median	First quartile (25%)	Third quartile (75%)	Minimum	Maximum
Human labor	7.13E+02	1.39E+02	7.06E+02	5.98E+02	8.42E+02	4.60E+02	9.52E+02
Agricultural machinery	5.24E+01	1.12E+01	5.23E+01	4.24E+01	6.24E+01	3.39E+01	7.09E+01
Lubricating oil	6.59E+00	1.37E+00	6.57E+00	5.38E+00	7.92E+00	4.25E+00	8.86E+00
Nitrogen	4.26E+02	8.94E+01	4.22E+02	3.39E+02	5.05E+02	2.75E+02	5.70E+02
Phosphate	1.26E+02	2.55E+01	1.25E+02	1.04E+02	1.49E+02	8.07E+01	1.68E+02
Potassium	1.13E+02	2.34E+01	1.13E+02	9.02E+01	1.33E+02	7.35E+01	1.54E+02
FYM	3.87E+02	7.49E+01	3.84E+02	3.23E+02	4.56E+02	2.54E+02	5.16E+02
Pesticides	1.75E+00	3.64E-01	1.76E+00	1.39E+00	2.06E+00	1.14E+00	2.37E+00
Fungicides	2.00E+00	4.18E-01	2.00E+00	1.64E+00	2.37E+00	1.29E+00	2.68E+00
Diesel fuel	5.47E+01	1.13E+01	5.44E+01	4.55E+01	6.53E+01	3.56E+01	7.33E+01
Electricity	2.60E+02	5.51E+01	2.59E+02	2.09E+02	3.05E+02	1.72E+02	3.51E+02
Yield	2.74E+04	5.76E+03	2.71E+04	2.26E+04	3.27E+04	1.76E+04	3.66E+04

[1] Missing values: 0

Table 8 represents the values of descriptive statistics of the input variables and yield for hazelnut production. A total of 180 observations were registered for all variables in Table 8, except electricity which is not used for hazelnut production. Among the crops analyzed, hazelnut production is the least labor-intensive, with an average of 175 hours. The low demand for other inputs, such as nitrogen (mean is approximately 14.8 kg/ha) which is substantially lower than for watermelon (426 kg/ha) and kiwi (259 kg/ha) and FYM (mean is approximately 385 kg/ha), indicates a less intensive production process. Hazelnut farming still requires pesticide and fungicide application. The average yield is 450 kg/ha, the lowest among all products, with relatively low variability. This suggests that the cultivation process is simpler or less resource-intensive but generates significantly lower output.

Table 8. Descriptive statistics of the input variables and yield for hazelnut.

Variable	Mean	Standard Deviation	Median	First quartile (25%)	Third quartile (75%)	Minimum	Maximum
Human labor	1.75E+02	3.65E+01	1.77E+02	1.40E+02	2.01E+02	1.15E+02	2.38E+02
Agricultural machinery	9.44E+00	1.99E+00	9.48E+00	7.57E+00	1.12E+01	6.20E+00	1.28E+01
Lubricating oil	1.89E+00	3.87E-01	1.88E+00	1.54E+00	2.24E+00	1.23E+00	2.53E+00
Nitrogen	1.48E+01	3.14E+00	1.50E+01	1.16E+01	1.73E+01	9.80E+00	2.03E+01
Phosphate	1.89E+01	3.95E+00	1.90E+01	1.54E+01	2.24E+01	1.23E+01	2.57E+01
Potassium	2.59E+01	5.40E+00	2.59E+01	2.11E+01	3.09E+01	1.70E+01	3.49E+01
FYM	3.85E+02	7.96E+01	3.79E+02	3.15E+02	4.60E+02	2.49E+02	5.15E+02
Pesticides	3.94E+00	7.84E-01	3.98E+00	3.31E+00	4.61E+00	2.59E+00	5.36E+00
Fungicides	4.52E+00	9.64E-01	4.52E+00	3.68E+00	5.31E+00	2.93E+00	6.15E+00
Diesel fuel	8.18E+00	1.74E+00	8.26E+00	6.68E+00	9.62E+00	5.35E+00	1.11E+01
Yield	4.50E+02	9.29E+01	4.44E+02	3.76E+02	5.41E+02	2.86E+02	5.96E+02

[1] Missing values: 180 in Electricity

Table 9 represents the values of descriptive statistics of the input variables and yield for tea production. The total number of observations (N) is 230.

Tea production requires high labor input, with an average of 6620 hours, 1370 hours standard deviation and small IQR that shows the similarity among labor force required. This table represents a production system with a relatively high demand for human labor, compared with products 1,3 and 4. However, the variation in inputs suggests some impact in those results as the presented tea orchard does not require phosphate or electricity.

Yield averages 10524.34 kg/ha, with a standard deviation of 2060.98 kg/ha. The interquartile range (8851.97 – 12250.15 kg/ha) indicates that most of the production values are concentrated in this range, with a minimum of 6840.05 and a maximum of 14078.48 kg/ha, reflecting moderate variability in productivity.

Table 9. Descriptive statistics of the input variables and yield for tea.

Variable	Mean	Standard Deviation	Median	First quartile (25%)	Third quartile (75%)	Minimum	Maximum
Human labor	6.62E+03	1.37E+03	6.62E+03	5.40E+03	7.79E+03	4.30E+03	8.92E+03
Agricultural machinery	1.25E+01	2.67E+00	1.26E+01	1.00E+01	1.49E+01	8.16E+00	1.70E+01
Lubricating oil	2.28E+00	4.54E-01	2.29E+00	1.89E+00	2.67E+00	1.49E+00	3.08E+00
Nitrogen	3.55E+02	7.14E+01	3.55E+02	2.93E+02	4.12E+02	2.32E+02	4.79E+02
Potassium	6.64E+01	1.30E+01	6.69E+01	5.52E+01	7.68E+01	4.35E+01	9.05E+01
FYM	3.51E+02	6.99E+01	3.50E+02	2.91E+02	4.10E+02	2.28E+02	4.72E+02
Pesticides	1.54E+00	3.15E-01	1.56E+00	1.27E+00	1.81E+00	1.00E+00	2.12E+00
Fungicides	1.78E+00	3.58E-01	1.78E+00	1.48E+00	2.09E+00	1.14E+00	2.35E+00
Diesel fuel	2.47E+01	5.05E+00	2.45E+01	2.03E+01	2.92E+01	1.59E+01	3.31E+01
Yield	1.05E+04	2.06E+03	1.04E+04	8,85E+03	1,23E+04	6.84E+03	1.41E+04

[1] Missing values: 230 in Phosphate and Electricity

There is a variation in human labor across the products. Citrus and tea demand higher labor input than the other products. This could indicate that these products are more labor-intensive or require more manual harvesting.

Hazelnut has the lowest mean machinery use. This could imply either more manual processes or smaller-scale operations for certain products. The low variance suggests consistent agricultural machinery usage, but the range still shows differences across products.

Citrus and tea have the highest nitrogen inputs, which might be attributed to the need for higher soil fertility or larger crop yields. The same products require higher levels of potassium and organic manure, for optimal growth. In contrast, hazelnut uses less FYM.

Table 10 provides descriptive statistics for the LCI variables corresponding to emissions related to the input variables in the dataset. The total number of observations is 865. The table includes key metrics such as minimum, maximum, mean, and standard deviation for each variable. These statistics provide a general analysis, offering insight into the dataset's characteristics before further modeling.

Table 10. Descriptive statistics of the remaining variables of the dataset.

Variable	Mean	Standard Deviation	Minimum	Maximum
NH_3 (B.1)	1.52E-03	1.11E-03	1.56E-04	3.84E-03
C_6H_6 (B.1)	5.56E-04	4.05E-04	5.70E-05	1.40E-03
Benzo(a)pyrene (B.1)	2.00E-06	2.00E-06	0.00E+00	6.00E-06
Cd (B.1)	1.00E-06	1.00E-06	0.00E+00	2.00E-06
CO_2fossil (B.1)	2.37E+02	1.73E+02	2.43E+01	6.00E+02
CO fossil (B.1)	4.47E-01	3.26E-01	4.57E-02	1.13E+00
Cr (B.1)	4.00E-06	3.00E-06	0.00E+00	1.00E-05
Cu (B.1)	1.29E-04	9.40E-05	1.30E-05	3.26E-04
N_2O (B.1)	9.14E-03	6.67E-03	9.35E-04	2.31E-02
Heat waste (B.1)	3.46E+03	2.52E+03	3.53E+02	8.73E+03
CH_4 fossil (B.1)	9.84E-03	7.18E-03	1.01E-03	2.49E-02
Ni (B.1)	5.00E-06	4.00E-06	1.00E-06	1.30E-05
Nox (B.1)	2.97E+00	2.17E+00	3.03E-01	7.49E+00
NMVOC (B.1)	1.61E-01	1.18E-01	1.65E-02	4.08E-01
PAH (B.1)	2.50E-04	1.82E-04	2.60E-05	6.30E-04
Particulates <2,5um (B.1)	3.77E-01	2.75E-01	3.86E-02	9.53E-01
Se (B.1)	1.00E-06	1.00E-06	0.00E+00	2.00E-06
SO2 (B.1)	7.54E-02	5.50E-02	7.71E-03	1.91E-01
Zn (B.1)	7.60E-05	5.60E-05	8.00E-06	1.92E-04
Cd (B.2)	1.40E-05	1.00E-05	1.00E-06	3.40E-05
Pb (B.2)	6.00E-05	4.40E-05	6.00E-06	1.51E-04
Zn (B.2)	3.78E-02	2.76E-02	3.87E-03	9.56E-02
N_2O (B.3)	3.47E+00	1.73E+00	1.03E+00	7.89E+00
NH_3 by chemical fertilizers (B.3)	2.79E+01	1.88E+01	1.19E+00	6.92E+01
NH_3 by FYM (B.3)	5.23E+00	4.60E+00	1.11E+00	1.71E+01
N_2O by chemical fertilizers (B.4)	3.61E-01	2.44E-01	1.54E-02	8.96E-01
N_2O by FYM (B.4)	6.77E-02	5.95E-02	1.43E-02	2.21E-01
Nitrate (B.5)	2.93E+01	1.46E+01	8.67E+00	6.67E+01

Cd (B.7)	7.64E+03	5.54E+03	1.36E+03	1.87E+04
Cu (B.7)	4.43E+04	3.37E+04	8.50E+03	1.31E+05
Zn (B.7)	1.98E+06	1.69E+06	4.49E+05	6.01E+06
Pb (B.7)	5.70E+04	4.84E+04	1.30E+04	1.73E+05
Ni (B.7)	3.01E+03	2.56E+03	3.36E+02	9.39E+03
Cr (B.7)	2.70E+05	1.79E+05	1.10E+04	5.77E+05
Hg (B.7)	1.63E+03	1.11E+03	2.97E+01	3.87E+03
CO_2 (B.8)	1.56E+03	1.94E+03	8.02E+01	6.24E+03

[1] Missing values: 0

Analysis of Figure 11 shows the relationship between diesel fuel consumption (kg/ha) and average NH_3 and N_2O emissions. The line representing the average value of N_2O emissions shows an increase in emissions as diesel consumption increases. Although the blue line also shows an increase in NH_3 emissions as diesel consumption increases it is less pronounced compared to N_2O. The graph shows that N_2O emissions increase much faster than NH_3 emissions as diesel consumption increases. As these are polluting gases, the high value of emissions associated with agricultural practices suggests implementing strategies to reduce the use of diesel and more sustainable practices in the application of fertilizers, pesticides, and insecticides to reduce them. Figure 12 analysis is very similar to Figure 11 as the diesel fuel consumption (kg/ha) leads to more CO_2 fossil emissions and more heat waste. In addition to the increase in CO_2 emissions, the contribution to the loss of heat waste is related to the increase in diesel consumption. Variables Heat Waste and N_2O show more pronounced growth compared to the other two variables shown in Figures 11 and 12. These graphs aim to show the behavior of the variables and that although they are similar, not all emissions grow at the same rate depending on the increase in inputs, in this case, the use of diesel.

