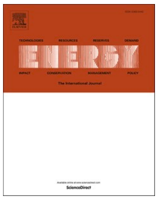
Energy 292 (2024) 130638

Contents lists available at ScienceDirect

Energy

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



Application of machine learning techniques to predict biodiesel iodine value

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

Handling Editor: Henrik Lund

Keywords:
Machine learning
Prediction models
Iodine value
Biodiesel

ABSTRACT

Biodiesel is a good alternative to fossil fuels for conventional engines, but determining the properties of biodiesel can be a time-consuming and resource-intensive process. Therefore, the development of models capable of predicting these properties would be of great importance. In this work, different machine learning models were investigated for predicting the Iodine Value (IV) based on the distribution of fatty acid methyl esters (FAME). For this purpose, a database with 266 examples of biodiesel from different feedstocks (1st, 2nd and 3rd generation) was used along the leave-one-out methodology. The main results of the work are: the double bonds and the distribution of FAMEs are the best attributes for predicting IV and the XGBoost algorithm gives an absolute mean error of 11.4 units; the machine learning models for predicting biodiesel properties need to be trained on a large number and variety of biodiesel examples to better predict and generalize; the use of both ANNs and the hold-out approach of dividing the dataset into train/validate/test are not recommended due to the risk of overfitting and the algorithm's dependence on which examples form each group given the currently available data. The leave-one-out method is most appropriate for estimating model performance.

1. Introduction

The importance of clean energy generation is growing over the years. In 2015 the United Nations (UN) promoted an affordable, reliable, sustainable and modern energy transition for all and the decarbonization of fuels plays a crucial role in the process [1]. Around 99.8% of transport is currently powered by internal combustion engines (ICEs) [2] and according to Senecal et al. [3], despite government policies, half of the vehicle fleet is still expected to be powered by ICEs by 2050. Therefore, one of the best options for reducing greenhouse gas emissions would be to find alternatives to fossil fuels that can be used in the current vehicle fleet. Considering the above facts, further research in the field of fuels such as biodiesel, it's an appropriate approach [4].

Biodiesel can be produced by a variety of methods, but the most viable method is the transesterification of oils derived from a variety of feedstocks. In this process, lipids extracted from the feedstock react with methanol or ethanol in the presence of a catalyst (acid or basic) to convert triglycerides to fatty acid methyl or ethyl esters (FAME or FAEE). Depending on the origin of the feedstock, biodiesel can be classified as "Biodiesel of 1st generation" when derived from edible crops [5] and "Biodiesel of 2nd generation" when derived from non-edible ones [6,7]. These feedstocks have been widely discussed as they affect the food chain and the use of arable land [8]. As a result of this concern, biodiesel of 3rd generation [8,9] and 4th generation [10,11] can be obtained from microalgae resources. Microalgae are photosynthetic microorganisms that can grow in freshwater, seawater or wastewater, and therefore do not require arable land for their cultivation, and are great carbon sequestrators [12].

Biodiesel must meet the requirements outlined in the EN 14214:2012 + A2:2019 [13] and ASTM D6751 [14] standards. Both standards specify that several important properties must be within certain limits, as shown in Table 1. Testing for these properties is a time-consuming and costly process [15]. One of the critical conditions specified in the EN 14214:2012 + A2:2019 standard is that the fatty acid methyl esters (FAME) content must be greater than 96.5%. Since the molecular characteristics determine many of the remaining biodiesel properties, some researchers have attempted to predict the value of biodiesel properties from the FAME distribution to avoid performing all the tests included in the above standards [16,17].

Among all the mentioned properties, iodine value (IV) represents the unsaturated methyl esters constituents of biodiesel, which affect oxidative stability [18] and cold filter plugging point [19]. The IV is the

* Corresponding author.
E-mail address: aehernandez@uniovi.es (A. Hernández Battez).

https://doi.org/10.1016/j.energy.2024.130638
Received 15 September 2023; Received in revised form 2 February 2024; Accepted 6 February 2024
Available online 7 February 2024

Table 1
Range of biodiesel properties extracted from EN 14214:2012 + A2:2019 [13] and ASTM D6751 [14] standards.

Property	Units	EN 14214	Test methods	ASTM D6751	Test methods
FAME content	% (m/m)	≥96.5	EN 14103	–	
Density at 15 °C	kg/m^3	860–900	ISO 12185	–	
Viscosity at 40 °C	mm^2/s	3.50–5.00	EN 16896	1.9–6	D445
Flash Point	°C	≥101	ISO 3679	≥130	D93
Cetane Number (CN)	–	≥51.0	EN 16175	≥47	D613
Copper corrosion	Class	≥1a	ISO 2160	≤3	D130
Oxidative stability (110 °C)	h	≥8	EN 14112	≥3	EN 14112
Acid value	mg KOH/g	≤0.50	EN 14104	≤0.5	D664
Iodine Value (IV)	g I_2/100 g	≤120	EN 14111	–	
Methyl ester linoleic acid	% (m/m)	≤12	EN 14103	–	
Methyl ester + 4 double bonds	% (m/m)	≤1	EN 15779	–	
Methanol content	% (m/m)	≤0.2	EN 14110	≤0.2	EN 14110
Monoglycerides content	% (m/m)	≤0.7	EN 14105	–	
Diglycerides content	% (m/m)	≤0.2	EN 14105	–	
Triglycerides content	% (m/m)	≤0.2	EN 14105	–	
Free Glycerine	% (m/m)	≤0.02	EN 14105	≤0.02	D6584
Total Glycerine	% (m/m)	≤0.25	EN 14105	≤0.24	D6584
Water content	% (m/m)	≤0.050	ISO 12937	≤0.05	D2709
Total contamination	mg/kg	≤24	EN 12662	–	
Sulphated ash content	% (m/m)	≤0.02	ISO 3987	≤0.02	D874
Sulphur content	mg/kg-% (m/m)	≤10	ISO 20846	≤15	D5453
Group I metals (Na + K)	mg/kg	≤5	EN 14108	≤5	EN 14538
Group II metals (Ca + Mg)	mg/kg	≤5	EN 14538	≤5	EN 14538
Phosphorus content	mg/kg-% (m/m)	≤4	EN 14107	≤10	D4951

mass of iodine absorbed in 100 g of biodiesel through the reaction between the I_2 and the carbon-carbon double bonds, providing a measure of the degree of unsaturation [20]. These unsaturated methyl esters can react with atmospheric oxygen during storage and handling leading to the formation of primary oxidation products like peroxides and hydroperoxides, which react forming numerous secondary oxidation products [21]. Subsequently, gums and sediments are formed by polymerization reactions [22]. All these changes result in variations of molecular composition and deterioration of the fuel quality. In addition, the IV is inversely related to cetane number (CN) and the NOx emissions increase with the IV [23]. Because of these facts the IV is limited to a maximum of 120 g I_2/100 g in the EN 14214 standard.

Models to predict the IV of biodiesel help to understand a priori the oxidative stability and predict other properties, and also allow the selection of feedstocks to produce biodiesel. The approach followed by the proposed models to predict the IV of biodiesel has been mainly based on biodiesel composition (mass % of FAME) [24,25], composition-based indicators (molecular weight, modified degree of unsaturation) [26] or both [27]. Despite the high accuracy reported for these models, they are very limited by the range and type of FAME in the model calibration and have not been validated with additional datasets. In addition, these models have a limited scope as they are usually linear, and the problem seems to be more complex [26]. Therefore, techniques based on machine learning [28] or deep learning [29] seem to be appropriate as they can model complex problems. These techniques have already been implemented with great results in other disciplines, such as the prediction of a gas turbine performance by Liu and Karimi [30] or an energy forecasting strategy by Ahmad et al. [31]. Although the have also been used to predict the properties of the biodiesel, the models developed lack generalization. Some do not use the proper performance metrics [32], others fail to ensure a robust data set as they generate artificial examples [33], or the methodology is not adequate for the number of examples used [34].

