

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EM GESTÃO E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO DO
CAMPUS SOROCABA

JOÃO VICTOR ENCIDE SALLA

**MODELO DE PREDIÇÃO DE IMPACTOS AMBIENTAIS DE PRODUTOS
BASEADO NA AVALIAÇÃO DO CICLO DE VIDA INTEGRADA À INTELIGÊNCIA
ARTIFICIAL**

Sorocaba
2025

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EM GESTÃO E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO DO
CAMPUS SOROCABA

JOÃO VICTOR ENCIDE SALLA

**MODELO DE PREDIÇÃO DE IMPACTOS AMBIENTAIS DE PRODUTOS
BASEADO NA AVALIAÇÃO DO CICLO DE VIDA INTEGRADA À INTELIGÊNCIA
ARTIFICIAL**

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção do *campus* Sorocaba, como parte dos requisitos para obtenção do título de Mestre em Engenharia de Produção.

Orientação: Prof. Dr. Diogo Aparecido Lopes Silva

Coorientação: Prof. Dr. Tiago Agostinho de Almeida

Financiamento: FAPESP, processo nº 2022/15134-1.

Sorocaba

2025

Salla, João Victor Encide

Modelo de predição de impactos ambientais de produtos baseado na avaliação do ciclo de vida integrada à inteligência artificial / João Victor Encide Salla -- 2025. 192f.

Dissertação (Mestrado) - Universidade Federal de São Carlos, campus Sorocaba, Sorocaba
Orientador (a): Diogo Aparecido Lopes, Silva
Banca Examinadora: Marília Ieda da Silveira Folegatti
Matsuura, Thiago Oliveira Rodrigues
Bibliografia

1. Inteligência artificial. 2. Desenvolvimento sustentável.
3. Predição de impactos. I. Salla, João Victor Encide. II. Título.

Ficha catalográfica desenvolvida pela Secretaria Geral de Informática (SIn)

DADOS FORNECIDOS PELO AUTOR

Bibliotecário responsável: Maria Aparecida de Lourdes Mariano -
CRB/8 6979



UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências em Gestão e Tecnologia
Programa de Pós-Graduação em Engenharia de Produção

Folha de Aprovação

Defesa de Dissertação de Mestrado do candidato João Victor Encide Salla, realizada em 22/04/2025.

Comissão Julgadora:

Prof. Dr. Diogo Aparecido Lopes Silva (UFSCar)

Prof. Dr. Thiago Oliveira Rodrigues (IBICIT)

Profa. Dra. Marília Ieda da Silveira Folegatti Matsuura (EMBRAPA)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Engenharia de Produção.

Dedico esse trabalho à minha família, especialmente, ao meu irmão, mãe e pai que me apoiaram a vida toda.

AGRADECIMENTO

Primeiramente, agradeço a Deus pelo dom da vida, por minha saúde e pela saúde de todos que amo.

Gostaria de agradecer a toda minha família pelo suporte e amor durante todos esses anos, especialmente, ao meu irmão, minha Mãe, meu Pai, minhas avós, meu afilhadinho, meus tios/tias e aos meus primos e primas.

Um agradecimento especial ao meu orientador e ao meu coorientador por todo apoio, por cada conhecimento compartilhado, pela paciência e pela grande disponibilidade em ajudar. Sou muito grato a vocês por tudo e desejo tudo de melhor em suas vidas.

Agradeço também ao grupo de pesquisa em Engenharia da Sustentabilidade (EngS) pela oportunidade e a todos os amigos do laboratório, do departamento e dos demais programas de pós, que tornaram esse processo inesquecível. Agradeço cada risada, cada assunto sério trocada e cada cafezinho pós-almoço. Agradeço a todos os Professores, técnicos e colaboradores que foram imprescindíveis nessa jornada.

Um agradecimento especial aos meus amigos fora do ambiente acadêmico (do jamba e da vida) que tornaram a vida mais leve através de risadas e de muita companhia nos momentos de dificuldade e também compartilhando sorriso nos momentos de felicidade e de lazer.

Amo todos vocês.

Por fim, agradeço muito à FAPESP, processo nº 2022/15134-1, pelo apoio durante esse período de estudo e pesquisa, pelo suporte nas atividades e pelo apreço da entidade para com o desenvolvimento da pesquisa no Estado de São Paulo.

“O que fazemos na vida ecoa na eternidade”

(Máximus Décimus Meridius)

RESUMO

SALLA, João Victor Encide. Modelo de predição de impactos ambientais de produtos baseado na avaliação do ciclo de vida integrada à inteligência artificial. 2025. Dissertação (Mestrado em nome do PPGEPP-So) – Universidade Federal de São Carlos, Sorocaba, 2025.

Estimam-se que 78% dos consumidores americanos e 70% dos brasileiros valorizam estilos de vida mais sustentáveis, principalmente em razão do bem-estar pessoal e por preocupação com gerações futuras. Os riscos climáticos são uma realidade e complexos de serem resolvidos, exigindo respostas rápidas e eficazes. Assim, surge a importância de integrar técnicas como a Avaliação do Ciclo de Vida (ACV) e Aprendizado de Máquina (AM) para apoiar decisões ambientais de forma rápida e confiável. A ACV, regulamentada pela ISO 14040 e ISO14044, demanda tempo e grandes volumes de dados para sua execução. Por outro lado, o AM, a partir de sua capacidade de processamento de dados e de geração de modelos preditivos, oferece uma solução eficaz para reduzir esse tempo de análise. Este estudo teve como objetivo propor e testar modelos de suporte à decisão, utilizando AM treinado com conjuntos de dados de ACV, para prever impactos ambientais. Esta pesquisa exploratória está dividida em formato de dois artigos principais. O primeiro revisou a literatura sobre técnicas computacionais e linguagens de modelagem para a predição de impactos ambientais, com o objetivo de avaliar o estado da integração ACV+AM. O segundo artigo aplicou os modelos de AM em três testes diferentes, utilizando conjuntos de dados de ACV e comparando os resultados das predições. O principal desafio desta pesquisa foi lidar com a heterogeneidade dos dados de uma ACV, uma vez que diferentes bases de dados utilizam formatos variados, como ecoSpold e ILCD, impactando a interoperabilidade. Softwares de código aberto, como o OpenLCA, foram utilizados para garantir a compatibilidade e portabilidade dos 1.219 conjuntos de dados avaliados por AM. O principal resultado foi um modelo refinado de suporte à decisão que permite executar todo o *pipeline* de treinamento dos modelos, desde o processo de geração de amostras a partir da exportação de dados em bases de ACV até a análise de desempenho dos modelos de AM. Espera-se que o modelo de suporte desenvolvido oriente usuários da ACV e AM, oferecendo uma abordagem replicável para outros setores e contextos.

Palavras-chave: avaliação do ciclo de vida; aprendizado de máquina; impacto ambiental; predição; regressão.

ABSTRACT

It is estimated that 78% of American consumers and 70% of Brazilian consumers already value more sustainable lifestyles, primarily due to personal well-being and concerns for future generations. However, climate risks are already a reality and complex to address, requiring swift and effective responses. This underscores the importance of integrating techniques such as Life Cycle Assessment (LCA) and Machine Learning (ML) to support environmental decision-making in a rapid and reliable manner. LCA, regulated by ISO 14040 and ISO 14044, requires significant time and large volumes of data for its execution. On the other hand, ML, with its data processing capabilities and predictive modeling, offers a solution to significantly reduce this analysis time. This study aimed to propose and test decision-support models using ML trained with LCA datasets to predict environmental impacts. This exploratory research is structured into two main articles. The first reviews the literature on computational techniques and modeling languages for environmental impact prediction, with the objective of assessing the current state of LCA+ML integration. The second article applies ML models in three different tests, utilizing LCA datasets and comparing prediction results. The primary challenge of this research was managing the heterogeneity of LCA data, as different databases use varied formats, such as ecoSpold and ILCD, impacting interoperability. Open-source software, such as OpenLCA, was employed to ensure compatibility and portability across the 1,219 datasets evaluated using ML. The key outcome was a refined decision-support model that enables the entire training pipeline, from sample generation through data export from LCA databases to ML model performance analysis. It is expected that the developed support model will guide LCA and ML users, offering a replicable approach for other sectors and contexts.

Keywords: life cycle assessment; machine learning; environmental impact; prediction; regression.

LISTA DE FIGURAS

Figura 1. 1 - Estrutura do trabalho.	19
Figure 2. 1 - Methodological procedure.	23
Figure 2. 2 - PRISMA Flowchart: stages of the review process.	26
Figure 2. 3 - Main LCA databases.	27
Figure 2. 4 - Geographical distribution of publications.	29
Figure 2. 5 - Areas of application of the studies and the main predicted variables.	32
Figure 2. 6 - Arrangement of ML methods by sector.	34
Figure 2. 7 - Arrangement of LCA system boundaries by sector (a) and by ML method (b).	38
Figure 2. 8 - Analysis of the number of independent variables by sector.	39
Figure 2. 9 - Analysis of the number of samples (a) per sector and (b) per machine learning method.	42
Figure 2. 10 - Analysis of the number of target attributes by sector (a) and by ML method (b).	44
Figure 2. 11 - Arrangement of ML methods by sector for full LCA (own authorship) - in the number of occurrences.	46
Figura 3. 1 - Pipeline do processo da ACV+AM realizado neste trabalho.	81
Figura 3. 2 - Pipeline de treinamento dos modelos de AM para os Testes 1 e 2.	86
Figura 3. 3 - Pipeline de treinamento dos modelos de AM para o Teste 3.	89
Figura 3. 4 - Fatores de caracterização zerados.	91
Figura 3. 5 - Sistema de produto sem a conexão com os <i>providers</i>	91
Figura 3. 6 - Matriz de correlação dos atributos de entrada.	93
Figura 3. 7 - Matriz de correlação dos atributos de entrada após as remoções.	95
Figura 3. 8 - Estrutura para predição de potenciais impactos ambientais para o Teste 2.	101
Figura 3. 9 - Matriz de correlação dos atributos de entrada (<i>Aluminium e Steel</i>).	102
Figura 3. 10 – Disposição dos dados disponíveis por versão da base.	109
Figura 3. 11 - Seleção da categoria <i>acidification</i> e detecção de <i>outliers</i>	122
Figura 3. 12 - Desempenho dos modelos para a categoria <i>acidification</i>	123
Figura 3. 13 - Desempenho dos modelos (R2 e MSE teste) para a categoria <i>acidification</i>	124
Figura 3. 14 - Testagem do melhor modelo (valor real vs valor predito).	124

LISTA DE TABELAS

Table 2. 1 - Search strings: Scopus and WoS.....	24
Table 2. 2 - Synthesis table – The most relevant ML models in the context of LCA.	52
Tabela 3. 1 - Inventários de milho utilizados do SICV Brasil.....	90
Tabela 3. 2 - Tabela com os fluxos contidos nos <i>datasets</i> (atributos).....	92
Tabela 3. 3 - Análise descritiva dos atributos-alvo para o teste 1.	96
Tabela 3. 4 - Parâmetros independentes utilizados em cada <i>dataset</i>	97
Tabela 3. 5 - Parâmetros dependentes utilizados em cada dataset.	98
Tabela 3. 6 – Análise descritiva dos atributos-alvo para o teste 2.	103
Tabela 3. 7 - Dados de inventário usados para geração amostras e suas respectivas fontes.	104
Tabela 3. 8 - Padronização da nomenclatura das amostras do Teste 3.....	107
Tabela 3. 9 – Padronização da nomenclatura das amostras do teste 3.	110
Tabela 3. 10 - Resultados para o Teste 1.....	116
Tabela 3. 11- Análise de <i>overfitting</i> para o teste 1.....	117
Tabela 3. 12 - Resultados para o teste 2.....	118
Tabela 3. 13 - Resultados para o teste 2.....	119
Tabela 3. 14 - Resultados preliminares para o teste 3.	126
Tabela 3. 15 - Resultados para o Teste 3.....	130

LISTA DE ABREVIATURAS E SIGLAS

PORTUGUÊS

ACV	Avaliação do ciclo de vida
ACV+AM	Avaliação do ciclo de vida + Aprendizado de máquina
AICV	Avaliação de impacto do ciclo de vida
AM	Aprendizado de máquina
DAP	Declaração Ambiental de Produto
FGV	Fundação Getúlio Vargas
GEE	Gases do efeito estufa
IA	Inteligência artificial
ICV	Inventário do ciclo de vida
PIB	Produto interno bruto

INGLÊS

AdaBoost	AdaBoost machine learning method
ANFIS	Adaptive Neuro-Fuzzy Inference System
ANFIS-FCM	Adaptive Neuro-Fuzzy Inference System with Fuzzy C-Means
ANN	Artificial neural network
BAYSEIAN	Bayesian machine learning
CatBoost	Category boosting
CNN	Convolutional neural network
DELM	Deep extreme learning machine
DF	Decision forest
DL	Deep learning architecture
DT	Decision tree
E_net	Elastic net
EPR	Evolutionary polynomial regression
FCM	Fuzzy C-Means Clustering

GBDT	Gradient-boosted decision trees
GBM	Gradient Boosting Machine
GBR	Gradient Boosting Regressor
GBRT	Gradient Boosted Regression Trees
GPR	Gaussian process regression
GRNN	Generalized regression neural network
KNN	K-Nearest Neighbors
KRR	Kernel ridge regression
LASSO	Lasso Regression
LR	Linear regression
MLP	Multi-layer perceptron
MLR	Multiple linear regression
MPR	Multivariate polynomial regression
PCR	Principal component regression
RF	Random forest
RR	Ridge Regression
SGD	Stochastic gradient descent
SVM	Support vector machine
SVR	Support vector regressor
XGBoost	eXtreme gradient boosting
EPD	Environmental Product Declaration
IPCC	Intergovernmental Panel on Climate Change
ISO	International Organization for Standardization

SUMÁRIO

1 INTRODUÇÃO	15
1.1 CONTEXTUALIZAÇÃO E JUSTIFICATIVA.....	15
1.2 OBJETIVOS.....	18
1.3 ESTRUTURA DO TRABALHO	19
2 ARTIGO 1.....	20
2.1 INTRODUCTION.....	21
2.2 METHODS.....	23
2.3 RESULTS.....	28
2.4. DISCUSSION.....	48
2.5. CONCLUSIONS	50
REFERENCES	56
APENDIX – 2A.....	63
3. ARTIGO 2.....	75
3.1 INTRODUÇÃO.....	76
3.2 MATERIAIS E MÉTODOS.....	78
3.2.1 ABORDAGENS METODOLÓGICAS UTILIZADAS.....	84
3.2.1.1 TESTES 1 E 2	84
3.2.1.2 TESTES 3	87
3.3 RESULTADOS E DISCUSSÃO	90
3.3.1 DESCRIÇÃO DOS RESULTADOS OBTIDOS COM OS FLUXOS METODOLÓGICOS.....	90
3.3.1.1 RESULTADOS DO TESTE 1	90
3.3.1.2 RESULTADOS DO TESTE 2	96
3.3.1.3 RESULTADOS DO TESTE 3	104
3.3.2 RESULTADOS PARA O TREINAMENTO DOS MODELOS DE AM.....	112
3.3.2.1 TESTE 1	115
3.3.2.2 TESTE 2	118
3.3.2.3 TESTE 3	120
3.4 CONCLUSÕES.....	133
APÊNDICE – 3A	136
APÊNDICE – 3B.....	143
APÊNDICE – 3C.....	168
REFERÊNCIAS	176
4 CONSIDERAÇÕES FINAIS.....	182
5 CONCLUSÕES.....	186
REFERÊNCIAS	189
GLOSSÁRIO.....	191

1 INTRODUÇÃO

O capítulo de introdução é formado pela contextualização e justificativa (Seção 1.1), pelos objetivos (Seção 1.2) e pela descrição da estrutura adotada neste trabalho (Seção 1.3). Vale destacar que esta dissertação foi elaborada no formato de artigo, sendo composta por dois trabalhos conforme a estrutura apresentada na Seção 1.3.

1.1 CONTEXTUALIZAÇÃO E JUSTIFICATIVA

O mundo está se tornando cada vez mais dinâmico, e, como resultado, o processo de tomada de decisões nas empresas precisou acelerar para acompanhar esse ritmo. A preocupação com o meio ambiente e a busca por produtos sustentáveis estão se tornando fatores cada vez mais valorizados pelos consumidores no momento da compra.

Um estudo recente da NielsenIQ, empresa líder mundial em inteligência do consumidor, identificou que 78% dos consumidores norte-americanos consideraram importante adotar um estilo de vida mais sustentável (McKinsey&Company, 2023). De forma similar, é observado tal mudança de comportamento nos consumidores brasileiros em uma pesquisa feita pela Fundação Getúlio Vargas (FGV) com 1.034 entrevistados de classes diversas, na qual 70% dos mesmos afirmaram que optaram por tais mudanças por preocupações com o próprio bem-estar, e que 65% incluíram tais questões nas suas vidas em virtude do temor e preocupação para com as gerações futuras (Hypeness, 2019).

Estudos apontam que riscos climáticos e não-climáticos serão eminentes, múltiplos e interagirão entre si, resultando em riscos cumulativos e em cascata para diversos setores e regiões (Calvin et al., 2023). De acordo com o *Intergovernmental Panel on Climate Change* (IPCC), os riscos atuais relacionados ao clima são mais elevados do que nas avaliações anteriores à 2023, sendo os impactos à longo prazo, inúmeras vezes maiores que os atualmente observados (Calvin et al., 2023).

Nesse sentido, surge a importância de técnicas que permitam conciliar às questões ambientais à elevada velocidade de resposta requerida pelos *stakeholders* no processo de tomada de decisão, seja no momento de escolher um produto com melhor performance ambiental (caso do consumidor) ou até mesmo para definir um material de menor carga ambiental no desenvolvimento de produto (caso das empresas).

Assim, se destaca a importância da integração ACV+AM (Avaliação do ciclo de vida + aprendizado de máquina) ou LCA+ML (*life cycle assessment + machine learning*) para tais questões, conforme destacado nas revisões da literatura feitas recentemente por Romeiko et al. (2024), Martínez-Ramón et al. (2024) e Salla et al. (2025).

A ACV, como o próprio nome sugere, consiste em avaliar de forma quantitativa os potenciais impactos ambientais de um bem, serviço ou processo dentro em uma dada fronteira de análise, ou seja, ao longo do ciclo de vida (Oliveira et al., 2021). Por se tratar de estudos com elevada gama de informações e grande volume de dados, uma ACV deve seguir normas e diretrizes que regem sua realização. Com base nos princípios da norma 14040 da *International Organization for Standardization* (ISO, 2006) há quatro etapas na ACV:

- i. Definição de objetivo e escopo: necessária para reconhecer os limites do sistema e a unidade funcional;
- ii. A elaboração do inventário do ciclo de vida (ICV), contendo as entradas e saídas (fluxos de materiais e energia) relevantes ao longo do ciclo de vida estudado. Etapa que costumeiramente é a que demanda maior dispêndio de tempo e recursos para sua consolidação;
- iii. Etapa de avaliação de impacto do ciclo de vida (AICV), objetivando quantificar a contribuição do ICV para potenciais impactos ambientais em dadas categorias de impacto escolhidas e;
- iv. Fase de interpretação dos resultados, na qual são feitas análises, recomendações e conclusões do estudo.

O processo de condução de uma ACV costuma ser lento e levar meses para finalização, o que acaba representando um limitante de tempo para as empresas. Nesse sentido, surge a importância do aprendizado de máquina (AM) ou *machine learning* (ML), área dentro da inteligência artificial (IA), que possibilita a utilização de máquinas para propor soluções rápidas à sistemas e decisões que exijam análises complexas (Faceli et al., 2022). Nesse sentido, modelos matemáticos e estatísticos são projetados para aprender a partir de dados existentes, a fim de melhorar o desempenho futuro, sendo uma das principais vias para permitir parear a capacidade das máquinas ao cérebro humano (Koyampambath et al., 2022).

O processo de aprendizado, ocorre em formato supervisionado ou não-supervisionado. O primeiro, é caracterizado pelo uso de conjuntos de dados rotulados para treinar algoritmos que aprendem padrões para classificar dados (classificação) ou para prever resultados com precisão (regressão), enquanto o segundo, por sua vez, usa algoritmos para analisar e propor

sugestões de agrupamento a conjuntos de dados não rotulados, permitindo o estabelecimento de padrões sem a necessidade de intervenção humana (IBM, 2024b, 2024a).

Existem várias técnicas de AM que podem ser utilizadas para a aprendizagem das máquinas integrada a ACV no contexto da predição de potenciais impactos (regressão), possuindo cada uma delas benefícios e limitações. Para possibilitar a integração entre inteligência artificial e ACV, é necessário que haja um volume considerável de dados a fim de permitir amostras o suficiente para os algoritmos aprenderem.

No entanto, a quantidade de conjuntos de dados disponíveis soma apenas na casa dos milhares de unidades, sendo que tipicamente estudos de AM trabalham com até milhões de dados para estudos de aprendizado profundo (Brownlee, 2019). Além disso, sabe-se que grande parte das bases de dados de ACV são fechadas (e.g., Ecoinvent, Gabi, Agrifootprint) impedindo a obtenção de amostras suficientes para tal e dificultando o processo de AM.

Outro fator que acaba dificultando esse processo está relacionado a natureza dos arquivos em si, que a depender da base dados varia a extensão. Salla et al. (2025) observou que esses conjuntos de dados (contidos nas bases) variam em formato, como ecoSpold (ecoinvent, PeruLCA), ILCD (base de dados SICV Brasil) e JSON-LD (OpenLCA), dependendo da base de dados ou do software de ACV utilizado, o que pode afetar a interoperabilidade e complicar a compilação de inventários para o treinamento de algoritmos de AM.

O Sphera's LCA for Experts (antigo GaBi - <https://lca-software.org/sphera-gabi/>) é um dos softwares de ACV mais amplamente utilizado, com suporte para a importação de pacotes de dados no formato ILCD. No entanto, recentemente apresentou constantes problemas no processo de importação completa dos dados de ACV da Tiangong, base de dados desenvolvida na China, que fornece dados ambientais específicos para processos e produtos industriais chineses (Qi, 2023). Especialistas suspeitam que tal processo tenha ocorrido devido aos novos fluxos criados não estarem no banco de dados do Gabi ou até mesmo devido alguma desconfiguração de formato durante a importação (Qi, 2023a).

Nesse sentido, recomenda-se o uso de softwares com código aberto e gratuito. De acordo com experimentos realizados, o OpenLCA tem a melhor compatibilidade com o formato ILCD e as funções de importação e exportação mais utilizáveis quando comparados aos softwares de código fechado, como SimaPro e GaBi (Qi, 2023b). Além disso, o OpenLCA não possui projetos significativos de barreira à migração de dados e sua compatibilidade e portabilidade de dados são relativamente boas (Qi, 2023b), além de possuir funcionalidades de desenvolvedor aos próprios usuários.

Diante dessa oportunidade de integração e dos desafios encontrados em tal processo se destaca a proposta do presente trabalho, conforme os objetivos de pesquisa a seguir expostos.

1.2 OBJETIVOS

O objetivo geral desta pesquisa é:

- Propor e testar modelos de suporte à decisão baseados em aprendizado de máquina e treinados com conjuntos de dados de ACV para realizar a predição de potenciais impactos ambientais.

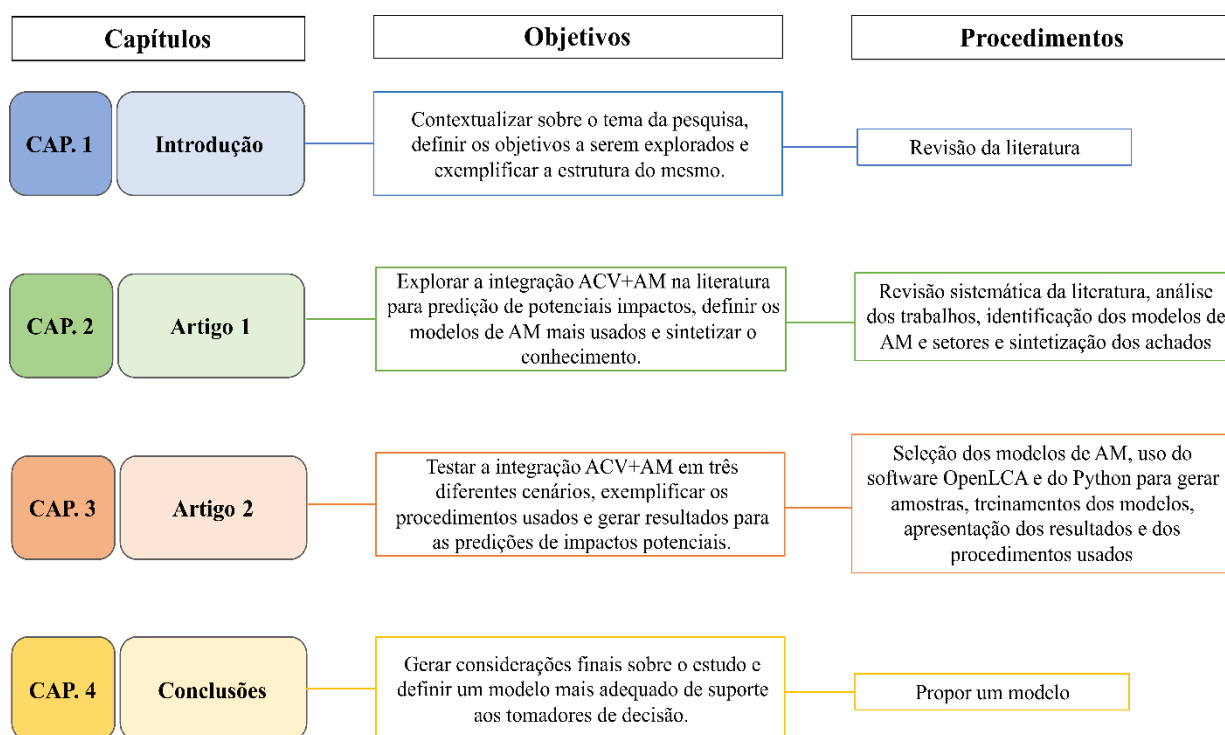
A partir do objetivo geral, são elencados como objetivos específicos:

- Revisar a literatura sobre técnicas e linguagens de modelagem computacional, com o objetivo de definir as que sejam mais adequadas para a predição de potenciais impactos ambientais - Artigo 1.
- Desenvolver um procedimento de compilação e normalização de conjuntos de dados de estudos de ACV para aplicação em AM - Artigo 2.
- Realizar testes com modelos de predição de AM em conjuntos de dados nacionais de ACV - Artigo 2.
- Propor um modelo de suporte a decisão com base na eficiência do melhor modelo de AM testado.

1.3 ESTRUTURA DO TRABALHO

Este trabalho é dividido em quatro capítulos que, exceto o capítulo introdutório e o capítulo final (conclusões), estão formatados como artigos. O Capítulo 2 aborda uma revisão bibliográfica sistemática sobre o tema de pesquisa, publicado na revista “*The International Journal of Life Cycle Assessment*” na chamada especial “*How Digital Technologies Could Empower Life Cycle Assessment Studies*”. O artigo seguinte será submetido, posteriormente, em alguma revista indexada na área. A Figura 1.1 apresenta a estrutura geral do trabalho, destacando o objetivo e a metodologia de cada capítulo desenvolvido.

Figura 1. 1 - Estrutura do trabalho.



*Legenda: ACV: avaliação do ciclo de vida; AM: aprendizado de máquina.

Fonte: autoria própria.

2 ARTIGO 1

Reproduzido com permissão da Springer Nature: Salla, J.V.E., de Almeida, T.A. & Silva, D.A.L., *Integrating machine learning with life cycle assessment: a comprehensive review and guide for predicting environmental impacts*, International Journal of Life Cycle Assessment, 2025, Springer Nature, reproduzido com permissão da SNCSC.

Integrating Machine Learning with Life Cycle Assessment: A Comprehensive Review and Guide for Predicting Environmental Impacts

Abstract:

Purpose

Life Cycle Assessment (LCA) is an essential tool for evaluating the environmental impacts of products and processes, yet its integration with Machine Learning (ML) remains underexplored. This paper addresses the critical gap in the literature by analyzing how AI-enabled tools are incorporated into LCA methodology to predict potential environmental impacts. Our research aims to provide clear guidance on selecting the most appropriate ML methods tailored to the unique goals and specificities of LCA studies.

Methods

The research methodology involves a comprehensive literature review using targeted search strings to identify relevant studies. A bibliometric analysis was conducted to understand the publication trends. The identified studies were screened and categorized based on ML methods and their applications in LCA. A detailed content analysis was performed to evaluate the integration of ML in different phases of LCA, focusing on regression, classification, clustering, and other tasks. The findings were compiled into a synthesis table to highlight the effectiveness and computational efficiency of various ML models.

Results and discussion

The analysis revealed that Artificial Neural Networks (ANN) and Random Forest (RF) are the most frequently used ML models in LCA integration, primarily for regression and classification tasks. The study also found a significant gap in real-time dynamic LCA models, with existing models largely static. The synthesis table provides a comparative overview of ML models, indicating their effectiveness, sensitivity, and computational efficiency. ANN emerged as the most effective model, while Bayesian networks showed potential for further exploration due to their specific characteristics.

Conclusions

Integrating ML with LCA enhances the accuracy and efficiency of environmental impact predictions. However, the current use of ML in LCA is limited by data quality and the static nature of traditional LCA models. Future research should focus on developing dynamic LCA models and improving data quality to leverage the full potential of ML. Recommendations include exploring underutilized ML models and extending AI-enabled LCA applications to various sectors, including social contexts and the tertiary sector.

Keywords: environmental impact prediction; life cycle management; artificial intelligence; neural networks; life cycle analysis; model selection; sustainability.

2.1 INTRODUCTION

The increasing frequency and severity of disasters erode the resilience of socio-ecological systems due to the depletion of assets and livelihoods. According to GRO (2024), if there is no urgent and coordinated action to change how resources are utilized, resource extraction could increase by almost 60% by 2060 compared to 2020 levels, resulting in serious risks to ecosystems and human health. These destabilizing effects on long-term prosperity and environmental protection highlight the need for developing relevant autonomous and AI-enabled tools to promote sustainable development (Vinuesa et al., 2020).

A notable signal of change is the surge in fossil-fuel subsidies, which reached a record \$US7 trillion in 2023 as governments responded to the global spike in energy prices, directing finance towards sustainable resource use and the cost-competitiveness of renewable energy (IMF, 2023). According to a Harvard Business Review study, products with sustainability claims exhibit twice the sales growth compared to their traditional counterparts, and 65% of interviewed consumers prefer products that advocate sustainability (White et al., 2019).

In this context, the importance of developing techniques that enable machines to provide rapid solutions to complex systems and decisions, particularly those addressing the growing demand for sustainable products, becomes evident (Faceli et al., 2022). Life Cycle Assessment (LCA) quantifies the potential environmental aspects and impacts of a product, process, or organization throughout its life cycle, facilitating performance comparisons and informed decision-making (Oliveira et al., 2021). However, traditional life cycle models are static and do not permit real-time prediction of impacts using the current LCA software tools and databases. Dynamic LCAs are being studied (da Costa et al., 2024), but they are still scarce and need thorough evaluation before supporting the development of life cycle inventories with real-time monitoring technologies.

Artificial intelligence (AI), as defined by Das and Behera (2017), encompasses processes aimed at enabling machines to exhibit intelligent behavior. Machine Learning (ML), a subfield of AI, leverages machines' ability to learn from large datasets to solve tasks. The emergence of ML marks a significant advancement in AI, facilitating predictive modeling, pattern recognition, and decision-making (Faceli et al., 2022). AI-enabled resource management can optimize the use of natural resources through advanced algorithm applications for demand prediction, uncertainty detection, and sustainable decision-making (Chen et al., 2023; Kamyab et al., 2023; Vinuesa et al., 2020; Wu and Shang, 2020). However, the efficacy of these technologies hinges on the quality of available data, which can impede the rapid

quantification and analysis of environmental management for decision-making purposes (Romeiko et al., 2024).

ML algorithms can be categorized into three groups based on their learning nature: supervised learning, unsupervised learning, and reinforcement learning (Faceli et al., 2022; Ledziński and Grzešek, 2023). In the context of LCA integration (LCA+ML), various ML models are used for tasks such as regression, classification, and clustering (Akhshik et al., 2022).

Supervised learning methods, such as classification and regression, are commonly used for predictive tasks, where algorithms learn from training datasets to make predictions faster than traditional LCAs (Faceli et al., 2022). A review by Romeiko et al. (2024) identified 40 studies integrating LCA with ML between 2010 and 2020, with 87.5% focused on value prediction (regression) and the remainder divided into dimensionality reduction, attribute ranking, classification, and clustering tasks.

The literature demonstrates that ML models can enhance three of the four phases of LCA: Life Cycle Inventory (LCI), Life Cycle Impact Assessment (LCIA), and interpretation. ML aids in estimating missing or incomplete primary and secondary data in the LCI phase, assessing potential impacts and characterization factors in the LCIA phase, and optimizing parameters, identifying influential attributes, and evaluating result uncertainties in the interpretation phase (Romeiko et al., 2024).

With the increasing need to comprehend product environmental impacts, stakeholders across product value chains seek structures capable of rapidly and reliably estimating environmental impacts and making decisions, underscoring the importance of LCA+ML integration (Koyamparambath et al., 2022). Consequently, developing and applying more techniques that merge the need for speed in decision-making with environmental concerns is crucial for moving towards more dynamic LCA applications than static ones (da Costa et al., 2024; Romeiko et al., 2024).

This paper addresses the limited number of critical reviews on LCA+ML currently exploring the following research question: how are autonomous and AI-enabled tools integrated with LCA methodology to predict potential environmental impacts? This question is posed due to the need for more guidance on selecting the best ML method based on literature reviews tailored to the goals and specificities of LCA studies.

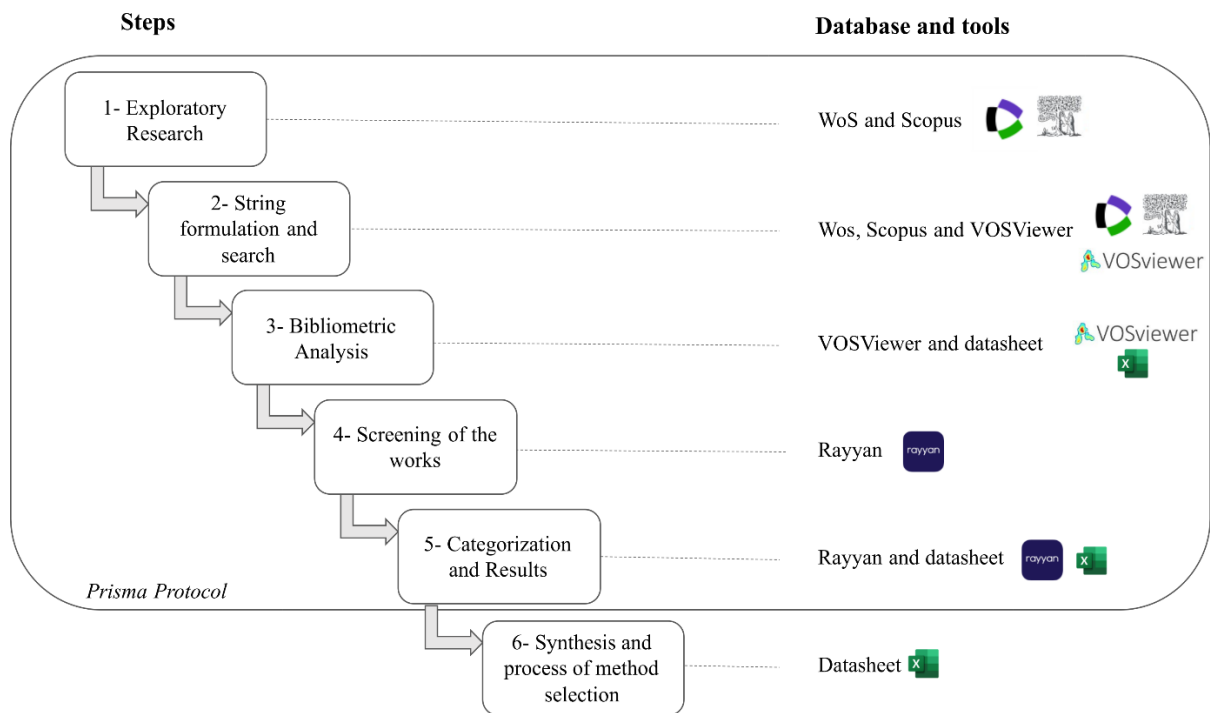
The remainder of the paper is organized as follows: Section 2 details the research methodology and search strings; Section 3 presents the main results, including bibliometric analysis, screening, categorization, and content analysis of existing studies; Section 4 discusses

the results, encompassing the main findings and their compilation in a synthesis table; and Section 5 summarizes the main conclusions, challenges, and provides directions for future research.

2.2 METHODS

We conducted a comprehensive systematic literature review (Denyer and Tranfield, 2009) following the methodological procedure outlined in Figure 2.1, employing the PRISMA methodology to ensure transparent, comprehensive, and accurate reporting (Page et al., 2021). The Web of Science (WoS) and Scopus databases were selected as primary sources, and supplementary materials were made available online to enable access to the complete datasets used in this study.

Figure 2. 1 - Methodological procedure.



Source: own authorship.

2.2.1 Search strategy

Preliminary searches used the terms "Life Cycle Assessment" and "Artificial Intelligence" along with "predict" in the title and abstract fields. Subsequently, an interactive process with the VOSViewer tool was employed to include terms correlated with the initial string. VOSViewer facilitates the construction of keyword maps based on co-occurrence data and uses the visualization of similarities technique to create distance-based maps (Filho and

Russo, 2018; Van Eck and Waltman, 2010). This process was repeated until no further correlated terms were identified, resulting in the final search strings (Table 2.1) and a similarity map (Figure 2.2) provided as Supplementary Material.

Table 2. 1 - Search strings: Scopus and WoS.

Database	String	Data
Scopus	TITLE-ABS (("life cycle assessment*" OR "life-cycle assessment*" OR "life cycle analysi*" OR ("life cycle* " AND "environmental impact*") OR ("life cycle* " AND "environmental-impact*") OR ("greenhouse gas*" AND "life cycle* ") OR ("greenhouse-gas*" AND "life cycle* ") OR ("GHG emission*" AND "life cycle* ") OR ("carbon emission*" AND "life cycle* ") OR ("carbon footprint*" AND "life cycle* ")) AND ("artificial intelligence" OR "machine learning" OR "artificial neural network*" OR "machine-learning" OR "learning algorithm*" OR "neural network*" OR "artificial neural-network*" OR "neural-network*" OR "random forest*" OR "decision-tree*" OR "decision tree*" "random-forest*" OR "fuzzy system*" OR "deep learning" OR "deep-learning") AND ("predict*" OR "forecast*" OR "prognosis*"))	Feb/ 2024
WoS	(TI=((("life cycle assessment*" OR "life-cycle assessment*" OR "life cycle analysi*" OR ("life cycle* " AND "environmental impact*") OR ("life cycle* " AND "environmental-impact*") OR ("greenhouse gas*" AND "life cycle* ") OR ("greenhouse-gas*" AND "life cycle* ") OR ("GHG emission*" AND "life cycle* ") OR ("carbon emission*" AND "life cycle* ") OR ("carbon footprint*" AND "life cycle* ")) AND ("artificial intelligence" OR "machine learning" OR "artificial neural network*" OR "machine-learning" OR "learning algorithm*" OR "neural network*" OR "artificial neural-network*" OR "neural-network*" OR "random forest*" OR "decision-tree*" OR "decision tree*" "random-forest*" OR "fuzzy system*" OR "deep learning" OR "deep-learning") AND ("predict*" OR "forecast*" OR "prognosis*")))) OR AB=((("life cycle assessment*" OR "life-cycle assessment*" OR "life cycle analysi*" OR ("life cycle* " AND "environmental impact*") OR ("life cycle* " AND "environmental-impact*") OR ("greenhouse gas*" AND "life cycle* ") OR ("greenhouse-gas*" AND "life cycle* ") OR ("GHG emission*" AND "life cycle* ") OR ("carbon emission*" AND "life cycle* ") OR ("carbon footprint*" AND "life cycle* ")) AND ("artificial intelligence" OR "machine learning" OR "artificial neural network*" OR "machine-learning" OR "learning algorithm*" OR "neural network*" OR "artificial neural-network*" OR "neural-network*" OR "random forest*" OR "decision-tree*" OR "decision tree*" "random-forest*" OR "fuzzy system*" OR "deep learning" OR "deep-learning") AND ("predict*" OR "forecast*" OR "prognosis*"))))	Feb/ 2024

Source: own authorship.

The search encompassed articles in all languages and from any year.

2.2.2 Screening of the works

We utilized Rayyan software to support the screening of articles, enhancing the efficiency of selecting relevant papers integrating LCA+ML (Ouzzani et al., 2016). Rayyan facilitates article classification and inclusion/exclusion based on predefined criteria, streamlining peer review.

The search returned 35 works from Scopus and 123 from WoS. After excluding 24 duplicates, 134 articles remained for title and abstract screening, and 77 works were removed due to the following factors:

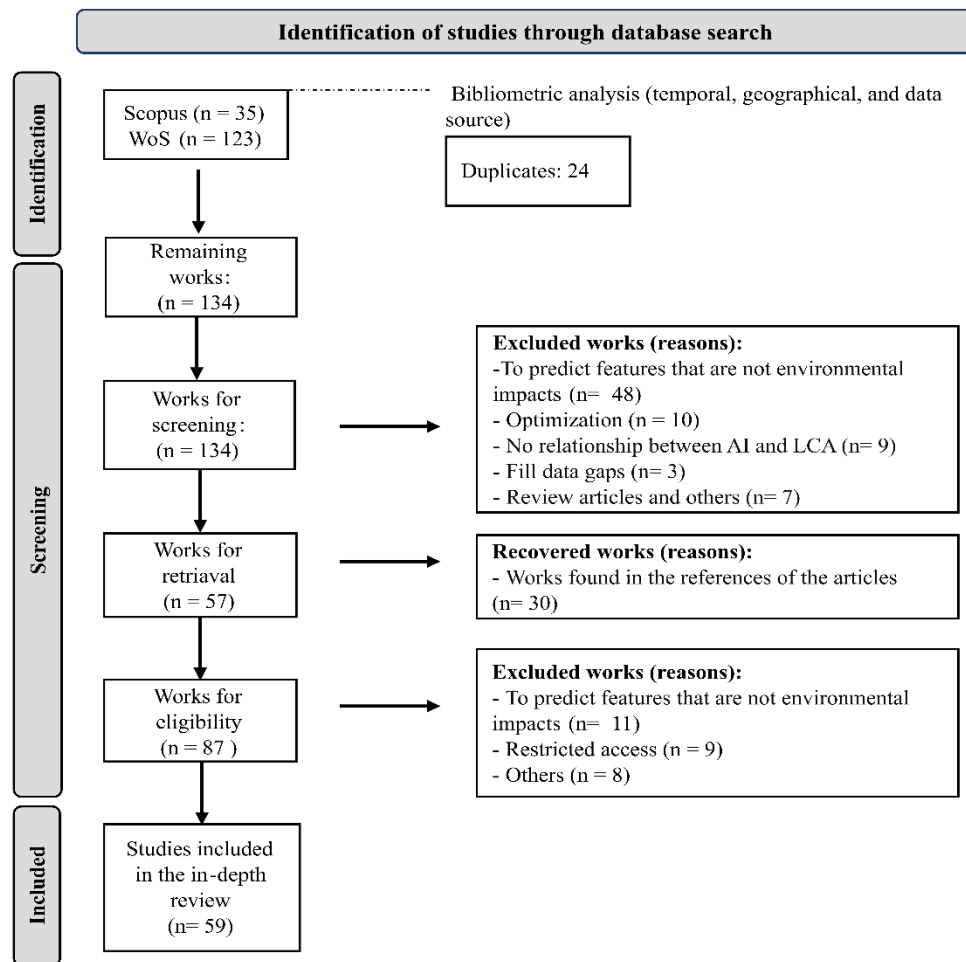
- i. Focus on predicting characteristics/parameters rather than potential environmental impacts (48 results);
- ii. Focus on LCA+ML integration for variable optimization (10 results);
- iii. Lack of integration between AI and LCA (9 results);
- iv. Review articles (5 results);
- v. Focus on filling inventory data gaps (3 results); and
- vi. Aim to verify process compliance (2 results).

Additionally, 30 articles were added through reference analysis. Further screening based on introduction, objectives, and conclusions led to the removal of 28 more articles: 11 focusing on predicting characteristics/parameters rather than potential environmental impacts, 9 with restricted access, and 8 for different reasons (duplicate (4), focusing on finding the most impactful variable within a group of variables (2), focusing on optimizing variables/parameters (1), and clustering (1)). After further analysis, 59 studies were selected for in-depth review in Section 2.3.

Figure 2.2 shows an overview of the PRISMA methodology adopted. Only 37.3% of the initial samples were considered relevant to the current use of AI and LCA to predict environmental impacts. There is a divergent and heterogeneous use of autonomous and AI-enabled tools integrated with LCA so far, resulting in fewer studies to be analyzed.

As highlighted in the review by Romeiko et al. (2024), only 25 studies were identified addressing the LCA+ML interface during the Life Cycle Impact Assessment phase. These studies were divided between those focused on estimating potential environmental impacts and those aimed at calculating characterization factors. Consequently, the present work includes studies that analyzed ML techniques to predict potential environmental impacts, even without a complete LCA. In this way, we can expand the data sample and allow compilation of more relevant results and discussion.

Figure 2. 2 - PRISMA Flowchart: stages of the review process.



Source: adapted from the PRISMA methodology.

2.2.3 Categorization of results and synthesis

We conducted a content analysis of the selected articles, categorizing them from LCA and AI perspectives. Key information was captured related to temporal coverage, study sector, AI and LCA methods used, main impact categories, system boundaries, AI models, number of samples, target attributes, independent variables, data partitioning, and data feeding.

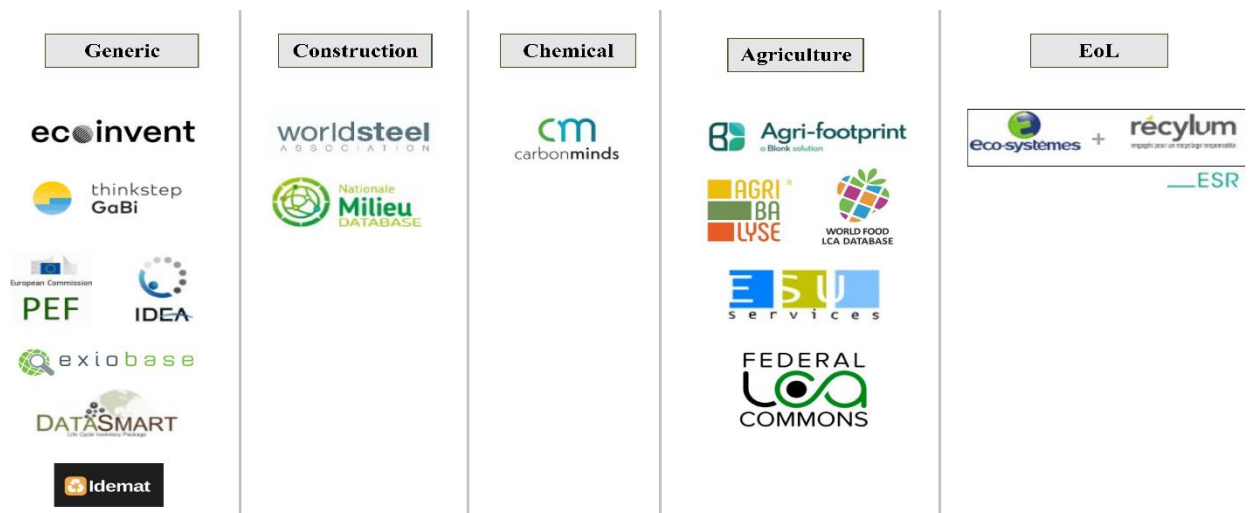
Data availability is crucial for LCA+ML studies, with LCA databases serving as reliable repositories for obtaining secondary data (Ecochain, 2024). There are several databases, with specific data for countries, sectors (agriculture, construction, chemicals, etc.), life cycle stages of interest (End of Life - EoL, use phase, materials supply, etc.), among others, as shown in Figure 2.3 (Ecochain, 2024).

Over 100,000 datasets are currently available across 16 LCA databases. The database with more datasets is the ecoinvent, followed by Sphera (Gabi databases) and agri-footprint.

The higher the number of datasets, the greater the possibilities for implementing ML in LCA, as already discussed by Romeiko et al. (2024), Koyamparambath et al. (2022) and Fang et al. (2021). However, some databases pose challenges due to the limited number of datasets.

The GLAD initiative, hosted by the United Nations Environment Programme, aggregates over 100,000 datasets from various providers (GLAD, 2024), including (until February 2024): Agribalyse (3,747), Amaru (7), Cobalt Institute (5),ecoinvent (69,472),ecoinvent SRI-LCI data (108),ecosystem (957),ESU-services (2,572),European Solvents Industry Group (ESIG) (6),SICV Brazil database by IBICT (21),IDEA (3,756),Japan Iron and Steel Federation (16),KEITI (185),LCA Society of Japan (JLCA) (107),NorEnviro (36),PeruLCA (55),Plastics Europe (50),RSE SpA (1),Sphera (14,082),US Federal LCA Commons (1,738),VitalMetrics Group (389),and worldsteel (81) (GLAD, 2024). These datasets vary in format such as ecoSpold (ecoinvent, PeruLCA), ILCD (SICV Brazil database), and JSON-LD (OpenLCA) depending on the database or LCA software used. This can impose constraints on LCA+ML integration because differences in format can affect interoperability and complicate the compilation of inventories for ML training.

Figure 2. 3 - Main LCA databases.



Source: adapted from (Ecochain, 2024).

GLAD facilitates dataset searches with filters for data providers, sectors, data categories, geographical coverage, inventory unit processes, and types of critical reviews (GLAD, 2024). Given GLAD's importance, we analyzed dataset availability by country to identify potential opportunities for LCA+ML integration, as detailed in Section 3.2.

The analyses were compiled into a synthesis table presented in Section 4 to identify

research gaps and streamline method selection for AI based on its prevalent use in the field and the objectives of LCA applications. This synthesis is intended to assist LCA practitioners in selecting AI-enabled tools and methods, address the research question, and support informed decision-making.

2.3 RESULTS

The results encompass two types of analysis: bibliometric analysis followed by a content analysis for the 59 articles selected for in-depth review.

2.3.1 Bibliometric analysis

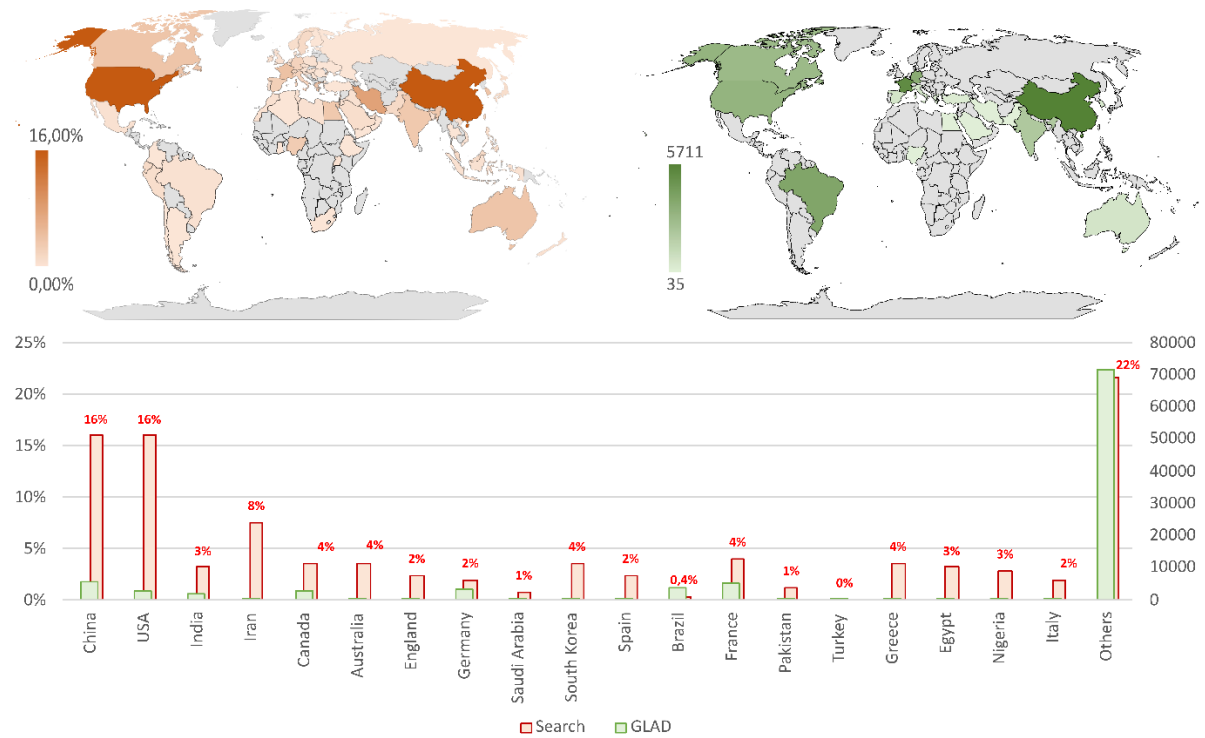
Our bibliometric analysis, detailed in the Supplementary Material, includes a temporal distribution (Figure S2), geographical distribution (Figure 2.4 and Figure S3), and an examination of major journals in the field (Figure S4). By analyzing the trends and patterns in these publications, we can better understand the evolving landscape of LCA+ML integration and identify key areas for future research and development.

2.3.1.1 Temporal distribution

The temporal analysis revealed a significant increase in research integrating ML to predict potential impacts over the last five years. Over 78% of the papers were published from 2020 to 2024, indicating an exponential trend. Conversely, there was less steep growth in works focused on integrating prediction and ML with a focus on LCA use, and this can be explained due to the widespread purpose of ML in LCA, since the creation of inventory data (Cheng et al., 2020), compliance analysis (Gu et al., 2023), and optimization correlation variables (López-Flores et al., 2023), for instance. Notably, 2023 saw a surge in publications, with a significant trend continuing into 2024, expected to exceed 60 publications.

2.3.1.2 Geographical distribution

Figure 2. 4 - Geographical distribution of publications.



Source: own authorship.

In search, publications were concentrated in a few sets of countries, with 19 nations accounting for over 78% of the papers, indicating a higher concentration of predictive LCA studies, particularly in developed and advanced countries. East Asian countries had the highest number of publications (20.4%), followed by North America (20%), Western Europe, the Middle East, Africa, and South America. The geographical scope regarding citations (Figure S3 in Supplementary Material) corroborated these findings, with China and the United States leading in influence.

2.3.1.3 Available datasets

Figure 2.4 also shows the quantity of available LCA datasets into the GLAD platform, identifying potentials for LCA+ML exploration in countries such as France, Brazil, and Germany (with 5,200, 3,894, and 3,537 datasets). However, these datasets numbers are relatively low compared to the sample sizes required by most ML models for application, like the case addressed in Software Engineering, where a small dataset (< 5,000 samples) and a medium-sized dataset (< 100,000 samples) are established to use modern ML methods (Prenner and Robbes, 2022).

Studies integrating LCA+ML have shown feasibility with smaller sample sizes

(typically less than 1,500), as Romeiko et al. (2024) observed in over 70% of the evaluated articles. This can be achieved by concatenating information from various data sources, such as EPD databases (Koyamparambath et al., 2022) and LCI datasets (Sun et al., 2023; Zhu et al., 2020), generating new samples through simulation tools like Monte Carlo Simulation (Ahmed et al., 2023) or optimizing ML model parameters to improve predictions with available samples.

2.3.1.4 Journal analysis

Our analysis of journal data sources revealed a predominance of five journals for search: Journal of Cleaner Production (26%), Energy Conversion and Management (14%), Science of the Total Environment (11%), Automation in Construction (4%), and Sustainability (4%), totaling over 59% of the citations.

The International Journal of LCA showed 16 citations (0.55%) for search. This can be explained by the fact that other journals have already issued calls for papers in this area, such as the call for "Transportation, Logistics and Emergent Technology within the Circular Economy" in the Journal of Cleaner Production (Journal of Cleaner Production, 2024) and for "Toward Sustainability: Reliable Transportation Network Design and Analysis" and "Applications of Artificial Intelligence in the Study of Land Use and Land Cover Change" in Sustainability (Sustainability, 2024). All of these calls also focus on applying artificial intelligence in sustainability.

2.3.1.5 Key themes

The journals highlighted for search cover themes of cleaner production, environment, and sustainability in an interdisciplinary manner (Journal of Cleaner Production, Science of the Total Environment, and Sustainability). Other focuses include ecology, management, and conservation of agricultural and forestry systems (Forest Ecology and Management), generation, use, quantification, and mitigation of environmental impacts related to energy resource use (Applied Energy, Energies, and Energy Conversion and Management), and even research focused on engineering, management, and recycling practices (Automation in Construction).

2.3.2 Content analysis (categorization and results)

In total, 59 articles were selected for inclusion in terms of the following classification criteria: study area, LCA system boundaries, number of independent variables covered, size samples, target attributes, as well as the machine learning techniques used, totaling the

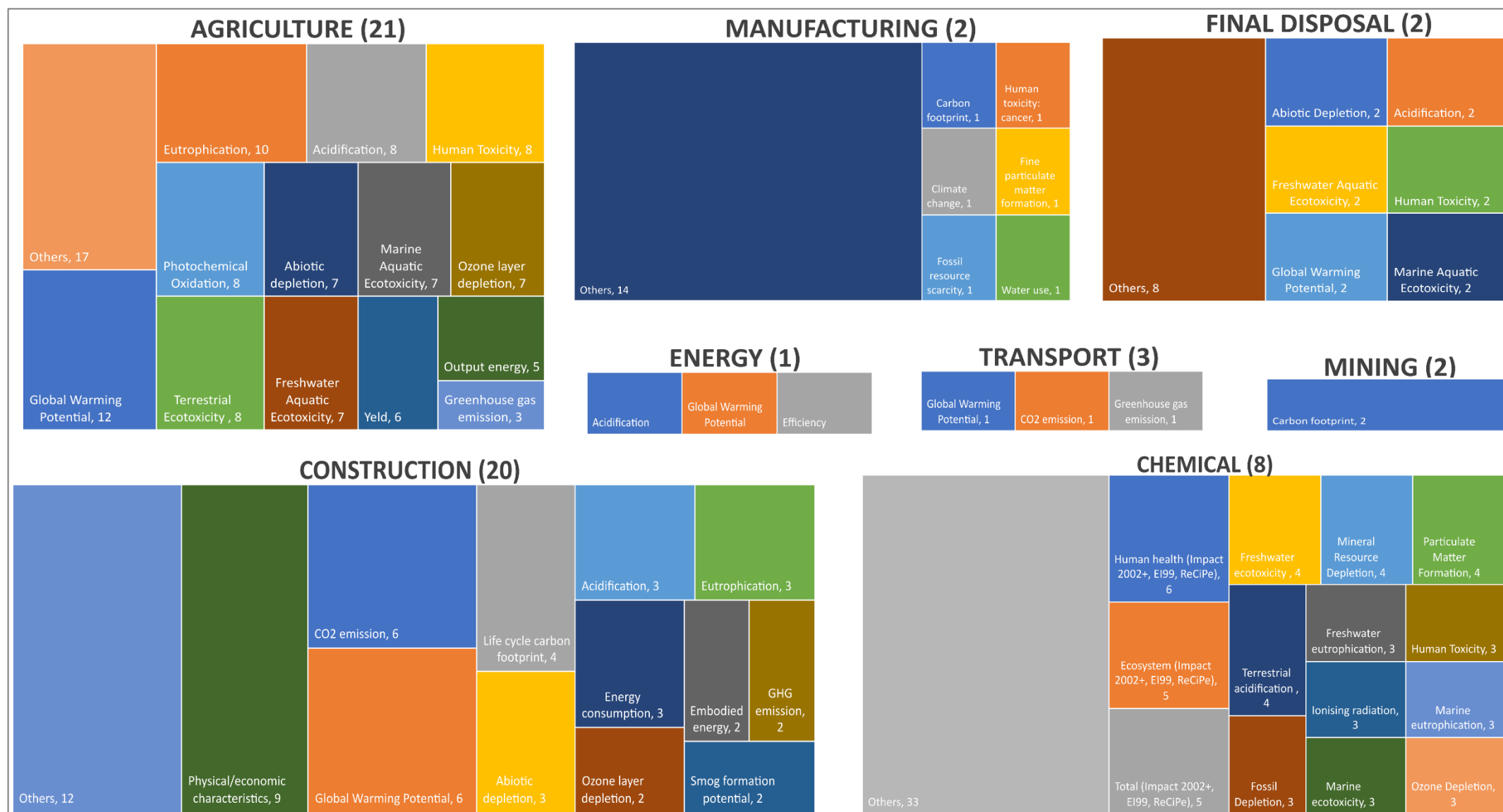
application of over 110 ML models as some studies applied more than five distinct ML methods.

