Machine Learning-Based Prediction of Bio-Oil Yields from Pyrolysis of Biomass: A Comparative Study of Imputation Algorithms and Model Benchmarking

Antonio Elia Pascarella[1,*,†]

[1]Department of Electrical Engineering and Information Technology (DIETI), University of Naples Federico II, Via Claudio 21, 80125 Naples, Italy

Abstract

The exhaustion of non-renewable fossil fuels has heightened awareness about environmental issues. As a result, biomass energy has come into the spotlight as a promising renewable alternative, particularly in the context of bio-oil production through pyrolysis from waste biomasses. Unfortunately, physics-aware models pose difficulties when modelling bio-oil production, prompting researchers to lean towards data-centric approaches. To cope with this problem, this paper showcases a comprehensive dataset of nearly one thousand records sourced from prior literature about bio-oil production. Besides collecting, cleaning and organising the gathered data, we also used machine learning techniques to evaluate the resulting dataset, with the most promising result yielding a mean absolute error of 2.6 and an adjusted R-squared of 0.9 in prediction of bio-oil yield. To the best of our knowledge, this paper delivers introduces the most comprehensive dataset ever collated in this domain. The assembly of such an exhaustive dataset is pivotal for sustainable process engineering because it fosters precise modelling, thus better-addressing uncertainties inherent in the process.

Keywords

Machine Learning, Missing data imputation, Renewable energy.

1. Introduction

The accelerated depletion of non-renewable fossil fuel reserves has led to critical issues related to environmental degradation. Consequently, biomass energy, which is abundant and sustainable, has attracted considerable attention [1]. Among the various conversion techniques for biomass, the production of bio-oil through pyrolysis, and particularly its modeling, is the focus of this work. If appropriately upgraded, bio-oil could serve as an alternative to fossil fuels.

The uncertainty in the yield of bio-oil production and its qualities and the optimization of process variables can be managed with mathematical models. Simulating pyrolysis for bio-oil production using physics-based models is challenging. As artificial intelligence, and in particular, the subfield of machine learning evolves, many data-centric approaches have been proposed

AIxIA'23: 22nd International Conference of the Italian Association for Artificial Intelligence, November 06–09, 2023, Rome, Italy

[*] Corresponding author.

[†] This research is part of a doctoral project supervised by Professor Carlo Sansone (University of Naples Federico II) and Professor Piero Salatino (University of Naples Federico II).

antonioelia.pascarella@unina.it (A. E. Pascarella)
0000-0002-1079-7741 (A. E. Pascarella)

for estimating bio-oil yields from waste biomass properties and plant operational conditions, acknowledging the non-linear correlations [2], overcoming in this way the difficulties of physical-based modelling. As a significant contribution, the research has embarked upon extensive data collection from existing literature to further machine-learning efforts for predicting bio-oil yields. This dataset exhibits the challenge of missing values, and an analysis for predicting bio-oil yields comparing different imputation methods to handle missing data is provided.

Section 2, ”Materials and Methods”, introduces the collected dataset related to waste biomasses and bio-oils needed to enforce data-centric research and the procedures used for handling missing data; Section 3, ”Results”, presents the comparison between different machine learning algorithms for predicting the yield of bio-oil; Section 4, ”Discussion with Conclusions” presents the conclusions elucidating the novelty of this work about the existing literature.

2. Material and methods

The following section introduces the dataset, which has been gathered from the literature to contribute to AI and renewable energy research in the context of waste biomasses. The missing data in the dataset has been addressed before establishing a machine-learning benchmark for predicting bio-oil yield, as shown in section 3. Section 2.1 presents the details of the dataset, while Section 2.2 describes the frameworks used to handle the missing data.

2.1. Data collection

This study’s collected dataset, consisting of 1057 entries, includes proximate analysis of biomass like ash fixed-carbon and volatiles, ultimate analysis of biomass like carbon hydrogen oxygen and nitrogen, lignocellulosic content, and plant operative conditions as temperature, heating rate, particle size, and nitrogen flow rate. The target variable in this research work is the bio-oil yield.

The extent of missing data varies across different variables. For the lignocellulosic content, missing data account for between 40 and 50 percent of total entries, whereas for ultimate and proximate analyses, the missing data are close to 5 percent and around 10 percent, respectively. This prevalence of missing data necessitates the implementation of data imputation techniques before applying models to estimate bio-oil yield.