This work aims to develop a model for predicting iodine value and provide guidance on the experimental methodology required to create a generalized model, determining which is the best experimental methodology taking into account the size of the datasets commonly used in the literature. This work also intends to determine the importance of different attributes used to make the IV prediction. Various machine learning techniques will be applied to this end, and it is expected that new possibilities and deeper insights into the optimization of biodiesel production and use will be achieved.

2. Material and methods

This section describes the data base and several machine learning algorithms suitable for prediction. Since the IV is a continuous variable, the selection should focus on algorithms capable of predicting numeric values.

2.1. Database description

This work includes a database[1] of 266 examples of biodiesel collected from the available literature from 2002 to 2022. The biodiesel come from more than 100 different feedstocks (edible and non-edible crops, microalgae …) to minimize bias. In the dataset the biodiesel came from the first, second and third generations with a distribution of 153 examples of first generation, 83 of second generation and 30 of third generation. Similar works, such as those presented by Azam [35] or Wang [26], use smaller datasets, including from 10 to 46 biodiesel samples.

The dataset presented in this work is composed out of 36 identified FAME. The column “Other” represent those FAMEs that could not be identified. The different FAMEs that appear in the dataset are: methyl butyric (C4:0), methyl hexanoate (C6:0), methyl octanoate (C8:0), methyl decanoate (C10:0), methyl laureate (C12:0), methyl tridecanoleate (C13:0), methyl myristate (C14:0), methyl meristoleate (C14:1), methyl pentadecanoate (C15:0), methyl pentadecenoate (C15:1), methyl palmitate (C16:0), methyl palmitoleate (C16:1), methyl hexadecadienoate (C16:2), methyl hexadecatrienoate (C16:3), methyl hexadecatetraenoate (C16:4), methyl margarate (C17:0), methyl heptadecenoate (C17:1), methyl stearate (C18:0), methyl oleate (C18:1), methyl linoleate (C18:2), methyl linolenate (C18:3), methyl nonadecanoate (C19:0), methyl nonadecenoate (C19:1), methyl nonadecetrinoate (C19:3), methyl arachidate (C20:0), methyl eicosenoate (C20:1), methyl eicosadienoate (C20:2), methyl arachidoniate (C20:4), methyl eicosapentanoate (C20:5), methyl heneicosanoate (C21:0), methyl behenate (C22:0), methyl erucate (C22:1), methyl decosatrienoate (C22:3), methyl decosapentaenoate (C22:5), methyl decosahexaenoate (C22:6) and methyl lignocerate (C24:0).

The frequency of occurrence of the different FAMEs is shown in

[1] The database is publicly available at <link> (link will be posted when article is published).

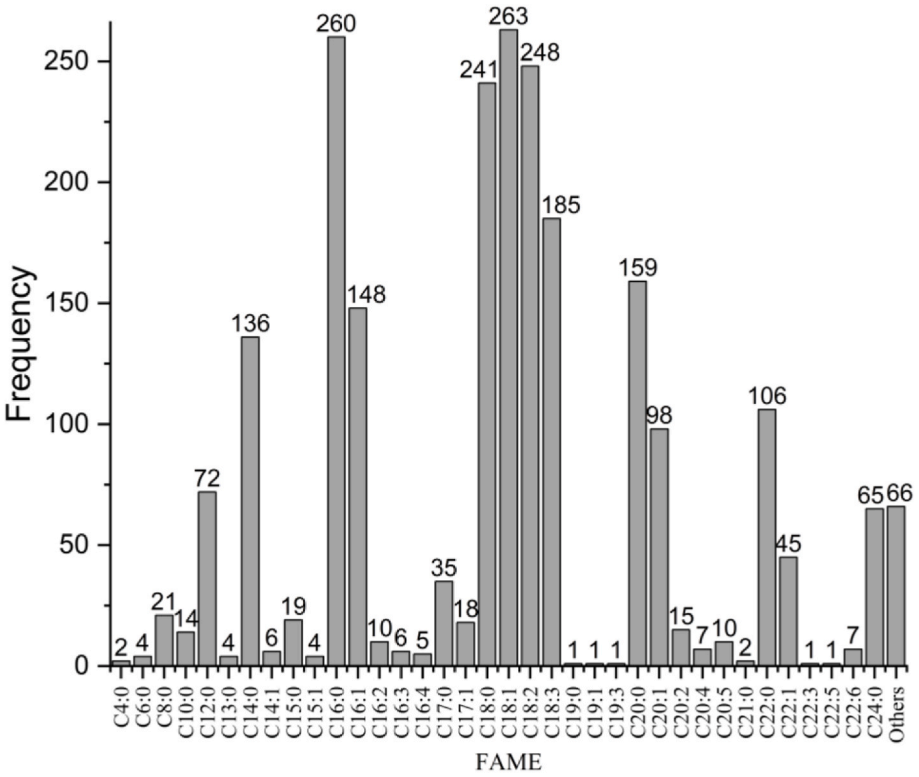


Fig. 1. Frequency distribution of the FAMEs presented in the current dataset.

Fig. 1. The FAMEs with the higher frequency are the methyl esters of the following fatty acids: C16:0, C18:0, C18:1, C18:2 and C18:3, whose presence ranged from 69.5 to 98.9% of the biodiesel simples. These four fatty acids are the most typical in commercial biodiesel according to Bukkarapu and Krishnasamy [36]. Barradas et al. [23] also found these fatty acids mainly in the 98 biodiesel samples used to predict viscosity, iodine number and induction period by applying artificial neural networks. However, the dataset used in this work shows that many more FAMEs can be found in biodiesel.

2.2. Data preparation

2.2.1. Data filtering

As mentioned in previous section our current dataset consists of 266 biodiesel samples. The total concentration of FAME in each of them varied from 50% to 100%. The data were pre-filtered to include only biodiesels with at least a 90% of FAME content, this reduced the dataset to 241 biodiesel samples. According to the standard EN 14214:2012 + A2:2019 [13], at least 96.5% of the total content must be FAME, but it is also known that most of the samples have to be distilled to reach this concentration [37]. The increase in FAME content due to distillation (reduction of impurities) is not expected to affect the prediction of IV.

2.2.2. Attribute selection

The aim of the model is to predict the IV using the FAME distribution as an input because this distribution is relatively easy to obtain. However, throughout the literature, other parameters have been used to make the prediction: the molecular weight (Mw), the length of the FAME represented by the number of carbons (Cn) [38], the number of double bonds (DB), the non-polarity index (NPI) (Eq. (1)), saturated fatty acids (SFA), monounsaturated fatty acids (MUFA), polyunsaturated fatty acids (PUFA) [39] or “other” [26] referring to what gas chromatography recognizes as unidentified FAMEs.

$$NPI = \frac{Cn \bullet Mw}{\sum FAME} \tag{1}$$

Considering these factors, experiments will be conducted to determine the attributes that best characterize the data and their importance in the prediction.

2.3. Regression algorithms in machine learning

There are many algorithms capable of predicting numeric values, and the most representative ones have been selected to be used in the experiments of this article. All these methods, which are described below, are available in the Python library Scikit-Learn [40].

