energies

MDPI

Article

Machine Learning-Based Predictive Modelling of Biodiesel Production—A Comparative Perspective

Krishna Kumar Gupta [1], Kanak Kalita [2,*], Ranjan Kumar Ghadai [3], Manickam Ramachandran [4] and Xiao-Zhi Gao [5]

[1] Department of Mechanical Engineering, MPSTME, SVKM's Narsee Monjee Institute of Management Studies (NMIMS), Shirpur Campus, Dhule 425 405, India; krishnakumar.gupta@nmims.edu
[2] Department of Mechanical Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi 600 062, India
[3] Department of Mechanical Engineering, Sikkim Manipal Institute of Technology, Sikkim Manipal University, Majitar 737 136, India; ranjan.g@smit.smu.edu.in
[4] Data Analytics Lab, REST Labs, Kaveripattinam, Krishnagiri 635 112, India; ramachandran.manickam@restlabs.in
[5] School of Computing, University of Eastern Finland, FI-70211 Kuopio, Finland; xiao-zhi.gao@uef.fi
* Correspondence: drkanakkalita@veltech.edu.in

Abstract: Owing to the ever-growing impetus towards the development of eco-friendly and low carbon footprint energy solutions, biodiesel production and usage have been the subject of tremendous research efforts. The biodiesel production process is driven by several process parameters, which must be maintained at optimum levels to ensure high productivity. Since biodiesel productivity and quality are also dependent on the various raw materials involved in transesterification, physical experiments are necessary to make any estimation regarding them. However, a brute force approach of carrying out physical experiments until the optimal process parameters have been achieved will not succeed, due to a large number of process parameters and the underlying non-linear relation between the process parameters and responses. In this regard, a machine learning-based prediction approach is used in this paper to quantify the response features of the biodiesel production process as a function of the process parameters. Three powerful machine learning algorithms—linear regression, random forest regression and AdaBoost regression are comprehensively studied in this work. Furthermore, two separate examples—one involving biodiesel yield, the other regarding biodiesel free fatty acid conversion percentage—are illustrated. It is seen that both random forest regression and AdaBoost regression can achieve high accuracy in predictive modelling of biodiesel yield and free fatty acid conversion percentage. However, AdaBoost may be a more suitable approach for biodiesel production modelling, as it achieves the best accuracy amongst the tested algorithms. Moreover, AdaBoost can be more quickly deployed, as it was seen to be insensitive to number of regressors used.

Keywords: biodiesel; machine learning; linear regression; random forest regression; AdaBoost regression

Citation: Gupta, K.K.; Kalita, K.; Ghadai, R.K.; Ramachandran, M.; Gao, X.-Z. Machine Learning-Based Predictive Modelling of Biodiesel Production—A Comparative Perspective. *Energies* **2021**, *14*, 1122. https://doi.org/10.3390/en14041122

Academic Editors: Amir Mosavi and Annamária R. Várkonyi-Kóczy

Received: 25 January 2021
Accepted: 14 February 2021
Published: 20 February 2021

1. Introduction

Biodiesel is a kind of nontoxic and biodegradable fuel, which is aromatic and sulfur free. The production of biodiesel is due to the transesterification reaction between the waste oil and different alcohols like methanol or ethanol [1,2]. In recent years, the importance of biodiesel has been increasing due to the reduction of total petroleum reserves in the world and the increase of global warming issues. It has been observed that each year, the use of petroleum is increasing exponentially, which can lead to complete depletion of its reserves by 2042 [3]. This alarming situation is boosting the interest of researchers to actively work on the substitute of petroleum products. Biodiesel can be considered as an alternative to fossil fuels used in the diesel engine without any major modification. Nowadays, researchers are showing interest in the use of vegetable oil to produce biodiesel, which is

less polluting compared to petroleum. If biodiesels are used effectively for the generation of energy, it would be extremely beneficial for the environment as well as for the job creation of local people. Since many biodiesel sources are indigenous (for example, *Calophyllum inophyllum* is native to Asia [4]) or abundant (for example, Brazil and US leads in soybean production [5]) to specific regions, its usage will motivate the local farmers to cultivate such sources. In recent years, various processes have been used to produce biodiesel from different oils like soybean oil, canola oil, palm oil, sunflower seed oil etc. [6–8].

Various operating parameters such as reaction time, reaction temperature, the molar ratio of oil/alcohol, catalyst yield percentage etc. can significantly affect the yield of biodiesel using transesterification reaction [9]. The optimization of various factors of transesterification reaction needs the conduction of many experiments, as well as the suitable statistical tool which can predict the effect of each factor over the reaction and their interactions. A response surface methodology (RSM) approach has been used widely for the optimization of process factors to get the required output by conducting a smaller number of experiments [10,11]. Various designs of experiment (DOE), like Taguchi experimental design, factorial design, central composite design (CCD), box-banging design (BBD) etc., are used to perform the number of experiments. Different authors have used different DOE and statistical techniques to optimize process factors. Hameed et al. used central composite design (CCD) to optimize different operating factors such as reaction time, catalyst amount and methanol/oil molar ratio for making the biodiesel from palm oil. It is observed that 9.72 h of reaction time, 11.43% of methanol/oil ratio and 5.52% of catalyst amount can maximize the biodiesel yield to 89.23% [12]. Ahmad et al. [13] used the face-centred central composite design (FCCD) method to optimize the operating parameters, like reaction temperature, the volume ratio of methanol/oil, reaction time and catalyst weight percentage for getting the highest yield of biodiesel performing through a transesterification reaction. Their results showed that using flaxseed oil at 33 min reaction time, 0.51% catalyst, 59 °C reaction temperature and 1:5.9 molar ratio of flaxseed oil to methanol can maximize the yield of biodiesel to 99.5% [13]. Jayaprabakar et al. developed biodiesel by using sheepskin fat of New Zealand origin (SSFNO) and 7% free fatty acid (FFA) content and used the response surface methodology (RSM) to optimize the esterification of SSNFO. Around 92% yield of SSFNO was achieved by optimizing different operating factors like reaction temperature, methanol: SSFNO mole ratio and reaction time [14]. Matinja et al. used BBD technique to optimize the process factors like palm oil mill effluent (POME)/methanol ratio, the weight of the immobilized beads, reaction time and agitation speed to produce biodiesel from POME. The highest yield was achieved at a 300-rpm agitation speed, 5 h of incubation, methanol/oil molar ratio (6:1), and 2 kg weight of the immobilized beads [15].

In this work, three machine learning algorithms—linear regression, random forest regression and AdaBoost regression—are used to predict the biodiesel production process responses based on the process parameters. To illustrate the methodology and carry out a robust study, two separate examples from the literature are presented. In the first example, the biodiesel production yield is estimated based on the process parameters. In the second example, the biodiesel free fatty acid conversion percentage is estimated based on the production process parameters. The machine learning regression predictive models are comprehensively compared using several error metrics.

2. Materials and Methods

2.1. Machine Learning Algorithms

2.1.1. Linear Regression

The linear regression predicts the dependent variable (y) by expressing it as a linear function of the predictor variable (x). In general, for a single predictor variable regression problem, the y is monotonically linked to the x i.e., any change in x will increase/decrease y. A scatter plot between x and y reveals the general trend, on which then a best line fit is done of the following form [16],

The general formula for the linear regression is:

$$y = \mathrm{a} + \mathrm{b}x \tag{1}$$

where a and b represent y-intercept and slope of the fitting line of the linear regression, respectively. In case the best fit line passes through origin, the equation becomes

$$y = \mathrm{b}x \tag{2}$$

Generalized linear regression model for any response (y) and its predictors ($x_1, x_2, \ldots, x_n$), i.e., biodiesel production process parameters in this case can be stated as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_n \tag{3}$$

2.1.2. Random Forest Regression

Random forest regression is an ensemble technique based on the concept of bagging and random subspace methods. Bagging or bootstrap aggregation is responsible for creating an ensemble of learner trees. These learner trees are trained on separate and independent bootstrap samples drawn from the original training dataset. From an original training dataset, D with N examples a bootstrap sample ($\mathrm{D_b}$) is constituted by randomly drawing n examples from D with replacement. $\mathrm{D_b}$ is approximately two-third of D, without any duplicate examples. K number of independent regression trees are created for the bootstrap samples with input vector x. These regression trees generally have low bias and high variance. The random forest ensemble prediction is then obtained by calculating the mean of the prediction of the K regression trees, $h_k(x)$ [17].

$$\text{Random forest prediction} = \frac{1}{K}\sum_{k=1}^{K} h_k(x) \tag{4}$$

Bagging reduces the variance in the ensemble model as compared to the individual decision trees. It also prevents overfitting of the model. It is important that the regression trees are not correlated. Samples other than those selected for training the kth regression tree during bagging are grouped as another subset called out-of-bag data (OOB). OOB data constitute roughly one-third of D. The kth regression tree's performance is evaluated using the OOB data using the following equation,

$$MSE_{OOB} = \frac{1}{n}\sum_{i=i}^{n}(y_i - \overline{y}_{iOOB})^2 \tag{5}$$

where MSE_{OOB} is the mean squared error, y_i is the ith prediction and $\overline{y}_{iOOB}$ is the mean of ith prediction from all the trees.

