Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 11, n. 1, 2025.

Trabalho apresentado no XLIII CNMAC, Centro de Convenções do Armação Resort - Porto de Galinhas - PE, 2024

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Classificação do Tipo de Pirólise da Biomassa Através do Algoritmo Floresta Aleatória

Sabrinna R. de O. de Souza[1], Vinicius L. Xavier[2]
IME/UERJ, Rio de Janeiro, RJ
Aderval S. Luna[3]
QUI/UERJ, Rio de Janeiro, RJ
Raquel E. Guedes[4], Alexandre R. Torres[5]
FAT/UERJ, Rio de Janeiro, RJ
Marcello M. Provenza[6]
IME/UERJ, Rio de Janeiro, RJ

Resumo. Este estudo aborda um problema de classificação do tipo de pirólise de biomassa. Tem como objetivo identificar as variáveis mais relevantes para a classificação do tipo de pirólise de biomassa, especificamente, para o reator do tipo Contínuo. Para isto, o algoritmo de Floresta Aleatória foi aplicado, obtendo uma exatidão em torno de 98%, indicando a capacidade em realizar previsões exatas. As variáveis de mais relevância foram obtidas utilizando os métodos de permutação e índice de Gini. As variáveis mais relevantes, dentre as doze analisadas, foram: *Porcentagem de cinzas em base seca da matéria-prima, Porcentagem de carbono em base seca livre de cinza na matéria-prima e Porcentagem de oxigênio em base seca livre de cinza na matéria-prima.*

Palavras-chave. Biomassa, Programa R, Floresta Aleatória, Aprendizado de Máquina

1 Introdução

Neste trabalho, analisa-se dados relacionados à biomassa, o qual engloba todo material orgânico proveniente de plantas. A biomassa é uma fonte alternativa de energia, destacando-se o bio-óleo obtido pela pirólise, útil como combustível e na produção de produtos químicos e biomateriais [5]. A pirólise é a decomposição térmica de material orgânico em condições de baixo oxigênio ou em ambientes onde a gaseificação é limitada pela concentração de oxigênio. Normalmente realizada entre 400°C e o início da gaseificação [13].

Na pirólise *Slow*, há uma temperatura mais baixa, aquecimento menor e tempo de residência mais longo, resultando na produção de carvão. Já na pirólise *Flash*, o tempo de reação é extremamente curto, apenas alguns segundos, com um aquecimento muito alto. A pirólise *Fast*, favorece a formação de bio-óleo, com temperatura moderada, curto tempo de residência do vapor e aquecimento rápido, embora não tão rápido quanto na pirólise *Flash*. Por último, a pirólise *Catalytic* é utilizada para aprimorar a qualidade do óleo produzido [5]. Esses tipos de pirólise são classificados nas respectivas classes (*Fast, Slow, Flash* e *Catalytic*) de acordo com doze variáveis químicas, representadas na Tabela 1.

[1]sabrinnardo123@gmail.com
[2]viniciuslx@ime.uerj.br
[3]adsluna@gmail.com
[4]raquelescrivane@yahoo.com.br
[5]artorres@fat.uerj.br
[6]mprovenza@gmail.com

O banco de dados compila vários processos de pirólise e suas condições operacionais, visando entender como esses fatores afetam a composição e o rendimento dos produtos resultantes da pirólise. Baseia-se em dados experimentais de mais de 200 estudos desde 1984, incluindo artigos de referência e pesquisas recentes sobre pirólise de biomassa [5]. Os tipos de pirólise são classificados com base em variáveis químicas, conforme detalhado na Tabela 1:

Tabela 1: Variáveis que compõem os tipos de pirólise.