2.3.1. Dummy regressor

When faced with a new problem, it is advisable to start by using a very basic method to predict. The results obtained with this method, used as a baseline, represent the performance that the more sophisticated algorithms should outperform to verify that they are learning something useful from the data [41]. In the case of a problem where we need to predict a numerical value, it is typically used as a baseline to predict the mean value of the target variable observed in the data [41]. This algorithm is so simple that it does not use the attributes describing the examples and is implemented in Scikit-Learn under the name DummyRegressor.

2.3.2. Linear regression

Multivariable linear regression is a useful method when trying to solve a regression problem [42]. Linear regression fits a linear model in which a coefficient is learned for each of the attributes describing the examples along with an independent term. Linear regression performs well when there is a linear solution and the learned coefficients can be used to determine the relevance of the attributes, as long as the attributes have been previously scaled to the same scale [42], which is recommended. Obviously, if the problem has a nonlinear character, linear regression performs worse. Within Scikit Learn, linear regression is implemented under the name LinearRegression.

2.3.3. *Decision tree*

Decision tree is an algorithm that performs successive splits of the input space until the examples are grouped into clusters that can share the same prediction [43]. At each step, the attribute that best discriminates the examples is selected and a thresholded split is applied to it, resulting in a new branch in the tree. A decision tree usually gives good results and is able to adapt to nonlinear problems [43]. Another advantage of a decision tree is that the trained model is human understandable, so in addition to a prediction, there will be register of parameters and their values used by the algorithm to make successive splits. Fig. 2 shows an example of a decision tree. The decision tree recursively splits the input space until it is able to find groups of examples for which it can predict the same value. In the figure, the decision tree will predict the value 0 if a example has a value less than or equal to 0.402 in attribute ×1 and a value greater than 0.708 in attribute x2. In the Scikit Learn library, the variant of decision trees for regression is implemented under the name DecisionTreeRegressor.

2.3.4. *Ensemble methods*

Combining several models can lead to a better model, this is the principle of ensemble methods [44]. The two most popular ensemble methods are bagging and boosting.

2.3.4.1. Bagging. Bagging consists of randomly creating N training sets of equal size from the original dataset (using sampling with replacement). Then, a model is trained separately from each of these training sets, so that there are as many models as there are training subsets [45]. The prediction for a new example is computed by averaging the predictions of all the models. The most representative algorithm of this type is Random Forest [46], which uses decision trees as the base algorithm to train from the N training sets. In Scikit Learn, the implementation is called RandomForestRegressor.

2.3.4.2. Boosting. Boosting always uses the original dataset but performs model training using weak learners. Weak learners are algorithms that are not allowed to learn as much as they could learn from the data by introducing some limiting condition [47]. An example of a weak learner is a decision tree that is not allowed to generate more than two branches. The predictions of this weak learner will be strong for some examples in the training set and poor for others. Thus, a weight will be assigned to each of the examples in the training set, so that the next iteration will focus on the poorly predicted samples. This process will be repeated as many times as necessary. To obtain the prediction of a new example, all models are applied and the average is calculated. The most representative algorithm of this type is XGBoost [48], which uses restricted decision trees as weak learners. In scikit Learn, the implementation is called XGBRegressor.

2.3.5. *Support vector machine (SVM)*

SVM [49] has the ability to adapt to any type of problem by using different strategies or kernels. SVM for regression searches for the hyperplane that best fits the examples in a continuous space. This is achieved by projecting the examples into a higher dimensional space using kernel functions, and in that space will attempt to minimize the error while maximizing the margin [50]. The polynomial kernel and the radial basis function (RBF) kernel are the most popular nonlinear kernels. In Scikit Learn, the implementation of support vector machines is called SVR.

2.3.6. *Artificial neural network (ANN)*

ANN [51] is very popular because it is able to adapt to any type of problem no matter how difficult it may be, which means that it is difficult to train, and takes the risk of reaching overfitting situations [52]. The larger the training dataset, the lower the risk of overfitting.

ANNs have a layered architecture where the number of layers, the number of neurons in each layer and the activation functions are modified to obtain different models. All ANNs must have an input layer, an output layer and the number of intermediate layers can vary. Fig. 3 shows a diagram of one possible architecture for a 2-hidden layers ANN. The neural networks update the weights of the neurons in their different layers using a back-propagation mechanism based on the errors obtained. Within Scikit Learn ANNs for regression problems are implemented under the name MLPRegressor.

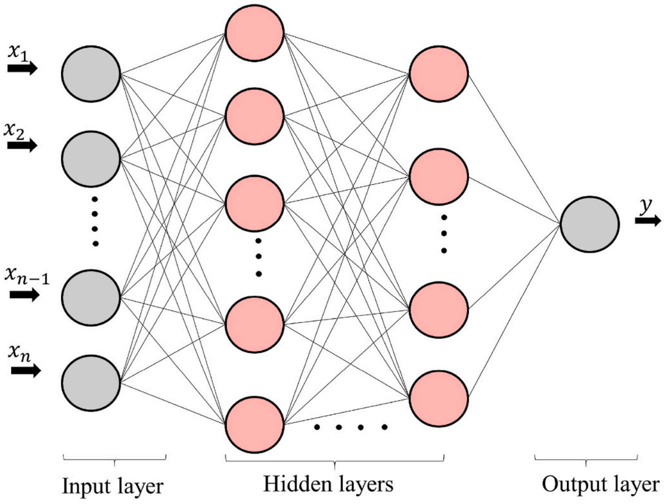


Fig. 3. Example of a 2-hidden layers ANN architecture. x_n represents the input variables and y the target variable.

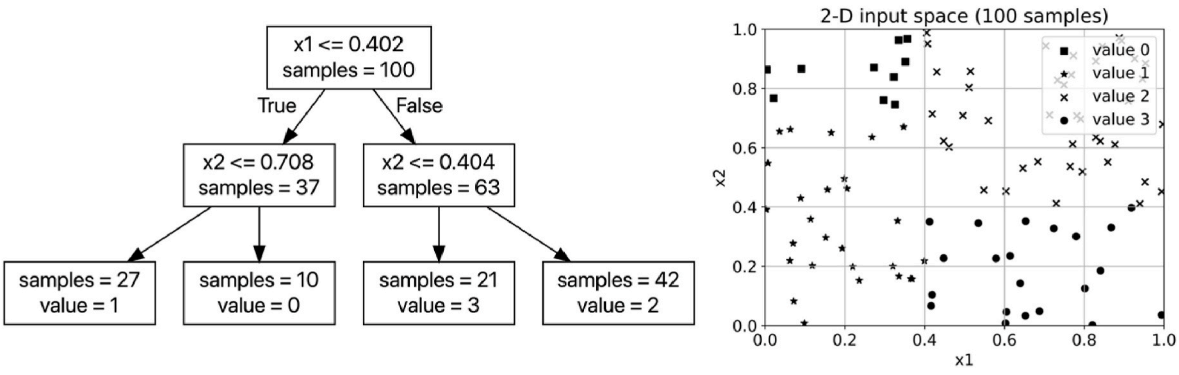


Fig. 2. Example of a decision tree.

2.4. *Performance of a model*

Having introduced the main regression algorithms, let's see how these algorithms should be trained and how the estimate of their performance is calculated. To calculate this estimate, some of the biodiesel in the database must be reserved for evaluation (test set), that is, they cannot be used to train the different algorithms [53]. The method used to reserve examples for testing can affect the estimation of model performance. This has led to the development of different methods, among which cross-validation and leave-one-out stand out the hold-out method (also known as the train/test experiment), which is strongly influenced by the examples that have been selected for the test set [53].

