Case Studies in Thermal Engineering 50 (2023) 103419

Contents lists available at ScienceDirect

Case Studies in Thermal Engineering

journal homepage: www.elsevier.com/locate/csite

Optimization of thermal biofuel production from biomass using CaO-based catalyst through different algorithm-based machine learning approaches

Jiangbo Tang [a,b,*], Ali Kareem Abbas [c], Nisar Ahmad Koka [d], Naiser Sadoon [e], Jamal K. Abbas [f], Rasha Ali Abdalhuseen [g], Munther Abosaooda [h], Naked Mahmood Ahmed [i], Ali Hashim Abbas [j]

[a] School of Engineering, Guangzhou College of Technology and Business, Foshan, Guangdong, 528100, China
[b] Institute of New Generation Electronic Information Technology, Guangzhou College of Technology and Business, Foshan, Guangdong, 528100, China
[c] Intelligent Medical Systems Department, Al-Mustaqbal University College, 51001, Hillah, Babil, Iraq
[d] Department of English, Faculty of Languages and Translation, King Khalid University, Abha, Kingdom of Saudi Arabia
[e] Medical Lab. Techniques Department, College of Medical Technology, Al-Farahidi University, Iraq
[f] AL-Nisour University College, Baghdad, Iraq
[g] Department of Pharmacy, AlNoor University College, Nineveh, Iraq
[h] College of Pharmacy, The Islamic University, 54001, Najaf, Iraq
[i] National University of Science and Technology, Dhi Qar, Iraq
[j] College of Information Technology, Imam Ja'afar Al-Sadiq University, Al-Muthanna, 66002, Iraq

ARTICLE INFO

Handling Editor: Huihe Qiu

Keywords:
Biodiesel
Optimization
Modeling
Machine learning
Energy

ABSTRACT

Optimization of biofuel production from algal oil through utilizing a CaO-based catalyst was carried out in this study. The optimal point for the highest yield of the reactions was determined using machine learning. To implement the optimization task, and to make predictions, we used three different methods, including Quantile regression, Logistic regression, and Gradient Boosted Decision Trees. The regression problem includes the amount of Catalyst, Reaction time, and Methanol/oil as input features, and FAME (fatty acid methyl ester) yield is the single output. We tuned the boosted version of these models with their important hyper-parameters and selected their best combination. Then different standard metrics are employed to assess their performance of them. Considering R^2 score, Quantile regression, Logistic regression, and Gradient Boosted Decision Trees have error rates of 0.934, 0.996, and 0.998, and with MAE, they have 1.94, 1.68, and 1.17 errors, respectively. Also, Considering MAPE 2.14×10^{-2}, 1.89×10^{-2}, and 1.29×10^{-2} values obtained. Gradient Boosting is selected as the most appropriate model finally. Furthermore, the optimal output value with the proposed approach is 97.50, with the input vector being (x1 = 153, x2 = 0.625, x3 = 20).

1. Introduction

Optimization of biofuel production has been a subject of great interest for sustainable development and expansion of renewable

* Corresponding author. School of Engineering, Guangzhou College of Technology and Business, Foshan, Guangdong, 528100, China.
E-mail address: ggstjb@126.com (J. Tang).

https://doi.org/10.1016/j.csite.2023.103419
Received 18 March 2023; Received in revised form 20 August 2023; Accepted 23 August 2023
Available online 23 August 2023

Table 1
List of the experimental data [4,20].

Run	X1 = Reaction time	X2 = Catalyst amount	X3 = Methanol:oil	Y= FAME yield (%)
1	60	0	40	17.88
2	60	1	20	23.14
3	120	1	40	93.34
4	120	1	40	92.45
5	120	0	20	5.82
6	180	0	40	16.94
7	60	2	40	9.09
8	180	1	20	96.15
9	180	2	40	80.83
10	120	0	60	29.65
11	60	1	60	32.65
12	120	2	60	23.35
13	120	1	40	95.65
14	120	2	20	83.65
15	180	1	60	28.46
16	120	1	40	91.67
17	120	1	40	88.49

energy sources for the society. Basically, biofuel can be produced from different sources such as biomass which is considered as sustainable feedstock for production of biodiesel. FAME is known as the main component of biodiesel which can be obtained utilizing esterification or/and transesterification reaction in a reactor operating in either batch or continuous mode. The reaction features need to be controlled to achieve the best yield of biodiesel production. Biodiesel production relies heavily on four main input parameters: temperature, time, catalyst content, and methanol to oil ratio [1–8].

Indeed, the relationship between the input and outputs parameters must be determined to optimize the process. This task can be implemented by optimization and development of process models such as mechanistic models or machine learning models. These process models need to be precisely tuned in order to obtain the best description and optimization of the process. Furthermore, numerical schemes are required to be developed for optimization of the process and solution of the governing equations. Machine learning (ML) can be considered a branch of computer science mainly used in experimental science. This discipline is a natural consequence of Computer Science and Statistics junction and tries to extract useful information from any data set. So, ML is applicable anywhere with some experimental data and Motivation to find some relationship between some features and some targets [9–11].

Logistic regression [12] is a kind of generalized linear regression analysis that is particularly well suited for multivariable control applications. In contrast to common linear regression models, the logistic regression model limits the output value to the range (0,1) [13].

Quantile regression is effective when predicting an interval rather than a single point. Prediction intervals are sometimes calculated under the assumption that the standard deviation of the error in the prediction is zero and constant. Even for errors with non-constant variance or a non-normal distribution, quantile regression provides reasonable prediction intervals [14,15].

Ensemble methods, especially Tree-based models are also strong and popular methods. As an ensemble method we use gradient boosting on the top of decision trees. Gradient boosting is a powerful ML model with numerous successful usages in classification and regression problems in various domains similar to our problem [16,17]. This is an ensemble-based method comprised of several basic predictors. Based on data from a bootstrap sample, we built each base predictor as an individual tree model, which was then divided into regions and a basic model was fitted to each region [17–19].