2.3.2.1 Analysis of study areas

Our analysis identified studies integrating LCA+ML for predicting potential impacts in eight areas: agriculture (21), civil construction (20), chemicals (8), transportation (3), manufacturing (2), mining (2), disposal (2), and energy (1), as shown in Figure 2.5 and Table S1.

From the agricultural sector perspective, we identified studies aimed at predicting the potential environmental impacts of various crops, such as corn (Lee et al., 2020; Romeiko et al., 2020; Yan et al., 2018), rice (Khoshnevisan et al., 2014a; Nabavi-Pelesaraei et al., 2018), wheat (Ghasemi-Mobtaker et al., 2022), sugarcane (Kaab et al., 2019), strawberry (Khoshnevisan et al., 2013a), mushroom (Nabavi-Pelesaraei et al., 2023), kiwi (Nabavi-Pelesaraei et al., 2016a), watermelon (Nabavi-Pelesaraei et al., 2016b), potato (Khoshnevisan et al., 2013b), soybean (Mohammadi Kashka et al., 2023), oilseed (Mousavi-Avval et al., 2017), lentil (Elhami et al., 2017), apple (Taghavifar and Mardani, 2015), tomato (Pishgar-Komleh et al., 2020), and tomato and cucumber (Khoshnevisan et al., 2014b). Additionally, some studies focused on tea processing (Khanali et al., 2017), rice milling (Nabavi-Pelesaraei et al., 2019), and the analysis of agricultural sector emissions by country within the European Union context (Vlontzos and Pardalos, 2017). Among the most investigated impact categories are global warming potential (GWP), eutrophication (EU), acidification (AC), human toxicity (HT), and photochemical oxidation (PO). For LCA+ML integration, as indicated by Kaab et al. (2019), they use ANN (artificial neural network) and ANFIS (Adaptive Neuro-fuzzy Inference System) models.

Figure 2. 5 - Areas of application of the studies and the main predicted variables.



Source: own authorship.

Within the construction sector, fifteen studies focused on predicting potential environmental impacts of the lifespan of buildings and supporting the definition of parameters for sustainable construction in the design phase (Azari et al., 2016; D'Amico et al., 2019; Duprez et al., 2019; Fang et al., 2021; Feng et al., 2019; Hamida et al., 2021; Koyamparambath et al., 2022; Liu et al., 2023; Martínez-Rocamora et al., 2021; Pan and Wu, 2023; Ploszaj-Mazurek et al., 2020; Tsay et al., 2021; Xikai et al., 2019; Zhang et al., 2022; Zheng et al., 2024), while three studies focused on predicting potential impacts of concrete mixtures composition (Faridmehr et al., 2021; Onyelowe et al., 2022; Onyelowe et al., 2023), one study on the construction of a metro station in China (Zheng et al., 2022), and one study on the impact of choosing different types of slabs in constructions (Teodosio et al., 2023). Among the impact categories, these studies focused on predicting GWP and the lifetime emissions of constructions (LCCO₂). For example, Duprez et al. (2019) evaluated the use of ANNs trained by sensitivity analysis data to predict the GWP of new design alternatives in a short period, which resulted in an R² greater than 0.9, while Xikai et al. (2019) compared four regression techniques to estimate building LCCO₂, with SVR (support vector regressor) identified as the best method with an R² of 0.8.

Within the chemical sector, two studies focused on choosing more sustainable chemicals, one to define cleaner substitutes (Zhu et al., 2020) and one for selecting new solvents (Fleitmann et al., 2021). On the other hand, two studies focused on predicting the life cycle impacts of chemicals (Song et al., 2017; Sun et al., 2023), as well as studies applied to chemical processes (Kleinekorte et al., 2023), biomass processing (Karka et al., 2019), biorefineries routes (Karka et al., 2022), and estimates of hazardous concentrations of ecotoxicity in manufacturing processes (Hou et al., 2020). Unlike the agricultural and construction sectors, the chemical sector presented studies focused on predicting endpoint categories such as human health and ecosystem total impacts. Zhu et al. (2020) used 224 non-ionic organic chemical products from ecoinvent v3.5 to train a deep ANN and subsequently predict the following endpoints: EI99-ecosystem, EI99-human health, EI99-resources, EI99-Total, ReCiPe-ecosystem, ReCiPe-human health, ReCiPe-resources, and ReCiPe-total endpoint impacts.

Furthermore, studies involving the potential impacts of incineration and landfill EoL disposals (Nabavi-Pelesaraei et al., 2017a), recycling (Nabavi-Pelesaraei et al., 2017b), air pollution management in mines (Asif and Chen, 2020), potential impacts of metal mining (Asif et al., 2019), influences of parameters in product design (Singh and Sarkar, 2023), additive manufacturing (Naser et al., 2023), the automotive sector processes (Akhshik et al., 2022; Chen et al., 2021), the use of tugs in the maritime context (Ahmed et al., 2023), and the use of nuclear

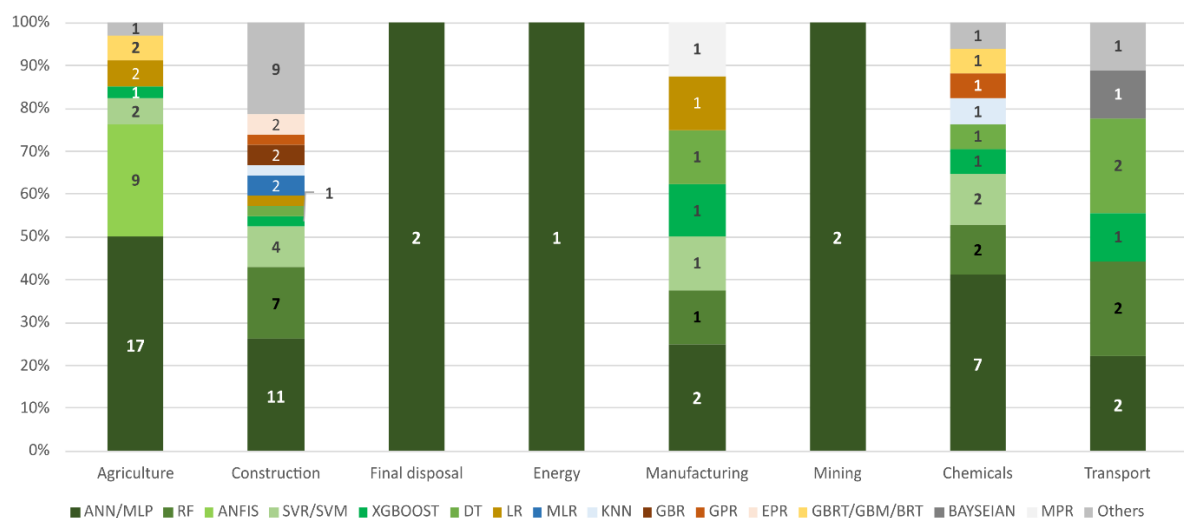
hydrogen in the energy sector (Ozbilen et al., 2013) were identified as other emerged examples.

The agricultural and construction sectors represented approximately 70% of the sample, with 41 out of 59 studies analyzed. This is likely due to the high environmental contribution these sectors exert on the planet and the complexity of measuring environmental impacts and suggesting new, more sustainable production technologies for such activities. This is in light of the GRO (2024) report, where over 50% of biodiversity impacts, 60% of water impacts, and 21% of climate impacts are related to food production, and about 18% of total climate impacts and global pollution are related to the built environment. Together, resource-intensive supply systems in the built environment (construction), mobility (transportation), food, and energy represent approximately 90% of global material demand, 70% of climate impacts, and over 80% of biodiversity loss and water stress (GRO, 2024). Despite such significance, the results point to limitations regarding the use of ML in these sectors due to the restricted number of publications in the area and even the availability of data to train ML models and ensure that predictions made have assured accuracy, which ultimately leads to potential uncertain results and hinders LCA+ML integration.

2.3.2.2 Analysis of ML methods

Figure 2.6 illustrates the different ML methods applied across key economic sectors to predict potential environmental impacts. Our analysis identified a range of ML techniques, from simple linear regression models to more complex artificial neural networks (ANNs) and advanced ensemble methods.

Figure 2. 6 - Arrangement of ML methods by sector.



Source: own authorship.

In sectors such as waste disposal, energy, and mining, the exclusive use of ANNs and Multilayer Perceptron (MLP) was observed for environmental impact predictions. Similarly, ANNs were predominant in other sectors, with 17 studies in agriculture, 11 in construction, 2 in manufacturing, and 7 in the chemical industry.

The Adaptive Neuro-Fuzzy Inference System (ANFIS) was exclusively used in about 50% of agricultural sector studies. The ANFIS model leverages fuzzy logic using IF-ELSE rules to manage imprecise and uncertain data, making it ideal for the agricultural context where variables like weather conditions and the dissipation of chemicals are inherently uncertain (Figueroa-García et al., 2015; Zadeh, 2015). Thus, using such logic, the ANFIS model can handle imprecise and uncertain data to establish relationships between variables approximately. This justifies its predominance in the agricultural context, as the sector presents uncertain/imprecise data in its inventories, such as the lack of control over weather conditions, the dissipation of fertilizers, pesticides, and even biocides in the soil, among other uncertain variables present in agricultural studies.

The Random Forest (RF) method was employed across multiple sectors, including seven studies in construction (Fang et al., 2021; Koyamparambath et al., 2022; Liu et al., 2023; Martínez-Rocamora et al., 2021; Pan and Wu, 2023; Xikai et al., 2019; Zheng et al., 2024), one in manufacturing (Naser et al., 2023), two in the chemical sector (Hou et al., 2020; Sun et al., 2023), and in two in transport (Ahmed et al., 2023; Akhshik et al., 2022). Decision Trees (DT), which follow a similar logic to RF, were found in up to two studies within these sectors.

Gradient Boosting Regression Trees (GBRT) and XGBoost were notable for their use in sequential decision structures, allowing them to capture complex, non-linear relationships between variables. GBRT was used in two agricultural studies (Lee et al., 2020; Romeiko et al., 2020) and one chemical study (Hou et al., 2020). XGBoost was applied across agriculture (Zheng et al., 2024), construction (Pan and Wu, 2023), manufacturing (Naser et al., 2023), chemical, and transport sectors (Ahmed et al., 2023). These methods enhance model robustness by reducing overfitting and improving generalization (James Montantes, 2020). Ahmed et al. (2023) observed such behavior when using RFR (random forest regression) to predict GWP in the context of supply vessels and offshore anchor handling vessels (AHTS), reducing the overfitting of the model and achieving an R^2 of 0.96 in prediction by increasing the sample size.

Linear Regression (LR) and Multiple Linear Regression (MLR) were less frequently used due to their inability to model non-linear relationships effectively. LR was used in agriculture (Ghasemi-Mobtaker et al., 2022; Romeiko et al., 2020), construction (Onyelowe et

al., 2023), and manufacturing studies (Naser et al., 2023). MLR appeared in construction studies (Duprez et al., 2019; Zheng et al., 2024). K-nearest neighbors (KNN) were used in construction (Pan and Wu, 2023) and chemical sector studies (Hou et al., 2020). These models were less effective in complex natural problems due to their simplistic approach. Onyelowe et al. (2023) used LR in part of their analyses, enabling the prediction of "28 days compressive strength (Fc28, MPa) and the environmental impact factor (P/GHG emission)" by relating the amount of ash in the cement ash composition, which returned an R2 of 0.48 and 0.91, respectively, through the use of linear equations. Conversely, Pan and Wu (2023) analyzed predictions of carbon emissions from buildings using 15 design factors, hyperparameter optimization with the Bayesian Optimization (BO) algorithm, and comparing four ML methods. The results showed that the BO-XGBoost model fits well with the calculated values. In contrast, the BO-KNN showed fluctuations around the calculated values, having weak prediction performance when analyzing complex natural problems.

Bayesian models were applied in the transport sector (Akhshik et al., 2022). Support Vector Regression (SVR), adapted from Support Vector Machines (SVM), was used across agriculture (Nabavi-Pelesaraei et al., 2023; Romeiko et al., 2020), construction (Duprez et al., 2019; Pan and Wu, 2023; Xikai et al., 2019; Zheng et al., 2024), manufacturing (Naser et al., 2023), and chemicals (Hou et al., 2020; Sun et al., 2023). Bayesian models, though simpler, showed higher error rates in handling complex data compared to SVR, which excels in optimizing weight vectors and bias for non-linear data relationships (Zheng et al., 2024).

Other methods like Multivariate Polynomial Regression (MPR), Gradient Boosting Regressor (GBR), Gaussian Process Regression (GPR), and Evolutionary Polynomial Regression (EPR) were used less frequently but are emerging in various sectors. While still in their early stages, these methods show promise in environmental LCA studies due to their ability to model complex relationships and optimize predictions.

The analysis highlights the diverse application of ML methods across sectors, emphasizing the suitability of specific techniques based on the nature of the data and the complexity of the relationships involved. Although useful, ANNs and ensemble methods like RF and XGBoost demonstrate robustness and generalizability, while simpler models like LR and Bayesian methods are less effective in complex scenarios. Emerging methods such as GPR and EPR offer new avenues for enhancing predictive accuracy in environmental impact assessments.

2.3.2.3 Analysis of LCA boundaries

Analyzing system boundaries is crucial in LCA as it defines the scope, including all relevant processes and flows to the technosphere and ecosphere, while excluding items unrelated to the study objectives or by cut-off rules. This section evaluates the system boundaries used in the analyzed studies based on sector and the most employed ML models (Figure 2.7).

Of the studies analyzed, 20 employed a cradle-to-gate approach, 9 adopted cradle-to-grave, 4 utilized gate-to-gate, and only 2 used gate-to-grave boundaries. Notably, 24 studies did not specify their LCA system boundaries.

Most studies addressing the agricultural sector employed a cradle-to-gate approach, modeling product systems from raw material supply up to the farm gate level in 13 out of 15 studies. In contrast, the construction sector showed a higher incidence of cradle-to-grave studies, reflecting a focus on the entire lifecycle, including use, maintenance, and demolition or recovery phases (Azari et al., 2016; Pan and Wu, 2023; Tsay et al., 2021; Zheng et al., 2024). For end-of-life (EoL) strategies, gate-to-grave studies analyzed the impacts of products post-use and urban waste management (Nabavi-Pelesaraei et al., 2017a; Nabavi-Pelesaraei et al., 2017b).

When correlating LCA system boundaries with ML methods, it is evident that ANNs are versatile and are used across all types of scopes. Other ML methods, such as LR, DT, GPR, and SVM/SVR, were also employed in both cradle-to-grave and cradle-to-gate studies.

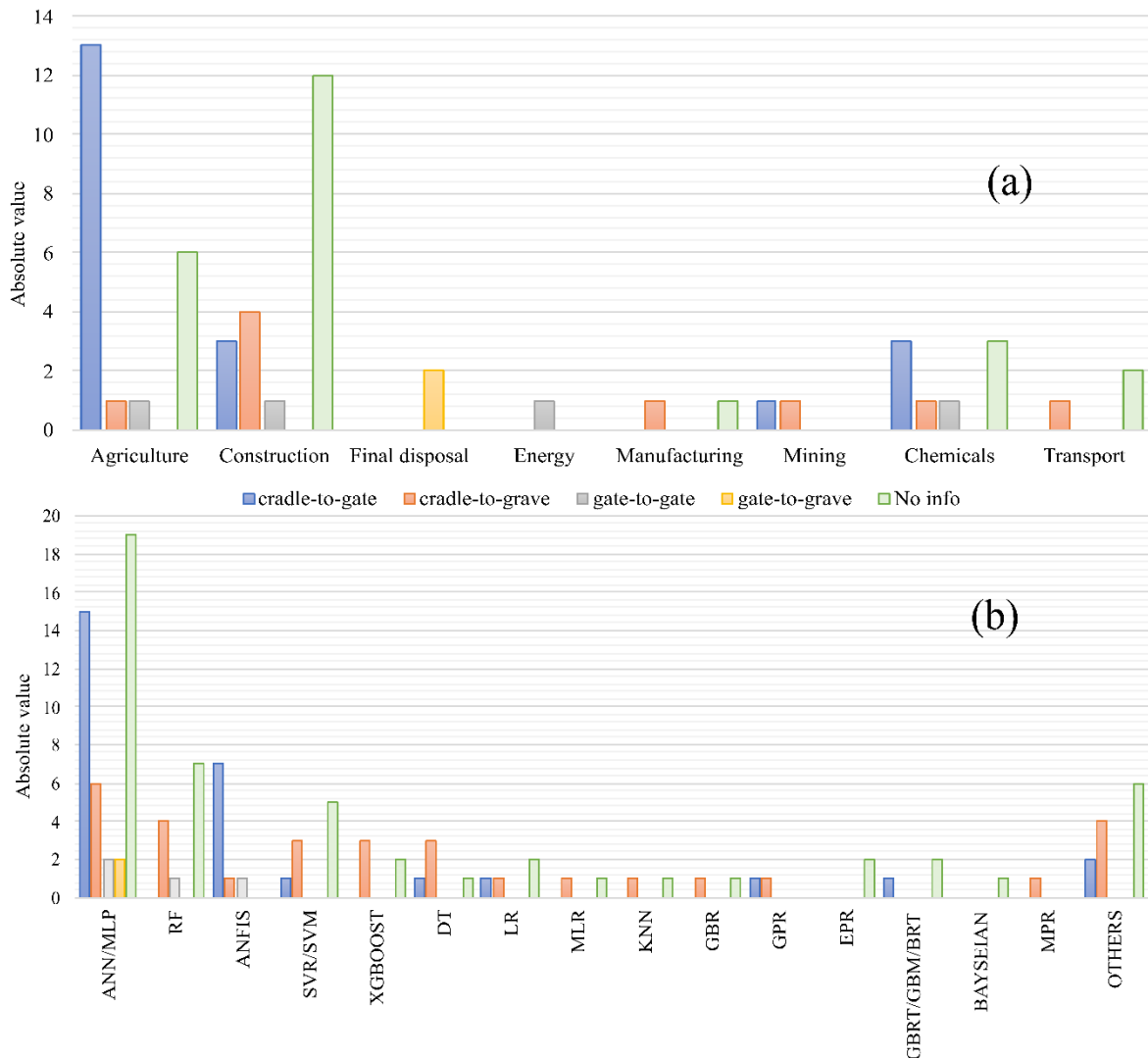
Cradle-to-gate LCA studies typically focus on intermediate products rather than final outputs, thus excluding use and post-use phases, as well as customizations in consumer goods production. Broader scopes, such as cradle-to-grave, demand more customization and additional data related to final goods manufacturing, use stage modeling, and end-of-life options. This adds complexity and uncertainty to both LCA modeling and ML training.

Studies with broader scopes face challenges due to the need for detailed data on product use and end-of-life stages, which are variable and less standardized than intermediate product data. Consequently, cradle-to-gate approaches are more common in sectors with well-documented and commoditized processes, such as agriculture (e.g., soybeans, corn, and wheat), energy (e.g., energy generation), mining (e.g., iron ore), and chemicals (e.g., caustic soda, sulfuric acid) (Research XP, 2024).

In conclusion, the choice of LCA system boundaries is closely tied to the sector and the availability of detailed lifecycle data. While cradle-to-gate studies are prevalent in agriculture and commoditized sectors, cradle-to-grave approaches are essential in construction due to their

comprehensive lifecycle focus. ML methods like ANNs, LR, DT, GPR, and SVM/SVR are adaptable to various scopes, but the complexity and data requirements increase significantly with broader system boundaries. Understanding these nuances is critical for effectively integrating ML into LCA studies to predict potential environmental impacts.

Figure 2. 7 - Arrangement of LCA system boundaries by sector (a) and by ML method (b).

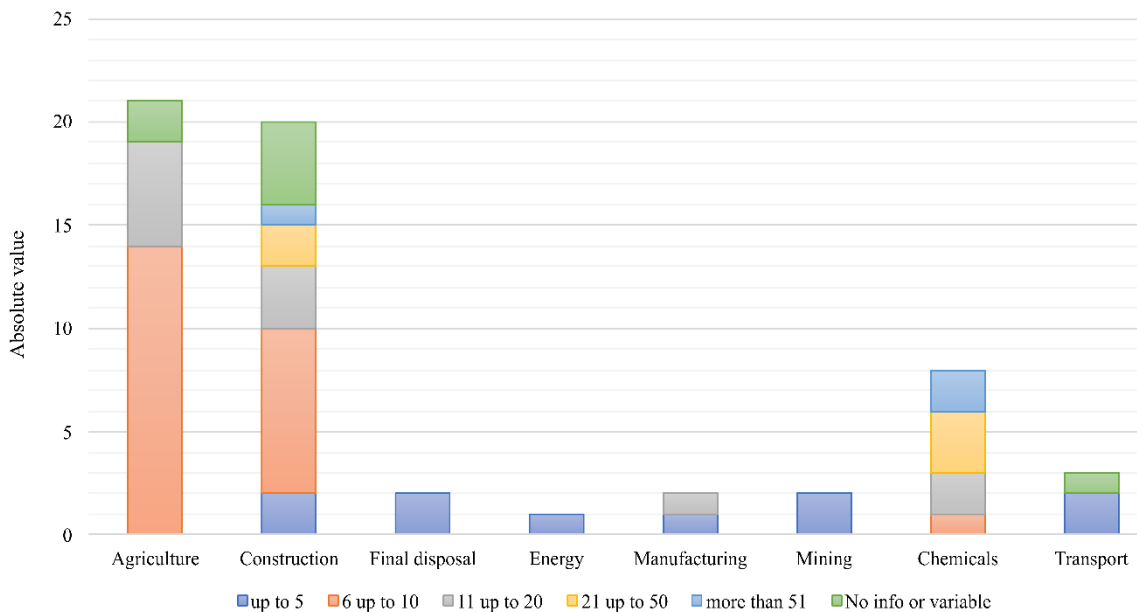


Source: own authorship.

2.3.2.4 Analysis of independent variables

Analyzing the number of independent variables allows us to identify the attributes commonly used to train ML models for predicting potential impacts in LCA studies (Figure 2.8).

Figure 2. 8 - Analysis of the number of independent variables by sector.



Source: own authorship.

Final disposal, energy, and mining typically used a maximum of five independent variables to carry out the desired predictions. Attributes such as human labor, water, electricity, natural gas, and transportation are often used. For example:

- **Recycling**: Attributes like human labor, water, electricity, natural gas, and transportation are often used (Nabavi-Pelesaraei et al., 2017b).
- **Nuclear Hydrogen Production**: Key variables include production plant capacity, plant lifetime, and heat and electrical work input (Ozbilen et al., 2013).
- **Mining Pollution**: Fuel consumption, CO₂, N₂O, CH₄ emission rates, and equipment operating hours are frequently used (Asif et al., 2019).

Studies in the agricultural sector often employ 6 to 10 attributes for model training due to the consolidated nature of commodity inputs. Common variables include machinery, human labor, diesel consumption, pesticide application, fertilizer and manure consumption, natural gas and electricity consumption, seed usage, and water usage. Examples include:

- **Strawberry Cultivation**: Farmyard manure, chemical fertilizers, pesticides, natural gas, irrigation water, electricity, diesel fuel, and machinery (Khoshnevisan et al., 2013a).
- **Tomato/Cucumber Cultivation**: Similar variables as above (Khoshnevisan et al., 2014b).

- **Rice Production**: N, P, and K-based fertilizers, agricultural machinery, pesticides, electricity, farmyard manure, diesel fuel, and farm size (Khoshnevisan et al., 2014a).
- **Tea Processing**: Energy equivalents of fresh tea leaves, human labor, diesel fuel, electricity, adhesive, cartons, corrugated paper boxes, and transportation (Khanali et al., 2017).
- **Oilseed Production**: Labor, machinery, diesel fuel, chemicals, nitrogen fertilizer, other chemical fertilizers, farmyard manure, and seed (Mousavi-Avval et al., 2017).
- **Apple Cultivation**: Human labor, machinery, diesel fuel, fertilizer, herbicide, manure, irrigation water, and electricity (Taghavifar and Mardani, 2015).

Similarly, the construction sector uses a wide range of 5 to 51 independent variables, with some studies employing as few as three variables (e.g., fly ash to binder ratio, fine aggregates to binder ratio, and coarse aggregates to binder ratio for concrete composition) (Onyelowe et al., 2023). Other studies use over 1,000 attributes extracted through natural language processing from Environmental Product Declarations (EPDs) to predict various impacts such as Global Warming Potential (GWP), abiotic depletion potential for fossil resources, acidification potential, and photochemical ozone creation potential (Koyamparambath et al., 2022). For example:

- **Waffle Slabs**: Attributes like surface characteristic movement, beam depth, area of residential slab, diameter of bottom and top steel reinforcements, area and concentrated loads applied to the footing, and active depth zone are used to predict embodied energy, carbon emissions, life cycle cost, and deflection (Teodosio et al., 2023).
- **Future Carbon Emissions**: Population size, urbanization rate, per capita GDP, added-value of tertiary industry, steel production, average haul distance of highway transportation, and labor productivity of construction enterprises are used to predict carbon emissions trends in Jiangsu Province, China (Zhang et al., 2022).

This chemical sector demonstrates a high demand for independent variables, with most studies using more than 21 attributes. For instance, Song et al. (2017) started with 3,839

descriptors. They reduced them to 60 attributes using Principal Component Analysis (PCA) to train a model predicting six impact categories, including cumulative energy demand, GWP, acidification potential, and three endpoint impact categories: eco-indicator 99, ecosystem quality, and human health.

In conclusion, various ML methods such as ANN, RF, and SVM exhibit great versatility across different sectors and numbers of independent variables (Figure S5). These methods are adept at handling a wide range of input variables, from a few essential attributes to thousands of complex descriptors, showcasing their flexibility in modeling diverse environmental impacts in LCA studies.

The number and type of independent variables used in LCA studies vary significantly across sectors. While final disposal, energy, and mining sectors rely on a limited set of variables, agriculture and construction often require a broader range. The chemical sector tends to use the most extensive set of variables due to the specificity of its studies. ML methods like ANN, RF, and SVM prove to be versatile tools capable of handling this diversity, making them valuable for predicting potential environmental impacts across different LCA contexts.

2.3.2.5 Analysis of the number of samples

LCA studies typically require extensive time for inventory analysis and impact assessment interpretation due to the substantial data compilation and modeling efforts needed. This time-intensive process often limits the production of new datasets. In contrast, ML methods demand relatively large numbers of samples to adjust the models and establish relationships between input variables (independent attributes) and output variables (target attributes).

Models such as ANNs, deep learning networks, and sequential decision algorithms (e.g., tree-based models) require substantial data volumes to capture complex relationships between variables. These models mimic human brain processing by traversing different nodes during this process. Contrariwise, models such as SVMs (with linear kernel), KNN, and LR can perform with smaller data volumes due to their reliance on linear relationships between inputs and outputs. Understanding the integration of LCA and ML to predict potential impacts is thus crucial (Figure 2.9).

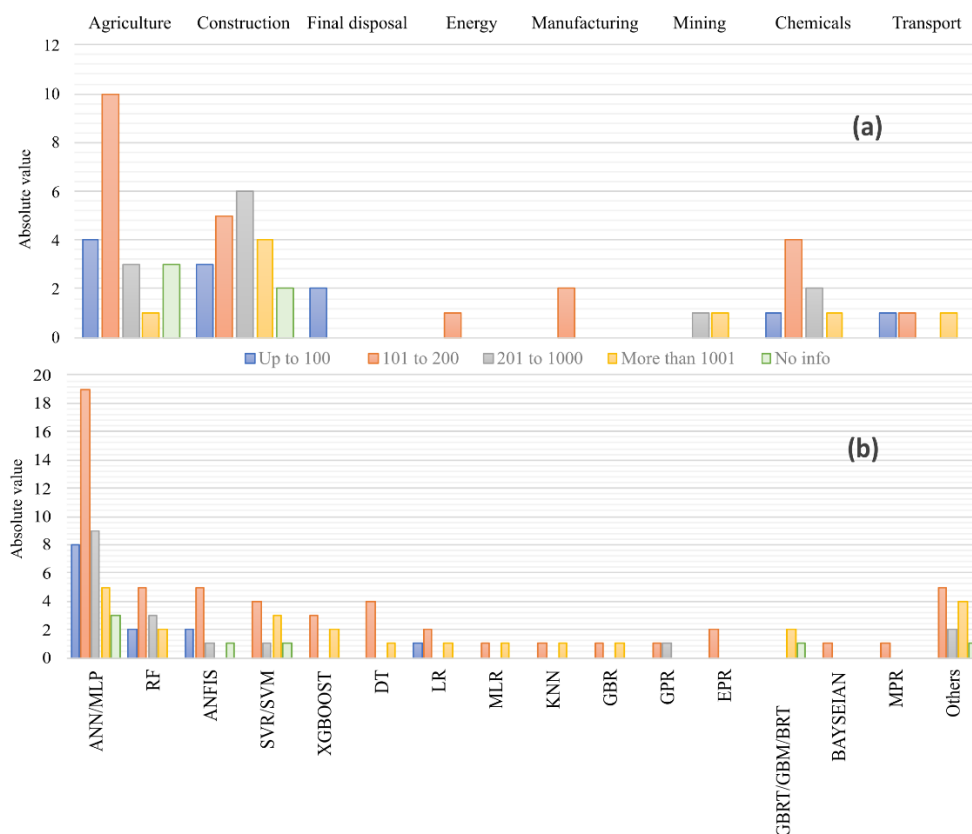
Most studies in the Agricultural, Final Disposal, Energy, Manufacturing, Chemical, and Transport sectors used ML models with up to 200 samples, indicating the challenges of employing LCA data for ML training (Figure S6). Specifically:

- Agricultural Sector: 14 studies

- Final Disposal Sector: 2 studies
- Energy Sector: 1 study
- Manufacturing Sector: 2 studies
- Chemical Sector: 5 studies
- Transport Sector: 2 studies

These findings highlight the difficulties in data collection and the need for more comprehensive datasets in these sectors. In contrast, the construction industry exhibits a more mature application of life cycle thinking, supported by more robust databases. This is evidenced by several studies using over 1,000 samples, suggesting higher data availability and integration for ML model training in this sector.

Figure 2. 9 - Analysis of the number of samples (a) per sector and (b) per machine learning method.



Source: own authorship.

ANN and RF methods demonstrate significant flexibility, being used across all sample intervals with varying data volumes. This flexibility underscores their robustness and

adaptability in different contexts. ANFIS, SVM, and LR also show considerable flexibility, with studies spanning three of the four analyzed sample intervals. Their adaptability makes them valuable tools for a range of LCA+ML integration scenarios. XGBoost, DT, MLR, KNN, and GBR are predominantly used in studies with 101 to 200 samples and those with more than 1,001 samples. This indicates their capability to handle moderate and extensive datasets, highlighting their versatility.

Overall, our results align with the findings of Romeiko et al. (2024), which revealed that over 70% of analyzed studies used fewer than 1,500 datasets for training ML models. However, it is important to note that smaller datasets often result in higher uncertainties. This underscores the need for more extensive and comprehensive datasets to enhance the reliability of LCA+ML models in predicting environmental impacts.

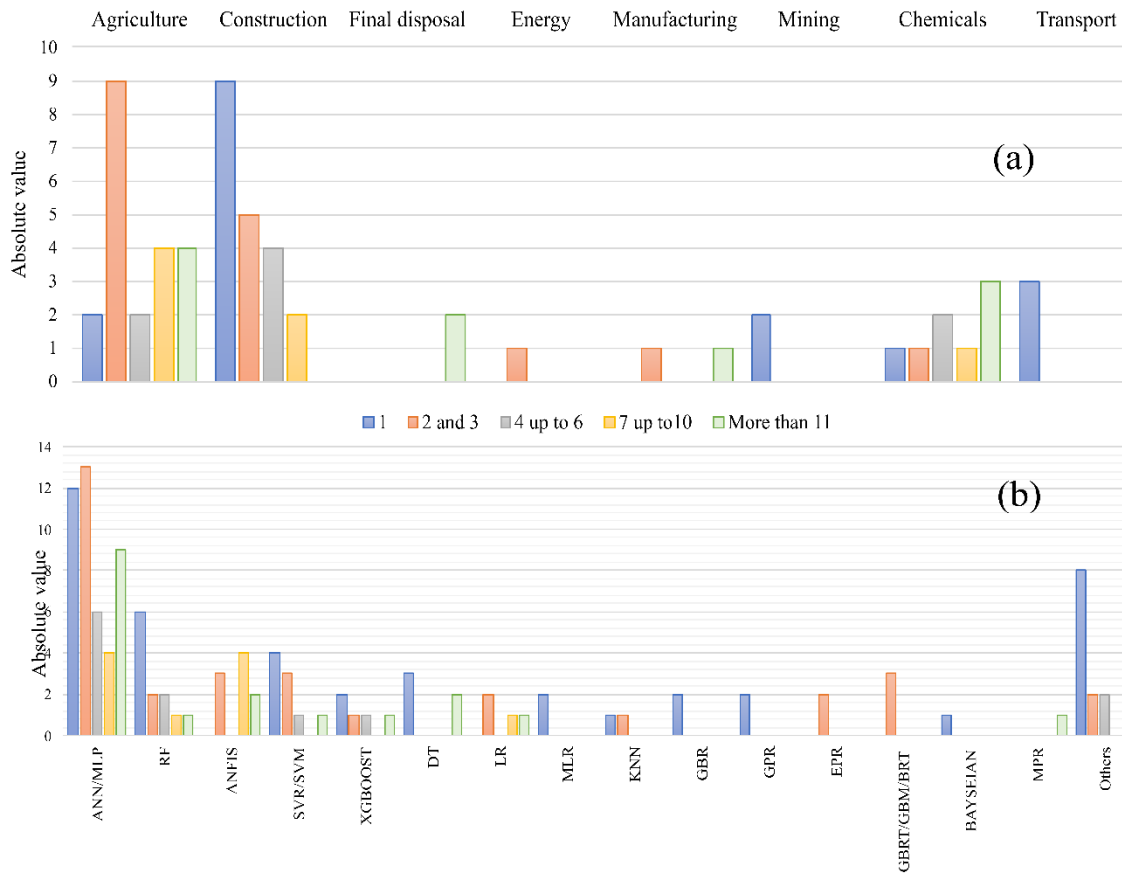
In conclusion, the analysis of sample sizes in LCA+ML studies reveals significant sectoral differences and highlights the flexibility of various ML models. While sectors such as agriculture and chemicals face challenges due to limited data, the construction sector benefits from more extensive datasets. Flexible ML models like ANN, RF, ANFIS, SVM, and LR can handle varying data volumes, making them suitable for diverse applications. The need for larger datasets to reduce uncertainties remains critical for improving the accuracy and reliability of environmental impact predictions in LCA+ML studies.

2.3.2.6 Analysis of the number of target attributes

LCA studies have diverse focuses, which are determined by the initially defined goal and scope. Consequently, the impact assessments can support calculations in various categories, such as climate change, ozone layer depletion (OLD), photochemical ozone formation, respiratory inorganics, ionizing radiation, acidification (AC), eutrophication, human toxicity (HT), ecotoxicity, land use, and resource depletion (ILCD, 2011). This analysis examined the number of impact categories as attributes for ML models (Figure 2.10).

Our results indicate notable versatility in studies within the agricultural and chemical sectors, enabling predictions for a range of potential impacts. For example, studies have assessed carbon emissions from corn production (Yan et al., 2018), greenhouse gas emissions from the agricultural sector in European Union countries (Vlontzos and Pardalos, 2017), and global warming impact of organic molecules (Kleinekorte et al., 2023), with some studies encompassing more than 11 distinct potential impacts.

Figure 2. 10 - Analysis of the number of target attributes by sector (a) and by ML method (b).



Source: own authorship.

In the agricultural sector, studies over 11 target attributes predicted impacts such as GWP, acidification (AD), eutrophication (EU), OLD, HT, freshwater aquatic ecotoxicity (FAE), marine aquatic ecotoxicity (MAE), terrestrial ecotoxicity (TE), photochemical ozone formation, and yield. Examples include tea processing (Khanali et al., 2017), lentil cultivation (Elhami et al., 2017), rice production (Nabavi-Pelesaraei et al., 2018), and sugarcane production, which considered ten impact categories along with output energy (Kaab et al., 2019). In the chemical sector, predictions were made for 17 midpoint impact categories using the ReCiPe method for solvent selection (Fleitmann et al., 2021). Additionally, 23 LCA metrics were employed to facilitate the environmental analysis of chemical choices during the design phase for biorefineries (Karka et al., 2022) and biomass production (Karka et al., 2019).

In contrast, studies in the construction sector predominantly predict a single potential impact. Examples include GWP from design parameter choices in the design phase (Duprez et al., 2019; Feng et al., 2019), life-cycle carbon emissions (LCCO₂) from construction projects (Pan and Wu, 2023; Xikai et al., 2019), and whole-life carbon emissions (WLCE) (Zheng et al.,

2024), carbon emissions from subway station construction (Zheng et al., 2022), building carbon emissions from buildings (Zhang et al., 2022), and even the total carbon footprint of a building project based on essential building characteristics and urban layout (Ploszaj-Mazurek et al., 2020). The emphasis on carbon in construction studies likely stems from international decarbonization policies and goals specific to the construction sector, pursued by most countries through innovations and more sustainable technologies. This focus is corroborated by a 2023 report from the World Green Building Council, which aims to support governments in transforming existing legislation to ensure that national policies provide development integrating measures that simultaneously address social issues, energy security, health, and well-being (WorldGBC, 2023). In agriculture, due to the involvement of numerous variables necessitates a focus on multiple impact categories, mainly due to the significant effects of crop production on different environmental compartments, such as soil dissipation, pesticide and biocide leaching, and methane emissions from residual biomass.

From the perspective of ML models, we observe that MLR, GBR, GPR, and Bayesian methods are predominantly used in studies predicting a single potential impact. Examples include predicting WLCE in construction studies using MLR and GBR (Zheng et al., 2024), global warming impact in the chemical sector using GPR (Kleinekorte et al., 2023), and greenhouse gas emissions in the automotive sector (Akhshik et al., 2022).

In contrast, ANN and RF methods are utilized in studies predicting both single and multiple attributes. Examples include predicting construction carbon emissions in the design phase using RF (Fang et al., 2021), analyzing the carbon footprint in the mining sector with ANN (Asif et al., 2019), predicting 23 attributes of interest using ANN for environmental analysis and chemical selection in biorefinery projects (Karka et al., 2022), and evaluating 18 impact categories using RF for product-process co-design in manufacturing (Naser et al., 2023).

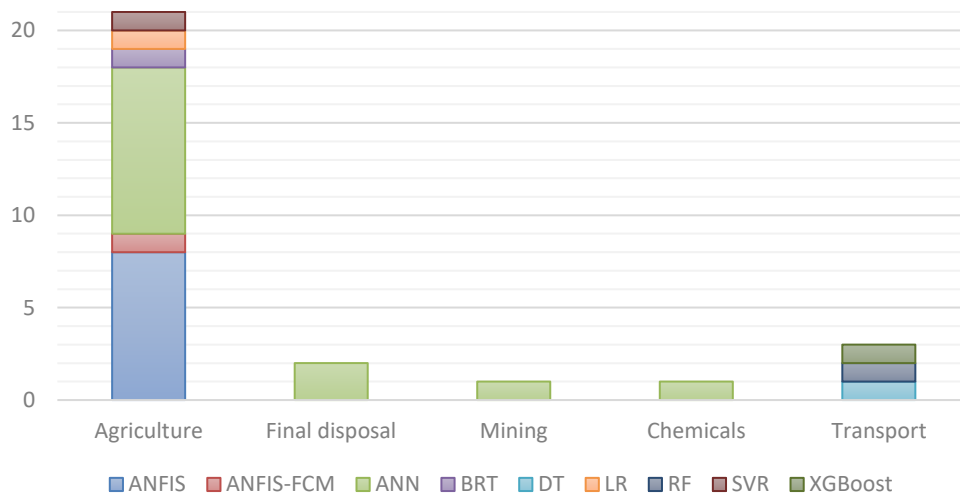
2.3.3 Content analysis (categorization and results) – full LCA studies

To identify the studies that conducted a full LCA, we employed a Supplementary Datasheet. Out of the 59 articles initially selected for content analysis, 18 studies were identified as meeting the criteria for full LCA. In the following, we present a focused analysis of these studies, representing a subset of the broader content analysis conducted in Section 3.2. The results here offer insights specifically into studies that performed a complete LCA.

The analysis applied the same methodological framework as in Section 3.2, covering various dimensions: ML methods used by sector (Fig. 2.11), LCA boundaries (Fig. S7), independent variables (Fig. S8), number of samples used (Fig. S9), number of target attributes

(Fig. S10), and LCIA methods (Fig. S11).

Figure 2. 11 - Arrangement of ML methods by sector for full LCA (own authorship) - in the number of occurrences.



Source: own authorship.

A consistent pattern emerged in the application of ML models across sectors. The agricultural sector remained the most representative, accounting for 13 studies, followed by the final disposal sector with 2 studies, while the other sectors presented only one study each. No full LCA studies were found for the construction, energy, or manufacturing sectors due to the absence of information on coverage, LCIA method, and LCA boundaries, as indicated in the provided Supplementary Datasheet.

Artificial Neural Networks (ANN) were the most prevalent, especially in the agricultural sector, and were exclusively used in mining, chemicals, and end-of-life (EoL) strategy sectors. Notably, Adaptive Neuro-Fuzzy Inference Systems (ANFIS) were exclusively employed in agricultural studies that conducted full LCAs. This highlights ANFIS's capability to handle the uncertainties inherent in agriculture, such as variables related to rainfall, leaching, nutrient absorption, and pest emergence.

The LCA boundaries varied by sector. A cradle-to-gate scope was dominant in agriculture and chemicals, offering a more focused analysis of upstream processes. In contrast, cradle-to-grave boundaries were exclusively applied in transportation and mining sectors, ensuring comprehensive life cycle impact assessments. Gate-to-grave boundaries were prevalent in EoL strategy studies, reflecting their focus on post-consumer stages.

From a model perspective, cradle-to-gate boundaries allowed for greater flexibility in

model selection, with ANFIS, ANFIS-FCM, ANN, BRT, LR, and SVR models applied. This suggests that more constrained boundaries facilitate the use of diverse ML models, including linear, neural network-based, and margin-maximizing models.

The analysis of independent variables revealed notable trends. Studies in final disposal and mining, both employing ANN models, typically used fewer than five attributes for training. Conversely, agricultural studies frequently used 6–10 independent variables, reflecting the need for richer datasets in this sector. The chemicals sector stood out, with one study employing over 21 attributes to train an ANN model, underlining the complexity of this domain. For instance, Karka et al. (2019) employed 30 predictor variables to assess potential impacts of biomass processing.

Most studies used ML models trained on fewer than 200 samples, reaffirming the data limitations highlighted in Section 3.2.5. However, transportation and mining sectors required larger datasets, with over 1,001 samples used for models like DT, ANN, RF, and XGBoost. This reflects the greater data requirements and uncertainties associated with cradle-to-grave LCAs.

Given the full LCA scope of these studies, a detailed analysis of LCIA methods by sector and ML model was possible. The most common method in agricultural and EoL studies was CML, used to assess impacts across multiple environmental compartments. Examples include lentil cultivation (Elhami et al., 2017), tea processing (Khanali et al., 2017), and incineration versus landfill impacts (Nabavi-Pelesaraei et al., 2017a, 2017b).

In the chemicals sector, two LCIA methods were prominent: Eco-indicator 99 (EI99) and ReCiPe (both endpoint and midpoint approaches). These methods supported the use of neural networks to predict multiple impact categories, such as EI99-ecosystem, EI99-human health, EI99-resources, and their ReCiPe equivalents. This dual-method approach facilitated the evaluation of greener substitute chemicals (Zhu et al., 2020). In short, it was observed that studies conducting full LCAs are less frequent in the context of LCA+ML integration because full LCAs are time-consuming and must strictly adhere to the defined steps for their execution (goal and scope definition; life cycle inventory; life cycle impact assessment; and interpretation). Moreover, the results suggest a growing trend in LCA+ML for the agricultural sector, driven by the multiple environmental impacts associated with food production and consumption, along with the ongoing need to increase food production by approximately 50% between 2019 and 2050 to meet the demands of an estimated population of 10 billion people by 2050 (World Resources Institute, 2019).

2.4. DISCUSSION

We critically analyzed the integration of ML methods with LCA to predict potential environmental impacts. We focused on the 11 most frequently used ML models, ranging from simpler models like Bayesian networks to more complex ones such as ANN, XGBoost, and SVM. Our goal was to provide a comprehensive understanding of the current state of LCA+ML studies, identify research gaps, and offer consolidated information to guide future research.

Our analysis, detailed in Table 2, ranks the ML models based on their frequency and effectiveness in LCA+ML integration studies as follows: ANN > RF > SVM > LR > XGBoost > DT > ANFIS > GBM > KNN > MLR > Bayesian. ANN emerged as the most commonly used model, indicating its robustness and reliability in environmental impact prediction. In contrast, though less frequently used, Bayesian networks show potential for further exploration due to their specific characteristics.

ANN and RF models demonstrate exceptional versatility across multiple sectors, including agriculture, construction, energy, and chemicals. Their widespread applicability indicates their robust algorithmic structure, which can handle diverse data types and complex relationships. This versatility makes ANN and RF particularly valuable for comprehensive LCA studies that span various industries and environmental contexts.

Conversely, while highly effective in specific scenarios such as transportation, Bayesian models exhibit limited applicability across different sectors. This suggests that Bayesian methods may be more suitable for niche applications where the data characteristics align well with the model's assumptions and strengths.

Our findings align with the work of Barros and Ruschel (2021), who identified ANN, SVM, Bayesian networks, and genetic algorithms as the primary ML techniques in recent LCA studies within the construction sector (Koyampambath et al., 2022). Additionally, our results partially corroborate the review by Romeiko et al. (2024), which highlighted ANN as the most frequently applied ML method, followed by ANFIS, LR, and RF.

ANN, RF, and XGBoost models are adequate across different LCA system boundaries, including cradle-to-gate, cradle-to-grave, and gate-to-grave. This flexibility underscores their capability to manage varying scopes of environmental assessments, from initial resource extraction to end-of-life disposal.

In terms of sample size, these models also excel in handling both small (up to 100 samples) and large datasets (more than 1,001 samples). Their scalability is a significant advantage, allowing them to be applied to extensive datasets typical in large-scale

environmental studies without compromising performance.

The ability of ANN and RF models to manage a wide range of features and target attributes further cements their utility in LCA. These models perform well with datasets that include numerous features (up to and beyond 21 attributes) and can predict multiple target attributes simultaneously. This makes them particularly suitable for complex LCA studies that require comprehensive environmental impact predictions across multiple dimensions.

The integration of ML with LCA heavily relies on robust databases and dataset repositories, which provide essential data for training ML models. However, variations in dataset formats (e.g., ILCD, JSON-LD, EcoSpold) can impede data interoperability, posing challenges for LCA+ML integration. To overcome this, standardizing dataset formats or employing format converters, such as Lavoisier (for converting “.spold” to “.ilcd”) and the OpenLCA converter (for various format conversions), is crucial. These tools facilitate the seamless use of different datasets, enhancing the effectiveness and efficiency of LCA+ML studies.

Bayesian models are noted for their low computational processing time, making them efficient for scenarios where computational resources are limited. However, ANN, RF, and XGBoost models effectively handle nonlinear data and establish complex relationships. These models' ability to capture intricate patterns in data allows for more accurate and reliable predictions of environmental impacts.

Sensitivity to outliers and the ability to generalize well to new data are critical characteristics for effective ML models in LCA. ANN, RF, and GBM models exhibit high sensitivity and generalization capabilities, making them adept at handling real-world data variability and ensuring robust predictive performance. Their potential for automation further enhances their practicality, enabling streamlined and efficient environmental assessments.

The construction sector predominantly benefits from KNN and SVM models, while the agricultural and chemical sectors are better served by ANN and RF models. This sector-specific preference highlights the importance of selecting appropriate ML models based on each sector's unique characteristics and requirements.

Despite the strengths of current ML models, there are notable gaps and opportunities for future research. The lower incidence of cradle-to-grave studies in LCA+ML integration suggests a need for more comprehensive assessments encompassing the entire product and process lifecycle. Additionally, extending LCA+ML applications to the tertiary sector, including commerce and service provision, represents an unexplored area with significant potential for impactful research.

Moreover, the lack of LCA+ML integration in the social context points to a significant research gap. This behavior might be due to the lack of social data samples available for training the models or it could be due to a natural research trend, as social LCA was developed after environmental LCA. Thus, addressing social dimensions alongside environmental impacts could provide a more holistic understanding of sustainability and inform more effective decision-making.

In conclusion, the integration of ML with LCA offers powerful tools for predicting environmental impacts, with ANN and RF models standing out for their versatility, scalability, and effectiveness. The insights from this analysis provide valuable guidance for researchers and practitioners in selecting appropriate ML models based on specific study requirements and highlight critical areas for future research to advance the field of sustainable environmental assessment.

2.5. CONCLUSIONS

The integration of Life Cycle Assessment (LCA) and Machine Learning (ML) across various sectors has demonstrated significant potential for predicting environmental impacts. Our analysis indicates that the volume of publications focused exclusively on LCA is smaller than that of those incorporating predictive AI, with a marked increase in publications over the past six years.

Our content analysis has established clear relationships between ML methods and the specificities of LCA studies. This is encapsulated in the synthesis table presented Section 4, which compiles key ML methods and summarizes their occurrences across different research sectors, LCA boundaries, numbers of attributes, samples, and target attributes. This compilation aims to facilitate the selection of appropriate models and prediction algorithms, providing insights into the most consolidated models and identifying gaps and opportunities for future research.

Our findings establish a descending order of ML model usage in LCA as follows: ANN > RF > SVM > LR > XGBoost > DT > ANFIS > GBM > KNN > MLR > Bayesian. Artificial Neural Networks (ANN) being the most prevalent and explored model. However, significant opportunities for future research still need to be explored, particularly in the context of cradle-to-grave studies, which are less common in LCA+ML integration. Expanding these studies to other economic sectors, such as the tertiary sector, could also enable new insights into commerce and service provision. Notably, we identified a critical gap in the integration of

LCA+ML within the social context, underscoring an important area for future exploration.

We generally observed a need for systematic studies documenting the integration possibilities between LCA and ML methodologies. Our proposition of Table 2 aims to address this gap by compiling advancements in the field, outlining the most commonly used ML methods based on research objectives, and identifying areas for future exploration.

The scarcity of datasets in main databases, combined with the diversity of saving formats (e.g., EcoSpold, ILCD, JSON-LD), poses significant challenges. These limitations hinder progress due to the restricted number of datasets available for training ML models. They may affect data integrity and interoperability, complicating the concatenation of samples for model training, testing, and validation.

In summary, our analysis highlights the potential of integrating LCA and ML to enhance the accuracy and efficiency of environmental impact predictions. However, to fully leverage the potential of ML in LCA, future research should focus on developing dynamic LCA models, improving data quality, and standardizing data formats. This will facilitate broader application and more robust integration of AI-enabled tools in LCA across diverse sectors, including social and tertiary contexts.

Conflicts of interest: The authors have no relevant financial or non-financial interests to disclose.

Acknowledgments: The authors are grateful to the “Fundação de Amparo à Pesquisa do Estado de São Paulo” (FAPESP), process number 2022/15134-1.

Data availability statement: All data generated or analyzed during this study are included in this published article and its supplementary information files.

Table 2. 2 - Synthesis table – The most relevant ML models in the context of LCA.

Criteria ↓		Models →										
		Bayesian	KNN	LR	MLR	DT	RF	GBM	XGBoost	ANN	ANFIS	SVM
Model based on:		Conditional probability	Distance	Linearity	Linearity	Decision structure	Decision structure	Decision structure	Sequential decision structures	Neural network	Neural networks and fuzzy logic	Maximize margins (optimization)
Applicati on sectors	Use in agricultural predictions?	×	×	■	×	×	×	■	■	■	■	■
	Use in construction predictions?	×	■	■	■	■	■	×	■	■	×	■
	Use in final disposal predictions?	×	×	×	×	×	×	×	×	■	×	×
	Use in energy sector predictions?	×	×	×	×	×	×	×	×	■	×	×
	Use in manufacturing predictions?	×	×	■	×	■	■	×	■	■	×	■
	Use in mining forecasts?	×	×	×	×	×	×	×	×	■	×	×
	Use in chemical sector predictions?	×	■	×	×	■	■	■	■	■	×	■
	Use in transport predictions?	■	×	×	×	■	■	×	■	■	×	×

LCA system boundary	Use in cradle-to-gate studies?	✗	☑	☑	☑	✗	✗	☑	✗	☑	✗	☑
	Use in cradle-to-grave studies?	✗	✗	☑	☑	✗	☑	☑	☑	☑	☑	☑
	Use in gate-to-gate studies?	✗	✗	✗	✗	✗	✗	✗	✗	☑	☑	☑
	Use in gate-to-grave studies?	✗	✗	✗	✗	✗	✗	✗	✗	☑	✗	✗
Number of samples	Studies with up to 100 samples?	✗	✗	☑	✗	✗	☑	✗	✗	☑	☑	✗
	Studies of 101 to 200 samples?	☑	☑	☑	☑	☑	☑	✗	☑	☑	☑	☑
	Studies from 201 to 1,000 samples?	✗	✗	✗	✗	✗	☑	✗	✗	☑	☑	☑
	Studies with more than 1,001 samples?	✗	☑	☑	☑	☑	☑	☑	☑	☑	✗	☑
Number of attributes	Studies with up to 5 attributes?	☑	✗	☑	✗	☑	☑	✗	✗	☑	✗	✗
	Studies of 6 to 10 attributes?	✗	✗	☑	✗	✗	☑	☑	✗	☑	☑	☑
	Studies of 11 to 20 attributes?	✗	☑	☑	✗	☑	☑	☑	☑	☑	☑	☑
	Studies of 21 to 50 attributes?	✗	✗	✗	☑	☑	☑	✗	✗	☑	✗	☑

	Studies with more than 51 attributes?	✗	✗	✗	✗	✗	■	✗	■	■	✗	■
Number of target attributes	Prediction of 1 target attribute?	■	■	✗	■	■	■	✗	■	■	✗	■
	Predictions of 2 to 3 target attributes?	✗	■	■	■	✗	■	■	■	■	■	■
	Predictions of 4 to 6 target attributes?	✗	✗	✗	✗	✗	■	✗	■	■	✗	■
	Predictions of 7 to 10 target attributes?	✗	✗	■	✗	✗	■	✗	✗	■	■	✗
	Predictions of more than 11 target attributes?	✗	✗	■	✗	■	■	✗	■	■	■	■
Data require	Effectiveness with linear type data	●	●	■	■	●	●	●	●	■	✗	■
	Effectiveness with nonlinear data	●	●	✗	✗	■	■	■	■	■	●	■
	Computational processing time	■	✗	■	●	●	●	●	●	✗	✗	●
Other info	Model sensitivity to outliers?	■	■	✗	✗	●	●	●	●	■	■	●
	Establish complex relationships?	✗	●	✗	✗	■	■	■	■	■	●	■
	Possibility to automate data?	■	■	■	■	●	●	●	●	●	●	●

Total	✗ Not applicable in the question / Not effective / Insensitive / High computational processing time	23	19	15	21	15	7	18	12	1	17	8
	⦿ Average effectiveness / Average sensitivity / Average computational processing time	2	3	0	1	4	4	4	4	1	3	3
	■ Occurs in the question / Effective / Sensitive / Low computational processing time	7	10	17	10	13	21	10	16	30	12	21

Source: own authorship.

***Table legend: KNN: K - Nearest Neighbors; LR: linear regression; MLR: Multiple linear regression; DT: decision tree; RF: random forest; GBM: Gradient Boosting Machine; XGBoost: eXtreme Boosting Machine; ANN: artificial neural network; ANFIS: Adaptive Network-Based Fuzzy Inference System; SVM: Support vector machine.

REFERENCES

- Ahmed, S., Li, T., Li, S. Y., & Chen, R. (2023). Comparative life cycle impact assessment of offshore support vessels powered by alternative fuels for sustainable offshore wind operations using machine learning. *Journal of Ocean Engineering and Science*. <https://doi.org/10.1016/j.joes.2023.10.005>
- Akhshik, M., Bilton, A., Tjong, J., Singh, C. V., Faruk, O., & Sain, M. (2022). Prediction of greenhouse gas emissions reductions via machine learning algorithms: Toward an artificial intelligence-based life cycle assessment for automotive lightweighting. *Sustainable Materials and Technologies*, 31. <https://doi.org/10.1016/j.susmat.2021.e00370>
- Asif, Z., Chen, Z. (2020). A life cycle based air quality modeling and decision support system (LCAQMS) for sustainable mining management. *Journal of Environmental Informatics*, 35(2), 103–117. <https://doi.org/10.3808/jei.201900406>
- Asif, Z., Chen, Z., & Zhu, Z. H. (2019). An integrated life cycle inventory and artificial neural network model for mining air pollution management. *International Journal of Environmental Science and Technology*, 16(4), 1847–1856. <https://doi.org/10.1007/s13762-018-1813-9>
- Azari, R., Garshasbi, S., Amini, P., Rashed-Ali, H., & Mohammadi, Y. (2016). Multi-objective optimization of building envelope design for life cycle environmental performance. *Energy and Buildings*, 126, 524–534. <https://doi.org/10.1016/j.enbuild.2016.05.054>
- Barros, N. N. Ruschel, R. C. (2021). Machine Learning for Whole-Building Life Cycle Assessment: A Systematic Literature Review (pp. 109–122). https://doi.org/10.1007/978-3-030-51295-8_10
- Chen, Z., Liu, L., & Li, C. (2021). Prediction and control of carbon emissions of electric vehicles based on BP neural network under carbon neutral background. *International Conference on Neural Networks*, 110. <https://doi.org/10.1117/12.2615330>
- Chen, L., Chen, Z., Zhang, Y., Liu, Y., Osman, A. I., Farghali, M., Hua, J., Al-Fatesh, A., Ihara, I., Rooney, D. W., & Yap, P.-S. (2023). Artificial intelligence-based solutions for climate change: a review. *Environmental Chemistry Letters*, 21(5), 2525–2557. <https://doi.org/10.1007/s10311-023-01617-y>
- Cheng, F., Porter, M.D., Colosi, L.M., 2020. Is hydrothermal treatment coupled with carbon capture and storage an energy-producing negative emissions technology? *Energy Convers. Manag.* 112252.
- D'Amico, A., Ciulla, G., Traverso, M., Lo Brano, V., & Palumbo, E. (2019). Artificial Neural Networks to assess energy and environmental performance of buildings: An Italian case study. *Journal of Cleaner Production*, 239. <https://doi.org/10.1016/j.jclepro.2019.117993>
- da Costa, T. P., da Costa, D. M. B., & Murphy, F. (2024). A systematic review of real-time data monitoring and its potential application to support dynamic life cycle inventories. *Environmental Impact Assessment Review*, 105, 107416. <https://doi.org/10.1016/j.eiar.2024.107416>
- Das, K., Behera, R. N. (2017). A Survey on Machine Learning: Concept, Algorithms and Applications. *International Journal of Innovative Research in Computer and Communication Engineering*, 5(2), 1301–1309.
- Denyer, D., Tranfield, D. (2009). Producing a systematic review. In *The sage handbook of organizational research methods* (Vol. 1, pp. 671–689). SAGE.
- Duprez, S., Fouquet, M., Herreros, Q., & Jusselme, T. (2019). Improving life cycle-based exploration methods by coupling sensitivity analysis and metamodels. *Sustainable Cities and Society*, 44, 70–84. <https://doi.org/10.1016/j.scs.2018.09.032>

Elhami, B., Khanali, M., & Akram, A. (2017). Combined application of Artificial Neural Networks and life cycle assessment in lentil farming in Iran. *Information Processing in Agriculture*, 4(1), 18–32. <https://doi.org/10.1016/j.inpa.2016.10.004>

European Commission-Joint Research Centre - Institute for Environment and Sustainability: International Reference Life Cycle Data System (ILCD). (2011). *Handbook- Recommendations for Life Cycle Impact Assessment in the European context*.