2.4.1. *Cross validation*

In a cross-validation, the data are randomly divided into N partitions and N models are trained. In each model, the training set is formed by N-1 partitions, and the excluded partition is the test set. The performance estimate is obtained by calculating the average performance of the N models [54]. Fig. 4 shows an example of 5-fold cross-validation. As can be seen in the figure, in a 5-fold cross-validation, the data set is divided into 5 partitions of equal size and 5 models are trained. Each of these 5 models is trained using the samples contained in 4 folds, leaving the remaining fold to evaluate the performance of the model. Each model uses a different fold as a test set, so all examples are used once to calculate performance. In addition, all examples are used N-1 times as part of the training set.

Cross-validation is a widely used method because all examples are part of the test set in some of the training and because the number of models trained equals the number of partitions generated, which allows to control the time spent training the models [54].

2.4.2. *Leave-one-out*

If the number of partitions in the cross-validation is the same as the number of examples in the dataset, this is called a leave-one-out experiment. This experiment is the best for evaluating the performance of a model, since each model is trained on all available examples except one (which is used as a test set) [55]. However, it is computationally very expensive because it requires training as many models as there are examples in the dataset.

This method represents an improvement in terms of the model's learning ability and error generalization when compared to the traditional train/test or hold-out approach. In the traditional approach, the results are heavily influenced by the examples selected for the test set when the data is scarce [53].

Considering the number of biodiesel examples available in the database and the current computing power of computers, the leave-one-out method is the most appropriate to estimate the performance of the models.

2.5. *Hyperparameter optimization*

Some algorithms have configurable mechanisms that allow them to work in different ways, such as the number of hidden layers or the activation function in an ANN. These mechanisms, known as hyperparameters, cause learned models to behave differently and, in many cases, need to be tuned in order for the models to perform well [56].

Hyperparameter tuning must be performed only on the data used to train the model, not on the data used to estimate model performance. Otherwise, the performance estimation will be optimistic. This means that, for example, in a train/test experiment, only the examples in the training set can be used for hyperparameter tuning [57]. The simplest solution in this case is to split the training set into two new subsets, which are usually called the training set (about 70 or 80% of the samples) and the validation set. This new training set will be used to train models with different hyperparameter settings that will be applied to the validation set.

This simple solution is strongly influenced by the examples selected for the validation set, so instead of performing a training/validation split for the hyperparameter search, it is advisable to perform another cross-validation using only the examples that were in the initial training set [56].

The number of hyperparameter combinations that can be tested for an algorithm can be very large, so the combinations are often sampled in an organised fashion, with sampling using a *grid search* strategy being very popular [57]. The hyperparameter configuration chosen will be that of the model that performs best.

2.6. *Normalization of the data*

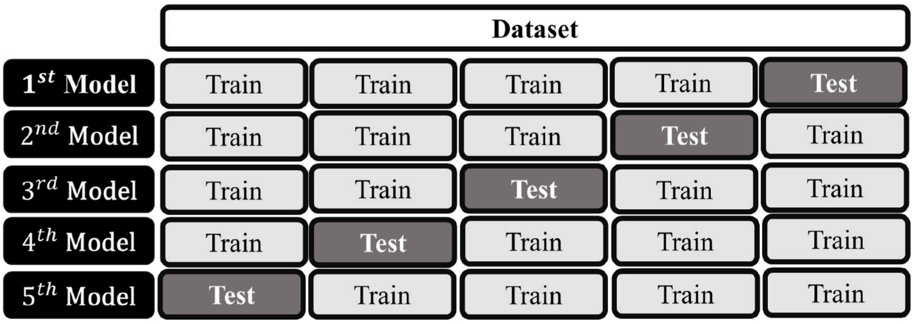
When the attributes used to describe the examples have different scales, the training of some algorithms becomes more difficult. To achieve a better performance, the attributes can be standardized before training the model for those algorithms that need it. The standardization follows equation (2) [58], where z is the value after the standardization, x is the value of the attribute to be standardized, $\bar{x}$ is the mean of the attribute, and σ is the standard deviation of the attribute.

$$z = \frac{x - \bar{x}}{\sigma} \tag{2}$$

2.7. *Performance metrics*

The performance of the different models presented in this paper was computed using the most common metrics in regression tasks [59]. These are the mean absolute error (MAE), the mean squared error (MSE) and the coefficient of determination (R^2), whose formulas are shown in the following equations:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$



	Dataset				
1^{st} Model	Train	Train	Train	Train	Test
2^{nd} Model	Train	Train	Train	Test	Train
3^{rd} Model	Train	Train	Test	Train	Train
4^{th} Model	Train	Test	Train	Train	Train
5^{th} Model	Test	Train	Train	Train	Train

Fig. 4. Conceptual diagram of a 5-fold cross-validation.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{5}$$

where n represents the number of samples, y_i is the actual value of the target variable, $\widehat{y}_i$ is the predicted value and $\overline{y}$ is the mean of the observed data. A perfect prediction will have an MAE and an MSE equal to 0, since they calculate the difference between the true value and the predicted value (in one case by calculating the absolute value and in the other by squaring the difference). If the prediction is not perfect, both the MAE and the MSE will have values greater than 0, with no known maximum value for the error, which sometimes makes it difficult to interpret the error obtained. The coefficient of determination solves this problem as its values are usually bounded between 0 and 1. If the prediction is perfect an R^2 of 1 will be obtained, worse predictions will have values of R^2 less than 1, obtaining 0 if the prediction is equivalent to predicting the observed average value.

2.8. Procedure to train a machine learning model

There are several steps involved in developing a predictive machine learning model. The first is to train the model and obtain an estimate of how the model will perform when presented with examples it did not see during the training phase. Fig. 5 illustrates this process.

1. From the raw data, the examples (biodiesels) that are useful for training the model and the appropriate attributes (the characteristics that describe the biodiesels) are selected, thus obtaining the data set.
2. From the data set, several divisions are made according to a cross-validation or leave-one-out strategy, so that there are several training and test sets. Since there are 241 examples of biodiesel, a leave-one-out experiment is chosen to train as many models as there are examples.
3. To train each of the models, hyperparameter optimization will be performed using only the examples contained in the corresponding training set. A grid search strategy will be used to obtain the different combinations of hyperparameters, and a 5-fold cross-validation is used to obtain the performance of each combination.
4. Each of the trained models is evaluated with its corresponding test set and the performance of each model is averaged to obtain the estimated performance for that algorithm.

Finally, a model (with the obtained hyperparameters) can be trained using the entire data set. This model, which will be used in the future for prediction based on new biodiesels, will have the estimated performance previously obtained.

3. Results and discussion

The experiments described in section 2 were performed on the 241 biodiesel samples to analyze different aspects for the prediction of IV. All the results were obtained by performing leave-one-out experiments. The experiments analyzed: i) the performance of popular machine learning algorithms in predicting IV, ii) the quality of different subsets of attributes, iii) the performance of other state-of-the-art methods, and iv) the relevance of the selected attributes for predicting IV.