2. Data set

With only 17 data points in total, we have a very small dataset with only three inputs and one output in this study. There are three features to consider: X1 = reaction time, X2 = catalyst amount, then X3 = ratio of methanol to oil. The aim is to produce FAME (fatty acid methyl ester). The entire data is depicted in Table 1 [4,20].

3. Methodology

3.1. Quantile regression

In 1978, Koenker and Bassett extended the traditional regression model by developing quantile regression as an extension of that model. Panel quantile regression is type of this model utilized to a panel of data [21,22]. Considered a linear model of the τ^{th} quantile:

$$y_i = x_i^T \beta_\tau + e_i, \mathrm{i} \in \{1, 2, \ldots, \mathrm{n}\}$$

the τ^{th} quantile of e_i is equal to 0. The estimator of β_τ gained through below equation:

$$\widehat{\beta_\tau} = argmin \sum_{i=1}^{n} \rho_\tau \left(y_i - x_i^T \beta\right)$$

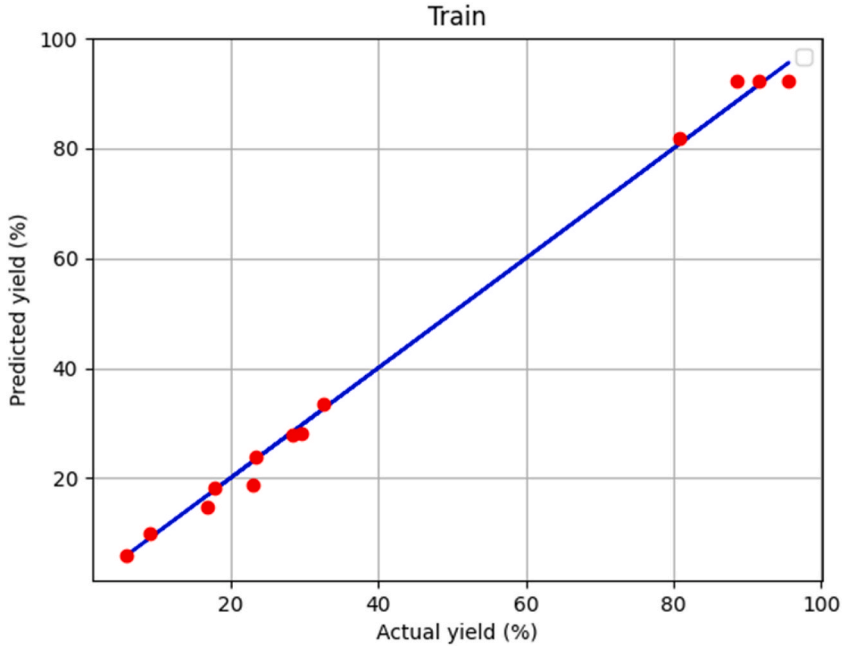

Fig. 1. Quantile regression train phase.

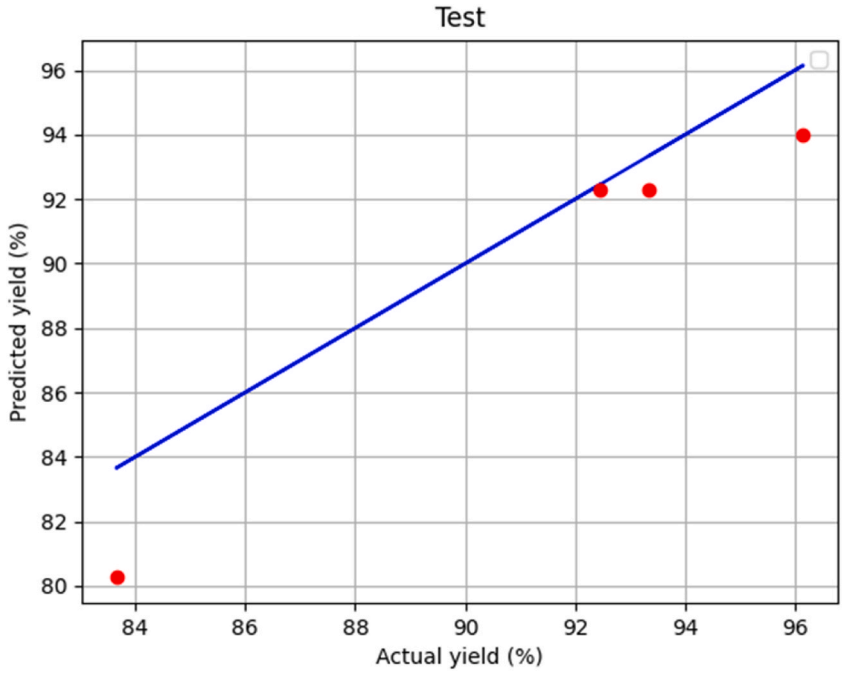

Fig. 2. Quantile regression test phase.

Instead of optimizing the sum of squared residuals, as ordinary least squares (OLS) does, quantile regressions employ the conditional quantiles of the dependent items to do so [23].

Quantile models, in comparison with linear methods, are typically less biased to skewed data and are able to get a wider range of outcomes. Thus, the approach fit non-normally distributed data very well and generate stronger outputs [24].