Faceli, K., Lorena, A. C., Gama, J., Almeida, T. A. de, & Carvalho, A. C. P. L. F. de. (2022). *Inteligência artificial: uma abordagem de aprendizado de máquina* (LTC, Ed.; 2nd ed.).

Fang, Y., Lu, X., & Li, H. (2021). A random forest-based model for the prediction of construction-stage carbon emissions at the early design stage. *Journal of Cleaner Production*, 328. <https://doi.org/10.1016/j.jclepro.2021.129657>

Faridmehr, I., Nehdi, M. L., Nikoo, M., Huseien, G. F., & Ozbakkaloglu, T. (2021). Life-cycle assessment of alkali-activated materials incorporating industrial byproducts. *Materials*, 14(9). <https://doi.org/10.3390/ma14092401>

Feng, K., Lu, W., & Wang, Y. (2019). Assessing environmental performance in early building design stage: An integrated parametric design and machine learning method. *Sustainable Cities and Society*, 50. <https://doi.org/10.1016/j.scs.2019.101596>

Figuerola-García, J. C., Ochoa-Rey, C. M., & Avellaneda-González, J. A. (2015). Rule generation of fuzzy logic systems using a self-organized fuzzy neural network. *Neurocomputing*, 151, 955–962. <https://doi.org/10.1016/j.neucom.2014.09.079>

Filho, A. R. M. Russo, S. L. (2018). *Biblios* (pp. 50–67).

Fleitmann, L., Kleinekorte, J., Leonhard, K., & Bardow, A. (2021). COSMO-susCAMPD: Sustainable solvents from combining computer-aided molecular and process design with predictive life cycle assessment. *Chemical Engineering Science*, 245. <https://doi.org/10.1016/j.ces.2021.116863>

Ghasemi-Mobtaker, H., Kaab, A., Rafiee, S., & Nabavi-Pelesaraei, A. (2022). A comparative of modeling techniques and life cycle assessment for prediction of output energy, economic profit, and global warming potential for wheat farms. *Energy Reports*, 8, 4922–4934. <https://doi.org/10.1016/j.egyr.2022.03.184>

Gu, X., Cook, N. J., Metcalfe, A. V, & Aldrich, C. (2023). Change detection in drill core images based on local binary pattern. *MODSIM2023, 25th International Congress on Modelling and Simulation*. <https://doi.org/10.36334/modsim.2023.gu67>

GLAD. (2024). About. <https://www.globalcadataaccess.org/about>

Hamida, A., Alsudairi, A., Alshaibani, K., & Alshamrani, O. (2021). Environmental impacts cost assessment model of residential building using an artificial neural network. *Engineering, Construction and Architectural Management*, 28(10), 3190–3215. <https://doi.org/10.1108/ECAM-06-2020-0450>

Hou, P., Jolliet, O., Zhu, J., & Xu, M. (2020). Estimate ecotoxicity characterization factors for chemicals in life cycle assessment using machine learning models. *Environment International*, 135. <https://doi.org/10.1016/j.envint.2019.105393>

International Monetary Fund (IMF). (2023). *IMF Fossil Fuel Subsidies Data: 2023 Update*.

James Montantes. (2020). 3 Reasons to Use Random Forest Over a Neural Network—Comparing Machine Learning versus Deep Learning. <https://towardsdatascience.com/3-reasons-to-use-random-forest-over-a-neural-network-comparing-machine-learning-versus-deep-f9d65a154d89>

Journal of Cleaner Production. (2024). Call for papers. <https://www.sciencedirect.com/journal/journal-of-cleaner-production/about/call-for-papers>

Kaab, A., Sharifi, M., Mobli, H., Nabavi-Pelesaraei, A., & Chau, K. wing. (2019). Combined life cycle assessment and artificial intelligence for prediction of output energy and environmental impacts of sugarcane production. *Science of the Total Environment*, 664, 1005–1019. <https://doi.org/10.1016/j.scitotenv.2019.02.004>

Kamyab, H., Khademi, T., Chelliapan, S., SaberiKamarposhti, M., Rezanian, S., Yusuf, M., Farajnezhad, M., Abbas, M., Hun Jeon, B., & Ahn, Y. (2023). The latest innovative avenues for the utilization of artificial Intelligence and big data analytics in water resource management. *Results in Engineering*, 20, 101566. <https://doi.org/10.1016/j.rineng.2023.101566>

Karka, P., Papadokostantakis, S., & Kokossis, A. (2019). Environmental impact assessment of biomass process chains at early design stages using decision trees. *International Journal of Life Cycle Assessment*, 24(9), 1675–1700. <https://doi.org/10.1007/s11367-019-01591-0>

Karka, P., Papadokostantakis, S., & Kokossis, A. (2022). Digitizing sustainable process development: From ex-post to ex-ante LCA using machine-learning to evaluate bio-based process technologies ahead of detailed design. *Chemical Engineering Science*, 250. <https://doi.org/10.1016/j.ces.2021.117339>

Khanali, M., Mobli, H., & Hosseinzadeh-Bandbafha, H. (2017). Modeling of yield and environmental impact categories in tea processing units based on artificial neural networks. *Environmental Science and Pollution Research*, 24(34), 26324–26340. <https://doi.org/10.1007/s11356-017-0234-5>

Khoshnevisan, B., Rafiee, S., & Mousazadeh, H. (2013a). Environmental impact assessment of open field and greenhouse strawberry production. *European Journal of Agronomy*, 50, 29–37. <https://doi.org/10.1016/j.eja.2013.05.003>

Khoshnevisan, B., Rafiee, S., Omid, M., Mousazadeh, H., & Sefeedpari, P. (2013b). Prognostication of environmental indices in potato production using artificial neural networks. *Journal of Cleaner Production*, 52, 402–409. <https://doi.org/10.1016/j.jclepro.2013.03.028>

Khoshnevisan, B., Rafiee, S., Omid, M., Mousazadeh, H., & Clark, S. (2014b). Environmental impact assessment of tomato and cucumber cultivation in greenhouses using life cycle assessment and adaptive neuro-fuzzy inference system. *Journal of Cleaner Production*, 73, 183–192. <https://doi.org/10.1016/j.jclepro.2013.09.057>

Khoshnevisan, B., Rajaeifar, M. A., Clark, S., Shamahirband, S., Anuar, N. B., Mohd Shuib, N. L., & Gani, A. (2014a). Evaluation of traditional and consolidated rice farms in Guilan Province, Iran, using life cycle assessment and fuzzy modeling. *Science of the Total Environment*, 481(1), 242–251. <https://doi.org/10.1016/j.scitotenv.2014.02.052>

Kleinekorte, J., Kleppich, J., Fleitmann, L., Beckert, V., Blodau, L., & Bardow, A. (2023). APPROPRIATE Life Cycle Assessment: A PROcess-Specific, PRedictive Impact Assessment Method for Emerging Chemical Processes. *ACS Sustainable Chemistry and Engineering*, 11(25), 9303–9319. <https://doi.org/10.1021/acssuschemeng.2c07682>

Koyamparabath, A., Adibi, N., Szablewski, C., Adibi, S. A., & Sonnemann, G. (2022). Implementing Artificial Intelligence Techniques to Predict Environmental Impacts: Case of Construction Products. *Sustainability (Switzerland)*, 14(6). <https://doi.org/10.3390/su14063699>

- Ledziński, Ł. Grzešek, G. (2023). Artificial Intelligence Technologies in Cardiology. *Journal of Cardiovascular Development and Disease*, 10(5), 202. <https://doi.org/10.3390/jcdd10050202>
- Lee, E. K., Zhang, W. J., Zhang, X., Adler, P. R., Lin, S., Feingold, B. J., Khwaja, H. A., & Romeiko, X. X. (2020). Projecting life-cycle environmental impacts of corn production in the U.S. Midwest under future climate scenarios using a machine learning approach. *Science of the Total Environment*, 714. <https://doi.org/10.1016/j.scitotenv.2020.136697>
- Lena Nickel (ECOCHAIN). The most used LCI Databases for Life Cycle Assessment (LCA). Retrieved April 15, 2024, from <https://ecochain.com/blog/lci-databases-in-lca/>
- Liu, Y., Li, T., Xu, W., Wang, Q., Huang, H., & He, B. J. (2023). Building information modelling-enabled multi-objective optimization for energy consumption parametric analysis in green buildings design using hybrid machine learning algorithms. *Energy and Buildings*, 300. <https://doi.org/10.1016/j.enbuild.2023.113665>
- López-Flores, F. J., Sánchez-Zarco, X. G., Rubio-Castro, E., & Ponce-Ortega, J. M. (2023). A machine learning approach for optimizing the water-energy-food-ecosystem nexus: a resilience perspective for sustainability. *Environment, Development and Sustainability*. <https://doi.org/10.1007/s10668-023-04257-y>
- Martínez-Rocamora, A., Rivera-Gómez, C., Galán-Marín, C., & Marrero, M. (2021). Environmental benchmarking of building typologies through BIM-based combinatorial case studies. *Automation in Construction*, 132. <https://doi.org/10.1016/j.autcon.2021.103980>
- Mohammadi Kashka, F., Tahmasebi Sarvestani, Z., Pirdashti, H., Motevali, A., Nadi, M., & Valipour, M. (2023). Sustainable Systems Engineering Using Life Cycle Assessment: Application of Artificial Intelligence for Predicting Agro-Environmental Footprint. *Sustainability (Switzerland)*, 15(7). <https://doi.org/10.3390/su15076326>
- Mousavi-Avval, S. H., Rafiee, S., Sharifi, M., Hosseinpour, S., & Shah, A. (2017). Combined application of Life Cycle Assessment and Adaptive Neuro-Fuzzy Inference System for modeling energy and environmental emissions of oilseed production. In *Renewable and Sustainable Energy Reviews* (Vol. 78, pp. 807–820). Elsevier Ltd. <https://doi.org/10.1016/j.rser.2017.05.002>
- Moutik, B., Summerscales, J., Graham-Jones, J., & Pemberton, R. (2023). Life Cycle Assessment Research Trends and Implications: A Bibliometric Analysis. *Sustainability*, 15(18), 13408. <https://doi.org/10.3390/su151813408>
- Nabavi-Pelesaraei, A., Abdi, R., & Rafiee, S. (2016b). Neural network modeling of energy use and greenhouse gas emissions of watermelon production systems. *Journal of the Saudi Society of Agricultural Sciences*, 15(1), 38–47. <https://doi.org/10.1016/j.jssas.2014.05.001>
- Nabavi-Pelesaraei, A., Bayat, R., Hosseinzadeh-Bandbafha, H., Afrasyabi, H., & Berrada, A. (2017b). Prognostication of energy use and environmental impacts for recycle system of municipal solid waste management. *Journal of Cleaner Production*, 154, 602–613. <https://doi.org/10.1016/j.jclepro.2017.04.033>
- Nabavi-Pelesaraei, A., Bayat, R., Hosseinzadeh-Bandbafha, H., Afrasyabi, H., & Chau, K. wing. (2017a). Modeling of energy consumption and environmental life cycle assessment for incineration and landfill systems of municipal solid waste management - A case study in Tehran Metropolis of Iran. *Journal of Cleaner Production*, 148, 427–440. <https://doi.org/10.1016/j.jclepro.2017.01.172>
- Nabavi-Pelesaraei, A., Ghasemi-Mobtaker, H., Salehi, M., Rafiee, S., Chau, K. W., & Ebrahimi, R.

- (2023). Machine Learning Models of Exergo Environmental Damages and Emissions Social Cost for Mushroom Production. *Agronomy*, 13(3). <https://doi.org/10.3390/agronomy13030737>
- Nabavi-Pelesaraei, A., Rafiee, S., Hosseinzadeh-Bandbafha, H., & Shamshirband, S. (2016a). Modeling energy consumption and greenhouse gas emissions for kiwifruit production using artificial neural networks. *Journal of Cleaner Production*, 133, 924–931. <https://doi.org/10.1016/j.jclepro.2016.05.188>
- Nabavi-Pelesaraei, A., Rafiee, S., Mohtasebi, S. S., Hosseinzadeh-Bandbafha, H., & Chau, K. wing. (2018). Integration of artificial intelligence methods and life cycle assessment to predict energy output and environmental impacts of paddy production. *Science of the Total Environment*, 631–632, 1279–1294. <https://doi.org/10.1016/j.scitotenv.2018.03.088>
- Nabavi-Pelesaraei, A., Rafiee, S., Mohtasebi, S. S., Hosseinzadeh-Bandbafha, H., & Chau, K. wing. (2019). Comprehensive model of energy, environmental impacts and economics in rice milling factories by coupling adaptive neuro-fuzzy inference system and life cycle assessment. *Journal of Cleaner Production*, 217, 742–756. <https://doi.org/10.1016/j.jclepro.2019.01.228>
- Naser, A. Z., Defersha, F., Xu, X., & Yang, S. (2023). Automating life cycle assessment for additive manufacturing with machine learning: Framework design, dataset buildup, and a case study. *Journal of Manufacturing Systems*, 71, 504–526. <https://doi.org/10.1016/j.jmsy.2023.10.009>
- Oliveira, J. A. de, Silva, D. A. L., Puglieri, F. N., & Saavedra, Y. M. B. (2021). *Life Cycle Engineering and Management of Products: theory and practice* (Springer Cham, Ed.; 1st ed., Vol. 1).
- Onyelowe, K. C., Ebid, A. M., Mahdi, H. A., Soleymani, A., Jahangir, H., & Dabbaghi, F. (2022). Optimization of Green Concrete Containing Fly Ash and Rice Husk Ash Based on Hydro-Mechanical Properties and Life Cycle Assessment Considerations. *Civil Engineering Journal (Iran)*, 8(12), 3912–3938. <https://doi.org/10.28991/CEJ-2022-08-12-018>
- Onyelowe, K. C., Ebid, A. M., Mahdi, H. A., Onyelowe, F. K. C., Shafieyoon, Y., Onyia, M. E., & Onah, H. N. (2023). AI Mix Design of Fly Ash Admixed Concrete Based on Mechanical and Environmental Impact Considerations. *Civil Engineering Journal (Iran)*, 9(Special Issue), 27–45. <https://doi.org/10.28991/CEJ-SP2023-09-03>
- Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews*, 5(1), 210. <https://doi.org/10.1186/s13643-016-0384-4>
- Ozbilen, A., Aydin, M., Dincer, I., & Rosen, M. A. (2013). Life cycle assessment of nuclear-based hydrogen production via a copper-chlorine cycle: A neural network approach. *International Journal of Hydrogen Energy*, 38(15), 6314–6322. <https://doi.org/10.1016/j.ijhydene.2013.03.071>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, n71. <https://doi.org/10.1136/bmj.n71>
- Pan, H. Wu, C. (2023). Bayesian optimization + XGBoost based life cycle carbon emission prediction for residential buildings—An example from Chengdu, China. *Building Simulation*, 16(8), 1451–1466. <https://doi.org/10.1007/s12273-023-1024-2>
- Pishgar-Komleh, S. H., Akram, A., Keyhani, A., Sefeedpari, P., Shine, P., & Brandao, M. (2020). Integration of life cycle assessment, artificial neural networks, and metaheuristic optimization algorithms for optimization of tomato-based cropping systems in Iran. *International Journal of Life Cycle Assessment*, 25(3), 620–632. <https://doi.org/10.1007/s11367-019-01707-6>

Ploszaj-Mazurek, M., Ryńska, E., & Grochulska-Salak, M. (2020). Methods to optimize carbon footprint of buildings in regenerative architectural design with the use of machine learning, convolutional neural network, and parametric design. *Energies*, 13(20). <https://doi.org/10.3390/en13205289>

Prenner, J. A. A. Robbes, R. (2022). Making the most of small Software Engineering datasets with modern machine learning. *IEEE Transactions on Software Engineering*, 1–1. <https://doi.org/10.1109/TSE.2021.3135465>

Research XP. (2024, February 23). O que são commodities? Confira o guia completo. <https://conteudos.xpi.com.br/aprenda-a-investir/relatorios/commodities/>

Romeiko, X. X., Guo, Z., Pang, Y., Lee, E. K., & Zhang, X. (2020). Comparing machine learning approaches for predicting spatially explicit life cycle global warming and eutrophication impacts from corn production. *Sustainability (Switzerland)*, 12(4). <https://doi.org/10.3390/su12041481>

Romeiko, X. X., Zhang, X., Pang, Y., Gao, F., Xu, M., Lin, S., & Babbitt, C. (2024). A review of machine learning applications in life cycle assessment studies. In *Science of the Total Environment* (Vol. 912). Elsevier B.V. <https://doi.org/10.1016/j.scitotenv.2023.168969>

Singh, P. K. Sarkar, P. (2023). An artificial neural network tool to support the decision making of designers for environmentally conscious product development. *Expert Systems with Applications*, 212. <https://doi.org/10.1016/j.eswa.2022.118679>

Song, R., Keller, A. A., & Suh, S. (2017). Rapid Life-Cycle Impact Screening Using Artificial Neural Networks. *Environmental Science and Technology*, 51(18), 10777–10785. <https://doi.org/10.1021/acs.est.7b02862>

Sun, Y., Wang, X., Ren, N., Liu, Y., & You, S. (2023). Improved Machine Learning Models by Data Processing for Predicting Life-Cycle Environmental Impacts of Chemicals. *Environmental Science and Technology*, 57(8), 3434–3444. <https://doi.org/10.1021/acs.est.2c04945>

Sustainability. (2024). Special Issues. https://www.mdpi.com/journal/sustainability/special_issues?section_id=&search=artificial+intelligence&sort=deadline&view=all&page_count=50&query=artificial+intelligence

Taghavifar, H. Mardani, A. (2015). Prognostication of energy consumption and greenhouse gas (GHG) emissions analysis of apple production in West Azarbayjan of Iran using artificial neural networks. *Journal of Cleaner Production*, 87(C), 159–167. <https://doi.org/10.1016/j.jclepro.2014.10.054>

Teodosio, B., Wasantha, P. L. P., Yaghoubi, E., Guerrieri, M., Fragomeni, S., & van Staden, R. (2023). Environmental, economic, and serviceability attributes of residential foundation slabs: A comparison between waffle and stiffened rafts using multi-output deep learning. *Journal of Building Engineering*, 80. <https://doi.org/10.1016/j.jobbe.2023.107983>

Tsay, Y. S., Yeh, C. Y. Chen, Y. H., Lu, M. C., Lin, Y. C. (2021). A machine learning-based prediction model of lcco2 for building envelope renovation in Taiwan. *Sustainability (Switzerland)*, 13(15). <https://doi.org/10.3390/su13158209>

United Nations Environment Programme, & International Resource Panel. (2024). Global Resources Outlook (GRO) 2024 - Bend the trend: Pathways to a Liveable Planet as Resource Use Spikes. <https://wedocs.unep.org/20.500.11822/44901zadeh>

Van Eck, N. J. Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric

mapping. *Scientometrics*, 84(2), 523–538. <https://doi.org/10.1007/s11192-009-0146-3>

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, 11(1), 233. <https://doi.org/10.1038/s41467-019-14108-y>

Vlontzos, G. Pardalos, P. M. (2017). Assess and prognosticate greenhouse gas emissions from agricultural production of EU countries, by implementing DEA Window analysis and artificial neural networks. In *Renewable and Sustainable Energy Reviews* (Vol. 76, pp. 155–162). Elsevier Ltd. <https://doi.org/10.1016/j.rser.2017.03.054>

White, K., Hardisty, D. J., & Habib, R. (2019, April 10). The Elusive Green Consumer. <https://hbr.org/2019/07/the-elusive-green-consumer>

WorldGBC. (2023). GLOBAL POLICY PRINCIPLES FOR A SUSTAINABLE BUILT ENVIRONMENT. https://worldgbc.org/wp-content/uploads/2023/04/WorldGBC-Global-Policy-Principles_FINAL.pdf

Wu, J. Shang, S. (2020). Managing Uncertainty in AI-Enabled Decision Making and Achieving Sustainability. *Sustainability*, 12(21), 8758. <https://doi.org/10.3390/su12218758>

Xikai, M., Lixiong, W., Jiwei, L., Xiaoli, Q., & Tongyao, W. (2019). Comparison of regression models for estimation of carbon emissions during building's lifecycle using designing factors: a case study of residential buildings in Tianjin, China. *Energy and Buildings*, 204. <https://doi.org/10.1016/j.enbuild.2019.109519>

Yan, Z., Li, W., Yan, T., Wang, J., Chen, L., Lu, Y., Liu, H., Tang, J., Zhang, L., Chen, Y., Chang, S., & Hou, F. (2018). Application and validity of BP neural networks on prediction of carbon emissions from corn production in Hexi Oasis. *Chinese Journal of Eco-Agriculture*, 26(8), 1100–1106. <https://doi.org/10.13930/j.cnki.cjea.180084>

Zadeh, L. A. (2015). Fuzzy logic—a personal perspective. *Fuzzy Sets and Systems*, 281, 4–20. <https://doi.org/10.1016/j.fss.2015.05.009>

Zhang, S., Huo, Z., & Zhai, C. (2022). Building Carbon Emission Scenario Prediction Using STIRPAT and GA-BP Neural Network Model. *Sustainability (Switzerland)*, 14(15). <https://doi.org/10.3390/su14159369>

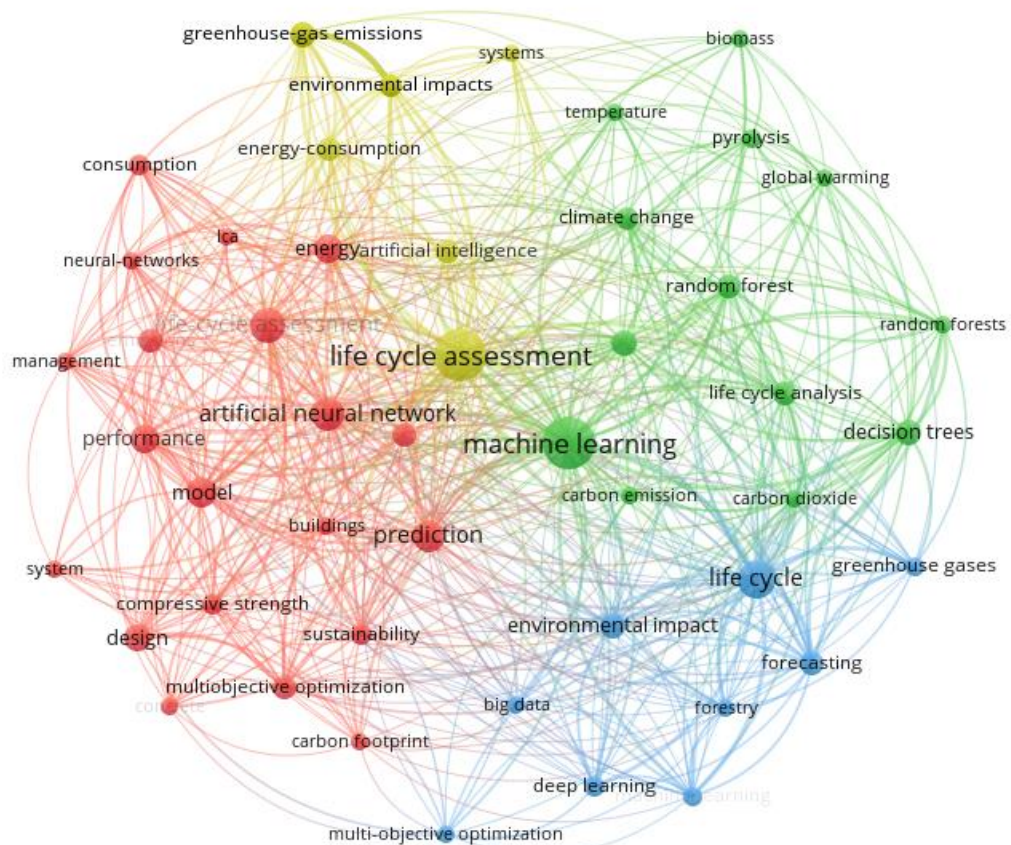
Zheng, C., Yalin, G., & Chun, G. (2022). Greenhouse Gas Prediction Using Whale Optimization Algorithm Deep Extreme Learning Machine in Chengdu Metro Construction Stage. *Tunnel Construction*, 42(12), 2048–2063.

Zheng, L., Mueller, M., Luo, C., & Yan, X. (2024). Predicting whole-life carbon emissions for buildings using different machine learning algorithms: A case study on typical residential properties in Cornwall, UK. *Applied Energy*, 357. <https://doi.org/10.1016/j.apenergy.2023.122472>

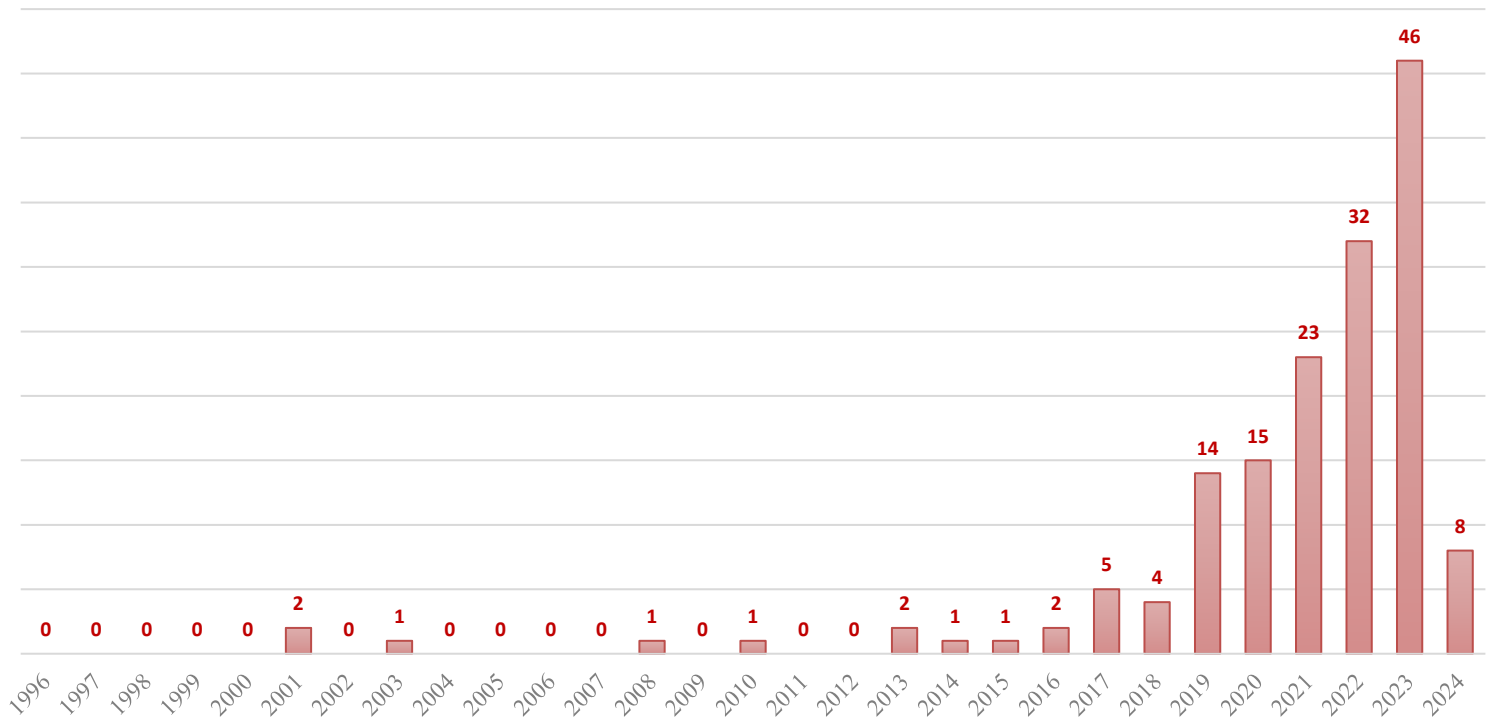
Zhu, X., Ho, C. H., & Wang, X. (2020). Application of Life Cycle Assessment and Machine Learning for High-Throughput Screening of Green Chemical Substitutes. *ACS Sustainable Chemistry and Engineering*, 8(30), 11141–11151. <https://doi.org/10.1021/acssuschemeng.0c02211>

APENDIX – 2A

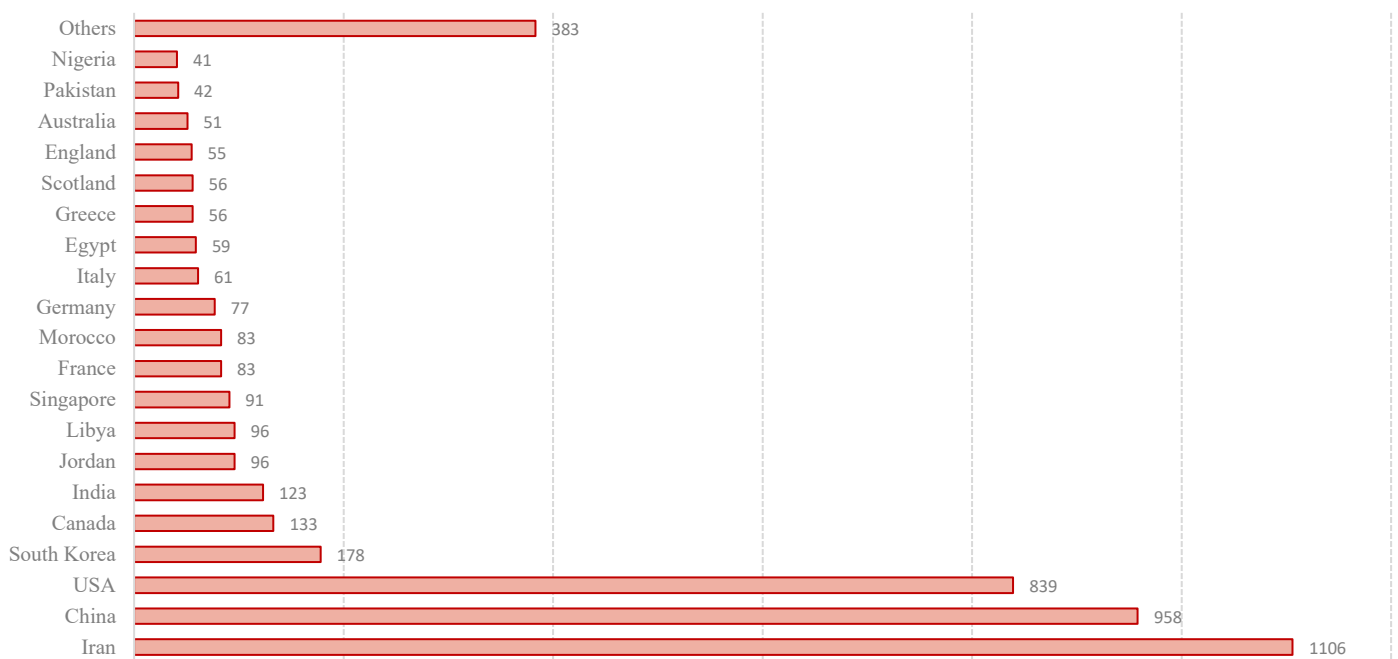
Fig. S1 Similarity visualization map: Scopus and Web of Science (50 most frequent words).



Source: own authorship, VOSViewer software.

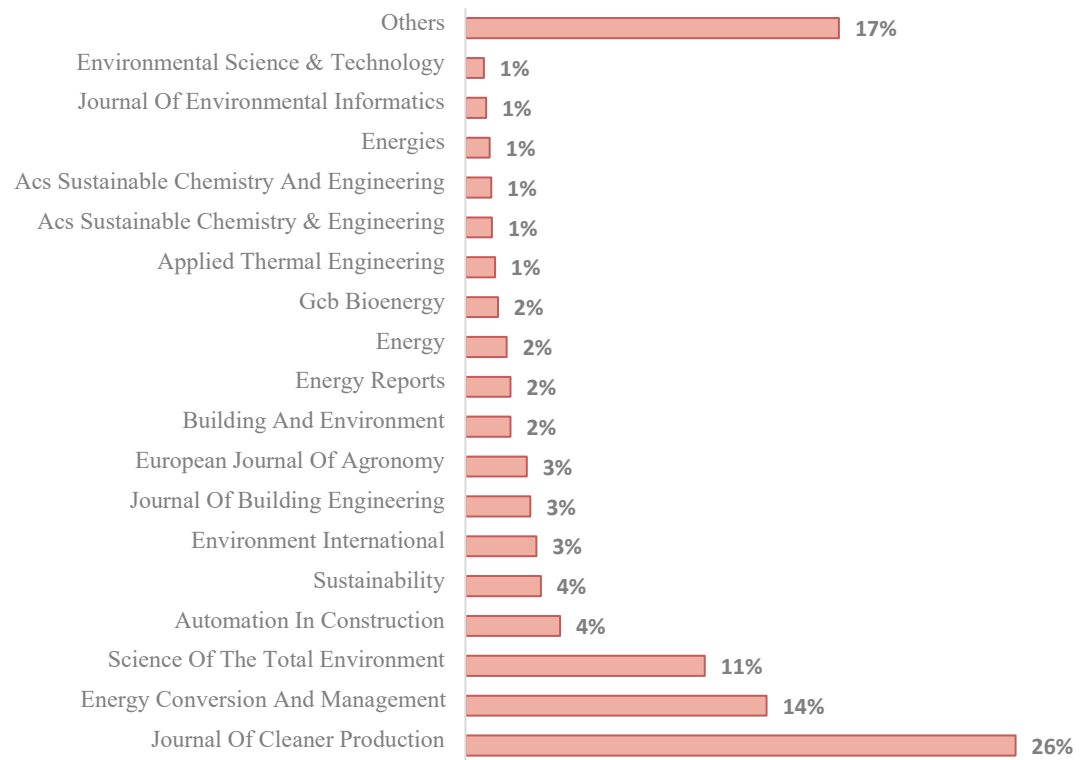
Fig. S2 Annual distribution of works (own authorship) – in number of works.

Source: own authorship.

Fig. S3 Geographical distribution of citations (own authorship) – in number of citations.

Source: own authorship.

Fig. S4 Distribution of citations by data sources/journals – in %.



Source: own authorship.

Table S1 Publication areas of the analyzed works.

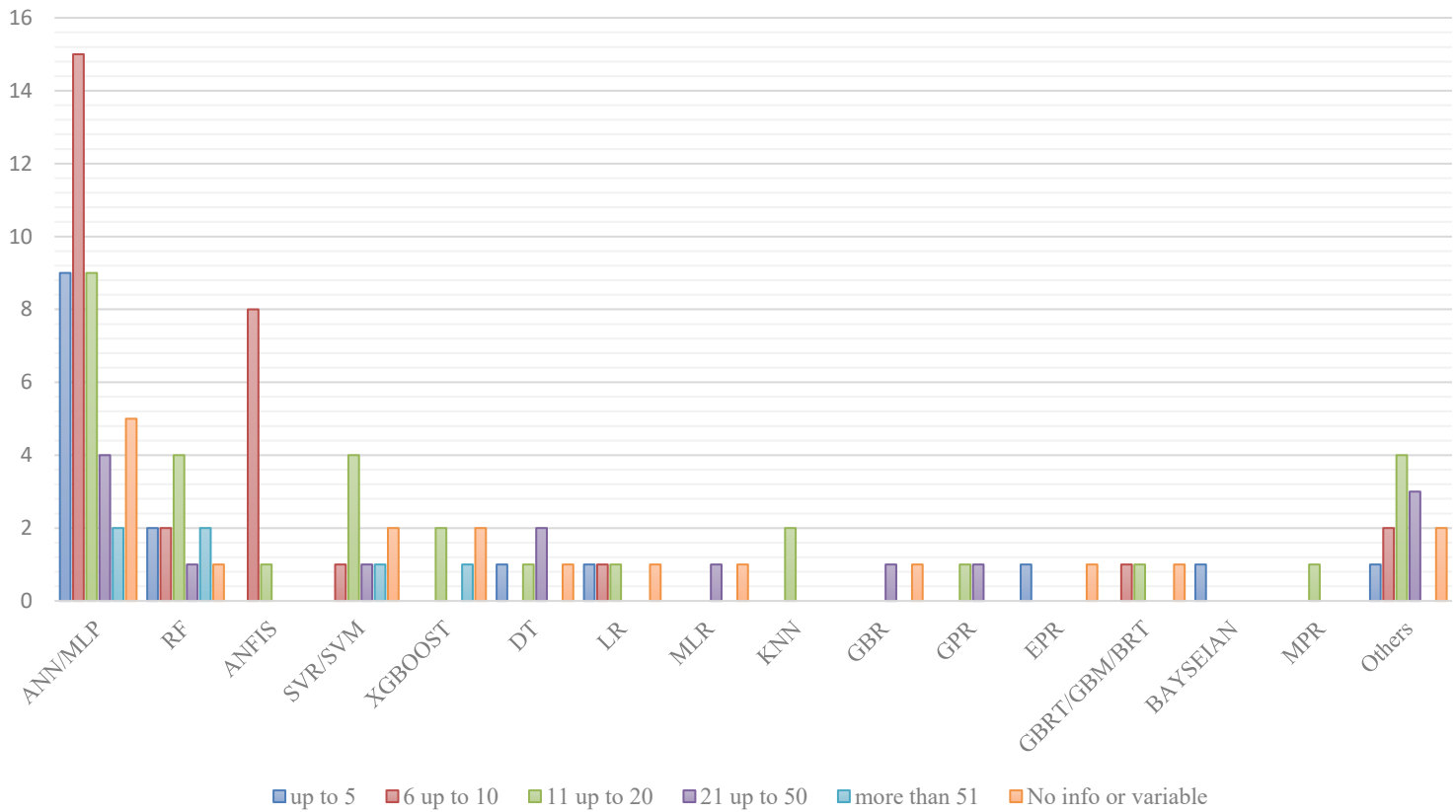
Sector	Sub-sector	N.	References
Agriculture		21	
	Crop cultivation	18	(Ghasemi-Mobtaker et al., 2022); (Yan et al., 2018); (Kaab et al., 2019); (Romeiko et al., 2020); (Khoshnevisan et al., 2013a); (Khoshnevisan, et al., 2014a); (Khoshnevisan et al., 2014b); (Nabavi-Pelesaraei et al., 2018); (Nabavi-Pelesaraei et al., 2023); (Nabavi-Pelesaraei et al., 2016a); (Nabavi-Pelesaraei et al., 2016b); (Khoshnevisan et al., 2013b); (Lee et al., 2020); (Mohammadi Kashka et al., 2023); (Mousavi-Avval et al., 2017); (Pishgar-Komleh et al., 2020); (Elhami et al., 2017); (Taghavifar & Mardani, 2015)
	Processing	2	(Nabavi-Pelesaraei et al., 2019); (Khanali et al., 2017)
	Sector emissions analysis by country	1	(Vlontzos and Pardalos, 2017)
Construction		20	
	Building emissions and support in the design phase	15	(Tsay et al., 2021); (Fang et al., 2021); (D'Amico et al., 2019); (Feng et al., 2019); (Pan and Wu, 2023); (Zhang et al., 2022); (Liu et al., 2023); (Martínez-Rocamora et al., 2021); (Hamida et al., 2021); (Koyamparambath et al., 2022b); (Duprez et al., 2019); (L. Zheng et al., 2024); (Płoszaj-Mazurek et al., 2020); (Azari et al., 2016); (Xikai et al., 2019)
	Concrete composition	3	(Onyelowe et al., 2023); (Faridmehr et al., 2021); (Onyelowe et al., 2022)
	Subway construction	1	(C. Zheng et al., 2022)
	Slabs	1	(Teodosio et al., 2023)
Chemicals		8	
	Choosing chemicals	2	(Zhu et al., 2020); (Fleitmann et al., 2021)

	Chemical processes	1	(Kleinekorte et al., 2023)
	Biorefineries	1	(Karka et al., 2022)
	Biomass	1	(Karka et al., 2019)
	Hazardous concentrations of ecotoxicity	1	(Hou et al., 2020)
	Life cycle impacts	2	(Sun et al., 2023); (Song et al., 2017)
Final disposal		2	
	Landfill and incineration	1	(Nabavi-Pelesaraei et al., 2017a)
	Recycling	1	(Nabavi-Pelesaraei et al., 2017b)
Energy		1	
	Nuclear hydrogen	1	(Ozbilen et al., 2013)
Manufacturing		2	
	Products design	1	(Singh and Sarkar, 2023)
	Additive manufacturing	1	(Naser et al., 2023)
Mining		2	
	Mine management (air pollution)	1	(Asif and Chen, 2020)
	Metal mining	1	(Asif et al., 2019)
Transport		3	

	Automotive	2	(Chen et al., 2021); (Akhshik et al., 2022)
	Tug (AHTS - Anchor Handling Tug Supply)	1	(Ahmed et al., 2023)

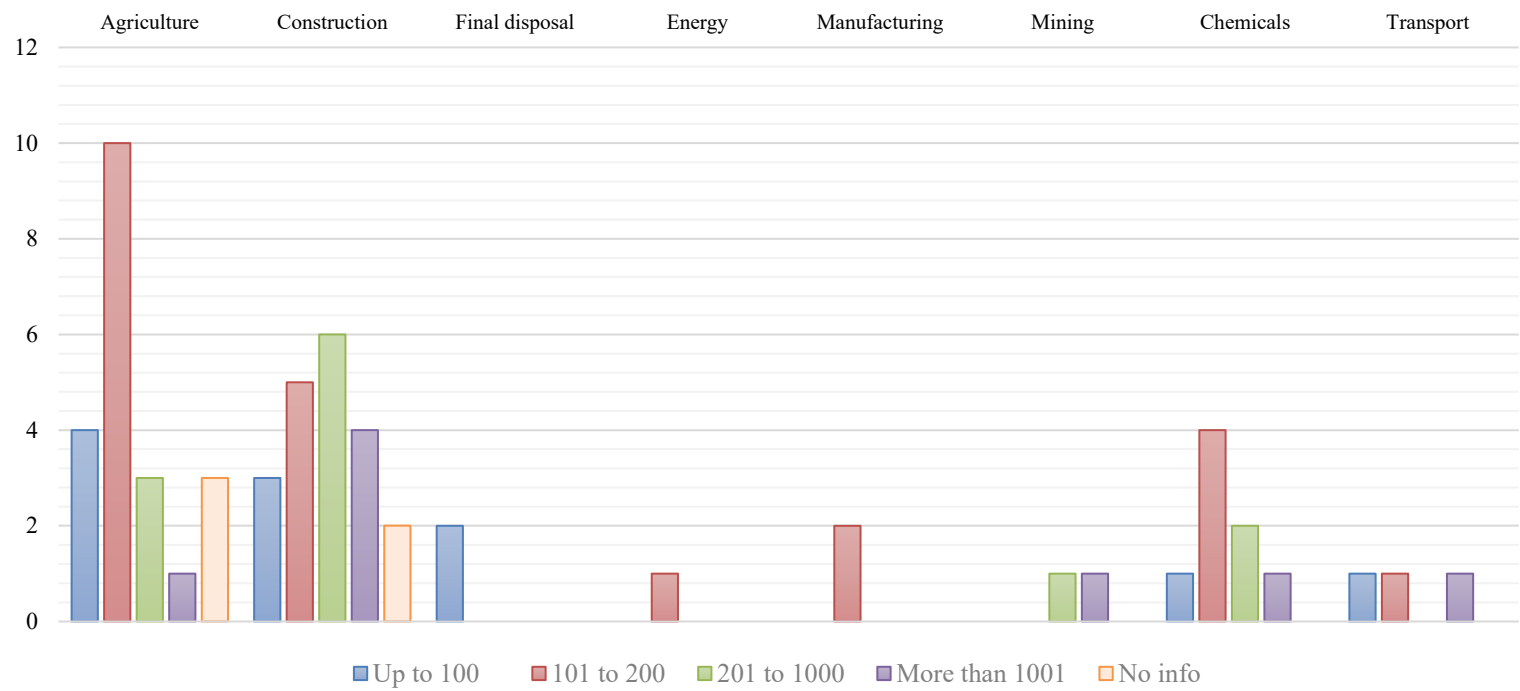
Source: own authorship.

Fig. S5 Analysis of independent variables by ML method – in number of occurrences.



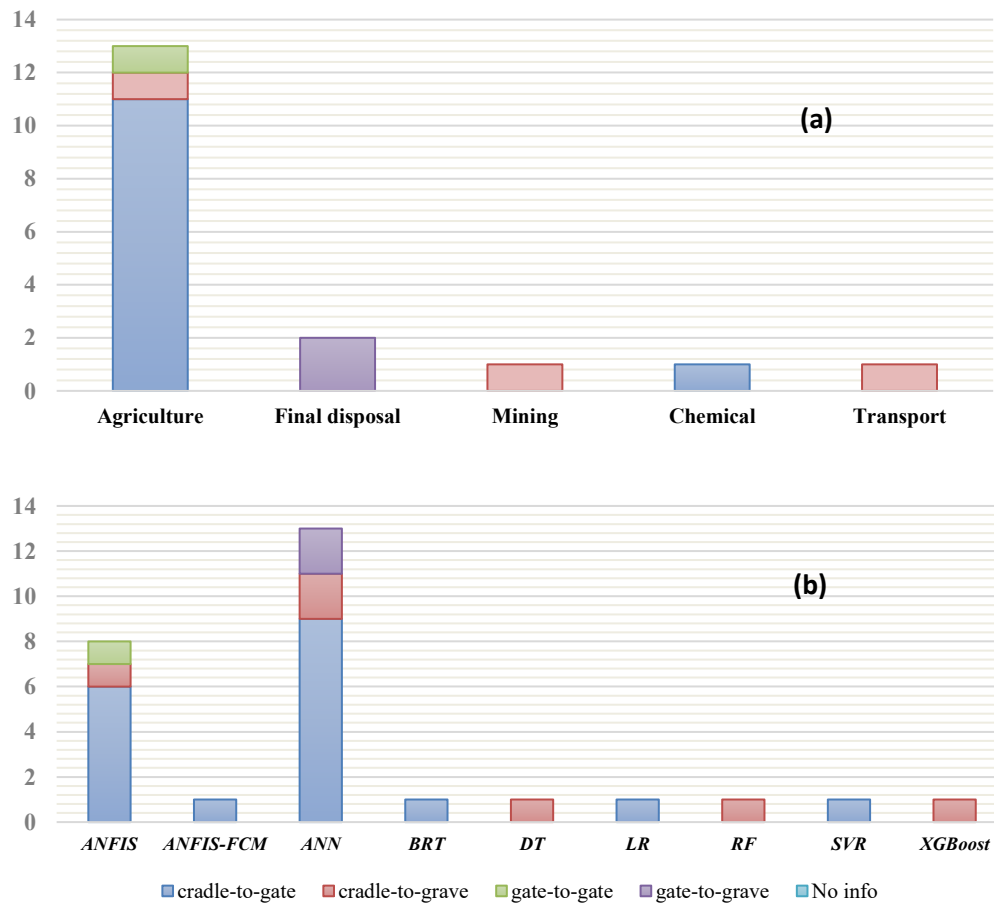
Source: own authorship.

Fig. S6 Analysis of the number of samples by study sector – in number of occurrences.



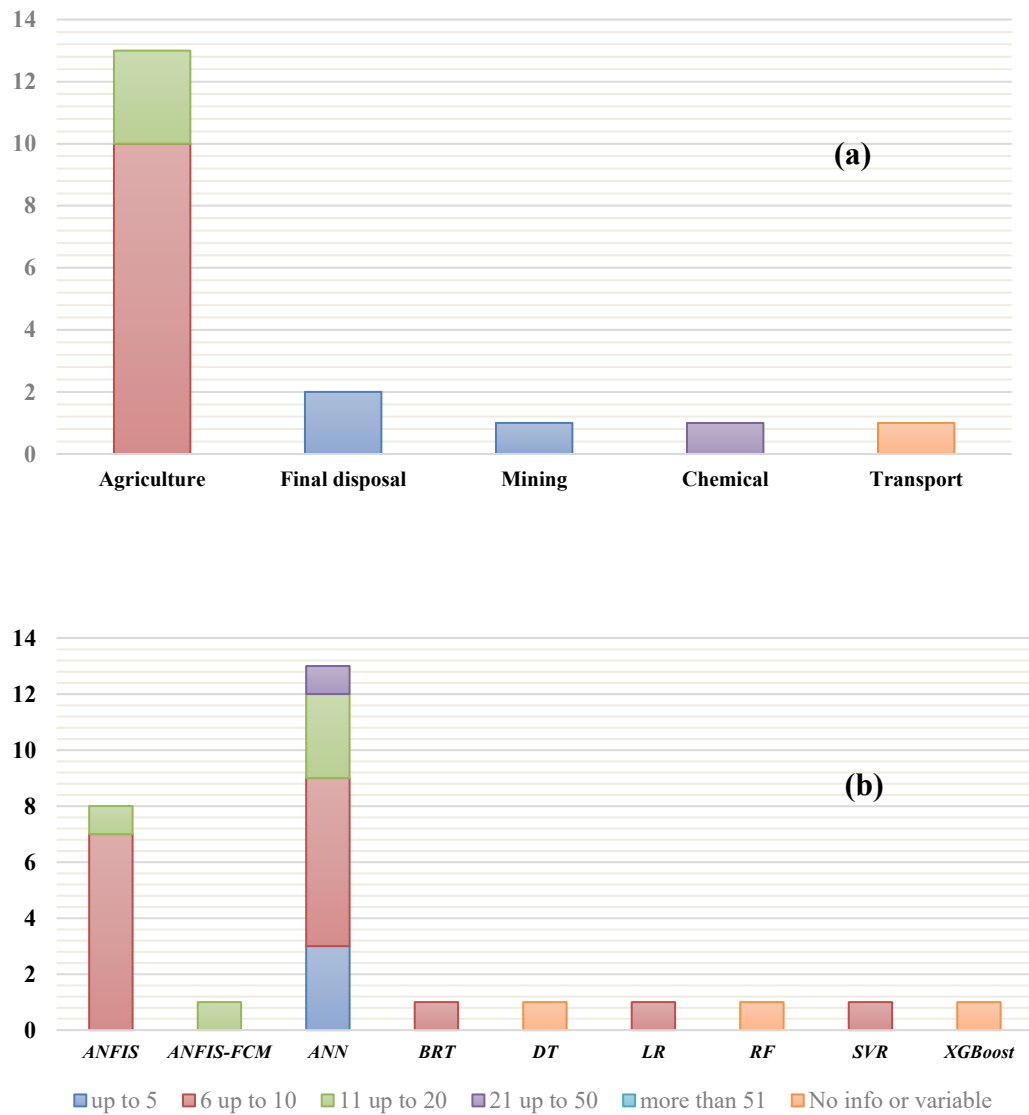
Source: own authorship.

Fig. S7 Arrangement of full LCA system boundaries by sector (a) and by ML method (b)
(own authorship) - in the number of occurrences



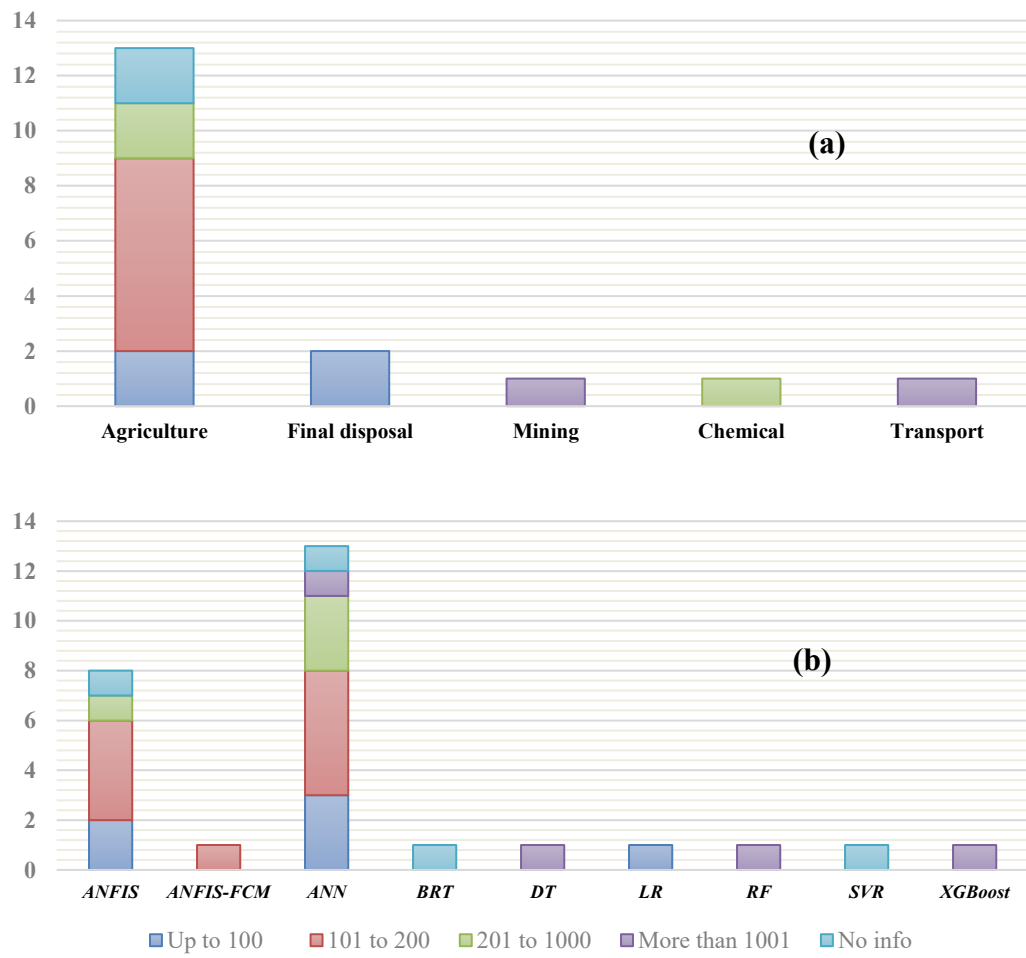
Source: own authorship.

Fig. S8 Analysis of the number of independent variables by sector (a) and by ML method for full LCA (own authorship) - in the number of occurrences



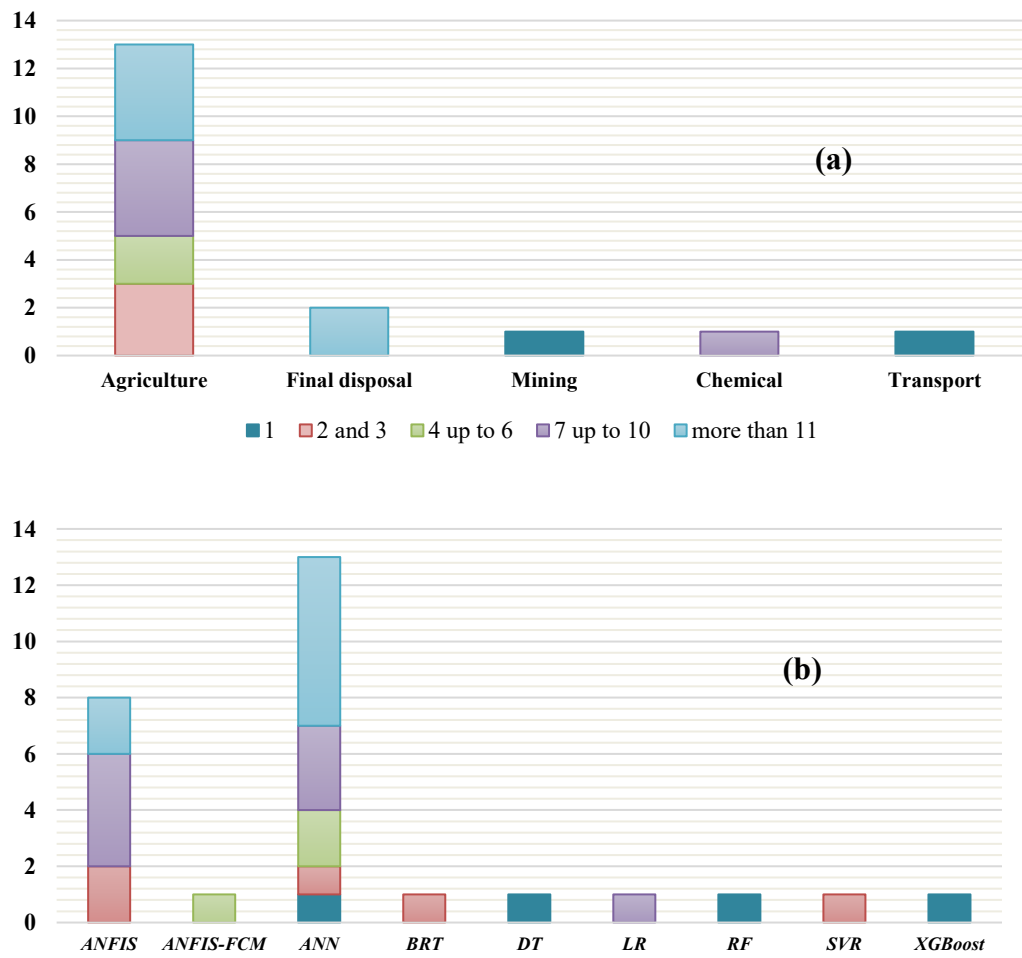
Source: own authorship.

