Journal of Molecular Liquids 400 (2024) 124555

Contents lists available at ScienceDirect

Journal of Molecular Liquids

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

Prediction of the cold flow properties of biodiesel using the FAME distribution and Machine learning techniques

G. Díez-Valbuena [a], A. García Tuero [a,*], J. Díez [b], E. Rodríguez [a], A. Hernández Battez [a]

[a] Department of Construction and Manufacturing Engineering, University of Oviedo, Pedro Puig Adam s/n, 33203 Gijón, Spain
[b] Artificial Intelligence Center, Universidad de Oviedo, Campus de Gijón, Gijón 33203, Spain

ARTICLE INFO

Keywords:
Biodiesel
FAME
Cold Flow Properties
Machine Learning

ABSTRACT

Burning fossil fuels is a significant contributor to global warming due to CO2 emissions. To mitigate these emissions, alternative bio-based fuels, such as biodiesel, have been developed. The cold flow properties of biodiesel, including pour point (PP), cold filter plugging point (CFPP), and cloud point (CP), are crucial. Predicting these properties can aid in selecting bio-oils for biodiesel production. Machine learning techniques were utilized to reveal intricate connections between the content of fatty acid methyl esters (FAME) in biodiesel and its cold flow properties. This study created three machine learning models based on a database of over 200 biodiesel samples to predict the aforementioned cold flow properties. The models' performance was assessed using three standard regression metrics: mean absolute error, mean squared error, and coefficient of determination. The experimental results show that the optimal algorithm for PP, CFPP, and CP has an average error of 4.51 °C, 3.56 °C, and 4.17 °C, respectively. The study also investigated the significance of various biodiesel attributes in making precise predictions, revealing that the distribution of FAME and the number of double bonds in the biodiesel are crucial factors for accurate predictions.

1. Introduction

Fossil fuels currently dominate the energy market, resulting in significant CO_2 emissions [1]. While more environmentally friendly forms of energy generation, such as wind, solar or nuclear have been implemented [2–4], they are not always suitable for certain types of energy consumption. For example, road and marine transport heavily rely on conventional internal combustion engines (ICEs) as their primary propulsion systems. These engines are specifically designed and optimized to operate on fossil fuels such as gasoline or diesel and are not compatible with many alternative energy sources. Despite the ongoing development of new models of ships, cars or trucks that can integrate alternative energy sources, it is likely to be a long time before these innovations become widespread [5]. Therefore, it makes sense to explore alternative fuels that can be used in existing engines. Of the various biofuels available, ethanol and biodiesel are the most widely used [6], and the latter appears to hold the most promise for widespread adoption [7].

Biodiesel is produced from oil by a variety of methods, including pyrolysis and emulsification, but the most important and widely used method is the transesterification reaction [8]. During transesterification, the fatty acids present in the oil react with alcohol, typically methanol or ethanol, to form fatty acid methyl/ethyl esters (FAME/FAEE). The resulting biodiesel contains these FAMEs along with potential impurities, the nature of which can vary depending on the production process and the source of the oil feedstock. Based on the origin and nature of the biodiesel, it can be classified into four distinct generations [9]. "1st generation biodiesel" is derived from oils suitable for human consumption [10], "2nd generation biodiesel" is derived from non-edible oils but still requires arable land [11] and "3rd generation biodiesel" [12] and "4th generation biodiesel" [13] are derived from microalgae.

Biodiesel must meet the requirements of the EN 14214:2012 + A2:2019 [14] and ASTM D6751 [15] standards. These standards specify the acceptable range for certain properties and how they should be measured, as shown in Table 1. While most of the critical properties are covered in the standard, there are some important properties that are addressed in national legislation as they are highly dependent on the country or region. These properties are related to biodiesel behaviour at low temperatures and are commonly referred as "cold flow" properties:

- Pour Point (PP). This refers to the point at which biodiesel ceases to flow [16]. It is measured in accordance with ISO 3016:2019 [17].

* Corresponding author.
E-mail address: garciatalejandro@uniovi.es (A. García Tuero).

https://doi.org/10.1016/j.molliq.2024.124555
Received 15 January 2024; Received in revised form 23 February 2024; Accepted 21 March 2024
Available online 22 March 2024

Table 1
Range of biodiesel properties extracted from the EN 14214:2012 + A2:2019 [14] and ASTM D6751 [15] standards.

Property	Units	EN 14,214	Test method	ASTM D6751	Test method
FAME content	%(m/m)	≥96.5	EN 14,103	–	
Density at 15 °C	kg/m^3	860—900	ISO 12,185	–	
Viscosity at 40 °C	mm^2/s	3.50—5.00	EN 16,896	1.9—6	D445
Flash Point	°C	≥101	ISO 3679	≥130	D93
Cetane Number (CN)	–	≥51.0	EN 16,175	≥47	D613
Copper corrosion	Class	≥1a	ISO 2160	≤3	D130
Oxidative stability (110 °C)	h	≥ 8	EN 14,112	≥3	EN 14,112
Acid value	mg KOH/g	≤0.50	EN 14,104	≤0.5	D664
Iodine Value (IV)	g I_2/100 g	≤120	EN 14,111	–	
Methyl ester linoleic acid	%(m/m)	≤12	EN 14,103	–	
Methyl ester + 4 double bonds	%(m/m)	≤1	EN 15,779	–	
Methanol content	%(m/m)	≤0.2	EN 14,110	≤0.2	EN 14,110
Monoglycerides content	%(m/m)	≤0.7	EN 14,105	–	
Diglycerides content	%(m/m)	≤0.2	EN 14,105	–	
Triglycerides content	%(m/m)	≤0.2	EN 14,105	–	
Free Glycerine	%(m/m)	≤0.02	EN 14,105	≤0.02	D6584
Total Glycerine	%(m/m)	≤0.25	EN 14,105	≤0.24	D6584
Water content	%(m/m)	≤0.050	ISO 12,937	≤0.05	D2709
Total contamination	mg/kg	≤24	EN 12,662	–	
Sulphated ash content	%(m/m)	≤0.02	ISO 3987	≤0.02	D874
Sulphur content	mg/kg-%(m/m)	≤10	ISO 20,846	≤15	D5453
Group I metals (Na + K)	mg/kg	≤5	EN 14,108	≤5	EN 14,538
Group II metals (Ca + Mg)	mg/kg	≤5	EN 14,538	≤5	EN 14,538
Phosphorus content	mg/kg-%(m/m)	≤4	EN 14,107	≤10	D4951

- Cold Filter Plugging Point (CFPP). This refers to the highest temperature at which a 20 mL sample of fuel fails to pass through a specified filter within 60 s under certain test conditions, including a specified pressure and cooling temperature [16]. It is measured in accordance with EN 116:2015 [18].
- Cloud Point (CP) refers to the temperature at which the heavier components of the biodiesel begin to crystallize and the solution becomes cloudy [16]. It is measured accordance with ISO 3015:2019 [19].

These requirements will vary from country to country or region to region, as the average temperature varies. Regions with hotter conditions may be more permissive with the limits, while colder regions should be more demanding, as these properties are critical to proper engine performance [20,21].