3.2. Logistic regression

Logistic Regression is a multivariable control approach that is based on generalized linear regression analysis. In contrast to

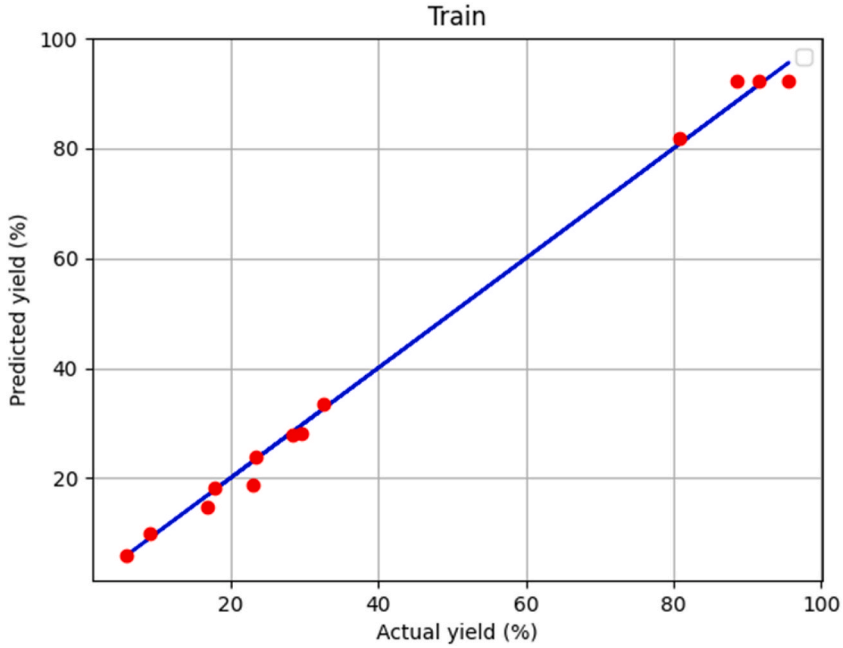

Fig. 3. Logistic regression train phase.

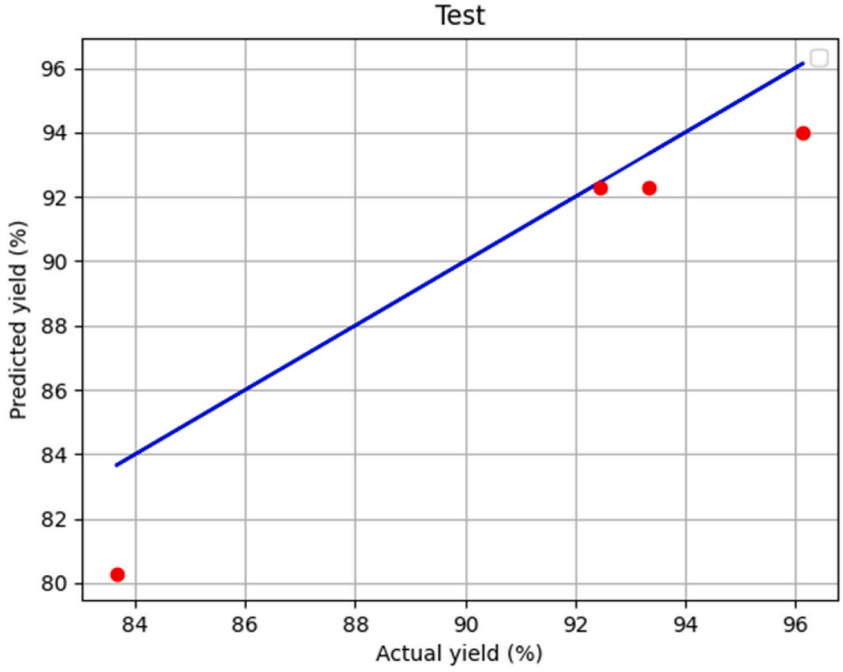

Fig. 4. Logistic regression test phase.

traditional linear models, the Logistic model uses a sigmoid function to limit the target value to the interval (0,1). For the purposes of this study, we will refer to the random variables at hand as "features," while Y will serve as the binary response variable of interest. To assess the provisional presumption, $P(Y = 1 \mid X_1, X_2, \ldots, X_p)$, the logistic regression approach uses $X_1, \ldots, Z_p$ [25,26]:

$$P(Y = 1 \mid X_1, \ldots, X_p) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}$$

Estimated from the data set via maximum likelihood, the regression coefficients β_0, β_1, ..., β_p are simply referred to as regression

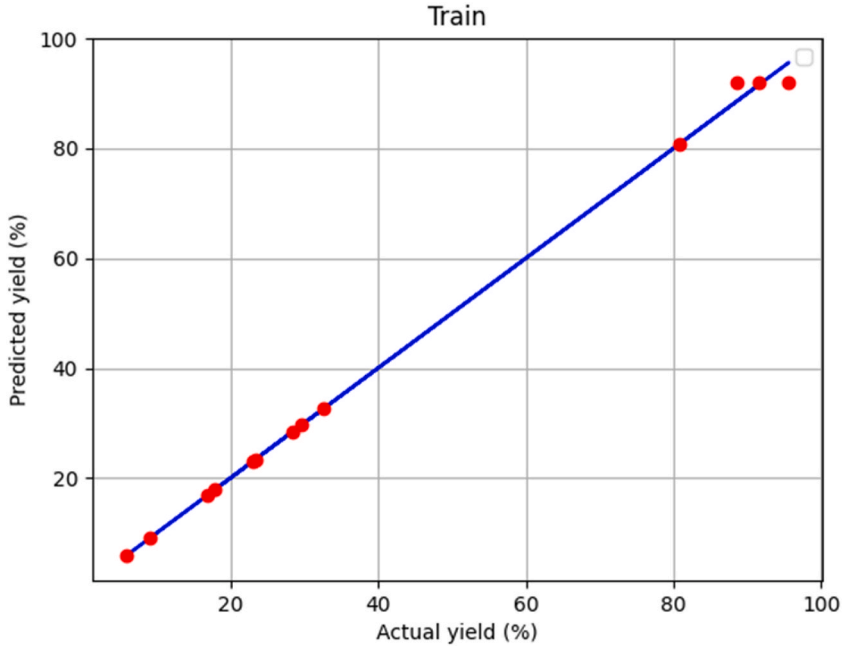

Fig. 5. Gradient boosting train phase.

