FRAMEWORK PARA PADRONIZAÇÃO DE DADOS EM BIORREFINO DE BIOMASSA APLICADO A MODELAGEM COM MACHINE LEARNING

JOÃO PEDRO LOPES[1]; TIAGO VOLKMER[2]; MATEUS FERRER[3]; ANDRÉ LUIZ MISSIO[4]

[1]*UFPel – lopes.a.joaopedro@gmail.com*
[2]*UFPel – tiago.volkmer@ufpel.edu.br*
[3]*UFPel – mmferrer@ufpel.edu.br*
[4]*UFPel – andre.missio@ufpel.edu.br*

1. INTRODUÇÃO

O avanço de biorrefinarias capazes de converter biomassa em biocombustíveis, produtos químicos e materiais de valor agregado é um eixo estratégico na transição para uma bioeconomia sustentável (RESHMY, 2022). Entretanto, os processos de conversão de biomassa permanecem marcados por baixa eficiência no uso de recursos e elevada complexidade, o que dificulta sua competitividade frente ao mercado consolidado de derivados do petróleo.

A otimização desses processos envolve múltiplas variáveis e relações não-lineares, um desafio no qual a aplicação de *Machine Learning* (ML) tem se destacado como ferramenta promissora. Modelos como Redes Neurais Artificiais (ANN), *Random Forest* (RF) e *Support Vector Machines* (SVM) já demonstraram alta acurácia na predição de rendimentos e propriedades de produtos (AKINPELU, 2023; BRUNTON, 2022).

Entretanto, a maioria dos estudos de ML na área se baseiam em conjuntos de dados fragmentados, não padronizados e, muitas vezes, extraídos manualmente de gráficos e tabelas de publicações anteriores. Conforme apontado por Zhang et al. (2023), não existem bancos de dados estruturados e reutilizáveis para muitos processos de conversão de biomassa. Essa prática limita drasticamente a robustez, a reprodutibilidade e a escalabilidade dos modelos desenvolvidos. Essa dependência de dados secundários e não estruturados gera consequências diretas: (i) baixo poder de generalização; (ii) alto risco de *overfitting*; (iii) dificuldades de reprodutibilidade e comparabilidade entre algoritmos; e (iv) barreira elevada à adoção em escala industrial. Somam-se a esses desafios as variações intrínsecas das biomassas: mesmo amostras provenientes de uma mesma origem podem apresentar propriedades físico-químicas heterogêneas, o que amplia as dificuldades de normalização e a consistência das análises de ML.

A prática atualmente adotada para a geração de dados estabelece um ciclo vicioso no qual estudos isolados resultam em modelos desconectados entre si que, embora possuam relevância acadêmica, apresentam contribuição limitada para a aplicação industrial. Este estudo adota a seguinte pergunta central: é possível estruturar um framework de dados padronizados que permita a aplicação consistente e escalável de modelos de ML em biorrefinarias?