Fig. S9 Analysis of the number of samples (a) per sector and (b) per machine learning method for full LCA (own authorship) – in the number of occurrences



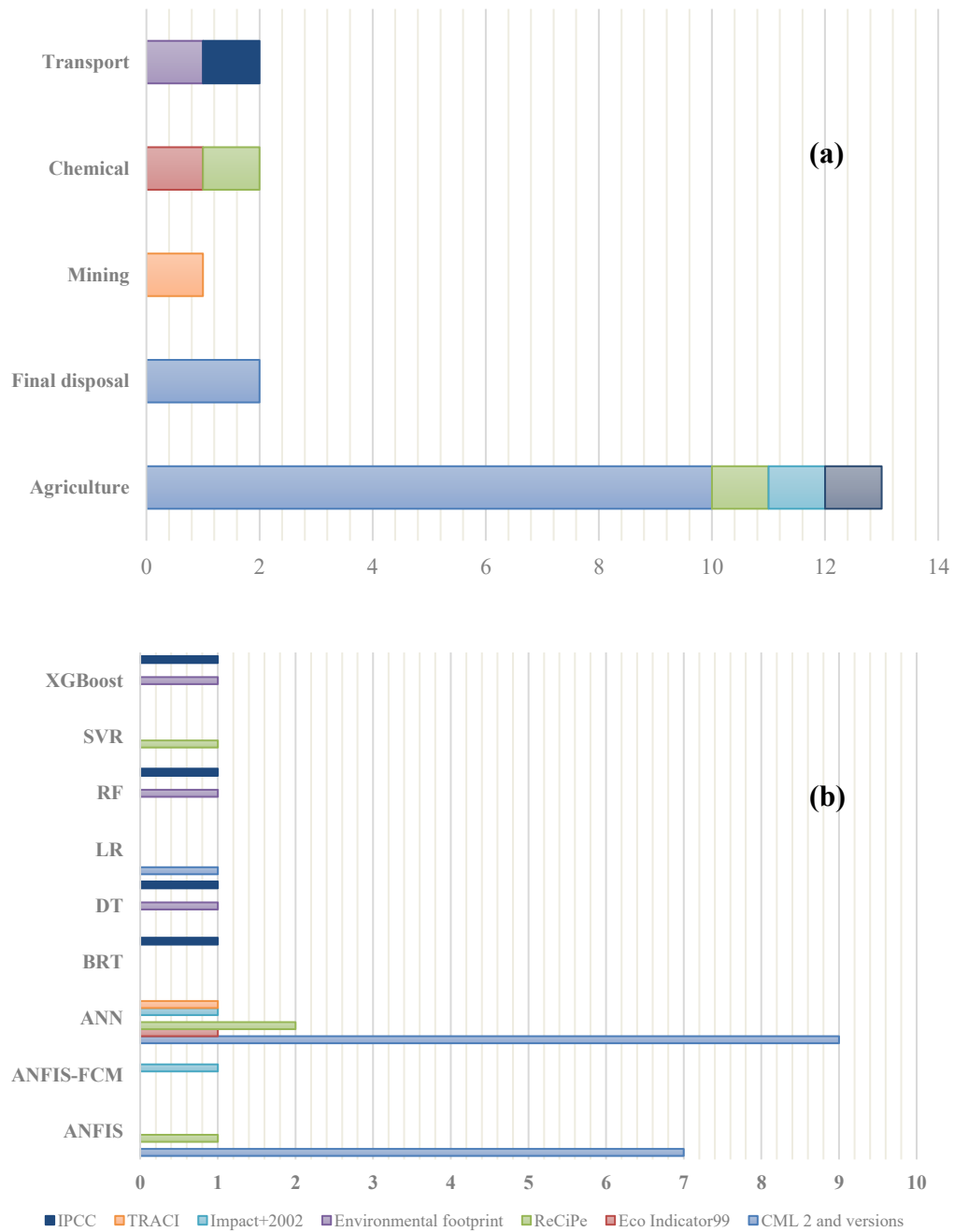
Source: own authorship.

Fig. S10 Analysis of the number of target attributes by sector (a) and by ML method (b) for full LCA (own authorship) – in the number of occurrences



Source: own authorship.

Fig. S11 Analysis of the use of Life Cycle Impact Assessment (LCIA) methods by sector (a) and by ML method (b) for full LCA (own authorship) - in the number of occurrences



Source: own authorship.

** It is important to highlight that there are studies that used more than one LCIA method; therefore, the methods were counted based on the number of occurrences rather than the number of studies.*

3. ARTIGO 2

Predição de impactos ambientais utilizando aprendizado de máquina: uma análise exploratória por diferentes abordagens

Resumo

A extração e o processamento de recursos materiais representam mais de 60% das emissões globais de gases de efeito estufa. As atividades florestais e a agricultura são grandes contribuintes dessas emissões, com destaque para o Brasil, que desempenha um papel significativo em ambos os setores. Este estudo visou explorar a integração da Avaliação do Ciclo de Vida (ACV) com Aprendizado de Máquina (AM) para prever impactos ambientais de maneira eficiente e rápida no setor do agronegócio nacional. A pesquisa enfrentou o desafio de lidar com o volume de dados e a complexidade nata da ACV, que requer uma coleta abrangente de informações ao longo do ciclo de vida dos produtos. Para resolver essa questão, foram utilizadas bases de dados nacionais e globais de ACV, tanto parametrizadas quanto não parametrizadas, para treinar modelos de AM. O processo foi dividido em três abordagens, que testaram a adoção de diferentes fontes de dados e suas respectivas integrações com o AM, indo desde dados de processos discretos (*gate-to-gate*) até inventários de culturas agrícolas completas (*cradle-to-farm gate*). A estrutura deste trabalho é composta por duas partes: 1) descrição do fluxo metodológico para se fazer a integração ACV+AM; e 2) o desenvolvimento de modelos de suporte à decisão orientados à integração ACV+AM. A primeira parte focou em apresentar os passos, técnicas e metodologias disponíveis que foram usadas no processo de geração e tratamento das amostras dos estudos de caso selecionados, enquanto a segunda aplicou os modelos de AM em diferentes bases de dados de ACV, buscando identificar as melhores soluções para previsões ambientais. Os resultados mostraram que aumentar o tamanho do conjunto de amostras melhora significativamente a capacidade preditiva dos modelos. Modelos baseados em *gradient boosting*, como XGBoost, GBM e CatBoost, alcançaram os melhores desempenhos, apresentando altos coeficientes de determinação (R^2 de 99,86%) e erros relativamente baixos (MSE de até 0,0032). O tempo de aprendizagem variou de 3h40 min a 12h48 min, e a predição dos impactos por cada modelo aprendido variou de 8,5s a 3147,8s, sendo os menores tempos associados ao modelo KNN (k-vizinhos mais próximos). De forma geral, processo de escolha do melhor *pipeline* de treinamento dos modelos depende do equilíbrio entre capacidade preditiva, generalização e custo computacional. Assim, o modelo de suporte à decisão criado foi capaz de estabelecer um passo a passo a ser utilizado para empregar conjuntos de dados de ACV no treinamento de modelos de AM a fim de realizar previsões futuras de impactos com alto nível de confiabilidade.

3.1 INTRODUÇÃO

A extração e o processamento de recursos materiais, incluindo combustíveis fósseis, minerais, não metálicos e biomassa, respondem por mais de 55% das emissões globais de gases de efeito estufa (GEE), e por 40% dos impactos na saúde decorrentes de partículas em suspensão (UNEP & IRP, 2024). Se a mudança no uso da terra for considerada, a biomassa se torna a maior contribuinte das emissões de GEE (28%), seguida pelos combustíveis fósseis (18%) e, depois, pelos não metálicos e metais (juntos 17%). Além disso, a biomassa proveniente de atividades agrícolas e florestais responde por mais de 90% da perda total de biodiversidade associada ao uso da terra e ao estresse hídrico (UNEP & IRP, 2024).

Nessas estatísticas, a produção de alimentos corresponde por um quarto das emissões globais de GEE (Our World in Data, 2024). O Brasil se destaca como um *player* global, produzindo alimentos suficientes para atender às necessidades internas e uma parte substancial da demanda internacional, contribuindo significativamente para a exportação de itens como soja, açúcar, café, suco de laranja, milho e fibra de algodão (Vargas et al., 2023)

A agricultura também desempenha um papel significativo no país, contribuindo com cerca de 30% do produto interno bruto (PIB) e sendo a segunda maior fonte de contribuição para as emissões de GEE no país, responsáveis por mais de 1,3 bilhões de toneladas de CO₂ emitidas em 2022 (Vargas et al., 2023).

A emissão de GEE anual per capita proveniente de mudanças no uso da terra, e vinculadas às práticas agrícolas e ao desmatamento, em 2022, foi destaque para o Brasil, totalizando 5,17 toneladas/pessoa*ano (Our World in Data, 2024). Para fins de comparação, é válido destacar que a média global foi de cerca de 0,54 toneladas/pessoa*ano, na China 0,69 toneladas/pessoa*ano, EUA 0,42 toneladas/pessoa*ano e na América do Sul, em média, 2,81 toneladas/pessoa*ano (Our World in Data, 2024).

Com efeitos potencialmente desestabilizadores na prosperidade a longo prazo para sociedade e no contexto da proteção ambiental, há uma necessidade de desenvolver ferramentas autônomas relevantes baseadas, por exemplo, em inteligência artificial (IA) para dar suporte ao desenvolvimento sustentável (Vinuesa et al., 2020). Esse tipo de solução poderia auxiliar na recuperação ecológica do planeta e auxiliar no processo de desacoplagem do desenvolvimento socioeconômico dos impactos ambientais adversos associados à atividade humana. Nesse sentido, destaca-se a importância de ferramentas e práticas quantitativas e que apoiem a transformação dos atuais sistemas produtivos, ou que ao menos permitam entender com certo

nível de confiabilidade os potenciais impactos das diferentes ações de um tomador de decisão de forma rápida.

Assim, alinhar estudos de Avaliação do Ciclo de Vida (ACV) com técnicas modernas de IA, sobretudo o aprendizado de máquina (AM), mostrou-se propícia para proporcionar velocidade de resposta requerida aos *Stakeholders* no:

- Contexto da manufatura aditiva: apoiando no “ecodesign” de produto-processo para itens de fabricação de filamento fundido (Naser et al., 2023);
- Indústria química: apoiando no processo de escolha de químicos ‘verdes’ substitutos para a produção de ativos farmacêuticos (Zhu et al., 2020);
- Construção civil: gerando *insights* para a definição do tipo de fundação mais vantajosa em termos de atributos de sustentabilidade e de desempenhos de manutenção, considerando para tal decisão a energia incorporada das lajes, a emissão de gases de efeito estufa, o custo do ciclo de vida entre outros (Teodosio et al., 2023).

A combinação/integração da IA com a ACV é recente, e no caso do Brasil é ainda embrionária (Salla et al., 2025). Portanto, trata-se de um tópico de pesquisa que ainda pode ser muito explorado acadêmica e empresarialmente. Sobre a avaliação do Ciclo de Vida (ACV), como o próprio nome sugere, envolve a avaliação dos impactos ambientais de um bem ou serviço dentro de um ciclo de vida (Oliveira et al., 2021). Devido ao seu escopo abrangente, estudos de ACV têm sido utilizados para apoiar o desenvolvimento de tecnologias, análises de políticas e até marketing verde para novos negócios (Romeiko et al., 2024), sendo uma ferramenta promissora para apoiar nesse processo de transição.

No entanto, por serem estudos com uma ampla gama de informações e um grande volume de dados, uma ACV deve seguir normas e diretrizes, exigindo especialistas para sua condução e um tempo considerável para sua conclusão (Oliveira et al., 2021). Notavelmente, a abrangência da ACV implica em uma extensa coleta de dados de diversas fontes ao longo da cadeia de suprimentos, tanto de processos à montante (Ex: extração de recursos, produção e transporte) quanto à jusante (Ex: uso e descarte do produto) (Romeiko et al., 2024). Nesse sentido, possuir dados com alta qualidade são essenciais para ACVs credíveis.

Estudos recentes com foco em revisão literária realizados por Romeiko et al. (2024), Martínez-Ramón et al. (2024) e Salla et al. (2025) permitiram estabelecer um panorama atual dessa integração ACV+AM:

- Romeiko et al. (2024): avaliaram 40 estudos com integração ACV+AM englobando as fases de ACV usadas, os algoritmos de AM usados, os *inputs*, *outputs* e fonte de dados usados. Tudo isso, considerando as possibilidades de integração nas etapas de inventário do ciclo de vida (ICV), avaliação do impacto do ciclo de vida (AICV) e interpretação do impacto do ciclo de vida.
- Martínez-Ramón et al. (2024): selecionaram 19 estudos para uma avaliação aprofundada com o objetivo de abrir caminho para futuros estudos baseados na implementação de AM para prever ICVs, especialmente, para processos ou sistemas que apresentam desafios em termos de complexidade ou carga computacional. Por fim, foram propostas três maneiras de permitir tal integração a partir da natureza dos dados.
- Salla et al. (2025): analisaram 59 estudos que de alguma forma utilizaram modelos de AM para a predição de potenciais impactos ambientais (fase de AICV). A partir da análise, foi proposta uma tabela-síntese que compila os principais modelos de AM usados e suas respectivas especificidades em termos de setor, fronteira do sistema, n° de amostras, atributos-alvo, variáveis independentes e outros requisitos. Permitindo, assim, consolidar o conhecimento existente e a fácil identificação de *gaps* a serem explorados em pesquisas futuras.

A predição de potenciais impactos ainda é incipiente na literatura, não havendo um procedimento delimitado que possa ser utilizado para tal. Em virtude disso, o presente capítulo explora diferentes maneiras de fazer essa integração, a fim de identificar potenciais caminhos, benefícios e elencar as principais dificuldades encontradas. Para tal, foi desenvolvida uma proposta de modelo de suporte à decisão com base em aprendizado de máquina e treinado com conjuntos de dados de ACV. O intuito foi avançar no tema de previsão de impactos via ACV de forma rápida e confiável e de facilitar o processo de tomada de decisão aos stakeholders; seja nas empresas, na definição do insumo de menor impacto para um produto em desenvolvimento; seja para os consumidores, no momento de definirem a escolha por determinados produtos em detrimento do impacto ambiental.

3.2 MATERIAIS E MÉTODOS

Do ponto de vista do método, foram usadas três diferentes abordagens. A primeira avaliou a integração ACV+AM em *datasets* (conjuntos de dados) nacionais não-parametrizados e de acesso gratuito; a segunda avaliou *datasets* nacionais parametrizados e de acesso gratuito;

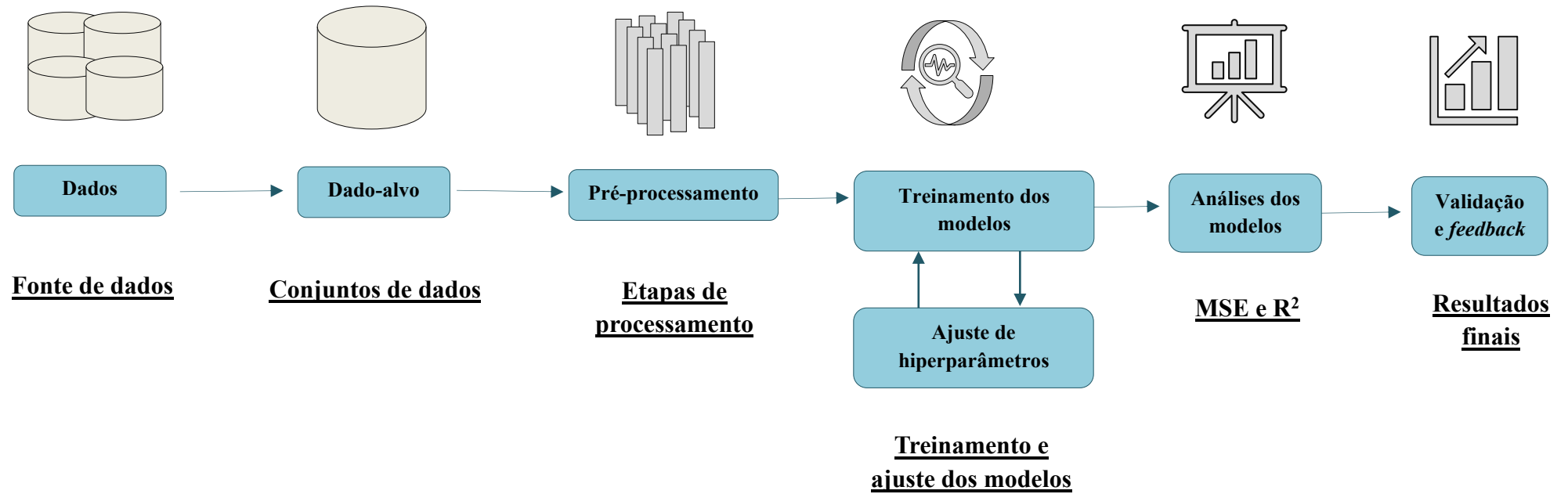
e a terceira empregou *datasets* globais não-parametrizados e sem a necessidade de serem de acesso gratuito.

Para a realização de cada um dos testes foi seguido um *pipeline* a fim de mapear o processo de obtenção e tratamento das amostras e, posteriormente, o treinamento e validação dos modelos de AM e possibilitar o estabelecimento de um fluxo metodológico que seja mais adequado para estudos dessa integração ACV+AM. Para isso, foram seguidas seis etapas (Figura 3.1), sendo:

- i. Dados: definição das fontes de dados (bases de dados de ACV) utilizadas em cada um dos testes. Bases de dados de ACV são conjuntos de dados organizados com a finalidade de gerar informações por meio de conexões entre diferentes conjuntos (Rodrigues et al., 2021). Bancos de dados de ACV variam globalmente, geralmente representam sistemas de produção nacionais ou até mesmo englobam fluxos de outros países/níveis (*Global, Rest-of-world, Europe-RER*, etc) (Rodrigues et al., 2021).
- ii. Dados-alvo: escolha dos *datasets* (conjuntos de dados) usados dentro de cada uma das bases de dados. Um conjunto de dados representa os metadados (informações descritivas, qualitativas sobre o escopo de cada inventário) e os dados quantitativos, i.e., os fluxos de entrada/saída do respectivo inventário (Rodrigues et al., 2021);
- iii. Pré-processamento: o pré-processamento envolve um conjunto de atividades para a preparação, organização e estruturação dos dados que foram gerados (Faceli et al., 2022). Existem várias técnicas que podem ser utilizadas para tal, entre elas a limpeza de dados (atributos com dados faltantes e dados com ruído), a transformação de dados (envolvendo a normalização e a seleção de atributos) e a redução de dados (englobando a clusterização, redução de dimensionalidade, amostragem, entre outros) (Faceli et al., 2022; Gomes, 2019). Desse modo, parte dessas técnicas foi utilizada para o tratamento dos dados presentes em cada um dos testes;
- iv. Treinamento dos modelos e ajustes dos parâmetros: definição dos métodos de AM, bem como o ajuste dos parâmetros para potencializar os resultados. Sipper (2022) destaca que não se deve esperar ganhos consideráveis com o ajuste de hiperparâmetros; no entanto, pode haver alguns conjuntos de dados, especialmente, dados complexos em que o ajustes dos hiperparâmetros pode ajudar a reduzir o desempenho ruim dos algoritmos;

- v. Análise dos modelos: processo de análise dos resultados obtidos e da performance de cada um dos modelos de AM. A performance de cada modelo foi analisada com base em dois indicadores: R2 (coeficiente de determinação) e MSE (erro quadrático médio). Assim, possibilitou analisar os erros inerentes a cada um dos modelos (MSE), bem como observar a capacidade de ajuste dos modelos aos dados observados, medindo a qualidade da predição de cada modelo (coeficiente de determinação R2);
- vi. Validação e feedback: etapa de validação dos resultados obtidos e definição dos melhores modelos de AM. Para isso, a tabela-síntese de Salla et al. (2025) será adotada como referência para a conferência e atribuição dos modelos mais bem pontuados. Nesse sentido, foi realizada uma validação cruzada entre os resultados obtidos na tabela-síntese de Salla et al. (2025), na qual $ANN > RF > SVM > LR > XGBoost > DT > ANFIS > GBM > KNN > MLR > Bayesian$, e os resultados obtidos para cada um dos testes, a fim de verificar se a ordem dos melhores modelos de AM foi mantida no presente trabalho

Figura 3. 1 - Pipeline do processo da ACV+AM realizado neste trabalho.



Fonte: autoria própria.

Foi utilizado o OpenLCA (disponível para download em: <https://www.openlca.org/download-form/>), principal software de Avaliação do Ciclo de Vida de código aberto do mundo. O OpenLCA é um software gratuito, o que permite ganho de flexibilidade, evitando a preocupação com códigos do software e deixar o usuário limitado por um software de código fechado/ caixa preta (GreenDelta, 2024a). Prova disso, são as “*developer tools*” disponíveis aos usuários, na qual é possível fazer a integração com a linguagem de programação Python na própria plataforma; o uso de códigos computacionais pelo *prompt* do próprio computador através da funcionalidade IPC Server, além de buscas mais sofisticadas por meio da funcionalidade SQL (*Structured Query Language*).

O OpenLCA foi utilizado como interface para a utilização dos processos (conjuntos de dados) disponíveis nas bases de dados selecionadas: SICV Brasil (Banco Nacional de Inventários do Ciclo de Vida), *Agribalyse*, *Agri-footprint* e *ecoinvent*. Sendo assim, o mesmo foi fundamental para viabilização das duas primeiras etapas do pipeline, ou seja, no processo de geração e compilação das amostras de dados de ACV.

Sobre o método de Avaliação de impacto do ciclo de vida (AICV), foi empregado o método “*Ecoinvent - CML v4.8 2016*” para gerar os potenciais impactos para cada conjunto de dados no OpenLCA, visto que o método foi o mais recorrente nos estudos analisados por Salla et al. (2025) sobre a integração ACV+AM. O método em questão foi adicionado a partir da versão 3.8 da base de dados *ecoinvent* e conta com 11 categorias de impacto:

- **Acidificação (AC):** medida em kg SO₂-Eq;
- **Mudanças Climáticas (CC):** medida em kg CO₂-Eq;
- **Ecotoxicidade em Água Doce (EF):** medida em kg 1,4-DCB-Eq;
- **Ecotoxicidade Marinha (EM):** medida em kg 1,4-DCB-Eq;
- **Ecotoxicidade Terrestre (ET):** medida em kg 1,4-DCB-Eq;
- **Recursos Energéticos não-renováveis (ER):** medido em MJ;
- **Eutrofização (EU):** medida em Kg de PO₄-Eq;
- **Toxicidade Humana (HT):** medida em kg 1,4-DCB-Eq;
- **Recursos Materiais metais/minerais (MR):** medida em kg Sb-Eq;
- **Depleção da Camada de Ozônio (OD):** medida em kg CFC-Eq;
- **Formação de Oxidantes Fotoquímicos (PHO):** medida em kg ethylene-Eq.

As etapas de definição das bases de dados e dos inventários variaram de teste para teste, logo, as especificações de cada teste (passos metodológicos) encontram-se detalhadas na Seção 3.2.1. De forma similar, é válido destacar a etapa de pré-processamento dos dados, etapa mais importante e crucial de todo o *pipeline*, cujos resultados encontram-se detalhados na Seção 3.3.1.

Após a etapa de pré-processamento dos dados, foi realizado o treinamento dos modelos de AM e o ajuste dos parâmetros. O processo de definição dos métodos baseou-se no estudo de Salla et al. (2025), na qual 9 métodos foram escolhidos para serem incluídos nos testes: Bayesian regression, K-vizinhos mais próximos (KNN), regressão linear (LR), árvores de decisão (DT), florestas aleatórias (RF), Gradient Boosting Machine (GBM), Extreme Gradient Boosting (XGBoost), redes neurais artificiais (ANN) e Máquina de Vetores de Suporte (SVM) e adicionados quatro outros métodos de regressão (CatBoost, Elastic Net, Stochastic Gradient Descendent – SGD e Kernel Ridge Regression- KRR). No Apêndice 3A, é oferecido um breve detalhamento acerca do funcionamento de cada um dos métodos.

Por fim, foram realizadas as duas últimas etapas do *pipeline*. A penúltima possibilitou analisar os erros inerentes a cada um dos modelos, além de observar a capacidade de ajuste aos dados observados a partir dos indicadores: R^2 e MSE.

O coeficiente de determinação (R^2) indica a proporção da variância na variável dependente que é explicada pela variável independente, geralmente variando de 0 (nenhuma variância explicada) a 1 (variância completamente explicada) (Learn Statistics Easily, 2023).

O erro quadrático médio (MSE) é a média das diferenças ao quadrado entre os valores observados em um estudo estatístico e os valores previstos por um modelo. Assim, as observações têm a mesma probabilidade de serem maiores ou menores que os valores previstos, por isso deve-se elevar ao quadrado (Stewart, 2024).

Para calcular cada um dos indicadores foi utilizada a biblioteca “*sklearn.metrics*” através das funções “*mean_squared_error*” e “*r2_score*” através da própria interface com o *python* por meio do *google colab (colab notebooks)* no mesmo código responsável pelo processo de treinamento dos modelos.

Em seguida, foi realizada a validação dos modelos, a qual permitiu legitimar os resultados obtidos e selecionar os modelos de AM de melhor *performance*. Nesse sentido, foi utilizado os resultados obtidos por Salla et al. (2025), com objetivo de verificar se a *performance* dos modelos em cada teste (teste 1, teste 2 e teste 3) segue a estrutura hierárquica proposta pelo autor a partir dos achados na literatura, na qual: na qual ANN > RF > SVM > LR > XGBoost > DT > ANFIS > GBM > KNN > MLR > rede Bayesiana.

3.2.1 Abordagens metodológicas utilizadas

O subtópico em questão teve por objetivo concentrar e exemplificar os passos que foram seguidos em cada um dos três testes para gerar e compilar as amostras usadas para treinamento dos modelos de AM. De modo geral, as amostras foram definidas da seguinte forma:

- a) Teste 1: dados de inventário do ciclo de vida para a produção de milho no Brasil disponíveis no SICV Brasil. A cultura do milho foi selecionada porque é a segunda mais importante do país (Vittia, 2024) e a segunda mais produzida no mundo, ficando apenas atrás da cana-de-açúcar (CropLife, 2020). Além disso, o milho é uma cultura muito versátil, servindo para alimentação, produção de ração animal e até para a produção de biocombustível (CropLife, 2020).
- b) Teste 2: dados de inventário do ciclo de vida parametrizados para processos de usinagem disponíveis no SICV Brasil. Foram selecionados os processos de usinagem da base porque os mesmos foram modelados seguindo a metodologia UPLCI (*unit process life-cycle inventory*) proposta por Kellens et al., (2012a, 2012b). Logo, as entradas e saídas disponíveis nos dois conjuntos de dados foram parametrizadas em função do tempo de operação de cada um dos modos das máquinas, o que possibilitou avaliar apenas os potenciais impactos dentro da indústria (*gate-to-gate*).
- c) Teste 3: dados de inventário do ciclo de vida não-parametrizados para processos de produção agrícola nacionais e internacionais disponíveis no *ecoinvent*, *Agribalyse* e *Agri-footprint*. Ao todo 1.219 *datasets* de diferentes culturas foram selecionados a partir dessas três bases escolhidas.

3.2.1.1 Testes 1 e 2

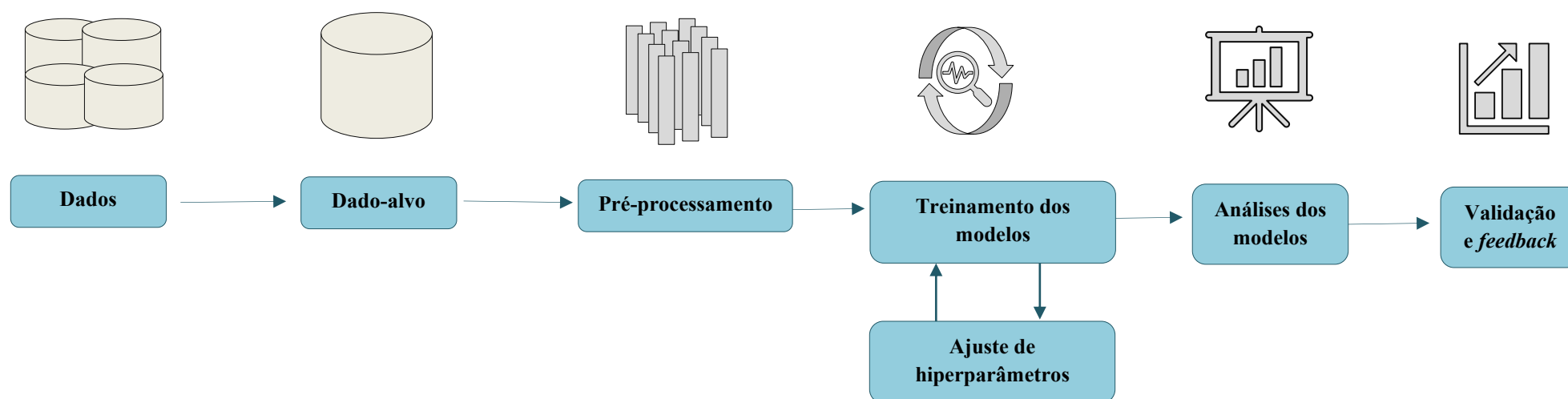
Os dados para os Testes 1 e 2 foram retirados do Banco Nacional de Inventários do Ciclo de Vida de Produtos Brasileiros - SICV Brasil (<https://sicv.acv.ibict.br/>), base de dados brasileira desenvolvida pelo IBICT (Instituto Brasileiro de Informação em Ciência e Tecnologia) e que funciona sob a perspectiva da ciência aberta, oferecendo acesso gratuito a centenas de inventários de produtos da economia brasileira, promovendo transparência ambiental e apoiando decisões sustentáveis (IBICT, 2024).

Para permitir tal integração, os conjuntos de dados foram obtidos no SICV Brasil no formato ILCD (*International Reference Life Cycle Data System*) e importados ao OpenLCA. O formato em questão é XML, sendo desenvolvido, inicialmente, para a base de dados europeia

de ICV, a *European Reference Life Cycle Database* (ELCD) e de caráter aberto, assim como, os princípios defendidos pelo SICV Brasil (IBICT, 2024).

A condução dos Testes 1 e 2 seguiu o *pipeline* sugerido na Figura 3.1, havendo apenas algumas adaptações de modo a representar o que de fato foi feito em cada um dos dois testes (Figura 3.2). A condução dos seguintes testes e os respectivos resultados encontram-se detalhados nas Seções 3.3.2.1 (para o Teste 1) e 3.3.2.2 (para o Teste 2).

Figura 3. 2 - Pipeline de treinamento dos modelos de AM para os Testes 1 e 2.



Teste 1	SICV Brasil	Dados nacionais não-parametrizados para o milho (15 amostras).	Compilação das amostras, normalização dos dados, remoção de <i>outliers</i> e separação dos dados para treino e teste.	Modelos: regressão bayesiana, KNN, LR, DT, RF, GBM, XGBoost, ANN, SVM, Catboost, E_net, SGD e KRR.	Coeficiente de determinação (R ²) e erro quadrático médio (MSE).	Validação dos resultados obtidos e definição dos melhores modelos.
Teste 2	SICV Brasil	Dados nacionais parametrizados de usinagem (200 amostras).		Hiperparâmetros: pré-definição de um <i>range</i> de parâmetros.		

*Legenda: KNN: K - Nearest Neighbors; LR: linear regression; DT: decision tree; RF: random forest; GBM: Gradient Boosting Machine; XGBoost: eXtreme Boosting Machine; ANN: artificial neural network; SVM: Support vector machine; E_net: ElasticNet; CatBoost: Category Boosting; SGD: Stochastic gradient descent e KRR: kernel ridge regression.

Fonte: autoria própria.

3.2.1.2 Testes 3

Para o terceiro teste, foram utilizados conjuntos de dados de diferentes culturas agrícolas para o *pipeline* de treinamento dos modelos de AM na predição de potenciais impactos (Figura 3.3). Assim, foram empregadas três das bases de dados mais renomadas e utilizadas para essa natureza: *Ecoinvent* (Wernet et al., 2016), *Agribalyse* (Colomb et al., 2015) e *Agri-footprint* (Durlinger et al., 2014) alinhada às funcionalidades disponíveis no *open-source* do próprio OpenLCA para a geração e compilação de amostras (contendo os ICVs e seus respectivos potenciais impactos).

a) **Base de dados:**

O *ecoinvent* é o banco de dados de Inventário do Ciclo de Vida (ICV) líder em nível mundial, possuindo mais de 17.000 *datasets* que são apresentados de forma consistente e mantendo a transparência necessária (Hillege, 2024). O *Ecoinvent* é conhecido por ser uma fonte de dados de dados “multiáreas”, ou seja, seus conjuntos de dados cobrem uma ampla gama de produtos, serviços e processos, desde materiais de construção até alimentos, e desde a extração de recursos até a gestão de resíduos (Hillege, 2024), isso tudo em nível global.

Agribalyse é a base de dados pública francesa mais exaustiva em termos de indicadores ambientais para produtos agrícolas e alimentares baseados na ACV, sendo uma referência nacional francesa aos potenciais impactos ambientais dos produtos agrícolas e alimentares (Agribalyse, 2024). A base contém dados sobre mais de 200 produções agrícolas e mais de 2.500 alimentos prontos para consumo, incluindo desde dados sobre ICVs agrícolas (Ex: maçã) até dados de produtos alimentícios já processados (ex: geleia de maçã). O foco da base de dados são os itens produzidos na França, contudo, a mesma ainda inclui produtos alimentares produzidos no mercado estrangeiro e que são importados (Agribalyse, 2024).

O *Agrifootprint* é um banco de dados abrangente de ICVs, incluindo uma gama de produtos agrícolas, tais como: rações, alimentos e biomassa, sendo um repositório de abrangência global. A base de dados contém aproximadamente 4.800 produtos e processos específicos para itens de natureza agrícola, entre eles: culturas, produtos, rações compostas, produtos alimentícios, sistemas de produção animal, transporte, pesticidas, fertilizantes, entre outros (Simapro, 2024).

b) **Processo de geração de amostras:**

Para fins de comparação, também foi utilizado o software OpenLCA, de modo a permitir a visualização dos conjuntos de dados, a geração dos sistemas de produto e o cálculo dos

potenciais impactos ambientais de acordo com o método CML. Para tal, foi utilizada a funcionalidade “*developer tools*”, a qual garante ferramentas de caráter de desenvolvedor aos usuários.

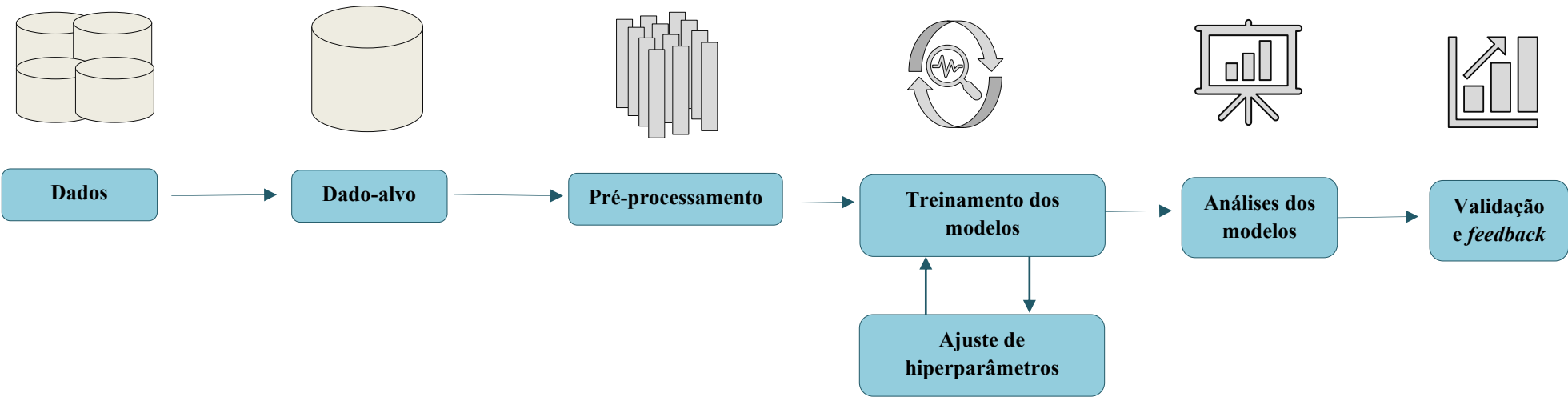
Desde a versão 1.4.1, o OpenLCA instituiu o campo “*developer tools*”, o qual oferece suporte à execução de programas em Python e JavaScript, além de consultas SQL dentro da interface do próprio software (GreenDelta, 2024b). Com a implementação de tal recurso, é possível o usuário automatizar cálculos no OpenLCA, seja criando importadores ou exportadores de dados ou até mesmo fazendo cálculos de análise de sensibilidade variando parâmetros de forma mais simples e automatizada (GreenDelta, 2024b).

Dentro do campo “*developer tools*” na versão 2.0.3 do OpenLCA, foi utilizada a funcionalidade IPC Server. O Servidor IPC no OpenLCA é uma interface de troca de dados independente de plataforma, que permite que serviços do OpenLCA sejam executados através da biblioteca padrão do Python. O OpenLCA fornece uma implementação de um protocolo baseado em JSON-RPC para comunicação entre processos (IPC). Com isso, é possível chamar funções no OpenLCA e processar seus resultados fora do software por meio de ambientes de programação (Pypi, 2019).

Assim, foi utilizado como prompt o *jupyter notebook* por meio da interface do Anaconda (<https://anaconda.org/>) para a implementação dos códigos em Python. Além disso, foram utilizadas duas bibliotecas já consolidadas e com interface com o OpenLCA, a biblioteca *olca-ipc* 2.0.3 (<https://pypi.org/project/olca-ipc/>) e a biblioteca *olca-schema* 2.2.1 (<https://pypi.org/project/olca-schema/>).

Com isso, utilizando funções já consolidadas nas mesmas e incluindo algumas adaptações é possível automatizar o processo tedioso de geração de amostras. Assim, o código Python utilizado foi capaz de selecionar os conjuntos de dados a serem exportados, gerar os sistemas de produto para cada conjunto de dados, rodar os respectivos sistemas de produto e exportar os resultados obtidos de forma automatizada.

Figura 3. 3 - Pipeline de treinamento dos modelos de AM para o Teste 3.



Teste 3	Ecoinvent, Agribalyse e Agrifootprint.	Dados internacionais não-parametrizados para os grãos: cevada, milho, canola, arroz, soja, girassol e trigo (1.219 amostras).	Compilação das amostras, redução de dimensionalidade (6 bases de dados diferentes geradas), normalização dos dados, remoção de <i>outliers</i> e separação dos dados para treino e teste.	Modelos: regressão bayesiana, KNN, LR, DT, RF, GBM, XGBoost, ANN, SVM, Catboost, E_net, SGD e KRR. Hiperparâmetros: pré-definição de um <i>range</i> de parâmetros.	Coefficiente de determinação (R2) e erro quadrático médio (MSE).	Validação dos resultados obtidos e definição dos melhores modelos.
---------	--	---	---	--	--	--

**Legenda: KNN: K - Nearest Neighbors; LR: linear regression; DT: decision tree; RF: random forest; GBM: Gradient Boosting Machine; XGBoost: eXtreme Boosting Machine; ANN: artificial neural network; SVM: Support vector machine; E_net: ElasticNet; CatBoost: Category Boosting; SGD: Stochastic gradient descent e KRR: kernel ridge regression.*

Fonte: autoria própria.

3.3 RESULTADOS E DISCUSSÃO

A Seção referente aos resultados e discussões foi subdividida em duas subseções, sendo a primeira para descrever de forma qualitativa os principais passos utilizados em cada um dos três testes e a segunda, por sua vez, para apresentar de forma quantitativa os resultados de treinamento de cada um dos modelos de AM em cada um dos testes.

3.3.1 Descrição dos resultados obtidos com os fluxos metodológicos

A subseção em questão teve por objetivo exemplificar os principais resultados dos passos metodológicos que foram seguidos em cada um dos três testes, salientando os principais desafios encontrados e descrevendo as etapas executadas para a compilação das bases de dados (conjunto de amostras totais) usadas em cada teste.

3.3.1.1 Resultados do Teste 1

Após a importação dos dados do SICV Brasil ao OpenLCA, foram gerados os respectivos sistemas de produto para os inventários de milho (*maize grain* do tipo “*first crop*” e “*second crop*”) das diferentes regiões brasileiras, conforme Tabela 3.1:

Tabela 3. 1 - Inventários de milho utilizados do SICV Brasil.

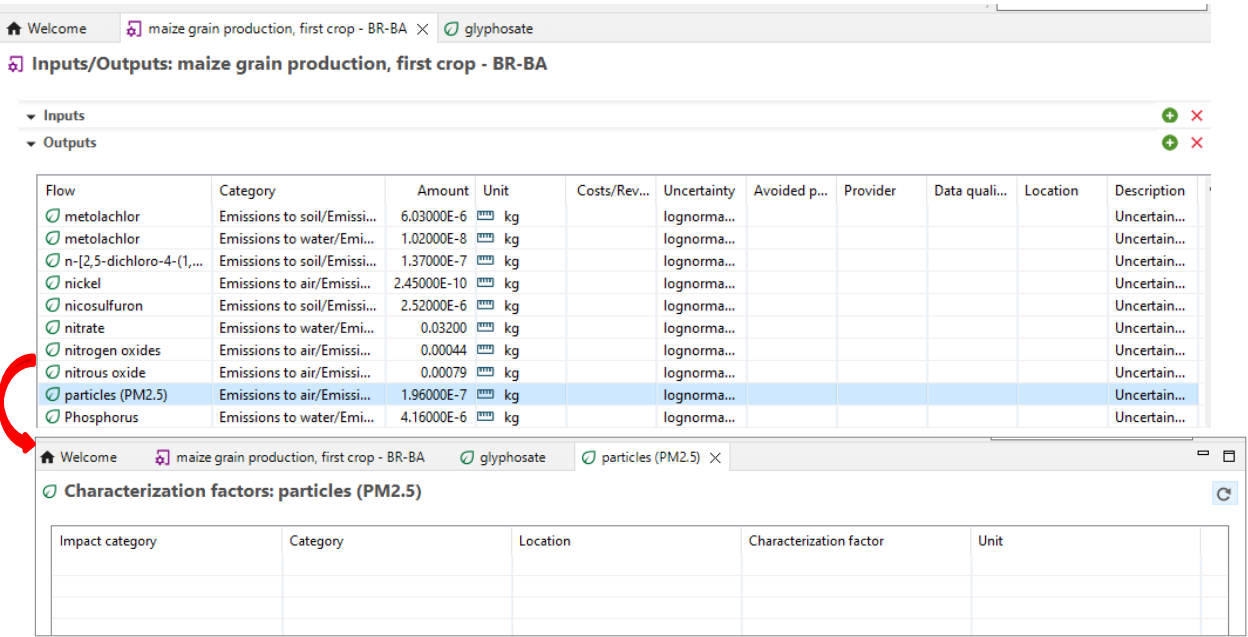
Nº	Nome do dataset	Nº	Nome do dataset
11	maize grain production, first crop-BR-BA	19	maize grain production, second crop-BR-MA
12	maize grain production, first crop-BR-GO	20	maize grain production, second crop-BR-MG
13	maize grain production, first crop-BR-MA	21	maize grain production, second crop-BR-MS
14	maize grain production, first crop-BR-MG	22	maize grain production, second crop-BR-MT
15	maize grain production, first crop-BR-PI	23	maize grain production, second crop-BR-PR
16	maize grain production, first crop-BR-PR	24	maize grain production, second crop-BR-SP
17	maize grain production, first crop-BR-RS	25	maize grain production, second crop-BR-TO
18	maize grain production, second crop-BR-GO		

Fonte: autoria própria.

Durante essa etapa, ocorreu um erro ao gerar os resultados para os sistemas de produto, visto que os fluxos elementares disponíveis nos processos do SICV Brasil estavam com fatores de caracterização zerados (Figura 3.4) e os fluxos de produto (Figura 3.5) não conseguiam ligar com os respectivos *providers* (provedores de dado), logo, não estava gerando nenhuma contribuição para as categorias de impacto ao rodar o sistema de produto para o método de

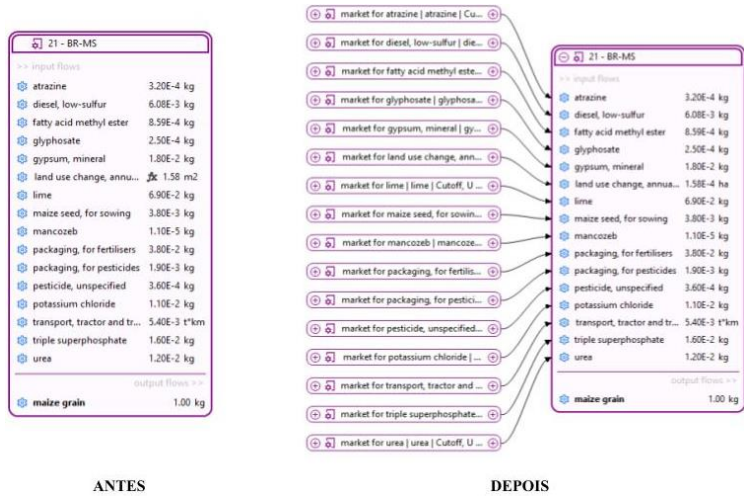
avaliação do impacto do ciclo de vida (AICV) escolhido. Essa ocorrência vai em direção ao observado por Qi (2023), na qual a importação da base Tiangong ao software GaBi também apresentou constantes problemas, provavelmente devido à inexistência de fluxos elementares semelhantes ou até mesmo pela diferença no formato dos arquivos (ILCD). De forma similar, Silva et al (2019) identificou que diferentes softwares de ACV implementam fatores de caracterização de forma distinta, afetando nos resultados e gerando problema com o intercâmbio de arquivos.

Figura 3. 4 - Fatores de caracterização zerados.



Fonte: autoria própria (OpenLCA).

Figura 3. 5 - Sistema de produto sem a conexão com os providers.



Fonte: autoria própria (OpenLCA).

Para contornar tal problema, foram usados fluxos do Ecoinvent. Nesse sentido, os fluxos de entrada (fluxos de produto e elementares) e saída (fluxos elementares – cor verde) foram substituídos em cada um dos 15 *datasets*, de modo a garantir a geração de potenciais impactos. Em seguida, foi realizada a etapa de avaliação dos atributos juntamente com o pré-processamento dos dados. Isso tudo, com o objetivo de diminuir a quantidade de fatores a serem informados pelo usuário (tomador de decisão), 27 atributos até então (Tabela 3.2), e ir em consonância aos achados obtidos por Salla et al. (2025), na qual os estudos de ACV+AM com foco agrícola possuem em sua maioria até 15 atributos para treinamento dos modelos de AM.

Tabela 3. 2 - Tabela com os fluxos contidos nos *datasets* (atributos).

Nome dos fluxos	Nome dos fluxos
maize grain 2,4-dichlorophenol	
arable, non-irrigated, intensive	maize seed, for sowing
atrazine	mancozeb
carbon dioxide (biogenic)	metolachlor
diammonium phosphate	monoammonium phosphate
diesel, low-sulfur	packaging, for fertilisers
dolomite	packaging, for pesticides
drying of maize grain	pesticide, unspecified
fatty acid methyl ester	potassium chloride
from arable, non-irrigated, intensive	to arable, non-irrigated, intensive
glyphosate	transport, tractor and trailer, agricultural
gypsum, mineral	triple superphosphate
land use change, annual crop	urea
lime	water

Fonte: autoria própria.

Análise dos atributos e pré-processamento dos dados (Teste 1)

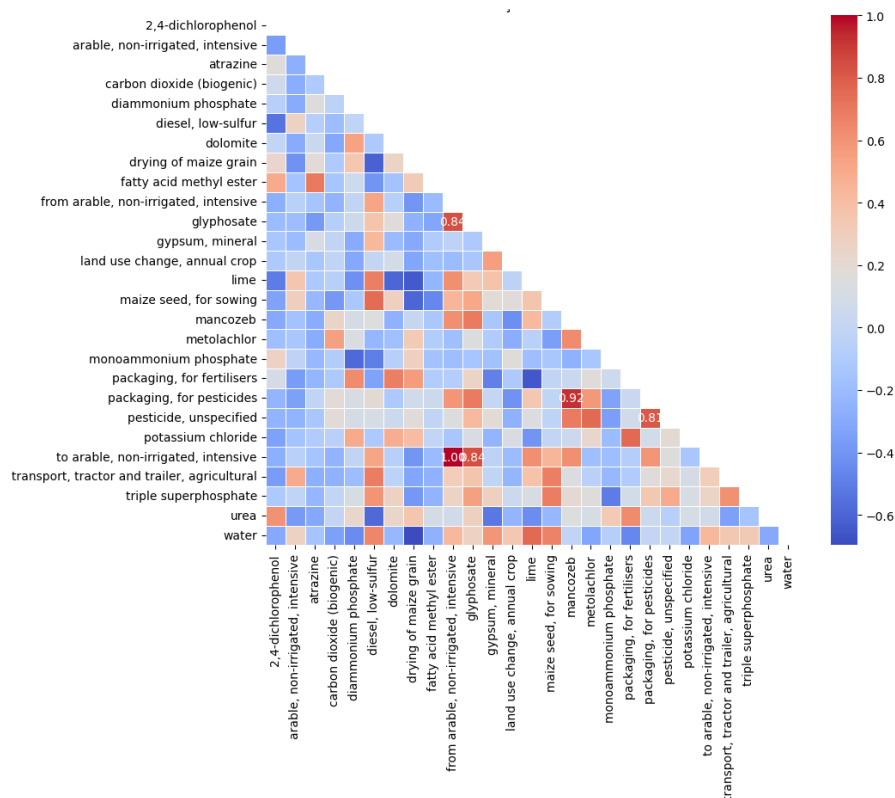
Inicialmente, foi efetuada uma análise dos 27 atributos que compõem as amostras. Para tal, optou-se por utilizar a matriz correlação, visto que a mesma permite estabelecer relações de proximidade entre as variáveis (Whitfield, 2024). Na matriz de correlação, cada célula da matriz representa o coeficiente de correlação entre duas variáveis, e esse valor indica o tipo e a força

da relação entre elas, variando de -1 a +1 e assumindo as seguintes considerações (Whitfield, 2024):

- -1 a 0: correlação negativa. Logo, quando uma variável aumenta a outra diminui de forma proporcional e quanto mais próximo de -1 mais forte a relação negativa entre elas.
- 0: sem relação linear entre as variáveis.
- 0 a +1: correlação positiva. Logo, quando uma variável aumenta a outra aumenta de forma proporcional e quanto mais próximo de 1 mais forte a relação entre elas.

Em processos de aprendizado de máquina, a matriz de correlação ajuda a identificar quais variáveis estão fortemente correlacionadas, permitindo uma seleção de variáveis mais eficiente e reduzindo a dimensionalidade, ou seja, o número de atributos (Murel & Kavlakoglu, 2025). Para os testes em questão, foi feita tal redução utilizando os valores estabelecidos na matriz de correlação, conforme Figura 3.6.

Figura 3. 6 - Matriz de correlação dos atributos de entrada.



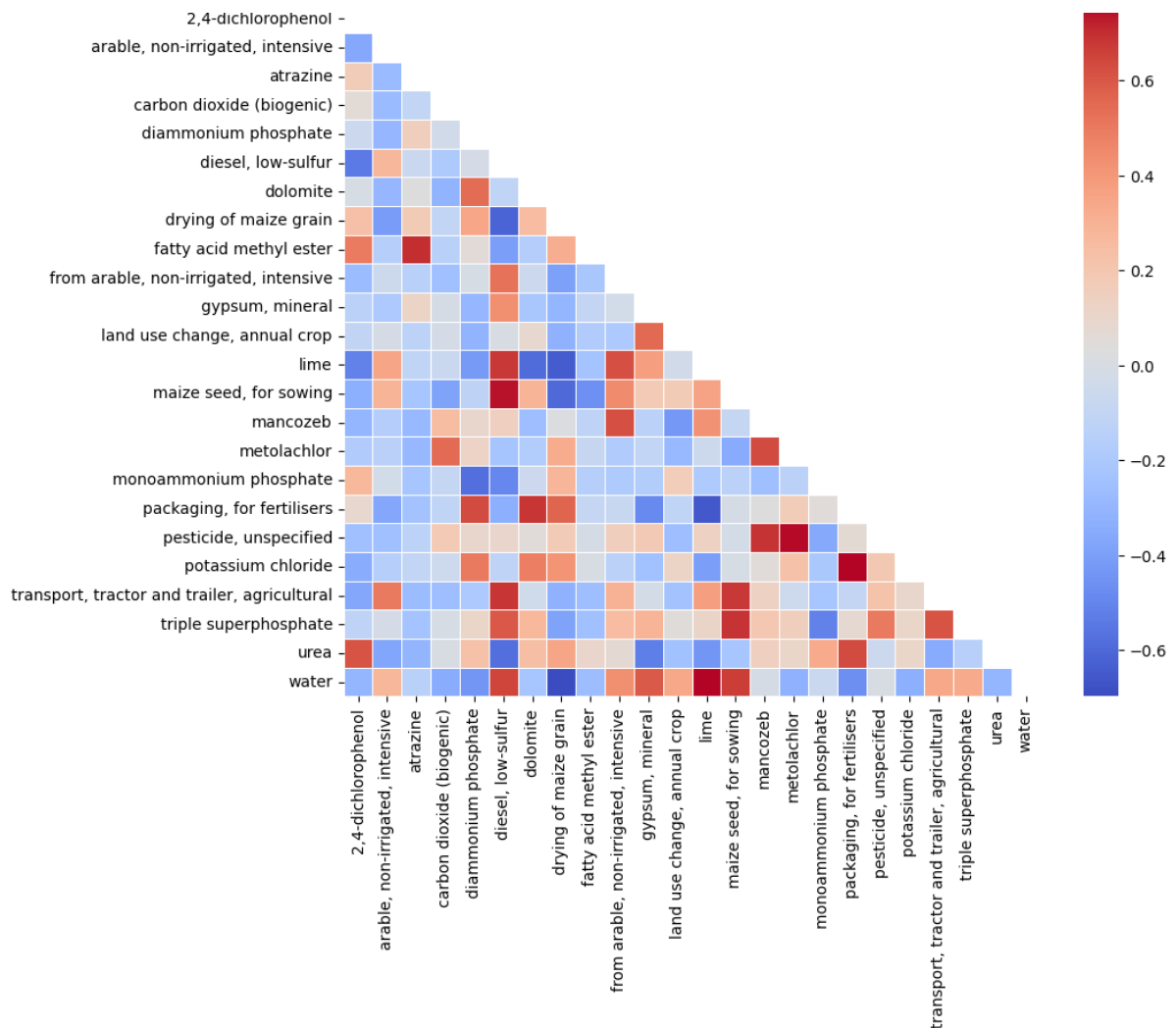
Fonte: autoria própria.

A partir da análise da matriz de correlação dos atributos, foram observados cinco cenários em que a relação entre os atributos é superior a 0,7, indicando uma correlação linear positiva relativamente forte entre eles (Lopes, 2016). Nesse sentido, foram feitos os seguintes ajustes ao conjunto de atributos:

- Remoção do atributo “glyphosate” visto que o mesmo apresenta uma elevada relação linear positiva com os atributos “to arable, non-irrigated, intensive” (0,84) e “from arable, non-irrigated, intensive” (0,84). Logo, sua remoção, não trará perdas significativas ao aprendizado dos modelos.
- Remoção do atributo “packaging, for pesticides” por possuir elevada correlação linear positiva com os atributos “mancozeb” (0,92) e “pesticide, unspecified” (0,81). Logo, sua remoção, não trará perdas significativas ao aprendizado dos modelos.
- Remoção do atributo “to arable, non-irrigated, intensive” visto que o mesmo apresenta uma elevada relação linear totalmente positiva com o atributo “from arable, non-irrigated, intensive” (1,00). Logo, sua remoção, não trará perdas ao aprendizado dos modelos.

Após essas remoções, os dados utilizados para treinar os modelos foram sintetizados em 24 atributos, com coeficiente de correlação máximo de cerca de 0,68 entre os atributos e negativo de cerca de -0,65, conforme novos resultados gerados na Figura 3.7.

Figura 3. 7 - Matriz de correlação dos atributos de entrada após as remoções.



Fonte: autoria própria.

Após a seleção dos atributos mais representativos, foi realizada a análise descritiva dos dados, a fim de identificar o comportamento dos atributos-alvo (Tabela 3.3), ou seja, o comportamento das categorias com impactos potenciais relativo ao cultivo de 1kg de milho, e permitir a identificação de potenciais *outliers* no processo. Para a identificação de *outliers*, foi utilizado o método IQR (*Interquartile Range*). O método calcula o intervalo interquartil (IQR) como a diferença entre o terceiro quartil (Q3), valor abaixo do qual estão 75% dos dados e o primeiro quartil (Q1), valor abaixo do qual estão 25% dos dados (Vinutha et al., 2018).

Com base nesse intervalo, um valor é considerado um *outlier* ou ponto fora da curva se estiver abaixo de Q1 menos 1,5 vezes o IQR ou acima de Q3 mais 1,5 vezes o IQR. Esse critério é amplamente utilizado para identificar valores atípicos em conjuntos de dados, pois ajuda a

detectar pontos que desviam significativamente da distribuição principal (Pastega, 2023) e que podem atrapalhar a taxa de aprendizado dos modelos de AM (Faceli et al, 2022).

Tabela 3. 3 - Análise descritiva dos atributos-alvo para o teste 1.

Atributos-alvo	Unidade	Valor mínimo	Valor máximo	Média	Desvio padrão (absoluto)	Outliers identificados
AC	kg SO ₂ -Eq	2,41 x 10 ⁻⁰³	3,12 x 10 ⁻⁰²	7,98 x 10 ⁻⁰³	6,86 x 10 ⁻⁰³	2
CC	kg CO ₂ -Eq	2,75 x 10 ⁻⁰¹	2,69 x 10 ⁺⁰⁰	8,75 x 10 ⁻⁰¹	6,97 x 10 ⁻⁰¹	1
EF	kg 1,4- CB-Eq	1,17 x 10 ⁻⁰¹	3,32 x 10 ⁺⁰⁰	8,19 x 10 ⁻⁰¹	1,09 x 10 ⁺⁰⁰	2
EM	kg 1,4-DCB-Eq	9,88 x 10 ⁺⁰¹	3,84 x 10 ⁺⁰²	2,13 x 10 ⁺⁰²	9,09 x 10 ⁺⁰¹	0
ET	kg 1,4-DCB-Eq	-8,25 x 10 ⁻⁰²	6,89 x 10 ⁻⁰²	1,04 x 10 ⁻⁰²	3,46 x 10 ⁻⁰²	3
ER	MJ	1,22 x 10 ⁺⁰⁰	3,08 x 10 ⁺⁰⁰	2,16 x 10 ⁺⁰⁰	5,71 x 10 ⁻⁰¹	0
EU	kg PO ₄ -Eq	5,27 x 10 ⁻⁰³	1,73 x 10 ⁻⁰²	8,89 x 10 ⁻⁰³	3,88 x 10 ⁻⁰³	0
HT	kg 1,4-DCB-Eq	1,82 x 10 ⁻⁰¹	5,85 x 10 ⁺⁰⁰	1,40 x 10 ⁺⁰⁰	1,77 x 10 ⁺⁰⁰	1
MR	kg Sb-Eq	1,12 x 10 ⁻⁰⁶	5,12 x 10 ⁻⁰⁶	2,99 x 10 ⁻⁰⁶	1,38 x 10 ⁻⁰⁶	0
OD	kg CFC-11-Eq	4,00 x 10 ⁻⁰⁹	1,50 x 10 ⁻⁰⁸	7,50 x 10 ⁻⁰⁹	2,93 x 10 ⁻⁰⁹	1
PO	kg ethylene-Eq	4,05 x 10 ⁻⁰⁵	2,96 x 10 ⁻⁰³	6,32 x 10 ⁻⁰⁴	8,99 x 10 ⁻⁰⁴	1

* *Legendas: AC: acidification; CC: climate change; EF: ecotoxicity: freshwater; EM: ecotoxicity: marine; ET: ecotoxicity: terrestrial; ER: energy resources; EU: eutrophication; HT: human toxicity; OD: ozone depletion; PO: photochemical oxidant formation.*

Fonte: autoria própria.

A partir da tabela acima, notou-se a existência de outliers em pelo menos 7 das 11 categorias de impactos analisadas. Nesse sentido, as respectivas amostras com outliers foram suprimidas do pipeline de treinamento dos modelos de AM nos respectivos testes. Por exemplo, para o pipeline de treinamento para a categoria de acidificação foram utilizadas apenas 13 amostras, visto que duas delas foram removidas do campo amostral por serem *outliers*, conforme os dados da Tabela 3.3. Assim, o pipeline de treinamento, teste e validação para cada um dos 11 atributos-alvo (categorias de impacto) contou com número variável de amostras a depender da quantidade de *outliers* existentes por categoria de impacto.

3.3.1.2 Resultados do Teste 2

De forma similar ao Teste 1, foram gerados os respectivos sistemas de produto para os inventários: “*Machining of aluminium parts*” (Teste 2.1) e “*Machining of steel alloy DIN 20MnCr5, hobbing process, at plant, for industrial use, BR 2021 (00.00.119)*” (Teste 2.2)

utilizando a base de dados brasileira. Os processos em questão se referem às atividades de usinagem realizadas no âmbito fabril para o alumínio (Firmino et al., 2022) e para o fresamento de engrenagens automotivas de aço (Soares, 2021). Logo, a fronteira de ACV dos estudos é do tipo *gate-to-gate*, englobando apenas as operações que ocorrem dentro dos portões da fábrica para que o processo de usinagem seja realizado.