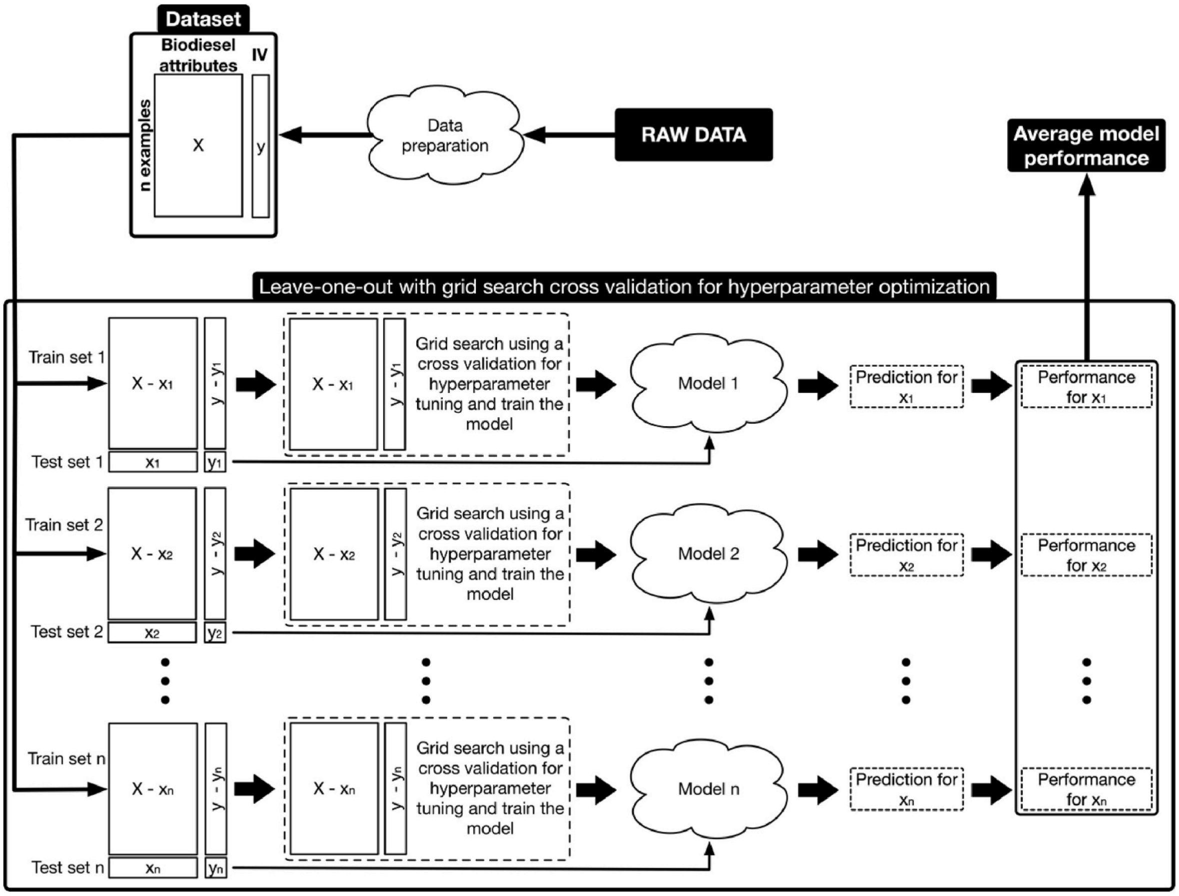


Fig. 5. Steps followed to obtain the performance of a model using leave-one-out and optimizing its hyperparameters through a grid search with cross-validation.

3.1. Attributes and algorithm performance

It is critical to select the attributes that can adequately model the problem because some attributes may be irrelevant, redundant, or even misleading [60]. Reducing the number of attributes speeds up the learning process and prevents the algorithm from being fooled by meaningless parameters. According to the literature and the experimental method for obtaining IV, the attribute that must have the most importance in the IV prediction is the number of double bonds. The performance of the tested algorithms, with different attribute configurations, is summarized in Table 2. As expected, all the algorithms are able to learn using only the DB attribute (combination 1), as shown by the comparison between the baseline (to predict the mean observed value of IV) and the other algorithms.

However, most of these algorithms performed better when making predictions based on the FAME distribution alone (combination 2). The poor performance of linear regression when using FAMEs is striking. This behavior is due to the fact that, as already seen in Fig. 1, the presence of some FAMEs is very rare (there is only one case), with which linear regression is not capable of learning an adequate coefficient for these FAMEs, making very large errors in the prediction for those biodiesels with rare FAMEs. The other algorithms are able to detect these situations and ignore these FAMES.

Since the DB number and FAME distribution are known to be important [26], they were used together in the combination 3. This combination of attributes is the one that provides the best results, with significantly better performance than the other combinations. All algorithms (except linear regression, as mentioned above) perform significantly better than the baseline. The worst performing models are the SVMs with linear and polynomial kernels, with an R^2 of 0.592 and 0.586 respectively, although this may be due to an inadequate choice of hyperparameters. Despite its simplicity, the decision tree performs well with an R^2 of 0.702, which is not significantly different from more complex models such as ANNs with an R^2 of 0.740. The three algorithms with the best results are SVM with an RBF kernel, RandomForest, and XGBoost, with R^2 values of 0.770, 0.772, and 0.784, respectively.

In addition, an attempt was made to use the DB and other common attributes to characterize biodiesel (combination 4). The results obtained using this combination of attributes are slightly better than those obtained using double bonds alone, but they are worse than those obtained using FAMEs explicitly. From these results it can be concluded that although there is a direct correlation between the number of double bonds and the iodine value since IV measures the amount of iodine needed to break the double bonds, DB is not the only parameter needed for prediction. Thus, not only the number of DB is important for predicting IV, but also the type of FAME in which these DBs are located.

3.2. Comparison of the proposed methodology with previous reported ones

A comparison with 6 different methods presented in the literature was made measuring the performance of those proposals using the dataset and the methodology presented in this work.

- **System 1**. This system was proposed by Azam et al. [35] and it is an empirical correlation which follows equation (6). In this formula, D represents the number of double bonds, A_i the percentage of the i-th FAME and Mw_i and the molecular weight of the i-th FAME. In the dataset the product of D multiplied by A_i is equal to DB × 100 in the dataset used in this work.

$$IV = \sum_{i=1}^{n} \frac{(254 \bullet D \bullet A_i)}{Mw_i} \tag{6}$$

- **System 2**. This system was proposed by Wang et al. [26] and follows equation (7). In this expression DU represents the average number of double bonds in the biodiesel as a percentage. DU is equal to DB × 100 in the dataset used in this work.

$$IV = 0.6683 \bullet DU + 25.0364 \tag{7}$$

- **System 3.** This system that follows equation (8) was proposed by Knothe [61]. In this expression D represents the number of double bonds, A_i the ratio of the i-th FAME and $\mathrm{Mw_i}$ and the molecular weight of the i-th FAME. In the dataset the product of D multiplied by A_i is equal to the DB used in this work.

$$IV = \sum_{i=1}^{n} 100 \bullet \frac{(253.81 \bullet D \bullet A_i)}{Mw_i} \tag{8}$$

- **System 4.** This system was proposed by Hoekman et al. [20] which follows equation (9). In this expression DU represents the average number of double bonds in the biodiesel as a percentage. It is equivalent to the DB attribute in the dataset used in this work.

$$IV = 74.373 \bullet DU + 12.71 \tag{9}$$

- **System 5.** This system was proposed by Oliveira et al. [23] and it is an Artificial Neural Network that proposes an architecture of 13 neurons in the input layer, 24 neurons in the first hidden layer with the hyperbolic tangent as activation function, 6 neurons in the second hidden layer with the logistic activation function, and one neuron for the output layer. The neural network was trained using the 13 attributes proposed by Oliveira et al.: C8:0, C10:0, C12:0, C14:0, C16:0, C18:0, C18:1, C18:1 OH, C18:2, C18:3, C20:0, C20:1 and C22:1.
- **System 6.** This system was proposed by Mostafaei [38] and is an Adaptive Neuro-Fuzzy Interference System (ANFIS). The ANFIS system is a neural network in which the number of neurons, the number of layers, and their activation functions are predefined. In

Table 2
Performance of the algorithms with different combination of attributes.