The biodiesel characterization process, including the extraction of biodiesel from the feedstock itself and the measurement of cold flow properties, is a time-consuming and costly process. Therefore, it would be desirable to have tools to predict these properties. According to EN 14214, the product must have a FAME content of at least 96.5 % to qualify as biodiesel. Taking this into account, it seems that the prediction of properties should be made by studying the content and distribution of FAME, as these affect the rest of the biodiesel properties, and the FAME determination test is easy to perform. To date, empirical equations based on a limited number of parameters have been the predominant methods used to predict cold flow properties [22–27]. These parameters are typically a combination of the FAME distribution and other parameters such as degree of unsaturation or average chain length. These empirical equations provide a linear approach to the problem, but this approach limits their ability to model more complex cases. While they are simple and easy to use, they do not yield accurate results when tested on a diverse set of biodiesels, due to their inability to model the interactions between FAMEs. Therefore, to show the improvement that this study represents, the results will be compared to those of Bolonio et al., Sarin et al., Dunn et al., Alviso et al. and Serrano et al. [28–32]. However, with the advent of artificial intelligence, we are no longer limited to using only a few parameters, but can use the entire FAME distribution along with other data to make the prediction.

This study aims to develop three different models for predicting the pour point (PP), the cold filter plugging point (CFPP) and the cloud point (CP) of biodiesel from its FAME distribution using an extensive database of more than 200 biodiesels. Improved performance evaluation and metrics to ensure generalization will be presented and an analysis of the relevance of each attribute will be conducted. Once developed, these models can help in the early stages of the production process, in which having a tool for screening the prospective bio-oils will be helpful. Currently, these predictive models cannot replace physical testing, as they are not covered by existing legislation. However, as these models improve, they could be used in place of physical characterization. In this paper several machine learning algorithms and performance methods are applied to build models for predicting biodiesel properties useful to the industry.

2. Material and methods

This section describes the database compiled from the existing literature, as well as several machine learning algorithms suitable for prediction. Since each of the cold flow properties is a continuous variable, the selection should focus on regression algorithms capable of predicting numerical values.

2.1. Databases description

A different database is presented for each cold flow property studied. The PP database contains 238 biodiesel examples, the CFPP database contains 248 biodiesel examples, and the CP database contains 282 biodiesel examples collected from the literature. Some examples are present in more than one database because the corresponding authors measured more than one property in the referenced work, resulting in a total of 372 different biodiesel examples across the 3 databases. These 372 different biodiesels come from more than 140 different feedstocks, making the biodiesel database as representative of current biodiesel trends as possible. On average, 68 % of the samples belonged to the 1st generation, 27 % to the 2nd generation and 5 % to the 3rd generation. Other studies addressing the same prediction problem, such as those presented by Alviso et.al [31] (48 biodiesel examples) and Razavi et.al [33] have used smaller datasets (48 and 44 biodiesel examples, respectively), which can lead to a generalization problem [34], developing models that would struggle to make accurate predictions for samples other than those used in the training set.

The datasets presented in this paper consist of 38 identified FAMEs. In addition, the group “others” includes the FAMEs that could not be identified. As shown in Fig. 1, the predominant FAMEs in the database are: methyl palmitate (C16:0), methyl stearate (C18:0), methyl oleate

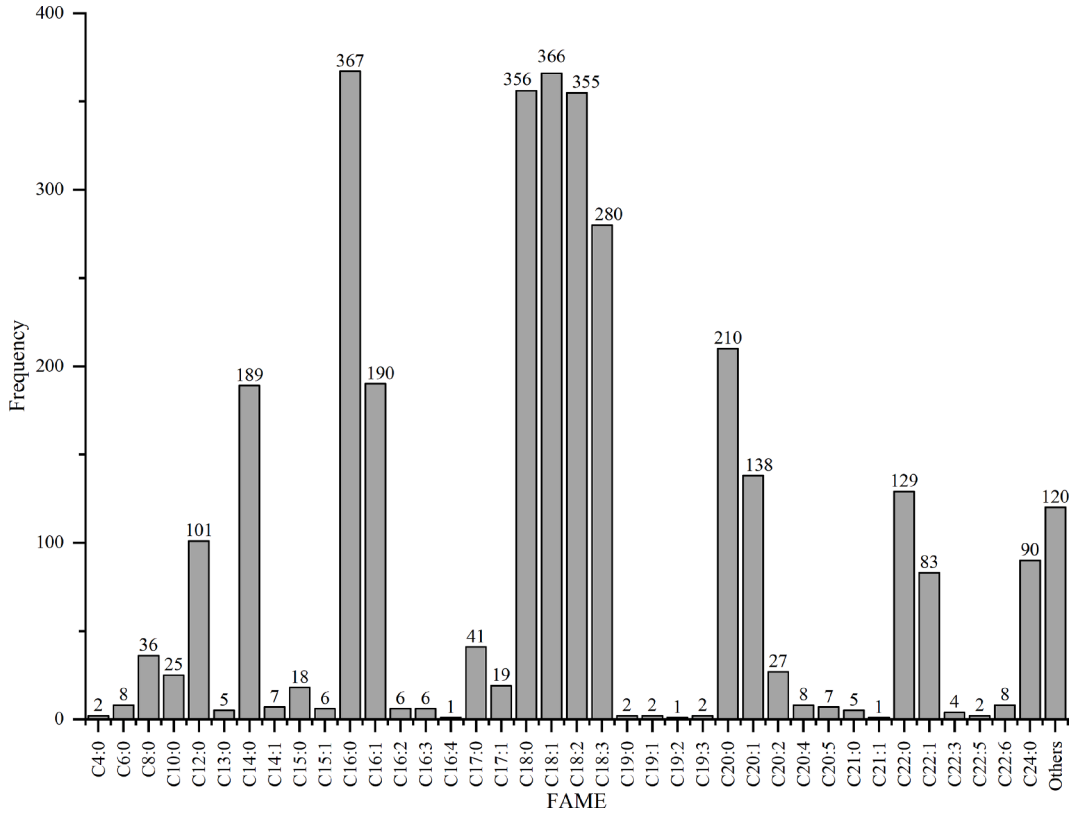

Fig. 1. Frequency distribution of the FAMEs presented in the current set.

(C18:1) and methyl linoleate (C18:2), which are present in over 95 % of the samples. Other FAMEs with a high abundance rate are: methyl myristate (C14:0), methyl palmitoleate (C16:1), methyl linolenate (C18:3), methyl arachidate (C20:0), methyl eicosenoate (C20:1) and methyl behenate (C22:0), whose presence ranged from 34 % to 76 %. This is consistent with the findings of Hoekman et al. [35] and Singh et al. [36], who showed that the majority of biodiesel is composed of C16 and C18 compounds.

2.2. Prediction methods

When approaching a prediction problem there are many algorithms capable of performing the task. This section presents a selection of the most representative regression algorithms. All of these methods are available in the Python library Scikit-Learn [37].

The *Mean Predictor* is used as a baseline for comparison. This algorithm predicts the mean value of the data and is the simplest algorithm that can be implemented. Another simple algorithm is the *Linear Regression*, which gives good (and interpretable) results when the problem can be solved linearly, but whose performance is poor when a nonlinear solution is required. In the latter case, algorithms that can adapt to nonlinear problems, such as *Decision Trees*, are desirable. This algorithm uses successive splits to group the data into clusters that have the same prediction and stores the learned model in a tree that can be easily interpreted [38].

Even if simplicity is preferred, it is likely that complex problems cannot be modelled by simpler algorithms. There are more sophisticated algorithms that can be useful for these complex problems, but they must be used carefully because they tend to overfit (the learned model can make accurate predictions using the training data, but cannot generalize sufficiently). Ensemble methods such as *Random Forest* [39] and *XGBoost* [40] have shown good performance on difficult problems. These algorithms train multiple decision trees using different methods and build a model by assembling the trained decision trees [41,42]. Other complex algorithms are the *Support Vector Machines (SVM),* which have the ability to model complex problems using different kernels [43] or *Artificial Neural Networks (ANN),* which can model any type of problem through a layered architecture [44].