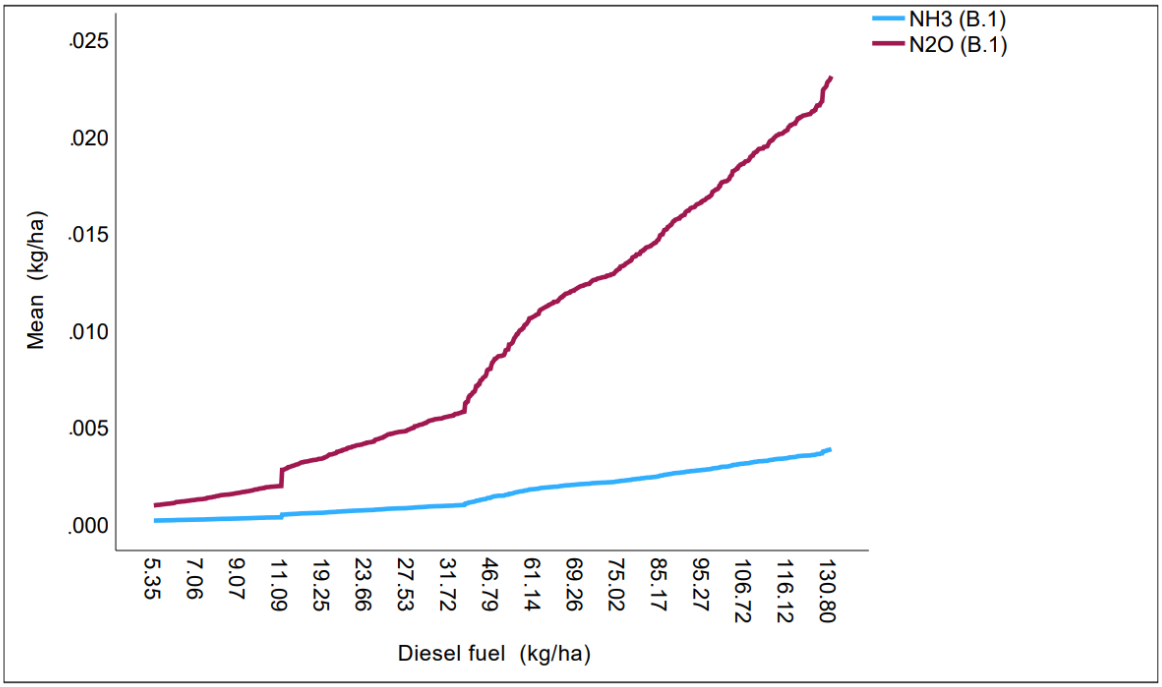

Figure 11. Diesel fuel consumption (Kg/Ha) and average NH_3 and N_2O emissions.

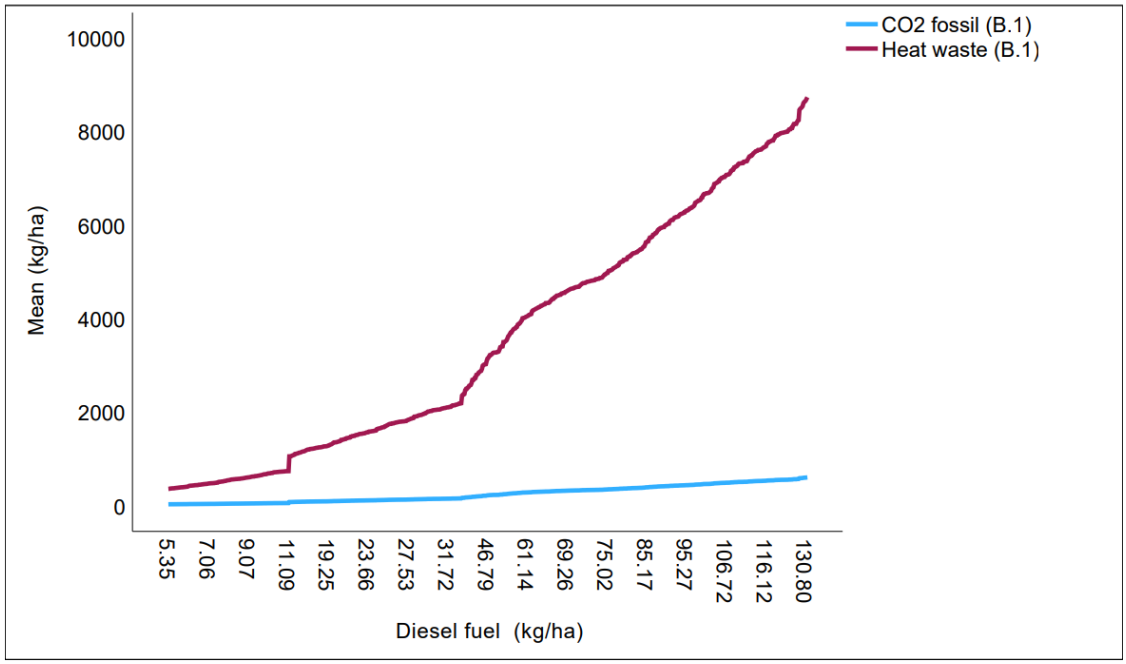

Figure 12. Diesel fuel consumption (Kg/Ha) and average CO_2 emissions and heat waste.

Figure 13 shows the amount of diesel fuel needed across the production of the different products, considering the yield output. Citrus and Kiwi have higher values, with the median around 100 kg/ha. In comparison, watermelon presents an intermediate value of diesel fuel

consumption, around 70 kg/ha, and the highest yield. Hazelnut is the product with the worst performance in harvest results.

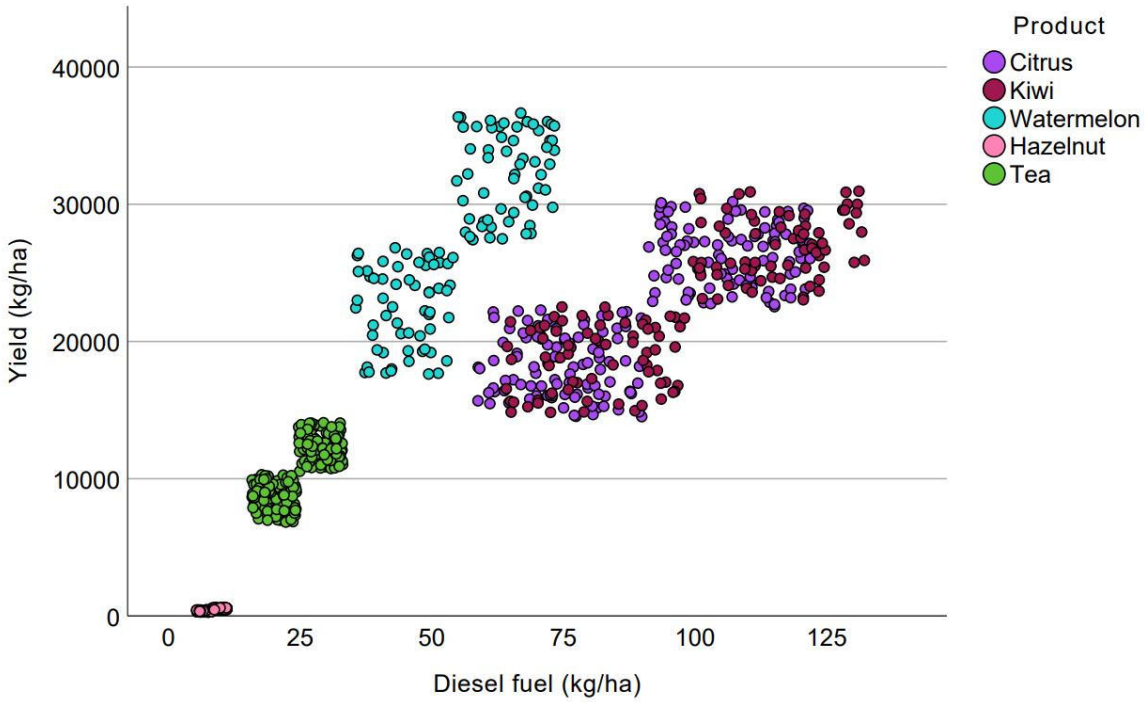

Figure 13.Diesel fuel use and yield relationship across the different products.

Figure 14 represents the quantity of agricultural machinery required for the different types of crops. Watermelon has the highest median and the widest range, indicating variability in machinery use.

Citrus has a relatively low median, with a smaller spread. Hazelnut and tea crops have the lowest and most consistent usage, with medians below 15 kg/ha and narrow interquartile ranges.

This form of representation makes it easier to interpret the data to quickly understand which products require the least use of agricultural machinery.