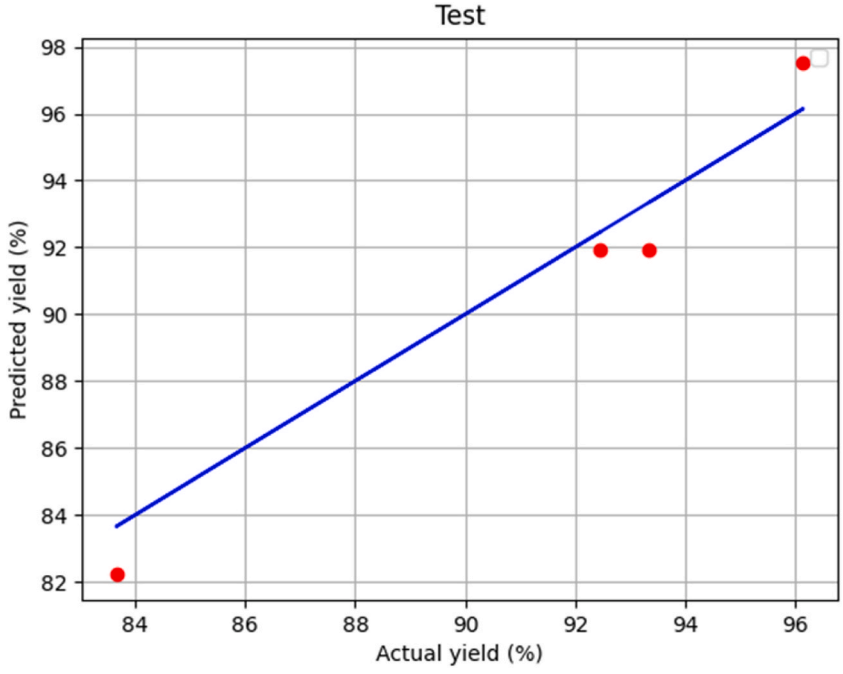

Fig. 6. Gradient boosting test phase.

Table 2
Outputs of developed approaches.

Models	MAE	R^2	MAPE
Gradient Boosting	1.17	0.998	1.29E-02
Logistic regression	1.68	0.996	1.89E-02
Quantile Regression	1.94	0.934	2.14E-02

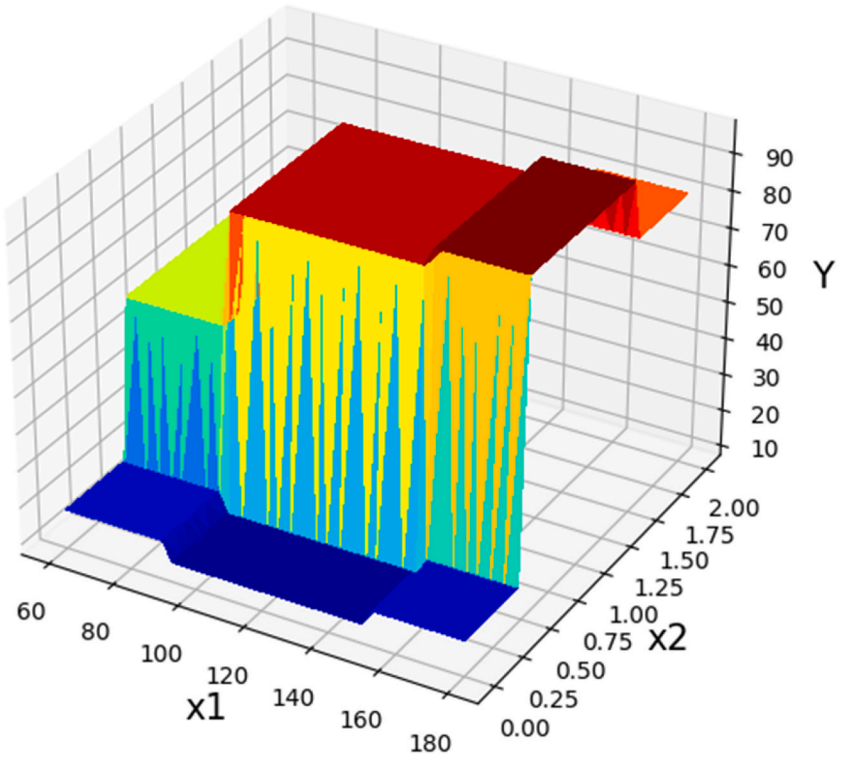

Fig. 7. Projection of X1 and X2 with estimation surface in the final GBRT method. X3 = 40 is regarded as Constant. The optimal value of y is 96.86 when x1 = 153 and x2 = 0.529.