2.2.1. Normalization of the data

Some algorithms have difficulty during the training phase when the data is in different scales. To overcome this, it is usually necessary to standardize the data. The chosen standardization in this research is shown in **Eq.** (1), where z_i is the value after the standardization, $\bar{x}$ is the mean of the attribute,x_i are the values observed in the attribute, and σ_i is the standard deviation of these values.

$$z_i = \frac{x_i - \bar{x}}{\sigma_i} \tag{1}$$

2.2.2. Performance of the model

A leave-one-out approach was chosen to evaluate the performance of the different models. The performance obtained from a leave-one-out experiment is the most accurate estimate of the models performance. The model is trained on all the available data except one example and then is asked to predict the value of that example. This process is repeated as many times as there are examples, omitting a different example each time. Also, some algorithms tend to be influenced by the hyperparameters chosen to train the model. To find the best hyperparameters for each algorithm, a grid search using a 5-fold cross-validation [45] was performed [46].

2.3. Performance metrics

In order to measure the performance of the different algorithms presented in this study, along with other models available in the literature, three common metrics have been used. These metrics are the Mean Absolute Error (MAE), the Mean Squared Error (MSE) and the

Coefficient of Determination (R^2), represented by the equations (2), (3) and (4) respectively:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4)$$

where n is the number of examples, y_i is the actual value of the target variable, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean of the observed data.

In summary, we use the Python Scikit-Learn library [37] for the experiments, where the algorithms selected for the comparison are available under the following names: *DummyRegressor* (Mean Predictor), *LinearRegressor*, *DecisionTreeRegressor*, *RandomForestRegressor*, *XGBRegressor*, *SVR* (Support Vector Machines) and *MLPRegressor* (Artificial Neural Networks). The performance of these algorithms is calculated using a leave-one-out experiment and the hyperparameter optimization using a grid search with 5-fold cross-validation.

3. Attribute determination and data filtering

3.1. Attribute determination

Building a model that could predict the various cold flow properties using only the FAME distributions would be ideal. However, as other authors have pointed out, there are some attributes that can be calculated from the FAME distribution that can help to build the model. These attributes found in the literature are:

- The attributes related to the degree of saturation of the FAME. These are the saturated fatty acids (SFA), the monounsaturated fatty acids (MUFA), and the polyunsaturated fatty acids (PUFA) [47,48].
- The molecular weight (Mw) of the biodiesel [49].
- The length of the biodiesel, expressed as the average number of carbons (Cn) [22].
- The average number of double bonds in the biodiesel (DB) [24].
- The non-polarity index of the biodiesel (NPI), calculated according to Eq. (5) [50].

$$NPI = \frac{Cn \bullet Mw}{\sum FAME} \quad (5)$$

3.2. Filtering the data

As mentioned in the first section, according to standard EN 14214:2012 + A2:2019 [14], biodiesel must have a minimum FAME content of 96.5 % to be available for commercialization. However, many of the research papers in this area do not aim to report a final marketable biodiesel, but to characterize possible feedstocks worthy of further research. This objective has resulted in many papers reporting FAME levels slightly below 96.5 %. It is necessary to strike a balance between working with the largest amount of data and working with data where the majority of the composition is FAME. As 96.5 % of the FAME content can be easily achieved after distillation [51], a filter of 90 % is set for the FAME content, as it is a high enough percentage to account for most of the composition. After applying this filter, the examples in each database change as follows: the 238 examples for PP become 223, the 248 examples for CFPP become 228, and the 282 examples for CP are reduced to 262.

4. Results and discussion

This section shows the best combination of attributes and their importance, together with the best algorithm to build the best model for each of the cold flow properties. In addition, a comparison between the models developed in this study and other models reported by other authors is made.

4.1. Best combination of attributes and algorithms

As mentioned in sections 2.2 and 3.1, different algorithms and attributes were tested to build the best possible model. Several experiments were performed combining different attributes, as shown in Table 2. Combination 1 uses only the FAME distribution of biodiesel and can be considered as the starting point. Combinations 2 through 7 use the FAME distribution along with other attributes calculated from the FAME distribution, and combination 7 includes all the possible attributes. Combination 8, which is included in this study because it is recommended by Mostafaei [52] to predict the PP and CP, uses the attributes SFA, MUFA, PUFA Mw and Cn. In order to select the attributes that contribute positively to the prediction and discard those that do not, all these combinations were tested for all the proposed algorithms and for the three cold flow properties.

The results of the experiments are collected in Table A.1, Table A.2 and Table A.3 in the Appendix. In these tables the *Linear Regression*, *SVM-Linear* and *SVM-Polynomic* perform poorly. This is due to the fact that these algorithms learn coefficients that are multiplied by the attribute to obtain the prediction and, as can be seen in Fig. 1, some FAMEs have a very low presence in the biodiesel of the database. These particular algorithms will perform better if there are more biodiesels containing these unusual FAMEs. As expected, the *Mean Predictor* does not perform well. However, this algorithm allows us to distinguish between those algorithms that are able to learn from the data and those that are not (any algorithm that performs worse than the *Mean Predictor*). With this in mind, Fig. 2 shows the MAE of the algorithms that outperformed the *Mean Predictor* for each attribute combination in Table 2. For the MSE and R^2 metrics, graphs with similar tendencies are obtained, which can be checked in the Appendix (Fig A.1 and Fig A.2).

Overall, the best algorithms in the three problems considered are the *XGBoost* and the *Random Forest*, followed by the *SVM-RBF*. The *ANN* did not perform as well, especially for the combination of attributes proposed by Mostafaei [52], characterized by not including the FAME distribution. The *Decision Trees* show the worst performance for the prediction of the three properties.

Looking at the performance of the best algorithms (*Random Forest* and *XGBoost*), we can see that some combinations of attributes perform

Table 2
Different attribute combinations tested for each algorithm.

	FAME Distribution	Others	SFA/ MUFA/ PUFA	Mw	Cn	DB	NPI
Combination 1	✔	✖	✖	✖	✖	✖	✖
Combination 2	✔	✔	✖	✖	✖	✖	✖
Combination 3	✔	✔	✔	✖	✖	✖	✖
Combination 4	✔	✖	✔	✖	✖	✖	✖
Combination 5	✔	✖	✔	✖	✖	✔	✔
Combination 6	✔	✖	✖	✖	✖	✔	✖
Combination 7	✔	✔	✔	✔	✔	✔	✔
Combination 8	✖	✖	✔	✔	✔	✖	✖