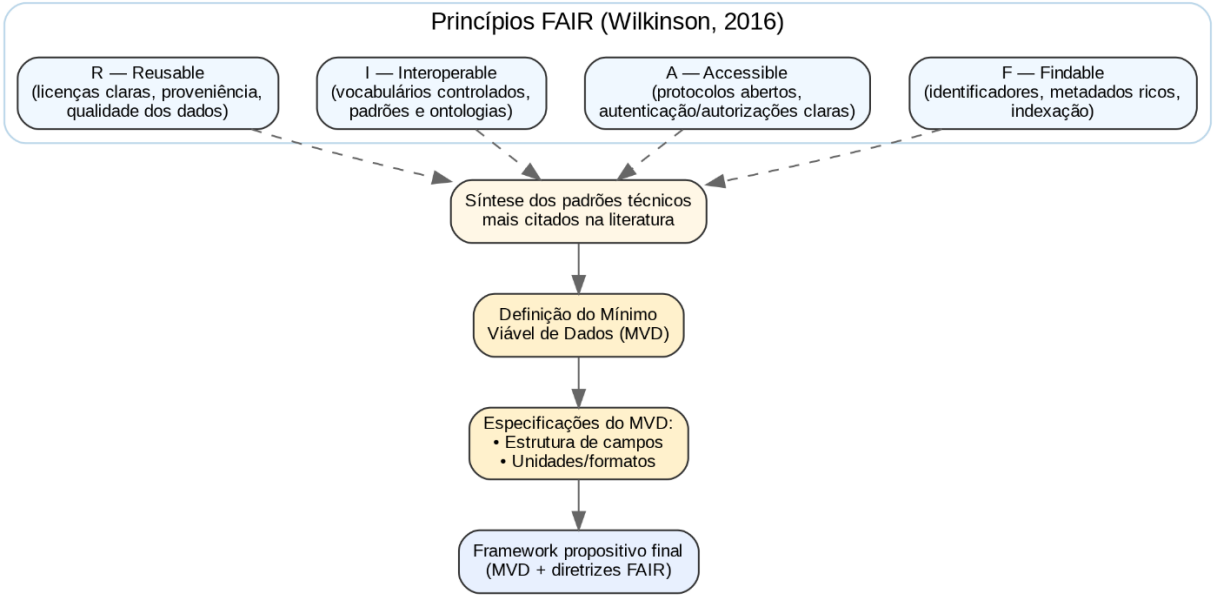
O presente trabalho objetiva dimensionar o impacto da fragmentação de dados e propor um framework inicial destinado a orientar a padronização e o compartilhamento de dados em pesquisas de biorrefino de biomassa.

2. METODOLOGIA

Este trabalho adota uma abordagem qualitativa e exploratória, estruturada como uma revisão narrativa da literatura científica recente sobre a aplicação de técnicas de ML em processos de biorrefino de biomassa. A escolha pela revisão narrativa, em vez de sistemática, justifica-se pelo objetivo de identificar tendências, limitações e lacunas conceituais, mais do que quantificar evidências. A metodologia foi organizada em quatro etapas principais:

- Revisão de literatura seletiva: foram consultadas as bases de dados *Web of Science e Scopus* para a seleção de artigos no período de 2015 a 2024. Como critério de seleção, priorizaram-se artigos publicados em periódicos de elevado impacto (*Impact Factor 5 years* > 15) e revisões relevantes citadas nesses estudos;
- Identificação de lacunas: a partir da análise dos estudos, foram sistematizadas as principais limitações relacionadas à fragmentação, padronização e reprodutibilidade dos dados;
- Construção de um framework propositivo: com base nos princípios FAIR (*Findable, Accessible, Interoperable e Reusable*) (WILKINSON, 2016) e nos padrões técnicos mais frequentemente citados na literatura foi estruturado um conjunto mínimo de dados (MVD) a ser adotado por pesquisas futuras com foco em interoperabilidade e ciência aberta, conforme a Figura 1.

Figura 1 – Fluxograma do Framework proposto



Fonte: autor

3. RESULTADOS E DISCUSSÃO

A literatura científica apresenta um número limitado de estudos que tratam de forma sistemática o armazenamento e a estruturação de dados em biorrefinarias. Trabalhos recentes, como os de He et al. (2023) e Osman et al. (2024), mencionam a necessidade de extração de dados da literatura por meio de *data mining*, mas não avançam na definição de padrões mínimos que assegurem reprodutibilidade e interoperabilidade.

A adoção dos princípios FAIR visa mitigar a fragmentação informacional, permitindo a integração entre diferentes conjuntos de dados e aumentando a confiabilidade dos modelos desenvolvidos. Resumidamente: (i) *Findable* requer identificadores persistentes, como DOIs, e depósito em repositórios públicos; (ii) *Accessible* demanda protocolos abertos para acesso por humanos e máquinas; (iii) *Interoperable* exige vocabulários controlados e formatos padronizados; e (iv) *Reusable* depende de metadados completos sobre origem, metodologia e licenciamento.

A análise da literatura indica que a ausência de padrões mínimos de dados compromete a escalabilidade dos modelos de ML e limita sua aplicação industrial. Para enfrentar essa limitação, propõe-se um MVD que pode servir como ponto de partida para pesquisadores e desenvolvedores de modelos. Esse MVD busca garantir comparabilidade entre estudos, reduzir redundâncias e viabilizar reuso em diferentes contextos de biorrefino.