Similarly, the $R^2{}_{OOB}$ is computed using the following equation,

$$R^2{}_{OOB} = 1 - MSE_{OOB}/Var_y \tag{6}$$

where V_{ar_y} is the total variance of the response.

2.1.3. Adaptive Boosting Regression

AdaBoost or adaptive boosting is a sequential ensemble technique which is based on the principle of developing several weak learners using different training sub-sets drawn randomly from the original training dataset. During each training, weights are assigned which is used when learning each hypothesis. The weights are used for computation of the error of the hypothesis on the dataset and is an indicator of the comparative importance of each instance. The weights are recalculated after every iteration, such that incorrectly classified instances by the last hypothesis receive higher weights. This enables the algorithm to focus on more difficult to learn instances. Assigning revised weights to the incorrectly

classified instances is the most vital task of the algorithm. Unlike in classification, in regression, the instances are not correct or incorrect, rather they constitute a real-value error. By comparing the computed error to a predefined threshold prediction error, it can be labelled as an error or not error and thus, the AdaBoost classifier can be used. Instances with larger error on previous learners are more likely (i.e., higher probability) to be selected for training the subsequent base learner. Finally, weighted average or median is used to provide an ensemble prediction of the individual base learner predictions [17].

2.1.4. Predictive Model Performance Evaluation Metrics

The residue, ε_i between the ith original value, y_i and predicted value, $\hat{y}_i$ is calculated as,

$$\varepsilon_i = y_i - \hat{y}_i \tag{7}$$

The coefficient of determination, R^2 is computed using the y_i, $\hat{y}_i$ and the mean of the dataset, $\overline{y}$.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{8}$$

The mean-absolute-error, MAE, is computed based on the number of the samples, n as

$$\text{MAE} = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \tag{9}$$

The mean-squared-error, MSE, is computed as

$$\text{MSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} \tag{10}$$

The root-mean-squared-error, RMSE, is computed as

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{11}$$

The maximum error is computed as

$$\text{Max. Error}\ (y,\ \hat{y}) = \text{max.}\ (|y_i - \hat{y}_i|) \tag{12}$$

The median error is computed as

$$\text{Med. Error}\ (\text{y}, \hat{y}) = \text{median}\ (|\text{y}_1 - \hat{y}_1|, \ldots, |\text{y}_\text{n} - \hat{y}_\text{n}|) \tag{13}$$

3. Results and Discussion

3.1. Example 1: Biodiesel Production Yield Estimation

The experimental data (presented in Appendix A, Table A1) are taken from Ahmad et al. [13,18], where alkali (KOH) is used as a catalyst. The production of biodiesel from flaxseed oil followed by two steps transesterification reaction. The flaxseed oil was purchased from Bale-Robe, Ethiopia. KOH, sulfuric acid (H_2SO_4) and methanol (CH_3OH) were purchased from Sigma Aldrich. To remove the moisture content, the oil was pre-heated for 30 min at 110 °C. For the preparation of various percentage of catalyst, a certain amount of KOH pellets was dissolved in the methanol. The preheated flaxseed oil was cooled in a specific temperature (30, 50 and 65 °C) and after that potassium methoxide (CH_3KO) solution was mixed. The CH_3KO solution was mixed with the pre-heated oil in a 250 mL three-neck glass reactor and fixed over a magnetic stirrer with constant stirring at the above specific temperature. Then, the mixture was kept at three different reaction times (30, 45 and 60 min). For proper settling, the mixture was kept overnight in a different funnel. For the purification of the biodiesel, water washing method was considered. To prevent the formation of foam, the mixture was stirred slowly. After that, the mixture was put down

overnight to settle into two different phases i.e., biodiesel phase and water impurity phase. To ensure the elimination of contaminates from the biodiesel, the above process was carried out for three times. After that, the biodiesel was heated for 1 h at 100 °C for the evaporation of residual water. For quantifying the biodiesel titration of biodiesel, fraction was used with 0.1 N sulphuric acid [19]. For the physicochemical characterization of flaxseed oil and biodiesel, standard procedures were considered, and it has been explained by Ahmad et al. [13,18]. For finding the fatty acid constituents, flaxseed oil was analysed, and the complete procedures have been discussed by Ahmad et al. [13,18]. In the present case, central composite design (CCD) has been considered for the design of experiments. Two different ways are there to perform CCD; one is face-centred central composite design (FCCD) and the other is rotatable central composite design (RCCD). Here, FCCD is used to find the influence of different process parameters for the maximum biodiesel yield. For the design of the experiment, two levels and four factors with five different centred point values have been considered. A total of 29 experiments have been performed by considering the FCCD. The process parameter chosen for the experiments is reaction time, methanol/oil molar ratio (M_r), reaction temperature and catalyst weight percentage, and this has been discussed by Ahmad et al. [13,18].

3.1.1. Effect of Process Parameters

The training data based on the experiments conducted by Ahmad et al. [13,18] are shown in the form of scatter plots in Figure 1 and in Table A1. Since a central composite design was used for the experimental design, the process parameters for biodiesel production are seen to be distributed primarily across three levels in each case. The yield versus methanol/oil molar ratio (M_r) scatter plot indicates that the yield is better at higher M_r. However, from the other three scatter plots of yield versus process parameters, it is non-conclusive of the effect of various levels of the process parameters. The histogram of the yield indicates that the median of all the experimentally calculated yields is around 96%.

The correlation between the various process parameters and the yield is studied in Figure 2. There is a negligible correlation among the process parameters, which confirms the lack of multicollinearity. Multicollinearity, if present, could lower the statistical power of the regressions. The methanol/oil molar ratio (M_r) shows a moderately strong positive correlation (0.68) to yield. The yield is also seen to have a negligible correlation with catalyst weight and reaction time. However, as seen from Figure 2, a very weak positive correlation (0.19) exists between the temperature and yield. This is indicative that low levels of temperature, on an average, will not lead towards very high biodiesel yield. Thus, the analysis of the experimental data on biodiesel yield, as shown in Figures 1 and 2, suggests that a brute force approach to find the best process parameter combination to maximize the yield will be extremely cost intensive. Due to the lack of clear trend in the scatter plots in Figure 2, an exhaustive experimental analysis of all possible process parameter combinations may be needed. This justifies the need for advanced machine learning algorithms to build predictive models to quantify the yield as function of the process parameters.