Variável	Descrição	Número
p_c_mp	Porc. carbono em base seca livre de cinza na matéria-prima	1
p_h_mp	Porc hidrogênio em base seca livre de cinza na matéria-prima	2
p_n_mp	Porc. nitrogênio em base seca livre de cinza na matéria-prima	3
p_o_mp	Porc. oxigênio em base seca livre de cinza na matéria-prima	4
p_umid_mp	Porc. umidade na matéria-prima	5
p_cfix_mp	Porc. carbono fixo em base seca da matéria-prima	6
p_cinz_mp	Porc. cinzas em base seca da matéria-prima	7
p_vola_mp	Porc. voláteis em base seca da matéria-prima	8
tp_med_reator	Tamanho da partícula média no reator	9
t_res_vap_reator	Tempo de residência médio no reator para o gás e arraste	10
temp_reator	Tempreatura de operação do reator	11
rend_gp	Rendimendo em gás - porcentagem	12
tipo_piro_reator	Tipo de pirólise (classes)	13

Este estudo considera casos de regime de operação do reator igual a Contínuo e o tipo do reator igual a **Leito Fixo** que contemplam os tipos de pirólise: *Fast*, *Flash* e *Catalytic*. A amostra consiste em 171 observações, sendo 135 para tipo de pirólise *Fast*, 35 para tipo de pirólise *Flash* e 1 para tipo de pirólise *Catalytic*. Após experimentos, a classe *Catalytic* foi excluída devido à sua única amostra. Além disso, por se tratar de uma amostra do tipo de reator Contínuo, não cabe a utilização da taxa de aquecimento e tempo de reação, pois a biomassa entra no processo de forma continua, então ela já entra na temperatura de pirólise. Em outras palavras, no reator Contínuo, a biomassa é alimentada continuamente no processo já na temperatura de pirólise. Então, não existe a taxa de aquecimento e tempo de reação.

Alguns estudos na literatura abordam análises de dados de biomassa, incluindo aplicações de redes neurais artificiais. Por exemplo, o autor Moscato (2019) investigou análises exergéticas de uma caldeira de biomassa usando redes neurais artificiais [9]. Outro estudo é o do autor Merdun (2018), com aplicações de dois métodos de redes neurais artificiais na modelagem dos rendimentos de produtos de pirólise, utilizando nove tipos de biomassa e dois parâmetros de processo como variáveis de entrada para os modelos [8]. Outro estudo é dos autores Oliveira de Souza et al. (2023), onde são realizadas aplicações de diversos tipos de métodos de agrupamentos em dados de biomassa para o reator do tipo Batelada, com objetivo de avaliar se as classes dos tipos de pirólise são suficientes para caracterizar esse processo químico [11].

No entanto, há uma escassez de estudos que relacionam as variáveis mais relevantes na classificação do tipo de pirólise para o reator do tipo Contínuo. Este estudo tem como diferencial a identificação das variáveis mais relevantes para a classificação dos tipos de pirólise dos dados de biomassa para o reator do tipo Contínuo. Essa identificação é feita com o algoritmo Floresta Aleatória, considerando a Exatidão e o índice de Gini, conforme desenvolvido no estudo de Oliveira de Souza (2023) [10]. As técnicas de aprendizado de máquina, mesmo quando os critérios para classificação são conhecidos, podem enriquecer o processo ao oferecer automação, eficiência, identificação de padrões complexos, aprimoramento contínuo, detecção de anomalias, entre outros fatores. Por isso, é pertinente a aplicação dessas técnicas, como forma de contribuir para a ciência,

DOI: 10.5540/03.2025.011.01.0362

o que abre caminho para alguns ganhos através dessa automação.

1.1 Floresta Aleatória (Random Forest)

Neste trabalho, é aplicado o método supervisionado de Floresta Aleatória. O aprendizado supervisionado utiliza exemplos previamente rotulados. Nesse tipo de aprendizado, estima-se a taxa de acerto e taxa de erro obtidas por um classificador e os dados são separados em dois subconjuntos: treinamento e teste. O subconjunto de treinamento é utilizado no aprendizado do classificador e o de teste mede a capacidade de generalização na predição [7].