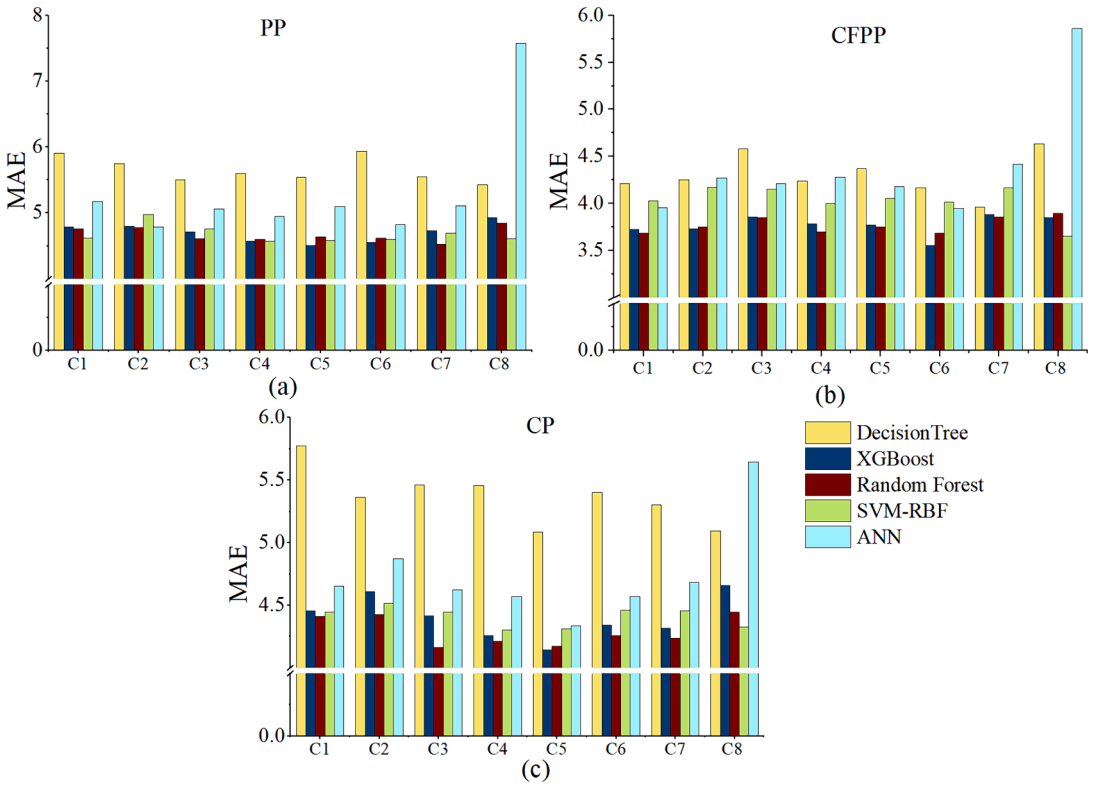

Fig. 2. MAE performance of the algorithms that exceed the Mean Predictor for every attribute combination and property.

better than others. When trying to predict the CFPP, good performance can be achieved with combination 6 (FAME and DB), which is the only combination that was able to improve on the base combination (FAME). It seems that the number of double bonds is important for predicting the CFPP and for this particular problem it is the only attribute that provides relevant information. For predicting the CP there are several combinations of attributes that are better than just using the FAME distributions. However, the best one seems to be combination 5 (FAME, SFA/MUFA/PUFA, DB and NPI). To predict the CP not only the DB is important but also the rest of the added attributes as they improve the performance of the algorithm. Although Mostafaei [52] proposed attribute combination 8 (SFA/MUFA/PUFA, Mw and Cn) as the best for predicting CP and PP, in our experiments this does not seem to be the case since the performance obtained with the Mostafaei attributes is worse than that obtained with the FAME distribution alone. For the PP prediction there are several attribute combinations that improve the performance, although this improvement is not so marked. Combinations 4, 5, 6 and 7 present the best results, which reinforces the importance of the number of double bonds and the SFA/MUFA/PUFA.

In general, using the FAME distribution (combination 1) for the prediction shows a better performance than not using the FAME distribution (combination 8). The performance increases when the number of double bonds (combination 6) is included, and for the PP and CP in particular, the performance also increases when the SFA/MUFA/PUFA and NPI (combination 5) are included. The Mw and CN do not seem to be important in predicting of these properties.

After considering the results for the three different metrics, the best model is chosen for each property. The best algorithms are the XGBoost for the PP and the CFPP, while for the CP it is the Random Forest. For the PP the best combination of attributes is 5 (MAE: 4.51, MSE: 44.31, R2: 0.48), for the CFPP the best combination is 6 (MAE: 3.56, MSE: 25.76, R2: 0.60) and for CP the best combination is 5 again (MAE: 4.17, MSE: 37.76, R2: 0.43).

4.2. Comparison of built models with previously reported models

The performance of the best model for each property was compared with six different models found in the literature, whose performance was measured using the datasets, metrics and methodology presented in this work.

- **Model 1.** This model proposed by Bolonio et al. [28] presents three equations, each of which could predict one of the properties: PP **[Eq. (6)]**, CFPP **[Eq. (7)]**, and CP **[Eq. (8)]** (referred to in the figures as $Model1_{PP}$, $Model1_{CFPP}$, and $Model1_{CP}$, respectively). In these equations, U_{FAME} is the sum of unsaturated FAMEs, expressed as a percentage and N_C is the weighted average number of carbons (equivalent to the Cn attribute in this article).

$$PP(^{\circ}C) = -125.04 - 0.62 \bullet U_{FAME} + 8.61 \bullet N_C \quad (6)$$

$$CFPP(^{\circ}C) = -103.47 - 0.59 \bullet U_{FAME} + 7.3 \bullet N_C \quad (7)$$

$$CP(^{\circ}C) = -81.62 - 0.45 \bullet U_{FAME} + 5.87 \bullet N_C \quad (8)$$

- **Model 2.** This model presented by Sarin et al. [29] proposes two different equations that have been able to predict PP **[Eq. (9)]** and CP **[Eq. (10)]** (named as $Model2_{PP}$ and $Model2_{CP}$, respectively). This model uses only the variable of palmitic acid methyl ester (PAME C16:0) concentration in percent and it is limited to 45 %.

$$PP(^{\circ}C) = 0.571 \bullet (PAME) - 12.240 \quad (9)$$

$$CP(^{\circ}C) = 0.526 \bullet (PAME) - 4.992 \quad (10)$$

- **Model 3.** This model presents an equation for each property: PP **[Eq. (11)]**, CFPP **[Eq. (12)]**, and CP **[Eq. (13)]**, named as $Model3_{PP}$, $Model3_{CFPP}$, and $Model3_{CP}$, respectively. It was proposed by Dunn

[30] and it includes in the total saturated fatty acid methyl esters *SFAME* (equivalent to the SFA attribute).

$$PP(^{\circ}C) = 1.41 \bullet SFAME - 29.4 \quad (11)$$

$$CFPP(^{\circ}C) = 1.44 \bullet SFAME - 29.3 \quad (12)$$

$$CP(^{\circ}C) = 1.44 \bullet SFAME - 24.8 \quad (13)$$

- **Model 4.** This model proposed by Alviso et al. [31] presents an equation, designed using genetic programming, for each of the cold flow properties. All of them depend on some specific FAME and the equations are (**14**), (**15**) and (**16**) for PP, CFPP and CP, named as $Model4_{PP}$, $Model4_{CFPP}$, and $Model4_{CP}$, respectively. In these equations x_{PA} is the percentage of palmitic acid ester (C16:0), x_{ST} is the percentage of stearic acid ester (C18:0), x_{LI} is the percentage of lignoceric acid ester (C24:0), x_{LN} is the percentage of linoleic acid ester (C18:2), and x_{BE} is the percentage of behenic acid ester (C22:0).

$$PP(K) = 267.303 + 0.3x_{PA} + 0.505x_{ST} - 0.1x_{LI} - 0.1x_{LN} \quad (14)$$

$$CFPP(K) = 259.051 + 0.72834x_{PA} + 0.5x_{ST} + 7.71255x_{BE} \quad (15)$$

$$CP(K) = 268.444 + 0.2x_{PA} + 0.666x_{ST} \quad (16)$$

- **Model 5.** The model proposed by Serrano et al. [32] states that the CFPP can be predicted using three parameters according to **Eq. (17)**, named as $Model5_{CFPP}$. The three parameters chosen by the authors are the FAME content of the saturated compounds from C4:0 to C14:0 (SAT_{C4-C14}), the FAME content of the saturated compounds from C16:0 to C24:0 ($SAT_{C16-C24}$), and the FAME content of the unsaturated compounds (*UNSAT*), all expressed in percent. The authors also proposed some applicability conditions for the correlation, which are given below the equation.

$$CFPP(^{\circ}C) = 0 - 0.12 \bullet SAT_{C4-C14} + 0.47 \bullet SAT_{C16-C24} - 0.14 \bullet UNSAT \quad (17)$$

$$0 < SAT_{C4-C14} \leq 81.1$$

$$7.4 \leq SAT_{C16-C24} \leq 44.4$$

$$7.1 \leq UNSAT \leq 92.2$$

- **Model 6.** This model is an ANN that uses the FAME distribution to make the prediction and it was proposed by Al-Shanableh et al. [53] (referred to as "ANN Al-Shanableh" in the figures). The architecture of the ANN has 9 neurons in the input layer (corresponding to the number of most common FAME compounds in their database), 6 neurons in the hidden layer using a sigmoid activation function, and one neuron in the output layer.