2.2. Missing data imputation

In this work, the issue of missing data was addressed by leveraging four distinct imputation techniques, utilizing an iterative method in a round-robin fashion as explained in the scikit-learn iterative-imputer doc[1], a framework inspired on [3]. This procedure involves a systematic approach to filling in the missing values for each variable iteratively, utilizing the remaining variables as predictors. Each missing value was initially provisionally filled with rudimentary estimates such as the variable’s mean, median, or mode. After that, one variable containing missing values was chosen (Variable A for this instance), and its preliminarily filled values

[1] https://scikit-learn.org/stable/modules/impute.html#iterative-imputer

were treated as missing again. The remaining variables, now including the initially presenting missing values but substituted with estimates in the first step, were employed to predict the missing values for Variable A. This phase uses various imputation models: Random Forest as [4], k-nearest Neighbors (kNN) as [5], and Support Vector Regressor (SVR).

Lastly, a fourth approach, the Variational Autoencoder (VAE), was used for data imputation purposes, as seen in [6] and [7]. As a type of artificial neural network, the VAE generates complex data distribution models by utilizing principles of variational inference to address the problem of missing values.

3. Results

Four versions of the complete dataset were created without missing values from the original dataset using the four imputation algorithms above. Machine learning algorithms were then employed to give a benchmark on bio-oil yields on this newly collected dataset. In Tables 1, 2, 3, and 4, the comparison of machine learning algorithms to predict the bio-oil yield is shown on the imputed versions of the datasets using KNN, random forest, support vector regressor, and variational autoencoder, respectively.

Models	Mean Absolute Percentage Error	Mean Signed Difference	Root Mean Squared Error	Mean Squared Error	Mean Absolute Error	Adjusted R^2	R^2
Tree	0.1	0.2	6.1	37.5	3.2	0.6	0.6
Linear Regression	0.2	0.8	9.0	80.9	7.1	0.2	0.2
GBTree	**0.1**	**0.1**	**4.8**	**23.2**	**3.0**	**0.8**	**0.8**
RF	0.1	0.1	4.9	23.5	3.2	0.8	0.8
KNN7	0.2	0.6	7.3	52.7	5.2	0.5	0.5
MLP	0.2	0.9	8.4	71.1	6.4	0.3	0.3
LibSVM	0.2	0.8	9.7	94.4	7.8	0.1	0.1
RBFRegressor	0.2	0.5	8.6	73.4	6.7	0.3	0.3
KNN3	0.1	0.1	6.8	46.8	4.6	0.5	0.5
KNN5	0.1	0.7	7.0	49.4	4.9	0.5	0.5
AdditiveRegression	0.2	0.5	8.1	65.6	6.3	0.4	0.4

Table 1
Benchmark on bio-oil yield on KNN imputed dataset.

In Tables 1, 2, 3, and 4, it is shown that the handling of missing values with the iterative imputation using Random Forest combined with random forest for regression on bio-oil yield produces the best results, resulting in the mean absolute error of 2.6 and adjusted R-squared of 0.9. The models were trained using a hold-out split of 70-15-15 for training, validation, and testing.

Models	Mean Absolute Percentage Error	Mean Signed Difference	Root Mean Squared Error	Mean Squared Error	Mean Absolute Error	Adjusted R^2	R^2
Tree	0.1	-0.2	5.0	24.7	2.7	0.8	0.8
Linear Regression	0.2	0.8	9.3	85.7	7.3	0.2	0.2
GBTree	0.1	**0.3**	4.0	16.0	2.7	0.8	0.8
RF	**0.1**	0.4	**3.8**	**14.7**	**2.6**	**0.9**	**0.9**
KNN7	0.2	0.8	7.1	50.6	5.2	0.5	0.5
MLP	0.2	1.8	10.1	102.4	7.0	0.0	0.0
LibSVM	0.2	0.7	9.7	94.9	7.8	0.1	0.1
RBFRegressor	0.2	0.8	9.1	82.7	7.4	0.2	0.2
KNN3	0.1	0.9	6.7	44.8	4.5	0.6	0.6
KNN5	0.1	1.0	6.8	45.8	4.9	0.6	0.6
AdditiveRegression	0.2	0.8	8.3	68.5	6.5	0.3	0.3

Table 2
Benchmark on bio-oil yield on Random Forest imputed dataset.

Models	Mean Absolute Percentage Error	Mean Signed Difference	Root Mean Squared Error	Mean Squared Error	Mean Absolute Error	Adjusted R^2	R^2
Tree	0.1	0.9	5.4	29.2	2.9	0.7	0.7
Linear Regression	0.2	0.8	9.1	82.3	7.3	0.2	0.2
GBTree	0.1	0.6	4.7	22.0	**3.0**	0.8	0.8
RF	**0.1**	**0.4**	**4.4**	**19.2**	3.0	**0.8**	**0.8**
KNN7	0.2	0.9	7.0	49.2	5.1	0.5	0.5
MLP	0.2	2.5	9.1	83.2	6.6	0.2	0.2
LibSVM	0.2	0.8	9.8	95.8	7.9	0.1	0.1
RBFRegressor	0.2	0.7	8.2	67.6	6.6	0.3	0.3
KNN3	0.1	1.0	6.5	41.9	4.4	0.6	0.6
KNN5	0.1	0.9	6.6	43.8	4.7	0.6	0.6
AdditiveRegression	0.2	0.8	8.1	65.1	6.4	0.4	0.4

Table 3
Benchmark on bio-oil yield on Support Vector Regressor imputed dataset.