A Floresta Aleatória é um modelo baseado em árvores de decisão, que são métodos utilizados como uma forma de dividir os dados. Nos modelos baseados em árvores, as amostras são divididas de acordo com os critérios estabelecidos para cada variável: *"sim"* leva a amostras para um lado, por exemplo, esquerda, enquanto *"não"* as direciona para o lado oposto. O processo de seleção e divisão das amostras continua em cada novo nó subsequente até que uma regra de decisão seja alcançada [4].

O algoritmo gera uma coleção de centenas a milhares de árvores, cada uma construída com uma amostra dos dados originais usando o método de *Bootstraping* [3]. Os dados excluídos da amostra *Bootstraping* são denominados *out-of-bag (OOB)* e servem para estimar o desempenho do modelo. Utilizando esses dados, é viável calcular a taxa de erro da previsão do modelo [2]. Neste estudo o método é utilizado como classificador e, também, para selecionar as variáveis de maior relevância.

1.1.1 Indice Gini

É possível encontrar as variáveis de maior relevância por meio do índice *Gini* que avalia a impureza de um nó [6]. Segundo Barbosa, Carneiro e Tavares (2012), o índice *Gini* em um determinado nó é dado pela seguinte equação:

$$Gini = 1 - \sum_{i=1}^{c} p_i^2 \tag{1}$$

onde, p_i é a frequência relativa de cada classe em cada nó e c é o número de classes [1].

Quando este índice é igual a zero, o nó é considerado puro. Mas quando ele se aproxima de um, o nó é considerado impuro, pois aumenta o número de classes uniformemente distribuídas neste nó [1].

2 Análise Exploratória dos Dados

Em toda a análise de dados foi utilizado o programa R, livre e *open source* [12]. Antes de aplicar ferramentas e algoritmos, é crucial compreender a natureza dos dados, incluindo medidas de posição, dispersão e distribuições.

2.1 Análise do Subconjunto para Reator do Tipo Contínuo

Na Tabela 2 são apresentadas as seguintes estatísticas descritivas das variáveis: valor mínimo, primeiro quartil (Q1), mediana, média, terceiro quartil (Q3), valor máximo e variância. Na Tabela 2, é perceptível que as variáveis 3, 5, 6, 7, 9 e 10 possuem suas observações bem próximas a zero no valor mínimo. Já as variáveis 10 e 11 possuem os maiores valores máximos. A média está abaixo da mediana no caso das variáveis 4, 5 e 11, indicando uma assimetria negativa [10].

DOI: 10.5540/03.2025.011.01.0362

Tabela 2: Estatística descritiva para reator do tipo Contínuo.

Variável	Mínimo	Q1	Mediana	Média	Q3	Máximo	Variância
1	42,49	45,10	47,36	48,44	51,40	55,80	10,744
2	5,10	5,79	6,14	6,27	6,85	8,10	0,52
3	0,10	0,39	0,74	1,00	1,30	4,54	0,80
4	34,70	40,46	44,41	44,13	48,80	50,63	17,57
5	0,68	6,00	7,90	7,15	8,64	12,30	7,22
6	0,11	10,78	14,55	14,55	18,77	30,55	51,76
7	0,55	2,64	3,60	4,36	6,60	13,98	6,68
8	64,07	77,70	81,50	80,80	83,86	98,06	49,89
9	0,12	0,43	0,53	1,04	1,50	10,00	1,40
10	0,50	1,27	4,10	13,19	7,54	101,40	590,99
11	300,00	450,00	495,00	483,30	546,00	700,00	5222,07
12	5,00	14,50	19,00	20,84	24,90	60,00	75,829

Com a compreensão de como estão distribuídos os valores das variáveis, pode-se averiguar as relações entre as variáveis, assim como a interdependência entre elas. Na Figura 1 podemos visualizar pelo mapa de calor das correlações que as variáveis que estão dentro da escala de 1 a 0,7 (+ ou -), na tonalidade azul escuro (no caso das positivas) e na tonalidade vermelho escuro (no caso das negativas), possuem uma forte correlação. Já as variáveis que estão entre 0,7 a 0,5 (+ ou -) na tonalidade azul claro (no caso das positivas) e na tonalidade laranja (no caso das negativas), possuem correlação moderada. As variáveis que possuem escala de 0,5 a 0,25 (+ ou -) possuem baixa correlação.