Fig. 3 shows a comparison between the models proposed in the literature and the prediction model for PP. The first model proposed by Bolonio et al. ($Model1_{PP}$) used only 16 biodiesel distributions to build their equations, 6 of which are blends of at least 2 of the 10 base biodiesels. Therefore, their correlation may adequately predict the cold flow properties for biodiesels similar to those used in the study but has problems generalizing to other biodiesels with different composition.

In a similar manner, Dunn proposed a correlation ($Model3_{PP}$) that depends only on the saturated fatty acids and used only 9 examples of biodiesel to build the model. With such a small database, this model has problems predicting PP for biodiesels from the database used in this study that have a different saturated fatty acid content from those used in Dunńs study. In contrast to $Model1_{PP}$ and $Model3_{PP}$, the model proposed by Sarin et al. [29] ($Model2_{PP}$) is also based on one attribute but shows better results. The reason for this is that the attribute selected for prediction (palmitic acid ester C16:0) appears in almost every biodiesel example. Insisting on the approach of using the most representative FAMEs to predict PP, the model proposed by Alviso et al. [31] ($Model4_{PP}$) considers a selection of the most representative FAMEs, leading to better results.

The model proposed by Al-Shanableh et al. (ANN Al-Shanableh) has a good performance, but not as good as that reported in their paper. This may be due to overfitting, as the ANN structure they proposed seems to predict more accurately biodiesels similar to those selected in their study and predicts many of the rest poorly.

Fig. 4 shows the comparison between the best CFPP model trained in the present investigation and the models proposed in the literature. The performance obtained by the models 1, 3, 4, and 6 ($Model1_{CFPP}$, $Model3_{CFPP}$, $Model4_{CFPP}$, ANN [53]) is similar to that obtained for the same models to predict the PP, and the reasons for these performances are the

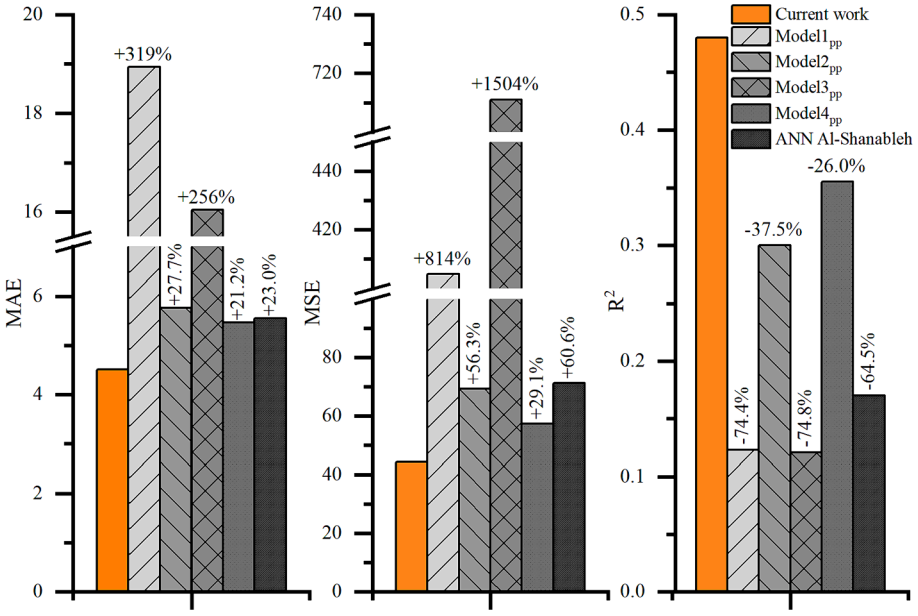

Fig. 3. Comparison between the best PP model and 5 different models proposed in the literature.

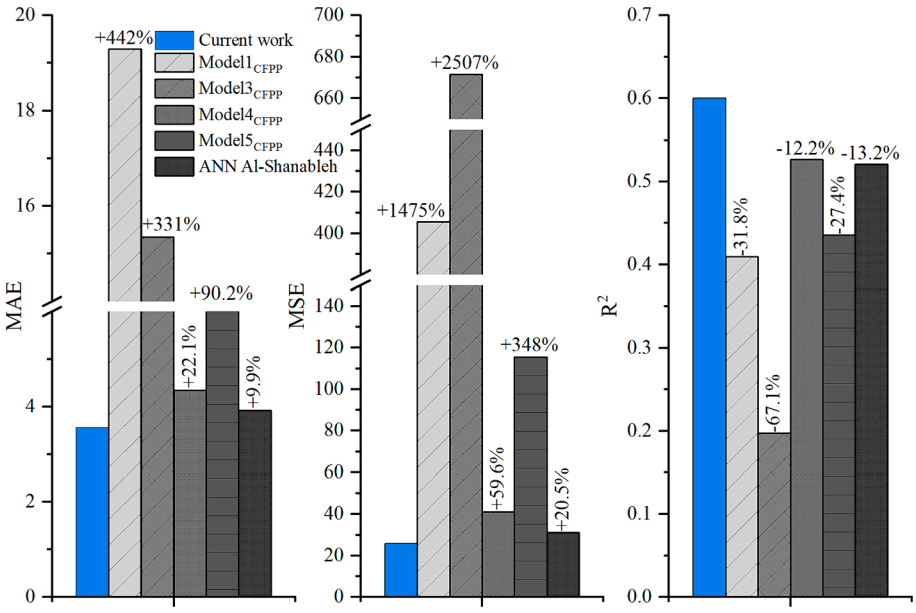

Fig. 4. Comparison between the best CFPP model and 5 different models proposed in the literature.

same.

Model 5, proposed by Serrano et al. (2014), is included in the comparison (Model5$_{CFPP}$). This model uses a selection of attributes that discriminate between the short chain length FAMEs and the long chain length FAMEs. This gave better results than Model1$_{CFPP}$ and Model3$_{CFPP}$, but the restrictions applied to the possible values of the attributes could make it impossible to use this correlation. In the case of the database presented in this work, the correlation could not be applied to 20.6 % of the biodiesel examples. For all the other models and the algorithms proposed in this work, all the biodiesels were included in the performance evaluation. This distinction gave an advantage to Model5$_{CFPP}$, since some biodiesels that are considered to give bad results are excluded.

A similar analysis for CP is shown in Fig. 5. The selected models are the same as those selected for the PP property and have the same deficiency as those mentioned above, and the figure shows that the model proposed in this study outperforms all the other models.

Figs. 3, 4 and 5 show that the performance results for the R^2 metric are not very good for any of the models. However, this does not mean that the models presented in this paper are not useful, as a closer look at the MAE metric leads to several conclusions. The MAE for PP, CFPP, and CP are 4.73 °C, 3.68 °C, and 4.16 °C, respectively, and these results provide a confidence interval where the prediction should lie. Using the prediction in conjunction with the confidence interval, the model can be used to estimate whether or not it meets the requirements of the regulation. Since the standards specify an upper limit value for each property, it is sufficient that the predicted value and its confidence interval is below this limit to ensure the corresponding standard compliance.