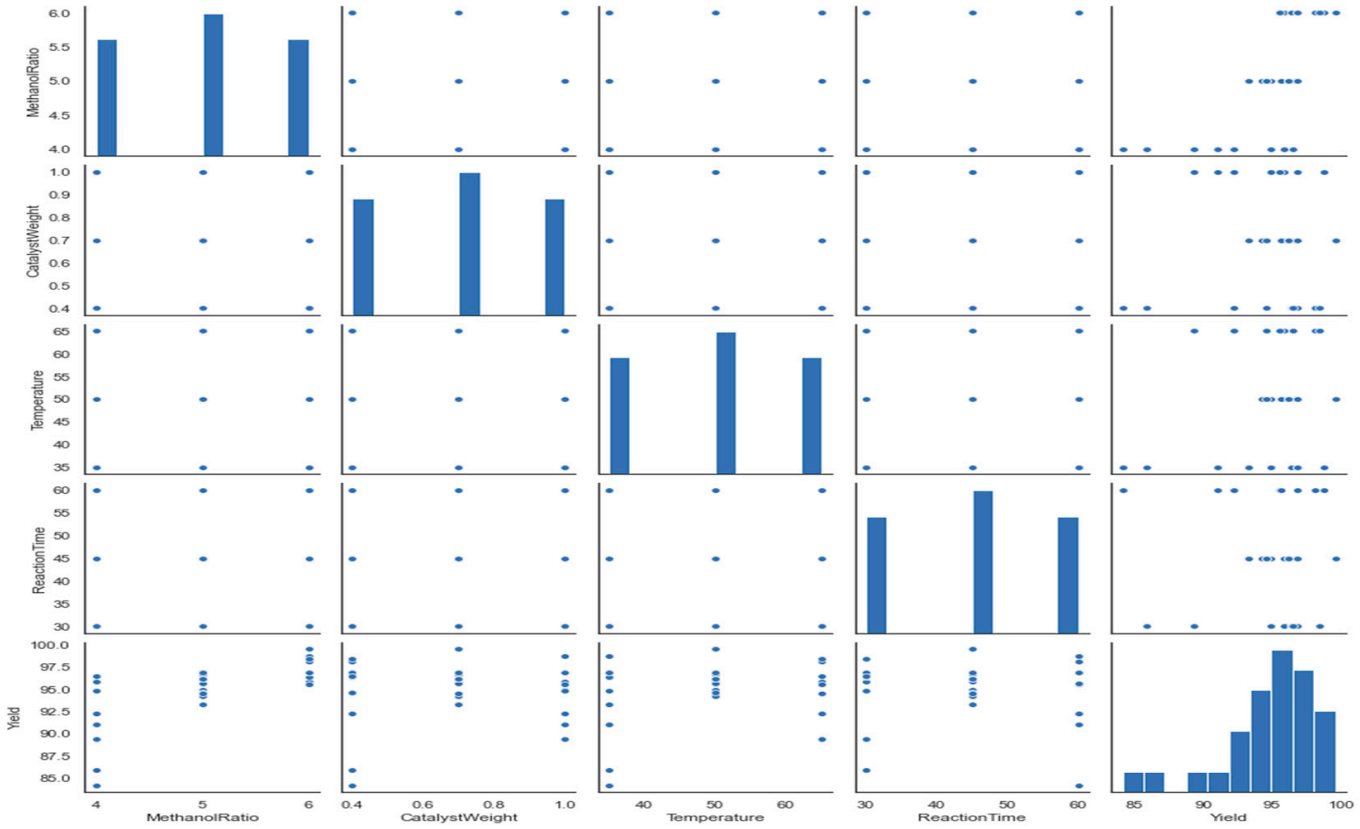

Figure 1. Scatter plot of the various biodiesel production process parameters and yield.

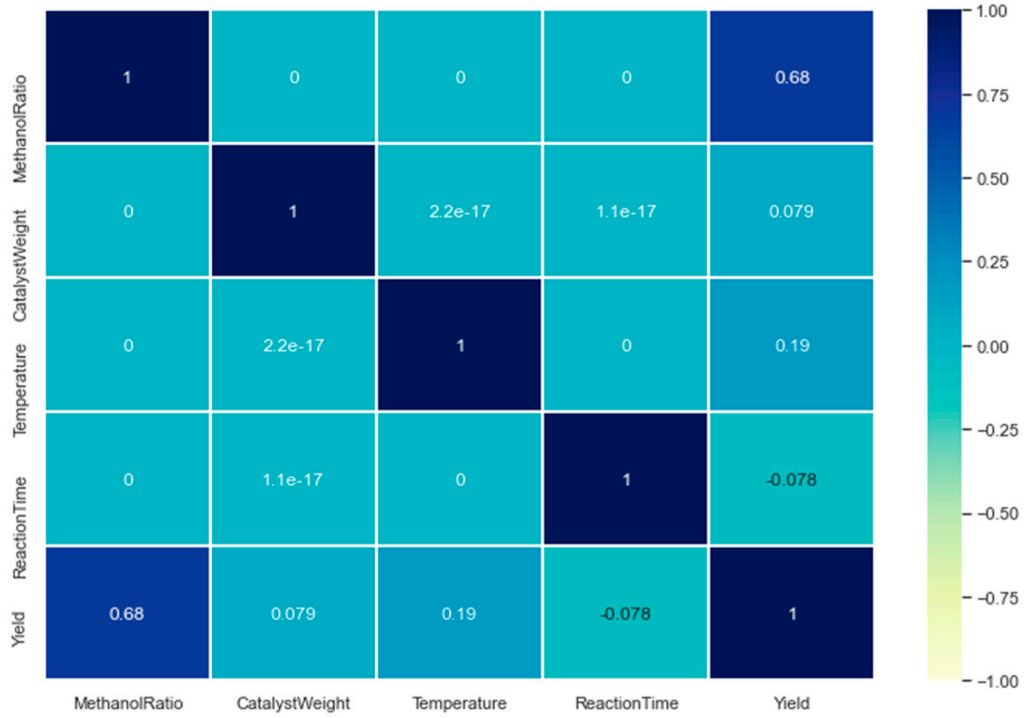

Figure 2. Correlation between the various biodiesel production process parameters and yield.

3.1.2. Linear Regression Predictive Model

Based on the training data presented in Table A1, a linear regression predictive model of the yield % is developed. An empirical relation describing the yield % as a function of

the methanol/oil molar ratio (M_r), catalyst weight (W_c), temperature (T) and reaction time (T_r) is developed as shown in Equation (14).

$$Yield\ \% = \beta_0 + \beta_1 M_r + \beta_2 W_c + \beta_3 T + \beta_4 T_r \quad (14)$$

Figure 3 shows the predictive performance of the linear regression predictive model. The actual yield % is plotted against linear regression predictive model-based predicted yield % in Figure 3a. The data points lying above the identity line indicate that the linear regression predictive model has overpredicted the yield %, while data points lying below the identity line indicate underprediction. The best line fit for the data points is generated and the R^2 is calculated which indicates the goodness of fit of the predictive model on the training data. It is seen that the predictive power of the linear regression predictive model is moderate, with an R^2 of about 52%. The performance of the linear regression predictive model is further analysed by using a predicted versus residuals plot in Figure 3b. It is seen that at low predicted values, the residuals are highest.

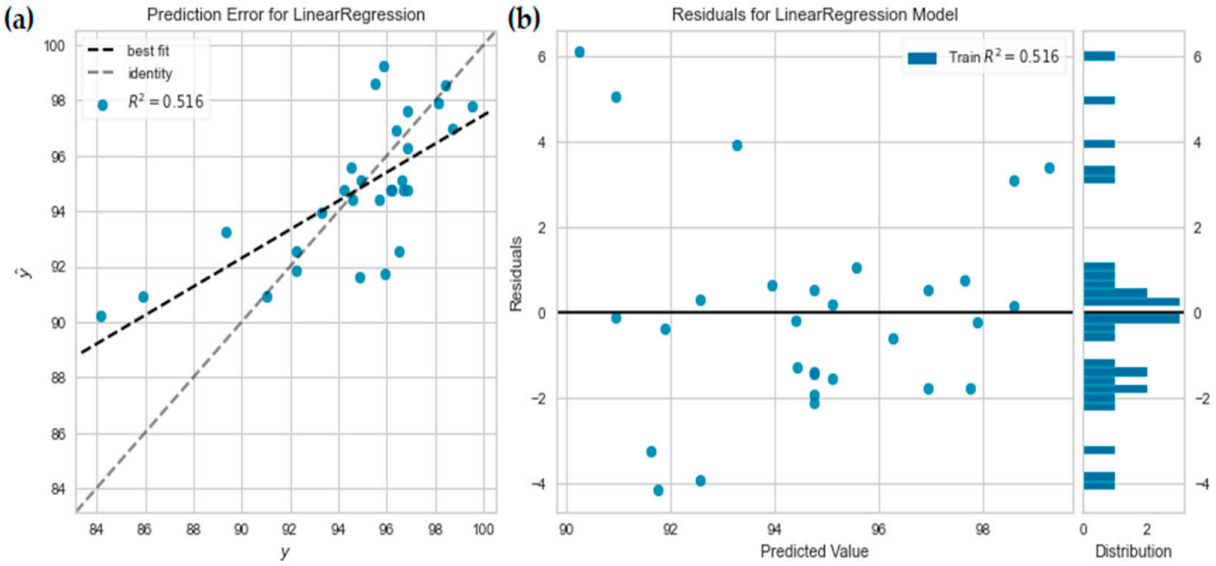

Figure 3. Predictive performance of the linear regression predictive model in biodiesel production yield estimation (**a**) R^2; (**b**) predicted value versus residuals.

3.1.3. Random Forest Regression Predictive Model

Using the training data, random forest regression predictive models are developed to quantify the yield % as a function of methanol/oil molar ratio (M_r), catalyst weight (W_c), temperature (T) and reaction time (T_r). Since the number of regressors in the random forest has a significant effect on the accuracy and performance of the predictive model, a sensitivity study on the effect of regressors is carried out. The number of regressors in the random forest approach is varied from 100 to 900 and the various error metrics are recorded. Figure 4 shows the effect of regressors on R^2, MAE, MSE, RMSE, Max. Error and MedAE. It is seen from Figure 4a that the R^2 is best for 800 regressors. Similarly, for 800 regressors, lowest MAE, MSE, RMSE and Max. Error are achieved. However, the MedAE is found to be lowest for the 900 regressors random forest predictive model.

Figure 5a shows the predictive pattern of the selected random forest predictive model. A high R^2 of approximately 89% is achieved with 800 regressors random forest predictive model. Most of the data points are seen to lie on the identity line, indicating the high correlation between actual and predicted results. The residuals plot is shown in Figure 5b, showing that the error in random forest predictive model is much lower than the linear regression predictive model.