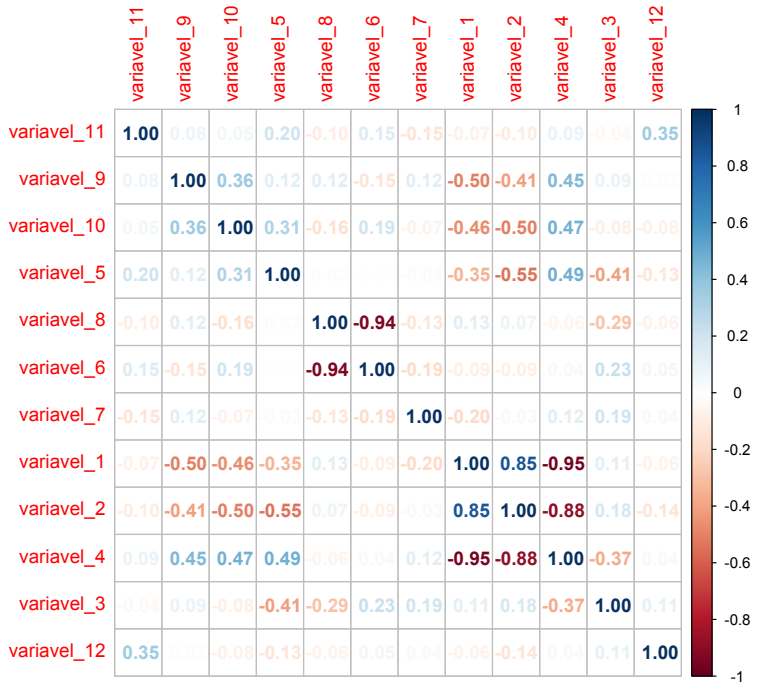

Figura 1: Correlação das variáveis considerando reator do tipo Contínuo. Fonte: dos autores.

DOI: 10.5540/03.2025.011.01.0362

3 Resultados e Análises

3.1 Resultados do Algoritmo Floresta Aleatória

Por meio do algoritmo Floresta Aleatória, foi feito um ajuste do número de variáveis e número de árvores e, em seguida, foi realizada a validação cruzada 10-*fold*. Para o número ideal de árvores, obteve-se um pico registrado em 1.500 árvores, com uma variável, ou seja, o número de variáveis que são consideradas para cada divisão das árvores de decisão.

Analisando também os valores de exatidão e *kappa*, obteve-se em torno de 98% para exatidão e em torno de 98% para o indicador *kappa*. Também foram gerados os valores da exatidão balanceada, 98% para *Fast* e 99,8% para *Flash*. Foram obtidos também os valores de *F1-score*, que é uma média ponderada entre a precisão e o *recall*, sendo 98% para *Fast* e 99,9% para *Flash*.

Por meio do erro *Out of bag (OOB)* obtido no gráfico da Figura 2, é possível observar o erro geral na linha preta iniciando bem próximo de 0,05, em seguida decresce e se mantém com uma sazonalidade bem imperceptível a partir de 100 árvores no eixo x. Sendo assim, foi obtido um erro *Out of bag* geral de 0,01%, com 1.500 árvores e uma variável.