Algorithm	Combination 1			Combination 2			Combination 3			Combination 4		
	MAE	MSE	R^2	MAE	MSE	R^2	MAE	MSE	R^2	MAE	MSE	R^2
Baseline	31.1	1513.9	−0.008	30.5	1446.1	−0.008	30.5	1446.1	−0.008	31.1	1513.9	−0.008
Linear Regression	13.4	497.0	0.669	2.53E+12	1.43E+27	-1E+24	6.82E+12	3.28E+27	−2.3E+24	12.5	465.2	0.690
Decision Tree	15.4	787.4	0.476	17.2	640.0	0.554	13.4	428.1	0.702	17.8	856.0	0.430
Random Forest	14.3	572.4	0.619	13.6	425.7	0.703	11.5	326.4	0.772	13.7	496.3	0.670
XGBoost	13.5	491.0	**0.673**	14.0	407.2	0.716	11.4	309.7	**0.784**	13.9	488.3	0.675
SVR-linear	12.2	514.1	0.658	13.6	726.9	0.493	12.7	585.3	0.592	12.6	481.0	0.680
SVR-POLY	12.2	514.1	0.658	14.4	663.5	0.537	13.2	593.6	0.586	12.0	461.1	0.693
SVR-RBF	12.3	513.2	0.658	11.6	340.6	**0.762**	11.2	329.7	0.770	12.1	458.6	**0.695**
ANN	13.6	499.0	0.668	11.9	365.6	0.745	12.2	372.5	0.740	16.1	580.9	0.613

Combination 1) Only Double Bound
Combination 2) Only FAME distribution
Combination 3) FAME and DB
Combination 4) SFA, MUFA, PUFA, Cn, Mw, DB and NPI

particular, ANFIS has a fuzzification layer and a defuzzification layer that scale the data to apply fuzzy logic rules. The authors proposed to use 3 attributes for training the neural network, which are DB, MUFA, and PUFA.

Since there is no available implementation of ANFIS algorithm in Scikit-Learn, the implementation available in MatLab within the Fuzzy Logic Toolbox was used.

Fig. 6 shows a comparison between the results of the best model XGBoost (MAE = 11.4 [g I_2/100 g], MSE = 309.7 and R^2 = 0.784) against the results of the six different systems using a leave-one-out experiment. As the results reported in the previous section, correlations that depend entirely or strongly on the number of double bonds cannot adequately generalize a prediction. The model proposed in this paper shows an improvement of about 11.6% in MAE and 65.1% in MSE compared to the average of the empirical equations (systems 1 to 4). In addition, according to the MSE values, the use of these equations leads to larger failures when they occur.

The model presented in this study shows an improvement of 50.9% and 28.1% in MAE, 113.9% and 318.5% in MSE and 28.5% and 48.3% in R^2 with respect to system 5 [23] and system 6 [38], respectively. The results obtained with these systems are worse than those published by their authors in their articles. This may be due to the type of experiment they performed (in this article, leave-one-out was used for the reasons explained in section 2.2.7.2) or to the number of characteristics of the biodiesels used (the dataset of the current work includes 266 biodiesel examples from more than 100 different feedstocks, spanning the years 2002–2022).

3.3. Attributes relevance

Based on the results explained in the previous sections, a final model was trained with the XGBoost algorithm using all the data from the dataset. After tuning its hyperparameters, the XGBoost was trained with 100 estimators, a learning rate of 0.1, and using decision trees with a maximum depth of 2 (the meaning of these hyperparameters is explained in Ref. [48]).

The relevance of the attributes to the algorithm is shown in Fig. 7. As expected, in Combination 3 the DB parameter is essential, accounting for 70% of the prediction, while the distribution of FAMEs accounts for the remaining 30%. The right side of the figure shows the most important FAMEs for the prediction, representing only those whose significance exceeds 1%, while the rest are grouped under “Other FAME”. In Combination 2, by removing the DB attribute, the algorithm must extract the information about the number of double bonds from the remaining attributes. This is the reason why C18:1 and C18:2 become very important as these attributes appear in almost every biodiesel (see Fig. 1) and inform about the presence of double bonds. The rest of the FAMEs that are important and add information to the Combination 3 also appear in the Combination 2. For the Combination 4, as expected, the DB and the FSA/MUFA/PUFA account for the 88% of the attribute relevance, which means that little information can be extracted from NPI, Mw and Cn to predict the IV value.

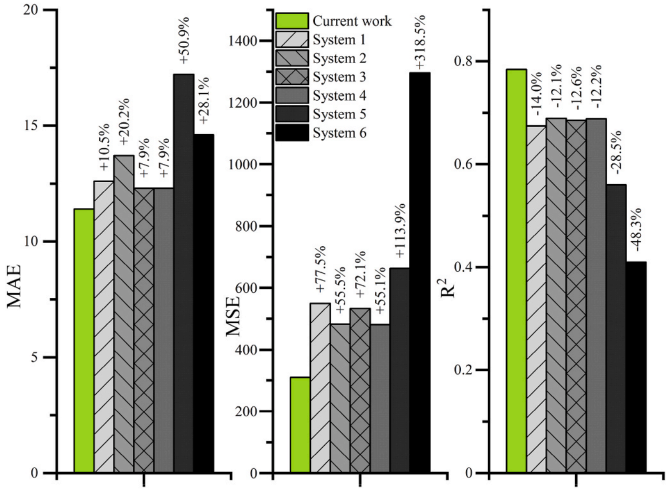


Fig. 6. Comparison between the current work and other IV prediction methods.

These artificial intelligence techniques could be a breakthrough for the production biodiesel industry. If different models for predicting biodiesel properties were ensembled together, they could serve as a feedstock selection tool. In this way, only those feedstocks for which the model predicts good properties would proceed to development, saving time and money in performing all the characterization tests.

On the other hand, although these systems allow us to obtain a relevance of the attributes in the prediction and thus to know what to pay attention to, further work should be done on their interpretability. Therefore, they should be combined with physicochemical laws to form hybrid systems [62], where it is not only known which attributes have the most information about the problem, but also how they affect the problem. In this sense we know that the IV is related to the number of double bonds of the FAMEs, but it is not yet known with certainty how it is affected by other features such as the number of double bonds in the same chain, or the length of the chain.

4. Conclusions

A total of 266 biodiesels were used for this study, covering samples from the years 2002–2022. Several algorithms were tested to predict the Iodine Value (IV) property. Different combinations of attributes were used, and the following conclusions were reached.

- Machine learning models for predicting biodiesel properties need to be trained on a large number and variety of examples (e.g., biodiesel from different generations and with different FAME distributions than typical) to better predict and generalize.
- The XGBoost algorithm provides the best results with the biodiesels included in our dataset. Although SVR with RBF kernel has a slightly smaller error in MAE, XGBoost is the algorithm with the highest R^2 and lowest MSE.
- The use of ANNs is not recommended due to the risk of overfitting considering the currently available data. These models would require a larger dataset to perform adequately. In addition, algorithms such as ANFIS, which cannot work with more than 4 or 5 attributes, are also discouraged as they cannot cover the full spectrum of FAMEs.
- With the currently available data, the hold-out approach of dividing the dataset into train/validate/test is also discouraged. Due to the limited number of examples, the performance of the algorithm depends on which examples form each group. Therefore, the use of a leave-one-out experiment is recommended.
- Although the number of double bonds is the most important parameter for predicting IV, the results are much better when the distribution of FAMEs is also included as input data.