4.3. *Attribute relevance in the prediction*

The *XGBoost* and *Random Forest* algorithms are not only able to build

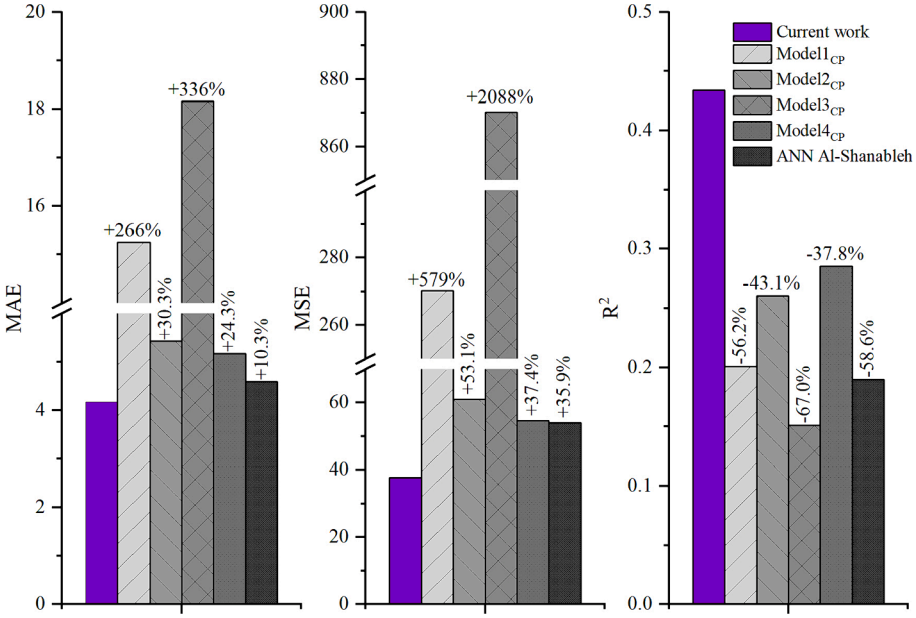

Fig. 5. Comparison between the best CP model and 5 different models proposed in the literature.

a predictive model, but also to determine the significance of each attribute in making the prediction. Fig. 6 shows the importance of each attribute, expressed in percentage, for the properties analysed in this work, and it can be seen that the FAME distribution and the number of double bonds play an important role in all three models.

For the PP prediction, the attribute combination 5 (FAME, SFA/MUFA/PUFA, DB and NPI) was selected and the relative importance of these attributes is shown in Fig. 6 (a), where the monounsaturated fatty acids (MUFA) gain some importance. For the CFPP prediction, the attribute combination 6 (FAME and DB) was selected and among the FAMEs the C20:1 seems to be of particular relevance. Finally, for CP, where the attribute combination 5 was also used, the saturated fatty acids (SFA) become important.

This relationship was expected, as previous studies have shown a correlation between the degree of saturation of biodiesel and its properties at low temperatures. It is well known that higher degrees of saturation result in higher PP, CFPP, and CP points [54]. Additionally, longer chain lengths also lead to higher PP, CFPP, and CP [55]. However, the interactions between the different FAMEs are not fully understood. Therefore, IA-based models that can capture the complexity of the problem are needed.

Regarding the FAME distribution, it is necessary to mention that the algorithms generally assign more relevance to those FAMEs with greater presence in the database (see Fig. 1). If a larger database with more homogeneous distribution is used, the relevance of these FAMEs is likely to vary.

5. Conclusions

In the present study, a selection of frequently used machine learning algorithms as well as 3 databases with more than 200 examples of biodiesel for each property (pour point, cold filter plugging point, and cloud point) were presented and made publicly available. Different models were built with different combinations of attributes and several conclusions were reached.

The use of machine learning applications improves previous models that relied on empirical equations, resulting in better generalization. Similarly, to achieve optimal generalization, it is crucial to use databases that accurately represent the biodiesel set. Therefore, creating examples through artificial mixtures and using a limited number of examples is discouraged. The FAME distribution and the number of double bonds are revealed as the primary factors that affect the low temperature properties of biodiesels. Other potentially useful attributes for prediction include SFA, MUFA, PUFA, and NPI.

Although the R^2 correlation coefficients of the models are not high, they are still useful for the industry. These correlations may have a low R^2 for several reasons, such as the limited amount of data, inaccuracies in the data, the use of different equipment or test configuration, or other factors that have not been taken into consideration but that could have provided relevant information. They can be used in conjunction with the

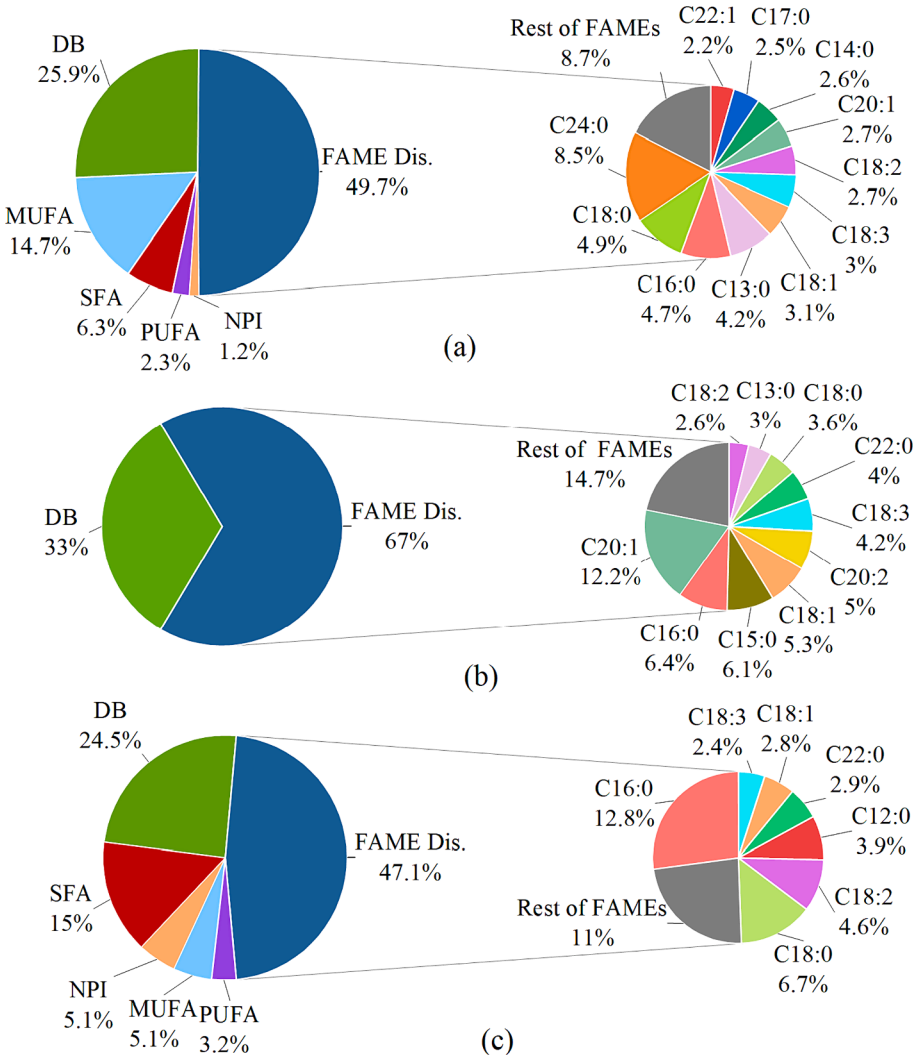

Fig. 6. Attribute relevance in the prediction of: (a) PP, (b) CFPP, and (c) CP.

obtained MAE values to establish predictions and confidence intervals. This way, if the predicted value plus the interval are within the specification, the biodiesel can be assumed to meet the cut-off.