A partir disso, com o número de variáveis e o número de árvores definidos, também foi possível identificar as variáveis de mais relevância por meio do algoritmo Floresta Aleatória. As variáveis de mais relevância foram obtidas por meio do método de permutação que avalia a diminuição média na exatidão e pelo índice de Gini que avalia a diminuição média na impureza do nó. Conforme o gráfico na Figura 3 e por meio da Tabela 3, é possível perceber essas variáveis, as quais foram: *Porcentagem de cinzas em base seca da matéria-prima, Porcentagem de carbono em base seca livre de cinza na matéria-prima e Porcentagem de oxigênio em base seca livre de cinza na matéria-prima.*

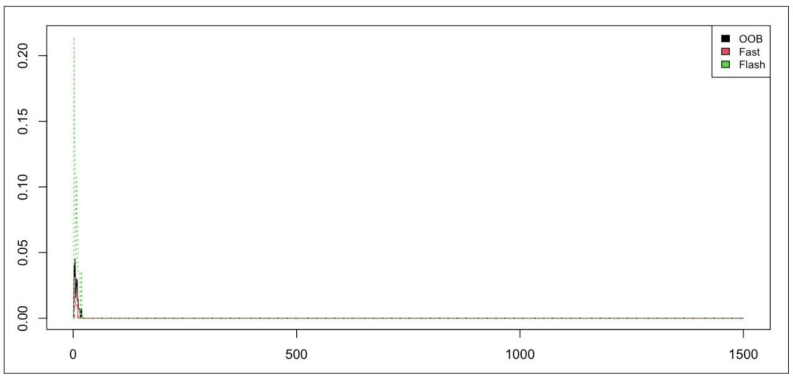

Figura 2: Representação gráfica da taxa de erro *Out of bag* em relação ao número de árvores para reator do tipo Contínuo. Fonte: dos autores.

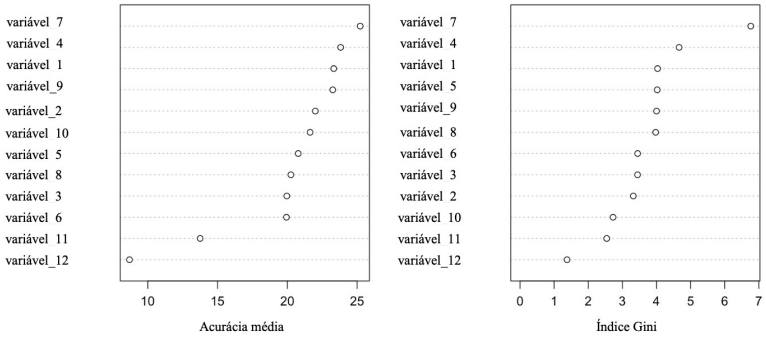

Figura 3: Representação gráfica da relevância de cada variável na classificação considerando a Exatidão e o ìndice de Gini para reator do tipo Contínuo. Fonte: dos autores.

Tabela 3: Variáveis de mais relevância para reator do tipo Contínuo.

Variável	Exatidão	Índice de Gini
7	25,22	6,76
4	23,81	4,66
1	23,32	4,03

Portanto, os resultados obtidos neste estudo, reforçam a eficácia do modelo de Floresta Aleatória na análise dos dados de biomassa para o reator do tipo Contínuo o qual foi analisado.

4 Considerações Finais

Neste artigo foi tratado o problema de análise de dados dos tipos de pirólise de biomassa, especificamente, para reator do tipo Contínuo. As amostras são agrupadas nas classes *Fast*, *Flash* e *Catalytic*, de acordo com as doze variáveis elementares, complementares e de processo.

Inicialmente, foi realizada uma análise exploratória dos dados para compreender o perfil e o comportamento geral das variáveis. Em seguida, o foco foi direcionado para o desafio de classificação do tipo de pirólise de biomassa, com o objetivo de identificar as variáveis de mais relevância para a classificação do tipo de pirólise de biomassa. Para isto, foi aplicado o algoritmo de Floresta Aleatória, uma técnica de aprendizado de máquina supervisionado.