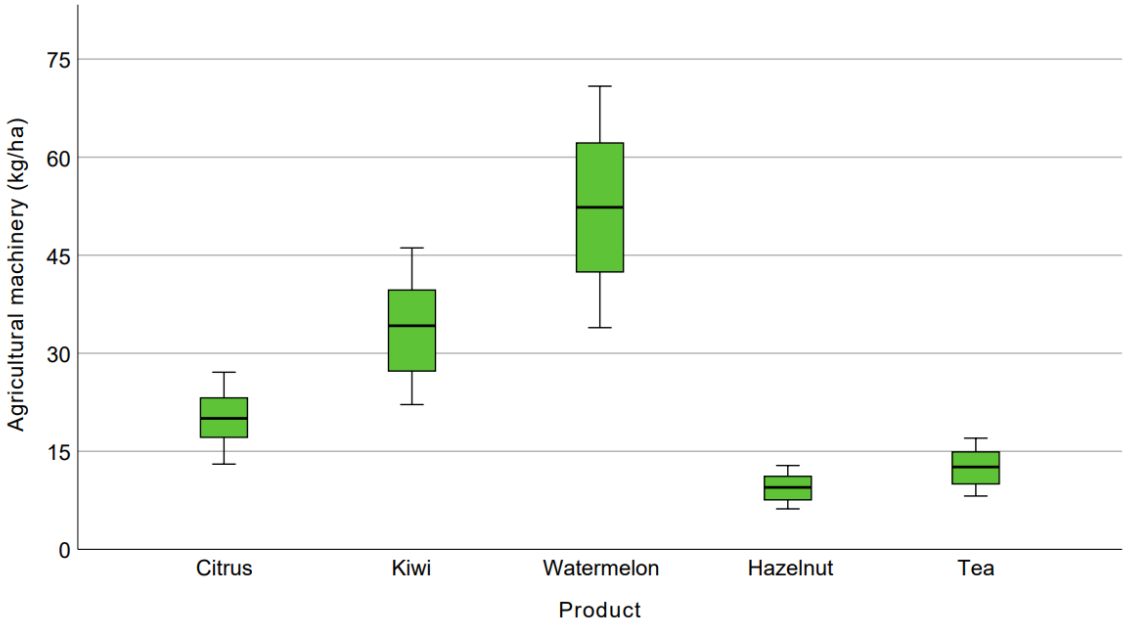

Figure 14. Agricultural machinery required across the different crops.

Figure 15 is a visual representation of fungicides applied to different crops. Hazelnut has the highest median value, demonstrating a higher need for fungicides in the production process. Citrus and Kiwi have close and lower values compared with hazelnut. Watermelon and tea have the lowest and most consistent usage, with medians around 2 kg/ha and relatively smaller interquartile ranges.

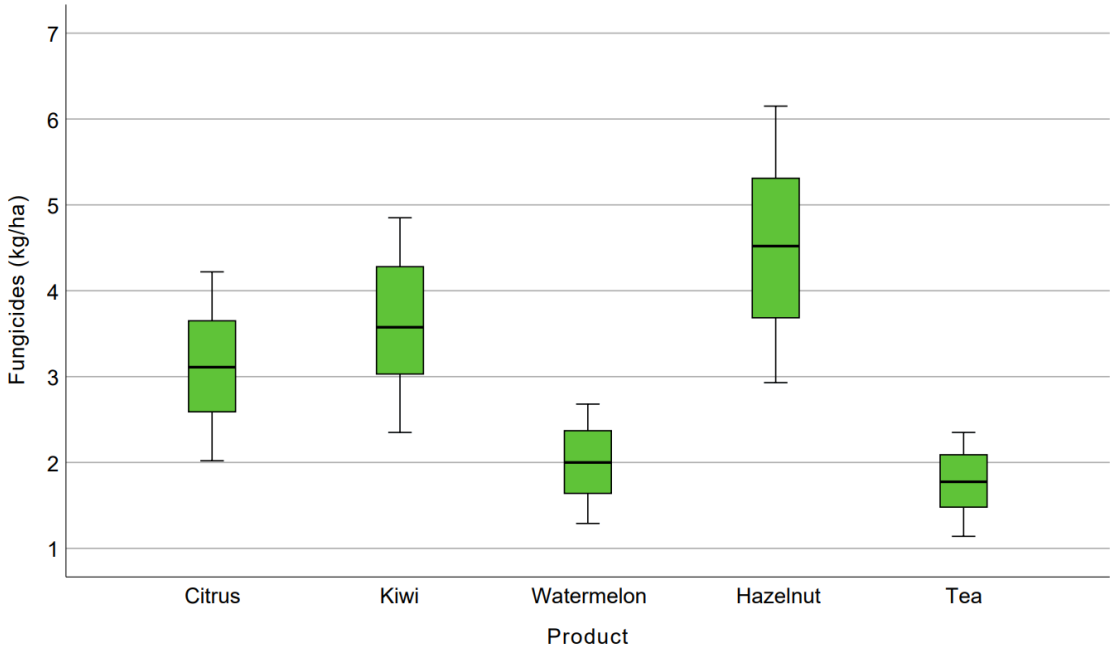

Figure 15. Fungicides required across the different crops.

Figure 16 illustrates the relationship between variables. This scatter plot visually represents how yield varies with pesticide use across different crops. Each product is color-coded according to the label presented in Figure 16.

Citrus, kiwi, watermelon, and tea data points are clustered, suggesting a more consistent relationship between pesticide use and yield. Hazelnut data points are more scattered, indicating variability in yield with different pesticide use levels. For each of these products, there are two notable groups which may be an indicator that it would be feasible to work on clusters to find out more about the behavior of the variables

Crops with more clustered points (like tea) may have more predictable outcomes, while those with widely spread points (like citrus) may have more variable results.

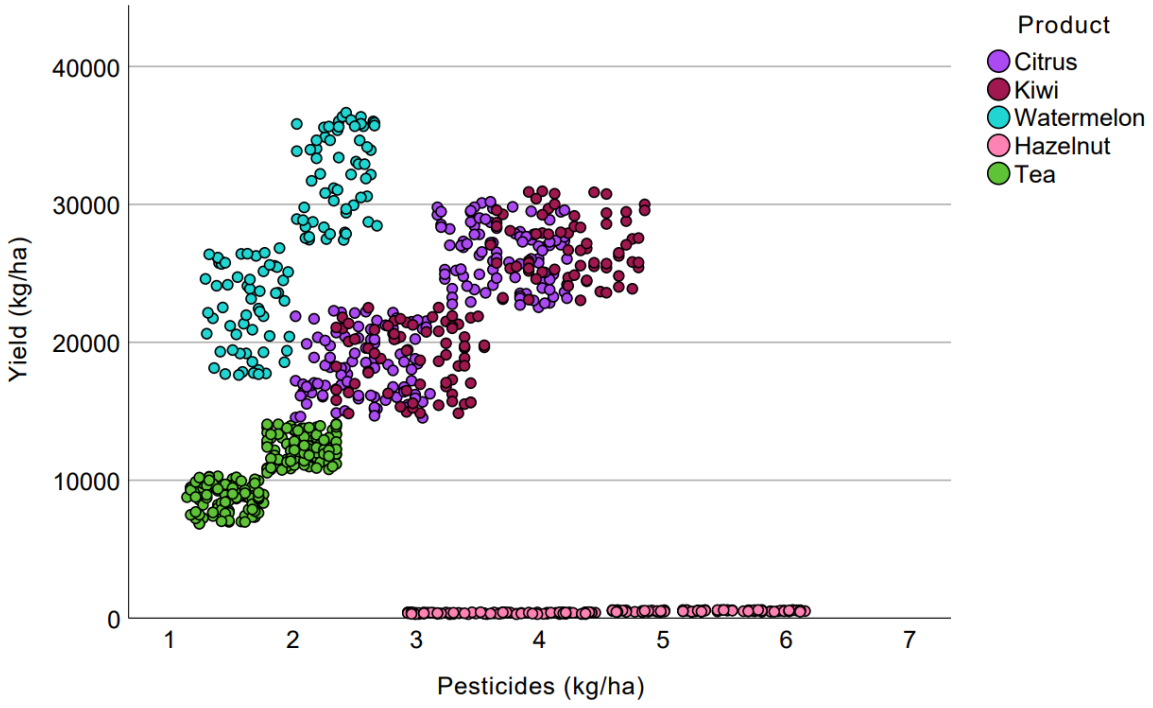

Figure 16. Pesticides use and yield relation considering the different products.

The relationship between FYM and yield is presented in Figure 17. Except from citrus and kiwi, the other products differ in their use of FYM and resulting yields. Citrus and kiwi are precisely the products that seem to depend more on FYM to present higher productivity results.

The low FYM usage for Hazelnut and Tea suggests that these products either do not require much FYM or there are other factors influencing their yield.

The scatter plot in Figure 17 provides hints about possible clusters on the dataset.

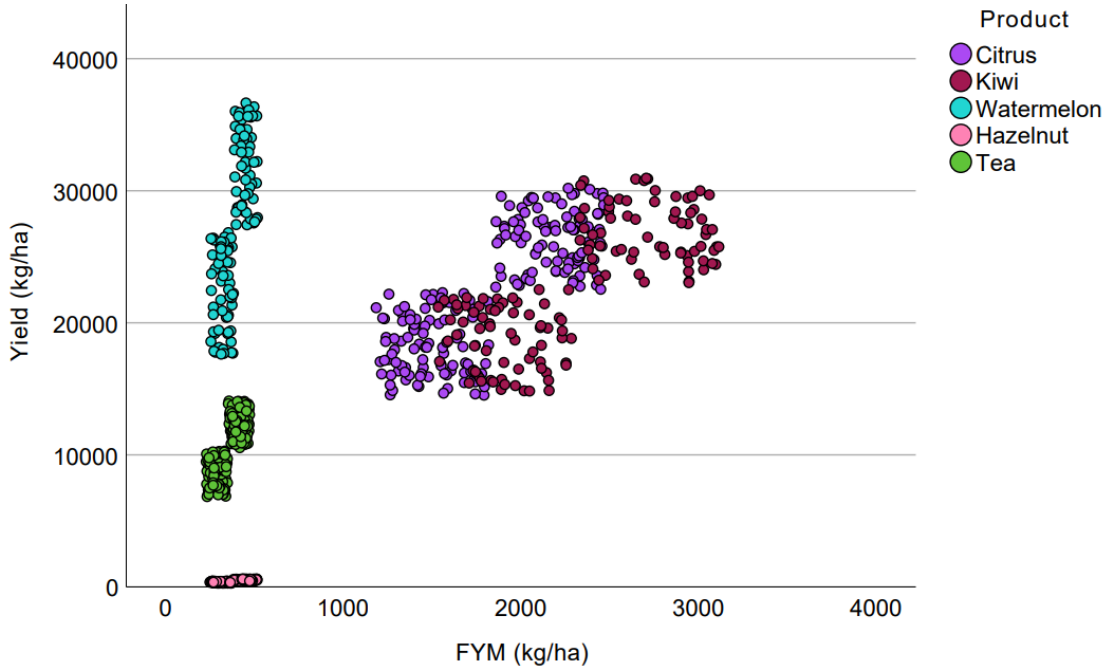

Figure 17. FYM application and yield considering the different crops.