To enhance these models, additional diverse examples should be added to the database. In addition, it is crucial to explore methods to improve their interpretability. One approach could be to transform the models into hybrid systems [56], which would identify not only the most important features, but also their interactions. Finally, a combined model should be developed by integrating these and other models for different properties to identify all biodiesel characteristics.

CRediT authorship contribution statement

G. Díez-Valbuena: Writing – original draft, Resources, Methodology, Investigation, Formal analysis. **A. García Tuero:** Writing – review & editing, Supervision. **J. Díez:** Writing – review & editing, Methodology, Conceptualization. **E. Rodríguez:** Writing – review & editing, Supervision, Conceptualization. **A. Hernández Battez:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

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

The database has been published in the repository https://doi.org/10.5281/zenodo.10670349.

Acknowledgements

This research was funded by the Ministry of Science, Innovation and Universities (Spain) and the State Investigation Agency, grant number: PID2022-136656NB-I00 (LubeMicroAlgae project), and by the Foundation for the Promotion of Applied Scientific Research and Technology in Asturias (Spain), which financed the contract of Guillermo Díez-Valbuena at the University of Oviedo (Spain), grant number: SV-PA-21-AYUD/2021/50987. Guillermo Díez-Valbuena acknowledges the support of the Government of the Principality of Asturias under Severo Ochoa predoctoral program (ref. BP22-153).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.molliq.2024.124555.

References

[1] N. Abas, A. Kalair, N. Khan, Review of fossil fuels and future energy technologies, Futures 69 (2015) 31–49, https://doi.org/10.1016/j.futures.2015.03.003.
[2] M. Lenzen, Life cycle energy and greenhouse gas emissions of nuclear energy: a review, Energy Convers. Manag. 49 (2008) 2178–2199, https://doi.org/10.1016/j.enconman.2008.01.033.
[3] M.D. Esteban, J.J. Diez, J.S. López, V. Negro, Why offshore wind energy? Renew. Energy 36 (2011) 444–450, https://doi.org/10.1016/j.renene.2010.07.009.
[4] E. Kabir, P. Kumar, S. Kumar, A.A. Adelodun, K.-H. Kim, Solar energy: potential and future prospects, Renew. Sustain. Energy Rev. 82 (2018) 894–900, https://doi.org/10.1016/j.rser.2017.09.094.
[5] A.C.T. Malaquias, N.A.D. Netto, F.A.R. Filho, R.B.R. Da Costa, M. Langeani, J.G. C. Baêta, The misleading total replacement of internal combustion engines by electric motors and a study of the brazilian ethanol importance for the sustainable future of mobility: a review, J. Braz. Soc. Mech. Sci. Eng. 41 (2019) 567, https://doi.org/10.1007/s40430-019-2076-1.
[6] A. Demirbas, Progress and recent trends in biofuels, Prog. Energy Combust. Sci. 33 (2007) 1–18, https://doi.org/10.1016/j.pecs.2006.06.001.
[7] M.V. Rodionova, R.S. Poudyal, I. Tiwari, R.A. Voloshin, S.K. Zharmukhamedov, H. G. Nam, B.K. Zayadan, B.D. Bruce, H.J.M. Hou, S.I. Allakhverdiev, Biofuel production: challenges and opportunities, Int. J. Hydrog. Energy 42 (2017) 8450–8461, https://doi.org/10.1016/j.ijhydene.2016.11.125.
[8] L. Meher, D. Vidyasagar, S. Naik, Technical aspects of biodiesel production by transesterification—a review, Renew. Sustain. Energy Rev. 10 (2006) 248–268, https://doi.org/10.1016/j.rser.2004.09.002.
[9] D. Singh, D. Sharma, S.L. Soni, S. Sharma, P. Kumar Sharma, A. Jhalani, A review on feedstocks, production processes, and yield for different generations of biodiesel, Fuel 262 (2020) 116553. Doi: 10.1016/j.fuel.2019.116553.
[10] G. Antolín, F.V. Tinaut, Y. Briceño, Optimisation of biodiesel production by sunflower oil transesterification, Bioresour. Technol. 83 (2002) 111–114, https://doi.org/10.1016/S0960-8524(01)00200-0.
[11] M.I. Al-Widyan, M.A. Al-Muhtaseb, Experimental investigation of jojoba as a renewable energy source, Energy Convers. Manag. 51 (2010) 1702–1707, https://doi.org/10.1016/j.enconman.2009.11.043.
[12] T.M. Mata, A.A. Martins, S. Nidia, Caetano, Microalgae for biodiesel production and other applications: a review, renew, Sustain. Energy Rev. 14 (2010) 217–232, https://doi.org/10.1016/j.rser.2009.07.020.
[13] O. Inganäs, V. Sundström, Solar energy for electricity and fuels, Ambio 45 (2016) 15–23, https://doi.org/10.1007/s13280-015-0729-6.
[14] CTN 51/SC 3, EN 14214:2013 V2+A2:2019 Liquid petroleum products - Fatty acid methyl esters (FAME) for use in diesel engines and heating applications - Requirements and test methods, (2019). https://www.une.org/encuentra-tu-norma/busca-tu-norma/norma?c=norma-une-en-14214-2013-v2-a2-2019-n0062687 (accessed October 2, 2023).
[15] D02.E0, ASTM D6751-20a Standard Specification for Biodiesel Fuel Blend Stock (B100) for Middle Distillate Fuels, (2023). https://www.astm.org/d6751-20a.html (accessed October 2, 2023).
[16] A. Wypych, G. Wypych, 2 - Information on data fields, in: A. Wypych, G. Wypych (Eds.), Datab. Rheol. Addit., ChemTec Publishing, 2022: pp. 3–17. Doi: 10.1016/B978-1-927885-91-8.50005-1.
[17] ISO/TC 28, ISO 3016:2019 Petroleum and related products from natural or synthetic sources. Determination of pour point, (2019). https://www.iso.org/standard/73386.html (accessed October 2, 2023).
[18] CEN/TC 19, EN 116:2015 - Diesel and domestic heating fuels - Determination of cold filter plugging point - Stepwise cooling bath method, (2015). https://standards.iteh.ai/catalog/standards/cen/f0b4bc72-e8b9-4969-b957-8e6ce72f489d/en-116-2015 (accessed October 3, 2023).
[19] ISO/TC 28, ISO 3015:2019 Petroleum and related products from natural or synthetic sources. Determination of cloud point, (2019). https://www.iso.org/standard/72765.html (accessed October 2, 2023).
[20] CTN 51, EN 590:2022 Automotive fuels - Diesel - Requirements and test methods, (2022). https://www.une.org/encuentra-tu-norma/busca-tu-norma/norma?c=N0069568 (accessed October 2, 2023).
[21] D02.E0, ASTM D975-21 Standard Specification for Diesel Fuel, (2022). https://www.astm.org/d0975-21.html (accessed October 3, 2023).
[22] Y.-C. Su, Y.A. Liu, C.A. Diaz Tovar, R. Gani, Selection of prediction methods for thermophysical properties for process modeling and product Design of Biodiesel Manufacturing, Ind. Eng. Chem. Res. 50 (2011) 6809–6836, https://doi.org/10.1021/ie102441u.
[23] A. Sarin, R. Arora, N.P. Singh, R. Sarin, R.K. Malhotra, S. Sarin, Blends of biodiesels synthesized from non-edible and edible oils: effects on the cold filter plugging point, Energy Fuels 24 (2010) 1996–2001, https://doi.org/10.1021/ef901131m.
[24] M.J. Ramos, C.M. Fernández, A. Casas, L. Rodríguez, Á. Pérez, Influence of fatty acid composition of raw materials on biodiesel properties, Bioresour. Technol. 100 (2009) 261–268, https://doi.org/10.1016/j.biortech.2008.06.039.
[25] S. Pinzi, D. Leiva, G. Arzamendi, L.M. Gandia, M.P. Dorado, Multiple response optimization of vegetable oils fatty acid composition to improve biodiesel physical properties, Bioresour. Technol. 102 (2011) 7280–7288, https://doi.org/10.1016/j.biortech.2011.05.005.
[26] M.-H. Yuan, Y.-H. Chen, J.-H. Chen, Y.-M. Luo, Dependence of cold filter plugging point on saturated fatty acid profile of biodiesel blends derived from different feedstocks, Fuel 195 (2017) 59–68, https://doi.org/10.1016/j.fuel.2017.01.054.
[27] L. Wang, H. Yu, X. He, R. Liu, Influence of fatty acid composition of woody biodiesel plants on the fuel properties, J. Fuel Chem. Technol. 40 (2012) 397–404, https://doi.org/10.1016/S1872-5813(12)60018-8.
[28] D. Bolonio, A. Llamas, J. Rodríguez-Fernández, A.M. Al-Lal, L. Canoira, M. Lapuerta, L. Gómez, Estimation of cold flow performance and oxidation stability of fatty acid ethyl esters from lipids obtained from *Escherichia coli*, Energy Fuels 29 (2015) 2493–2502, https://doi.org/10.1021/acs.energyfuels.5b00141.
[29] A. Sarin, R. Arora, N.P. Singh, R. Sarin, R.K. Malhotra, K. Kundu, Effect of blends of palm-Jatropha-Pongamia biodiesels on cloud point and pour point, Energy 34 (2009) 2016–2021, https://doi.org/10.1016/j.energy.2009.08.017.
[30] R.O. Dunn, Cold flow properties of biodiesel: a guide to getting an accurate analysis, Biofuels 6 (2015) 115–128, https://doi.org/10.1080/17597269.2015.1057791.
[31] D. Alviso, G. Artana, T. Duriez, Prediction of biodiesel physico-chemical properties from its fatty acid composition using genetic programming, Fuel 264 (2020) 116844, https://doi.org/10.1016/j.fuel.2019.116844.
[32] M. Serrano, R. Oliveros, M. Sánchez, A. Moraschini, M. Martínez, J. Aracil, Influence of blending vegetable oil methyl esters on biodiesel fuel properties: oxidative stability and cold flow properties, Energy 65 (2014) 109–115, https://doi.org/10.1016/j.energy.2013.11.072.
[33] R. Razavi, A. Bemani, A. Baghban, A.H. Mohammadi, S. Habibzadeh, An insight into the estimation of fatty acid methyl ester based biodiesel properties using a LSSVM model, Fuel 243 (2019) 133–141, https://doi.org/10.1016/j.fuel.2019.01.077.