A Tabela 1 apresenta a estrutura sugerida, organizada em três categorias: caracterização da matéria-prima, parâmetros do processo e caracterização do produto. As variáveis selecionadas correspondem às mais recorrentes nos estudos revisados, acompanhadas das unidades recomendadas e, quando disponíveis, dos padrões técnicos de referência (ASTM, NREL, TAPPI).

Tabela 1 - Proposta de um conjunto de dados mínimo para estudos de ML em biorrefinarias

Categoria	Variável	Unidade Sugerida	Padrão/Método
1. Caracterização da Matéria-Prima	Teor de Umidade	% (massa)	ASTM E871
	Voláteis	% (massa, base seca)	ASTM E872
	Cinzas	% (massa, base seca)	ASTM E1755
	Carbono (C), Hidrogênio (H) e Nitrogênio (N)	% (massa, base seca e livre de cinzas)	ASTM D5373
	Celulose, Hemicelulose e Lignina	% (massa, base seca)	NREL/TP-510-42618
2. Parâmetros do Processo	Tipo de Processo	Palavras-Chave	Pirólise, Gaseificação, HTL, etc.
	Temperatura de Reação	°C	-
	Taxa de Aquecimento	°C/min	-
	Tempo de Residência	min ou s	-
	Tipo de Carga de Catalisador	Palavras-Chave	NaOH, Na_2S, etc.
	Carga de Catalisador	% (massa)	-
3. Caracterização do Produto	Rendimento do produto final	% (massa, da biomassa seca)	TAPPI T 203, etc.

Fonte: Autor

4. CONCLUSÕES

A principal barreira para a aplicação de ML no campo das biorrefinarias não reside em limitações algorítmicas, mas sim na escassez, heterogeneidade e ausência de padronização dos dados experimentais. A prática corrente, baseada na utilização de informações dispersas e frequentemente extraídas de forma manual da literatura, dificulta o desenvolvimento de modelos preditivos e escaláveis para o ambiente industrial. Este trabalho argumenta que a superação deste gargalo requer uma reestruturação das práticas de geração, documentação e compartilhamento de dados pela comunidade científica. Nesse contexto, apresenta-se um framework fundamentado nos princípios FAIR e em um MVD como estratégia para promover a padronização e a interoperabilidade dos dados. A adoção sistemática dessa estrutura pode viabilizar a formação de bases de dados consistentes e de alta qualidade, capazes de sustentar o desenvolvimento de modelos mais robustos e generalizáveis. Trata-se de uma proposta inicial, que pode e deve ser ampliada, refinada e validada pela comunidade científica, constituindo um passo concreto para aproximar o uso de ML em biorrefinarias da esfera acadêmica às demandas de aplicação e escalabilidade industrial.

5. REFERÊNCIAS BIBLIOGRÁFICAS

AKINPELU, D. A., OLUWASEUN A., et al. Machine learning applications in biomass pyrolysis: from biorefinery to end-of-life product management. **Digital Chemical Engineering 8**, 2023.

BRUNTON, et al. Data-driven science and engineering: Machine learning, dynamical systems, and control. **Cambridge University Press**, 2022.

HE, H. et al. Functional carbon from nature: biomass-derived carbon materials and the recent progress of their applications. ***Advanced science* 10**, no. 16, 2023

RESHMY, R., et al. Updates on high value products from cellulosic biorefinery. **Fuel 308**, 2022.

OSMAN, I., et al. Optimizing biodiesel production from waste with computational chemistry, machine learning and policy insights: a review. **Environmental Chemistry Letters 22**, no. 3: 1005-1071, 2024.

WILKINSON, M., DUMONTIER, M., AALBERSBERG, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. **Sci Data 3**, 160018, 2016. DOI: https://doi.org/10.1038/sdata.2016.18

ZHANG, W., CHEN, Q., CHEN, J., et al. *Machine learning for hydrothermal treatment of biomass: A review***. Bioresource Technology**, 370, 128547, 2023.