É importante salientar que ambos os conjuntos disponíveis no SICV foram modelados seguindo a metodologia UPLCI (*unit process life-cycle inventory*) proposta por Kellens et al., (2012a, 2012b). De forma geral, a metodologia proposta funciona como uma sistemática para processos unitários de fabricação, permitindo que os conjuntos de dados de processos unitários sejam melhores aproveitados em bancos de dados e bibliotecas de inventário do ciclo de vida. Os inventários mencionados acima foram modelados seguindo a abordagem “*in-depth*”, portanto, os dados foram subdivididos em quatro módulos, incluindo um estudo de tempo, um estudo de consumo de energia, um estudo de consumíveis e um estudo de emissões (Kellens et al., 2012a, 2012b).

A modelagem dos conjuntos de dados feita por Firmino et al. (2022) e Soares (2021) foi segregada em parâmetros de modo a facilitar a customização dos dados a depender da especificidade dos equipamentos do usuário. Assim, o *dataset* modelado utiliza o tempo de cada um dos modos de operação da máquina como parâmetros independentes e que podem ser facilmente customizados para atualizar o inventário, conforme Tabela 3.4.

Tabela 3. 4 - Parâmetros independentes utilizados em cada *dataset*.

Machining of aluminium parts	Machining of steel alloy DIN 20MnCr5, hobbing process, at plant, for industrial use, BR 2021 (00.00.119)
<u>Time_Startup</u> : Corresponds to the total time to start the machine in seconds (s)	<u>Time_WarmUp</u> : Corresponds to the total time to the machine stay in warmup in seconds (s)
<u>Time_Standby</u> : Corresponds to the total machine wait time in seconds (s)	<u>Time_ReadforProcessing</u> : Corresponds to the total time to the machine stay in read for processing in seconds (s)
<u>Time_Full_load</u> : Corresponds to the total time of full load mode or machine operation in seconds (s)	<u>Time_Standby</u> : Corresponds to the total time to the machine stay in standby in seconds (s)
<u>Time_Shutdown</u> : Corresponds to the total machine shutdown time in seconds (s)	<u>Time_ExtStandby</u> : Corresponds to the total time to the machine stay in Extstandby in seconds (s)
<u>Elect_Consum_Full_load</u> : Corresponds to energy consumption during the Full load process in kwh/s	<u>Time_Processing</u> : Corresponds to the total time to the machine stay in processing in seconds (s)

Elect_Consum_Shutdown: Corresponds to energy consumption during the Shutdown process in kwh/s

Elect_Consum_Standby: Corresponds to energy consumption during the machine waiting process in Kwh/s

Elect_Consum_Startup: Corresponds to energy consumption during the process of starting the machine in Kwh/s

aluminium_consum: Corresponds to the entrance of aluminum tubes for the machining process in Kg/s

aluminium_parts: Represents the amount of everything machined in Kg / s

alumiun_chips: Corresponds to the amount of chip removed from machining in Kg / s

compressed_air_consum: Refers to compressed air consumption in Nm3 / s

Sintetic_fluid_consum: Refers the amount of cutting fluid consumed during machining in Kg/s

recovered_sintetic_fluid: Refers the amount of cutting fluid recovered after machining in kg/s

tool_wear_reuse: Refers to the amount of machining tool that is used after the process in units/s

tool_wear_start: Refers to the amount of machining tool that is placed before the process in units/s

waste_water: Corresponds to the amount of effluent that leaves the machining process in liters/s

Elect_Consum_ExtStandby: Corresponds to energy consumption during the process of Ext. standby the machine in 1s

Elect_Consum_Processing: Corresponds to energy consumption during the machine processing process in 1s

Elect_Consum_ReadforProcessing: Corresponds to energy consumption during the Read for Processing process in 1s

Elect_Consum_Standby: Corresponds to energy consumption during the process of standby the machine in 1s

Elect_Consum_WarmUp: Corresponds to energy consumption during the WarmUp process in 1s

Gear_machined: Represents the amount mass of the final product (machined parts per second) [Kg/s]

Gear_SemiFinished: Corresponds to the entrance of semi-finished geras for the machining process per second [Kg/s]

Integral_fluid_consum: Corresponds to the fluid consumption per second [liters for second]

Contaminated_oil: Corresponds to the total contaminated oil for second (kg for s)

Fonte: autoria própria.

Com base nos parâmetros independentes identificados, é válido destacar a existência de parâmetros dependentes (Tabela 3.5), ou seja, itens que dependem dos mesmos para representar o ciclo de vida dos processos de usinagem.

Tabela 3. 5 - Parâmetros dependentes utilizados em cada dataset.

Machining of aluminium parts		Machining of steel alloy DIN 20MnCr5, hobbing process, at plant, for industrial use, BR 2021 (00.00.119)	
<i>Parâmetro dependente</i>	<i>Itens que se relacionam</i>	<i>Parâmetro dependente</i>	<i>Itens que se relacionam</i>
<u>Consum_Total_Electricity</u> : electricity per cycle	Elect_Consum_Full_load , Time_Full_load,	<u>Consum_Total_Electricity</u> : electricity per cycle (kwh/s)	Elect_Consum_ExtStandby, Time_ExtStandby, Elect_Consum_Processing,

	Elect_Consum_Shutdown, Time_Shutdown, Elect_Consum_Standby, Time_Standby, Elect_Consum_Startup e Time_Startup		Time_Processing, Elect_Consum_ReadforProcessing, Time_ReadforProcessing, Elect_Consum_Standby, Time_Standby, Elect_Consum_WarmUp e Time_WarmUp
<u>Aluminium consum Total:</u> kg per cycle	aluminium_consum e Time_Full_load	<u>Contaminated oil total:</u> kg of contaminated oil per cycle	contaminated_oil e Time_Processing
<u>Aluminium parts total:</u> kg per cycle	aluminium_parts e Time_Full_load	<u>Gear machined total:</u> kg of steel sustained per cycle	gear_machined e Time_Processing
<u>Alumium chips total:</u> kg per cycle	alumium_chips e Time_Full_load	<u>Gear Semifinished total:</u> kg of gear steel come per cycle	gear_SemiFinished e Time_Processing
<u>Compressed air consum tota</u> <u>l:</u> Nm3 per cycle	compressed_air_consum e Time_Full_load	<u>Integral fluid consum tota</u> <u>l:</u> L of integral fluid used per cycle	integral_fluid_consum e Time_Processing
<u>Recovered sintetic fluid tota</u> <u>l:</u> kg per cycle	recovered_sintetic_fluid e Time_Full_load	<u>Steel chips gear:</u> kg of steel removed per cycle	gear_SemiFinished, Gear_machined e Time_Processing
<u>Sintetic fluid consum total:</u> kg per cycle	Sintetic_fluid_consum e Time_Full_load		
<u>Time total:</u> s	time_Full_load, time_Shutdown, time_Standby e time_Startup		
<u>Tool wear start total:</u> units per cycle	tool_wear_start e time_Full_load		
<u>Tool wear waste total:</u> units per cycle	tool_wear_start, tool_wear_reuse e time_Full_load		
<u>Waste water total:</u> liters per cycle	waste_water e Time_Full_load		

Fonte: autoria própria.

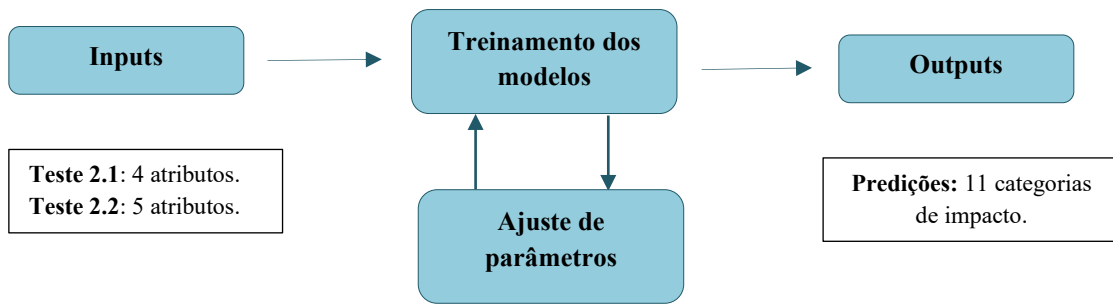
A observação dos parâmetros contidos nas tabelas acima possibilita confirmar a dependência dos fatores, sobretudo, para os fatores relativos ao tempo de cada um dos modos de operação das máquinas, os quais aparecem em todas as linhas de parâmetros dependentes. Esse fator, permite estabelecer uma relação entre os tempos de operação e as respectivas quantidades de material, energia e resíduos gerados em cada um dos processos. Portanto, ao variar os tempos de cada modo de operação (antes tidos como 1s) é possível gerar novos dados de entrada para o inventário (Ex: consumo de energia, consumo de fluido de corte, consumo de ar comprimido, etc) e, por conseguinte, rodar os respectivos sistemas de produto e gerar novos

potenciais impactos para cada um desses cenários. Assim, a parametrização foi utilizada para apoiar a geração de amostras.

Para apoiar nessa etapa de geração de amostras, foi utilizada a funcionalidade “aleatórioentre” do excel de modo a permitir a variação dos tempos de cada modo de operação da máquina de 1 a 600 segundos (10 min). Com base nos resultados obtidos por Salla et al. (2025) foram geradas 200 amostras para cada um dos processos de usinagem, visto que o intervalo de 100 a 200 amostras foi o mais representativo entre os 59 estudos que utilizaram a integração ACV+AM para a predição de potenciais impactos ambientais. Em seguida, esses dados de inventário gerados foram usados para alimentar a funcionalidade de “*Projects*” do OpenLCA e gerar os respectivos potenciais impactos ambientais de uma só vez utilizando o método de AICV definido.

Por tratar de amostras geradas a partir de mesmos *datasets* não foi realizada a etapa completa de pré-processamento dos dados, visto que as amostras geradas são apenas variações aleatórias dos conjuntos de dados existentes, logo, sabe-se que não há a presença de dados ruidosos ou de dados faltantes, sendo apenas explorado nessa análise a correlação entre os atributos e a questão dos *outliers* por categoria de impacto. Assim, cada uma das amostras geradas contou com 4 ou 5 variáveis independentes (*inputs*) e 11 atributos-alvo (*outputs*), para as quais foram feitos treinamentos considerando uma categoria de impacto por vez (Figura 3.8) e, cujos resultados estão detalhados na subseção 3.3.2.2.

Figura 3. 8 - Estrutura para predição de potenciais impactos ambientais para o Teste 2.



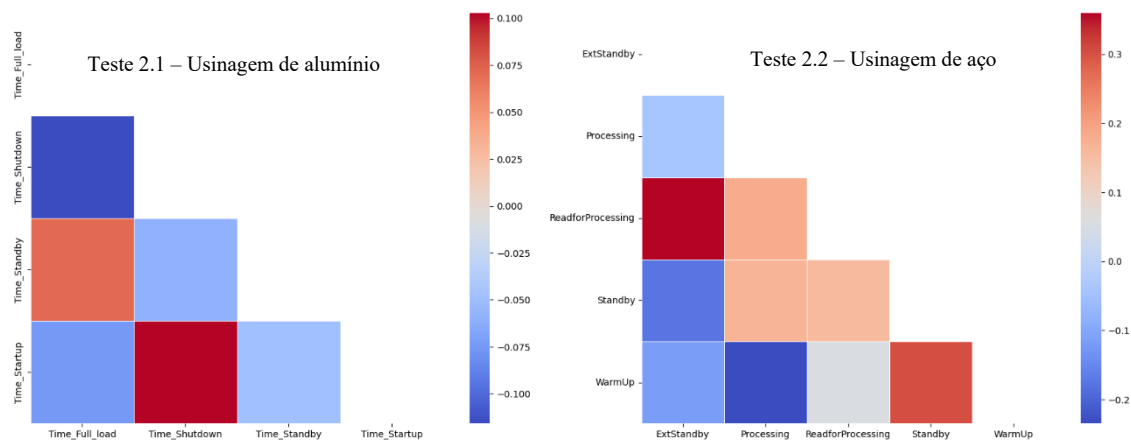
Teste 2.1	I. Time_Startup II. Time_Standby III. Time_Full_load IV. Time_Shutdown	Modelos: regressão bayesiana, KNN, LR, DT, RF, GBM, XGBoost, ANN, SVM, Catboost, E_net, SGD e KRR. Hiperparâmetros: pré-definição de um <i>range</i> de parâmetros.	AC: acidification; CC: climate change; EF: ecotoxicity: freshwater; EM: ecotoxicity: marine; ET: ecotoxicity: terrestrial; ER: energy resources: non-renewable; EU: eutrophication; HT: human toxicity; MR: material resources: metals/minerals; OD: ozone depletion; PO: photochemical oxidant formation.
Teste 2.2	I. Time_WarmUp II. Time_ReadforProcessing III. Time_Standby IV. Time_ExtStandby V. Time_Processing		

Fonte: autoria própria.

Análise dos atributos e pré-processamento dos dados (Teste 2)

A partir da análise da matriz de correlação dos atributos (Figura 3.9), observou-se que a maior correlação linear positiva ocorreu entre os atributos “time_ReadforProcessing” e “time_ExtStandby” para o conjunto amostral dos dados gerados a partir do dataset “*Machining of steel alloy*” para o Teste 2.2, com correlação máxima de 0,35. Por outro lado, a correlação positiva de maior intensidade para os dados gerados a partir do dataset “*Machining of aluminium parts*” ocorreu entre os atributos “Time_Startup” e “Time_ShutDown” com correlação de 0,1. Logo, tais resultados permitem concluir inexistência de relação entre as variáveis, confirmando a aleatoriedade no processo de geração de amostras utilizado e permitindo manter o mesmo número de atributos para o *pipeline* de treinamento de cada um dos *datasets*.

Figura 3. 9 - Matriz de correlação dos atributos de entrada (*Aluminium e Steel*).



Fonte: autoria própria.

Para a identificação de *outliers*, foi utilizado o método IQR (*Interquartile Range*) com fator multiplicador de 1,5, assim como no Teste 1 e resultados obtidos foram descritos na Tabela 3.6. O uso do fator 1,5 foi popularizado pelo estatístico Tukey em meados da década de 70. Tukey (1977) introduziu o conceito de *outliers* com base no IQR, propondo esse multiplicador como um critério para identificar pontos de dados que se afastam significativamente da distribuição central.

Tabela 3. 6 – Análise descritiva dos atributos-alvo para o teste 2.

Alvo	Unidade	Dataset correspondente	Valor mínimo	Valor máximo	Média	Desvio padrão	Outliers
AC	kg SO ₂ -Eq	Steel	$4,40 \times 10^{-06}$	$7,04 \times 10^{-05}$	$7,24 \times 10^{-06}$	$6,44 \times 10^{-06}$	29
		Aluminium	$1,22 \times 10^{-05}$	$9,00 \times 10^{-04}$	$4,39 \times 10^{-05}$	$9,16 \times 10^{-05}$	23
CC	kg CO ₂ -Eq	Steel	$6,00 \times 10^{-04}$	$9,59 \times 10^{-03}$	$9,85 \times 10^{-04}$	$8,77 \times 10^{-04}$	29
		Aluminium	$1,09 \times 10^{-03}$	$1,22 \times 10^{-01}$	$5,40 \times 10^{-03}$	$1,25 \times 10^{-02}$	23
EF	kg 1,4- CB-Eq	Steel	$1,21 \times 10^{-06}$	$1,94 \times 10^{-05}$	$1,99 \times 10^{-06}$	$1,77 \times 10^{-06}$	29
		Aluminium	$2,28 \times 10^{-06}$	$2,50 \times 10^{-04}$	$1,10 \times 10^{-05}$	$2,55 \times 10^{-05}$	23
EM	kg 1,4-DCB-Eq	Steel	$3,26 \times 10^{-01}$	$5,22 \times 10^{+00}$	$5,36 \times 10^{-01}$	$4,77 \times 10^{-01}$	29
		Aluminium	$5,92 \times 10^{-01}$	$6,64 \times 10^{+01}$	$2,94 \times 10^{+00}$	$6,78 \times 10^{+00}$	23
ET	kg 1,4-DCB-Eq	Steel	$7,29 \times 10^{-07}$	$1,17 \times 10^{-05}$	$1,20 \times 10^{-06}$	$1,07 \times 10^{-06}$	29
		Aluminium	$1,33 \times 10^{-06}$	$1,50 \times 10^{-04}$	$6,59 \times 10^{-06}$	$1,53 \times 10^{-05}$	23
ER	MJ	Steel	$0,00 \times 10^{+00}$	$0,00 \times 10^{+00}$	$0,00 \times 10^{+00}$	$0,00 \times 10^{+00}$	0
		Aluminium	$0,00 \times 10^{+00}$	$0,00 \times 10^{+00}$	$0,00 \times 10^{+00}$	$0,00 \times 10^{+00}$	0
EU	kg PO ₄ -Eq	Steel	$1,07 \times 10^{-07}$	$1,71 \times 10^{-06}$	$1,76 \times 10^{-07}$	$1,57 \times 10^{-07}$	29
		Aluminium	$7,23 \times 10^{-07}$	$2,23 \times 10^{-05}$	$1,49 \times 10^{-06}$	$2,23 \times 10^{-06}$	23
HT	kg 1,4-DCB-Eq	Steel	$8,87 \times 10^{-05}$	$1,42 \times 10^{-03}$	$1,46 \times 10^{-04}$	$1,30 \times 10^{-04}$	30
		Aluminium	$2,40 \times 10^{-04}$	$1,82 \times 10^{-02}$	$8,80 \times 10^{-04}$	$1,85 \times 10^{-03}$	23
MR	kg Sb-Eq	Steel	$0,00 \times 10^{+00}$	$0,00 \times 10^{+00}$	$0,00 \times 10^{+00}$	$0,00 \times 10^{+00}$	0
		Aluminium	$0,00 \times 10^{+00}$	$0,00 \times 10^{+00}$	$0,00 \times 10^{+00}$	$0,00 \times 10^{+00}$	0
OD	kg CFC-11-Eq	Steel	$3,00 \times 10^{-13}$	$4,70 \times 10^{-12}$	$4,81 \times 10^{-13}$	$4,33 \times 10^{-13}$	36
		Aluminium	$0,00 \times 10^{+00}$	$6,01 \times 10^{-11}$	$2,60 \times 10^{-12}$	$6,16 \times 10^{-12}$	22
PO	kg ethylene-Eq	Steel	$2,22 \times 10^{-07}$	$3,55 \times 10^{-06}$	$3,65 \times 10^{-07}$	$3,24 \times 10^{-07}$	29
		Aluminium	$6,03 \times 10^{-07}$	$4,53 \times 10^{-05}$	$2,20 \times 10^{-06}$	$4,61 \times 10^{-06}$	23

* *Legendas: AC: acidification; CC: climate change; EF: ecotoxicity: freshwater; EM: ecotoxicity: marine; ET: ecotoxicity: terrestrial; ER: energy resources; EU: eutrophication; HT: human toxicity; OD: ozone depletion; PO: photochemical oxidant formation.*

Fonte: autoria própria.

Com base nos resultados reportados na tabela acima, notou-se a existência de *outliers* em praticamente todas as categorias de impactos analisadas. Assim, as respectivas amostras com *outliers* foram suprimidas do *pipeline* de treinamento dos modelos de AM em suas respectivas etapas de execução. Por exemplo, para o *pipeline* de treinamento da categoria de acidificação, para o conjunto de dados de alumínio, foram utilizadas apenas 177 amostras, visto que 23 delas foram removidas do campo amostral por serem *outliers*, conforme dados da tabela acima.

Não foram encontradas amostras com valores de potenciais impactos para as categorias “*energy resources: non-renewable*” e “*material resources: metals/minerals*”, portanto, não foi

seguido o *pipeline* em nenhuma dessas categorias. Assim, o pipeline de treinamento foi executado apenas para cada uma das 9 categorias de impacto restantes no campo amostral para as bases de dados provenientes dos *datasets* de alumínio e de aço, Testes 2.1 e 2.2, respectivamente.

3.3.1.3 Resultados do teste 3

Para a escolha dos dados agrícolas do terceiro teste, foi definida a utilização de conjuntos de dados para as culturas que estivessem disponíveis nas três bases (Agribalyse, Agrifootprint e Ecoinvent) e com números relevantes de amostras. Assim, foram selecionadas algumas culturas para compor o campo amostral do teste em questão, conforme Tabela 3.7.

Tabela 3. 7 - Dados de inventário usados para geração amostras e suas respectivas fontes.

Cultura /Base de dados	ecoinvent_391_cutoff_upr _regionalized	agribalyse_v3_0_1	agrifootprint_5_200127
Barley grain	13	18	198
Grain maize	26	16	159
Rape seed	10	7	96
Rice	8	8	84
Soybean	20	4	75
Sunflower	7	18	69
Wheat	22	37	324
Total (n°)	106	108	1005

Fonte: autoria própria.

Com base na quantidade de *datasets* disponíveis nas bases de dados, foram selecionadas duas diferentes abordagens para a geração de amostras (contendo entradas dos processos, saídas e respectivos impactos potenciais gerados para o método). Para as bases do Ecoinvent e Agribalyse o processo de exportação dos inventários, geração dos sistemas de produto, rodagem dos sistemas de produto usando o método CML e exportação dos resultados dos potenciais impactos (categorias de impacto) foram feitos de forma manual (um a um) em virtude do tamanho reduzido das amostras, desse modo, optou-se por não utilizar as ferramentas do “*developer tools*” disponíveis no OpenLCA.

Por outro lado, para os dados da base Agrifootprint foi necessário utilizar a funcionalidade IPC Server, mencionada anteriormente, em virtude do alto volume de dados disponíveis. Desse modo, foi utilizado um procedimento automatizado utilizando códigos com linguagem de programação Python, conforme detalhamento nos subtópicos a seguir.

A) Processo de geração de amostras do Agrifootprint:

O processo de geração de amostras para os dados do Agrifootprint seguiu duas etapas, sendo:

- 1) Exportação dos inventários: na etapa em questão os inventários foram exportados utilizando a própria função de exportar do OpenLCA. Nesse sentido, todos os processos da pasta de dados agrícolas foram exportados de uma única vez em diversos arquivos com a extensão “.xlsx”, contendo 17 abas cada, sendo as abas: inputs, outputs, parâmetros e general information as abas de interesse. Isso tudo, por conterem dados quantitativos de entradas e saídas, os respectivos parâmetros usados (caso haja) e as informações qualitativas (metadados) do processo de cultivo agrícola das diferentes culturas.
- 2) Exportação dos potenciais impactos: na etapa em questão, os resultados dos potenciais impactos ambientais foram gerados para cada um dos processos selecionados da base dados. Para tal, foi elaborado um código em Python utilizando as bibliotecas *olca-schema*, *olca-ipc*, *pandas* (usada para trabalhar com conjuntos de dados) e *logging* (rastrear eventos que acontecem quando algum software executa) para apoiar tal processo, sendo o mesmo dividido em cinco subetapas:
 - a) acesso a cada um dos conjuntos de dados de interesse (conforme Tabela 3.7);
 - b) geração dos respectivos sistemas de produto utilizando provedores do tipo padrão (*providers default*);
 - c) rodagem de cada um dos sistemas de produto utilizando o método *Ecoinvent - CML v4.8 2016* e garantindo que os resultados refletissem a produção de 1kg de cada cultura;
 - d) exportação dos resultados para os potenciais impactos de cada sistema de produto e;
 - e) exclusão dos sistemas de produto gerados a fim de evitar sobrecarregar o funcionamento do software OpenLCA.

Tudo isso foi realizado, operando em formato de *looping*, ou seja, gerando os resultados de impacto de forma contínua para um processo de cada vez, o que possibilitou gerar 1.005 arquivos do tipo “.xlsx” com uma aba contendo os potenciais impactos de cada processo agrícola.

B) Processo de compilação das amostras:

Como destacado anteriormente, cada amostra gerada, independente da base de dados usada (*Ecoinvent*, *Agribalyse* e *Agrifootprint*), foi composta por dois arquivos “.xlsx” sendo:

- Um arquivo referente ao inventário em si, composto por 17 diferentes abas, sendo elas: “General information”, “Modeling and validation”, “Administrative information”, “Inputs”, “Outputs”, “Allocation”, “Parameters”, “Flows”, “Flow property factors”, “Flow properties”, “Unit groups”, “Units”, “Currencies”, “Locations”, “Actors”, “Sources” e “Providers”;
- Um arquivo composto por uma aba contendo os resultados para os potenciais impactos ambientais para a produção de 1kg de cada cultura de acordo com as 11 categorias de impacto propostas pelo método *Ecoinvent - CML v4.8 2016*.

Agrupamento dos dados para cada uma das bases

Em virtude dos dados virem de diferentes bases e de contarem com dois arquivos “.xlsx” diferentes, foi necessário fazer o processo de compilação das amostras. O processo em questão acabou sendo trabalhoso e relativamente lento, já que foi preciso fazer diferentes adaptações aos dados para a compilação das amostras advindas das diferentes bases de dados.

No caso das bases do *Agribalyse* e *Ecoinvent*, o processo de compilação das amostras de cada uma das bases foi feito de forma manual em virtude da quantidade de dados disponíveis. Para isso, utilizou-se de planilhas para organizar as entradas das diferentes culturas em atributos que fossem de fato comuns, chegando ao final em uma planilha agrupada (com diferentes culturas) para cada base e com seus respectivos atributos.

Para a base do *Agrifootprint*, o processo utilizado diferiu um pouco, sendo necessário a realização de algumas etapas adicionais e o uso de códigos em Python para a compilação de seus dados. Inicialmente, foi realizada a etapa de (a) renomeação dos nomes dos arquivos gerados (inventário e potenciais impactos) para cada um dos conjuntos de dados, com o objetivo de garantir que o identificador do processo (ID) fosse mantido o mesmo no momento de compilar potenciais impactos e entradas/saídas do processo. Esse processo foi necessário, pois o ID do arquivo de inventário exportado (contendo 17 abas) é diferente do ID dos potenciais impactos exportado para a mesma cultura, o que impossibilitava relacionar o arquivo contendo os atributos de entrada (entradas agrícolas) com o arquivo contendo os atributos-alvo (resultados para as categorias de impacto).

Em seguida, foi realizada a etapa de (b) verificação da presença de dados não numéricos na aba *inputs*, sobretudo, para analisar a existência ou não de fluxos de produto com dados em

formato de valor parametrizado (Ex: Input_0 ao invés do valor numérico). Isso tudo, pois esse “valor não-numérico/categórico” (valor em formato de parâmetro) acaba não sendo substituído pelo respectivo dado numérico durante o processo de exportação, sendo assim, representa um ponto de atenção e que precisa ser tratado para quem pretende fazer essa integração. Logo, é necessário observar as abas “*inputs*” e “*outputs*” para verificar tais existências e se for observado, substituí-las pelos respectivos valores dos parâmetros a partir da observação da aba “*parameters*”.

Após verificar a inexistência de dados categóricos e fazer as devidas substituições, foi realizada a etapa de (c) compilação das abas de *inputs* e os potenciais impactos em uma única planilha para cada um dos conjuntos de dados. Por fim, (d) os dados das 1005 planilhas diferentes foram compilados em uma planilha única e agrupada para a base Agrifootprint, utilizando os passos do tópico a seguir.

Integração dos dados das diferentes bases em uma base única

A etapa de integração dos conjuntos de dados das diferentes bases em uma base única mostrou-se tediosa. Isso se justifica pela necessidade de tratar os fluxos de entrada de cada um dos inventários um a um, visto que bases de dados diferentes usam fluxos de produto com nomes diferentes e que de forma geral podem significar coisas muito parecidas ou até as mesmas coisas. Como por exemplo, os fluxos “[*sulfonyl*]urea-compound {RER}| production” do Agribalyse, “[*sulfonyl*]urea-compound” do Ecoinvent e “Urea, as 100% CO(NH₂)₂ (NPK 46.6-0-0)” do Agrifootprint que apesar de possuírem nomes diferentes, referem-se à utilização de ureia nas culturas agrícolas.

Antes de seguir para a transformação e integração dos atributos, foi padronizada a nomenclatura das amostras com o objetivo de facilitar a rastreabilidade da cultura e da base de dados relativa à mesma. Logo, a primeira amostra de milho vinda do Ecoinvent recebeu o identificador EI_MA_1 e assim por diante, conforme Tabela 3.8.

Tabela 3. 8 - Padronização da nomenclatura das amostras do Teste 3.

Sigla	Base de dados relativa	Sigla	Cultura relativa	Tradução
EI	Ecoinvent	BA	Barley grain	Cevada
AG	Agrifootprint	MA	Grain maize	Milho
AY	Agribalyse	RA	Rape seed	Colza / canola
		RI	Rice	Arroz
		SO	Soybean	Soja
		SU	Sunflower	Semente de girassol
		WH	Wheat	Trigo

Fonte: autoria própria.

Após a padronização dos identificadores das amostras foi realizada a etapa de integração dos atributos para agrupar as bases. Todos os fluxos de entrada (atributos ou variáveis independentes) mapeados nos processos das três bases de dados foram tratados e transformados de modo a permitir a integração das bases.

Inicialmente, a base de dados Agribalyse contava com 314 atributos distintos para as entradas e 16 atributos distintos representando as saídas, ou seja, 16 diferentes saídas representando as culturas em si (Ex: grãos, grãos orgânicos e até mesmo palha da cultura). A base de dados do Ecoinvent tinha 147 atributos de entrada para 16 culturas distintas, enquanto, que a base do Agrifootprint contava com 81 atributos como entradas do processo e 17 representando as saídas das culturas. Ao todo, as três bases juntas contavam com cerca de 570 atributos com nomes distintos e que precisavam ser transformados e integrados em itens que fossem comuns.

O processo de transformação inicial dos atributos foi realizado em duas etapas subsequentes, possibilitando chegar em 169 atributos que permitiram a integração das três bases de dados, sendo 146 atributos referentes as entradas e 23 representando as diferentes variações das culturas. O Anexo - 3B ressalta como foi feito esse processo de transformação e indica as principais alterações necessárias para compilar as amostras em atributos que fossem de fato comuns entre elas.

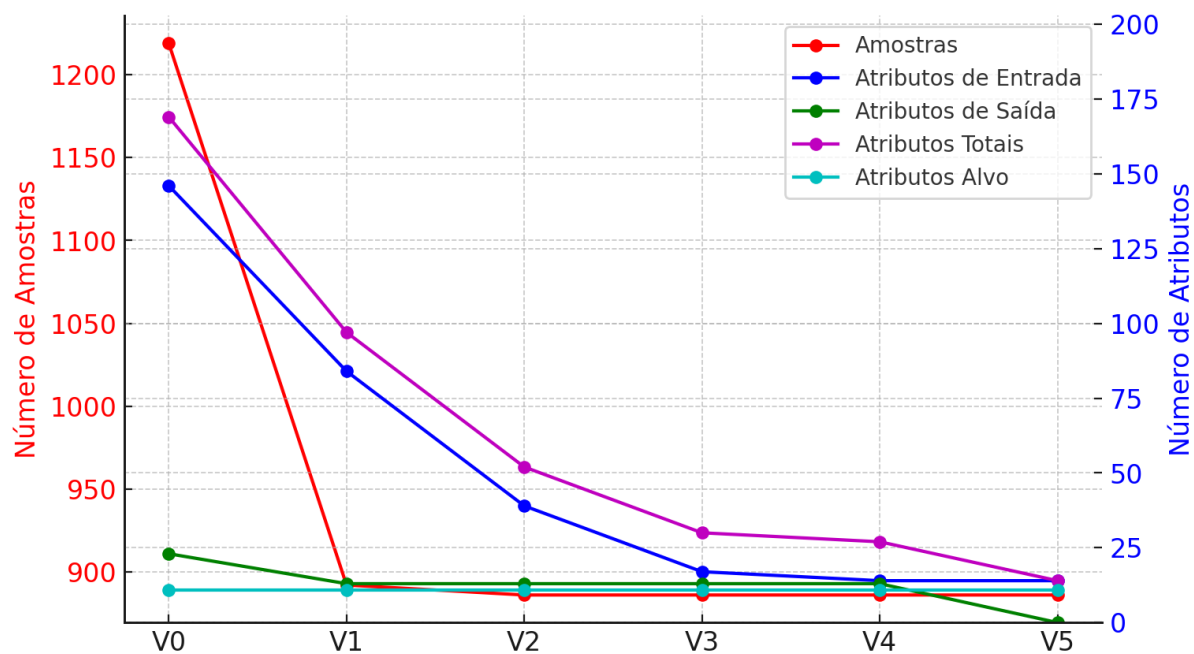
Após a integração das amostras das três bases de dados, a planilha resultante com 1.219 (108 da Agribalyse, 106 do Ecoinvent e 1005 do Agrifootprint) linhas com 169 atributos de entrada e 11 atributos-alvo (categorias de impactos) seguiu para a etapa de pré-processamento dos dados a fim de preparar, organizar e estruturar de melhor forma os dados que foram agrupados para o, posterior, treinamento dos modelos (Faceli et al., 2022).

Pré-processamento dos dados

Existem diversas técnicas para a preparação dos dados, entre elas a limpeza de dados (amostras com ruído), a transformação de dados (envolvendo a normalização e a seleção de atributos) e a redução de dados (englobando a redução de dimensionalidade, amostragem, entre outros) (Faceli et al., 2022; Gomes, 2019). Parte das técnicas citadas acima foi utilizada para melhorar a qualidade dos dados em quatro rodadas distintas de preparação, o que possibilitou gerar cinco diferentes bases de dados (Figura 3.10).

A Tabela 3.9 concentra todos ajustes realizados nas quatro rodadas de pré-processamento dos dados, indo desde a base de dados compilada inicialmente com 169 atributos e 1.219 amostras (V0) até a base de dados final com 14 atributos e 886 amostras (V5).

Figura 3. 10 – Disposição dos dados disponíveis por versão da base.



Fonte: autoria própria (gerada com auxílio do ChatGPT).

Tabela 3. 9 – Padronização da nomenclatura das amostras do teste 3.

Versão da base	Características iniciais		Ajustes realizados			Características finais	
	Nº Atributos	Nº Amostras	Limpeza de dados	Exclusão de atributos	Combinação de atributos	Nº Atributos	Nº Amostras
V0	169	1219	-	-	-	169	1219
V1	169	1219	Agrifootprint: exclusão de 306 amostras por serem de secagem (só há a entrada de energia elétrica e vapor de água e a saída de grão seco). Agribalyse: exclusão de 7 amostras de processo de secagem (só a entrada de grão e a saída de grão seco). Ecoinvent: exclusão de 14 amostras, visto que a saída principal deles era a palha ao invés da cultura em si.	Atributos de entrada: - Exclusão de 23 atributos que estavam presentes em um número baixo de amostras. - Exclusão de 2 atributos de entrada com dados potencialmente errados (unidade em toneladas). Atributos de saída: - Exclusão de 1 atributo com valor zerado em todas as amostras. - Exclusão de 1 atributo com valor em apenas uma amostra. - Exclusão de 6 atributos cujo foco era a produção da palha e não da cultura em si. Detalhes no Anexo – 3C.	Atributos de entrada: - Compilação de 52 atributos de entrada em 13 atributos de mesma natureza. Detalhes no Anexo – 3C.	97	892
V2	97	892	Agribalyse: exclusão de 6 por estarem com entradas negativas para o diesel e esterco verde (dado ruidoso).	Atributos de entrada: - Exclusão de 37 atributos que estavam presentes em um número baixo de amostras ou não estavam presente em alguma das bases. Detalhes no Anexo – 3C.	Atributos de entrada: - Compilação de 13 atributos de entrada em 5 atributos, fazendo as devidas transformações. Detalhes no Anexo – 3C.	52	886

V3	52	886	-	-	Atributos de entrada: - Compilação de 28 atributos de entrada em 6 atributos, fazendo as devidas transformações. Detalhes no Anexo – 3C.	30	886
V4	30	886	-	Atributos de entrada: - Exclusão de 3 atributos por possuírem correlação linear (acima de 0,7) forte com outros atributos. Detalhes no Anexo – 3C.	-	27	886
V5	27	886	-	Atributos de saída: - Exclusão dos 13 atributos relativos às culturas, visto que as amostras das diferentes culturas mostraram estáveis. Detalhes no Anexo – 3C.	-	14	886

Fonte: autoria própria.

A Tabela 3.9 sumariza os dados após a aplicação de técnicas que envolveram, principalmente, a limpeza de dados com ruído, a exclusão de atributos com pouca influência ou presente em poucas amostras e a etapa de transformação dos atributos, a qual possibilitou agrupar atributos de natureza semelhante.

A etapa em questão foi realizada com o objetivo de melhorar a qualidade dos dados, seja excluindo amostras com ruído, seja reduzindo a dimensionalidade do problema (atributos). Isso tudo, com o intuito de respeitar os achados obtidos por Salla et al. (2025), em que 90% dos estudos com integração LCA+ML para dados agrícolas (em 19 das 21 ocorrências) utilizaram de 6 a 20 atributos para realizar a predição de potenciais impactos ambientais.

Nesse sentido, a etapa de pré-processamento dos dados foi realizada com o objetivo de deixar o campo amostral com amostras sem dados ruidosos e ainda deixar os atributos em um formato mais simples de serem informados por usuários sem experiência prévia em ACV. Em seguida, foi realizado o treinamento dos modelos juntamente com o processo de ajuste de hiperparâmetros, seguido pelas etapas de interpretação e validação dos resultados obtidos.

3.3.2 Resultados para o treinamento dos modelos de AM

Esta seção apresenta os principais resultados quantitativos obtidos em cada um dos três testes realizados, sendo:

- Teste 1: resultados obtidos para a base de dados com 24 atributos e no máximo 15 amostras de milho para o Brasil (amostras a serem removidas em caso de existência de *outliers* percebidos durante o *pipeline* de treinamento dos modelos, conforme Tabela 3.3).
- Teste 2: resultados obtidos para duas bases de dados diferentes:
 - Teste 2.1: contendo 4 atributos e no máximo 178 amostras de usinagem de alumínio (amostras a serem removidas em caso de existência de *outliers* percebidos durante o *pipeline* de treinamento dos modelos, conforme Tabela 3.6);
 - Teste 2.2: contendo 5 atributos e no máximo 171 amostras de usinagem de aço (amostras a serem removidas em caso de existência de *outliers* percebidos durante o *pipeline* de treinamento dos modelos, conforme Tabela 3.6);
- Teste 3: resultados obtidos para as bases de dados com atributos variando de 169 a 14 e amostras variando de 1.219 até 886 (base de dados versão v0 até v5).

A divisão dos dados em conjuntos de treino e teste é um passo essencial no desenvolvimento de modelos de aprendizado de máquina, pois possibilita avaliar seu desempenho em dados que não foram utilizados durante o treinamento (Faceli et al., 2022). Essa etapa é muito importante no *pipeline* de treinamento dos modelos, sendo assim, foram seguidas abordagens diferentes nos três testes.

De acordo com a documentação do Scikit-Learn (Pedregosa et al., 2011), a função “*train_test_split*” é amplamente empregada para essa tarefa, assegurando que uma parte dos dados seja utilizada para treinamento e outra para validação. Em situações onde a quantidade de amostras é limitada, uma alternativa eficiente é a validação cruzada *Leave-One-Out*, que permite maximizar o aproveitamento dos dados.

O método *Leave-One-Out Cross-Validation* (LOO ou LOOCV) tem suas raízes na estatística clássica e na teoria da validação cruzada, que foi introduzida por Seymour Geisser em 1975 (Geisser, 1975). Apesar de ter sido inicialmente proposta para modelos estatísticos, tornou-se amplamente utilizada em aprendizado de máquina e modelagem preditiva, principalmente quando os conjuntos de dados são pequenos. Esse método funciona retirando uma única amostra para teste enquanto o restante é usado para treino, repetindo esse processo até que todas as amostras tenham sido testadas individualmente.

Nos dois primeiros testes realizados (Teste 1 e Teste 2), foi utilizada a estratégia LOO em conjunto com o GridSearch para a otimização dos hiperparâmetros. O Grid Search é um método que realiza uma busca exaustiva dentro de um espaço pré-definido de combinações possíveis, identificando aquela que apresenta o melhor desempenho segundo métricas de validação (Bergstra & Bengio, 2012). Como os primeiros testes envolvem um conjunto de dados reduzido, essa abordagem se mostrou adequada.

No entanto, no terceiro teste, em que o volume de amostras é maior, foi aplicada a validação cruzada *k-folds*, técnica que permite dividir os dados em *k* subconjuntos (*folds*). Assim, o modelo é treinado “*k-1*” vezes e testado no subconjunto restante, repetindo o processo para cada *fold*. Essa metodologia melhora a estimativa do desempenho do modelo ao minimizar a variação gerada por uma única divisão dos dados (Chugani, 2024).

Para a otimização dos hiperparâmetros no terceiro teste, optou-se pelo uso do Optuna, uma biblioteca avançada de otimização bayesiana. Segundo Akiba et al. (2019), essa ferramenta facilita a busca por hiperparâmetros ideais utilizando métodos como TPE (*Tree-structured Parzen Estimator*) e *Pruning*, acelerando o processo ao eliminar testes pouco promissores. Diferente do GridSearch disponível no Scikit-Learn, que avalia todas as combinações possíveis, o Optuna ajusta dinamicamente o espaço de busca, tornando a otimização mais eficiente e

reduzindo o tempo computacional necessário (Akiba et al., 2019). Assim, cada abordagem adotada foi escolhida levando em consideração o tamanho das bases de dados e a necessidade de manter certa eficiência computacional para obter o melhor ajuste dos modelos.

Nos três testes realizados, foram utilizadas duas métricas principais para avaliar o desempenho dos modelos de regressão: o coeficiente de determinação (R^2) e o erro quadrático médio (MSE). A utilização conjunta dessas métricas fornece uma visão abrangente do desempenho dos modelos de regressão, permitindo avaliar tanto a proporção da variabilidade explicada pelo modelo (R^2) quanto a magnitude média dos erros de previsão (MSE), permitindo avaliar questões relativas a modelos muito generalistas/simplistas (*underfitting*) e também modelos muito ajustados (*overfitting*) (Scikit-learn developers, 2025; Facelli et al, 2022).

O coeficiente de determinação é uma medida estatística que indica a proporção da variância da variável dependente (categoria de impacto) que é explicada pelas variáveis independentes (dados de entrada e tipo de cultura) no modelo. Ele varia de 0 a 1, onde valores mais próximos de 1 indicam que o modelo explica grande parte da variabilidade dos dados (Harrison, 2020). Por exemplo, um R^2 de 0,8234 significa que o modelo explica 82,34% da variância da variável dependente a partir dos regressores incluídos.

O erro quadrático médio é uma métrica que quantifica a média dos erros quadráticos entre os valores previstos pelo modelo e os valores reais observados (Bruce & Bruce, 2019; Faceli et al, 2022). Ele é calculado pela média das diferenças ao quadrado entre os valores previstos e os valores observados, penalizando mais severamente os erros maiores. Um MSE menor indica que as previsões do modelo estão mais próximas dos valores reais, refletindo um melhor desempenho (Scikit-learn developers, 2025; Bruce & Bruce, 2019; Faceli et al, 2022).

Ao todo 13 diferentes métodos de aprendizado de máquina foram avaliados em cada um dos testes (detalhes no apêndice 3A), sendo eles:

1. Regressão Bayesiana;
2. K-vizinhos mais próximos ou *K-nearest neighbors* (KNN);
3. Regressão linear ou *linear regression* (LR);
4. Árvores de decisão ou *decision tree* (DT);
5. Florestas aleatórias ou *random forest* (RF);
6. *Gradient boosting machine* (GBM);
7. eXtreme gradiente boosting (XGBoost);
8. Redes neurais artificiais ou artificial neural network (ANN);
9. Máquina de vetores de suporte ou *support vector machine* (SVM);
10. Regressão de kernel ou *kernel ridge regression* (KRR);

11. Rede de regressão elástica ou *elastic net regression* (E_net);
12. Método do gradiente estocástico ou *stochastic gradient descent* (SGD);
13. *Categorical Boosting* (CatBoost).

Os nove primeiros métodos vão em consonância aos achados de Salla et al. (2025), sendo os mesmos encontrados com certa frequência em estudos com integração entre avaliação do ciclo de vida e aprendizado de máquina para a predição de potenciais impactos ambientais. Além disso, foram incluídos 4 outros modelos com o objetivo de deixar mais robusto o campo de testes.

É importante destacar que os 13 modelos foram testados em formato de *looping* para cada uma das 11 categorias de impacto e em cada um dos três testes. Nesse sentido, por ser um grande volume de resultados, optou-se por focar as análises apenas nos modelos com melhores desempenhos.

3.3.2.1 Teste 1

Esta subseção apresenta os principais resultados obtidos para as 11 categorias de impactos utilizando a abordagem *leave-one-out* (LOO) para a divisão dos dados em treino e teste e busca de hiperparâmetros usando GridSearch, ambas ferramentas disponíveis na biblioteca no Scikit-Learn.

Os resultados foram analisados utilizando duas métricas: o R^2 (coeficiente de determinação) a fim de mensurar a qualidade da predição de cada modelo e o MSE (erro quadrático médio) para observar a capacidade de ajuste dos modelos aos dados observados.

O R^2 calculado no código está considerando a média de todos os modelos gerados no LOO. Isso ocorre porque a métrica é calculada usando “*cross_val_predict*” do Scikit-Learn para obter as predições de cada amostra quando ela é deixada de fora. Depois, a métrica R^2 é calculada comparando as predições LOO com os valores reais da variável alvo. Esse procedimento equivale a calcular um único R^2 com base em todas as previsões obtidas no processo de LOO, conforme resultados da Tabela 3.10.

Tabela 3. 10 - Resultados para o Teste 1.

Categorias de impacto	Melhor modelo		
	MSE	R ²	Nome do modelo
acidification	0,2276	0,7723	Elastic_net
climate change	0,5032	0,4967	Decision tree
ecotoxicity: freshwater	0,5820	0,4179	Decision tree
ecotoxicity: marine	0,0732	0,9267	Elastic_net
ecotoxicity: terrestrial	0,9209	0,0790	Decision tree
energy resources: non-renewable	0,0882	0,9117	Elastic_net
eutrophication	0,3336	0,6663	KNN
human toxicity	0,8442	0,1557	Elastic_net
material resources: metals/minerals	0,0674	0,9325	Elastic_net
ozone depletion	0,1274	0,8725	Decision tree
photochemical oxidant formation	0,5942	0,4057	Decision tree

* Legendas: KNN: K - Nearest Neighbors.

Fonte: autoria própria.

Os resultados para o R² e MSE indicam uma predominância de melhores resultados sendo obtidos para os modelos que utilizaram a regressão elástica (E_Net). Elastic-Net é um modelo de regressão linear treinado com regularização das normas ℓ_1 e ℓ_2 dos coeficientes. O Elastic-Net é útil quando há múltiplas variáveis que são correlacionadas entre si e para realizar a seleção de variáveis, sendo uma técnica valiosa no arsenal da modelagem (AEF, 2025).

De forma geral, é conclusivo que os modelos funcionaram bem para a predição de algumas categorias de impacto (*ecotoxicity: marine; energy resources: non-renewable; material resources: metals/minerals; ozone depletion*) utilizando poucas amostras, mas falharam para outras (*ecotoxicity: terrestrial; human toxicity; photochemical oxidant formation*).

Para a predição de categorias relacionadas a *acidification, climate change, ecotoxicity: freshwater* e *eutrophication* foi observado um comportamento intermediário dos modelos. Logo, observou-se que os modelos estão capturando razoavelmente a variabilidade desses atributos, mas há margem para melhoria.

Com base nos valores de MSE e R² obtidos, há hipóteses sobre a possibilidade de *underfitting* (subajustamento) e *overfitting* (sobreajustamento) nos modelos. Os resultados sugerem indícios de *underfitting* na predição das categorias com piores desempenhos (*ecotoxicity: terrestrial; human toxicity* e *photochemical oxidant formation*), visto que a predição para tais atributos apresentaram altos erros associados (MSE elevado) e baixa capacidade explicativa (R² próximo de 0), sugerindo que o modelo não está aprendendo com os padrões de dados, podendo se tratar de um modelo simplista.

Para as categorias com as melhores previsões (*ecotoxicity: marine; energy resources: non-renewable; material resources: metals/minerals; ozone depletion*) foram feitas análises avaliando os erros de treino e teste para avaliar a possibilidade de ocorrência *overfitting*. Os resultados obtidos encontram-se detalhados na Tabela 3.11.

Tabela 3. 11- Análise de *overfitting* para o teste 1.

Categorias de impacto	Melhor modelo			
	MSE_LOO	R ²	MSE_treino	Dif (%)
ecotoxicity: marine	0,0732	0,9267	0,0254	65,3%
energy resources: non-renewable	0,0882	0,9117	0,0276	68,7%
material resources: metals/minerals	0,0674	0,9325	0,0195	71,0%
ozone depletion	0,1274	0,8725	0,0057	95,5%

Fonte: autoria própria.

Em todos os casos, a diferença percentual entre o MSE de validação LOO e o MSE de treinamento é positiva, variando de 65,3% a 95,5%. Isso indica que o modelo se ajustou de maneira muito eficiente aos dados de treinamento, porém apresentou um desempenho inferior na validação, sugerindo uma má generalização dos modelos aos novos dados.

Embora o R² nos resultados de treinamento seja alto (variando de 0,8725 a 0,9325), refletindo uma boa explicação da variabilidade dos dados, a discrepância entre os erros de treinamento e validação representa um ponto de atenção, sendo um indicativo da existência de *overfitting* nos modelos. Esse comportamento ocorre quando o modelo é excessivamente ajustado aos dados de treinamento (sobre-ajustamento), perdendo a capacidade de generalizar (gerar previsões contundentes) para novos conjuntos de dados.

A análise dos resultados e o comportamento observado nos modelos indicam que tanto o *overfitting* quanto o *underfitting* podem estar relacionados ao baixo número de amostras disponíveis para treinamento e teste (15 amostras no máximo). Com um número tão reduzido de amostras, o modelo tende a se ajustar excessivamente aos dados específicos do treinamento (*overfitting*), resultando em um desempenho significativamente melhor nas amostras de treino em comparação com a validação (como visto nos valores de MSE e R²). Por outro lado, os modelos também podem acabar não conseguindo estabelecer padrões confiáveis a partir de uma base de dados pequena, resultando em *underfitting* (desempenho insatisfatório tanto no treino quanto na validação).

Esse equilíbrio entre *overfitting* e *underfitting* reflete as limitações de um número muito baixo de amostras. Logo, uma possível solução seria aumentar a quantidade de dados para o

treinamento, o que permitiria ao modelo aprender de maneira mais sólida e generalizar melhor para dados desconhecidos. Nesse sentido, é destacada a importância dos Testes 2 e 3 realizados.

3.3.2.2 Teste 2

De forma similar ao teste anterior, esta subseção explora os principais resultados obtidos para as 9 categorias de impacto (categorias sem impacto: *energy resources: non-renewable*” e *material resources: metals/minerals*, tópico 3.3.1.2) utilizando a abordagem *leave-one-out* (LOO) e a busca de hiperparâmetros usando GridSearch do Scikit-Learn, para as bases de dados de usinagem de alumínio e usinagem de aço.

Como indicadores de desempenho, também foram utilizadas as métricas R^2 e MSE. Os resultados foram compilados em uma tabela única (Tabela 3.12) permitindo a comparação de resultados obtidos em cada uma das bases.

Tabela 3. 12 - Resultados para o teste 2.

Alvo	Machining of aluminium			Machining of steel alloy		
	MSE	R^2	Modelo	MSE	R^2	Modelo
AC	0,0393	0,9606	ANN	0,0506	0,9493	ANN
CC	0,0337	0,9662	ANN	0,0677	0,9322	ANN
EF	0,0427	0,9572	ANN	0,0899	0,9100	ANN
EM	0,0311	0,9688	ANN	0,0698	0,9301	ANN
ET	0,0372	0,9627	ANN	0,0725	0,9274	ANN
EU	0,0379	0,9620	ANN	0,0931	0,9068	ANN
HT	0,0309	0,9690	ANN	0,1125	0,8874	ANN
OD	0,2771	0,7228	ANN	0,2851	0,7148	ANN
PO	0,0404	0,9595	ANN	0,0680	0,9319	ANN

Fonte: autoria própria.

* *Legendas: AC: acidification; CC: climate change; EF: ecotoxicity: freshwater; EM: ecotoxicity: marine; ET: ecotoxicity: terrestrial; EU: eutrophication; HT: human toxicity; OD: ozone depletion; PO: photochemical oxidant formation.*

Os resultados demonstram que o modelo de rede neural artificial (ANN) foi o melhor entre os modelos de regressão testados, alcançando desempenho superior em relação às outras abordagens. Para o campo amostral relativo aos dados de *“machining of aluminium”*, o modelo ANN obteve resultados notáveis, com valores de MSE variando de 0,0311 a 0,2771 e R^2 entre 0,7228 e 0,9690. Tais resultados indicam que o modelo se tornou capaz de ajustar adequadamente a maioria das previsões, exceto em um caso onde o MSE foi maior (0,2771).

No segundo campo amostral (amostras geradas a partir do dataset “*machining of steel alloy*”), o modelo ANN continuou a apresentar os melhores desempenhos, com MSE variando de 0,0506 a 0,1125 e R^2 entre 0,7148 e 0,9493. Embora os resultados sejam inferiores aos do primeiro conjunto amostral, o modelo ainda demonstrou capacidade de generalização. A redução no R^2 e o aumento no MSE em comparação ao primeiro conjunto pode indicar uma maior complexidade do segundo conjunto de dados. A superioridade do modelo ANN em relação aos outros modelos testados destaca sua eficácia em prever com certo nível de precisão, sugerindo que ajustes adicionais ou maior volume de dados poderiam melhorar ainda mais a generalização para este segundo conjunto. De forma similar ao que foi feito no primeiro teste, também foram feitas análises para verificar a possibilidade de existência de *overfitting* nos resultados obtidos (Tabela 3.13).

Tabela 3. 13 - Resultados para o teste 2.

Alvo	Machining of aluminium				Machining of steel alloy			
	MSE_Loo	R^2	MSE_treino	Dif (%)	MSE_Loo	R^2	MSE_treino	Dif (%)
AC	0,0393	0,9606	0,0163	58,5%	0,0506	0,9493	0,0134	73,5%
CC	0,0337	0,9662	0,0084	75,0%	0,0677	0,9322	0,0144	78,3%
EF	0,0427	0,9572	0,0216	49,4%	0,0899	0,9100	0,0240	73,3%
EM	0,0311	0,9688	0,0093	70,1%	0,0698	0,9301	0,0141	79,8%
ET	0,0372	0,9627	0,0176	52,6%	0,0725	0,9274	0,0244	66,3%
EU	0,0379	0,9620	0,0264	30,3%	0,0931	0,9068	0,0173	81,4%
HT	0,0309	0,9690	0,0139	55,0%	0,1125	0,8874	0,0215	80,8%
OD	0,2771	0,7228	0,2205	20,4%	0,2851	0,7148	0,3147	-10,3%
PO	0,0404	0,9595	0,0122	69,8%	0,0680	0,9319	0,0220	67,5%

Fonte: autoria própria.

* *Legendas: AC: acidification; CC: climate change; EF: ecotoxicity: freshwater; EM: ecotoxicity: marine; ET: ecotoxicity: terrestrial; EU: eutrophication; HT: human toxicity; OD: ozone depletion; PO: photochemical oxidant formation.*

A análise dos resultados destaca um padrão de diferenças entre os erros médios quadráticos (MSE_treino) dos conjuntos de treinamento e validação *leave-one-out* (MSE_Loo). Em quase todos os casos, o MSE_Loo é superior ao MSE_treino, o que sugere um possível *overfitting* dos modelos. Isso ocorre porque o modelo consegue se ajustar bem aos dados de treino, mas tem dificuldades para manter esse desempenho quando exposto a novos dados. Em alguns casos, essa diferença ultrapassa 70%, como nas categorias-alvo CC, EM e HT, que apresentam discrepâncias acima de 75% nos dois conjuntos amostrais. Esse comportamento

indica que o modelo pode estar capturando padrões específicos dos dados de treino sem conseguir generalizar adequadamente.

Por outro lado, um ponto fora do padrão aparece no alvo OD do conjunto amostral *Machining of steel alloy* (usinagem do aço), onde a diferença percentual é negativa (-10,38%). Isso significa que o erro no conjunto de validação foi menor que no treinamento, o que pode ter algumas explicações. Uma possibilidade é que o modelo tenha enfrentado dificuldades para ajustar-se aos dados de treino devido à alta variabilidade ou ruído nos exemplos, o que pode ser um indício de *underfitting*. Outra hipótese é que a amostra usada na validação tenha sido mais fácil de prever do que a de treino, reduzindo o erro de forma não esperada.

Os maiores valores de diferença percentual, alvos CC (75,07% e 78,73%), EM (70,10% e 79,80%) e HT (55,02% e 80,89%), reforçam a necessidade de estratégias para tentar reduzir a ocorrência de *overfitting*. Essa limitação pode estar relacionada ao tamanho reduzido da amostra, que restringe a capacidade do modelo de aprender padrões mais sólidos. Para mitigar esse problema, algumas soluções incluem o aumento da base de dados ou até mesmo ajustes na complexidade do modelo, a fim de deixá-lo mais generalista.

Embora haja indícios de *overfitting*, assim como no Teste 1, observa-se que os resultados obtidos no Teste 2 apresentam diferenças menos significativas e até cenários em que a diferença foi negativa. Esse comportamento sugere que os modelos do segundo teste estão menos ajustados aos dados de treinamento em comparação aos do Teste 1, o que indica uma maior capacidade de generalização, sobretudo, devido ao aumento do número de amostras e a natureza mais simples dos dados utilizados para o treinamento dos modelos (amostras geradas a partir de dados parametrizados e fronteiras de ACV mais restritas, do tipo *gate-to-gate*).

3.3.2.3 Teste 3

Esta subseção apresenta os principais resultados obtidos para as 11 categorias de impactos utilizando a base de dados compilada com processos do Ecoinvent, Agribalyse e Agrifootprint, segregando-os em testes preliminares e teste final. Por ser uma base de dados com mais de 1000 amostras foi utilizada a validação cruzada (*cross-validation*) ao invés da abordagem *leave-one-out*, sobretudo, devido à grande exigência computacional da segunda abordagem.

A validação cruzada é uma técnica utilizada para avaliar a capacidade de generalização de um modelo de aprendizado de máquina utilizando os dados da melhor forma possível. O método mais comum é o *k-fold cross-validation*, onde os dados são divididos em *k* subconjuntos

(folds). O modelo é treinado em $(k-1)$ folds e testado no fold restante, repetindo o processo k vezes. A média das métricas (R^2 e MSE) obtidas nessas iterações reflete o desempenho geral do modelo (Scikit-learn developers, 2025).

Essa abordagem evita o problema de *overfitting* causado pelo ajuste excessivo aos dados de teste, reduz a dependência da escolha aleatória dos conjuntos de treino e validação e evita o uso excessivo da exigência computacional (método LOO). Embora seja computacionalmente mais cara do que dividir os dados apenas uma vez (apenas treino e teste), a validação cruzada garante que todos os dados disponíveis sejam utilizados de forma eficiente para treinar e avaliar o modelo e permite uma análise mais completa dos dados (Scikit-learn developers, 2025).

A recomendação de 5 ou 10 *folds* é amplamente reconhecida na literatura e nas práticas de aprendizado de máquina. O uso de 10 *folds* tem um respaldo significativo na prática estatística, como descrito na referência de Kohavi (1995), que discute a validação cruzada em detalhe. Porém, por ser uma base de dados relativamente menor do que os modelos usuais de aprendizado de dados, conjunto de dados pequeno (< 5.000 amostras) e um conjunto de dados de tamanho médio (< 100.000 amostras) (Prenner & Robbes, 2022), optou-se por seguir com um número maior de *folds* ($k = 25$). Portanto, os resultados para o coeficiente de determinação e MSE são médios, sendo obtidos após a testagem das 25 diferentes combinações de fold para treino e teste.