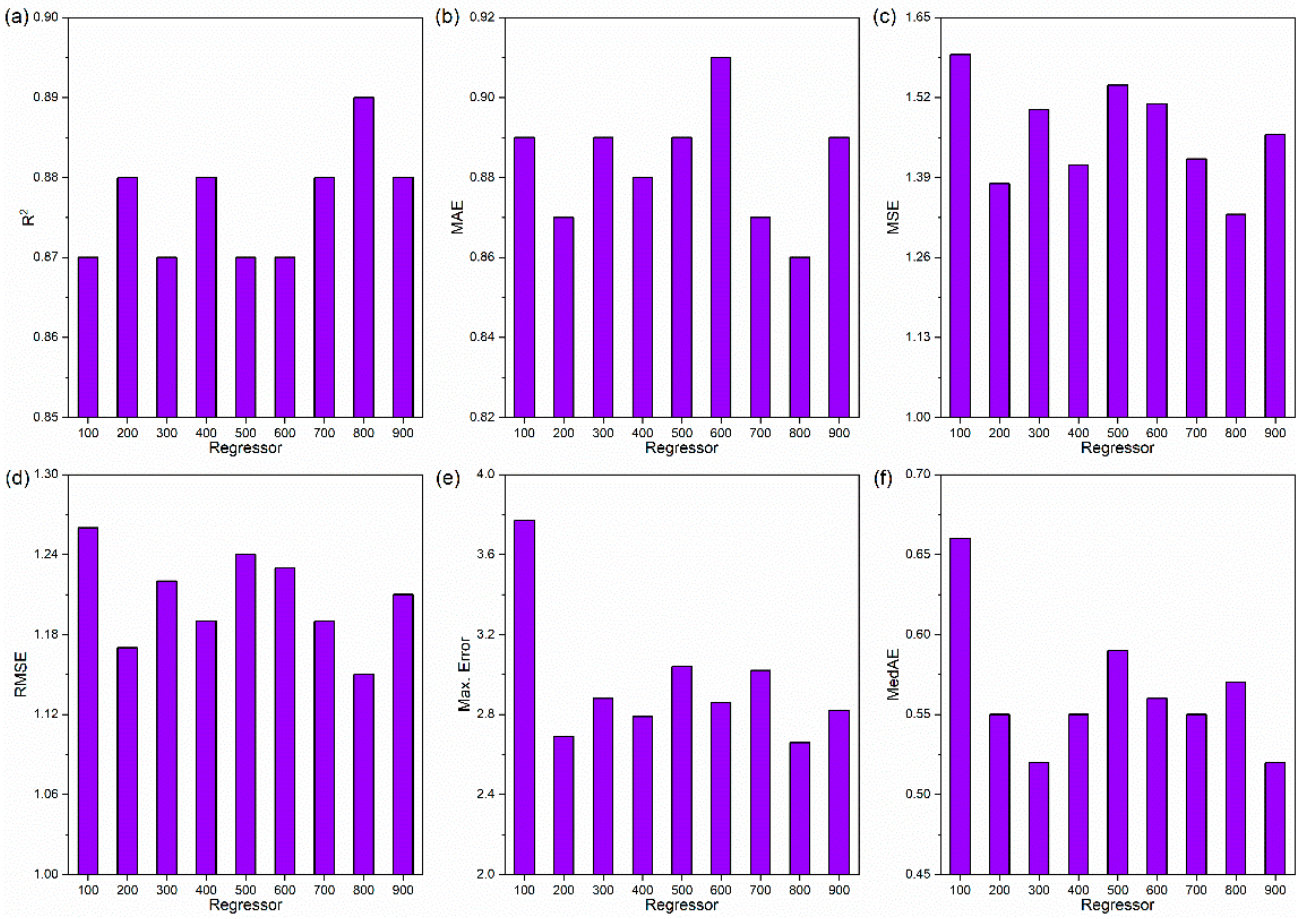

Figure 4. Effect of number of regressors in random forest regression on (**a**) R^2; (**b**) MAE; (**c**) MSE;(**d**) RMSE; (**e**) Max. Error; (**f**) MedAE. R^2: coefficient of determination; MAE: mean absolute error; MSE: mean squared error; RMSE: root mean squared error; Max. Error: maximum error; MedAE: median error.

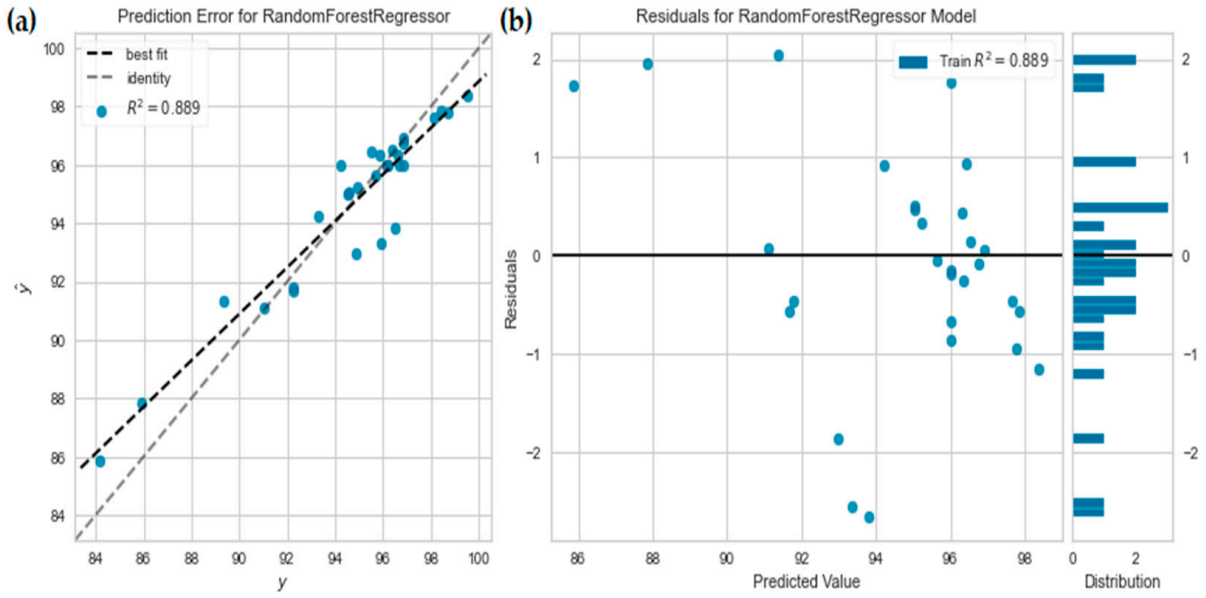

Figure 5. Predictive performance of the random forest regression predictive model in biodiesel production yield estimation (**a**) R^2; (**b**) predicted value versus residuals.

3.1.4. AdaBoost Predictive Model

AdaBoost regression predictive model is developed using 100 regressors. A sensitivity study shows that the number of regressors has a negligible impact on the predictive power of the AdaBoost regression predictive model. Figure 6a shows that the R^2 of the AdaBoost regression predictive model is approximately 91%, indicating that the AdaBoost regression predictive model can successfully model 91% variance in the training data. The residuals plot in Figure 6b shows almost a similar pattern in negative and positive residuals, indicating a balanced prediction for most of the cases, without much underprediction or overprediction.

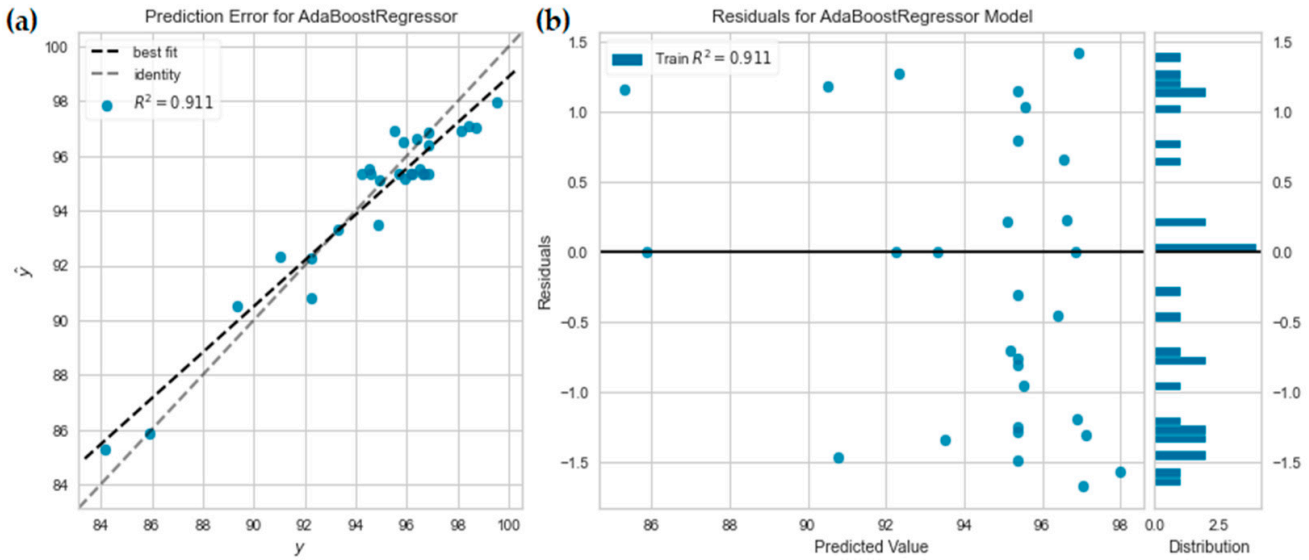

Figure 6. Predictive performance of the AdaBoost regression predictive model in biodiesel production yield estimation (**a**) R^2; (**b**) predicted value versus residuals.