CRediT authorship contribution statement

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

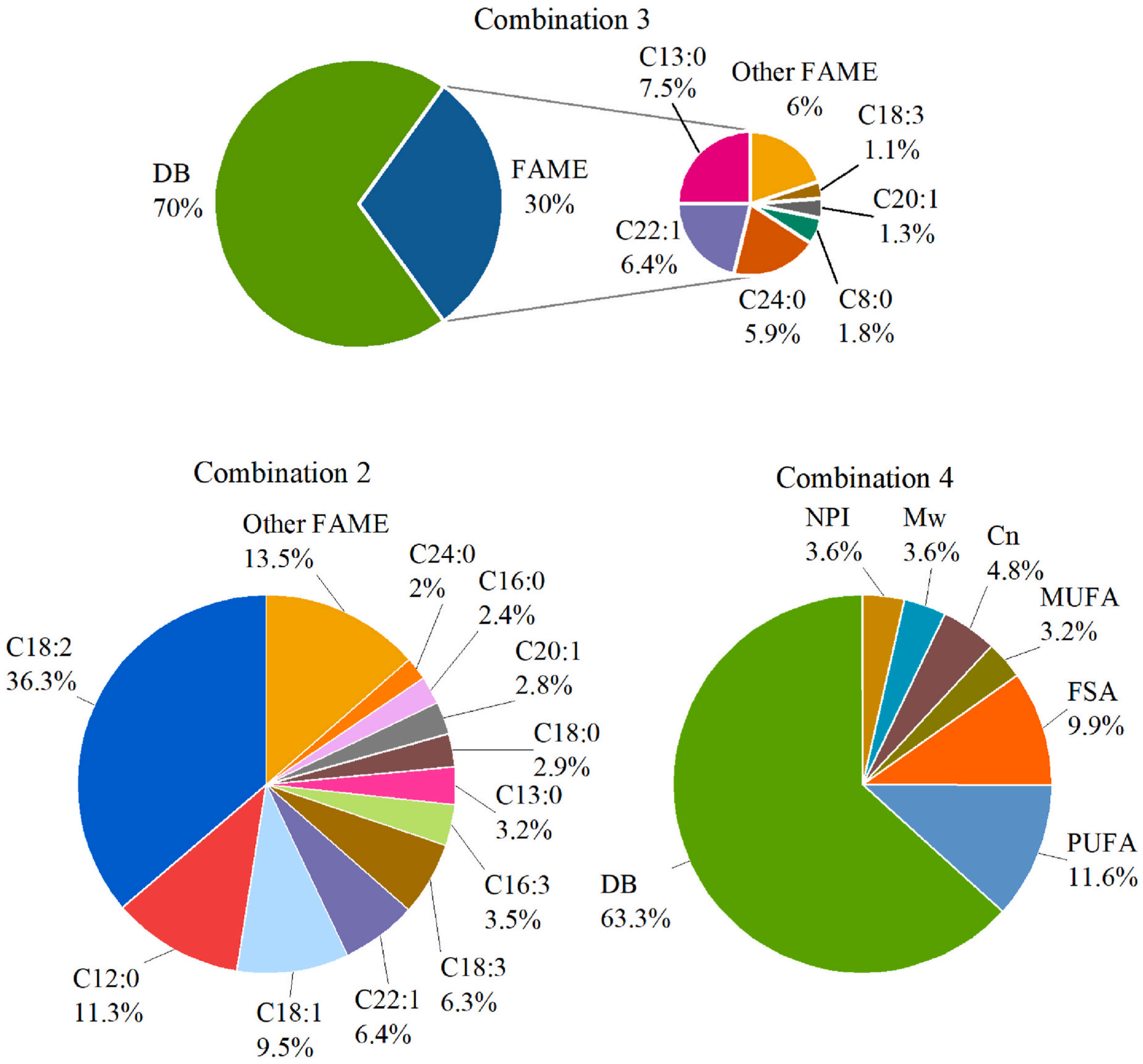


Fig. 7. Relevance of the attributes.

Declaration of competing interest

No conflict of interest is existing.

Data availability

Data will be made available on request.

Acknowledgements

This research was funded by the Ministry of Science, Innovation and Universities (Spain) and the State Investigation Agency, grant number: PID2022-136656NB-100 (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 at the University of Oviedo (Spain), grant number: SV-PA-21-AYUD/2021/50987. Guillermo Díez-Valbuena acknowledges the support of the Goverment of the Principality of Asturias under *Severo Ochoa* predoctoral program (ref. BP22-153).

Abbrevations

ANFIS	Artificial neuro-fuzzy interference system
ANN	Artificial neural network
ASME	American society of mechanical engineers
CN	Cetane number
EN	European norm
FAME	Fatty acid methyl ester
FAEE	Fatty acid ethyl ester
ICEs	Internal combustion engines
IV	Iodine value
LM	Levenberg-Marquardt
MUFA	Mono-unsaturated fatty acids
PUFA	Poly-unsaturated fatty acids
RBF	Radial basis function
SFA	Saturated fatty acids
SVM	Support vector machine
SVR	Support vector regressor
UN	United Nations

Nomenclature

Ca	Calcium
CO_2	Carbon dioxide
Cn	Carbon number
DB	Number of double bonds
I_2	Molecular iodine

K	Potassium
MAE	Mean absolute error
Mg	Magnesium
MSE	Mean squared error
Mw	Molecular weight
Na	Sodium
NPI	Non polarity index
NO_x	Nitrogen oxides
R^2	Coefficient of determination

References

[1] United Nations. Goal 7: Affordable and clean energy. United Nations Sustainable Development n.d. https://www.un.org/sustainabledevelopment/energy/(accessed January 8, 2024)..

[2] Malaquias ACT, Netto NAD, Filho FAR, Da Costa RBR, Langeani M, Baêta JGC. 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 2019;41:567. https://doi.org/10.1007/s40430-019-2076-1.

[3] Senecal PK, Leach F. Diversity in transportation: why a mix of propulsion technologies is the way forward for the future fleet. Results in Engineering 2019;4: 100060. https://doi.org/10.1016/j.rineng.2019.100060.

[4] Duarte Souza Alvarenga Santos N, Rückert Roso V, Teixeira Malaquias AC, Coelho Baêta JG. Internal combustion engines and biofuels: examining why this robust combination should not be ignored for future sustainable transportation. Renew Sustain Energy Rev 2021;148:111292. https://doi.org/10.1016/j.rser.2021.111292.

[5] Hosseinzadeh-Bandbafha H, Tan YH, Kansedo J, Mubarak NM, Liew RK, Yek PNY, et al. Assessing biodiesel production using palm kernel shell-derived sulfonated magnetic biochar from the life cycle assessment perspective. Energy 2023;282: 128758. https://doi.org/10.1016/j.energy.2023.128758.

[6] Badawy T, Mansour MS, Daabo AM, Abdel Aziz MM, Othman AA, Barsoum F, et al. Selection of second-generation crop for biodiesel extraction and testing its impact with nano additives on diesel engine performance and emissions. Energy 2021;237: 121605. https://doi.org/10.1016/j.energy.2021.121605.

[7] Chauhan BS, Kumar N, Cho HM. A study on the performance and emission of a diesel engine fueled with Jatropha biodiesel oil and its blends. Energy 2012;37: 616–22. https://doi.org/10.1016/j.energy.2011.10.043.

[8] Galadima A, Muraza O. Biodiesel production from algae by using heterogeneous catalysts: a critical review. Energy 2014;78:72–83. https://doi.org/10.1016/j.energy.2014.06.018.