Conforme detalhado na Seção 3.3.1.3, a base de dados do teste passou por um refinamento, partindo de 169 atributos na base inicial (v0) para uma base com 14 atributos em sua versão final (v5). Nesse sentido, o mesmo *pipeline* de treinamento dos modelos foi realizado para cada uma das versões da base, a fim de identificar potenciais perdas advindas da redução de dimensionalidade realizada.

Os hiperparâmetros foram definidos para cada um dos modelos utilizando o Optuna a partir de uma mesma pré-seleção de parâmetros nas seis versões da base de dados. O Optuna é uma abordagem de definição de hiperparâmetros um pouco diferente do *gridsearch* que foi utilizado nos Testes 1 e 2, pois estabelecer melhores resultados aos modelos a medida em que os testes são realizados através de um caminho de melhorias (Akiba et al., 2019).

No Optuna é possível definir a quantidade de interações de hiperparâmetros que serão feitas na busca da melhor combinação (Akiba et al., 2019). Em virtude do tempo de processamento e do volume de modelos a serem testados, optou-se por definir a quantidade de 20 combinações por modelo para cada uma das 11 categorias de impacto alvo. Logo, para rodar o *pipeline* para cada uma das bases (v0 a v5) foi necessário treinar cerca de 71.500 modelos com combinações diferentes de hiperparâmetros, isso porque foram executadas 20 combinações

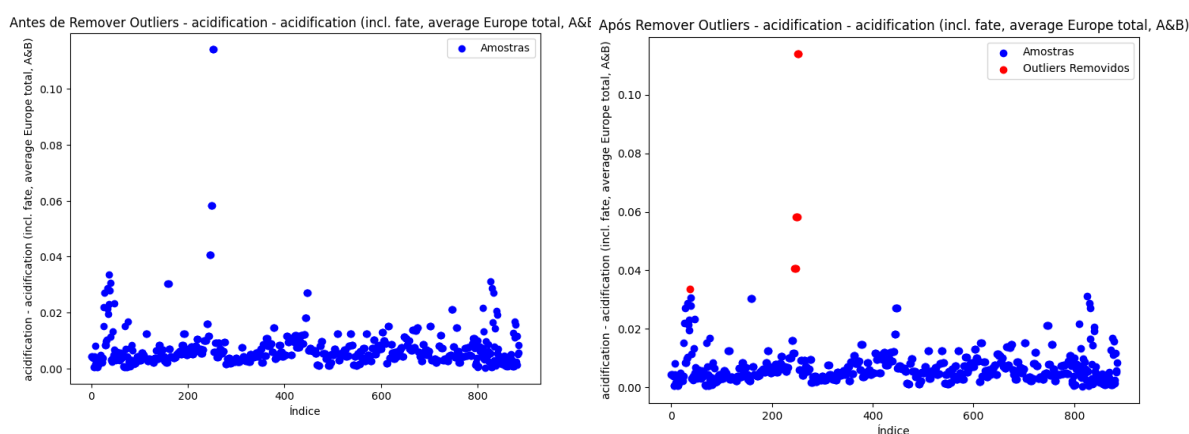
do Optuna para cada um dos 13 modelos de AM levando em consideração o *looping* contendo as 11 categorias de impacto em análise e o número de *folds* igual a 25.

Os tempos de execução dos códigos variaram de 7h52min para a base contendo 169 atributos até 3h40min para a base contendo apenas 14 atributos. Como mencionado anteriormente, o *pipeline* de execução dos códigos foi exatamente o mesmo e contou com as respectivas etapas:

1. Seleção da categoria de impacto (atributo-alvo) a ser avaliada.
2. Detecção e remoção dos *outliers* para a dada categoria de impacto (Figura 3.11).
3. Normalização dos dados (função *standscaler* do Scikit-learn) para o treinamento dos modelos.
4. Treinamento de cada um dos modelos e ajuste de parâmetros (biblioteca Optuna).
5. Escolha do melhor modelo com base no maior R^2 .
6. Verificação da diferença entre os valores reais e os valores preditos pelo melhor modelo.

Como saídas da execução do código é possível destacar gráficos e tabelas que refletem cada uma das etapas do *pipeline* de treinamento dos modelos e a mensuração de seus respectivos desempenhos. Seguem abaixo um exemplo, contendo as respectivas saídas para o treinamento da categoria “*acidification*” utilizando a base de dados com 30 atributos (v3).

Figura 3. 11 - Seleção da categoria *acidification* e detecção de *outliers*.



Fonte: autoria própria.

Os gráficos da figura acima destacam a seleção das amostras referente ao atributo-alvo para a categoria acidificação. Inicialmente, foram selecionadas 886 amostras, as quais passaram

pelo processo de remoção de *outliers*, sendo removidas 10 amostras e chegando ao final em 876 amostras para treinamento e testagem dos modelos. Em seguida, as 876 amostras foram divididas em treino (80% - 700 amostras) e teste (20% - 176 amostras) e executado o *pipeline* de treinamento para os 25 diferentes *folds*.

Após a divisão dos dados em conjuntos de treino e teste, procedeu-se para a análise de desempenho do processo de treinamento dos modelos a partir da avaliação em 25 diferentes *folds*. Dessa forma, a análise de desempenho dos modelos (Figura 3.12) apresenta os valores máximo, mínimo e médio dos coeficientes de determinação obtidos para as 25 diferentes repartições, bem como os respectivos tempos de execução e a seleção dos melhores hiperparâmetros após o teste em 20 diferentes combinações com o uso do Optuna.

Figura 3. 12 - Desempenho dos modelos para a categoria *acidification*.

					Hiperparâmetros
Modelo	R ² Mínimo	R ² Máximo	R ² Médio	Tempo de Treinamento	
0 Bayes	0.1963	0.9542	0.6835	81.0!	{'regressor__alpha_1': 0.004299899916041581, 'regressor__lambda_1': 0.009902959921067267}
1 KNN	0.2148	0.9992	0.7734	4.0:	{'regressor__n_neighbors': 11, 'regressor__weights': 'distance'}
2 LR	0.1075	0.9527	0.6752	0.2!	Parâmetros padrão
3 E_Net	0.2528	0.9455	0.6911	4.0:	{'regressor__alpha': 0.055439397934892, 'regressor__l1_ratio': 0.24800096938137775}
4 SGD	0.2210	0.9581	0.6883	4.4!	{'regressor__alpha': 0.00782276520409749, 'regressor__max_iter': 2264}
5 Kernel	0.1931	0.9541	0.6844	95.7:	{'regressor__alpha': 9.948326908468271, 'regressor__degree': 3}
6 DT	-0.7743	0.9872	0.6998	6.9!	{'regressor__max_depth': 20, 'regressor__min_samples_leaf': 4}
7 RF	-0.3440	0.9881	0.7971	1170.1!	{'regressor__n_estimators': 378, 'regressor__max_depth': 12}
8 GBM	0.2963	0.9973	0.8147	403.5!	{'regressor__n_estimators': 229, 'regressor__learning_rate': 0.12935345038419876}
9 XGBoost	-0.2074	0.9989	0.8028	98.2:	{'regressor__n_estimators': 204, 'regressor__max_depth': 4}
10 CatBoost	0.3938	0.9994	0.8601	1177.4!	{'regressor__iterations': 2034, 'regressor__learning_rate': 0.049676247697330776}
11 ANN	-0.1647	0.9885	0.6550	181.9:	{'regressor__learning_rate_init': 0.0868403223249573}
12 SVM	0.1562	0.9899	0.8120	21.3:	{'regressor__C': 40.41879956664856, 'regressor__gamma': 0.011260536961162443}

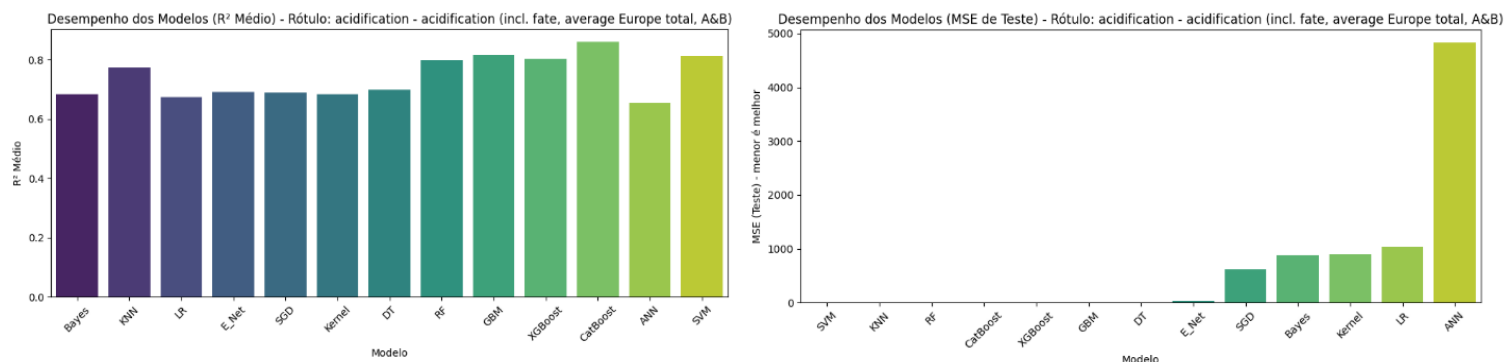
Fonte: autoria própria.

A análise dos resultados acima permite identificar os desempenhos obtidos em cada um dos modelos para a regressão da categoria alvo (acidificação). O modelo Catboost mostrou-se o de melhor desempenho, com R² médio de 0,8601, seguido pelo modelo SVM com R² médio de 0,8120. Os modelos de ANN e LR foram os de pior desempenho, com R² médio de 0,6550 e 0,6752, respectivamente.

Os resultados acima também possibilitam visualizar a seleção dos melhores hiperparâmetros por modelo e seus respectivos tempos de treinamento. O modelo Catboost apresentou o maior tempo de execução, exigindo cerca de 19 min (1177,48 segundos) de processamento computacional para chegar ao melhor modelo (R² médio de 0,8601, ou seja, o modelo explica 86,01% da variação dos dados em relação à variável alvo) com os respectivos parâmetros: “regressor__iterations: 2034” e “regressor__learning_rate: 0.04967624...”. Por outro lado, o modelo de regressão linear mostrou-se o de menor exigência computacional, levando apenas 0,28 segundos para chegar a melhor solução (R² médio de 0,6752) com os hiperparâmetros padrão do Scikit-learn.

Os *outputs* do código Python desenvolvido também possibilitaram visualizar os resultados para o coeficiente de determinação médio e MSE médio em formato gráfico, conforme Figura 3.13. De forma geral, os resultados permitem identificar que o modelo SVM foi de segundo melhor R^2 médio e o modelo com menor MSE associado.

Figura 3. 13 - Desempenho dos modelos (R^2 e MSE teste) para a categoria *acidification*.

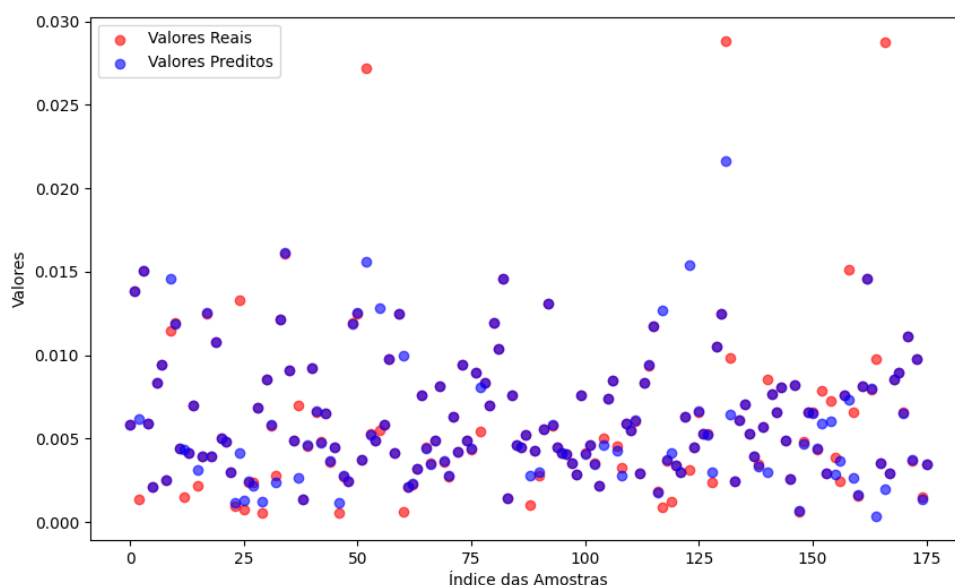


Fonte: autoria própria.

Por outro lado, o modelo baseado em redes neurais mostrou-se como o de pior R^2 médio e com o maior erro quadrático médio associado, indicando que o modelo em questão não foi capaz de aprender com o conjunto de dados disponíveis em comparação aos demais modelos testados.

Por fim, o código desenvolvido também permitiu observar as previsões realizadas (círculos em azul) pelo modelo de melhor desempenho em comparação aos valores reais das amostras de treino (círculos em vermelho), possibilitando analisar de forma visual os principais erros de previsão e as respectivas amostras, conforme Figura 3.14.

Figura 3. 14 - Testagem do melhor modelo (valor real vs valor predito).



Fonte: autoria própria.

É importante destacar que, inicialmente, o processo de detecção e remoção de *outliers* para o Teste 3 foi feito utilizando a funcionalidade Z-score do Scikit-learn. O Z-score é uma métrica que indica quantos desvios padrão um valor está acima ou abaixo da média (Iglewicz & Hoaglin, 1993; Khan & McCormick, 2014). Valores com Z-scores altos (positivos ou negativos) são considerados *outliers*, pois estão distantes da média (Iglewicz & Hoaglin, 1993; Khan & McCormick, 2014). Por exemplo, um Z-score de 3 indica que o valor está 3 desvios padrão acima da média, o que é frequentemente usado como um limiar para identificar *outliers*, portanto, foi utilizado como padrão inicial para os testes preliminares.

Testes preliminares (6 versões da base)

De forma similar ao *pipeline* exemplificado acima para a categoria de acidificação da base v3 (30 atributos), todas as etapas foram replicadas para cada um dos 11 atributos-alvo para as seis diferentes versões da base de dados (v0 com 169 atributos até v5 com 14 atributos). Assim, os melhores resultados identificados para um dos testes foram compilados na Tabela 3.14, a fim de facilitar a observação do desempenho para as diferentes categorias de impacto em virtude do treinamento com número diferentes de atributo.

Tabela 3. 14 - Resultados preliminares para o teste 3.

	Outliers removidos	169 atributos		97 atributos		52 atributos		30 atributos		27 atributos		14 atributos	
		R ²	Modelo	R ²	Modelo	R ²	Modelo	R ²	Modelo	R ²	Modelo	R ²	Modelo
AC	10	0,7288	DT	0,8078	CATBOOST	0,8231	CATBOOST	0,8601	CATBOOST	0,8385	CATBOOST	0,8468	CATBOOST
CC	24	0,8583	XGBOOST	0,8424	XGBOOST	0,8399	GBM	0,8370	XGBOOST	0,8254	KNN	0,8105	GBM
EF	16	0,5809	SVM	0,4549	SVM	0,5472	SVM	0,5342	SVM	0,5253	SVM	0,5672	KNN
EM	12	0,8745	XGBOOST	0,9119	GBM	0,8352	RF	0,8434	GBM	0,8314	RF	0,8351	RF
ET	18	0,8104	XGBOOST	0,8476	XGBOOST	0,7912	KNN	0,8343	KNN	0,8196	KNN	0,8552	KNN
ER	1	0,8861	GBM	0,8968	XGBOOST	0,9096	GBM	0,8894	GBM	0,8743	GBM	0,8799	KNN
EU	11	0,9032	GBM	0,9295	GBM	0,9077	GBM	0,9000	XGBOOST	0,9113	GBM	0,8987	GBM
HT	17	0,7394	GBM	0,7107	XGBOOST	0,6421	KNN	0,7025	KNN	0,7318	SVM	0,7413	SVM
MR	11	0,8543	GBM	0,8960	GBM	0,7572	SVM	0,7518	SVM	0,7041	SVM	0,7581	KNN
OD	10	0,8955	XGBOOST	0,8698	E_NET	0,8423	SVM	0,8274	SVM	0,8278	SVM	0,8144	SVM
PO	18	0,7625	RF	0,7371	DT	0,7693	RF	0,7782	RF	0,7734	KNN	0,5933	SVM
Med	13,5	0,8085		0,8095		0,7877		0,7962		0,7875		0,7818	

* Legendas: Med = média.

Categorias de impacto: AC: acidification; CC: climate change; EF: ecotoxicity: freshwater; EM: ecotoxicity: marine; ET: ecotoxicity: terrestrial; ER: energy resources: non-renewable; EU: eutrophication; HT: human toxicity; MR: material resources: metals/minerals; OD: ozone depletion; PO: photochemical oxidant formation.

Modelos: KNN: K - Nearest Neighbors; DT: decision tree; RF: random forest; GBM: Gradient Boosting Machine; XGBoost: eXtreme Boosting Machine; SVM: Support vector machine; E_net: ElasticNet; CatBoost: Category Boosting.

Fonte: autoria própria.

A análise dos resultados obtidos para a regressão revela um desempenho variável dos modelos conforme a quantidade de atributos utilizados. De maneira geral, foi observado que os testes com maior número de atributos (169 e 97) apresentam médias de R^2 superiores (0.8085 e 0.8095, respectivamente), sugerindo que a inclusão de mais informações pode contribuir para um ajuste mais preciso.

No entanto, à medida que os atributos são reduzidos, algumas variáveis-alvo mantêm um desempenho relativamente estável e/ou até melhoram (*acidification* e *ecotoxicity terrestrial*), enquanto outras sofrem quedas significativas (*photochemical oxidation formation*). Modelos como XGBoost, Gradient Boosting Machine (GBM) e CatBoost destacam-se como os mais consistentes, obtendo os melhores coeficientes de determinação em diversas situações.

Dentre os melhores desempenhos individuais, o CatBoost se destaca especialmente para a variável *acidification* (AC), alcançando um R^2 de 0.8601 com apenas 30 atributos, um dos maiores valores da análise. O GBM e o XGBoost mostraram ser capazes de prever potenciais impactos para as variáveis *energy resources: non-renewable* (ER); *eutrophication* (EU) e *material resources: metals/minerals* (MR), atingindo coeficientes acima de 0.90 nos cenários para as bases com 97 e 169 atributos.

Já para a variável *ecotoxicity terrestrial* (ET), o modelo KNN apresentou semelhante capacidade, especialmente nos cenários de menor dimensionalidade, como nos testes com 27 e 14 atributos, onde atingiu valores de 0.8196 e 0.8552, respectivamente. A persistência de bons resultados nesses casos sugere que a redução de atributos pode beneficiar alguns modelos ao minimizar o risco de *overfitting* e ainda sugere a possibilidade de utilizar modelos de AM com natureza de funcionamento menos complexas.

Entretanto, algumas variáveis-alvo mostraram-se mais desafiadores nos testes preliminares. A análise da variável *ecotoxicity: freshwater* “EF”, por exemplo, ressalta desempenhos inferiores em relação às demais, com valores de R^2 variando entre 0.4549 e 0.5809, sendo que o SVM domina na maioria dos cenários. Isso pode indicar que essa variável possui padrões mais difíceis de modelar ou que a escolha dos atributos não captura adequadamente suas relações.

Já a variável oxidação fotoquímica (PO) apresenta um comportamento peculiar, com oscilações dos valores de R^2 ao longo dos diferentes testes e com uma queda significativa para 0.5933 no cenário de 14 atributos, onde o SVM é o modelo predominante. Essa queda permite confirmar que informações acerca do tipo de cultura trabalhada são importantes para os modelos predizerem valores relacionados a oxidação fotoquímica. Portanto, ao suprimir as

informações do tipo de cultura (base v4 para a base v5) notou-se uma queda de cerca de 23% de desempenho dos modelos no tocante ao coeficiente de determinação.

A análise das médias dos coeficientes de determinação sugere que a escolha do modelo é tão importante quanto a seleção de atributos. Apesar da redução progressiva da dimensionalidade, os modelos GBM, XGBoost e CatBoost mantêm um desempenho sólido na maioria dos casos, enquanto SVM e KNN se destacam para variáveis específicas. A partir desses resultados, pode-se concluir que não há uma única configuração ótima para todos os alvos, mas sim um equilíbrio entre a quantidade de atributos e o modelo utilizado.

Além disso, os testes preliminares permitiram observar do comportamento dos modelos de aprendizado de máquina. Modelos como redes neurais artificiais (ANN), gradiente estocástico (SGD) e regressão Kernel (KRR) demonstraram sensibilidade à escolha dos hiperparâmetros, frequentemente apresentando desempenhos negativos no coeficiente de determinação (R^2). De forma semelhante, o mesmo padrão de desempenho foi identificado nos modelos de regressão linear (LR) e regressão bayesiana, os quais podem ter exibido tais resultados em função da simplicidade estrutural desses modelos, que se mostraram muitas vezes incapazes de capturar a complexidade dos dados e estabelecer relações sofisticadas entre as variáveis de entrada e saída (categorias de impacto).

Assim, a escolha final do modelo dependerá da necessidade de interpretar os resultados e do compromisso entre complexidade e desempenho. Dessa forma, por ser um código que pretende ser utilizado por usuários sem conhecimento prévio em avaliação do ciclo de vida (ACV) e aprendizado de máquina (AM), optou-se por aprofundar as análises na base de dados com menor número de atributos (v5 – 14 atributos), indo em consonância aos achados de Salla et al (2025) na qual os estudos de ACV+AM com foco em estudos agrícolas utilizavam em sua maioria até 15 atributos para treinamento dos modelos de AM.

Testes final (base v5 -14 atributos)

Nesse sentido, foi executado um novo *pipeline* de treinamento usando a base de dados v5 com 14 atributos, sendo eles:

- Amônia e afins (kg),
- Calcário e gesso (kg),
- Emissões advindas de pesticidas (kg),
- Energia proveniente da biomassa (MJ),
- Energia elétrica consumida (kwh),
- Esterco animal (kg),

- Esterco verde (kg),
- Quantidade usada de fungicida, herbicida e inseticida (kg),
- Quantidade usada de macronutrientes (kg): nitrogênio, potássio, fósforo, cálcio, magnésio e enxofre,
- Quantidade usada de micronutrientes (kg): boro, cloro, molibdênio, cobre, ferro, zinco e manganês,
- Ocupação total de terra ($m^2 \cdot \text{ano}$),
- Quantidade de semente usada (kg),
- Transformação total de terra (m^2),
- Ureia (kg).

Para isso, foram feitas algumas mudanças em comparação aos testes preliminares a fim de melhorar o desempenho dos modelos, entre elas:

- Inclusão de mais *folds* para evitar a presença de *underfitting* e *overfitting* nos modelos (de 25 para 50 *folds*);
- Aumento do número de testes realizados para a escolha dos melhores hiperparâmetros (20 para 30 combinações testadas por modelos) a partir do Optuna;
- Inclusão de um maior range de hiperparâmetros a serem testados e escolhidos usando o Optuna;
- Substituição da funcionalidade Z-Score ($Z=3$) do Scikit-learn pelo IQR (*Interquartile Range*), com objetivo de deixar amostras mais padronizadas para a execução dos testes.

A realização do código que antes levava cerca de 3h40 min levou 12h48 min para sua execução completa, sendo gerados 214.500 modelos ao invés dos 71.500 modelos (testes preliminares). Após a execução de todo o *pipeline*, os resultados foram compilados na Tabela 3.15.

Tabela 3. 15 - Resultados para o Teste 3.

	Outliers removidos	R ²	MSE_treino	MSE_teste	Tempo (s)	Modelo
AC	45	0,7967	0,0342	0,1591	2342,8	RF
CC	87	0,8514	0,0002	0,1559	260,9	XGBoost
EF	130	0,7914	0,0004	0,2204	682,2	GBM
EM	140	0,8981	0,0001	0,0916	212,3	XGBoost
ET	118	0,8361	0,0001	0,2045	8,5	KNN
ER	202	0,8773	0,0001	0,0001	142,6	XGBoost
EU	45	0,9022	0,0026	0,0876	49,5	CatBoost
HT	90	0,8801	0,0152	0,0936	3147,8	RF
MR	187	0,9986	0,0003	0,0009	672,2	GBM
OD	195	0,9675	0,0016	0,0072	1216,9	GBM
PO	118	0,8819	0,0033	0,1405	63,5	CatBoost
Média	123,3	0,8801	0,0052	0,1055	799,9	

* *Legendas:*

Categorias de impacto: AC: acidification; CC: climate change; EF: ecotoxicity: freshwater; EM: ecotoxicity: marine; ET: ecotoxicity: terrestrial; ER: energy resources: non-renewable; EU: eutrophication; HT: human toxicity; MR: material resources: metals/minerals; OD: ozone depletion; PO: photochemical oxidant formation.

Modelos: KNN: K - Nearest Neighbors; RF: random forest; GBM: Gradient Boosting Machine; XGBoost: eXtreme Boosting Machine; CatBoost: Category Boosting.

Fonte: autoria própria.

Os resultados obtidos evidenciam um desempenho robusto dos modelos XGBoost, CatBoost e GBM, que consistentemente apresentaram altos coeficientes de determinação (R²), indicando sua capacidade de capturar relações complexas nos dados, assim como observado nos testes preliminares.

Entre eles, destaca-se o modelo GBM na predição de MR, que obteve um R² de 0,9986, sugerindo um ajuste quase perfeito. Contudo, esse valor pode indicar um possível *overfitting*, dado que o erro médio quadrático (MSE) de teste é consideravelmente menor que o de treino. Essa tendência também é observada na predição de OD, onde o GBM alcançou um R² de 0,9675, reforçando sua capacidade preditiva, mas levantando questionamentos sobre sua generalização para novos dados.

A questão do *overfitting* também se faz presente no modelo Random Forest (RF) aplicado às predições de AC e de HT, ambas com R² elevados (0,7967 e 0,8801, respectivamente), mas com MSE de teste relativamente mais altos em comparação a outros modelos. O tempo de execução desses modelos, especialmente para HT (3147,8 segundos),

sugere um custo computacional elevado, o que pode impactar sua viabilidade em cenários onde eficiência temporal é um fator crítico. Já no caso do KNN, utilizado na predição ET, observa-se um desempenho moderado (R^2 de 0,8361), porém com um tempo de execução significativamente inferior (8,5 segundos), o que o torna uma alternativa interessante em situações que demandam rapidez, ainda que em detrimento de uma leve perda de precisão.

A análise de *underfitting* pode ser observada em alguns casos, especialmente quando há uma discrepância entre os MSE de treino e teste associada a um R^2 relativamente baixo. O modelo GBM na predição de EF (R^2 de 0,7914) e o RF na predição de AC (R^2 de 0,7967) indicam que o modelo pode não estar capturando toda a variabilidade dos dados, resultando em uma performance inferior.

Além disso, a predição de CC, realizada com XGBoost, apresentou um R^2 alto (0,8514), mas um MSE de treino extremamente reduzido (0,0002), o que pode sugerir que o modelo está se ajustando demasiadamente aos dados de treino sem garantir uma generalização eficiente, sugerindo possibilidades de *overfitting*.

Os modelos mais equilibrados entre desempenho e eficiência computacional foram o XGBoost aplicado às predições de EM e ER e o CatBoost na predição de EU. Com R^2 superiores a 0,87 e tempos de execução relativamente baixos, esses modelos se destacam como boas opções para aplicações que exigem tanto precisão quanto eficiência. Em especial, a predição ER com XGBoost obteve um R^2 de 0,8773 e MSE de treino e teste idênticos (0,0001), sugerindo um ajuste estável e generalizável. Da mesma forma, o CatBoost na predição EU alcançou um R^2 de 0,9022, o maior entre os modelos analisados, com um tempo de execução de apenas 49,5 segundos, indicando um excelente equilíbrio entre precisão e eficiência.

De modo geral, é válido destacar que a escolha do modelo ideal dependerá do contexto da aplicação. O GBM demonstrou um desempenho excepcional em predições específicas, mas levanta preocupações quanto ao *overfitting*. O XGBoost e o CatBoost foram os mais consistentes, entregando previsões precisas sem tempos de execução excessivamente longos. Já o KNN, embora com um desempenho inferior, se mostrou uma alternativa viável para cenários onde a rapidez é um fator determinante. Assim, considerando o balanço entre desempenho e custo computacional, os modelos XGBoost e CatBoost emergem como as opções mais adequadas para aplicações que exigem tanto precisão quanto eficiência operacional.

Comparação teste preliminar vs teste final

Em termos absolutos, os resultados do teste final foram superiores ao do teste preliminar para base v5. A média geral do coeficiente de determinação (R^2) variou de 0,7818 no teste

inicial para 0,8801 no último teste. Em termos percentuais absolutos, a média do R^2 aumentou 12,57% entre o primeiro e o segundo teste, indicando uma melhoria na capacidade preditiva dos modelos através do aumento no número de *k-folds* (de 25 para 50) e no número de iterações da busca por hiperparâmetros no Optuna (de 20 para 30).

Essa variação sugere que a maior quantidade de divisões dos dados para validação cruzada (*k-folds*) reduziu a variabilidade das estimativas de desempenho, enquanto a ampliação da busca de hiperparâmetros permitiu encontrar configurações mais otimizadas, resultando em um ajuste mais preciso dos modelos.

No teste preliminar, os modelos foram avaliados utilizando 25 *folds* na validação cruzada e 20 iterações de busca com Optuna para otimização de hiperparâmetros. Os resultados indicaram um ajuste equilibrado, com os modelos CatBoost e GBM apresentando os melhores desempenhos. O KNN e o SVM, por outro lado, exibiram maior variação nos valores de R^2 , o que sugere que sua performance pode ser mais sensível à natureza dos dados.

A ausência do XGBoost limita comparações diretas, mas os modelos baseados em árvores, como Random Forest e GBM, demonstraram boa generalização sem valores extremos de R^2 . Essa configuração de teste permitiu obter modelos ajustados sem um risco excessivo de sobreajuste dos modelos (*overfitting*), garantindo que as métricas de desempenho fossem mais confiáveis.

No teste final, a validação cruzada foi mais rigorosa, com 50 *folds*, e a busca por hiperparâmetros foi ampliada para 30 iterações com Optuna. Isso resultou em desempenhos superiores para a maioria dos modelos, com destaque para o GBM (0,9986) e o XGBoost (0,8981), que alcançaram valores de R^2 maiores. No entanto, esses números podem indicar a existência de *overfitting*, especialmente no caso do GBM, onde a discrepância entre o MSE de treino e de teste sugere um ajuste excessivo aos dados de treinamento.

O CatBoost manteve um bom desempenho, confirmando sua robustez em diferentes cenários. O KNN, apesar de alguma melhoria, permaneceu atrás dos modelos baseados em árvores. O aumento na quantidade de *folds* e iterações do Optuna pode ter ajudado a encontrar hiperparâmetros mais ajustados, mas também pode ter exacerbado o problema de sobreajuste em alguns modelos.

Comparando os dois testes, percebe-se que o aumento no número de *folds* e de iterações da busca por hiperparâmetros trouxe melhorias em termos de R^2 , mas potencialmente comprometeu a capacidade de generalização de alguns modelos, como o GBM. Enquanto no teste preliminar houve um equilíbrio maior entre desempenho e robustez, no teste final, a busca

intensiva por hiperparâmetros pode ter levado alguns modelos a se ajustarem excessivamente aos dados de treino.

Dessa forma, a escolha do melhor teste dependerá do objetivo final: se a prioridade for um ajuste equilibrado e generalizável, o primeiro teste é mais adequado (utilizando menor volume de teste); por outro lado, se a intenção for obter o melhor desempenho absoluto, mesmo com o risco de *overfitting*, o segundo teste pode ser mais vantajoso.

3.4 CONCLUSÕES

A pesquisa explorou o desempenho de diferentes modelos de Aprendizado de Máquina (AM) aplicados a bases de dados de Avaliação do Ciclo de Vida (ACV) para a predição de potenciais impactos ambientais. Os testes realizados demonstraram que a qualidade e o tamanho das bases de dados são fatores de atenção para a obtenção de resultados confiáveis, já que bases reduzidas podem levar a problemas de *overfitting* ou *underfitting*, comprometendo a capacidade preditiva dos modelos e limitando sua aplicação em cenários reais.

Os três testes conduzidos nesta pesquisa permitiram avaliar como diferentes volumes e estruturas de dados impactam os resultados dos modelos. O primeiro teste, realizado com apenas 15 amostras nacionais de milho, evidenciou dificuldades relacionadas ao pequeno volume de dados, resultando em modelos que se ajustaram excessivamente aos dados de treinamento, mas falharam em generalizar para novas observações.

O segundo teste, realizado com *datasets* nacionais parametrizados, mostrou-se uma alternativa viável para facilitar o processo de geração de amostras. Nesse sentido, alterar os valores dos parâmetros independentes utilizando a funcionalidade “Projects” disponível no OpenLCA permite a geração de potenciais impactos de uma só vez, o que permitiu gerar amostras facilmente. Em termos gerais, o *pipeline* de treinamento dos modelos para o segundo teste apresentou melhorias na capacidade de generalização, embora ainda houvesse desafios quanto ao potencial risco de *overfitting*.

O terceiro teste, que integrou três bases de dados distintas (Agribalyse, Agrifootprint e Ecoinvent) e totalizou 1.219 amostras, representou o maior desafio computacional e metodológico do estudo. No teste em questão, foi necessário o uso de técnicas avançadas de manipulação de dados, como a exportação de dados via OpenLCA utilizando o "IPC server" e o emprego das bibliotecas Python *olca-ipc* 2.0.3 e *olca-schema* 2.2.1. Além disso, houve a necessidade de padronização de identificadores de arquivos e atributos, garantindo a

consistência dos dados e permitindo a integração das bases heterogêneas em atributos que fossem minimamente comuns.

Um dos principais desafios enfrentados no terceiro teste foi a harmonização dos 570 atributos distintos presentes nas bases originais. Esse processo envolveu diversas etapas de transformação e redução, resultando em uma base inicial contendo 169 atributos distintos, posteriormente refinada para 14 atributos e 886 amostras de interesse. Tal refinamento foi essencial para garantir a compatibilidade dos dados, a compilação das amostras e possibilitar a aplicação eficaz dos modelos de AM.

Os resultados mostraram que aumentar o conjunto de amostras melhora significativamente a capacidade preditiva dos modelos. No entanto, é fundamental considerar o custo computacional envolvido, especialmente no ajuste de hiperparâmetros. O aumento no número de iterações da busca com Optuna e a utilização de k-folds mais elevados impactaram positivamente os resultados, mas também elevaram os tempos de processamento, tornando a execução computacionalmente mais demorada e, por conseguinte, consumindo mais energia e de demais recursos.

Os testes demonstraram que modelos baseados em gradient boosting, como XGBoost, GBM e CatBoost, alcançaram os melhores desempenhos, apresentando altos coeficientes de determinação (R^2) e erros relativamente baixos. No entanto, é necessário um cuidado especial para evitar sobreajustes, como observado no GBM, que apresentou divergências nos valores de erro de treino e teste.

Por outro lado, modelos como k-vizinhos mais próximos, regressão Linear e regressão bayesiana demonstraram-se mais restritivos, apresentando dificuldades para generalizar e aprender eficientemente com os dados existentes nos três testes. Os modelos baseados em florestas, como Random Forest (RF) e Decision Tree (DT), obtiveram resultados expressivos, apresentando melhores resultados para algumas categorias, contudo, exigem um custo computacional relativamente alto devido à sua estrutura de funcionamento.

Além disso, modelos como Kernel Ridge Regression (KRR), Stochastic Gradient Descent (SGD), Elastic Net (E_net) e redes neurais artificiais (ANN) mostraram-se sensíveis ao ajuste de hiperparâmetros, frequentemente resultando em valores negativos de R^2 . Portanto, um cuidado especial se faz necessário ao utilizar tais modelos, sobretudo, com o objetivo de garantir um ajuste prévio para evitar resultados inconsistentes.

De forma geral, foi possível identificar que o processo de escolha do melhor teste e modelo depende do equilíbrio entre capacidade preditiva, generalização e custo computacional. Logo, não há uma regra para a escolha do melhor modelo com base na categoria de impacto de

interesse ou melhor quantidade de atributo por tipo de fronteira do sistema (*cradle-to-gate* ou *gate-to-gate*) ou melhor modelo por setor do estudo, o que se faz necessário é entender as limitações de cada conjunto de dados e processar os dados de ACV da melhor forma possível a fim de potencializar o treinamento dos modelos.

Além disso, bases de dados mais robustas são preferíveis, pois permitem uma melhor aprendizagem dos modelos e reduzem os riscos de *overfitting* e *underfitting*, desde que o aumento da complexidade computacional seja devidamente gerenciado (ajustes dos parâmetros). Da mesma forma, bases de dados mais extensas oferecem maior liberdade aos usuários para processar as informações de maneira mais eficiente, além de mitigar o receio de remover amostras (outliers, dados ruidosos, etc.), que frequentemente surge quando se trabalha com bases reduzidas.

A consideração dos fatores regionais e locais também é um fator fundamental para a avaliação dos impactos ambientais e para o rendimento das culturas agrícolas, pois diferentes condições geográficas e climáticas podem influenciar de maneira significativa os resultados. No contexto do Teste 3, que utilizou dados globais, pode ter ocorrido uma limitação na precisão dos resultados devido à falta de parametrização específica para as variáveis locais, como a erosividade do solo, profundidade da raiz, precipitação, clima, radiação solar, umidade, entre outras.

Esses fatores têm grande impacto na produtividade das culturas, já que as condições do solo e do clima podem variar drasticamente de uma região para outra e afetam diretamente nos fluxos elementares que são emitidos aos diferentes compartimentos ambientais (ar, solo e água). Assim, é crucial o desenvolvimento de dados que sejam parametrizados de forma a permitir o intercâmbio de variáveis regionais de forma fácil e rápida, permitindo o rápido ajuste dos *datasets* às particularidades locais e proporcionando uma avaliação mais precisa e relevante para as condições específicas de cada região agrícola e facilitando o processo de geração de amostras mais condizentes para o treinamento de modelos de AM.

Em conclusão, os resultados reforçam a importância de bases de dados amplas e bem estruturadas para a aplicação eficaz de modelos de AM em ACV. O avanço da pesquisa nessa área depende tanto do desenvolvimento de modelos mais sofisticados quanto da construção de bases de dados mais abrangentes e padronizadas, permitindo análises mais precisas e confiáveis sobre os impactos ambientais de diferentes processos produtivos e em diferentes setores da economia.

APÊNDICE – 3A

O tópico em questão objetivou apresentar um breve detalhamento dos modelos de AM a serem usados em cada dos três testes.

1. *Bayesian regression*

O Teorema de Bayes é utilizado para calcular probabilidades condicionais, possuindo um papel fundamental para as abordagens Bayesianas de epistemologia, estatística e lógica indutiva. O insight central do teorema volta-se para a ideia de que uma hipótese é confirmada por qualquer conjunto de dados que a torne hipótese mais provável com base na sua veracidade (Joyce, 2003).

O algoritmo Naive Bayes é uma técnica probabilística de aprendizado de máquina baseada no Teorema de Bayes. Para isso, o mesmo faz uma suposição "ingênua" de independência entre as características e os rótulos das classes, assim, é comumente usado para tarefas de classificação e, em alguns casos, de regressão. O modelo Naive Bayes é valorizado por sua capacidade de lidar com grandes conjuntos de dados de maneira eficiente e por sua eficácia em cenários onde a independência das características é razoavelmente válida (Ambika, 2023).

A regressão linear bayesiana ou *Bayesian regresion* é um método de modelagem de regressão que utiliza a abordagem Bayesiana para a estimativa de parâmetros. Na abordagem Bayesiana, há três componentes principais: a distribuição a priori, a distribuição de verossimilhança e a distribuição posterior. Nesse sentido, a estimativa dos parâmetros (atributo-alvo) é realizada processando a distribuição posterior a partir da multiplicação da distribuição a priori pela verossimilhança (Permai & Tanty, 2018).

2. **K-vizinhos mais próximos**

O algoritmo *k-Nearest Neighbor* ou K-vizinhos mais próximos (KNN) possui um funcionamento reativamente simples e pode ser usado tanto para resolver problemas de classificação como de regressão. O KNN tem sido utilizado na estimativa estatística e no reconhecimento de padrões desde o início da década 1970 como uma técnica não paramétrica (Muhajir, 2019).

É considerado um algoritmo do tipo preguiçoso (“*lazy*”), visto que o mesmo não usa os pontos obtidos nos dados de treinamento para fazer qualquer generalização; assim, o que ele faz é apenas manter os dados obtidos na etapa de treinamento para apoiar nas predições

realizadas durante a fase de teste (Song et al., 2017). De modo geral, o KNN armazena todos os casos disponíveis e prevê o alvo numérico com base em uma medida de similaridade (por exemplo, funções de distância) (Muhajir, 2019).

3. Regressão Linear

Realizar uma análise de regressão linear é provavelmente a maneira mais simples e popular de medir as relações entre variáveis preditoras e de resposta contínuas. De modo geral, o método assume que as relações entre as variáveis preditoras e alvo são lineares, indicando que uma unidade constante de mudança em uma variável irá gerar uma unidade constante de mudança na outra. Esse modelo de funcionamento torna a regressão linear a escolha ideal para análises com pequenas amostras e para amostras fáceis de interpretar e entender (Hope, 2020).

Análises de regressão são amplamente utilizadas para previsão e predição, além de serem usadas para determinar relações causais entre um conjunto de variáveis independentes (atributos) e dependentes (atributos-alvo) (Maulud & Abdulazeez, 2020). De acordo com Maulud & Abdulazeez (2020), existem diferentes tipos de regressão, incluindo a regressão linear simples, que envolve uma única variável independente; a regressão linear múltipla ou *multiple linear regression* (MLR), que utiliza várias variáveis independentes para prever um resultado e até a regressão polinomial, que é uma forma especial de MLR que modela a relação entre variáveis através de uma equação polinomial, permitindo capturar interações curvilíneas entre as variáveis.

4. Árvore de decisão

A análise de árvores de decisão ou *decision tree* (DTs) é uma abordagem de "dividir e conquistar" usada para resolver problemas de classificação e de regressão (Myles et al., 2004) por meio de um método de aprendizado supervisionado não-paramétrico (Scikitlearn, 2024).

O objetivo deste método é criar um modelo que prediga o valor de uma variável alvo aprendendo regras de decisão inferidas a partir das características dos dados, sendo assim, úteis para descobrir características e padrões importantes em grandes bancos de dados para modelagem preditiva (Myles et al., 2004; Scikitlearn, 2024). As árvores de decisão operam por meio de regras de decisão do tipo *if-else*, portanto, quanto mais profunda a árvore, mais complexas são as regras de decisão e mais ajustado é o modelo (Scikitlearn, 2024).

5. Florestas aleatórias

O algoritmo de florestas aleatórias ou *random forest* (RF) é um algoritmo que pode ser aplicado tanto em problemas de classificação quanto de regressão devido a sua flexibilidade e facilidade de uso, criado por Breiman em 2001 (Breiman, 2001; IBM, 2023). RF é um tipo de algoritmo que combina os resultados de múltiplas árvores de decisão para alcançar um único resultado, logo, seu funcionamento se relaciona a uma combinação de preditores de árvore, de modo que cada árvore depende dos valores de um vetor aleatório amostrado independentemente e com a mesma distribuição para todas as árvores na floresta (Breiman, 2001).

O algoritmo RF é composto por uma coleção de árvores de decisão, onde cada árvore é formada a partir de uma amostra de dados retirada com reposição do conjunto de treinamento, com o objetivo de melhorar a precisão e reduzir o risco de *overfitting* (sobreajuste dos modelos aos atributos). Para tarefas de regressão, as previsões das árvores individuais são combinadas pela média, enquanto para tarefas de classificação, a classe prevista é determinada pelo voto da maioria (IBM, 2023).

6. *Gradient boosting machine*

O método de *Gradient Boosting machine* (GBM) foi introduzido por Jerome Friedman em meados dos anos 2000 (Friedman, 2001). O algoritmo em questão conta com técnicas poderosas de aprendizado de máquina do tipo supervisionado e pode ser utilizado em tarefas de classificação e regressão, especialmente apropriados para dados menos limpos. (Friedman, 2001; Natekin & Knoll, 2013)

Nesse tipo de algoritmo, as estimativas ou aproximações de funções são vistas sob a perspectiva da otimização numérica no espaço das funções, em vez de no espaço dos parâmetros, na qual é estabelecida uma conexão entre expansões aditivas etapa a etapa e a minimização por descida do gradiente (Friedman, 2001).

Eles são altamente personalizáveis, podendo ser ajustadas para diferentes funções de perda. A ideia principal por trás deste algoritmo é construir novos aprendizes-base que estejam correlacionados com o gradiente negativo da função de perda de forma maximizada. De modo geral, a escolha da função de perda depende do pesquisador e do objetivo do trabalho, havendo uma ampla variedade de funções de perda já existentes e a possibilidade de implementar funções específicas para a tarefa em questão. (Natekin & Knoll, 2013).

7. *Extreme gradient boosting*

O XGBoost (*Extreme Gradient Boosting*) é uma ferramenta avançada para aprendizado supervisionado, projetada para melhorar a performance dos métodos de *boosting* tradicionais.

O método foi desenvolvido em meados de 2006 por Tianqi Chen (Chen & Guestrin, 2016). De acordo com (Chen & Guestrin, 2016), o funcionamento do método consiste em utilizar árvores de decisão como modelos base, ajustando-as sequencialmente para melhorar as previsões, sendo assim, o mesmo pode ser considerado como um método aprimorado o algoritmo GBM tradicional.

Para que isso seja possível, o XGBoost conta com técnicas avançadas de otimização, incluindo regularização para evitar o *overfitting*, o que o torna eficiente em termos de tempo de processamento e memória utilizada. Nesse sentido, ele acaba se destacando por sua capacidade de processamento distribuído e *out-of-core*, permitindo a análise de grandes volumes de dados que excedem a capacidade da memória disponível (Chen & Guestrin, 2016).

8. Redes neurais artificiais

O primeiro modelo de rede neural artificial ou *artificial neural network* (ANN) foi introduzido em 1958 por Rosenblatt a partir uma forma básica de rede neural com uma única camada de neurônios, conhecida como *perceptron* (Rosenblatt, 1958).

As ANNs são inspiradas no funcionamento do sistema nervoso biológico humano, na qual os sistemas computacionais são projetados para simular o comportamento do cérebro humano na realização de tarefas de reconhecimento, generalização e aprendizado (Rosenblatt, 1958). Nesse sentido, o *perceptron* é composto por unidades sensoriais (ou "neurônios") que recebem sinais de entrada, processam esses sinais e produzem uma resposta (atributo-alvo/outputs) (Rosenblatt, 1958).

Após o desenvolvimento do *perceptron*, as redes neurais passaram por uma série de avanços a fim de expandirem suas capacidades e aplicações, sendo possíveis de serem usadas em modelos de classificação e regressão. Nesse sentido, é possível destacar as seguintes aplicações:

- i. ***Perceptron* Multicamadas ou *multi-layer perceptron* (MLP):** extensão do *perceptron* simples, com múltiplas camadas de neurônios (camadas de entrada, camadas ocultas e camada de saída).
- ii. **Algoritmo de Retropropagação ou *Backpropagation* (BP):** algoritmo que calcula o gradiente do erro da rede com relação aos pesos através da propagação do erro de volta pela rede, ajustando os pesos de maneira a minimizar o erro, possibilitando o uso de redes mais profundas (Rumelhart et al., 1986):
- iii. **Redes Neurais Convolucionais (CNNs):** algoritmos que utilizam camadas convolucionais para capturar características locais e camadas de *pooling* para

reduzir a dimensionalidade, sendo muito usadas para tarefas de reconhecimento de imagem e processamento de vídeo (Lecun et al., 1998).

- iv. **Aprendizado Profundo (*Deep Learning*):** o termo "aprendizado profundo" refere-se ao uso de redes neurais profundas, ou seja, modelos com inúmeras camadas ocultas. Isso tudo, permite que os algoritmos sejam usados em uma ampla gama de tarefas, indo desde a visão computacional mais simples até jogos complexos (Hinton et al., 2006).

9. Máquinas de vetores de suporte

O modelo de Máquinas de Vetores de Suporte ou *Support Vector Machine* (VSM) pode ser classificado como uma técnica de aprendizado supervisionado, podendo ser usado para classificação e regressão. O conceito central do SVM surgiu, inicialmente, com o objetivo de propor uma nova abordagem para problemas de classificação binária. Para isso, o objetivo foi encontrar um hiperplano ótimo que separasse os dados em diferentes classes com a maior margem possível (Cortes & Vapnik, 1995).

Os SVMs surgem como uma evolução nesse contexto, adotando uma abordagem distinta focada na maximização da margem entre as classes separadas, possibilitando uma maior capacidade de generalização. Para problemas de regressão, os SVMs se destacam pela capacidade de adaptar o modelo de superfície de decisão para uma função de predição, permitindo a criação de regressões robustas e que minimizam desvios dos dados previstos. Assim, é válido destacar que o mesmo oferece uma solução inovadora ao mapear vetores de entrada para um espaço de características de alta dimensão utilizando transformações não lineares (Cortes & Vapnik, 1995).

10. Stochastic Gradient Descent (SGD)

O algoritmo Stochastic Gradient Descent (SGD) é um método iterativo de otimização utilizado em problemas de regressão e classificação, amplamente aplicado em aprendizado de máquina devido à sua eficiência em grandes conjuntos de dados. Diferente do Gradiente Descendente tradicional, que atualiza os parâmetros do modelo após avaliar todos os exemplos de treinamento, o SGD realiza atualizações a cada amostra, tornando-o mais rápido e adequado para grandes volumes de dados (Bottou, 2010).

O SGD é altamente sensível à escolha da taxa de aprendizado e pode apresentar variações devido ao seu caráter estocástico. Para mitigar esse comportamento, técnicas como a

regularização L1 e L2 podem ser aplicadas, ajudando a reduzir o sobreajuste e melhorando a generalização do modelo (Scikit-learn Developers, 2024).

11. Elastic Net (E-Net)

O modelo Elastic Net (E-Net) combina as penalizações L1 (Lasso) e L2 (Ridge) para resolver problemas de regressão, sendo especialmente útil quando há muitas variáveis correlacionadas no conjunto de dados (Zou & Hastie, 2005). Diferente do Lasso, que pode zerar coeficientes e levar à seleção de variáveis, o Elastic Net mantém essa propriedade ao mesmo tempo em que evita instabilidades quando há um alto número de variáveis correlacionadas.

No contexto computacional, o Elastic Net é frequentemente utilizado quando o Lasso falha devido ao número de preditores ser maior do que o número de observações ou quando há forte multicolinearidade entre as variáveis explicativas (Scikit-learn Developers, 2024). Esse método pode ser ajustado por meio do hiperparâmetro que controla o equilíbrio entre as penalizações L1 e L2, permitindo maior flexibilidade na modelagem.

12. Kernel Ridge Regression (KRR)

O Kernel Ridge Regression (KRR) combina a regressão Ridge com o truque do kernel, permitindo a modelagem de relações não lineares entre variáveis. Diferente da regressão linear tradicional, o KRR utiliza funções de kernel, como o polinomial e o radial basis function (RBF), para transformar os dados em um espaço de maior dimensionalidade antes de aplicar a regressão Ridge (Saunders et al., 1998).

Uma das principais vantagens do KRR é sua capacidade de capturar padrões complexos sem a necessidade explícita de definir interações entre variáveis, sendo útil em problemas onde os dados apresentam distribuições não lineares (Scikit-learn Developers, 2024). Contudo, a escolha do kernel e seus hiperparâmetros é fundamental para garantir um bom desempenho do modelo, podendo afetar diretamente em seu desempenho.

13. CatBoost

O CatBoost é um algoritmo baseado em gradient boosting desenvolvido para lidar de maneira eficiente com variáveis categóricas, sendo amplamente utilizado em problemas de regressão e classificação. Diferente de outros algoritmos de boosting, o CatBoost emprega um método inovador de codificação de variáveis categóricas e técnicas para reduzir o efeito do sobreajuste (Dorogush et al., 2018).

Uma de suas principais vantagens é sua capacidade de lidar com dados tabulares sem a necessidade de pré-processamento intenso das variáveis categóricas, além de apresentar uma implementação altamente eficiente para grandes volumes de dados (Scikit-learn Developers, 2024). O CatBoost também incorpora processamento paralelo e estratégias avançadas de ordenação para melhorar a estabilidade das previsões.

APÊNDICE – 3B

O subtópico em questão teve por objetivo concentrar as duas transformações que foram realizadas nos atributos para compilar os dados das três bases em uma base única.