Por fim, obteve-se uma taxa de exatidão em torno de 98% quando aplicado o algoritmo de Floresta Aleatória. Além disso, foram gerados os valores da exatidão balanceada, 98% para *Fast* e 99,8% para *Flash* e também foram identificadas as variáveis mais relevantes para a classificação do tipo de pirólise, são elas: *Porcentagem de cinzas em base seca da matéria-prima, Porcentagem de carbono em base seca livre de cinza na matéria-prima e Porcentagem de oxigênio em base seca livre de cinza na matéria-prima.* Como trabalhos futuros, serão aplicados outros métodos de aprendizado de máquina com o intuito de comparar com os resultados obtidos neste trabalho.

Agradecimentos

Os autores agradecem à Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), ao CNPq e à universidade UERJ: E-10/2020 - Edital Inteligência Artificial,

processo E-26/290.023/2021; E-26/2021 – Auxílio Básico à Pesquisa (APQ1) EM ICTs ESTADUAIS UERJ, UENF e UEZO, processo E-26/211.958/2021; Bolsa CNE, processo 202.552/2019; Bolsa de Produtividade CNPq 1C, processo 304.581/2022-4; Bolsas Pró-Ciência (UERJ).

Referências

[1] M. J. Barbosa, T. G. S. Carneiro e A. I. Tavares. "Métodos de Classificação por Árvores de Decisão Disciplina de Projeto e Análise de Algoritmos". Em: **UFOP–Universidade Federal de Ouro Preto Ouro Preto, Minas Gerais–MG** (2012).

[2] L. Breiman. "Random forests". Em: **Machine learning** 45.1 (2001), pp. 5–32.

[3] X. Chen e H. Ishwaran. "Random forests for genomic data analysis". Em: **Genomics** 99.6 (2012), pp. 323–329. DOI: 10.1016/j.ygeno.2012.04.003.

[4] F. B. DE SANTANA. "Floresta aleatória para desenvolvimento de modelos multivariados de classificação e regressão em química analítica". Tese de doutorado. IQ/UNICAMP, 2020.

[5] R. E. Guedes, A. S Luna e A. R. Torres. "Operating parameters for bio-oil production in biomass pyrolysis: A review". Em: **Journal of analytical and applied pyrolysis** 129 (2018), pp. 134–149. DOI: 10.1016/S0165-2370(96)00956-4.

[6] G. James, D. Witten, T. Hastie e R. Tibshirani. **An introduction to statistical learning**. Vol. 112. Springer, 2013. ISBN: 978-1-4614-7137-0.

[7] A. C. Lorena e A. C. P. L. F. de Carvalho. **Introduçao aos Classificadores de Margens Largas**. 2003.

[8] H. Merdun. "Modeling of pyrolysis product yields by artificial neural networks". Em: **International Journal of Renewable Energy Research (IJRER)** 8.2 (2018), pp. 1178–1188. ISSN: 1309-0127.

[9] A. L. S. Moscato. "Análise exergética de uma caldeira de biomassa utilizando redes neurais artificiais". Tese de doutorado. FEB/UNESP, 2019.

[10] S. R. de Oliveira de Souza. "Métodos de inteligência artificial aplicados em dados de biomassa para a caracterização dos diferentes tipos de pirólise". Dissertação de mestrado. IME/UERJ, 2023.

[11] S. R. de Oliveira de Souza, V. L. Xavier, R. R. Guedes, A. R. Torres, A. S. Luna e M. M. Provenza. "Avaliação de métodos de agrupamentos em dados de biomassa considerando os diferentes tipos de pirólise". Em: **Revista Internacional de Ciências** 13.2 (2023), pp. 124–139. DOI: 10.12957/ric.2023.78860.

[12] R Core Team. **R Core Team R. R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.** 2016.

[13] G. E. G. Vieira, A. P. Nunes, L. F. Teixeira e A. G. N. Colen. "Biomassa: uma visão dos processos de pirólise". Em: **Revista Liberato** 15.24 (2014), pp. 167–178. ISSN: 2178-8820.