In Figure 18 hazelnut and tea are not presented because, as mentioned before, there is no electricity consumption registered for these two products. Each crop seems to form two different groups.

Citrus appears to cluster at the lower end of electricity consumption (around 100 kWh), with yields ranging up to 30000 kg. Kiwi is scattered over a comparatively wider range of electricity usage (from 200 to almost 500 kWh) and generally produces yields similar to citrus. Watermelon is mostly found in the 100–300 kWh consumption range, presenting higher yields when compared to the other crops.

This plot can be useful for analyzing the energy efficiency of different crops concerning yield output.

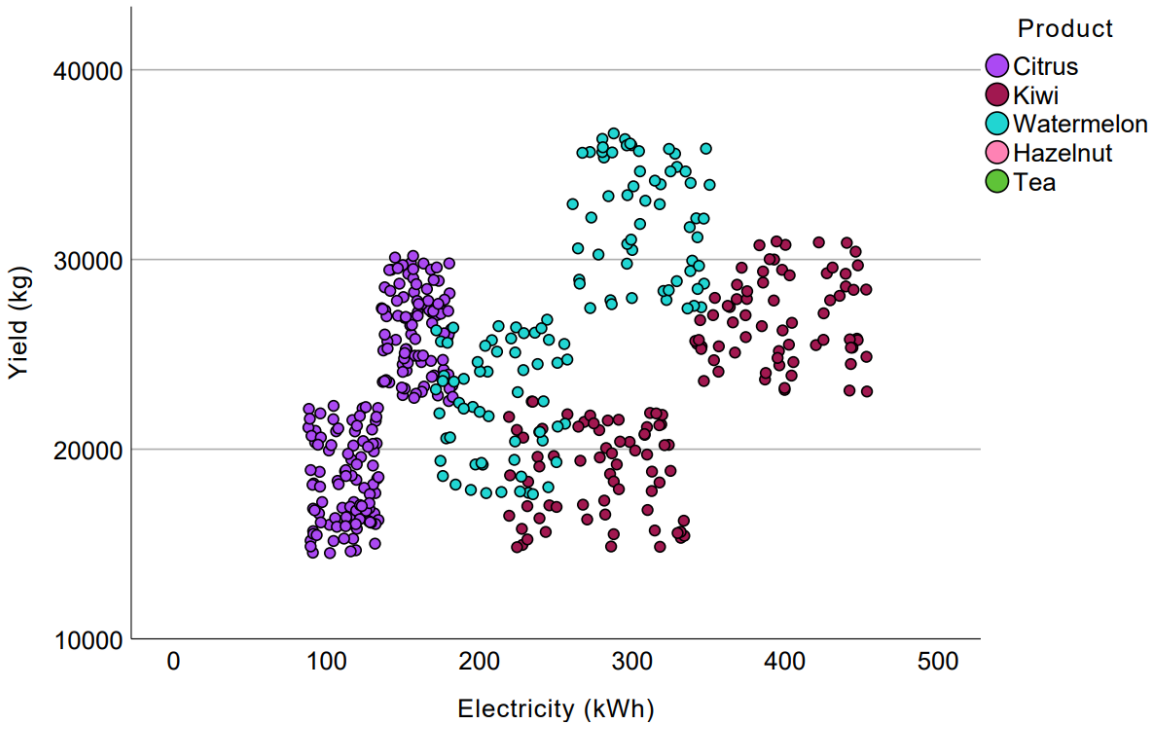

Figure 18. Electricity consumption and yield considering the different crops.

4.2. Performance Evaluation and Comparison of Models

The following paragraphs present the results obtained in DT and ANN models. As RMSE may be affected by outliers it is important to combine other performance metrics. MAE is an example of a more robust method. NAE allows for comparison across the full range, giving a relative measure of error and a more balanced comparison.

Performance metrics of a decision tree model without pre-pruning and depth 6 are presented in Table 11. After testing models at various depths, it became clear that between 6 and 10 there were smaller differences in performance results for most variables. As the aim is to obtain simplified models that do not fit the data too closely and are therefore easier to generalize, it is recommended to opt for the one with the least depth.

Almost all the results are indicators of good model accuracy. Some examples are the label variables NH_3 (B.1) and C_6H_6 (B.1) which show low RMSE, indicating good model accuracy.

For many variables, the absolute error is minimal, such as Benzo(a)pyrene (B.1) and Cd (B.1).

Most of the variables have a NAE around 0.03 or lower. Some variables like Cd (B.1) and Se (B.1) demonstrate higher values, indicating the model has more difficulty in predicting those, comparatively with the other variables.

For most variables, the correlation is extremely close to one, showing a high correlation, representing strength and direction between two variables, leading to a stronger model performance.

Benzo(a)pyrene (B.1), Cd (B.1), and Se (B.1) have zero absolute errors but higher normalized absolute errors, unveiling the relative performance of the model.

This table shows that, without pre-pruning, the DT model generally performs with high correlation and low errors.

The average execution time of the DT model is close to zero seconds.

Table 11. Decision Tree Models Results without pre-pruning.

Label variable	Root Mean Squared Error	Absolute Error	Normalized Absolute Error	Correlation
NH_3 (B.1)	4.20E-05	2.70E-05	2.79E-02	9.99E-01
C_6H_6 (B.1)	1.50E-05	1.00E-05	2.80E-02	9.99E-01
Benzo(a)pyrene (B.1)	0.00E+00	0.00E+00	1.14E-01	9.91E-01
Cd (B.1)	0.00E+00	0.00E+00	3.86E-01	9.01E-01
CO2 fossil (B.1)	6.56E+00	4.25E+00	2.81E-02	9.99E-01
CO fossil (B.1)	1.23E-02	7.96E-03	2.79E-02	9.99E-01
Cr (B.1)	0.00E+00	0.00E+00	9.81E-02	9.94E-01
Cu (B.1)	4.00E-06	2.00E-06	2.82E-02	9.99E-01
N_2O (B.1)	2.52E-04	1.62E-04	2.78E-02	9.99E-01
Heat waste (B.1)	9.51E+01	6.16E+01	2.80E-02	9.99E-01
CH4 fossil (B.1)	2.71E-04	1.76E-04	2.80E-02	9.99E-01
Ni (B.1)	0.00E+00	0.00E+00	8.32E-02	9.95E-01
Nox (B.1)	8.19E-02	5.31E-02	2.81E-02	9.99E-01
NMVOC (B.1)	4.45E-03	2.88E-03	2.81E-02	9.99E-01
PAH (B.1)	7.00E-06	4.00E-06	2.81E-02	9.99E-01
Particulates < 2,5um (B.1)	1.04E-02	6.71E-03	2.79E-02	9.99E-01
Se (B.1)	0.00E+00	0.00E+00	3.86E-01	9.01E-01
SO2 (B.1)	2.05E-03	1.34E-03	2.78E-02	9.99E-01
Zn (B.1)	2.00E-06	1.00E-06	2.87E-02	9.99E-01
Cd (B.2)	0.00E+00	0.00E+00	3.59E-02	9.99E-01

Pb (B.2)	2.00E-06	1.00E-06	2.86E-02	9.99E-01
Zn (B.2)	1.05E-03	6.77E-04	2.81E-02	9.99E-01
N_2O (B.3)	1.33E-01	9.55E-02	7.02E-02	9.97E-01
NH_3 by chemical fertilizers (B.3)	5.76E-01	4.15E-01	2.62E-02	1.00E+00
NH_3 by FYM (B.3)	2.84E-01	1.60E-01	3.98E-02	9.98E-01
N_2O by chemical fertilizers (B.4)	7.45E-03	5.37E-03	2.62E-02	1.00E+00
N_2O by FYM (B.4)	3.52E-03	2.03E-03	3.89E-02	9.98E-01
Nitrate (B.5)	1.13E+00	8.11E-01	7.04E-02	9.97E-01
Cd (B.7)	2.29E+02	1.55E+02	3.06E-02	9.99E-01
Cu(B.7)	1.59E+03	1.07E+03	3.92E-02	9.99E-01
Zn(B.7)	4.28E+04	2.96E+04	1.95E-02	1.00E+00
Pb(B.7)	1.25E+03	8.62E+02	1.99E-02	1.00E+00
Ni (B.7)	1.12E+02	6.90E+01	3.16E-02	9.99E-01
Cr (B.7)	1.37E+04	8.65E+03	5.54E-02	9.97E-01
Hg (B.7)	5.93E+01	3.85E+01	4.03E-02	9.99E-01
CO2 (B.8)	3.81E+01	2.37E+01	1.48E-02	1.00E+00

The DT in Figure 19 represents a part of the model that predicts N_2O (B.3). The nodes represent a decision based on a certain variable that is considered to predict the outcome variable. Figure 19 begins with a decision based on the fungicides variable. If the fungicide value is greater than 1.7750, the tree splits to the left; otherwise, it splits to the right. After this split, decisions are made based on Nitrogen and FYM until the value predicted to N_2O (B.3) is presented.

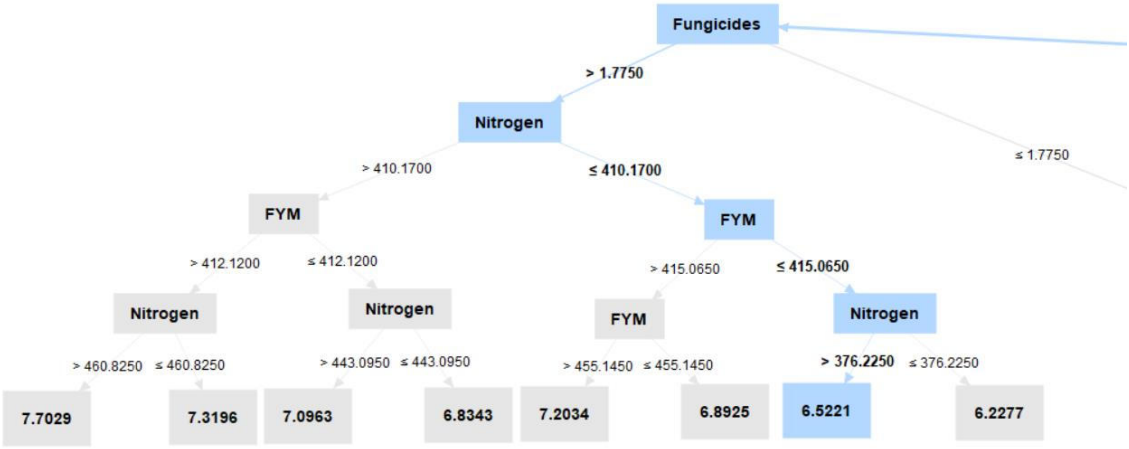

Figure 19. Partial example of decision tree structure to predict N_2O (B.3).