- Transformação 1 – atributos da base Agribalyse

Base	Atributos originais	Atributos transformados V1
Entradas		
AY	[sulfonyl]urea-compound {GLO} market for Cut-off, S - Copied from Ecoinvent	Urea
AY	[sulfonyl]urea-compound {RER} production Cut-off, S - Copied from Ecoinvent	Urea
AY	[thio]carbamate-compound {GLO} market for Cut-off, S - Copied from Ecoinvent	Herbicide
AY	[thio]carbamate-compound {RER} production Cut-off, S - Copied from Ecoinvent	Herbicide
AY	A. Average compost, from green waste, biowaste, sludge, manure, slurry	Esterco verde
AY	Acetamide-anilide-compound, unspecified {GLO} market for Cut-off, S - Copied from Ecoinvent	Herbicide
AY	Acetamide-anilide-compound, unspecified {RER} production Cut-off, S - Copied from Ecoinvent	Herbicide
AY	Acclonifen {GLO} market for Cut-off, S - Copied from Ecoinvent	Herbicide
AY	Agricultural machinery, unspecified {CH} production Cut-off, S - Copied from Ecoinvent	OP maquinário
AY	Ammonia (with 100% NH ₃), steam reforming process, at plant (WFLDB 3.5)/RER U	Ammonia e afins
AY	Ammonia, liquid {RER} market for Cut-off, S - Copied from Ecoinvent	Ammonia e afins
AY	Ammonium nitrate (AN), as N, at plant (WFLDB 3.5)/RER U	Ammonia e afins
AY	Ammonium nitrate phosphate (ANP), as P ₂ O ₅ , at plant (WFLDB 3.5)/RER U	Ammonia e afins
AY	Ammonium nitrate, as N {RER} ammonium nitrate production Cut-off, S - Copied from Ecoinvent	Ammonia e afins
AY	Ammonium nitrate, as N/FR U	Ammonia e afins
AY	Ammonium sulfate, as N {RER} ammonium sulfate production Cut-off, S - Copied from Ecoinvent	Ammonia e afins
AY	Application of liquid mineral fertilizer (WFLDB 3.1)/CH U	OP espalhar esterco
AY	Application of plant protection product, by field sprayer {CH} processing Cut-off, S - Copied from Ecoinvent	OP aplicação defensivo
AY	Atrazine {RER} production Cut-off, S - Copied from Ecoinvent	Herbicide
AY	Average mineral fertilizer, as K ₂ O, at regional storehouse/FR U	Potássio e afins
AY	Average mineral fertilizer, as N, at regional storehouse/FR U	Nitrogenados e afins
AY	Average mineral fertilizer, as P ₂ O ₅ , at regional storehouse/FR U	Nitrogenados e afins
AY	Bale loading {CH} processing Cut-off, S - Copied from Ecoinvent	OP carregar fardo
AY	Baling {CH} processing Cut-off, S - Copied from Ecoinvent	OP enfardar
AY	Baling, with round baler (straw)/FR U	OP enfardar
AY	Benzimidazole-compound {GLO} market for Cut-off, S - Copied from Ecoinvent	Fungicide
AY	Benzimidazole-compound {RER} production Cut-off, S - Copied from Ecoinvent	Fungicide
AY	Benzo[thia]diazole-compound {GLO} market for Cut-off, S - Copied from Ecoinvent	Fungicide

AY	Benzo[thia]diazole-compound {RER} production Cut-off, S - Copied from Ecoinvent	Fungicide
AY	Benzoic-compound {GLO} market for Cut-off, S - Copied from Ecoinvent	Herbicide
AY	Biowaste {CH} treatment of biowaste, industrial composting	Biowaste
AY	Bipyridylum-compound {GLO} market for Cut-off, S - Copied from Ecoinvent	Herbicide
AY	Boric acid, anhydrous, powder {RER} production Cut-off, S - Copied from Ecoinvent	Micronutrientes
AY	Calcium ammonium nitrate (CAN), as N, at plant (WFLDB 3.5)/RER U	Cálcio e afins
AY	Calcium ammonium nitrate, as N/FR U	Cálcio e afins
AY	Carbon dioxide, in air	Carbon dioxide, in air
AY	Carbon, organic, in soil or biomass stock	Carbon, organic, in soil or biomass stock
AY	Chlorothalonil {GLO} market for Cut-off, S - Copied from Ecoinvent	Fungicide
AY	Chlorotoluron {GLO} market for Cut-off, S - Copied from Ecoinvent	Herbicide
AY	Clover seed, Swiss integrated production, for sowing {GLO} market for Cut-off, S - Copied from Ecoinvent	Seed clover
AY	Combine harvesting {CH} processing Cut-off, S - Copied from Ecoinvent	OP colheita
AY	Compost {CH} treatment of biowaste, industrial composting	Composto
AY	Crop, default, heavy metals uptake (WFLDB 3.1)/GLO U	Crop, default, heavy metals uptake (WFLDB 3.1)/GLO U
AY	Crushing, with shredder or chipper, processing/RoW U	OP britagem
AY	Crushing, with shredder or chipper/FR U	OP britagem
AY	Cyclic N-compound {GLO} market for Cut-off, S - Copied from Ecoinvent	cyclic N-compound
AY	Cyclic N-compound {RER} production Cut-off, S - Copied from Ecoinvent	cyclic N-compound
AY	Diammonium phosphate (DAP), as N, at plant (WFLDB 3.5)/RER U	Fosfatos e afins
AY	Diazole-compound {GLO} market for Cut-off, S - Copied from Ecoinvent	Herbicide
AY	Diesel {Europe without Switzerland} market for Cut-off, S - Copied from Ecoinvent	Diesel
AY	Diesel combustion, in tractor/FR U	Diesel
AY	Diesel, burned in agricultural machinery (WFLDB 3.1)/kg/GLO U	Diesel
AY	Dimethenamide {GLO} market for Cut-off, S - Copied from Ecoinvent	Herbicide
AY	Dinitroaniline-compound {GLO} market for Cut-off, S - Copied from Ecoinvent	Fungicide
AY	Dinitroaniline-compound {RER} production Cut-off, S - Copied from Ecoinvent	Fungicide
AY	Diphenylether-compound {GLO} market for Cut-off, S - Copied from Ecoinvent	Herbicide
AY	Diphenylether-compound {RER} production Cut-off, S - Copied from Ecoinvent	Herbicide
AY	Dithiocarbamate-compound {GLO} market for Cut-off, S - Copied from Ecoinvent	Fungicide
AY	Drying of bread grain, seed and legumes {CH} processing Cut-off, S - Copied from Ecoinvent	Secagem dos grãos
AY	Drying of feed grain {CH} processing Cut-off, S - Copied from Ecoinvent	Secagem dos grãos
AY	Drying of sunflower in storing agency, processing/FR U	Secagem dos grãos
AY	Electricity, low voltage {FR} market for Cut-off, S - Copied from Ecoinvent	Energia elétrica, baixa voltagem
AY	Electricity, medium voltage {FR} market for Cut-off, S - Copied from Ecoinvent	Energia elétrica, média voltagem
AY	Emissions from fungicides, unspecified (WFLDB 3.1)/GLO S	Emissões fungicidas

AY	Emissions from herbicides, unspecified (WFLDB 3.1)/GLO S	Emissões herbicidas
AY	Emissions from insecticides, unspecified (WFLDB 3.1)/GLO S	Emissões inseticidas
AY	Emissions from pesticides, unspecified (WFLDB 3.1)/GLO S	Emissões pesticidas
AY	Energy, gross calorific value, in biomass	Energia da biomassa
AY	Fertilising, by broadcaster {CH} processing Cut-off, S - Copied from Ecoinvent	OP fertilização
AY	Fertilizing or plant protection, with sprayer, 2500 l/FR U	OP fertilização ou aplicação pesticida
AY	Fertilizing or plant protection, with sprayer, 2500l, processing/RoW U	OP fertilização ou aplicação pesticida
AY	Fertilizing, slurry, with tanker, processing/RoW U	OP fertilização
AY	Fertilizing, slurry, with tanker/FR U	OP fertilização
AY	Fertilizing, solid manure (charging and spreading) with 8-10t spreader, processing/RoW U	OP espalhar esterco
AY	Fertilizing, solid manure (charging and spreading), with 8-10t spreader/FR U	OP espalhar esterco
AY	Fertilizing, solid manure or compost (charging and spreading), with frontal bucket and 5t spreader/FR U	OP espalhar esterco
AY	Fertilizing, with spreader, 2500 l, processing/RoW U	OP fertilização
AY	Fertilizing, with spreader, 2500 l/FR U	OP fertilização
AY	Flooding of paddy fields, with motor cultivator/FR U	OP inundação
AY	Glyphosate {GLO} market for Cut-off, S - Copied from Ecoinvent	Herbicide
AY	Glyphosate {RER} production Cut-off, S - Copied from Ecoinvent	Herbicide
AY	Grain maize seed, organic, at farm gate/FR U	Seed Maize, organic
AY	Grass seed, Swiss integrated production, for sowing {CH} market for grass seed, Swiss integrated production, for sowing Cut-off, S - Copied from Ecoinvent	Seed Maize
AY	Harrowing, with harrow 12m/FR U	OP gradagem
AY	Harrowing, with rotary harrow (standard equipment), processing/RoW U	OP gradagem
AY	Harrowing, with rotary harrow (standard equipment)/FR U	OP gradagem
AY	Harrowing, with vibrating tine cultivator (standard equipment) 5m, processing/RoW U	OP gradagem
AY	Harrowing, with vibrating tine cultivator (standard equipment) 5m/FR U	OP gradagem
AY	Harvester {CH} production Cut-off, S - Copied from Ecoinvent	Colhedeira
AY	Harvesting, with combine harvester, processing/RoW U	OP colheita
AY	Harvesting, with combine harvester/FR U	OP colheita
AY	Hoeing, with 4-6m hoe (standard), processing/RoW U	OP capinar
AY	Hoeing, with 4-6m hoe (standard)/FR U	OP capinar
AY	Horn meal, at regional storehouse/FR U	Horn meal
AY	Irrigating, sprinkler, diesel powered (WFLDB 3.1)/GLO U	OP Irrigação
AY	Irrigating, sprinkler, electricity powered (WFLDB 3.1)/BR U	OP Irrigação
AY	Irrigating, sprinkler, electricity powered (WFLDB 3.1)/US U	OP Irrigação
AY	Irrigating, surface, diesel powered (WFLDB 3.1)/GLO U	OP Irrigação
AY	Irrigation {US} processing Cut-off, S - Copied from Ecoinvent	OP Irrigação
AY	Isoproturon {GLO} market for Cut-off, S - Copied from Ecoinvent	Herbicide
AY	Jasmine rice, North East, Irrigated, Dry season, at farm gate/TH U	Grain Rice

AY	Jasmine rice, North East, Irrigated, Wet season, at farm gate/TH U	Grain Rice
AY	Jasmine rice, North East, Rainfed, Wet season, at farm gate/TH U	Grain Rice
AY	Jasmine rice, North, Irrigated, Dry season, at farm gate/TH U	Grain Rice
AY	Jasmine rice, North, Irrigated, Wet season, at farm gate/TH U	Grain Rice
AY	Jasmine rice, North, Rainfed, Wet season, at farm gate/TH U	Grain Rice
AY	Land use change, annual crop (WFLDB 3.1)/AR U	land use change, annual crop
AY	Land use change, annual crop (WFLDB 3.1)/AU U	land use change, annual crop
AY	Land use change, annual crop (WFLDB 3.1)/BR U	land use change, annual crop
AY	Land use change, annual crop (WFLDB 3.1)/CA U	land use change, annual crop
AY	Land use change, annual crop (WFLDB 3.1)/FR U	land use change, annual crop
AY	Land use change, annual crop (WFLDB 3.1)/HU U	land use change, annual crop
AY	Land use change, annual crop (WFLDB 3.1)/IN U	land use change, annual crop
AY	Land use change, annual crop (WFLDB 3.1)/RU U	land use change, annual crop
AY	Land use change, annual crop (WFLDB 3.1)/UA U	land use change, annual crop
AY	Land use change, annual crop (WFLDB 3.1)/US U	land use change, annual crop
AY	Lime {CH} production, milled, loose Cut-off, S - Copied from Ecoinvent	Calcário
AY	Lime {GLO} market for Cut-off, S - Copied from Ecoinvent	Calcário
AY	Lime {RER} market for lime Cut-off, S - Copied from Ecoinvent	Calcário
AY	Liquid manure spreading, by vacuum tanker {CH} processing Cut-off, S - Copied from Ecoinvent	OP espalhar esterco
AY	Maize grain {US} production Cut-off, S - Copied from Ecoinvent	Grain Maize
AY	Maize grain, at farm (WFLDB 3.1)/AR U	Grain Maize
AY	Maize grain, at farm (WFLDB 3.1)/BR U	Grain Maize
AY	Maize grain, seed, conventional, national average, animal feed, at farm gate/FR U	Seed maize
AY	Maize grains, heavy metals uptake (WFLDB 3.1)/GLO U	Grain Maize
AY	Maize seed, at regional storehouse (WFLDB 3.1)/CH U	Seed maize
AY	Metaldehyde {GLO} market for Cut-off, S - Copied from Ecoinvent	Inseticide
AY	Metolachlor {GLO} market for Cut-off, S - Copied from Ecoinvent	Herbicide
AY	Napropamide {GLO} market for Cut-off, S - Copied from Ecoinvent	Herbicide
AY	Natural gas, high pressure {FR} market for Cut-off, S - Copied from Ecoinvent	Natural gas
AY	Nitrile-compound {GLO} market for Cut-off, S - Copied from Ecoinvent	Herbicide
AY	Nitrogen fertiliser, as N {GLO} nutrient supply from potassium nitrate	Nitrogenados e afins
AY	Nitrogen fertiliser, as N {RER} ammonium nitrate phosphate production	Nitrogenados e afins
AY	Nitrogen fertiliser, as N {RER} calcium ammonium nitrate production Cut-off, S - Copied from Ecoinvent	Nitrogenados e afins
AY	Nitrogen fertiliser, as N {RER} diammonium phosphate production	Nitrogenados e afins
AY	Nitrogen fertiliser, as N {RER} monoammonium phosphate production	Nitrogenados e afins
AY	Nitrogen fertiliser, as N {RER} urea ammonium nitrate production Cut-off, S - Copied from Ecoinvent	Nitrogenados e afins

AY	Occupation, annual crop	Occupation, annual crop
AY	Occupation, annual crop, non-irrigated	Occupation, annual crop, non-irrigated
AY	Oilseed processing, in seed centre (WFLDB 3.1)/CH U	OP processamento sementes oleaginosas
AY	Organic manure mix (empty process), as K2O/FR U	Esterco orgânico
AY	Organic manure mix (empty process), as N/FR U	Esterco orgânico
AY	Organic manure mix (empty process), as P2O5/FR U	Esterco orgânico
AY	Organic or farm manure (empty process), as K2O/FR U	Esterco orgânico
AY	Organic or farm manure (empty process), as N/FR U	Esterco orgânico
AY	Organic or farm manure (empty process), as P2O5/FR U	Esterco orgânico
AY	Organophosphorus-compound, unspecified {GLO} market for Cut-off, S - Copied from Ecoinvent	Herbicide
AY	Organophosphorus-compound, unspecified {RER} production Cut-off, S - Copied from Ecoinvent	Herbicide
AY	Packaging, liquid fertilisers or pesticides (WFLDB 3.1)/GLO U	packaging, for fertilisers
AY	Packaging, solid fertilisers or pesticides (WFLDB 3.1)/GLO U	packaging, for fertilisers
AY	Pea seed, for sowing {CH} production Cut-off, S - Copied from Ecoinvent	Seed pea
AY	Pendimethalin {GLO} market for Cut-off, S - Copied from Ecoinvent	Herbicide
AY	Pesticide, unspecified {GLO} market for Cut-off, S - Copied from Ecoinvent	Pesticide
AY	Pesticide, unspecified {RER} production Cut-off, S - Copied from Ecoinvent	Pesticide
AY	Phenoxy-compound {GLO} market for Cut-off, S - Copied from Ecoinvent	Herbicide
AY	Phenoxy-compound {RER} production Cut-off, S - Copied from Ecoinvent	Herbicide
AY	Phosphate fertiliser, as P2O5 {RER} ammonium nitrate phosphate production	Fosfatos e afins
AY	Phosphate fertiliser, as P2O5 {RER} basic slag fertiliser production Cut-off, S - Copied from Ecoinvent	Fosfatos e afins
AY	Phosphate fertiliser, as P2O5 {RER} diammonium phosphate production	Fosfatos e afins
AY	Phosphate fertiliser, as P2O5 {RER} monoammonium phosphate production	Fosfatos e afins
AY	Phosphate fertiliser, as P2O5 {RER} single superphosphate production Cut-off, S - Copied from Ecoinvent	Fosfatos e afins
AY	Phosphate fertiliser, as P2O5 {RER} triple superphosphate production Cut-off, S - Copied from Ecoinvent	Fosfatos e afins
AY	Phosphate rock, as P2O5, beneficiated, dry {MA} phosphate rock beneficiation, dry Cut-off, S - Copied from Ecoinvent	Fosfatos e afins
AY	Plant protection, spraying, with knapsack sprayer/TH U	OP aplicação defensivo
AY	Plant protection, weeding, with portable swinging scythe/FR U	OP aplicação defensivo
AY	Ploughing, with 5 or 6 soc plough, processing/RoW U	OP arar
AY	Ploughing, with 5 or 6 soc plough/FR U	OP arar
AY	Potassium chloride, as K2O {RER} potassium chloride production Cut-off, S - Copied from Ecoinvent	Potássio e afins
AY	Potassium chloride, as K2O, at plant (WFLDB 3.5)/RER U	Potássio e afins
AY	Potassium chloride, as K2O/FR U	Potássio e afins
AY	Potassium fertiliser, as K2O {GLO} nutrient supply from potassium nitrate	Potássio e afins
AY	Potassium sulfate, as K2O {RER} potassium sulfate production Cut-off, S - Copied from Ecoinvent	Potássio e afins

AY	Pyrethroid-compound {GLO} market for Cut-off, S - Copied from Ecoinvent	Inseticide
AY	Pyrethroid-compound {RER} production Cut-off, S - Copied from Ecoinvent	Inseticide
AY	Pyridine-compound {GLO} market for Cut-off, S - Copied from Ecoinvent	Inseticide
AY	Rape seed {DE} production Cut-off, S - Copied from Ecoinvent	Grain Rapeseed
AY	Rape seed {FR} production Cut-off, S - Copied from Ecoinvent	Grain Rapeseed
AY	Rape seed {US} production Cut-off, S - Copied from Ecoinvent	Grain Rapeseed
AY	Rape seed, at regional storehouse (WFLDB 3.1)/CH U	Grain Rapeseed
AY	Rape seed, heavy metals uptake (WFLDB 3.1)/GLO U	Grain Rapeseed
AY	Rapeseed, at farm (WFLDB 3.1)/CA U	Grain Rapeseed
AY	Rapeseed, conventional, 9% moisture, national average, animal feed, at farm gate, production/FR U	Grain Rapeseed
AY	Rapeseed, seed, conventional, national average, animal feed, at farm gate/FR U	Seed rapeseed
AY	Rice grains, heavy metals uptake (WFLDB 3.1)/GLO U	Grain Rice
AY	Rice seed, at regional storehouse (WFLDB 3.1)/GLO U	Seed rice
AY	Rolling, with roller 3m/FR U	OP rolling
AY	Rolling, with roller 9m, processing/RoW U	OP rolling
AY	Rolling, with roller 9m/FR U	OP rolling
AY	Soft wheat grain, conventional, breadmaking quality, 15% moisture, at farm gate/FR U	Grain Wheat
AY	Soft wheat grain, conventional, national average, animal feed, at farm gate, production/FR U	Grain Wheat
AY	Soft wheat grain, conventional, protein improved quality, 15% moisture, at farm gate/FR U	Grain Wheat
AY	Soft wheat grain, organic (model type), after Alfalfa, Central Region, at farm gate/FR U	Grain Wheat, organic
AY	Soft wheat grain, organic (model type), after fava beans, Central Region, at farm gate/FR U	Grain Wheat, organic
AY	Soft wheat, seed, conventional, national average, animal feed, at farm gate/FR U	Seed wheat
AY	Soft winter wheat seed, organic, at farm gate/FR U	Seed wheat, organic
AY	Soft winter wheat seed, organic, system n°1, at farm gate/FR U	Seed wheat, organic
AY	Soil decompaction, processing/RoW U	OP Soil decompaction
AY	Soil decompaction, with 4.5m chisel, processing/RoW U	OP Soil decompaction
AY	Soil decompaction, with 4.5m chisel/FR U	OP Soil decompaction
AY	Soil decompaction/FR U	OP Soil decompaction
AY	Soil maintenance, with cover crop 4m, processing/RoW U	OP Soil maintenance
AY	Soil maintenance, with cover crop 4m/FR U	OP Soil maintenance
AY	Soil preparation, with disc harrow/FR U	OP Soil preparation
AY	Soil preparation, with rotary tiller/FR U	OP Soil preparation
AY	Solid manure loading and spreading, by hydraulic loader and spreader {CH} processing Cut-off, S - Copied from Ecoinvent	OP espalhar esterco
AY	Sowing {CH} processing Cut-off, S - Copied from Ecoinvent	OP semeadura
AY	Sowing or planting, with classic seeder and harrow, processing/RoW U	OP semeadura
AY	Sowing or planting, with classic seeder and harrow/FR U	OP semeadura ou plantar
AY	Sowing or planting, with pneumatic seeder and harrow, processing/RoW U	OP semeadura ou plantar

AY	Sowing or planting, with pneumatic seeder and harrow/FR U	OP semeadura ou plantar
AY	Sowing or planting, with pneumatic seeder, 6 rows, processing/ RoW U	OP semeadura ou plantar
AY	Sowing or planting, with pneumatic seeder, 6 rows/FR U	OP semeadura ou plantar
AY	Soybean seed, at farm (WFLDB 3.1)/CH U	Seed soybean
AY	Soybean, at farm (WFLDB 3.1)/AR U	Grain Soybean
AY	Soybean, at farm (WFLDB 3.1)/BR U	Grain Soybean
AY	Soybean, at farm (WFLDB 3.1)/US U	Grain Soybean
AY	Soybeans, heavy metals uptake (WFLDB 3.1)/GLO U	Grain Soybean
AY	Spring barley seed, conventional, malting quality, animal feed, at farm gate/FR U	Seed barley
AY	Spring barley seed, organic, at farm gate/FR U	Seed barley, organic
AY	Spring barley, conventional, downgraded quality, animal feed, at farm gate/FR U	Grain Barley
AY	Spring barley, conventional, malting quality, animal feed, at farm gate/FR U	Grain Barley
AY	Spring barley, organic, system n°3, at farm gate/FR U	Grain Barley, organic
AY	Spring barley, organic, system n°4, at farm gate/FR U	Grain Barley, organic
AY	Spring pea seed, conventional, 15% moisture, animal feed, at farm gate, production/FR U	Seed pea
AY	Spring wheat grain, at farm (WFLDB 3.1)/CA U	Grain Wheat
AY	Storage of grain, temporary, processing/FR U	Silo para estocar
AY	Storing of cereals grain in ventilated bin in storing agency, processing/FR U	Silo para estocar
AY	Stubble ploughing, with stubble share 5.5m, processing/RoW U	OP aração
AY	Stubble ploughing, with stubble share 5.5m/FR U	OP aração
AY	Sunflower grain, conventional, 9% moisture, national average, animal feed, at farm gate, production/FR U	Grain Sunflower
AY	Sunflower seed, at farm (WFLDB 3.1)/CH U	Seed sunflower
AY	Sunflower seed, organic, at farm gate/FR U	Seed sunflower, organic
AY	Sunflower, at farm (WFLDB 3.1)/FR U	Grain Sunflower
AY	Sunflower, at farm (WFLDB 3.1)/HU U	Grain Sunflower
AY	Sunflower, at farm (WFLDB 3.1)/RU U	Grain Sunflower
AY	Sunflower, at farm (WFLDB 3.1)/UA U	Grain Sunflower
AY	Sunflower, seed, conventional, national average, animal feed, at farm gate, production/FR U	Seed sunflower
AY	Tillage, cultivating, chiselling {CH} processing Cut-off, S - Copied from Ecoinvent	OP arar, tillage
AY	Tillage, generic (WFLDB 3.1)/CH U	OP arar, tillage
AY	Tillage, harrowing, by spring tine harrow {CH} processing Cut-off, S - Copied from Ecoinvent	OP arar, tillage
AY	Tillage, ploughing {CH} processing Cut-off, S - Copied from Ecoinvent	OP arar, tillage
AY	Tillage, rotary cultivator {CH} processing Cut-off, S - Copied from Ecoinvent	OP arar, tillage
AY	Tractor, 4-wheel, agricultural {CH} production Cut-off, S - Copied from Ecoinvent	Trator
AY	Transformation, from annual crop	Transformation, from annual crop
AY	Transformation, from annual crop, non-irrigated	Transformation, from annual crop, non-irrigated

AY	Transformation, from forest, intensive, clear-cutting	Transformation, from forest, intensive, clear-cutting
AY	Transformation, from forest, unspecified	Transformation, from forest, unspecified
AY	Transformation, from grassland	Transformation, from grassland
AY	Transformation, from permanent crop	Transformation, from permanent crop
AY	Transformation, from shrub land, sclerophyllous	Transformation, from shrub land, sclerophyllous
AY	Transformation, to annual crop	Transformation, to annual crop
AY	Transformation, to annual crop, non-irrigated	Transformation, to annual crop, non-irrigated
AY	Transport, freight train {Europe without Switzerland} market for Cut-off, S - Copied from Ecoinvent	transport, train
AY	Transport, freight, inland waterways, barge {RER} processing	transport, barge
AY	Transport, freight, light commercial vehicle {Europe without Switzerland} processing Cut-off, S - Copied from Ecoinvent	transport, light commercial vehicle
AY	Transport, freight, lorry >32 metric ton, EURO3 {RER} transport, freight, lorry >32 metric ton, EURO3 Cut-off, S - Copied from Ecoinvent	transport, lorry
AY	Transport, freight, lorry >32 metric ton, EURO4 {GLO} market for Cut-off, S - Copied from Ecoinvent	transport, lorry
AY	Transport, freight, lorry >32 metric ton, EURO4 {RER} transport, freight, lorry >32 metric ton, EURO4 Cut-off, S - Copied from Ecoinvent	transport, lorry
AY	Transport, freight, lorry 16-32 metric ton, EURO3 {RER} transport, freight, lorry 16-32 metric ton, EURO3 Cut-off, S - Copied from Ecoinvent	transport, lorry
AY	Transport, freight, lorry, unspecified {GLO} market for Cut-off, S - Copied from Ecoinvent	transport, lorry
AY	Transport, freight, lorry, unspecified {RER} market for transport, freight, lorry, unspecified Cut-off, S - Copied from Ecoinvent	transport, lorry
AY	Transport, freight, sea, transoceanic ship {GLO} market for Cut-off, S - Copied from Ecoinvent	transport, container ship
AY	Transport, tractor and trailer, agricultural {CH} processing Cut-off, S - Copied from Ecoinvent	transport, tractor and trailer
AY	Transporting to farm, with 2 axle trailer (15 t), processing/RoW U	OP transporte para fazenda
AY	Transporting to farm, with 2 axle trailer (15 t)/FR U	OP transporte para fazenda
AY	Transporting to farm, with trailer (<15t) heavy tractor, processing/RoW U	OP transporte para fazenda
AY	Transporting to farm, with trailer (<15t) heavy tractor/FR U	OP transporte para fazenda
AY	Transporting, tractor with 2 axle trailer (15 t), processing/FR U	transport, tractor
AY	Transporting, with forage flatbed/FR U	OP transporte para fazenda
AY	Triazine-compound, unspecified {GLO} market for Cut-off, S - Copied from Ecoinvent	Herbicide
AY	Triazine-compound, unspecified {RER} production Cut-off, S - Copied from Ecoinvent	Herbicide
AY	Triple superphosphate, as P2O5, at plant (WFLDB 3.5)/RER U	triple superphosphate
AY	Triple superphosphate, as P2O5/FR U	triple superphosphate
AY	Urea ammonium nitrate (UAN), as N, at plant (WFLDB 3.5)/RER U	Urea
AY	Urea ammonium nitrate, as N/FR U	Urea
AY	Urea, as N {RER} production Cut-off, S - Copied from Ecoinvent	Urea
AY	Urea, as N, at plant (WFLDB 3.5)/RER U	Urea
AY	Urea, as N/FR U	Urea
AY	Vinasse, from fermentation of sugar beet {GLO} market for Cut-off, S - Copied from Ecoinvent	Vinhaça
AY	Water, rain	Water, rain

AY	Water, river	Water, river
AY	Water, river, AR	Water, river
AY	Water, river, BR	Water, river
AY	Water, river, FR	Water, river
AY	Water, river, IN	Water, river
AY	Water, river, US	Water, river
AY	Water, well, AR	Water, well
AY	Water, well, BR	Water, well
AY	Water, well, FR	Water, well
AY	Water, well, IN	Water, well
AY	Water, well, US	Water, well
AY	Wheat grain {FR} wheat production Cut-off, U - Copied from Ecoinvent	Grain Wheat
AY	Wheat grain {US} wheat production Cut-off, U - Copied from Ecoinvent	Grain Wheat
AY	Wheat grains, heavy metals uptake (WFLDB 3.1)/GLO U	Grain Wheat
AY	Wheat seed, at regional storehouse (WFLDB 3.1)/CH U	Seed wheat
AY	Wheat straw, heavy metals uptake (WFLDB 3.1)/GLO U	Straw Wheat
AY	Winter barley seed, conventional, malting quality, animal feed, at farm gate/FR U	Seed barley
AY	Winter barley seed, organic, at farm gate /FR U	Seed barley, organic
AY	Winter barley, conventional, malting quality, animal feed, at farm gate/FR U	Grain Barley
AY	Winter barley, organic, system n°1, at farm gate/FR U	Grain Barley, organic
AY	Winter barley, organic, system n°2, at farm gate/FR U	Grain Barley, organic
AY	Winter forage barley, conventional, animal feed, at farm gate/FR U	Grain Barley, forage
AY	Winter wheat grain, at farm (WFLDB 3.1)/AU U	Grain Wheat
AY	Winter wheat, organic, system n°12, at farm gate/FR U	Grain Wheat, organic
AY	Winter wheat, organic, system n°13, at farm gate/FR U	Grain Wheat, organic
AY	Winter wheat, organic, system n°14, at farm gate/FR U	Grain Wheat, organic
AY	Winter wheat, organic, system n°15, at farm gate/FR U	Grain Wheat, organic
AY	Winter wheat, organic, system n°16, at farm gate/FR U	Grain Wheat, organic
AY	Winter wheat, organic, system n°17, at farm gate/FR U	Grain Wheat, organic
AY	Winter wheat, organic, system n°18, at farm gate/FR U	Grain Wheat, organic
AY	Winter wheat, organic, system n°5, at farm gate/FR U	Grain Wheat, organic
AY	Winter wheat, organic, system n°6, at farm gate/FR U	Grain Wheat, organic
AY	Winter wheat, organic, system n°8, at farm gate/FR U	Grain Wheat, organic
AY	Winter wheat, organic, system n°9, at farm gate/FR U	Grain Wheat, organic
Saídas		
AY	barley	S_Grain Barley
AY	barley, organic	S_Grain Barley, organic

AY	Feed barley	S_Grain Barley, feed
AY	forage barley	S_Grain Barley, forage
AY	Grain maize	S_Grain Maize
AY	Grain maize, organic	S_Grain Maize, organic
AY	Jasmine rice	S_Grain Rice
AY	Rapeseed	S_Grain Rapeseed
AY	Rapeseed, organic	S_Grain Rapeseed, organic
AY	Rice	S_Grain Rice
AY	Soybean	S_Grain Soybean
AY	Sunflower	S_Grain Sunflower
AY	Sunflower dried	S_Grain Sunflower, dried
AY	Sunflower, organic	S_Grain Sunflower, organic
AY	wheat	S_Grain Wheat
AY	wheat, organic	S_Grain Wheat, organic

- Transformação 1 – atributos da base Ecoinvent

Base	Atributos originais	Atributos transformados V1
Entradas		
EI	[sulfonyl]urea-compound	Urea
EI	[thio]carbamate-compound	Herbicide
EI	2,4-dichlorophenol	Herbicide
EI	acetamide-anillide-compound, unspecified	Herbicide
EI	aclonifen	Herbicide
EI	ammonia, anhydrous, liquid	Ammonia e afins
EI	ammonium nitrate	Ammonia e afins
EI	ammonium sulfate	Ammonia e afins
EI	application of plant protection product, by field sprayer	OP aplicação defensivo
EI	atrazine	Herbicide
EI	bale loading	OP carregar fardo
EI	baling	OP enfardar
EI	barley seed, for sowing	Seed barley
EI	barley seed, organic, for sowing	Seed barley, organic
EI	barley seed, Swiss integrated production, for sowing	Seed barley
EI	benzimidazole-compound	Fungicide
EI	benzo[thia]diazole-compound	Fungicide

EI	benzoic-compound	Herbicide
EI	bipyridylium-compound	Herbicide
EI	borax, anhydrous, powder	Micronutrientes
EI	boric acid, anhydrous, powder	Micronutrientes
EI	calcium ammonium nitrate	Cálcio e afins
EI	calcium carbonate, precipitated	Cálcio e afins
EI	captan	Fungicide
EI	Carbon dioxide, in air	Carbon dioxide, in air
EI	chemical, inorganic	chemical, inorganic
EI	cobalt	Micronutrientes
EI	combine harvesting	OP colheita
EI	compost	Composto
EI	copper sulfate	Micronutrientes
EI	cyclic N-compound	cyclic N-compound
EI	diammonium phosphate	Fosfatos e afins
EI	diazole-compound	Herbicide
EI	diesel, burned in agricultural machinery	Energia do diesel
EI	diesel, low-sulfur	Diesel
EI	dimethenamide	Herbicide
EI	dinitroaniline-compound	Fungicide
EI	diphenylether-compound	Herbicide
EI	dolomite	Calcário
EI	drying of bread grain, seed and legumes	Secagem dos grãos
EI	drying of feed grain	Secagem dos grãos
EI	drying of maize grain	Secagem dos grãos
EI	drying of maize straw and whole-plant	Secagem dos grãos
EI	electricity, low voltage	Energia elétrica, baixa voltagem
EI	Energy, gross calorific value, in biomass	Energia da biomassa
EI	fatty acid methyl ester	fatty acid methyl ester
EI	fertilising, by broadcaster	OP fertilização
EI	glyphosate	Herbicide
EI	green manure, organic, until April	Esterco verde
EI	green manure, organic, until March	Esterco verde
EI	green manure, Swiss integrated production, until April	Esterco verde
EI	green manure, Swiss integrated production, until January	Esterco verde
EI	green manure, Swiss integrated production, until March	Esterco verde
EI	gypsum, mineral	Gesso

EI	heat, central or small-scale, natural gas	Aquecimento com gás natural
EI	hoeing	OP capinar
EI	inorganic nitrogen fertiliser, as N	Nitrogenados e afins
EI	inorganic phosphorus fertiliser, as P ₂ O ₅	Fosfatos e afins
EI	inorganic potassium fertiliser, as K ₂ O	Potássio e afins
EI	irrigation	OP Irrigação
EI	land use change, annual crop	land use change, annual crop
EI	lime	Calcário
EI	liquid manure spreading, by vacuum tanker	OP espalhar esterco
EI	magnesium oxide	Macronutrientes (kg)
EI	magnesium sulfate	Macronutrientes (kg)
EI	maize seed, for sowing	Seed Maize
EI	maize seed, organic, for sowing	Seed Maize, organic
EI	maize seed, Swiss integrated production, for sowing	Seed Maize
EI	mancozeb	Fungicide
EI	manganese sulfate	Micronutrientes
EI	manure, liquid, cattle	Esterco (vaca)
EI	manure, liquid, swine	Esterco (porco)
EI	manure, solid, cattle	Esterco (vaca), sólido
EI	metaldehyde	Inseticide
EI	metolachlor	Herbicide
EI	monoammonium phosphate	Fosfatos e afins
EI	naphtha	naphtha
EI	nitrile-compound	Herbicide
EI	Occupation, annual crop	Occupation, annual crop
EI	Occupation, annual crop, irrigated	Occupation, annual crop, irrigated
EI	Occupation, annual crop, irrigated, intensive	Occupation, annual crop, irrigated, intensive
EI	Occupation, annual crop, non-irrigated	Occupation, annual crop, non-irrigated
EI	Occupation, annual crop, non-irrigated, extensive	Occupation, annual crop, non-irrigated, extensive
EI	Occupation, annual crop, non-irrigated, intensive	Occupation, annual crop, non-irrigated, intensive
EI	organophosphorus-compound, unspecified	Herbicide
EI	packaging, for fertilisers	packaging, for fertilisers
EI	packaging, for pesticides	packaging, for pesticides
EI	pendimethalin	Herbicide
EI	pesticide, unspecified	Pesticide
EI	phenoxy-compound	Herbicide
EI	phosphate rock, beneficiated	Fosfatos e afins

EI	potassium chloride	Potássio e afins
EI	potassium sulfate	Potássio e afins
EI	poultry manure, dried	Esterco (frango), sólido
EI	poultry manure, fresh	Esterco (frango), sólido
EI	pyrethroid-compound	Inseticide
EI	pyridazine-compound	Herbicide
EI	rape seed, for sowing	Seed rapeseed
EI	rape seed, organic, for sowing	Seed rapeseed, organic
EI	rice seed, for sowing	Seed rice
EI	single superphosphate	single superphosphate
EI	solid manure loading and spreading, by hydraulic loader and spreader	OP espalhar esterco
EI	sowing	OP semeadura
EI	soybean seed, for sowing	Seed soybean
EI	soybean seed, organic, for sowing	Seed soybean, organic
EI	sulfite	sulfite
EI	sulfur	sulfur
EI	sunflower seed, for sowing	Seed sunflower
EI	swath, by rotary windrower	OP swath
EI	tillage, cultivating, chiselling	OP arar, tillage
EI	tillage, currying, by weeder	OP arar, tillage
EI	tillage, harrowing, by rotary harrow	OP arar, tillage
EI	tillage, harrowing, by spring tine harrow	OP arar, tillage
EI	tillage, ploughing	OP arar, tillage
EI	tillage, rolling	OP arar, tillage
EI	tillage, rotary cultivator	OP arar, tillage
EI	Transformation, from annual crop	Transformation, from annual crop
EI	Transformation, from annual crop, irrigated	Transformation, from annual crop, irrigated
EI	Transformation, from annual crop, non-irrigated	Transformation, from annual crop, non-irrigated
EI	Transformation, from annual crop, non-irrigated, extensive	Transformation, from annual crop, non-irrigated, extensive
EI	Transformation, from annual crop, non-irrigated, intensive	Transformation, from annual crop, non-irrigated, intensive
EI	Transformation, from arable land, unspecified use	Transformation, from arable land, unspecified use
EI	Transformation, from pasture, man made, intensive	Transformation, from pasture, man made, intensive
EI	Transformation, to annual crop	Transformation, to annual crop
EI	Transformation, to annual crop, irrigated	Transformation, to annual crop, irrigated
EI	Transformation, to annual crop, irrigated, intensive	Transformation, to annual crop, irrigated, intensive
EI	Transformation, to annual crop, non-irrigated	Transformation, to annual crop, non-irrigated
EI	Transformation, to annual crop, non-irrigated, extensive	Transformation, to annual crop, non-irrigated, extensive

EI	Transformation, to annual crop, non-irrigated, intensive	Transformation, to annual crop, non-irrigated, intensive
EI	Transformation, to arable land, unspecified use	Transformation, to arable land, unspecified use
EI	transport, freight train	transport, train
EI	transport, freight, inland waterways, barge	transport, barge
EI	transport, freight, lorry, unspecified	transport, lorry
EI	transport, freight, sea, container ship	transport, container ship
EI	transport, tractor and trailer, agricultural	transport, tractor and trailer
EI	triazine-compound, unspecified	Herbicide
EI	triple superphosphate	triple superphosphate
EI	urea	Urea
EI	urea ammonium nitrate mix	Urea
EI	Water, river, US	Water, river
EI	Water, unspecified natural origin	Water, unspecified
EI	Water, well, in ground	Water, well
EI	wheat seed, for sowing	Seed wheat
EI	wheat seed, organic, for sowing	Seed wheat, organic
EI	wheat seed, Swiss integrated production, for sowing	Seed wheat
EI	zinc monosulfate	Micronutrientes
EI	zinc oxide	Micronutrientes
Saídas		
EI	Barley grain	S_Grain Barley
EI	Barley straw	S_Straw Barley
EI	maize grain	S_Grain Maize
EI	maize grain, organic	S_Grain Maize, organic
EI	rape seed	S_Grain Rapeseed
EI	rape seed, organic	S_Grain Rapeseed, organic
EI	rice straw	S_Straw Rice
EI	rice, basmati	S_Grain Rice
EI	rice, non-basmati	S_Grain Rice
EI	soybean	S_Grain Soybean
EI	soybean, organic	S_Grain Soybean, organic
EI	straw	S_Straw Wheat
EI	straw, organic	S_Straw Wheat, organic
EI	sunflower seed	S_Grain Sunflower
EI	wheat grain	S_Grain Wheat
EI	wheat grain, organic	S_Grain Wheat, organic

- Transformação 1 – atributos da base Agrifootprint

Base	Atributos originais	Atributos transformados V1
Entradas		
AG	2,4 D	Herbicide
AG	Aliphatic organothiophosphate insecticides	Inseticide
AG	Ammonia, as 100% NH ₃ (NPK 82-0-0)	Ammonia e afins
AG	Ammonium nitrate, as 100% (NH ₄)(NO ₃) (NPK 35-0-0)	Ammonia e afins
AG	Ammonium sulfate, as 100% (NH ₄) ₂ SO ₄ (NPK 21-0-0)	Ammonia e afins
AG	Aryloxyphenoxypropionic herbicides	Herbicide
AG	Atrazine	Herbicide
AG	Bentazone	Herbicide
AG	Benzimidazole fungicides	Fungicide
AG	Calcium ammonium nitrate (CAN), (NPK 26.5-0-0)	Cálcio e afins
AG	Carbofuran	Inseticide
AG	Chloroacetanilide herbicides	Herbicide
AG	Chlorotriazine herbicides	Herbicide
AG	Cyanazine	Herbicide
AG	Cypermethrin	Inseticide
AG	Di ammonium phosphate, as 100% (NH ₃) ₂ HPO ₄ (NPK 22-57-0)	Fosfatos e afins
AG	Dicamba	Herbicide
AG	Dinitroaniline herbicides	Herbicide
AG	Dinitrophenol herbicides	Herbicide
AG	Dipropylthiocarbamic acid S-ethyl ester	Fungicide
AG	Electricity mix, AC, consumption mix, at consumer, < 1kV	Energia elétrica
AG	Energy, from diesel burned in machinery	Energia do diesel
AG	Fungicide	Fungicide
AG	Fungicide emissions	Emissões fungicidas
AG	Glyphosate	Herbicide
AG	Grain barley	Grain barley
AG	Grain Wheat	Grain Wheat
AG	Grain Wheat spring	Grain Wheat spring
AG	Grain Wheat winter	Grain Wheat winter
AG	Herbicide	Herbicide
AG	Herbicide emissions	Emissões herbicidas

AG	Infraestrutura básica fazenda	Infraestrutura básica fazenda
AG	Inseticide	Inseticide
AG	Inseticide emissions	Emissões inseticidas
AG	Lime	Calcário
AG	Linuron	Herbicide
AG	Liquid urea-ammonium nitrate solution (NPK 30-0-0)	Ammonia e afins
AG	Malathion	Inseticide
AG	Manure (cow)	Esterco (vaca)
AG	Manure (pig)	Esterco (porco)
AG	Manure (poultry)	Esterco (frango)
AG	Material inicial produção semente Barley	Material inicial produção semente Barley
AG	Material inicial produção semente Maize	Material inicial produção semente Maize
AG	Material inicial produção semente Rapeseed	Material inicial produção semente Rapeseed
AG	Material inicial produção semente Rice	Material inicial produção semente Rice
AG	Material inicial produção semente Soybean	Material inicial produção semente Soybean
AG	Material inicial produção semente Sunflower	Material inicial produção semente Sunflower
AG	Material inicial produção semente Wheat	Material inicial produção semente Wheat
AG	Metolachlor	Herbicide
AG	NPK compound (NPK 15-15-15)	NPK compound (NPK 15-15-15)
AG	Occupation, agriculture	Occupation, agriculture
AG	Occupation, annual crop	Occupation, annual crop
AG	Organophosphorus herbicides	Herbicide
AG	Parathion	Inseticide
AG	Phenoxyacetic herbicides	Herbicide
AG	Phenyl organothiophosphate insecticides	Inseticide
AG	Phosphate rock (32% P ₂ O ₅ , 50% CaO) (NPK 0-32-0)	Fosfatos e afins
AG	PK compound (NPK 0-22-22)	PK compound (NPK 0-22-22)
AG	Plant growth regulator	Plant growth regulator
AG	Potassium chloride (NPK 0-0-60)	Potássio e afins
AG	Potassium sulfate (NPK 0-0-50) (Mannheim)	Potássio e afins
AG	Propanil	Herbicide
AG	Pyrethroid ester insecticides	Inseticide
AG	Quaternary ammonium herbicides	Herbicide
AG	Seed Maize	Seed Maize
AG	Seed rice	Seed rice
AG	Silo, for grain storage	Silo para estocar
AG	Single superphosphate, as 35% Ca(H ₂ PO ₄) ₂ (NPK 0-21-0)	Single superphosphate, as 35% Ca(H ₂ PO ₄) ₂ (NPK 0-21-0)

AG	Thiocarbamate herbicides	Herbicide
AG	Transformation, from annual crop	Transformation, from annual crop
AG	Transformation, from forest, unspecified	Transformation, from forest, unspecified
AG	Transformation, from grassland	Transformation, from grassland
AG	Transformation, from permanent crop	Transformation, from permanent crop
AG	Transformation, to annual crop	Transformation, to annual crop
AG	Transport, truck 10-20t, EURO4	Transport, truck 10-20t, EURO4
AG	Triazinylsulfonyleurea herbicides	Herbicide
AG	Triple superphosphate, as 80% Ca(H ₂ PO ₄) ₂ (NPK 0-48-0)	Triple superphosphate, as 80% Ca(H ₂ PO ₄) ₂ (NPK 0-48-0)
AG	Unclassified herbicides	Herbicide
AG	Urea, as 100% CO(NH ₂) ₂ (NPK 46.6-0-0)	Urea
AG	Vapor a partir de gás natural	Vapor a partir de gás natural
AG	Water, unspecified natural origin	Water, unspecified natural origin
Saídas		
AG	Dried Grain Barley	S_Dried Grain Barley
AG	Dried Grain Wheat	S_Dried Grain Wheat
AG	Dried Grain Wheat spring	S_Dried Grain Wheat Spring
AG	Dried Grain Wheat winter	S_Dried Grain Wheat winter
AG	Grain Barley	S_Grain Barley
AG	Grain Maize	S_Grain Maize
AG	Grain Rapeseed	S_Grain Rapeseed
AG	Grain Rice	S_Grain Rice
AG	Grain Soybean	S_Grain Soybean
AG	Grain Sunflower	S_Grain Sunflower
AG	Grain Wheat	S_Grain Wheat
AG	Stover Maize	S_Stover Maize
AG	Straw Barley	S_Straw Barley
AG	Straw rapeseed	S_Straw rapeseed
AG	Straw Rice	S_Straw Rice
AG	Straw Soybean	S_Straw Soybean
AG	Straw Wheat	S_Straw Wheat

- Transformação 2 – atributos das três bases de dados

ENTRADAS			
Atributos originais V1	Unid	Transformação atributos V2	Unid
Ammonia e afins	kg	Ammonia e afins	kg

Aquecimento com gás natural	MJ	Energia gás natural (MJ)	MJ
Biowaste	kg	Esterco verde (kg)	kg
Calcário	kg	Calcário	kg
Cálcio e afins	kg	Macronutrientes (kg)	kg
Carbon dioxide, in air	kg	Carbon dioxide, in air	kg
Carbon, organic, in soil or biomass stock	kg	Carbon, organic, in soil or biomass stock	kg
chemical, inorganic	kg	chemical, inorganic	kg
Colheiteira	kg	Colheiteira	kg
Composto	kg	Composto	kg
Crop, default, heavy metals uptake (WFLDB 3.1)/GLO U	kg	Crop, default, heavy metals uptake (WFLDB 3.1)/GLO U	kg
cyclic N-compound	kg	Macronutrientes (kg)	kg
Diesel	kg	Diesel	kg
Emissões fungicidas	kg	Emissões fungicidas	kg
Emissões herbicidas	kg	Emissões herbicidas	kg
Emissões inseticidas	kg	Emissões inseticidas	kg
Emissões pesticidas	kg	Emissões pesticidas	kg
Energia da biomassa	MJ	Energia da biomassa	MJ
Energia do diesel	MJ	Energia do diesel	MJ
Energia elétrica	MJ	Energia elétrica	MJ
Energia elétrica, baixa voltagem	kWh	Energia elétrica (kwh)	kWh
Energia elétrica, média voltagem	kWh	Energia elétrica (kwh)	kWh
Esterco (frango)	kg	Esterco animal (kg)	kg
Esterco (frango), sólido	kg	Esterco animal (kg)	kg
Esterco (porco)	kg	Esterco animal (kg)	kg
Esterco (vaca)	kg	Esterco animal (kg)	kg
Esterco (vaca), sólido	kg	Esterco animal (kg)	kg
Esterco orgânico	kg	Esterco verde (kg)	kg
Esterco verde	kg	Esterco verde (kg)	kg
Esterco verde	ha	Esterco verde (ha)	ha
fatty acid methyl ester	kg	fatty acid methyl ester	kg
Fosfatos e afins	kg	Macronutrientes (kg)	kg
Fungicide	kg	Fungicide	kg
Gesso	kg	Gesso	kg
Grain Barley	kg	Grain Barley (kg)	kg
Grain barley	t	Grain barley (t)	t
Grain Barley, forage	kg	Grain Barley, forage	kg
Grain Barley, organic	kg	Grain Barley, organic	kg

Grain Maize	kg	Grain Maize	kg
Grain Rapeseed	kg	Grain Rapeseed	kg
Grain Rice	kg	Grain Rice	kg
Grain Soybean	kg	Grain Soybean	kg
Grain Sunflower	t	Grain Sunflower (t)	t
Grain Sunflower	kg	Grain Sunflower (kg)	kg
Grain Wheat	kg	Grain Wheat (kg)	kg
Grain Wheat	t	Grain Wheat (t)	t
Grain Wheat spring	t	Grain Wheat spring	t
Grain Wheat winter	t	Grain Wheat winter	t
Grain Wheat, organic	kg	Grain Wheat, organic	kg
Herbicide	kg	Herbicide	kg
Horn meal	kg	Horn meal	kg
Infraestrutura básica fazenda	ha	Infraestrutura básica fazenda	ha
Inseticide	kg	Inseticide	kg
land use change, annual crop	ha	land use change, annual crop	ha
Macronutrientes (kg)	kg	Macronutrientes (kg)	kg
Material inicial produção semente Barley	kg	Material inicial produção semente Barley	kg
Material inicial produção semente Maize	kg	Material inicial produção semente Maize	kg
Material inicial produção semente Rapeseed	kg	Material inicial produção semente Rapeseed	kg
Material inicial produção semente Rice	kg	Material inicial produção semente Rice	kg
Material inicial produção semente Soybean	kg	Material inicial produção semente Soybean	kg
Material inicial produção semente Sunflower	kg	Material inicial produção semente Sunflower	kg
Material inicial produção semente Wheat	kg	Material inicial produção semente Wheat	kg
Micronutrientes	kg	Micronutrientes	kg
naphtha	kg	naphtha	kg
Natural gas	m3	Natural gas	m3
Nitrogenados e afins	kg	Macronutrientes (kg)	kg
NPK compound (NPK 15-15-15)	kg	Macronutrientes (kg)	kg
Occupation, agriculture	m2*a	Occupation, agriculture	m2*a
Occupation, annual crop	m2*a	Occupation, annual crop	m2*a
Occupation, annual crop, irrigated	m2*a	Occupation, annual crop, irrigated	m2*a
Occupation, annual crop, irrigated, intensive	m2*a	Occupation, annual crop, irrigated, intensive	m2*a
Occupation, annual crop, non-irrigated	m2*a	Occupation, annual crop, non-irrigated	m2*a
Occupation, annual crop, non-irrigated, extensive	m2*a	Occupation, annual crop, non-irrigated, extensive	m2*a
Occupation, annual crop, non-irrigated, intensive	m2*a	Occupation, annual crop, non-irrigated, intensive	m2*a
OP aplicação defensivo	ha	OP aplicação defensivo (ha)	ha

OP aplicação defensivo	h	OP aplicação defensivo (h)	h
OP aração	h	OP aração	h
OP arar	h	OP arar	h
OP arar, tillage	ha	OP arar, tillage	ha
OP britagem	h	OP britagem	h
OP capinar	h	OP capinar (h)	h
OP capinar	ha	OP capinar (ha)	ha
OP carregar fardo	Item(s)	OP carregar fardo	Item(s)
OP colheita	ha	OP colheita (ha)	ha
OP colheita	h	OP colheita (h)	h
OP enfardar	Item(s)	OP enfardar (item)	Item(s)
OP enfardar	ha	OP enfardar (ha)	ha
OP espalhar esterco	ha	OP espalhar esterco (ha)	ha
OP espalhar esterco	h	OP espalhar esterco (h)	h
OP espalhar esterco	m3	OP espalhar esterco (m3)	m3
OP espalhar esterco	kg	OP espalhar esterco (kg)	kg
OP fertilização	ha	OP fertilização (ha)	ha
OP fertilização	h	OP fertilização (h)	h
OP fertilização ou aplicação pesticida	h	OP fertilização ou aplicação pesticida	h
OP gradagem	h	OP gradagem	h
OP inundação	h	OP inundação	h
OP Irrigação	m3	OP Irrigação	m3
OP maquinário	kg	OP maquinário	kg
OP processamento sementes oleaginosas	kg	OP processamento sementes oleaginosas	kg
OP rolling	h	OP rolling	h
OP semeadura	ha	OP semeadura (ha)	ha
OP semeadura	h	OP semeadura (h)	h
OP semeadura ou plantar	h	OP semeadura ou plantar	h
OP Soil decompaction	h	OP Soil decompaction	h
OP Soil maintenance	h	OP Soil maintenance	h
OP Soil preparation	h	OP Soil preparation	h
OP swath	ha	OP swath	ha
OP transporte para fazenda	h	OP transporte para fazenda	h
packaging, for fertilisers	kg	packaging, for fertilisers	kg
packaging, for pesticides	kg	packaging, for pesticides	kg
Pesticide	kg	Pesticide	kg
PK compound (NPK 0-22-22)	kg	Macronutrientes (kg)	kg

Plant growth regulator	kg	Plant growth regulator	kg
Potássio e afins	kg	Macronutrientes (kg)	kg
Secagem dos grãos	l	Secagem dos grãos	l
Seed barley	kg	Seed barley	kg
Seed barley, organic	kg	Seed barley, organic	kg
Seed clover	kg	Seed clover	kg
Seed Maize	kg	Seed Maize	kg
Seed Maize, organic	kg	Seed Maize, organic	kg
Seed pea	kg	Seed pea	kg
Seed rapeseed	kg	Seed rapeseed	kg
Seed rapeseed, organic	kg	Seed rapeseed, organic	kg
Seed rice	kg	Seed rice	kg
Seed soybean	kg	Seed soybean	kg
Seed soybean, organic	kg	Seed soybean, organic	kg
Seed sunflower	kg	Seed sunflower	kg
Seed sunflower, organic	kg	Seed sunflower, organic	kg
Seed wheat	kg	Seed wheat	kg
Seed wheat, organic	kg	Seed wheat, organic	kg
Silo para estocar	t	Silo para estocar	t
single superphosphate	kg	Macronutrientes (kg)	kg
Single superphosphate, as 35% Ca(H ₂ PO ₄) ₂ (NPK 0-21-0)	kg	Macronutrientes (kg)	kg
Straw Wheat	kg	Straw Wheat	kg
sulfite	kg	Macronutrientes (kg)	kg
sulfur	kg	Macronutrientes (kg)	kg
Transformation, from annual crop	m2	Transformation, from annual crop	m2
Transformation, from annual crop, irrigated	m2	Transformation, from annual crop, irrigated	m2
Transformation, from annual crop, non-irrigated	m2	Transformation, from annual crop, non-irrigated	m2
Transformation, from annual crop, non-irrigated, extensive	m2	Transformation, from annual crop, non-irrigated, extensive	m2
Transformation, from annual crop, non-irrigated, intensive	m2	Transformation, from annual crop, non-irrigated, intensive	m2
Transformation, from arable land, unspecified use	m2	Transformation, from arable land, unspecified use	m2
Transformation, from forest, intensive, clear-cutting	m2	Transformation, from forest, intensive, clear-cutting	m2
Transformation, from forest, unspecified	m2	Transformation, from forest, unspecified	m2
Transformation, from grassland	m2	Transformation, from grassland	m2
Transformation, from pasture, man made, intensive	m2	Transformation, from pasture, man made, intensive	m2
Transformation, from permanent crop	m2	Transformation, from permanent crop	m2
Transformation, from shrub land, sclerophyllous	m2	Transformation, from shrub land, sclerophyllous	m2
Transformation, to annual crop	m2	Transformation, to annual crop	m2

Transformation, to annual crop, irrigated	m2	Transformation, to annual crop, irrigated	m2
Transformation, to annual crop, irrigated, intensive	m2	Transformation, to annual crop, irrigated, intensive	m2
Transformation, to annual crop, non-irrigated	m2	Transformation, to annual crop, non-irrigated	m2
Transformation, to annual crop, non-irrigated, extensive	m2	Transformation, to annual crop, non-irrigated, extensive	m2
Transformation, to annual crop, non-irrigated, intensive	m2	Transformation, to annual crop, non-irrigated, intensive	m2
Transformation, to arable land, unspecified use	m2	Transformation, to arable land, unspecified use	m2
transport, barge	t*km	transport, barge	t*km
transport, light commercial vehicle	t*km	transport, light commercial vehicle	t*km
transport, container ship	t*km	transport, container ship	t*km
transport, lorry	t*km	transport, lorry	t*km
transport, tractor	t*km	transport, tractor	t*km
transport, tractor and trailer	t*km	transport, tractor and trailer	t*km
transport, train	t*km	transport, train	t*km
Transport, truck 10-20t, EURO4	t*km	transport, lorry	t*km
Trator	kg	Trator	kg
triple superphosphate	kg	Macronutrientes (kg)	kg
Triple superphosphate, as 80% Ca(H ₂ PO ₄) ₂ (NPK 0-48-0)	kg	Macronutrientes (kg)	kg
Urea	kg	Urea	kg
Vapor a partir de gás natural	MJ	Vapor a partir de gás natural	MJ
Vinhaça	kg	Vinhaça	kg
Water, unspecified	m3	Water (m3)	m3
Water, rain	m3	Water (m3)	m3
Water, river	m3	Water (m3)	m3
Water, unspecified natural origin	m3	Water (m3)	m3
Water, well	m3	Water (m3)	m3

- Atributos resultantes após as duas transformações:

Antes de compilar as três bases, foram excluídas 306 amostras do Agrifootprint por serem relativas apenas ao processo de secagem de grãos, não considerando a produção das culturas em fazendas. Essa exclusão permitiu remover 1 atributo de entrada (Vapor a partir de gás natural – MJ) e 4 atributos referentes as saídas (S_Dried Grain Barley, S_Dried Grain Wheat, S_Dried Grain Wheat spring e S_Dried Grain Wheat winter), totalizando 169 atributos ao todo.

Atributos	Unid	Atributos	Unid
-----------	------	-----------	------

ENTRADAS			
Ammonia e afins	kg	OP espalhar esterco (h)	h
Calcário	kg	OP espalhar esterco (ha)	ha
Carbon dioxide, in air	kg	OP espalhar esterco (kg)	kg
Carbon, organic, in soil or biomass stock	kg	OP espalhar esterco (m3)	m3
chemical, inorganic	kg	OP fertilização (h)	h
Colhedeira	kg	OP fertilização (ha)	ha
Composto	kg	OP fertilização ou aplicação pesticida	h
Crop, default, heavy metals uptake (WFLDB 3.1)/GLO U	kg	OP gradagem	h
Diesel	kg	OP inundação	h
Emissões fungicidas	kg	OP Irrigação	m3
Emissões herbicidas	kg	OP maquinário	kg
Emissões inseticidas	kg	OP processamento sementes oleaginosas	kg
Emissões pesticidas	kg	OP rolling	h
Energia da biomassa	MJ	OP semeadura (h)	h
Energia do diesel	MJ	OP semeadura (ha)	ha
Energia elétrica	MJ	OP semeadura ou plantar	h
Energia elétrica (kwh)	kWh	OP Soil decompaction	h
Energia gás natural (MJ)	MJ	OP Soil maintenance	h
Esterco animal (kg)	kg	OP Soil preparation	h
Esterco verde (ha)	ha	OP swath	ha
Esterco verde (kg)	kg	OP transporte para fazenda	h
fatty acid methyl ester	kg	packaging, for fertilisers	kg
Fungicide	kg	packaging, for pesticides	kg
Gesso	kg	Pesticide	kg
Grain Barley (kg)	kg	Plant growth regulator	kg
Grain barley (t)	t	Secagem do grãos	l
Grain Barley, forage	kg	Seed barley	kg
Grain Barley, organic	kg	Seed barley, organic	kg
Grain Maize	kg	Seed clover	kg
Grain Rapeseed	kg	Seed Maize	kg
Grain Rice	kg	Seed Maize, organic	kg
Grain Soybean	kg	Seed pea	kg
Grain Sunflower (kg)	kg	Seed rapeseed	kg
Grain Sunflower (t)	t	Seed rapeseed, organic	kg
Grain Wheat (kg)	kg	Seed rice	kg
Grain Wheat (t)	t	Seed soybean	kg

Grain Wheat, organic	kg	Seed soybean, organic	kg
Herbicide	kg	Seed sunflower	kg
Horn meal	kg	Seed sunflower, organic	kg
Infraestrutura básica fazenda	ha	Seed wheat	kg
Inseticide	kg	Seed wheat, organic	kg
land use change, annual crop	ha	Silo para estocar	t
Macronutrientes (kg)	kg	Straw Wheat	kg
Material inicial produção semente Barley	kg	Transformation, from annual crop	m2
Material inicial produção semente Maize	kg	Transformation, from annual crop, irrigated	m2
Material inicial produção semente Rapeseed	kg	Transformation, from annual crop, non-irrigated	m2
Material inicial produção semente Rice	kg	Transformation, from annual crop, non-irrigated, extensive	m2
Material inicial produção semente Soybean	kg	Transformation, from annual crop, non-irrigated, intensive	m2
Material inicial produção semente Sunflower	kg	Transformation, from arable land, unspecified use	m2
Material inicial produção semente Wheat	kg	Transformation, from forest, intensive, clear-cutting	m2
Micronutrientes	kg	Transformation, from forest, unspecified	m2
naphtha	kg	Transformation, from grassland	m2
Natural gas	m3	Transformation, from pasture, man made, intensive	m2
Occupation, agriculture	m2*a	Transformation, from permanent crop	m2
Occupation, annual crop, irrigated	m2*a	Transformation, from shrub land, sclerophyllous	m2
Occupation, annual crop	m2*a	Transformation, to annual crop	m2
Occupation, annual crop, irrigated, intensive	m2*a	Transformation, to annual crop, irrigated	m2
Occupation, annual crop, non-irrigated	m2*a	Transformation, to annual crop, irrigated, intensive	m2
Occupation, annual crop, non-irrigated, extensive	m2*a	Transformation, to annual crop, non-irrigated	m2
Occupation, annual crop, non-irrigated, intensive	m2*a	Transformation, to annual crop, non-irrigated, extensive	m2
OP aplicação defensivo (h)	h	Transformation, to annual crop, non-irrigated, intensive	m2
OP aplicação defensivo (ha)	ha	Transformation, to arable land, unspecified use	m2
OP aração	h	transport, barge	t*km
OP arar	h	transport, light commercial vehicle	t*km
OP arar, tillage	ha	transport, container ship	t*km
OP britagem	h	transport, lorry	t*km
OP capinar (h)	h	transport, tractor	t*km
OP capinar (ha)	ha	transport, tractor and trailer	t*km
OP carregar fardo	Item(s)	transport, train	t*km
OP colheita (h)	h	Trator	kg
OP colheita (ha)	ha	Urea	kg
OP enfardar (ha)	ha	Vinhaça	kg
OP enfardar (item)	Item(s)	Water (m3)	m3

SAÍDAS			
S_Grain Barley	kg	S_Grain Sunflower, dried	t
S_Grain Barley, feed	kg	S_Grain Sunflower, organic	kg
S_Grain Barley, forage	kg	S_Grain Wheat	kg
S_Grain Barley, organic	kg	S_Grain Wheat, organic	kg
S_Grain Maize	kg	S_Stover Maize	kg
S_Grain Maize, organic	kg	S_Straw Barley	kg
S_Grain Rapeseed	kg	S_Straw rapeseed	kg
S_Grain Rapeseed, organic	kg	S_Straw Rice	kg
S_Grain Rice	kg	S_Straw Soybean	kg
S_Grain Soybean	kg	S_Straw Wheat	kg
S_Grain Soybean, organic	kg	S_Straw Wheat, organic	kg
S_Grain Sunflower	kg		

APÊNDICE – 3C

O subtópico em questão tem por objetivo descrever os motivos para a exclusão de amostras, mescla de atributos e exclusão de atributos que foram realizadas desde a base de dados inicial (v0) até a base de dados final (v5).

Ajustes realizados: base v0 → base v1

Para a transformação da base de dados compilada inicial para a versão v1 foram feitas modificações que envolveram a exclusão de amostras, exclusão de atributos e mescla de atributos.

- **Exclusão de amostras (327):**

- 306 amostras do Agrifootprint foram excluídas por representarem processos de secagem (só havia a entrada de energia elétrica e vapor de água e saída de grão seco).
- 7 amostras foram excluídas do Agribalyse por representarem processos de secagem (entrada de grão e silo e saída de grão seco) ou processos de globalização dos dados (entrada de grãos de diferentes países para fazer a ponderação global).
- 14 amostras foram excluídas do Ecoinvent por representarem processos em que o foco é a saída da palha e não do grão em si (4 conjuntos de dados de cevada, 2 de arroz e 8 de trigo).