4. Discussion with Conclusions

The produced bio-oil yield from biomass waste was successfully predicted with random forest regression combined with an iterative imputation method based on Random Forest to tackle the problem of missing data, achieving an R-squared value of approximately 0.9 and a mean absolute error of about 2.6.

To the best of our knowledge, this is the benchmark on bio-oil yield on the more extensive dataset collected in literature in the context of pyrolysis of biomass, which hence contains a broader range of biomass properties and plant operative variables, concerning other studies on different and smaller datasets like [8], [9] and [10]. It is worth emphasizing that in the context of renewable energy from biomass waste, creating an extensive dataset, as was done for

Models	Mean Absolute Percentage Error	Mean Signed Difference	Root Mean Squared Error	Mean Squared Error	Mean Absolute Error	Adjusted R^2	R^2
Tree	0.1	0.6	5.7	32.0	3.2	0.7	0.7
Linear Regression	0.2	0.7	9.3	85.7	7.4	0.2	0.2
GBTree	**0.1**	0.7	4.6	21.0	**3.0**	0.8	0.8
RF	0.1	**0.6**	**4.5**	**20.5**	3.1	**0.8**	**0.8**
KNN7	0.2	0.8	7.3	53.8	5.4	0.5	0.5
MLP	0.2	3.0	8.4	71.0	6.4	0.3	0.3
LibSVM	0.2	0.8	9.8	96.2	7.9	0.0	0.0
RBFRegressor	0.2	0.3	8.7	75.1	6.8	0.3	0.3
KNN3	0.1	1.0	7.2	52.0	5.0	0.5	0.5
KNN5	0.2	0.9	7.1	51.0	5.1	0.5	0.5
AdditiveRegression	0.2	0.7	8.7	75.3	6.9	0.3	0.3

Table 4
Benchmark on bio-oil yield on Variational Auto-Encoder imputed dataset.

the specific process of pyrolysis on which an algorithm can be trained to predict bio-oil yield with good performance, provides a valuable tool for plant modelling, aiding in simulating new scenarios for optimizing operative conditions and facilitating the development of systems to handle uncertainties with their predictive capabilities, overcoming in this way the difficulties of modelling with physics models and boosting the research to data-centric methodologies.

References

[1] S. Wang, G. Dai, H. Yang, Z. Luo, Lignocellulosic biomass pyrolysis mechanism: A state-of-the-art review, Progress in Energy and Combustion Science 62 (2017) 33–86.

[2] Y. Sun, L. Liu, Q. Wang, X. Yang, X. Tu, Pyrolysis products from industrial waste biomass based on a neural network model, Journal of Analytical and Applied Pyrolysis 120 (2016) 94–102.

[3] S. Van Buuren, K. Groothuis-Oudshoorn, mice: Multivariate imputation by chained equations in r, Journal of Statistical Software 45 (2011) 1–67.

[4] D. Stekhoven, P. Bühlmann, Missforest—non-parametric missing value imputation for mixed-type data, Bioinformatics 28 (2012) 112–118.

[5] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R. Altman, Missing value estimation methods for dna microarrays, Bioinformatics 17 (2001) 520–525.

[6] Y. Qiu, H. Zheng, O. Gevaert, Genomic data imputation with variational auto-encoders, GigaScience 9 (2020) giaa082.

[7] G. Boquet, J. Vicario, A. Morell, J. Serrano, Missing data in traffic estimation: A variational autoencoder imputation method, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 2882–2886.

[8] T. Zhang, D. Cao, X. Feng, J. Zhu, X. Lu, L. Mu, H. Qian, Machine learning prediction

of bio-oil characteristics quantitatively relating to biomass compositions and pyrolysis conditions, Fuel 312 (2022) 122812.

[9] Q. Tang, Y. Chen, H. Yang, M. Liu, H. Xiao, Z. Wu, H. Chen, S. Naqvi, Prediction of bio-oil yield and hydrogen contents based on machine learning method: effect of biomass compositions and pyrolysis conditions, Energy & Fuels 34 (2020) 11050–11060.

[10] K. Yang, K. Wu, H. Zhang, Machine learning prediction of the yield and oxygen content of bio-oil via biomass characteristics and pyrolysis conditions, Energy 254 (2022) 124320.