3.1.5. Comparison of Various ML Predictive Models

The comparison of the machine learning predictive models in Figure 7 shows that the AdaBoost predictive model performs a little better than the random forest regression and comprehensively outperforms the linear regression predictive model. The performance of the machine learning predictive models on various error metrics like MSE, MAE, Max. Error and MedAE are shown in Figure 7b. It is observed that the MSE and the MAE are lowest for AdaBoost regression predictive model. However, Max. Error and MedAE are least for the random forest regression predictive model. Thus, overall, the performance of random forest regression predictive model and AdaBoost regression predictive model is seen to be on par with each other. However, AdaBoost regression predictive model has a distinct advantage in this study over random forest regression predictive model that it is almost insensitive to the number of regressors.

3.2. Example 2: Biodiesel FFA Conversion Percentage Estimation

The experimental data (presented in Appendix A, Table A2) are taken from Karmakar et al. [20], where castor oil of commercial grade has been considered as a raw material to produce biodiesel. To remove the moisture and impurities, the oil was dried by using the hot air within the temperature range of 100–110 °C. Propan-2-ol, sulphuric acid, methanol and potassium hydroxide were considered as the solvent. To get the deionized water, Arium 611 DI ultra-pure water system of Sartorius A. G was used. By using ASTM standardized experiments, castor oil was characterized. ASTM Method D974 was used to estimate the acid value of castor. In the present case, L16 orthogonal array was used to experiment. The esterification of castor oil was done either by using a homogeneous acid catalyst (H_2SO_4) or

heterogeneous sulfonated carbon catalyst. During esterification, one neck was fitted to the thermometer to measure the temperature. To decrease the evaporative loss of the methanol, the other neck was fixed with the reflux condenser. To add the reagent the middle neck was used. Then, the methanol was moved to the reactor after heating the oil up to the set temperature. Due to the transfer of methanol, the temperature within the reactor dropped. The adjustment of plate temperature and mixture temperature was done with the help of magnetic stirrer hotplate and the temperature was fixed according to the experimental design. The acid catalyst was added after getting the desired temperature and then the reaction was allowed for the fixed time duration for experimental design. The Whatman series 40 filter paper is used for the removal of sulfonated carbon catalyst. The distillation process is used to retrieve the unused methanol. After that, the mixture was kept for 10 h to separate it into two different layers. The top layer of the mixture is the FAME rich phase while the bottom layer of the mixture is an aqueous phase, which was drained out. Then, the FAME was washed several times in the deionized water to remove acid catalyst and residual alcohol. To purify the crude biodiesel according to international standard, refining was done.

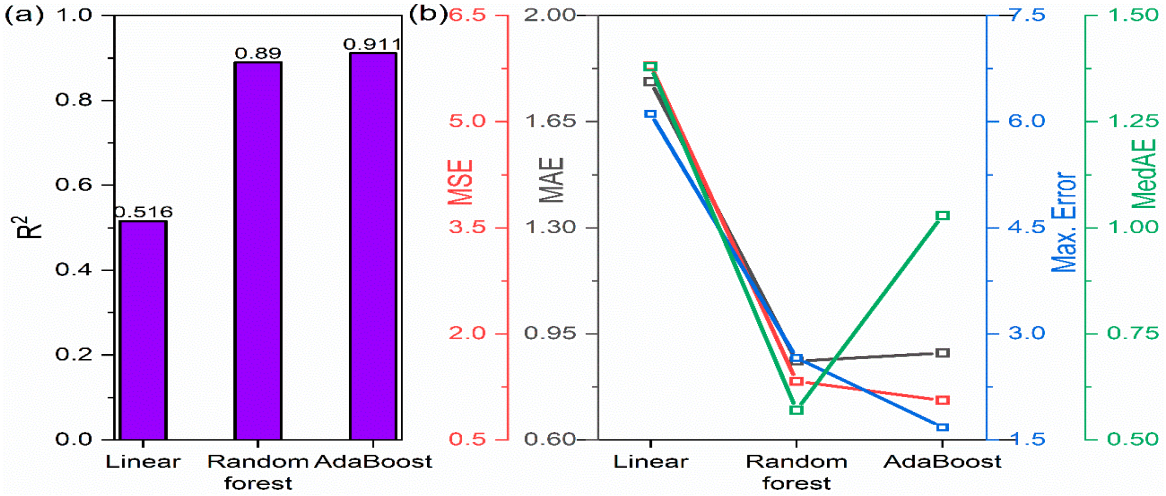

Figure 7. Performance of the machine learning predictive models in biodiesel production yield estimation (**a**) R^2; (**b**) MSE, MAE, Max. Error and MedAE.

3.2.1. Effect of Process Parameters

The training data presented in Table A2 are from the experiments conducted by Karmakar et al. [20] and are shown in the form of scatter plots in Figure 8. Since an L16 array was used for experimental design, the process parameters are considered at four equally spaced levels. The scatter of the response (i.e., FFA conversion percentage) is seen to be highly dependent on the levels of the process parameters. To further understand the effect of the process parameters on the biodiesel FFA conversion percentage, a correlation study is carried out as shown in Figure 9. It is seen that the process parameters do not correlate themselves, thus, multicollinearity is avoided. Correlation plot among the response and the process parameters show that methanol/oil molar ratio (M_r) has a moderately high correlation with the response. Agitation speed is seen to have a moderately low effect on the biodiesel FFA conversion percentage.

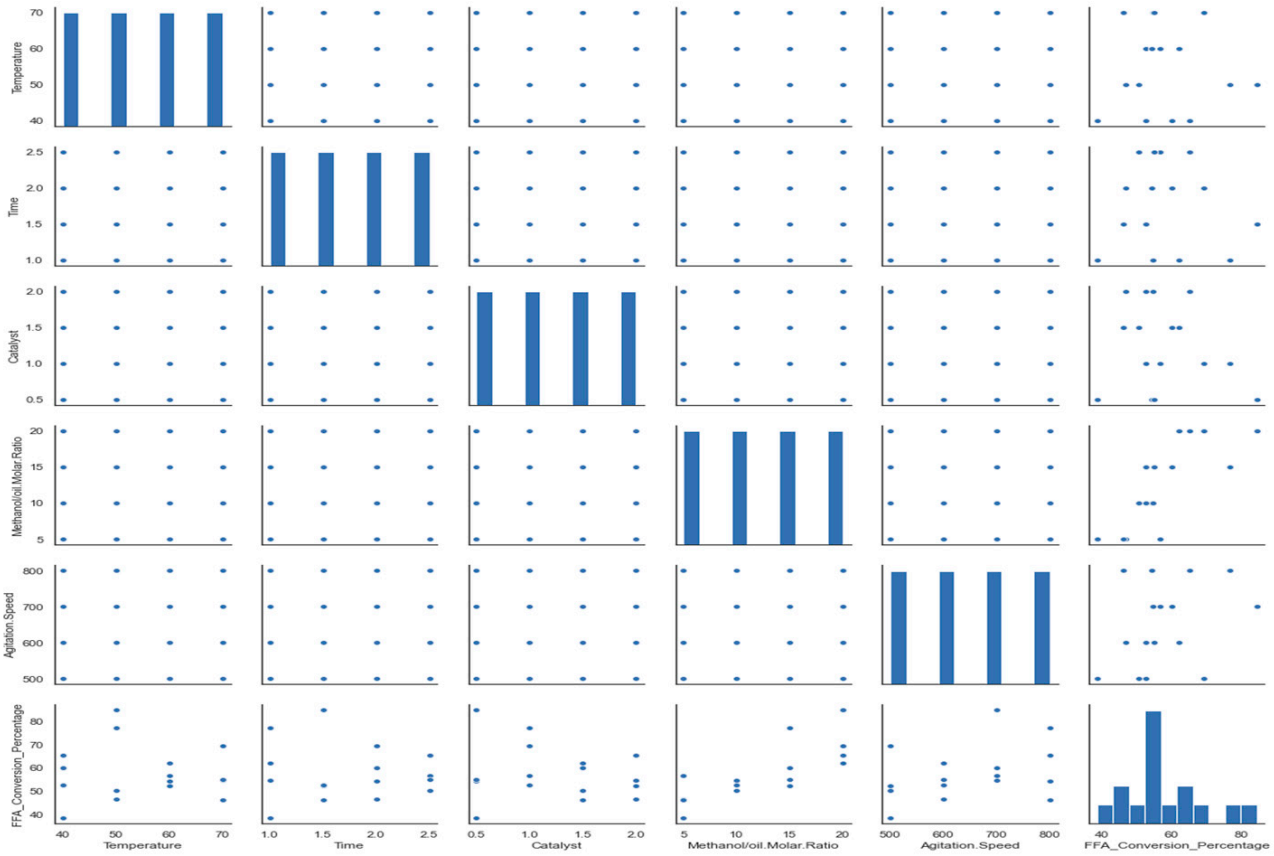

Figure 8. Scatter plot of the various biodiesel production process parameters and yield.