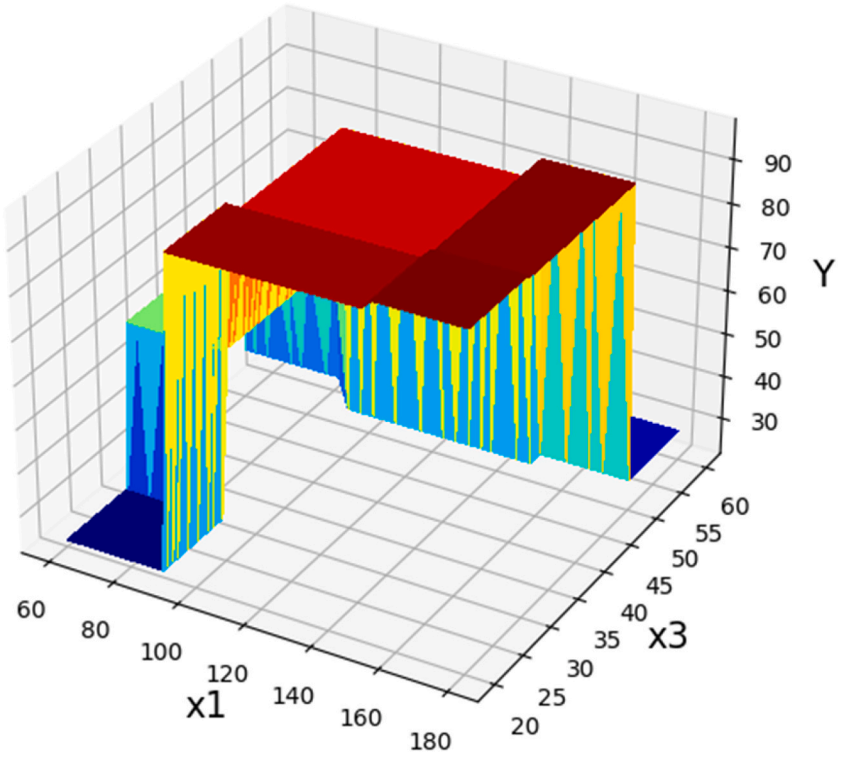

Fig. 8. Projection of X1 and X3 with estimation surface in the final GBRT approach. X2 = 1 is regarded as Constant. The optimal value of y is 97.46 when x1 = 153, x3 = 20.0.

coefficients. $Y=1$ probability for an unseen sample is then predicted by replacing the β's in the equation above with their predicted counterparts and the X's with their realizations for the new data point (sample) [27].

3.3. Gradient boosting

One of the superlative well-known approach is the Gradient Boosting Machine (GBM) [18]. This method has recently inspired the creation of a number of noteworthy ensembles, which we will discuss later [28].

Each iteration of the Gradient Boosting Machine represents the steepest descent reduction of a specific loss function, making the model a stage-wise additive one. Numerical optimization is used to compute the predictive function in the function space. GBM can employ a decision tree or a lineal regression as its basis learner, however most practitioners use Gradient Boosting Decision Trees (GBDT). The following algorithm is a pseudo-code description of the generic Gradient Boosting learning technique [17,18,29–31].

Function GBM (D=$\{(x_i y_i)\}_{j=1}^{N} - training\ set$, $n\ Estimators$ –

$number\ learners, L\big(y, F(x)\big) - loss\ function, v - learning\ rate$)

Initialize:$F_0\ (x)\ = argmin_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$

For t in [1,n Estimators] do

//Compute pseudo-residual $r_{i,m}$ for each sample x_i

For I in [1,N] do

$$r_i = -\left[\frac{\partial L(y,F(x_i))}{\partial F(x_i)}\right]_{F(x)=F(x)_{I-1}}$$

End for

T_t=Fit -Regression -Tree(D,r)

For i in [1,J] do

$R_j = Name - leaves(T_t)$

$//compute\ output\ \gamma_j of\ each\ leaf\ J$

$\gamma_j = argmin_\gamma \sum_{x_i \in R_J} L(y_i, F_{t-1}(x_i) + \gamma)$

End for

$$F_t(x) = F_{t-1}(x) + v\sum_{j=1}^{J} \gamma_j I(x \in R_j)$$

End for

Output: F(x)

End function

4. Results and discussion

The mentioned methods are tuned with their important hyper parameters. This tuning is done with the help of genetic algorithm. In fact, various combinations of possible amounts for hyper-parameters are considered as individuals in the genetic algorithm, and to prevent overfitting, a special fitting function is considered with the K-fold mechanism.

Then their performance are examined trough these metrics [32]:.

- R^2 score: $1 - \frac{\sum_{i=1}^{m}(Predicted\ Effort_i - Observed\ Effort_i)^2}{\sum_{i=1}^{m}(Observed\ Effort_i - \overline{Average\ Effort})^2}$.
- MAE is the arithmetic mean of errors between observed and expected effort: $MAE = \frac{1}{N}\sum_{i=1}^{N}|Observed\ Effort_i - PredictedEffort_i|$..

- RMSE: The square root of the mean square of observed and projected effort differences $RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Observed\ Effort_i - Predicted\ Effort_i)^2}{N}}$..

By comparing Figs. 1, 3 and 5, we can consider the GB model is the most accurate model in the training stage. Nevertheless, this must also be confirmed in the testing phase. Therefore, we compare Figs. 2, 4 and 6. By placing the information of these diagrams with Table 2, this fact is confirmed, so we choose the boosting Gradient model as the model with the best generality.

Fig. 7 shows the 3D demonstration for showing the simultaneous influence of the reaction time and catalyst amount on the efficiency of biofuel production when the value of methanol to oil ratio is constant. Fig. 8 demonstrates the 3D depiction for evaluating the simultaneous impact of the reaction time and methanol to oil ratio on the efficiency of biofuel production when the value of catalyst amount isn't changed [20]. Additionally, Fig. 9 presents the 3D projection for evaluating the simultaneous impact of the methanol to oil ratio and catalyst amount on the efficiency of biofuel production when the value of reaction time is constant. Figs. 10–12 show the impact of reaction time, catalyst amount and methanol to oil ratio as individual parameter on the efficiency of biofuel production. As can be seen from the figures, at the beginning, the biodiesel production efficiency improves instantly by increasing the catalyst amount, but later, it begins to decline. The increased efficiency of biodiesel may be rationalized by the fact that catalytic reactions involving triglycerides can be sped up by using an excessive enough reaction of catalyst. Conversely, increasing the reaction time has a beneficial impact on biofuel production efficiency by increasing the rate of fatty acid conversion. But by achieving an efficiency higher than the maximum amount of that, the biofuel production starts decreasing. Moreover, at the beginning of process, through enhancing the methanol to oil ratio, the efficiency of biofuel production significantly improves, but after that before reaching the highest efficiency point it slowly begins to decline. Therefore, it is worth noting that the efficiency of biodiesel production decreases by reducing the methanol quantity [33,34]. Table 3 presents the optimized values of the parameters for reaching the maximum efficiency of biofuel production.