[34] T. Hastie, R. Tibshirani, J. Friedman, The elements of Statistical Learning Data Mining, Interference and Prediction, Springter, New York, 2009.
[35] S.K. Hoekman, A. Broch, C. Robbins, E. Ceniceros, M. Natarajan, Review of biodiesel composition, properties, and specifications, Renew. Sustain. Energy Rev. 16 (2012) 143–169, https://doi.org/10.1016/j.rser.2011.07.143.
[36] D. Singh, D. Sharma, S.L. Soni, S. Sharma, D. Kumari, Chemical compositions, properties, and standards for different generation biodiesels: a review, Fuel 253 (2019) 60–71, https://doi.org/10.1016/j.fuel.2019.04.174.
[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, Scikit-learn: machine Learning in python, Mach. Learn. PYTHON 12 (2011) 2825–2830.
[38] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106, https://doi.org/10.1007/BF00116251.
[39] L. Breiman, Random Forest, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[40] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, in: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., ACM, San Francisco California USA, 2016: pp. 785–794. Doi: 10.1145/2939672.2939785.
[41] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140, https://doi.org/10.1007/BF00058655.
[42] R.E. Schapire, A brief introduction to boosting, in: Proc. 16th Int. Jt. Conf. Artif. Intell, - Vol. 2, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999, pp. 1401–1406.
[43] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297, https://doi.org/10.1007/BF00994018.
[44] J. Hertz, A. Krogh, R.G. Palmer, Introduction to the theory of neural computation, Addison-Wesley Pub. Co, Redwood City, Calif, 1991.
[45] R. Kohavi, A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, in: (Montreal) Canada, 1995: pp. 1137–1143.
[46] G. Díez Valbuena, A. García Tuero, J. Díez, E. Rodríguez, A. Hernández Battez, Application of machine learning techniques to predict biodiesel iodine value, Energy 292 (2024) 130638, https://doi.org/10.1016/j.energy.2024.130638.
[47] J.-Y. Park, D.-K. Kim, J.-P. Lee, S.-C. Park, Y.-J. Kim, J.-S. Lee, Blending effects of biodiesels on oxidation stability and low temperature flow properties, Bioresour. Technol. 99 (2008) 1196–1203, https://doi.org/10.1016/j.biortech.2007.02.017.
[48] M.E. González Gómez, R. Howard-Hildige, J.J. Leahy, B. Rice, Winterisation of waste cooking oil methyl ester to improve cold temperature fuel properties, Fuel 81 (2002) 33–39, https://doi.org/10.1016/S0016-2361(01)00117-X.
[49] G. Knothe, Structure indices in FA chemistry. how relevant is the iodine value? J. Am. Oil Chem. Soc. 79 (2002) 847–854, https://doi.org/10.1007/s11746-002-0569-4.
[50] M. Arif, Y. Bai, M. Usman, M. Jalalah, F.A. Harraz, M.S. Al-Assiri, X. Li, E.-S. Salama, C. Zhang, Highest accumulated microalgal lipids (polar and non-polar) for biodiesel production with advanced wastewater treatment: role of lipidomics, Bioresour. Technol. 298 (2020) 122299, https://doi.org/10.1016/j.biortech.2019.122299.
[51] C. Bachler, S. Schober, M. Mittelbach, Simulated distillation for biofuel analysis, ACS Publ. (2009), https://doi.org/10.1021/ef901295s.
[52] M. Mostafaei, Prediction of biodiesel fuel properties from its fatty acids composition using ANFIS approach, Fuel 229 (2018) 227–234, https://doi.org/10.1016/j.fuel.2018.04.148.
[53] F. Al-Shanableh, A. Evcil, M.A. Savaş, Prediction of cold flow properties of biodiesel fuel using artificial neural network, Procedia Comput. Sci. 102 (2016) 273–280, https://doi.org/10.1016/j.procs.2016.09.401.
[54] R.D. Lanjekar, D. Deshmukh, A review of the effect of the composition of biodiesel on NO x emission, oxidative stability and cold flow properties, Renew. Sustain. Energy Rev. 54 (2016) 1401–1411, https://doi.org/10.1016/j.rser.2015.10.034.
[55] J.F. Sierra-Cantor, C.A. Guerrero-Fajardo, Methods for improving the cold flow properties of biodiesel with high saturated fatty acids content: a review, Renew. Sustain. Energy Rev. 72 (2017) 774–790, https://doi.org/10.1016/j.rser.2017.01.077.
[56] S. Zendehboudi, N. Rezaei, A. Lohi, Applications of hybrid models in chemical, petroleum, and energy systems: a systematic review, Appl. Energy 228 (2018) 2539–2566, https://doi.org/10.1016/j.apenergy.2018.06.051.