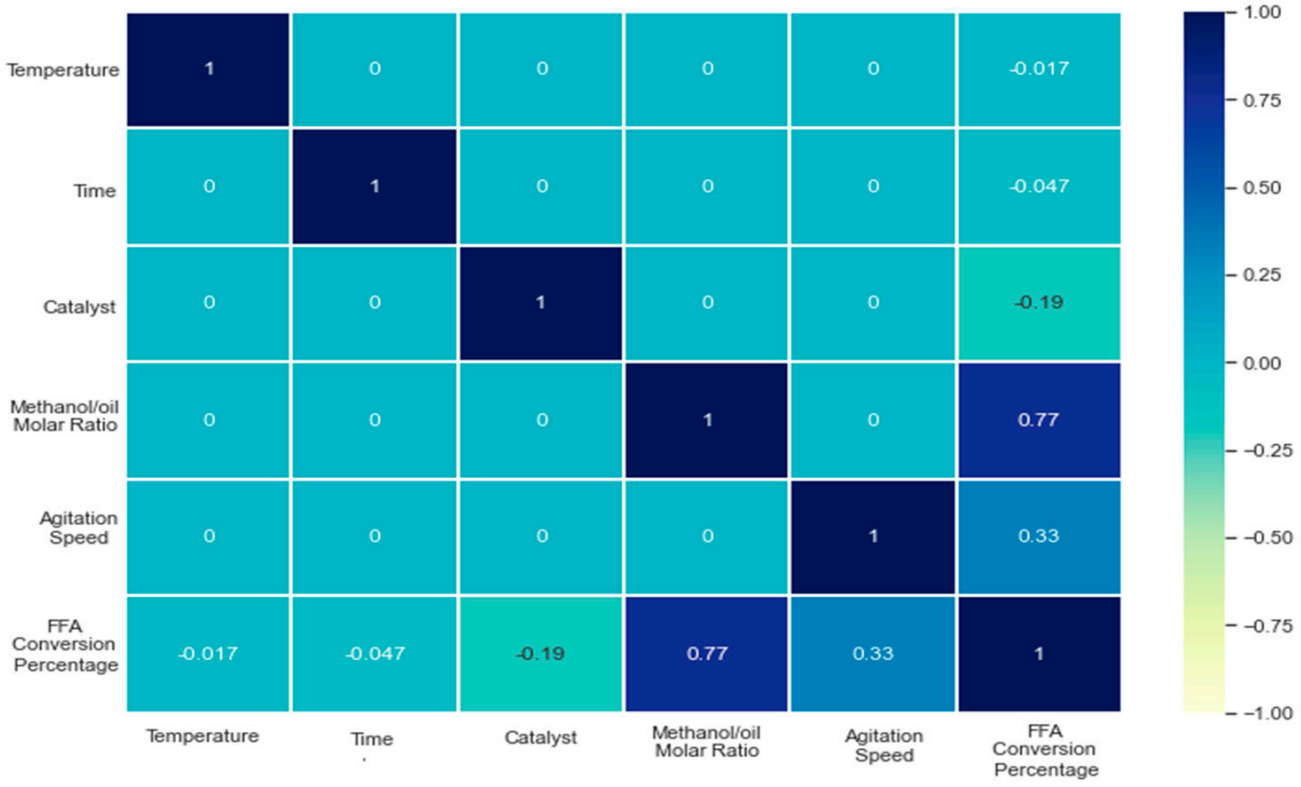

Figure 9. Correlation between the various biodiesel production process parameters and yield.

3.2.2. Linear Regression Predictive Model

Based on the training data listed in Table A2 from Karmakar et al. [20], a linear polynomial regression model of the following form is developed,

$$FFA\ conv.\% = 23.62 - 0.0175T - 0.9485T_r - 3.8585W_c + 1.5635M_r + 0.03407S_a \quad (15)$$

Figure 10a shows the scatter of the predicted versus the actual biodiesel FFA conversion percentage. The scatter of the residuals versus the predicted biodiesel FFA conversion percentage is shown in Figure 10b. It is seen that the model achieves moderately high accuracy in terms of R^2. The scatter of the residuals is random, indicating no ties in the data. Moreover, the random underprediction and overprediction i.e., data points lying randomly under and over the identity line, indicate that the model is not biased in a certain direction.

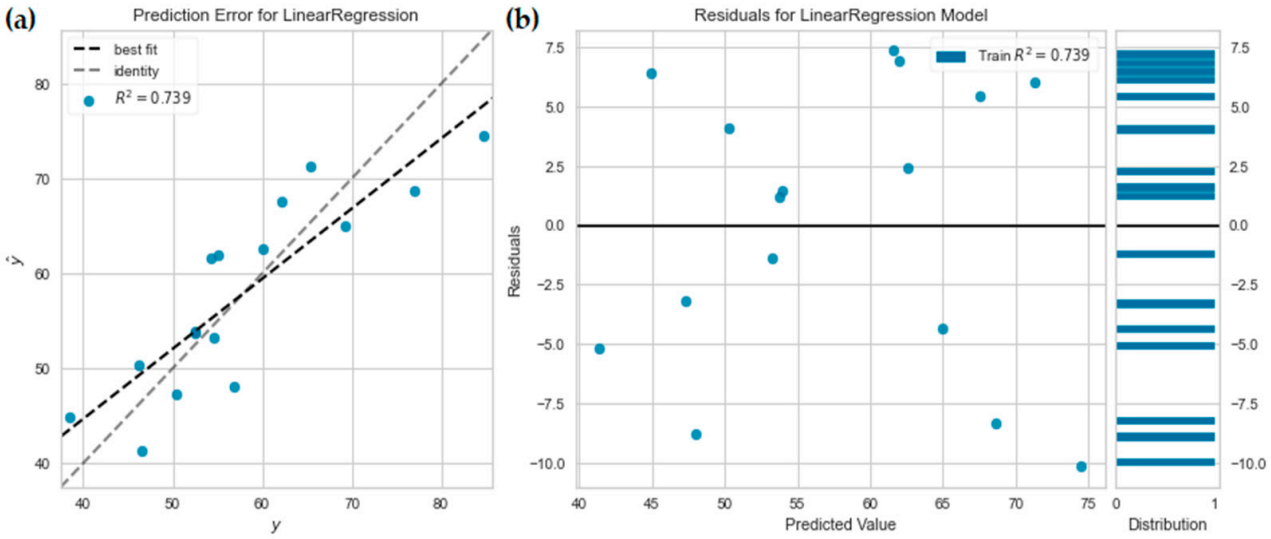

Figure 10. Predictive performance of the linear regression predictive model in biodiesel production yield estimation (**a**) R^2; (**b**) predicted value versus residuals.

3.2.3. Random Forest Regression Predictive Model

As established in Section 3.1.3, random forest regression is sensitive to the regressors. Thus, a pilot study is carried out to determine the optimal number of regressors, which in this case is 500. The predicted biodiesel FFA conversion percentage versus the actual one in Figure 11a shows that high accuracy is obtained by the random forest regression predictive model. However, Figure 11b shows that only baring a couple of datapoints, most other residues are within $\pm$ 4. This indicates that the three remaining outliers in Figure 11b are responsible for the loss in the prediction power of the predictive model.

3.2.4. AdaBoost Predictive Model

Before developing predictive models with AdaBoost, a sensitivity study is undertaken on the number of regressors which indicate a negligible change in the predictive power of AdaBoost Regression predictive model to the number of regressors. Figure 12a shows the performance of the AdaBoost regression predictive model. It is seen that the predictive model has almost ideal estimation. The residuals versus the predicted biodiesel FFA conversion percentage in Figure 12b reveals that the maximum residue is $\pm$2. Moreover, baring a couple of data points, the remaining residuals are within $\pm$0.5.