5. Conclusion

This study used three methods to make predictions: Quantile regression, Logistic regression, and Gradient-Boosted Decision Trees. Reaction time, Catalyst amount, Methanol/oil as input features, and FAME yield as the single output feature in the regression problem. We selected the best combination for these models' boosted versions using hyper-parameter tuning. Then, in order to gauge how well they are doing, various standard metrics are applied. Quantile regression, Logistic regression, and Gradient Boosted Decision Trees have R^2 (coefficient of determination) of 0.934, 0.996, and 0.998, respectively, and with MAE, they have 1.94, 1.68, and 1.17 errors. The MAPE values of 2.14×10^{-2}, 1.89×10^{-2}, and 1.29×10^{-2} are also considered. Finally, the most general and accurate model is Gradient Boosting. Using this approach, the best output value is 97.50, with inputs (x1 = 153, x2 = 0.625, x3 = 20).

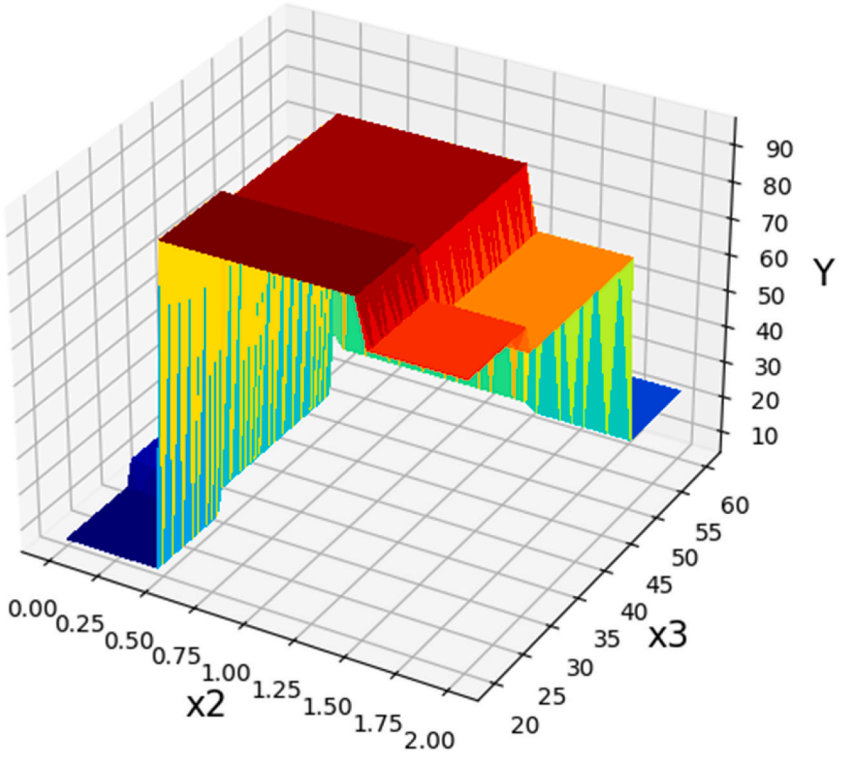

Fig. 9. Projection of X2 and X3 with estimation surface in the final GBRT approach. X1 = 120 is regarded as Constant. The optimal value of y is 97.46 when x2 = 0.529, x3 = 20.0.

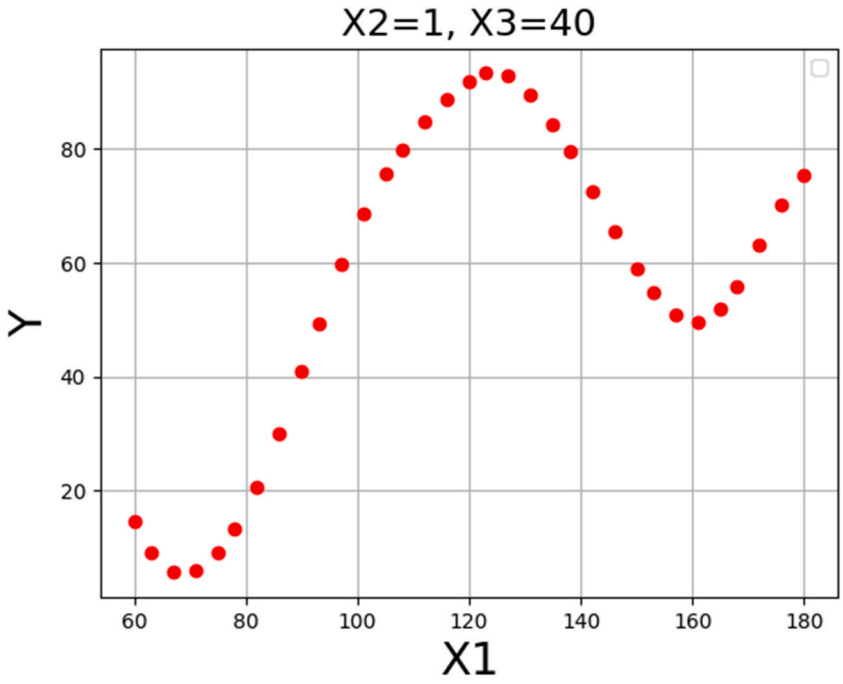

Fig. 10. Tendency graph of x1.