Tables 12 and 13 present the results of the same neural network model, varying the label to predict environmental emissions associated with production inputs. Performance metrics include Root Mean Squared Error, Absolute Error, Normalized Absolute Error, Squared Error, and Squared Correlation, allowing for an assessment of the prediction accuracy.

RMSE of variables such as NH_3 (B.1) and C_6H_6 (B.1) suggest extremely low RMSE, indicating high precision in these predictions. In contrast, variables like Cd (B.7), and Cu (B.7) have much higher RMSE, suggesting greater errors in these predictions.

A low absolute error, such as in NH_3 (B.1), demonstrates more accurate predictions, while higher values, like Zn (B.7), indicate larger deviations in the predictions.

For variables like Zn (B.7), Cr (B.7), Pb (B.7) and Cu (B.7), the RMSE values are comparatively higher. On the contrary, present lower NAE. The models for these variables may be facing challenges in making accurate predictions due to the intrinsic variability of the data. This justified the need to also compare the models using the NAE.

Correlation is presented to measure how well the predicted values follow the trend of the actual values. Values close to 1, indicate that the model is capturing the data pattern well. Most variables show a high correlation, suggesting that the models are appropriately fitted.

The average execution time of the model ANN1 is 12.61 seconds.

Table 12.Performance of Neural Network 1 model varying the label variable.

Label variable	RMSE	Absolute Error	Normalized Absolute Error	Squared Error	Squared Correlation
NH_3 B.1	8.40E-05	6.30E-05	6.41E-02	0.00E+00	9.94E-01
C_6H_6 (B.1)	3.10E-05	2.30E-05	6.41E-02	0.00E+00	9.94E-01
Benzo(a)pyrene (B.1)	0.00E+00	0.00E+00	6.41E-02	0.00E+00	9.94E-01
Cd (B.1)	0.00E+00	0.00E+00	6.41E-02	0.00E+00	9.94E-01
CO2 fossil (B.1)	1.31E+01	9.84E+00	6.41E-02	1.74E+02	9.94E-01
CO fossil (B.1)	2.47E-02	1.85E-02	6.41E-02	6.15E-04	9.94E-01
Cr (B.1)	0.00E+00	0.00E+00	6.41E-02	0.00E+00	9.94E-01
Cu (B.1)	7.00E-06	5.00E-06	6.41E-02	0.00E+00	9.94E-01
N_2O (B.1)	5.06E-04	3.79E-04	6.41E-02	0.00E+00	9.94E-01
Heat waste (B.1)	1.91E+02	1.43E+02	6.41E-02	3.68E+04	9.94E-01
CH4 fossil (B.1)	5.45E-04	4.08E-04	6.41E-02	0.00E+00	9.94E-01
Ni (B.1)	0.00E+00	0.00E+00	6.41E-02	0.00E+00	9.94E-01
Nox (B.1)	1.64E-01	1.23E-01	6.41E-02	2.71E-02	9.94E-01
NMVOC (B.1)	8.93E-03	6.69E-03	6.41E-02	8.00E-05	9.94E-01
PAH (B.1)	1.40E-05	1.00E-05	6.41E-02	0.00E+00	9.94E-01
Particulates < 2.5um (B.1)	2.09E-02	1.56E-02	6.41E-02	4.39E-04	9.94E-01
Se (B.1)	0.00E+00	0.00E+00	6.41E-02	0.00E+00	9.94E-01
SO2 (B.1)	4.17E-03	3.13E-03	6.41E-02	1.80E-05	9.94E-01
Zn (B.1)	4.00E-06	3.00E-06	6.41E-02	0.00E+00	9.94E-01
Cd (B.2)	1.00E-06	1.00E-06	6.41E-02	0.00E+00	9.94E-01
Pb (B.2)	3.00E-06	2.00E-06	6.41E-02	0.00E+00	9.94E-01
Zn (B.2	2.09E-03	1.57E-03	6.41E-02	4.00E-06	9.94E-01
N_2O B.3	3.12E-01	2.46E-01	1.80E-01	9.79E-02	9.68E-01
NH_3 by chemical fertilizers (B.3)	1.44E+00	1.13E+00	7.03E-02	2.08E+00	9.94E-01
NH_3 by FYM (B.3)	2.75E-01	1.99E-01	4.91E-02	7.63E-02	9.96E-01
N_2O by chemical fertilizers (B.4)	1.86E-02	1.46E-02	7.03E-02	3.48E-04	9.96E-01
N_2O by FYM (B.4)	3.56E-03	2.58E-03	4.91E-02	1.30E-05	9.96E-01

Nitrate (B.5)	2.69E+00	2.12E+00	1.83E-01	7.28E+00	9.66E-01
Cd (B.7)	3.94E+02	2.77E+02	5.45E-02	1.56E+05	9.95E-01
Cu(B.7)	1.95E+03	1.45E+03	5.29E-02	3.80E+06	9.97E-01
Zn(B.7)	9.77E+04	7.21E+04	4.73E-02	9.63E+09	9.97E-01
Pb(B.7)	2.77E+03	2.04E+03	4.69E-02	7.74E+06	9.97E-01
Ni (B.7)	1.66E+02	1.12E+02	5.10E-02	2.78E+04	9.96E-01
Cr (B.7)	1.18E+04	9.40E+03	5.97E-02	1.40E+08	9.96E-01
Hg (B.7)	9.22E+01	7.12E+01	7.43E-02	8.54E+03	9.93E-01
CO2 (B.8)	9.85E+01	7.65E+01	4.71E-02	9.73E+03	9.97E-01

Figure 20 represents the Neural Network 1 for the Benzo(a)pyrene (B.1) variable. It gives an example of the structure of the ANN with only 1 hidden layer.

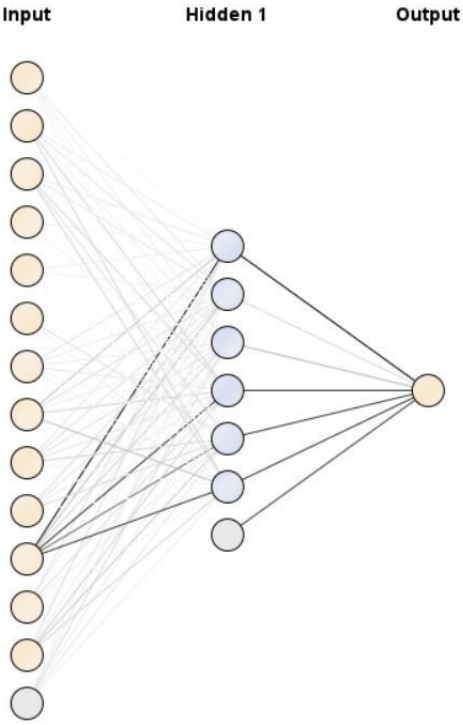

Figure 20. Neural Network 1 for the Benzo(a)pyrene (B.1) variable.

Generally speaking, and comparing Tables 12 and 13, the variables with the lowest NAE and RMSE are not the same in Table 13.

The average execution time of the model ANN2 is 1.86 seconds.

Analyzing Table 13 these are the variables with higher RMSE, values are higher than 1: NH3 by chemical fertilizers (B.3), Nitrate (B.5), CO_2 fossil (B.1), Hg (B.7), CO_2 (B.8), Ni (B.7), Heat waste (B.1), Cd (B.7), Cu(B.7), Pb(B.7), Cr (B.7), Zn(B.7). But, as mentioned above, due to the order of

magnitude of the values of the variables, it is prudent to also evaluate the NAE. Top performers in NAE are N2O by FYM (B.4) and Ni (B.7) with approximately 0.07. The highest correlations are found in N2O by FYM (B.4) and Ni (B.7).

On the other hand, models like Cd (B.1) and Zn (B.1) show higher errors and lower correlations, suggesting room for improvement.

Table 13. Performance of Neural Network 2 model varying the label variable.

Label variable	RMSE	Absolute Error	Normalized Absolute Error	Squared Error	Squared Correlation
NH_3 B.1	2.90E-04	2.48E-04	2.54E-01	0.00E+00	9.81E-01
C_6H_6 (B.1)	7.90E-05	6.20E-05	1.75E-01	0.00E+00	9.79E-01
Benzo(a)pyrene (B.1)	1.00E-06	0.00E+00	3.01E-01	0.00E+00	9.78E-01
Cd (B.1)	0.00E+00	0.00E+00	1.94E-01	0.00E+00	9.76E-01
CO2 fossil (B.1)	1.82E+01	1.51E+01	9.82E-02	3.89E+02	9.95E-01
CO fossil (B.1)	3.60E-02	2.86E-02	9.96E-02	1.38E-03	9.94E-01
Cr (B.1)	0.00E+00	0.00E+00	1.56E-01	0.00E+00	9.80E-01
Cu (B.1)	2.00E-05	1.70E-05	2.00E-01	0.00E+00	9.78E-01
N_2O (B.1)	9.42E-04	7.51E-04	1.27E-01	1.00E-06	9.87E-01
Heat waste (B.1)	2.60E+02	2.03E+02	9.03E-02	8.12E+04	9.94E-01
CH4 fossil (B.1)	7.45E-04	5.77E-04	9.11E-02	1.00E-06	9.94E-01
Ni (B.1)	1.00E-06	1.00E-06	2.48E-01	0.00E+00	9.73E-01
Nox (B.1)	2.43E-01	1.96E-01	1.03E-01	6.81E-02	9.93E-01
NMVOC (B.1)	1.22E-02	1.00E-02	9.60E-02	1.62E-04	9.93E-01
PAH (B.1)	4.30E-05	3.60E-05	2.19E-01	0.00E+00	9.77E-01
Particulates < 2.5um (B.1)	3.18E-02	2.53E-02	1.03E-01	1.49E-03	9.91E-01
Se (B.1)	0.00E+00	0.00E+00	2.36E-01	0.00E+00	9.72E-01
SO2 (B.1)	7.09E-03	5.85E-03	1.20E-01	5.70E-05	9.92E-01
Zn (B.1)	1.20E-05	1.00E-05	1.94E-01	0.00E+00	9.79E-01
Cd (B.2)	2.00E-06	1.00E-06	1.66E-01	0.00E+00	9.81E-01
Pb (B.2)	1.00E-05	8.00E-06	2.12E-01	0.00E+00	9.79E-01
Zn (B.2	2.88E-03	2.33E-03	9.45E-02	9.00E-06	9.94E-01
N_2O B.3	3.12E-01	2.47E-01	1.80E-01	1.01E-01	9.72E-01