- **Exclusão de atributos (33):**

- **Atributos de entrada (25):**
- Grain Sunflower (kg) / Grain Sunflower (t) / transport, tractor (t*km) – atributos com valores zerados em todas as amostras.
- Trator (kg) / colhedeira (kg) / natural gas (m3) / Carbon, organic, in soil or biomass stock (kg) / OP swath (ha) / Silo para estocar – atributos só usados em 1 amostra.
- Straw Wheat (kg) / naphta (kg) / Composto (kg) - atributos usados em apenas 2 amostras.
- horn meal (kg) / Crop, default, heavy metals uptake (WFLDB 3.1)/GLO U (kg) - atributos usados em apenas 4 amostras.
- Plant growth regulator (kg) – atributo usado em apenas 14 amostras de arroz do Agrifootprint.

- Vapor a partir de gás natural - atributo usado apenas em amostras de secagem de grão e não de produção em si.
- Carbon dioxide, in air (kg) / land use change, annual crop (ha) – atributos não presentes nos dados do Agrifootprint (base de dados de maior predominância).
- Grain Wheat (t) / Grain barley (t) – atributo com dado ruidoso, não havia mais nenhum atributo com dado em tonelada.
- Packaging, for fertilisers (kg) / packaging, for pesticides (kg) – atributo pouco relevante para o problema.
- Grain Barley, forage – atributo agregado no atributo Grain Barley (kg).
- Grain Barley, organic – atributo agregado no atributo Grain Barley (kg).
- Grain Wheat, organic – atributo agregado no atributo Grain Wheat (kg).

- **Atributos de saída (8):**
 - S Straw Wheat, organic – atributo com valor zerado em todas as amostras.
 - S Grain Sunflower, dried – atributo utilizado em apenas 1 amostra.
 - S Stover Maize / S Straw Barley / S Straw rapeseed / S Straw Rice / S Straw Soybean / S Straw Wheat – atributos cujo o foco central não é a produção do grão em si (objetivo do presente trabalho).

- **Mescla de atributos (52 -> 13):**
 - Seed barley – seed barley (kg) + Grain Barley (kg) + Material inicial produção semente Barley.
 - Seed Maize – seed maize(kg) + Grain maize (kg) + Material inicial produção semente maize.
 - Seed rapeseed – seed rapeseed (kg) + Grain rapeseed (kg) + Material inicial produção semente rapeseed.
 - Seed rice – seed rice (kg) + Grain rice (kg) + Material inicial produção semente rice.
 - Seed soybean – seed soybean (kg) + Grain soybean (kg) + Material inicial produção semente soybean.
 - Seed sunflower – seed sunflower (kg) + Material inicial produção semente sunflower.
 - Seed wheat – seed wheat (kg) + Grain wheat (kg) + Material inicial produção semente wheat.
 - transport, ship or barge (t*km) – transport, container ship (t*km) + transport, barge (t*km).
 - pesticide (kg) – pesticide (kg) + chemical, inorganic.
 - Occupation, total (m2*a) - Occupation, agriculture + Occupation, annual crop, irrigated + Occupation, annual crop + Occupation, annual crop, irrigated, intensive + Occupation, annual crop, non-irrigated + Occupation, annual crop, non-irrigated, extensive+ Occupation, annual crop, non-irrigated, intensive
 - Transformation, from Total (m2) - Transformation, from annual crop + Transformation, from annual crop, irrigated + Transformation, from annual

- crop, non-irrigated + Transformation, from annual crop, non-irrigated, extensive + Transformation, from annual crop, non-irrigated, intensive + Transformation, from arable land, unspecified use + Transformation, from forest, intensive, clear-cutting + Transformation, from forest, unspecified + Transformation, from grassland + Transformation, from pasture, man made, intensive + Transformation, from permanent crop + Transformation, from shrub land, sclerophyllous
- Transformation, to Total (m2) - Transformation, to annual crop + Transformation, to annual crop, irrigated + Transformation, to annual crop, irrigated, intensive + Transformation, to annual crop, non-irrigated + Transformation, to annual crop, non-irrigated, extensive + Transformation, to annual crop, non-irrigated, intensive + Transformation, to arable land, unspecified use
 - S_Grain Barley (kg) - S_Grain Barley, feed + S_Grain Barley, forage

Ajustes realizados: base v1 → base v2

Para a transformação da base de dados compilada v1 para a versão v2 foram feitas modificações que envolveram a exclusão de amostras, exclusão de atributos e mescla de atributos.

- **Exclusão de amostras (6):**

- 6 amostras da base de dados Agribalyse foram excluídas por apresentarem atributos com entrada negativa com relação ao consumo de diesel e/ou esterco verde.

- **Exclusão de atributos:**

- **Atributos de entrada (37):**
- Infraestrutura básica fazenda (ha) / Vinhaça (kg) – atributo com valor zerado em todas amostras do Agribalyse e Ecoinvent.
- Secagem dos grãos (l) – atributo com valor zerado em todas amostras do Agrifootprint.
- Seed clover (kg) – atributo com valor apenas em duas amostras do Agribalyse.
- Seed pea (kg) – atributo com valor apenas em três amostras do Agribalyse.
- Atributos com operações agrícolas (hora ou hectare de operação) – ao todo foram excluídos 32 atributos de operações agrícolas, visto que os mesmos estavam zerados em todas as amostras do Agrifootprint.

- **Mescla de atributos (13 -> 5):**

- Emissões pesticidas (kg) – Emissões fungicidas + Emissões herbicidas + Emissões inseticidas + Emissões pesticidas.
- Energia elétrica (kwh) – Energia elétrica (kwh) + Energia elétrica (MJ) * 0.277777778 (transformação feita).
- Energia do diesel (MJ) - Energia do diesel (MJ) + diesel (kg) * 43 (transformação feita) + fatty acid methyl ester (biodiesel) * 37,2 (transformação feita).
 - <https://www.antoniolima.web.br.com/arquivos/podercalorifico.htm>
- Água (m3) – water (m3) + OP irrigação (m3).
- Esterco verde (kg) – esterco verde (kg) + esterco verde (há) * 2300 (transformação feita, rendimento de esterco por ha).
 - <https://ecoquery.ecoinvent.org/3.11/cutoff/dataset/569/documentation>

Ajustes realizados: base v2 → base v3

Para a transformação da base de dados compilada v2 para a versão v3 foram feitas apenas modificações que envolveram a mescla de atributos.

- **Mescla de atributos (28 -> 6):**

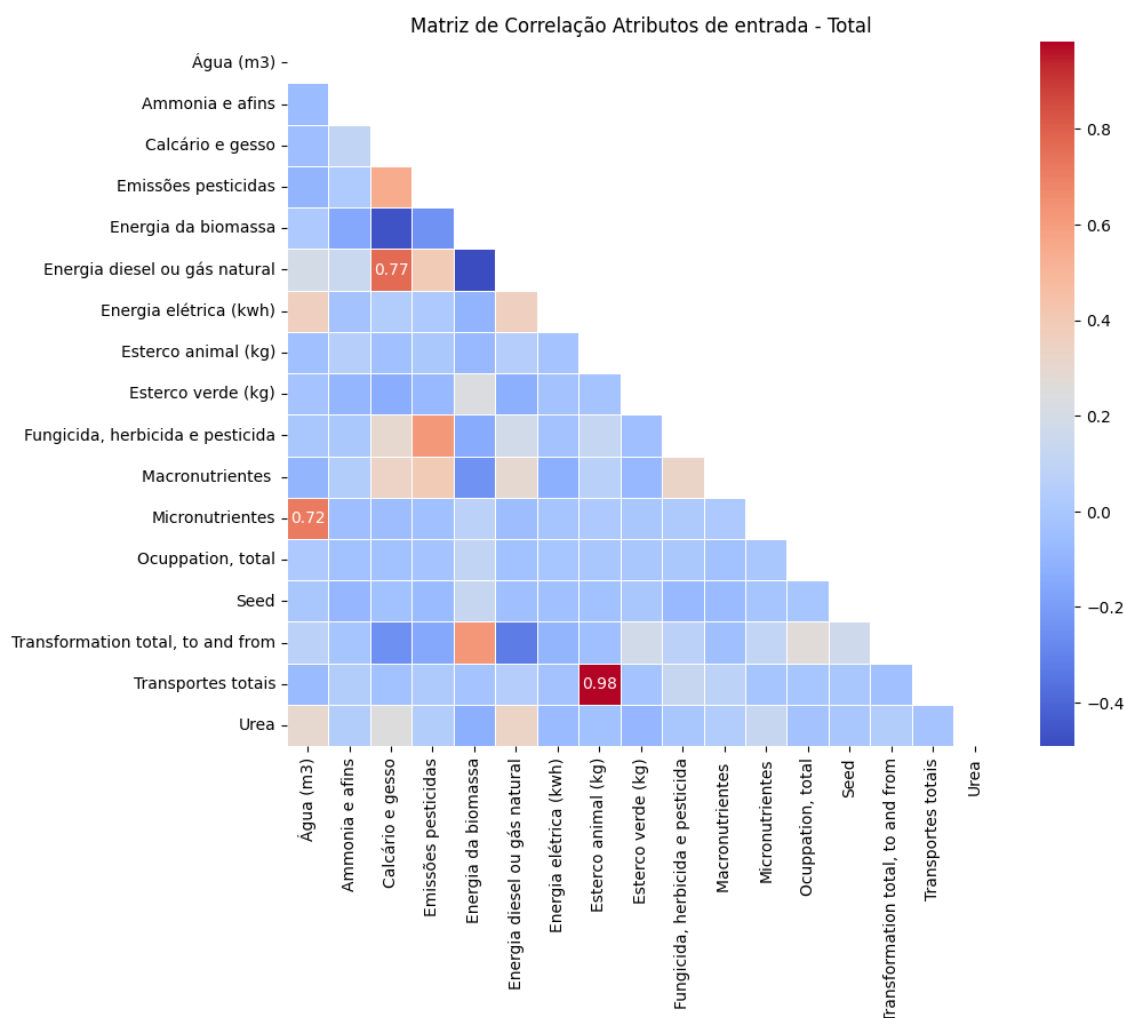
- Energia diesel ou gás natural (MJ) – energia do diesel (MJ) + energia do gás natural (MJ).
- Fungicida, herbicida e pesticida (kg) – fungicide + herbicide + inseticide + pesticide.
- Calcário e gesso (kg) – calcário + gesso.
- Seed (kg) – Seed barley + Seed barley, organic + Seed Maize + Seed Maize, organic + Seed rapeseed + Seed rapeseed, organic + Seed rice + Seed soybean + Seed soybean, organic + Seed sunflower + Seed sunflower, organic + Seed wheat + Seed wheat, organic.
- Transformation total, to and from (m2) – Transformation, from Total + Transformation, to Total.
- Transportes totais (t*km) – transport, light commercial vehicle + transport, ship or barge + transport, lorry + transport, tractor and trailer + transport, train.

Ajustes realizados: base v3 → base v4

Para a transformação da base de dados compilada v3 para a versão v4 foram feitas apenas modificações que envolveram a exclusão de atributos.

- **Exclusão de atributos (3):**

- Transportes totais (kg * km) – atributo excluído por apresentar correlação muito forte com o atributo “Esterco animal” (correlação de 0,98). Logo, optou-se por manter o atributo de esterco animal, por ser mais simples de ser estimado ao considerar as entradas da agricultura para produção de grãos do que estimar os transportes envolvidos.
- Energia diesel ou gás natural (MJ) – atributo excluído por apresentar correlação forte com o atributo “Calcário e gesso” (correlação de 0,77). Logo, optou-se por manter o atributo de calcário e gesso, por ser mais simples de ser estimado ao considerar as entradas da agricultura para produção de grãos do que estimar os o consumo de energia advinda do diesel ou gás natural.
- Água (m3) – atributo excluído por apresentar correlação forte com o atributo “micronutrientes” (correlação de 0,72). Logo, optou-se por manter o atributo de micronutrientes, por ser mais simples de ser estimado ao considerar as entradas da agricultura para produção de grãos do que estimar os o consumo de água.

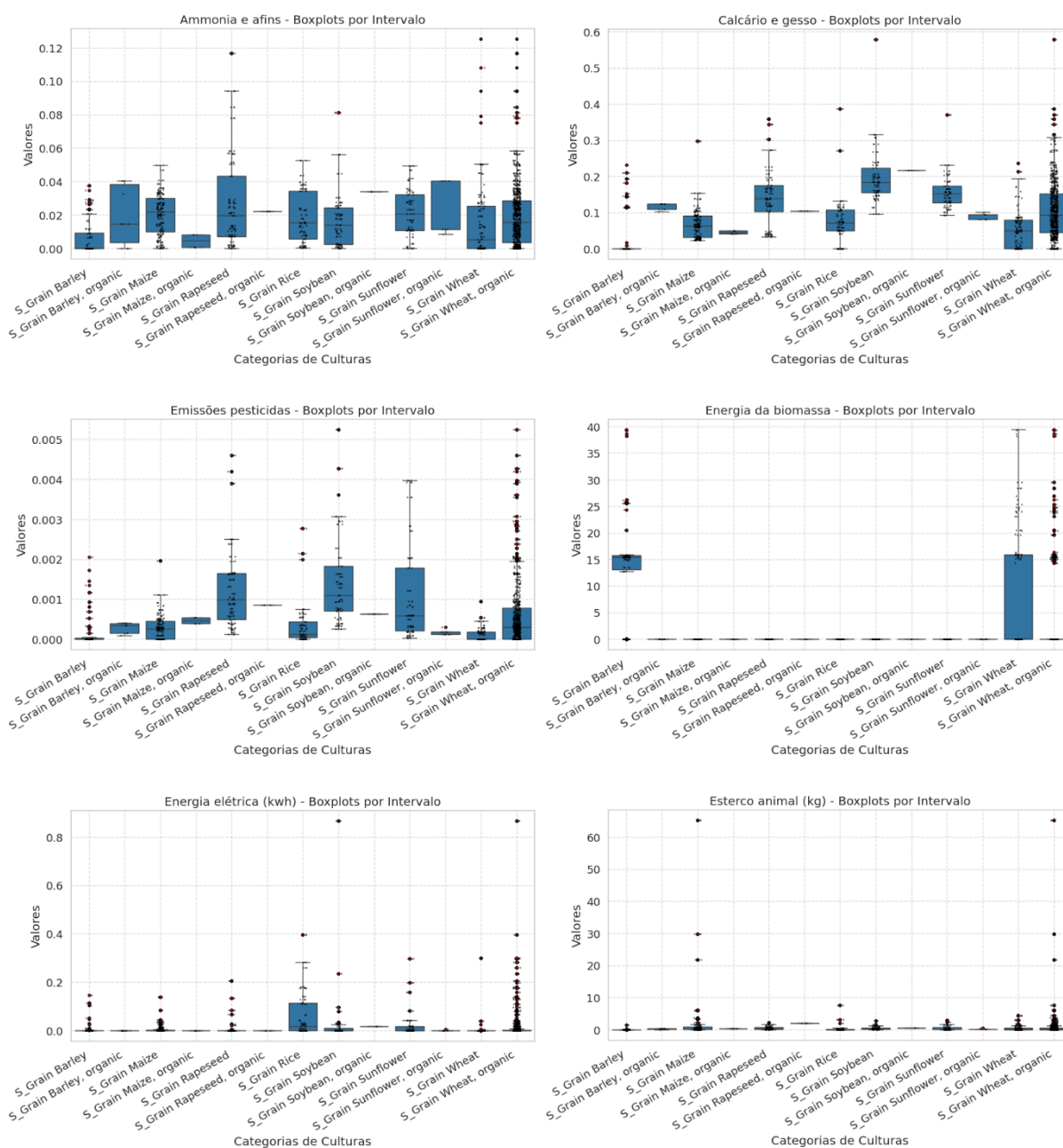


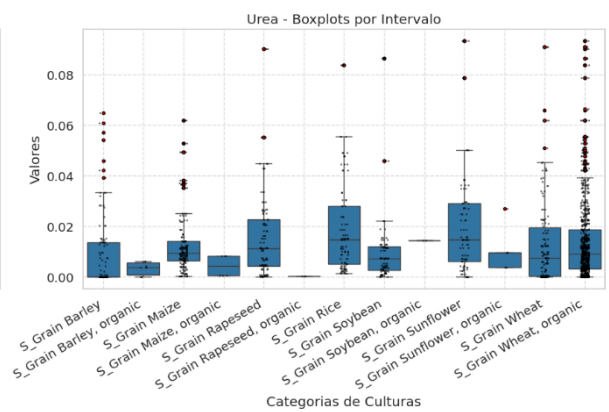
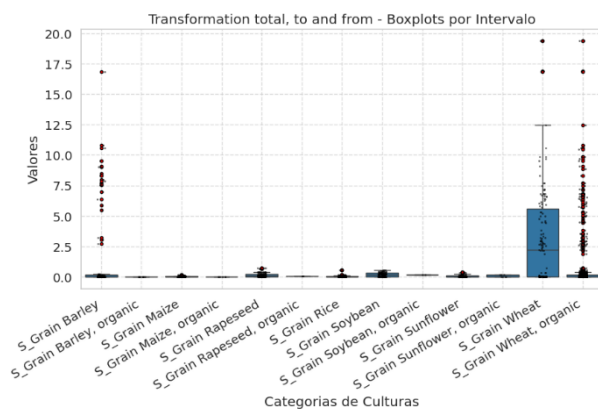
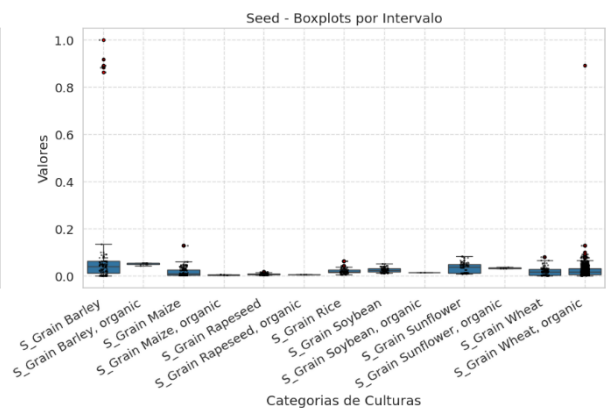
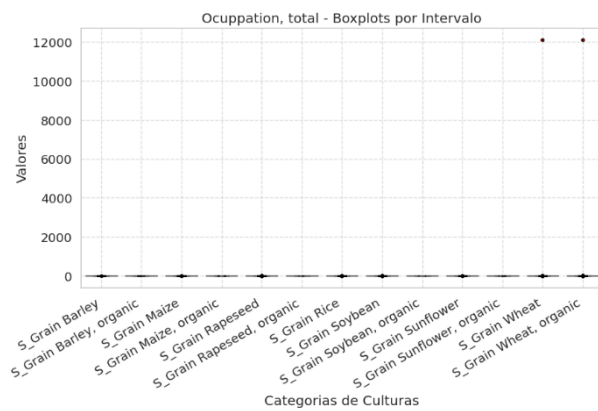
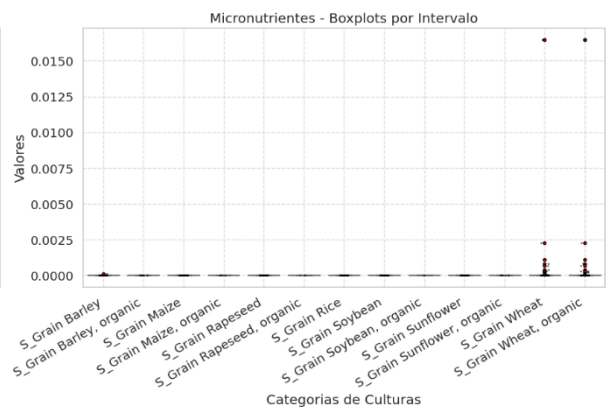
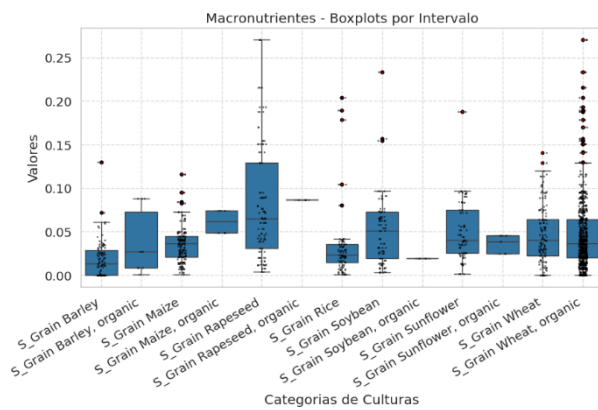
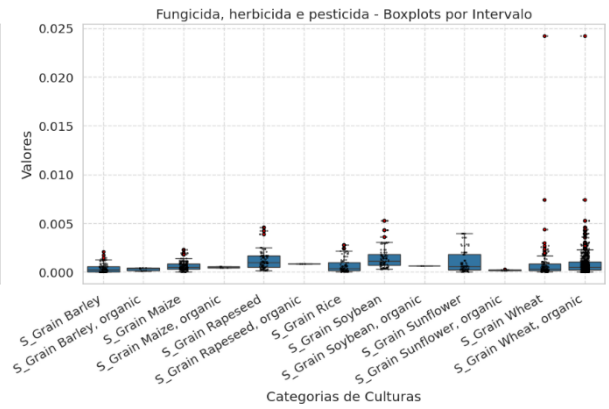
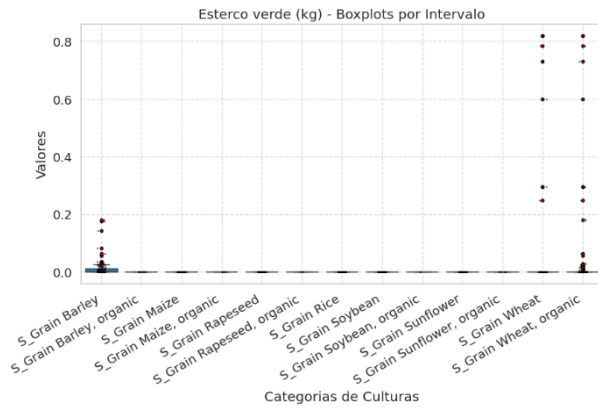
Ajustes realizados: base v4 → base v5

Para a transformação da base de dados compilada v4 para a versão v5 foram feitas apenas modificações que envolveram a exclusão de atributos de saída.

- **Exclusão de atributos (13):**

- 13 atributos binários – os *boxplots* feitos permitiram garantir estabilidade as amostras advindas de diferentes culturas, visto que a variação entre seus atributos não é significativa. Este fator atrelado ao número baixo de amostras por cultura (em torno de 100 amostras) permitiu consolidar a base de dados usando diferentes culturas em uma base única ao invés de segregá-la em bases diferentes por cultura.





A análise dos gráficos permitiu identificar certa estabilidade entre os 14 atributos usados pelas 13 diferentes culturas (27 atributos ao todo na versão da base v4). Nesse sentido, as 13 culturas (variáveis de saída) que estavam sendo utilizadas como atributos binários das amostras foram excluídas, permitindo a chegar em uma base final consolidada (v5) com 886 amostras e 14 atributos.

REFERÊNCIAS

- Aprender Estatística Fácil - AEF (2025). O que é: Regularização Elastic Net. <
<https://estatisticafacil.org/glossario/o-que-e-regularizacao-elastic-net/#:~:text=A%20Regulariza%C3%A7%C3%A3o%20Elastic%20Net%20%C3%A9%20uma%20t%C3%A9cnica%20essencial%20para%20analistas,arsenal%20de%20t%C3%A9nicas%20de%20modelagem>> Acesso em: 18 de março de 2025.
- Agribalyse. (2024). Qu'est-ce que la base de données AGRIBALYSE? <https://doc.agribalyse.fr/documentation/le-programme-agribalyse/introduction>
- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. Disponível em: <https://optuna.org/>
- Ambika. (2023). Naive Bayes in Machine Learning: Explained. Medium. <https://medium.com/@ambika199820/naive-bayes-in-machine-learning-explained-3b70427c6aa5>
- Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13, 281-305.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. *Proceedings of COMPSTAT*, 2010, 177–186.
- Brownlee, J. (2019, May 23). How Much Training Data is Required for Machine Learning? *Machine Learning Process*. <https://machinelearningmastery.com/much-training-data-required-machine-learning/>
- Bruce, Peter; Bruce, Andrew. *Estatística prática para cientistas de dados: 50 conceitos essenciais*. 1ª edição. Rio de Janeiro — RJ: Alta Books, 2019.
- Calvin, K., Dasgupta, D., Krinner, G., Mukherji, A., Thorne, P. W., Trisos, C., Romero, J., Aldunce, P., Barrett, K., Blanco, G., Cheung, W. W. L., Connors, S., Denton, F., Diongue-Niang, A., Dodman, D., Garschagen, M., Geden, O., Hayward, B., Jones, C., ... Ha, M. (2023). IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland. <https://doi.org/10.59327/IPCC/AR6-9789291691647>
- Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Chugani, V. (2024, June 21). A Comprehensive Guide to K-Fold Cross Validation. <https://www.datacamp.com/tutorial/k-fold-cross-validation>
- Colomb, V., Ait Amar, S., Mens, C. B., Gac, A., Gaillard, G., Koch, P., Mousset, J., Salou, T., Tailleur, A., & van der Werf, H. M. G. (2015). AGRIBALYSE the French LCI Database for agricultural products: high quality data for producers and environmental labelling. *OCL*, 22(1), D104. <https://doi.org/10.1051/ocl/20140047>
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. <https://doi.org/10.1007/BF00994018>

CropLife. (2020, August 19). Milho, a evolução de uma cultura milenar. [https://croplifebrasil.org/noticias/milho-a-evolucao-de-uma-cultura-milenar/#:~:text=Com%20uma%20safra%20global%20\(2019,produtor%20desse%20gr%C3%A3o%20no%20mundo.](https://croplifebrasil.org/noticias/milho-a-evolucao-de-uma-cultura-milenar/#:~:text=Com%20uma%20safra%20global%20(2019,produtor%20desse%20gr%C3%A3o%20no%20mundo.)

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems (NeurIPS)*.

Durlinger, B., Tyszler, M., Scholten, J., Broekema, R., Blonk, H. J., Schenck, R., & Huizen, D. (2014). Agri-footprint; A life cycle inventory database covering food and feed production and processing.

Faceli, K., Lorena, A. C., Gama, J., Almeida, T. A. de, & Carvalho, A. C. P. L. F. de. (2022). *Inteligência artificial: uma abordagem de aprendizado de máquina* (LTC, Ed.; 2nd ed.).

Firmino, A. S., Akim, E. K., de Oliveira, J. A., & Silva, D. A. L. (2022). Green machining of aluminum pipes: an integrated approach for eco-efficiency and life cycle assessment in manufacturing systems. *The International Journal of Advanced Manufacturing Technology*, 121(9–10), 6225–6241. <https://doi.org/10.1007/s00170-022-09737-5>

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5). <https://doi.org/10.1214/aos/1013203451>

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350), 320–328. DOI: 10.1080/01621459.1975.10479865

Gomes, P. C. T. (2019). Conheça as Etapas do Pré-Processamento de dados. <https://www.datageeks.com.br/pre-processamento-de-dados/>

GreenDelta. (2024a). About OpenLCA. <https://www.openlca.org/>

GreenDelta. (2024b). Developer tools. <https://www.openlca.org/project/developer-tools/>

Harrison, Matt. *Machine Learning: Guia de referência rápido*. 1ª edição. São Paulo — SP: Novatec Editora Ltda, 2020.

Hillege, L. (2024). What is ecoinvent? – Meet the LCI Database. <https://ecochain.com/blog/what-is-ecoinvent/>

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A Fast-Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7), 1527–1554. <https://doi.org/10.1162/neco.2006.18.7.1527>

Hope, T. M. H. (2020). Linear regression. In *Machine Learning* (pp. 67–81). Elsevier. <https://doi.org/10.1016/B978-0-12-815739-8.00004-3>

IBICT - Instituto Brasileiro de Informação em Ciência e Tecnologia. (2024, August). Home SICV Brasil. <https://sicv.acv.ibict.br/>

IBM. (2023). What is random forest? <https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,Decision%20trees>

Iglewicz, B., & Hoaglin, D. C. (1993). How to detect and handle outliers. This book discusses methods for detecting outliers, including the Z-score, and is a widely cited source on the topic.

Joyce, J. (2003). Bayes' Theorem. *Stanford Encyclopedia of Philosophy (SEP)*. <https://plato.stanford.edu/entries/bayes-theorem/>

Kellens, K., Dewulf, W., Overcash, M., Hauschild, M. Z., & Duflou, J. R. (2012a). Methodology for systematic analysis and improvement of manufacturing unit process life cycle inventory (UPLCI) CO2PE! initiative (cooperative effort on process emissions in manufacturing). Part 2: case studies. *The International Journal of Life Cycle Assessment*, 17(2), 242–251. <https://doi.org/10.1007/s11367-011-0352-0>

Kellens, K., Dewulf, W., Overcash, M., Hauschild, M. Z., & Duflou, J. R. (2012b). Methodology for systematic analysis and improvement of manufacturing unit process life-cycle inventory (UPLCI)—CO2PE! initiative (cooperative effort on process emissions in manufacturing). Part 1: Methodology description. *The International Journal of Life Cycle Assessment*, 17(1), 69–78. <https://doi.org/10.1007/s11367-011-0340-4>

Khan, S., & McCormick, W. (2014). Statistical methods for outlier detection. *Journal of Applied Statistics*. This paper provides an overview of statistical methods, including the Z-score, for outlier detection in data analysis.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence (IJCAI)*, 1137-1143.

Learn Statistics Easily. (2023, May 5). Coefficient of Determination vs. Coefficient of Correlation in Data Analysis. <https://statisticseasily.com/coefficient-of-determination-vs-coefficient-of-correlation/>

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324. <https://doi.org/10.1109/5.726791>

Lopes, L. F. D. (2016). *Métodos Quantitativos*. 1ª edição. Santa Maria: Universidade Federal de Santa Maria.

Martínez-Ramón, N., Calvo-Rodríguez, F., Iribarren, D., & Javier Dufour. (2024). Frameworks for the application of machine learning in life cycle assessment for process modeling. *Cleaner Environmental Systems*, 100221. <https://doi.org/10.1016/j.cesys.2024.100221>

Maulud, D. H., & Abdulazeez, A. M. (2020). Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*, 1(2), 140–147.

Muhajir, I. (2019). K-Neighbors Regression Analysis in Python. <https://medium.com/analytics-vidhya/k-neighbors-regression-analysis-in-python-61532d56d8e4>

Murel, J., & Kavlakoglu, E. (2025). O que é redução da dimensionalidade? IBM. <https://www.ibm.com/br-pt/think/topics/dimensionality-reduction>

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics*, 18(6), 275–285. <https://doi.org/10.1002/cem.873>

Naser, A. Z., Defersha, F., Xu, X., & Yang, S. (2023). Automating life cycle assessment for additive manufacturing with machine learning: Framework design, dataset buildup, and a case study. *Journal of Manufacturing Systems*, 71, 504–526. <https://doi.org/10.1016/j.jmsy.2023.10.009>

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7. <https://doi.org/10.3389/fnbot.2013.00021>

Oliveira, J. A. de, Silva, D. A. L., Puglieri, F. N., & Saavedra, Y. M. B. (2021). *Life Cycle Engineering and Management of Products: theory and practice* (Springer Cham, Ed.; 1st ed., Vol. 1).

Our World in Data. (2024). Datas (browse by topic). Oxford. <https://ourworldindata.org/>

Pastega, G. (2023, April 11). Outliers: Como definir, detectar e tratar. Medium.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Permai, S. D., & Tanty, H. (2018). Linear regression model using bayesian approach for energy performance of residential building. *Procedia Computer Science*, 135, 671–677. <https://doi.org/10.1016/j.procs.2018.08.219>

Prenner, J. A. A. Robbes, R. (2022). Making the most of small Software Engineering datasets with modern machine learning. *IEEE Transactions on Software Engineering*, 1–1. <https://doi.org/10.1109/TSE.2021.3135465>

Pypi. (2019). olca-ipc 0.0.8. [https://pypi.org/project/olca-ipc/0.0.8/#:~:text=openLCA%20provides%20an%20implementation%20of,from%20standard%20Python%20\(Cpython%20v3](https://pypi.org/project/olca-ipc/0.0.8/#:~:text=openLCA%20provides%20an%20implementation%20of,from%20standard%20Python%20(Cpython%20v3)

Qi, J. C. (2023). Use Guide: GaBi. <https://www.tiangong.earth/post/use-guide-gabi>

Rodrigues, T. O., Belizario-Silva, F., Braga, T. E. N., & da Silveira Folegatti Matsuura, M. I. (2021). LCA—Life Cycle Inventory Analysis and Database. In *Life Cycle Engineering and Management of Products* (pp. 71–93). Springer International Publishing. https://doi.org/10.1007/978-3-030-78044-9_4

Romeiko, X. X., Zhang, X., Pang, Y., Gao, F., Xu, M., Lin, S., & Babbitt, C. (2024). A review of machine learning applications in life cycle assessment studies. In *Science of the Total Environment* (Vol. 912). Elsevier B.V. <https://doi.org/10.1016/j.scitotenv.2023.168969>

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386–408. <https://doi.org/10.1037/h0042519>

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533–536. <https://doi.org/10.1038/323533a0>

Salla, J.V.E., de Almeida, T.A. & Silva, D.A.L. Integrating machine learning with life cycle assessment: a comprehensive review and guide for predicting environmental impacts. *Int J Life Cycle Assess* (2025). <https://doi.org/10.1007/s11367-025-02437-8>

Saunders, C., Gammerman, A., & Vovk, V. (1998). Ridge regression learning algorithm in dual variables. *ICML'98: Proceedings of the Fifteenth International Conference on Machine Learning*, 515–521.

Scikitlearn. (2024). 1.10. Decision Trees. <https://scikit-learn.org/stable/modules/tree.html#id2>

Scikit-learn Developers. (2024). Scikit-learn: Machine Learning in Python. Retrieved from <https://scikit-learn.org>

Scikit-learn developers. Underfitting vs. Overfitting. Disponível em: <https://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html>. Acesso em: 18 de março de 2025.

Scikit-learn developers. Regression metrics. Disponível em: <https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics>. Acesso em: 18 de março de 2025.

Scikit-learn developers. Cross-validation: evaluating estimator performance. Disponível em: <https://scikit-learn.org/stable/modules/cross_validation.html>. Acesso em: 18 de março de 2025.

Silva, D. A. L., Nunes, A. O., Piekarski, C. M., Moris, V. A. S., Souza, L. S. M., & Rodrigues, T. O. (2019). Why using different Life Cycle Assessment software tools can generate different results for the same product system? A cause–effect analysis of the problem. *Sustainable Production and Consumption*, 20, 304–315. <https://doi.org/10.1016/j.spc.2019.07.005>

Simapro. (2024). Agri-footprint database. <https://simapro.com/products/agri-footprint-database/>

Sipper, M. (2022). High Per Parameter: A Large-Scale Study of Hyperparameter Tuning for Machine Learning Algorithms. *Algorithms*, 15(9), 315. <https://doi.org/10.3390/a15090315>

Soares, L. A. P. (2021). ACV Comparativa entre Processos de Fresamento de Engrenagens Automotivas com Lubrificação Convencional e MQL. UFSCar .

Song, Y., Liang, J., Lu, J., & Zhao, X. (2017). An efficient instance selection algorithm for k nearest neighbor regression. *Neurocomputing*, 251, 26–34. <https://doi.org/10.1016/j.neucom.2017.04.018>

Stewart, K. (2024). mean squared error. <https://www.britannica.com/science/mean-squared-error>

Teodosio, B., Wasantha, P. L. P., Yaghoubi, E., Guerrieri, M., Fragomeni, S., & van Staden, R. (2023). Environmental, economic, and serviceability attributes of residential foundation slabs: A comparison between waffle and stiffened rafts using multi-output deep learning. *Journal of Building Engineering*, 80. <https://doi.org/10.1016/j.job.2023.107983>

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

United Nations Environment Programme, & International Resource Panel. (2024). *Global Resources Outlook 2024 - Bend the trend: Pathways to a Liveable Planet as Resource Use Spikes*. <https://wedocs.unep.org/20.500.11822/44901>

Vargas, V., Coser, T., Souza, T. R., Brentrup, F., Hesani, M., Neves, C., Bungener, S., Otto, R., Carvalho, M. C. S., Guelfi, D., & Cantarella, H. (2023). Carbon Footprint in Agriculture: Insights, towards neutral crop production and industry integration. <https://ainfo.cnptia.embrapa.br/digital/bitstream/doc/1161635/1/informacoes-agronomicas-2023.pdf>

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, 11(1), 233. <https://doi.org/10.1038/s41467-019-14108-y>

Vinutha, H. P., Poornima, B., & Sagar, B. M. (2018). Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset (pp. 511–518). https://doi.org/10.1007/978-981-10-7563-6_53

Vittia. (2024). Milho. <https://vittia.com.br/milho/#:~:text=A%20cultura%20do%20milho%20%C3%A9,grande%20destaque%20na%20economia%20agr%C3%ADcola.>

Wernet, G., Bauer, C., Steubing, B., Reinhard, J., Moreno-Ruiz, E., & Weidema, B. (2016). The ecoinvent database version 3 (part I): overview and methodology. *The International Journal of Life Cycle Assessment*, 21(9), 1218–1230. <https://doi.org/10.1007/s11367-016-1087-8>

Whitfield, B. (2024, November 20). Introduction to the Correlation Matrix. <https://builtin.com/data-science/correlation-matrix>

Zhu, X., Ho, C. H., & Wang, X. (2020). Application of Life Cycle Assessment and Machine Learning for High-Throughput Screening of Green Chemical Substitutes. *ACS Sustainable Chemistry and Engineering*, 8(30), 11141–11151. <https://doi.org/10.1021/acssuschemeng.0c02211>

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.

4 CONSIDERAÇÕES FINAIS

Os estudos da literatura analisados no primeiro artigo abordaram diferentes possibilidades de integração entre ACV e AM, utilizando diferentes técnicas de modelagem para prever impactos ambientais, eficiência energética e rentabilidade econômica em diversos setores da economia.

Lee et al. (2020) explorou formas rápidas de quantificar os impactos potenciais em acidificação e em aquecimento global para o milho nos EUA. Para isso, foram coletados dados em 874 a 974 condados do Meio-Oeste dos EUA, ao longo de um período de 9 anos (dados de 2000 e 2008), sobre clima, características do solo e práticas agrícolas, incluindo temperatura média mensal, precipitação, proporção de argila, areia, teor de matéria orgânica e práticas de manejo do solo, como atributos.

O estudo utilizou um modelo de árvore de regressão impulsionada (BRT) com distribuição Gaussiana para desenvolver o *pipeline* de treinamento e gerar as previsões para os impactos ambientais da produção de milho entre 2022 e 2100. Os resultados indicaram que os impactos ambientais do ciclo de vida do milho aumentarão sob todos os quatro cenários de emissões futuras, com desempenho para o coeficiente de determinação (R^2) para a previsão dos impactos do ciclo de vida no potencial de aquecimento global e a eutrofização variando de 0,78 a 0,82, indicando capacidade do modelo em aprender com os dados (Lee et al, 2020).

De forma similar, o primeiro teste detalhado no segundo artigo também optou por analisar a produção de milho. No entanto, o número de amostras foi bem mais restrito, utilizando 15 conjuntos de dados brasileiros sobre o cultivo do milho para o *pipeline* de treinamento dos modelos.

Embora o R^2 nos resultados de treinamento seja alto (variando de 0,8725 a 0,9325), refletindo uma boa explicação da variabilidade dos dados para as diversas categorias de impacto (*ecotoxicity: marine; energy resources: non-renewable; material resources: metals/minerals; ozone depletion*), a discrepância entre os erros de treinamento e validação representa um ponto de atenção. De modo geral, a análise dos resultados e o comportamento observado nos modelos indicam que tanto o *overfitting* quanto o *underfitting* podem estar relacionados ao baixo número de amostras disponíveis para treinamento e teste, sendo insuficiente para o aprendizado dos modelos.

Com uma fronteira mais restrita, abordagem *gate-to-gate*, Nabavi-Palesare et al. (2019) explorou os aspectos energéticos, econômicos e ambientais do processamento de arroz branco em fábricas da província de Guilan, no Irã. Para tal, utilizou informações obtidas em 60 fábricas

de beneficiamento de arroz obtidas por meio de entrevistas diretas com trabalhadores e especialistas, totalizando seis atributos (trabalho humano, maquinário, eletricidade, gás natural, nylon e transporte).

Para prever o lucro econômico, a energia gerada e o impacto no aquecimento global, foi desenvolvido um modelo baseado em sistema de inferência neuro-fuzzy adaptativo (ANFIS) utilizando um algoritmo híbrido de aprendizado. Os coeficientes de determinação na previsão da energia gerada, do lucro econômico e do aquecimento global foram de 0,911, 0,978 e 0,964, respectivamente. Tais resultados demonstraram a utilidade do modelo ANFIS para auxiliar o gerenciamento de longo prazo na previsão de diversos indicadores ambientais, energéticos e econômicos em sistemas de produção de alimentos em larga escala (Nabavi-Palesarei et al, 2019).

De forma similar ao realizado por Nabavi-Palesarei et al. (2019), o segundo teste tratado no segundo artigo também optou por explorar conjuntos de dados dentro do ambiente fabril (portão-a-portão ou *gate-to-gate*). Para isso, foram utilizados conjuntos de dados brasileiros parametrizados sobre os processos de usinagem de alumínio e aço. Por serem *datasets* parametrizados, foi possível gerar amostras de forma simples, apenas alterando os valores das variáveis independentes (atributos).

Os resultados demonstraram que o modelo de rede neural artificial (ANN) foi o melhor entre todos os modelos de regressão testados, com R^2 entre 0,7228 e 0,9690 (usinagem do alumínio) e entre 0,7148 e 0,9493 (usinagem do aço). A superioridade do modelo ANN em relação aos outros modelos permitiu destacar sua eficácia em prever com certo nível de precisão, sugerindo que ajustes adicionais ou maior volume de dados poderiam melhorar ainda mais a generalização para este segundo conjunto. Embora os resultados sejam positivos, o teste realizado sugeriu a necessidade de avaliar o sobreajuste dos modelos.

Modelos ANN também foram utilizados por Nabavi-Palesarei et al. (2016a), para estimar produtividade e impacto ambiental na produção de kiwi, com valores de R^2 variando entre 0,987 e 0,9992. Kaab et al. (2019) também utilizaram ANN para prever a geração de energia e impacto ambiental na cultura da cana-de-açúcar.

Ghasemi-Mobtaker et al (2022) utilizou a regressão linear, redes neurais artificiais (ANNs) e sistemas adaptativos de inferência neuro-difusa (ANFIS) para prever a energia de saída, o lucro econômico e o potencial de aquecimento global (GWP) da produção de trigo. Para tal, foram selecionadas aleatoriamente 75 fazendas de trigo localizadas na região central da província de Hamadan, Irã, e os dados foram coletados por meio de entrevistas orais. Foram considerados dez insumos: mão de obra humana (X1), maquinaria (X2), nitrogênio (X3),

fosfato (X4), esterco (X5), diesel (X6), biocidas (X7), água (X8), sementes (X9) e eletricidade (X10).

As variáveis independentes consideradas na análise foram os dez insumos utilizados no cultivo do trigo, enquanto as variáveis dependentes (atributos-alvo) foram sete resultados, incluindo produtividade (grão de trigo e palha), energia de saída (grão de trigo e palha), valor total da produção, retorno líquido e GWP. De forma geral, os resultados da modelagem ilustraram que o R^2 variou entre 0,264 e 0,978 na regressão linear, entre 0,313 e 0,954 na estrutura da ANN com duas camadas ocultas e entre 0,520 e 0,962 no ANFIS estruturado em três níveis (Ghasemi-Mobtaker et al, 2022), indicando oportunidades de realizar tal integração.

O terceiro teste realizado no segundo artigo permitiu uma análise mais completa da possibilidade de integração ACV+AM, integrando conjuntos de dados agrícolas de três bases diferentes (Agribalyse, Agrifootprint e Ecoinvent). Além disso, foram sugeridas abordagens para a compilação de amostras e redução de dimensionalidade sem perder a qualidade dos dados, indo de 169 atributos até 14 atributos com R^2 médio para as 11 categorias de impacto variando 0,8085 (v0 da base – 169 atributos) para 0,7818 (v5 da base – 14 atributos).

A partir dos resultados, observou-se que os modelos baseados em árvores, como GBM e XGBoost, alcançaram valores de R^2 superiores a 0,89, evidenciando a robustez dessas abordagens para prever variáveis de interesse, especialmente quando combinadas com estratégias de otimização como Optuna.

De forma geral, os resultados do terceiro teste permitem concluir que a robustez dos modelos baseados em árvores e a importância da otimização dos hiperparâmetros são fundamentais para balancear a precisão preditiva e a generalização dos modelos. Além disso, modelos como *Kernel Ridge Regression* (KRR), *Stochastic Gradient Descent* (SGD), *Elastic Net* (E_net) e redes neurais artificiais (ANN) mostraram-se mais sensíveis ao ajuste de hiperparâmetros, frequentemente resultando em valores negativos de R^2 .

Nesse sentido, foi possível observar que o processo de escolha das técnicas de modelagem precisa ser combinado, ou seja, levar em consideração estratégias avançadas para avaliação de desempenho nas etapas de validação e ajuste (avaliar *overfitting* e *underfitting*) e exigência computacional, a fim de definir um modelo com melhor performance nos diferentes sentidos.

Modelo de suporte a decisão

O presente subtópico pretende explicitar o modelo de suporte de decisão desenvolvido ao longo do trabalho. O modelo em questão pode ser descrito como um *pipeline* de treinamento e testagem dos modelos de aprendizado de máquina (AM) para ser integrado aos conjuntos de dados presentes em bases de avaliação do ciclo de vida (ACV).

Nesse sentido, foram seguidas as etapas propostas no *pipeline* da Figura 3.1:

- I. Dados: definição das fontes de dados (bases de dados de ACV) a serem utilizadas a depender do objetivo do estudo.
- II. Dados-alvo: escolha dos *datasets* (conjuntos de dados) a serem usados dentro de cada uma das bases de dados.
- III. Pré-processamento: conjunto de atividades para a preparação, organização e estruturação dos dados, podendo englobar a limpeza de dados (atributos com dados faltantes e dados com ruído), a transformação de dados (envolvendo a normalização e a seleção de atributos) e a redução de dados (englobando a clusterização, redução de dimensionalidade, amostragem, entre outros).
- IV. Treinamento dos modelos e ajustes dos parâmetros: definição dos modelos de AM a serem utilizados, bem como o ajuste dos parâmetros para potencializar os resultados.
 - a. Abordagens como *leave-one-out* (LOO) ou *Crossvalidation* (CV) podem ser definidas em virtude da quantidade de amostras disponíveis.
 - b. Abordagens para definição de hiperparâmetros como GridSearch e Optuna podem ser preferidas em virtude da exigência computacional.
- V. Análise dos modelos: processo de análise dos resultados obtidos e da performance de cada um dos modelos de AM utilizando o R^2 (coeficiente de determinação) e MSE (erro quadrático médio). Tal abordagem, possibilitou analisar os erros inerentes a cada um dos modelos (MSE), bem como observar a capacidade de ajuste dos modelos aos dados observados.
- VI. Validação e feedback: etapa de validação dos resultados obtidos e definição dos melhores modelos de AM em comparação aos achados da literatura.

5 CONCLUSÕES

A agricultura é um setor de suma importância no contexto mundial, sendo responsável pela produção de alimentos que são essenciais para a sobrevivência humana e também desempenhando um papel crucial na economia global, principalmente em países como o Brasil, onde a agricultura funciona como um dos principais pilares para o desenvolvimento econômico e social.

Além de sua relevância econômica, a agricultura tem uma estreita relação com as questões ambientais, uma vez que os impactos gerados por práticas agrícolas podem afetar tanto o meio ambiente quanto o desempenho das próprias culturas. Em um cenário global onde as mudanças climáticas são uma realidade crescente, a busca por soluções mais sustentáveis e eficientes para a produção agrícola é cada vez mais urgente. Neste contexto, este trabalho se propôs a explorar métodos de aprendizado de máquina (AM) para prever potenciais impactos ambientais de diferentes processos agrícolas, com o objetivo de proporcionar uma ferramenta mais precisa e ágil para a tomada de decisão.

Os objetivos centrais da pesquisa voltaram-se para atingir confiabilidade na predição de impactos ambientais potenciais. Isso tudo, por meio da integração entre Avaliação do Ciclo de Vida (ACV) e AM, aplicando o *pipeline* de treinamento dos modelos em diferentes conjuntos de dados. Os estudos realizados derivaram em dois artigos científicos. O primeiro abordou a revisão da literatura sobre as metodologias de AM mais adequadas à predição de impactos ambientais, fornecendo uma visão crítica sobre as ferramentas e identificando as lacunas existentes na integração dessas duas abordagens.

A partir dessa revisão, foi possível destacar as Redes Neurais Artificiais (ANN) e as Florestas Aleatórias (RF) como os métodos mais promissores e com uso crescente em pesquisas na área. O segundo artigo, por sua vez, focou sua aplicação na parte prática, realizando testes com diferentes modelos de AM em conjuntos de dados de ACV de diferentes origens e com diferentes características.

Este artigo não apenas validou as hipóteses levantadas no primeiro, mas também trouxe *insights* práticos sobre os desafios enfrentados ao lidar com bases de dados reduzidas, integrar bases de dados diferentes, passos metodológicos utilizados para tal e destacando sobre a importância de ajustar os dados para o treinamento e os hiperparâmetros para obter uma boa generalização dos modelos.

Com base nos resultados obtidos, ficou claro que a qualidade e a quantidade dos dados são fatores essenciais para o sucesso da aplicação de AM em ACV. O primeiro artigo concluiu

que a utilização de modelos de aprendizado de máquina oferece grande potencial para aprimorar a precisão das previsões dos impactos ambientais, especialmente quando os modelos são alimentados por bases de dados estruturadas e abrangentes, e ainda estabeleceu uma hierarquia para os modelos: ANN > RF > SVM > LR > XGBoost > DT > ANFIS > GBM > KNN > MLR > Regressão bayesiana.

No entanto, os resultados obtidos com o segundo artigo foram diferentes um pouco da hierarquia observada na literatura. Embora os modelos mais avançados, como o XGBoost e o GBM, tenham apresentado um bom desempenho, a escolha do modelo deve ser feita com cuidado, levando em consideração o equilíbrio entre a capacidade preditiva, custo computacional, a qualidade dos dados utilizados, bem como atenção às possibilidades de subajuste (*underfitting*) e sobreajustes (*overfitting*) dos modelos. Portanto, a aplicação de AM em ACV é promissora, mas exige a atenção para a qualidade dos dados, a escolha cuidadosa do modelo e a gestão dos recursos computacionais.

Além disso, a consideração dos fatores regionais e locais se mostrou um ponto importante para a melhoria da acurácia dos modelos. O segundo artigo destacou que o uso de bases de dados com *datasets* globais pode afetar a representatividade da realidade local, uma vez que variáveis como precipitação, tipo de solo, radiação solar e outros fatores regionais têm um grande impacto no desempenho das culturas.

Esse fator observado, reforça a necessidade de desenvolver modelos que possam ser parametrizados com parâmetros que permitam substituir variáveis regionais específicas, viabilizando ajustes rápidos e eficientes que melhorem a precisão das previsões. A adaptação dos dados de ACV para refletir as condições locais é essencial para garantir que os modelos de AM possam ser aplicados de forma eficaz em diferentes contextos agrícolas, fornecendo resultados mais realistas e confiáveis e facilitando o processo de geração de amostras, já que o volume de dados pode melhorar o desempenho dos modelos.

O avanço da pesquisa nesta área depende não apenas do aprimoramento das metodologias de modelagem, mas também do desenvolvimento de bases de dados mais robustas e padronizadas. A integração de ACV e AM se beneficia significativamente quando os dados são acessíveis, bem estruturados, atualizados com informações relevantes e possuem extensões (e.g., ILCD, J-SON) que facilitem sua importação e evite perdas no processo.

De forma geral, o trabalho traz uma contribuição para a área de avaliação ambiental e sustentabilidade agrícola, oferecendo novas perspectivas sobre o uso de AM para prever impactos ambientais e destacando a importância de dados de alta qualidade, modelos adaptáveis e parametrização de inventários. Ao integrar a análise de dados ambientais e a previsão de

impactos por meio do aprendizado de máquina, espera-se poiar o processo de criação de ferramentas que sejam utilizadas por usuários não-especialistas em ACV e AM para tomar decisões de forma rápida no contexto agrícola, com certo nível de confiabilidade e com respaldos ambientais.

Portanto, os objetivos e as hipóteses propostas no início desta pesquisa foram confirmados, especialmente no que se refere à utilidade dos modelos de AM para a predição de impactos ambientais em diferentes cenários. As conclusões dos artigos confirmam a importância de se trabalhar com bases de dados robustas e bem estruturadas, e demonstraram a necessidade de integrar variáveis regionais específicas e parametrização para melhorar a precisão das previsões.

REFERÊNCIAS

- EPD Brasil. (2024). Programa de declarações ambientais de tipo III. <https://www.epdbrasil.com.br/>
- Faceli, K., Lorena, A. C., Gama, J., Almeida, T. A. de, & Carvalho, A. C. P. L. F. de. (2022). *Inteligência artificial: uma abordagem de aprendizado de máquina* (LTC, Ed.; 2nd ed.).
- Ghasemi-Mobtaker, H., Kaab, A., Rafiee, S., & Nabavi-Pelesaraei, A. (2022). A comparative of modeling techniques and life cycle assessment for prediction of output energy, economic profit, and global warming potential for wheat farms. *Energy Reports*, 8, 4922-4934. <https://doi.org/10.1016/j.egy.2022.03.184>
- Hypeness. (2019). Preocupação ambiental já mudou hábitos de 75% dos consumidores, aponta levantamento. <https://www.hypeness.com.br/2019/12/preocupacao-ambiental-ja-mudou-habitos-de-75-dos-consumidores-aponta-levantamento/>
- IBM. (2024a). O que é Aprendizado não Supervisionado? <https://www.ibm.com/br-pt/topics/unsupervised-learning>
- IBM. (2024b). O que é aprendizado supervisionado? <https://www.ibm.com/br-pt/topics/supervised-learning>
- ISO 14040:2006 Environmental Management — Life Cycle Assessment — Principles and Framework (2006).
- Kaab, A., Sharifi, M., Mobli, H., Nabavi-Pelesaraei, A., & Chau, K. wing. (2019). Combined life cycle assessment and artificial intelligence for prediction of output energy and environmental impacts of sugarcane production. *Science of the Total Environment*, 664, 1005–1019. <https://doi.org/10.1016/j.scitotenv.2019.02.004>
- Koyamparambath, A., Adibi, N., Szablewski, C., Adibi, S. A., & Sonnemann, G. (2022). Implementing Artificial Intelligence Techniques to Predict Environmental Impacts: Case of Construction Products. *Sustainability (Switzerland)*, 14(6). <https://doi.org/10.3390/su14063699>
- Lee, E. K., Zhang, W.-J., Zhang, X., Adler, P. R., Lin, S., Feingold, B. J., Khwaja, H. A., & Romeiko, X. X. (2020). Projecting life-cycle environmental impacts of corn production in the U.S. Midwest under future climate scenarios using a machine learning approach. *Science of The Total Environment*, 714, 136697. <https://doi.org/10.1016/j.scitotenv.2020.136697>
- Martínez-Ramón, N., Calvo-Rodríguez, F., Iribarren, D., & Javier Dufour. (2024). Frameworks for the application of machine learning in life cycle assessment for process modeling. *Cleaner Environmental Systems*, 100221. <https://doi.org/10.1016/j.cesys.2024.100221>
- McKinsey&Company. (2023). Consumers care about sustainability—and back it up with their wallets. <https://www.mckinsey.com/industries/consumer-packaged-goods/our-insights/consumers-care-about-sustainability-and-back-it-up-with-their-wallets>
- Nabavi-Pelesaraei, A., Rafiee, S., Hosseinzadeh-Bandbafha, H., & Shamshirband, S. (2016a). Modeling energy consumption and greenhouse gas emissions for kiwifruit production using artificial neural networks. *Journal of Cleaner Production*, 133, 924–931. <https://doi.org/10.1016/j.jclepro.2016.05.188>
- Nabavi-Pelesaraei, A., Rafiee, S., Mohtasebi, S. S., Hosseinzadeh-Bandbafha, H., & Chau, K.-W. (2019). Comprehensive model of energy, environmental impacts and economic in rice milling factories by coupling adaptive neuro-fuzzy inference system and life cycle assessment. *Journal of Cleaner Production*, 217, 742–756. <https://doi.org/10.1016/j.jclepro.2019.01.228>

GLOSSÁRIO

O presente tópico objetiva explicitar de forma um pouco mais detalhada os principais termos técnicos usados no trabalho.

1. Termos relacionados ao aprendizado de máquina (AM):

- I. **Amostra:** no contexto do aprendizado de máquina, a amostra (*sample*) pode ser definida como um exemplo/conjunto de dados a ser usado para treinamento, validação e testagem do modelo. Uma amostra contém atributos (*features*) e atributo(s)-alvo (*target attributes*) em modelos de regressão.
- II. **Atributo:** no contexto do aprendizado de máquina, o atributo (*feature*) ou variável independente pode ser definido como uma característica presente em uma dada amostra. Também conhecido como variáveis de entrada, os atributos podem ser numéricos (Ex: peso), categóricos (Ex: cores) ou até mesmo binários (Ex: sim ou não). Os atributos são importantes pois será através deles que os modelos de aprendizado de máquina tentarão estabelecer padrões e relações entre as variáveis.
- III. **Atributo-alvo:** no contexto do aprendizado de máquina, o atributo-alvo (*target attributes*) ou variável-dependente pode ser definido como a característica a ser prevista pelos modelos com base nos dados de entrada.
- IV. **Hiperparâmetros:** no contexto do aprendizado de máquina, os hiperparâmetros podem ser definidos como os valores informados pelos usuários antes da etapa de treinamento dos modelos e que irão definir como os modelos aprendem. O ajuste dos hiperparâmetros afeta o aprendizado dos modelos e normalmente são ajustados via experimentação (Ex: *Grid Search*, *Random Search*, *Optuna*, etc...).
- V. **Modelos de aprendizado de máquina:** no contexto do aprendizado de máquina, os modelos podem ser definidos como algoritmos que são utilizados para estabelecer relação entre as variáveis de entrada (atributos) e as variáveis de saída (atributos-alvo). Podem ser do tipo supervisionado, não supervisionado ou por reforço. Após o treinamento, os modelos são capazes de gerar previsões a partir de novos dados.

2. Termos relacionados a avaliação do ciclo de vida (ACV):

- VI. **Dataset:** no contexto da ACV, o *dataset* ou conjunto de dados representa um conjunto estruturado de informações com entradas (inputs) e saídas (outputs) que descrevem fluxos ambientais, processos produtivos ou impactos associados a um sistema. Os *datasets* podem ser utilizados para descrever produtos, processos e/ou serviços. No contexto do presente trabalho, os conjuntos de dados com entradas e saídas representam as amostras a serem utilizadas para treinar os modelos de AM.
- VII. **Inputs:** no contexto da ACV, os *inputs* ou entradas representam os recursos (materiais, energia, consumíveis) utilizados por um dado processo (bem, processo ou serviço) em seu ciclo de vida. Os inputs são necessários para calcular os impactos ambientais potenciais associados ao sistema estudado. No contexto do presente trabalho, as entradas funcionam como os atributos a serem utilizados para treinar os modelos.
- VIII. **Outputs:** no contexto da ACV, os *outputs* ou saídas referem-se aos produtos/subprodutos gerados por um processo, bem como suas respectivas emissões para o ar, água ou solo, resíduos gerados ou produtos finais.
- IX. **Categoria de impacto:** no contexto da ACV, as categorias de impacto representam grupos de efeitos ambientais potenciais relacionados a áreas específicas de preocupação, que são usados para avaliar o impacto de um sistema ou processo. As categorias de impacto variam de acordo com o método de avaliação do impacto do ciclo de vida (AICV) escolhido. No contexto do presente trabalho, as categorias de impacto são utilizadas como atributos-alvo, ou seja, são os fatores a serem preditos após o treinamento dos modelos.