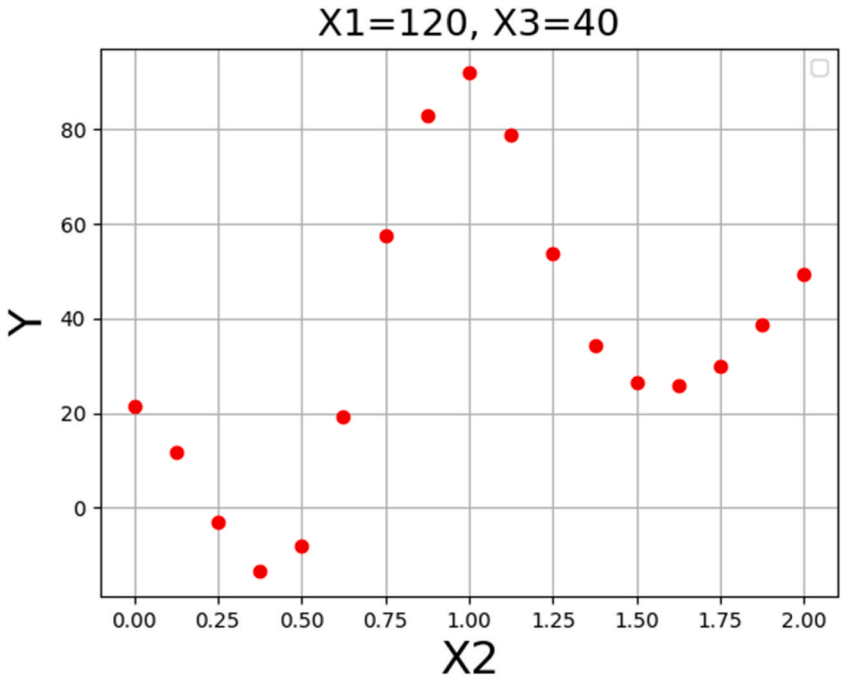

Fig. 11. Tendency graph of x2.

Author statement

Jiangbo Tang: Writing, editing, modeling and simulation, resources, validation.
Ali Kareem Abbas: Writing, resources, investigation, formal analysis.
Nisar Ahmad Koka: Writing, editing, analysis, validation, modeling.
Naiser Sadoon: Conceptualization, formal analysis, investigation, editing.

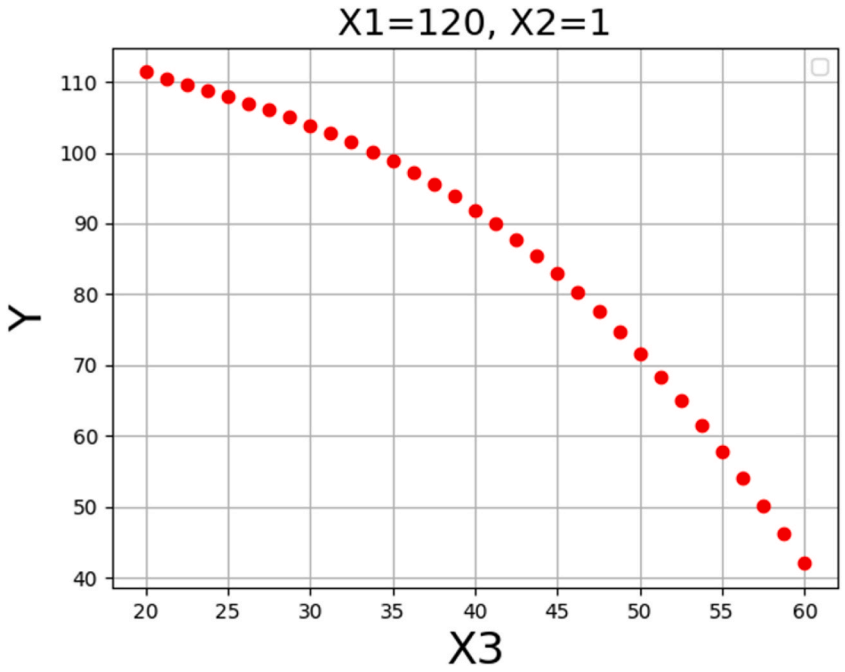

Fig. 12. Tendency graph of x3.

Table 3
Maximum response from the parameters at their optimal settings.

X1 = Reaction time	X2 = Catalyst amount	X3 = Methanol:oil	Y= FAME yield (%)
153	0.625	20	97.50

Jamal K. Abbas: formal analysis, investigation, editing, simulation, validation.
Rasha Ali Abdalhuseen: formal analysis, investigation, editing, simulation, validation.
Munther Abosaooda: Writing, editing, analysis, validation, modeling.
Naked Mahmood Ahmed: Writing, editing, analysis, validation, modeling.
Ali Hashim Abbas: formal analysis, investigation, editing, simulation, validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data are available within the published paper.

Acknowledgements

The authors extend their appreciation and thanks to the Deanship of Scientific Research at King Khalid University, Abha, KSA for funding this work under Research Grant number GRP/306/43.

This work is funded by Guangdong Educational Science Planning Project (Project No.: 2021GXJK488), 2022 Guangdong Undergraduate University Teaching Quality and Teaching Reform Project (Project No.: PX- 9322903), 2021 Ministry of Education Industry-University Cooperation Collaborative Education Project (Project No.: 202102211099), 2021 GGS Higher Education Teaching Reform Project (Project No.: ZL20211139); GGS Curriculum Ideology and Politics Project (Project No.: KCSZ202209).