NH_3 by chemical fertilizers (B.3)	1.81E+00	1.51E+00	9.35E-02	3.73E+00	9.95E-01
NH_3 by FYM (B.3)	4.87E-01	3.73E-01	9.20E-02	3.09E-01	9.92E-01
N_2O by chemical fertilizers (B.4)	2.17E-02	1.81E-02	8.72E-02	4.85E-04	9.96E-01
N_2O by FYM (B.4)	4.80E-03	3.76E-03	7.14E-02	2.50E-05	9.97E-01
Nitrate (B.5)	2.72E+00	2.13E+00	1.82E-01	7.85E+00	9.77E-01
Cd (B.7)	5.33E+02	4.40E+02	8.64E-02	3.24E+05	9.96E-01
Cu(B.7)	4.09E+03	3.40E+03	1.25E-01	1.88E+07	9.96E-01
Zn(B.7)	1.69E+05	1.32E+05	8.57E-02	3.24E+10	9.93E-01
Pb(B.7)	4.61E+03	3.68E+03	8.63E-02	2.28E+07	9.95E-01
Ni (B.7)	1.99E+02	1.62E+02	7.33E-02	4.37E+04	9.97E-01
Cr (B.7)	2.40E+04	1.89E+04	1.20E-01	6.63E+08	9.88E-01
Hg (B.7)	1.49E+02	1.23E+02	1.28E-01	3.01E+04	9.90E-01
CO2 (B.8)	1.59E+02	1.27E+02	7.87E-02	2.81E+04	9.96E-01

The results of the Friedman and Wilcoxon tests are presented below to see if there are any differences between the distribution of the three models. Table 14 presents the descriptive statistics of the Decision Tree Model Absolute Error (DT NAE), Artificial Neural Network 1 Normalized Absolute Error (ANN1 NAE), and Artificial Neural Network 2 Normalized Absolute Error (ANN2 NAE). The values represent the mean and corresponding standard deviation of the NAE obtained for the different variables and N=36.

Table 14. Descriptive Statistics of the samples.

Metric	Mean	Standard Deviation
DT NAE	5.77E-02	8.38E-02
ANN1 NAE	6.79E-02	2.87E-02
ANN2 NAE	1.41E-01	6.13E-02

Being the Friedman test p-value <0.001 which is lower than α (α=0.05) indicates that there is a statistically significant difference in the NAE values among at least one of the models

compared (DT, ANN1, and ANN2). The results of the Wilcoxon test for the three comparisons (ANN1 vs. DT, ANN2 vs. DT, and ANN2 vs. ANN1), the p-values <0.001 are lower than α (α=0.05) This means to reject the null hypothesis for all comparisons, indicating that there are significant differences in NAE among these models.

5. Discussion and research limitations

In this chapter, the study hypothesis formulated initially is answered: "DT and ANN models can estimate emissions, learn from data, and effectively predict data in the LCI phase?"

The results of this study show that DT, in the context of LCI, can be applied to estimate emissions. By training a DT it was possible to overcome the uncertainty around the calculation methods and equations used by the most common LCA software. In the future, with optimized models, it is expected to be possible to predict LCI variables based on similar input features. The developed DT allows to know the estimated emissions values associated with certain inputs.

DT provides insight into which features are most influential in determining emissions. This can help in understanding key variables in emissions calculations.

The work done in this study reunites validated trustworthy equations from open sources and recommended methodologies instead of using paid software that lacks transparency. The DT is important for gathering the data needed for the LCI phase. Now, those DT models can be applied to new data without requiring as much time for analysis, reducing manual data entry and speeding up the process.

Citrus emerges as the most environmentally friendly crop due to its strategic importance and efficient input use, while hazelnut production lags due to inefficiencies in input utilization and yield (78).

Citrus production highlights key areas where improvements can be made, particularly in resource usage and environmental efficiency. Furthermore, enhancing life cycle inventories, especially with regionalized datasets for inputs and improved water usage models, could lead to more accurate assessments and ultimately more sustainable practices. LCI data must be representative of different climates, soil types, farming practices, pesticides, and fertilizer use, among others. This aligns with literature findings that emphasize the need for regionally representative data and the inclusion of biodiversity and water scarcity impacts in assessments (95).

Kiwi production demonstrates a higher resource intensity compared to other crops. This could be attributed to factors like irrigation or more intensive pest management practices, frequently necessary for maintaining consistent production levels. Previous studies in the literature reveal that soil organic matter increased N_2O emissions and the availability of nitrogen and carbon in soil. Depending on the composition of the soil, this may or may not influence water

retention. Soils that tend to retain more water are more susceptible to higher N_2O emissions due to denitrification[2]. In addition, meteorological conditions also influence pollution resulting from the production process (20). For example, precipitation can promote the penetration of nutrients through soil layers and lead to leaching[3] and runoff[4]. It is important then that it is considered in LCA. In a study by Pergola et al., kiwi orchard direct and indirect CO_2 emissions were greater than those observed in peach production (16). This supports the possibility that characteristics inherent to the kiwi production process influence LCA.

Compared to other products, watermelon production requires greater Nitrogen application. Similar conclusions can also be found in the literature, including a study in the province of Guilan, which indicates that better agricultural management in terms of nitrogen application can lead to watermelon production with lower GHG emissions (98).

The low variability in yield and input use suggests that hazelnut cultivation is stable but less productive. In this study, hazelnut production performs worse in pesticide and fungicide application, when compared to the other products studied. Some studies in the literature concluded that it is reasonable to reduce or replace diesel fuel, pesticides, and agricultural machinery in hazelnut orchards. Ashkan et al. concluded that diesel fuel contributed the largest (33.84%) to the total CO_2 emissions in hazelnut production, followed by machinery at 26.54%, and nitrogen at 24.76%, highlighting the need for optimized machinery use, proper maintenance, and soil analysis to reduce CO_2 emissions and chemical fertilizer consumption (99,100).

Tea production in this study stands out for its high demand for human labor. This reflects the labor-intensive nature of tea cultivation, highlighting the potential for future mechanization or labor-saving technologies to enhance productivity without compromising quality (101,102).

Despite this work not including regional features, climate conditions, or geographical data, according to the literature, DT models may capture specific patterns and relationships relevant to a certain region or condition (20). For example, one model could be trained for urban areas while another for rural areas. These adaptations enable the decision trees to better capture

[2] Denitrification is a key microbial process in the nitrogen cycle, whereby nitrate (NO_3^-) is reduced to gaseous forms of nitrogen, such as nitric oxide (NO), nitrous oxide (N_2O), and dinitrogen (N_2). The process is influenced by factors such as soil moisture, temperature, and organic matter availability. It is carried out by facultative anaerobic bacteria, such as *Pseudomonas*, *Paracoccus*, and *Bacillus*, that use nitrate as an alternative electron acceptor in the absence of oxygen during cellular respiration. (96)

[3] Leaching occurs when chemicals are excessively applied or too much rainwater or irrigation occurs in a short period. It is when the chemicals flow into or out of the soil (97).

[4] Runoff is when chemical flow in the surface of the soil occurs due to precipitation (97).

and predict regional variations, leading to more accurate and relevant insights tailored to specific geographic locations or climate conditions.

One of the limitations of this work is that LCA and ML analytical models are expensive and depend on a large amount of structured training data. Computational cost and training time are other important parameters related to the accuracy of the results.

The study's reliance on data collected from Guilan province in northern Iran may not capture all the variability in emissions across different soil types and climatic conditions and even distinct agricultural practices.

The relationship between input levels and yields is evident, with more resource-intensive products yielding higher outputs, which could help in decision-making regarding cost efficiency and resource allocation.

6. Conclusions and Future Work

The present dissertation aimed to explore the potential of ML, specifically DT and ANN to estimate environmental emissions in the LCI phase, particularly for the production of kiwi, watermelon, citrus, tea, and hazelnut, in Iran. The study focused on whether these models can learn from existing data, predict emissions effectively, and compare their performance.

After developing and testing both DT and ANN models, it was found that both could successfully estimate emissions, meeting the study's objectives.

The DT model performed best when the depth was set between 6 and 10. The RMSE obtained with depth 6 was 1664.84 and NAE 1124.79. When depth was set to 10, RMSE was 877.98 and NAE 533.52. A lower-depth model was favored as it generalizes better, avoiding overfitting while maintaining a balance between good performance results and ensuring the model's applicability to new datasets. It was found that the DT model can adapt itself depending on the relevant variables to predict the outcome of the variable selected as the label.

Overall, ANN1 presented lower errors, mean NAE of 0.07, when compared to ANN2, mean NAE of 0.14. The ANN1 model was particularly accurate in predicting variables like NH_3 and C_6H_6, with RMSE close to zero and high correlation values, indicating high precision and fit to the data. However, variables such as Cd, Cu, and Zn, presented greater challenges reflected in higher RMSE values.

Both the DT and ANN models demonstrated the capability to estimate emissions effectively, confirming that ML can be applied to LCI phase to automate prediction tasks and learn from data. ANN1 was preferable to ANN2 as increasing the number of hidden layers did not result in a significant improvement in error reduction.

The relationship between diesel fuel consumption and emissions such as N_2O, NH_3, and CO_2, confirmed the ability of the models to predict emissions based on input variables. As expected, N_2O emissions increased with diesel consumption, while NH_3 and CO_2 emissions also rose but at different rates, demonstrating the models' ability to capture varying behaviors.

Data analysis also revealed natural clusters which can be further explored using clustering algorithms.

The application of ML in the LCI phase has the potential to reduce the effort, time, and costs typically associated with manual data collection.

However, it should be noted that while the models were generally successful, there were still some challenges with accurately predicting certain variables, such as Zn and Cu, which may

require further model adjustment or additional data to improve prediction accuracy. Despite these issues, the overall performance of the models supports the conclusion that ML can aid the LCI phase by providing emission estimations.

In conclusion, the integration of Decision Trees and Artificial Neural Networks into the LCI phase shows promise for automating emission estimations. While both models performed well in most scenarios, the ANN models displayed superior prediction accuracy in many cases. On the other hand, the DT model provided a simpler, more interpretable approach that still offered good predictive power. For practical implementation, a balance between model simplicity and accuracy is ideal, depending on the specific requirements of the study.

Future research could explore the environmental impacts of agricultural activities across different regions, considering varying soil and climatic conditions. The potential of ANN and other ML techniques could be further explored by building new algorithms, evaluating their performance, and comparing them with the performance of the algorithms developed in this study.

Despite the Pesticide Model found in Appendix II of Agri-footprint 6 the variables B.9, B.10, and B.11 – biocide emissions – were not considered for this work as explained before. Future work could address the emissions of pesticides. It might be helpful to focus on these variables in future work to build a robust model and estimate those missing values.