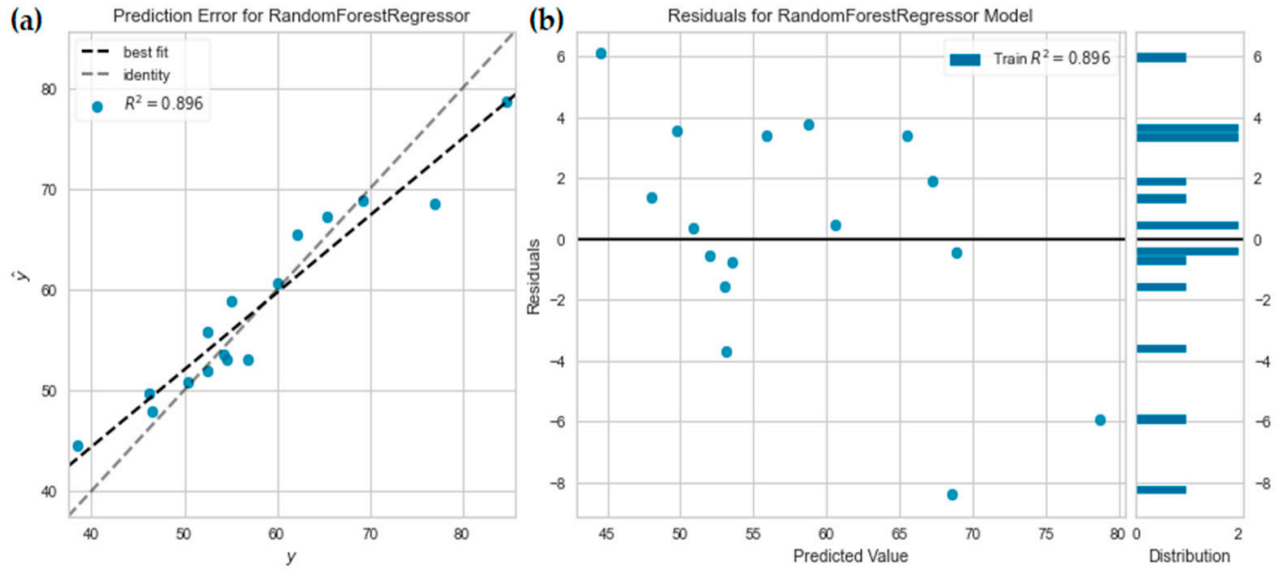

Figure 11. Predictive performance of the random forest regression predictive model in biodiesel production yield estimation (**a**) R^2; (**b**) Predicted value versus residuals.

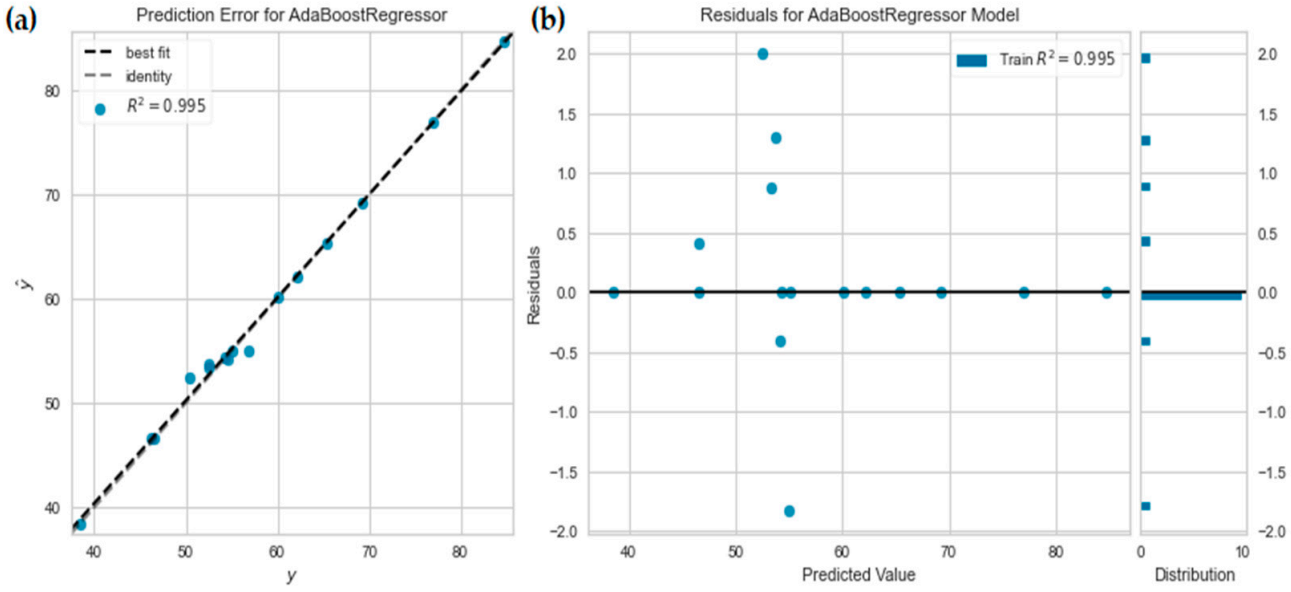

Figure 12. Predictive performance of the AdaBoost regression predictive model in biodiesel production yield estimation (**a**) R^2; (**b**) Predicted value versus residuals.

3.2.5. Comparison of Various ML Predictive Models

The comparison of the R^2 for the three predictive models in Figure 13a reveals that the AdaBoost regression predictive model has the best estimation power. The MSE of the AdaBoost regression predictive model is almost zero, whereas for random forest regression predictive model, it is around 13. Comparison of the predictive models based on all the error metrics, as shown in Figure 13b indicates that AdaBoost regression predictive model is comprehensively superior as compared to random forest regression predictive model and linear regression predictive model.

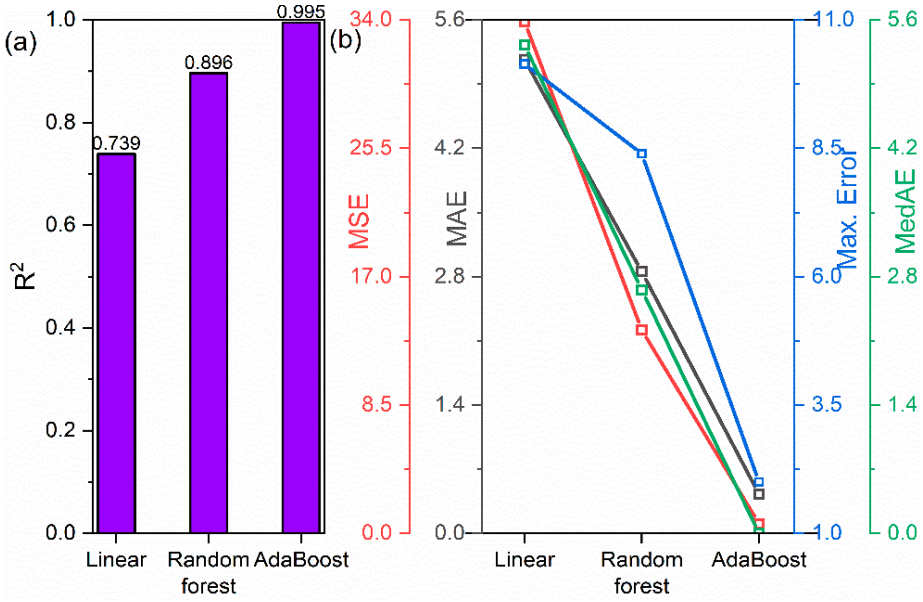

Figure 13. Performance of the machine learning predictive models on (**a**) R^2; (**b**) MSE, MAE, Max. Error and MedAE.

4. Conclusions

Building predictive models for the quantification of biodiesel yield based on the level of the process parameters is a realistic goal with a tremendous impact on sustainable development. In this paper, in one illustrative example, the biodiesel yield is modelled as a function of methanol/oil molar ratio (M_r), catalyst weight (W_c), reaction temperature (T) and reaction time (T_r). Similarly, in another illustrative example, the free fatty acid conversion percentage is estimated based on reaction temperature (T) and reaction time (T_r), catalyst weight (W_c), methanol/oil molar ratio (M_r) and agitation speed (S_a). A comprehensive comparative analysis is carried out to ascertain the utility of three machine learning techniques (linear regression, random forest regression and AdaBoost regression) for developing predictive models in biodiesel production. A wide range of accuracy and error metrics is used to quantify the efficacy of the machine learning algorithms. Based on the analysis it is seen that the linear regression approach is only able to achieve moderate accuracy, whereas both random forest regression and AdaBoost regression show very high accuracy in predictive modelling of the biodiesel yield. However, a sensitivity study on the effect of the regressors on the predictive performance of random forest regression and AdaBoost regression show that while AdaBoost may be non-sensitive to the number of regressors, the random forest may be significantly affected by the change in the number of regressors. Thus, AdaBoost regression may be preferred for the predictive modelling of biodiesel yield. This would lead to a significant saving in time and effort in identifying the optimum process parameters to increase the yield or FFA conversion % of biodiesel production process.