References

[1] V.G. Tacias-Pascacio, et al., Comparison of acid, basic and enzymatic catalysis on the production of biodiesel after RSM optimization, Renew. Energy 135 (2019) 1–9.
[2] M.K. Lam, K.T. Lee, A.R. Mohamed, Homogeneous, heterogeneous and enzymatic catalysis for transesterification of high free fatty acid oil (waste cooking oil) to biodiesel: a review, Biotechnol. Adv. 28 (4) (2010) 500–518.

[3] M. Collotta, et al., Life cycle analysis of the production of biodiesel from microalgae, in: R. Basosi, et al. (Eds.), Life Cycle Assessment of Energy Systems and Sustainable Energy Technologies: the Italian Experience, Springer International Publishing, Cham, 2019, pp. 155–169.
[4] V. Narula, et al., Low temperature optimization of biodiesel production from algal oil using CaO and CaO/Al2O3 as catalyst by the application of response surface methodology, Energy 140 (2017) 879–884.
[5] S. Yahya, S.K. Muhamad Wahab, F.W. Harun, Optimization of biodiesel production from waste cooking oil using Fe-Montmorillonite K10 by response surface methodology, Renew. Energy 157 (2020) 164–172.
[6] M. Mohadesi, et al., Production of biodiesel from waste cooking oil using a homogeneous catalyst: study of semi-industrial pilot of microreactor, Renew. Energy 136 (2019) 677–682.
[7] A.A. Albuquerque, et al., Reactive separation processes applied to biodiesel production from residual oils and fats: design, optimization and techno-economic assessment of routes using solid catalysts, Energy 240 (2022), 122784.
[8] W.K. Abdelbasset, et al., Development of multiple machine-learning computational techniques for optimization of heterogenous catalytic biodiesel production from waste vegetable oil, Arab. J. Chem. 15 (6) (2022), 103843.
[9] T.M. Mitchell, The Discipline of Machine Learning, 9, Carnegie Mellon University, School of Computer Science, Machine Learning, 2006.
[10] J.G. Carbonell, R.S. Michalski, T.M. Mitchell, An Overview of Machine Learning, Machine learning, 1983, pp. 3–23.
[11] I. Goodfellow, Y. Bengio, A. Courville, Machine learning basics, Deep learning 1 (7) (2016) 98–164.
[12] R.E. Wright, Logistic Regression, 1995.
[13] T.F. Jaeger, Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models, J. Mem. Lang. 59 (4) (2008) 434–446.
[14] Y. Yang, et al., Power load probability density forecasting using Gaussian process quantile regression, Appl. Energy 213 (2018) 499–509.
[15] J.W. Taylor, D.W. Bunn, A quantile regression approach to generating prediction intervals, Manag. Sci. 45 (2) (1999) 225–237.
[16] T. Dube, et al., Intra-and-inter species biomass prediction in a plantation forest: testing the utility of high spatial resolution spaceborne multispectral rapideye sensor and advanced machine learning algorithms, Sensors 14 (8) (2014) 15348–15370.
[17] Q. Xu, et al., PDC-SGB: prediction of effective drug combinations using a stochastic gradient boosting algorithm, J. Theor. Biol. 417 (2017) 1–7.
[18] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. (2001) 1189–1232.
[19] V.-H. Truong, et al., A robust method for safety evaluation of steel trusses using Gradient Tree Boosting algorithm, Adv. Eng. Software 147 (2020), 102825.
[20] Y. Liu, et al., Novel and robust machine learning model to optimize biodiesel production from algal oil using CaO and CaO/Al2O3 as catalyst: sustainable green energy, Environ. Technol. Innovat. 30 (2023), 103018.
[21] R. Koenker, G. Bassett Jr., *Regression Quantiles.* Econometrica, journal of the Econometric Society, 1978, pp. 33–50.
[22] M. Xie, Z. Huang, W. Zang, The inequality of health-income effect in employed workers in China: a longitudinal study from China Family Panel Studies, Int. J. Equity Health 19 (1) (2020) 1–15.
[23] L. Zhang, J.H. Gove, L.S. Heath, Spatial residual analysis of six modeling techniques, Ecol. Model. 186 (2) (2005) 154–177.
[24] Y. Tian, M. Tang, M. Tian, A class of finite mixture of quantile regressions with its applications, J. Appl. Stat. 43 (7) (2016) 1240–1252.
[25] D.G. Kleinbaum, et al., Logistic Regression, Springer, 2002.
[26] X. Yin, et al., A flexible sigmoid function of determinate growth, Ann. Bot. 91 (3) (2003) 361–371.
[27] D.W. Hosmer Jr., S. Lemeshow, R.X. Sturdivant, Applied Logistic Regression, 398, John Wiley & Sons, 2013.
[28] M.H.D.M. Ribeiro, L. dos Santos Coelho, Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series, Appl. Soft Comput. 86 (2020), 105837.
[29] S. Borra, A. Di Ciaccio, Improving nonparametric regression methods by bagging and boosting, Comput. Stat. Data Anal. 38 (4) (2002) 407–420.
[30] A.J. Ferreira, M.A. Figueiredo, Boosting Algorithms: A Review of Methods, Theory, and Applications, Ensemble machine learning, 2012, pp. 35–85.
[31] R. Maclin, D. Opitz, An Empirical Evaluation of Bagging and Boosting, AAAI/IAAI, 1997, pp. 546–551, 1997.
[32] A. Botchkarev, Evaluating Performance of Regression Machine Learning Models Using Multiple Error Metrics in Azure Machine Learning Studio, 2018. Available at: SSRN 3177507.
[33] V. Narula, et al., Process parameter optimization of low temperature transesterification of algae-Jatropha Curcas oil blend, Energy 119 (2017) 983–988.
[34] K. Boonmee, et al., Optimization of biodiesel production from Jatropha oil (Jatropha curcas L.) using response surface methodology, Agriculture and Natural Resources 44 (2) (2010) 290–299.