Another recommendation is to gather data not only on the amount of fertilizers applied to crops but also on the chemical composition of these fertilizers, in order to be able to calculate the emissions resulting from the use of these chemicals more accurately.

This study did not account for variable related to water use, waste, or consumption. Given that, water is a crucial resource in agriculture. Future research could focus on databases with information on water consumption in crops.

Future work may involve group producers by their production process profile, conducting a comparative analysis of environmental indicators to identify differences and similarities in their environmental performance. This approach would help promote strategies to reduce pollution and waste in agriculture, thus reducing the ecological footprint and enhancing food quality for human health. This way, producers can align their practices with broader environmental goals, contributing to achieving sustainable objectives established, particularly at the European level.

References

1. European Environment Agency. Is Europe on track towards climate resilience? Status of reported national adaptation actions in 2023. 2023.

2. ONU Portugal. Objetivos de Desenvolvimento Sustentável [Internet]. 2022 [cited 2024 Jan 17]. Available from: https://unric.org/pt/objetivos-de-desenvolvimento-sustentavel/

3. THE 17 GOALS | Sustainable Development [Internet]. [cited 2024 Jun 8]. Available from: https://sdgs.un.org/goals

4. FoodLossWaste [Internet]. [cited 2024 Jul 14]. Events detail | Technical Platform on the Measurement and Reduction of Food Loss and Waste | Food and Agriculture Organization of the United Nations. Available from: https://www.fao.org/platform-food-loss-waste/flw-events/events/events-detail/2024-food-loss-and-waste-masterclass--the-basics/en

5. Portal do INE [Internet]. [cited 2024 Jun 8]. Available from: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&indOcorrCod=0011470&contexto=bd&selTab=tab2

6. Cederberg C, Sonesson U. Global food losses and food waste: extent, causes and prevention; study conducted for the International Congress Save Food! at Interpack 2011, [16 - 17 May], Düsseldorf, Germany. Gustavsson J, editor. Rome: Food and Agriculture Organization of the United Nations; 2011. 29 p.

7. Tiseo I. Statista. [cited 2024 Jul 26]. Global greenhouse gas emissions shares 2023, by sector. Available from: https://www.statista.com/topics/10348/agriculture-emissions-worldwide/

8. Agriculture and food system [Internet]. 2024 [cited 2024 Jul 21]. Available from: https://www.eea.europa.eu/en/topics/in-depth/agriculture-and-food

9. Hawkes C, Ruel M. Understanding the Links Between Agriculture and Health. International Food Policy Research Institute (IFPRI), 2020 vision briefs. 2006 Jan 1;

10. Garnett T. Food sustainability: problems, perspectives and solutions. Proceedings of the Nutrition Society. 2013 Feb;72(1):29–39.

11. Dekeyser K, Rampa F, D'Alessandro C, Bizzotto Molina P. The food systems approach in practice: Our guide for sustainable transformation | Knowledge for policy [Internet]. [cited 2024 Jul 21]. Available from: https://knowledge4policy.ec.europa.eu/publication/food-systems-approach-practice-our-guide-sustainable-transformation_en

12. Heydari M. Cultivating sustainable global food supply chains: A multifaceted approach to mitigating food loss and waste for climate resilience. Journal of Cleaner Production. 2024 Feb 25;442:141037.

13. Kuwornu JKM, Khaipetch J, Gunawan E, Bannor RK, Ho TDN. The adoption of sustainable supply chain management practices on performance and quality assurance of food companies. Sustainable Futures. 2023 Dec 1;5:100103.

14. Romeiko XX, Zhang X, Pang Y, Gao F, Xu M, Lin S, et al. A review of machine learning applications in life cycle assessment studies. Science of The Total Environment. 2024 Feb 20;912:168969.

15. Casolani N, D'Eusanio M, Liberatore L, Raggi A, Petti L. Life Cycle Assessment in the wine sector: A review on inventory phase. Journal of Cleaner Production. 2022 Dec 15;379:134404.

16. Pergola M, Persiani A, D'Ammaro D, Pastore V, D'Adamo C, Palese AM, et al. Environmental and Energy Analysis of Two Orchard Systems: A Case Study in Mediterranean Environment. Agronomy. 2022 Oct;12(10):2556.

17. Coşkun AE, Erturgut R. How Do Uncertainties Affect Supply-Chain Resilience? The Moderating Role of Information Sharing for Sustainable Supply-Chain Management. Sustainability. 2024 Jan;16(1):131.

18. Cavalcante de Oliveira R, Souza e Silva RD. Artificial Intelligence in Agriculture: Benefits, Challenges, and Trends. Applied Sciences. 2023 Jan;13(13):7405.

19. Sharifani K, Amini M. Machine Learning and Deep Learning: A Review of Methods and Applications [Internet]. Rochester, NY: Social Science Research Network; 2023 [cited 2024 Jan 3]. Available from: https://papers.ssrn.com/abstract=4458723

20. Romeiko X, Guo Z, Pang Y, Lee EK, Zhang X. Comparing Machine Learning Approaches for Predicting Spatially Explicit Life Cycle Global Warming and Eutrophication Impacts from Corn Production. Sustainability. 2020 Feb 17;12:19.

21. Ghoroghi A, Rezgui Y, Petri I, Beach T. Advances in application of machine learning to life cycle assessment: a literature review. International Journal of Life Cycle Assessment. 2022 Mar 1;27(3):433–56.

22. Zhao B, Shuai C, Hou P, Qu S, Xu M. Estimation of Unit Process Data for Life Cycle Assessment Using a Decision Tree-Based Approach. Environ Sci Technol. 2021 Jun 15;55(12):8439–46.

23. RETAILL's Consortium. REtail using Technology based on Artificial InteLLigence. 2022.

24. about - EUREKA Portugal 2021-22 [Internet]. 2021 [cited 2024 Jan 17]. Available from: https://eurekaportugal2021-22.pt/about/

25. Ecochain. Life Cycle Assessment (LCA) - Complete Beginner's Guide [Internet]. 2024 [cited 2024 Jan 21]. Available from: https://ecochain.com/blog/life-cycle-assessment-lca-guide/

26. Link between climate change and pollution: health implications [Internet]. 2021 [cited 2024 Apr 20]. Available from: https://climate-adapt.eea.europa.eu/en/observatory/evidence/health-effects/pollution

27. European Climate and Health Observatory [Internet]. [cited 2024 Oct 13]. Available from: https://climate-adapt.eea.europa.eu/en/observatory

28. Sources and emissions of air pollutants in Europe — European Environment Agency [Internet]. [cited 2024 Mar 2]. Available from: https://www.eea.europa.eu/publications/air-quality-in-europe-2022/sources-and-emissions-of-air

29. European Environment Agency [Internet]. [cited 2024 Sep 17]. The European environment — state and outlook 2020. Available from: https://www.eea.europa.eu/publications/soer-2020

30. European Platform on LCA | EPLCA [Internet]. [cited 2024 May 1]. Available from: https://eplca.jrc.ec.europa.eu/lifecycleassessment.html

31. ISO. ISO 14040:2006. 2006.

32. ISO [Internet]. [cited 2024 Jun 29]. ISO - About ISO. Available from: https://www.iso.org/about

33. Environment UN. UNEP - UN Environment Programme. 2017 [cited 2024 Jun 29]. GOAL 13: Climate action. Available from: http://www.unep.org/explore-topics/sustainable-development-goals/why-do-sustainable-development-goals-matter/goal-13

34. ISO. ISO 14044:2006. 2006.

35. Finkbeiner M, Ackermann R, Bach V, Berger M, Brankatschk G, Chang YJ, et al. Challenges in Life Cycle Assessment: An Overview of Current Gaps and Research Needs. In: Klöpffer W, editor. Background and Future Prospects in Life Cycle Assessment [Internet]. Dordrecht: Springer Netherlands; 2014 [cited 2024 Jun 29]. p. 207–58. Available from: https://doi.org/10.1007/978-94-017-8697-3_7

36. SimaPro [Internet]. [cited 2024 Aug 29]. SimaPro | LCA software for informed changemakers. Available from: https://simapro.com/

37. openLCA modeling suite | openLCA.org [Internet]. [cited 2024 Aug 29]. Available from: https://www.openlca.org/openlca/

38. ecoinvent Version 3.6 [Internet]. [cited 2024 Jun 10]. Available from: https://support.ecoinvent.org/ecoinvent-version-3.6

39. SimaPro Flow Tutorial [Internet]. [cited 2024 Aug 30]. Available from: https://support.simapro.com/s/article/SimaPro-Flow-Tutorial

40. Nemecek T, Roesch A, Bystricky M, Jeanneret P, Lansche J, Stüssi M, et al. Swiss Agricultural Life Cycle Assessment: A method to assess the emissions and environmental impacts of agricultural systems and products. Int J Life Cycle Assess. 2024 Mar 1;29(3):433–55.

41. Mérieux NutriSciences | Blonk | Optimeal® 2.0: insights and solutions for food issues [Internet]. [cited 2024 Aug 31]. Available from: https://blonksustainability.nl/news/optimeal-2-0

42. PRé Sustainability [Internet]. 2022 [cited 2024 Aug 29]. Life Cycle Assessment (LCA) explained. Available from: https://pre-sustainability.com/articles/life-cycle-assessment-lca-basics/

43. Goedkoop M, Heijungs R, De Schryver A, Struijs J, Van Zelm R. ReCiPe 2008 A life cycle impact assessment method which comprises harmonised category indicators at the midpoint and the endpoint level First edition Report I: Characterisation Mark Huijbregts 3). 2009.

44. Elhami B, Khanali M, Akram A. Combined application of Artificial Neural Networks and life cycle assessment in lentil farming in Iran. Information Processing in Agriculture. 2017 Mar 1;4(1):18–32.

45. Chappell D. Introducing Azure Machine Learning.

46. scikit-learn [Internet]. [cited 2024 Aug 31]. 3.4. Metrics and scoring: quantifying the quality of predictions. Available from: https://scikit-learn/stable/modules/model_evaluation.html

47. Performance (Regression) - Altair RapidMiner Documentation [Internet]. [cited 2024 Sep 1]. Available from: https://docs.rapidminer.com/2024.0/studio/operators/validation/performance/predictive/performance_regression.html

48. scikit-learn [Internet]. [cited 2024 Jun 15]. 1.10. Decision Trees. Available from: https://scikit-learn/stable/modules/tree.html

49. Gupta B, Rawat A, Jain A, Arora A, Dhami N. Analysis of Various Decision Tree Algorithms for Classification in Data Mining. International Journal of Computer Applications. 2017 Apr 17;163:15–9.