Author Contributions: Conceptualization, methodology, software, writing—review and editing, K.K., R.K.G., X.-Z.G.; validation, formal analysis, investigation, K.K.G. and M.R.; data curation, writing, original draft preparation, visualization, K.K.G., K.K., R.K.G., M.R. K.K.G. and K.K. contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in the article.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

The dataset used in Sections 3.1 and 3.2 is presented in Tables A1 and A2 respectively. The experimental data presented on biodiesel production process presented in Tables A1 and A2 are taken from [18] and [20] respectively.

Table A1. Biodiesel production process parameters and experimental yield % [18].

Methanol/Oil Molar Ratio (M_r)	Catalyst Weight (W_c)	Reaction Temperature (T)	Reaction Time (T_r)	Yield (%)
5	0.7	35	45	93.3
6	1	65	30	95.88
5	0.7	50	30	96.62
5	1	50	45	94.9
6	0.4	35	30	96.4
4	1	65	60	92.26
6	0.4	35	60	96.84
5	0.7	50	45	94.22
4	1	35	60	91.04
4	0.4	35	60	84.14
4	0.7	50	45	95.9
6	0.4	65	60	98.1
6	1	35	30	96.86
6	1	35	60	98.72
5	0.7	50	45	96.66
6	1	65	60	95.5
5	0.4	50	45	94.58
6	0.7	50	45	99.54
5	0.7	50	60	95.68
4	1	35	30	94.86
6	0.4	65	30	98.41
4	0.4	35	30	85.88
5	0.7	50	45	96.86
4	0.4	65	30	96.48
5	0.7	65	45	94.52
4	0.4	65	60	92.26
5	0.7	50	45	96.14
4	1	65	30	89.32
5	0.7	50	45	96.18

Table A2. Biodiesel production process parameters and experimental FFA conversion % [20].

Reaction Temperature (T)	Reaction Time (T_r)	Catalyst Weight (W_c)	Methanol/Oil Molar Ratio (M_r)	Agitation Speed (S_a)	FFA Conversion %
50	2	2	5	600	46.6
50	1.5	0.5	20	700	84.66
70	2	1	20	500	69.27
60	1.5	2	15	500	52.46
60	2.5	1	5	700	56.8
40	2.5	2	20	800	65.35
60	2	0.5	10	800	54.3
40	1	0.5	5	500	38.48
40	2	1.5	15	700	60.1
60	1	1.5	20	600	62.1
70	1.5	1.5	5	800	46.18
70	1	2	10	700	54.6
70	2.5	0.5	15	600	55.06
40	1.5	1	10	600	52.5
50	1	1	15	800	76.97
50	2.5	1.5	10	500	50.46

References

1. Bastos, R.R.C.; Corrêa, A.P.D.L.; Da Luz, P.T.S.; Filho, G.N.D.R.; Zamian, J.R.; Da Conceição, L.R.V. Optimization of biodiesel production using sulfonated carbon-based catalyst from an amazon agro-industrial waste. *Energy Convers. Manag.* **2020**, *205*, 112457. [CrossRef]
2. Junior, W.A.P.; Takeno, M.L.; Nobre, F.X.; Barros, S.D.S.; Sá, I.S.; Silva, E.P.; Manzato, L.; Iglauer, S.; De Freitas, F.A. Application of water treatment sludge as a low-cost and eco-friendly catalyst in the biodiesel production via fatty acids esterification: Process optimization. *Energy* **2020**, *213*, 118824. [CrossRef]
3. Shafiee, S.; Topal, E. When will fossil fuel reserves be diminished? *Energy Policy* **2009**, *37*, 181–189. [CrossRef]
4. Naveenkumar, R.; Baskar, G. Optimization and techno-economic analysis of biodiesel production from Calophyllum inophyllum oil using heterogeneous nanocatalyst. *Bioresour. Technol.* **2020**, *315*, 123852. [CrossRef] [PubMed]
5. Martins, E.H.; Vilela, A.P.; Mendes, R.F.; Mendes, L.M.; Vaz, L.E.V.D.S.B.; Guimarães, J.B., Jr. Soybean waste in particleboard production. *Ciência e Agrotecnologia* **2018**, *42*, 186–194. [CrossRef]
6. Lertsathapornsuk, V.; Pairintra, R.; Aryusuk, K.; Krisnangkura, K. Microwave assisted in continuous biodiesel production from waste frying palm oil and its performance in a 100 kW diesel generator. *Fuel Process. Technol.* **2008**, *89*, 1330–1336. [CrossRef]
7. Leung, D.; Guo, Y. Transesterification of neat and used frying oil: Optimization for biodiesel production. *Fuel Process. Technol.* **2006**, *87*, 883–890. [CrossRef]
8. Georgogianni, K.; Kontominas, M.; Pomonis, P.; Avlonitis, D.; Gergis, V. Conventional and in situ transesterification of sunflower seed oil for the production of biodiesel. *Fuel Process. Technol.* **2008**, *89*, 503–509. [CrossRef]
9. Guo, M.; Jiang, W.; Chen, C.; Qu, S.; Lu, J.; Yi, W.; Ding, J. Process optimization of biodiesel production from waste cooking oil by esterification of free fatty acids using La^{3+}/ZnO-TiO_2 photocatalyst. *Energy Convers. Manag.* **2021**, *229*, 113745. [CrossRef]
10. Ortega, M.F.; Donoso, D.; Bousbaa, H.; Bolonio, D.; Ballesteros, R.; García-Martínez, M.-J.; Lapuerta, M.; Canoira, L. Optimized Production of Fatty Acid Ethyl Esters (FAEE) from Waste Frying Oil by Response Surface Methodology. *Waste Biomass Valorization* **2020**, 1–8. [CrossRef]
11. Shin, H.-Y.; Lim, S.-M.; Kang, S.C.; Bae, S.-Y. Statistical optimization for biodiesel production from rapeseed oil via transesterificaion in supercritical methanol. *Fuel Process. Technol.* **2012**, *98*, 1–5. [CrossRef]
12. Hameed, B.; Lai, L.; Chin, L. Production of biodiesel from palm oil (*Elaeis guineensis*) using heterogeneous catalyst: An optimized process. *Fuel Process. Technol.* **2009**, *90*, 606–610. [CrossRef]
13. Ahmad, T.; Danish, M.; Kale, P.; Geremew, B.; Adeloju, S.B.; Nizami, M.; Ayoub, M. Optimization of process variables for biodiesel production by transesterification of flaxseed oil and produced biodiesel characterizations. *Renew. Energy* **2019**, *139*, 1272–1280. [CrossRef]
14. Jayaprabakar, J.; Dawn, S.; Ranjan, A.; Priyadharsini, P.; George, R.; Sadaf, S.; Rajha, C.R. Process optimization for biodiesel production from sheep skin and its performance, emission and combustion characterization in CI engine. *Energy* **2019**, *174*, 54–68. [CrossRef]
15. Matinja, A.I.; Zain, N.A.M.; Suhaimi, M.S.; Alhassan, A.J. Optimization of biodiesel production from palm oil mill effluent using lipase immobilized in PVA-alginate-sulfate beads. *Renew. Energy* **2019**, *135*, 1178–1185. [CrossRef]
16. Hazra, A.; Gogtay, N. Biostatistics series module 6: Correlation and linear regression. *Indian J. Dermatol.* **2016**, *61*, 593–601. [CrossRef] [PubMed]
17. Seo, D.K.; Kim, Y.H.; Eo, Y.D.; Park, W.Y.; Park, H.C. Generation of Radiometric, Phenological Normalized Image Based on Random Forest Regression for Change Detection. *Remote Sens.* **2017**, *9*, 1163. [CrossRef]
18. Ahmad, T.; Danish, M.; Kale, P.; Geremew, B.; Adeloju, S.B.; Nizami, M.; Ayoub, M. Conversion of flaxseed oil into biodiesel using KOH catalyst: Optimization and characterization dataset. *Data in Brief* **2020**, *29*, 105225. [CrossRef]
19. Kumar, R.; Tiwari, P.; Garg, S. Alkali transesterification of linseed oil for biodiesel production. *Fuel* **2013**, *104*, 553–560. [CrossRef]
20. Karmakar, B.; Dhawane, S.H.; Halder, G. Optimization of biodiesel production from castor oil by Taguchi design. *J. Environ. Chem. Eng.* **2018**, *6*, 2684–2695. [CrossRef]