50. Tan PN, Steinbach M, Kumar V. Introduction to Data Mining. Pearson; 2014. 732 p.

51. Decision Tree - Altair RapidMiner Documentation [Internet]. [cited 2024 Jun 7]. Available from: https://docs.rapidminer.com/2024.0/studio/operators/modeling/predictive/trees/parallel_decision_tree.html

52. Google for Developers [Internet]. [cited 2024 Sep 3]. Árvores de decisão | Machine Learning. Available from: https://developers.google.com/machine-learning/decision-forests/decision-trees?hl=pt-br

53. Pekel E. Estimation of soil moisture using decision tree regression. Theor Appl Climatol. 2020 Feb 1;139(3):1111–9.

54. Costa VG, Pedreira CE. Recent advances in decision trees: an updated survey. Artif Intell Rev. 2023 May 1;56(5):4765–800.

55. Sathyadevan S, Nair RR. Comparative Analysis of Decision Tree Algorithms: ID3, C4.5 and Random Forest. In: Jain LC, Behera HS, Mandal JK, Mohapatra DP, editors. Computational Intelligence in Data Mining - Volume 1. New Delhi: Springer India; 2015. p. 549–62.

56. Quinlan JR. Induction of decision trees. Mach Learn. 1986 Mar 1;1(1):81–106.

57. Sabry F. Decision Tree Pruning: Fundamentals and Applications. One Billion Knowledgeable; 2023. 182 p.

58. Efficient Algorithms for Decision Tree Cross-validation. | Request PDF. ResearchGate [Internet]. 2024 Oct 22 [cited 2024 Dec 4]; Available from: https://www.researchgate.net/publication/220320021_Efficient_Algorithms_for_Decision_Tree_Cross-validation

59. ID3 – Altair RapidMiner Documentation [Internet]. [cited 2024 Jun 26]. Available from: https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/id3.html

60. Classification and Regression Trees (CART)-Classifier [Internet]. 2021 [cited 2024 Jun 26]. Available from: https://www.geo.fu-berlin.de/en/v/geo-it/gee/3-classification/3-1-methodical-background/3-1-1-cart/index.html

61. Schumacher D. SERP AI. 2023 [cited 2024 Aug 31]. Iterative Dichotomiser 3. Available from: https://serp.ai/iterative-dichotomiser-3/

62. Piasini E, Liu S, Chaudhari P, Balasubramanian V, Gold JI. How Occam's razor guides human decision-making. bioRxiv. 2023 Feb 8;2023.01.10.523479.

63. Wohlwend B. Decision Tree, Random Forest, and XGBoost: An Exploration into the Heart of Machine Learning [Internet]. Medium. 2023 [cited 2024 Jun 26]. Available from: https://medium.com/@brandon93.w/decision-tree-random-forest-and-xgboost-an-exploration-into-the-heart-of-machine-learning-90dc212f4948

64. What Is Random Forest? | IBM [Internet]. 2021 [cited 2024 Jun 26]. Available from: https://www.ibm.com/topics/random-forest

65. Faria BM. Mestrado de Bioestatística e Bioinformática Aplicadas à Saúde. 2021 Oct 29; Politécnico do Porto.

66. Dongare AD, Kharde RR, Kachare AD. Introduction to Artificial Neural Network. 2012;2(1).

67. Neural Net – Altair RapidMiner Documentation [Internet]. [cited 2024 Aug 26]. Available from: https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/neural_nets/neural_net.html

68. What is a Neural Network? | IBM [Internet]. 2021 [cited 2024 Sep 3]. Available from: https://www.ibm.com/topics/neural-networks

69. Qamar R, Zardari B. Artificial Neural Networks: An Overview. Mesopotamian Journal of Computer Science. 2023 Aug 2;2023:130–9.

70. Sharma S, Sharma S, Athaiya A. Activation Functions in Neural Networks. IJEAST. 2020 May 10;04(12):310–6.

71. Activation Functions: Sigmoid vs Tanh | Baeldung on Computer Science [Internet]. 2022 [cited 2024 Oct 7]. Available from: https://www.baeldung.com/cs/sigmoid-vs-tanh-functions

72. Hidden Layers in a Neural Network | Baeldung on Computer Science [Internet]. 2022 [cited 2024 Oct 14]. Available from: https://www.baeldung.com/cs/hidden-layers-neural-network

73. Coursera [Internet]. 2024 [cited 2024 Oct 14]. What Is a Hidden Layer in a Neural Network? Available from: https://www.coursera.org/articles/hidden-layer-neural-network

74. Emissões GEE | Agência Portuguesa do Ambiente [Internet]. [cited 2024 Mar 2]. Available from: https://apambiente.pt/clima/emissoes-gee

75. Sustainability B. Agri-footprint 6 Methodology Report.

76. Mérieux NutriSciences | Blonk | Agri-footprint [Internet]. [cited 2024 Aug 21]. Available from: https://blonksustainability.nl/tools-and-databases/agri-footprint

77. Huijbregts MAJ, Steinmann ZJN, Elshout PMF, Stam G, Verones F, Vieira M, et al. ReCiPe2016: a harmonised life cycle impact assessment method at midpoint and endpoint level. Int J Life Cycle Assess. 2017 Feb 1;22(2):138–47.

78. Mostashari-Rad F, Ghasemi-Mobtaker H, Taki M, Ghahderijani M, Kaab A, Chau K wing, et al. Exergoenvironmental damages assessment of horticultural crops using ReCiPe2016 and cumulative exergy demand frameworks. Journal of Cleaner Production. 2021 Jan 1;278.

79. Mostashari-Rad F, Ghasemi-Mobtaker H, Taki M, Ghahderijani M, Saber Z, Chau KW, et al. Data supporting midpoint-weighting life cycle assessment and energy forms of cumulative exergy demand for horticultural crops. Data Brief. 2020 Nov 4;33:106490.

80. Impact Categories (LCA) – Overview [Internet]. Ecochain. [cited 2024 Feb 13]. Available from: https://ecochain.com/blog/impact-categories-lca/

81. Intergovernmental Panel on Climate Change. The Intergovernmental Panel on Climate Change. 2006 [cited 2024 Feb 18]. IPCC Guidelines for National Greenhouse Gas Inventories Volume 4 Agriculture, Forestry and Other Land Use. Available from: https://www.ipcc-nggip.iges.or.jp/public/2006gl/index.html

82. Mérieux NutriSciences | Blonk | Previous Versions of Agri-footprint Methodology Reports [Internet]. [cited 2024 Aug 21]. Available from: https://blonksustainability.nl/previous-versions-of-agri-footprint-methodology-reports

83. Mousavi-Avval SH, Rafiee S, Sharifi M, Hosseinpour S, Notarnicola B, Tassielli G, et al. Application of multi-objective genetic algorithms for optimization of energy, economics and environmental life cycle assessment in oilseed production. Journal of Cleaner Production. 2017 Jan 1;140:804–15.

84. How much does human breathing contribute to climate change? [Internet]. [cited 2024 Jun 10]. Available from: https://www.sciencefocus.com/planet-earth/how-much-does-human-breathing-contribute-to-climate-change

85. Thomas Nemecek, Schnetzer J. Methods of assessment of direct field emissions for LCIs of agricultural production systems. 2011 Aug.

86. Software de Folha de Cálculo Online Gratuito: Excel | Microsoft 365 [Internet]. [cited 2024 Sep 11]. Available from: https://www.microsoft.com/pt-pt/microsoft-365/excel

87. IBM SPSS Software [Internet]. 2024 [cited 2024 Sep 11]. Available from: https://www.ibm.com/spss

88. RapidMiner Marketplace [Internet]. [cited 2024 Sep 11]. RapidMiner Studio. Available from: https://marketplace.rapidminer.com/UpdateServer/faces/facesContext.externalContext.requestURL

89. QuillBot: Your complete writing solution [Internet]. [cited 2024 Sep 11]. Available from: https://quillbot.com/

90.ChatGPT [Internet]. [cited 2024 Sep 11]. Available from: https://chatgpt.com/c/66e1f181-8414-800a-afd9-371877ba4616

91. Grammarly: Free AI Writing Assistance [Internet]. [cited 2024 Sep 11]. Available from: https://www.grammarly.com/

92. Pesticides use and trade, 1990–2022 [Internet]. [cited 2024 Oct 13]. Available from: https://openknowledge.fao.org/items/262b96c8-eef0-4810-9c23-d8639a5dbf1b

93. Deep Learning - RapidMiner Documentation [Internet]. [cited 2024 Sep 18]. Available from: https://docs.rapidminer.com/9.10/studio/operators/modeling/predictive/neural_nets/deep_learning.html

94. Maroco J. Análise Estatística Com Utilização do SPSS. 3ª. Lisboa: Edições Sílabo, Lda; 2007. 822 p.

95. Cabot MI, Lado J, Clemente G, Sanjuán N. Towards harmonised and regionalised life cycle assessment of fruits: A review on citrus fruit. Sustainable Production and Consumption. 2022 Sep 1;33:567–85.

96. Denitrification - an overview | ScienceDirect Topics [Internet]. [cited 2024 Sep 29]. Available from: https://www.sciencedirect.com/topics/earth-and-planetary-sciences/denitrification

97.UNL Water [Internet]. 2017 [cited 2024 Sep 29]. Pesticide Leaching & Runoff Management. Available from: https://water.unl.edu/article/crop-production/pesticide-leaching-runoff-management

98. Nabavi-Pelesaraei A, Abdi R, Rafiee S. Neural network modeling of energy use and greenhouse gas emissions of watermelon production systems. Journal of the Saudi Society of Agricultural Sciences. 2016 Jan 1;15(1):38–47.

99. Sabzevari A, Kouchaki-Penchah H, Nabavi-Pelesaraei A. Investigation of life cycle assessment of hazelnut production in Guilan province of I. R. Iran based on orchards size levels. Biological Forum-An International Journal. 2015 Apr 3;7.

100. Nabavi-Pelesaraei A, Sadeghzadeh A, Payman S, Ghasemi Mobtaker H. An analysis of energy use, CO2 emissions and relation between energy inputs and yield of hazelnut production in Guilan province of Iran. International Journal of Advanced Biological and Biomedical Research. 2013 Dec 19;1:1601–13.

101. Hicks A. Review of Global Tea Production and the Impact on Industry of the Asian Economic Situation. AU Journal of Technology [Internet]. 2001 [cited 2024 Sep 29];5(2). Available from: http://www.assumptionjournal.au.edu/index.php/aujournaltechnology/article/view/1174

102. Yang Z, Ma W, Lu J, Tian Z, Peng K. The Application Status and Trends of Machine Vision in Tea Production. Applied Sciences. 2023 Jan;13(19):10744.