[9] Al-Jabri H, Das P, Khan S, AbdulQuadir M, Thaher MI, Hoekman K, et al. A comparison of bio-crude oil production from five marine microalgae – using life cycle analysis. Energy 2022;251:123954. https://doi.org/10.1016/j.energy.2022.123954.

[10] Dutta K, Daverey A, Lin J-G. Evolution retrospective for alternative fuels: first to fourth generation. Renew Energy 2014;69:114–22. https://doi.org/10.1016/j.renene.2014.02.044.

[11] Aro E-M. From first generation biofuels to advanced solar biofuels. Ambio 2016;45 (Suppl 1):S24–31. https://doi.org/10.1007/s13280-015-0730-0.

[12] Shelare SD, Belkhode PN, Nikam KC, Jathar LD, Shahapurkar K, Soudagar MEM, et al. Biofuels for a sustainable future: examining the role of nano-additives, economics, policy, internet of things, artificial intelligence and machine learning technology in biodiesel production. Energy 2023;282:128874. https://doi.org/10.1016/j.energy.2023.128874.

[13] 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.

[14] D02.E0. ASTM D6751-20a standard specification for biodiesel fuel blend stock (B100) for middle distillate fuels. 2023.

[15] Ahmad J, Awais M, Rashid U, Ngamcharussrivichai C, Raza Naqvi S, Ali I. A systematic and critical review on effective utilization of artificial intelligence for bio-diesel production techniques. Fuel 2023;338:127379. https://doi.org/10.1016/j.fuel.2022.127379.

[16] Yang H, Ring Z, Briker Y, McLean N, Friesen W, Fairbridge C. Neural network prediction of cetane number and density of diesel fuel from its chemical composition determined by LC and GC–MS. Fuel 2002;81:65–74. https://doi.org/10.1016/S0016-2361(01)00121-1.

[17] Ramírez-Verduzco LF, Rodríguez-Rodríguez JE, Jaramillo-Jacob A del R. Predicting cetane number, kinematic viscosity, density and higher heating value of biodiesel from its fatty acid methyl ester composition. Fuel 2012;91:102–11. https://doi.org/10.1016/j.fuel.2011.06.070.

[18] Sorate KA, Bhale PV. Biodiesel properties and automotive system compatibility issues. Renew Sustain Energy Rev 2015;41:777–98. https://doi.org/10.1016/j.rser.2014.08.079.

[19] Sarin A. Biodiesel: production and properties. Royal Society of Chemistry; 2012.

[20] Hoekman SK, Broch A, Robbins C, Ceniceros E, Natarajan M. Review of biodiesel composition, properties, and specifications. Renew Sustain Energy Rev 2012;16: 143–69. https://doi.org/10.1016/j.rser.2011.07.143.

[21] Jain S, Sharma MP. Stability of biodiesel and its blends: a review. Renew Sustain Energy Rev 2010;14:667–78. https://doi.org/10.1016/j.rser.2009.10.011.

[22] Mittelbach M. Diesel fuel derived from vegetable oils, VI: specifications and quality control of biodiesel. Bioresour Technol 1996;56:7–11. https://doi.org/10.1016/0960-8524(95)00172-7.

[23] Barradas Filho AO, Barros AKD, Labidi S, Viegas IMA, Marques DB, Romariz ARS, et al. Application of artificial neural networks to predict viscosity, iodine value and induction period of biodiesel focused on the study of oxidative stability. Fuel 2015; 145:127–35. https://doi.org/10.1016/j.fuel.2014.12.016.

[24] Gopinath A, Puhan S, Nagarajan G. Theoretical modeling of iodine value and saponification value of biodiesel fuels from their fatty acid composition. Renew Energy 2009;34:1806–11. https://doi.org/10.1016/j.renene.2008.11.023.

[25] Alviso D, Artana G, Duriez T. Prediction of biodiesel physico-chemical properties from its fatty acid composition using genetic programming. Fuel 2020;264:116844. https://doi.org/10.1016/j.fuel.2019.116844.

[26] Wang L, Yu H, He X, Liu R. Influence of fatty acid composition of woody biodiesel plants on the fuel properties. J Fuel Chem Technol 2012;40:397–404. https://doi.org/10.1016/S1872-5813(12)60018-8.

[27] Kalayasiri P, Jeyashoke N, Krisnangkura K. Survey of seed oils for use as diesel fuels. J Am Oil Chem Soc 1996;73:471–4. https://doi.org/10.1007/BF02523921.

[28] Mahesh B. Machine Learning Algorithms - A Review 2018;9.

[29] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44. https://doi.org/10.1038/nature14539.

[30] Liu Z, Karimi IA. Gas turbine performance prediction via machine learning. Energy 2020;192:116627. https://doi.org/10.1016/j.energy.2019.116627.

[31] Ahmad T, Huanxin C, Zhang D, Zhang H. Smart energy forecasting strategy with four machine learning models for climate-sensitive and non-climate sensitive conditions. Energy 2020;198:117283. https://doi.org/10.1016/j.energy.2020.117283.

[32] Thangaraja J, Zigan L, Rajkumar S. A machine learning framework for evaluating the biodiesel properties for accurate modeling of spray and combustion processes. Fuel 2023;334:126573. https://doi.org/10.1016/j.fuel.2022.126573.

[33] Suvarna M, Jahirul MI, Aaron-Yeap WH, Augustine CV, Umesh A, Rasul MG, et al. Predicting biodiesel properties and its optimal fatty acid profile via explainable machine learning. Renew Energy 2022;189:245–58. https://doi.org/10.1016/j.renene.2022.02.124.

[34] Ghiasi MM, Mohammadzadeh O, Zendehboudi S. Reliable connectionist tools to determine biodiesel cetane number based on fatty acids methyl esters content. Energy Convers Manag 2022;264:115601. https://doi.org/10.1016/j.enconman.2022.115601.

[35] Mohibbe Azam M, Waris A, Nahar NM. Prospects and potential of fatty acid methyl esters of some non-traditional seed oils for use as biodiesel in India. Biomass Bioenergy 2005;29:293–302. https://doi.org/10.1016/j.biombioe.2005.05.001.

[36] Bukkarapu KR, Krishnasamy A. A critical review on available models to predict engine fuel properties of biodiesel. Renew Sustain Energy Rev 2022;155:111925. https://doi.org/10.1016/j.rser.2021.111925.

[37] Bachler C, Schober S, Mittelbach M. Simulated distillation for biofuel analysis. Energy Fuel 2010;24:2086–90. https://doi.org/10.1021/ef901295s.

[38] Mostafaei M. ANFIS models for prediction of biodiesel fuels cetane number using desirability function. Fuel 2018;216:665–72. https://doi.org/10.1016/j.fuel.2017.12.025.

[39] Giakoumis EG, Sarakatsanis CK. A comparative assessment of biodiesel cetane number predictive correlations based on fatty acid composition. Energies 2019;12: 422. https://doi.org/10.3390/en12030422.

[40] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. JMLR 2011;12:2825–30.

[41] Trenchevski A, Kalendar M, Gjoreski H, Efnusheva D. Prediction of air pollution concentration using weather data and regression models. 2020.

[42] Kenney JF, Keeping ES. Mathematics of statistics. vol. First. third ed. D. Van Nostrand Comppany, Inc.; 1962.

[43] Quinlan JR. Induction of decision trees. Mach Learn 1986;1:81–106. https://doi.org/10.1007/BF00116251.

[44] Dietterich TG. Ensemble methods in machine learning. In: Multiple classifier systems. Berlin, Heidelberg: Springer; 2000. p. 1–15. https://doi.org/10.1007/3-540-45014-9_1.

[45] Breiman L. Bagging predictors. Mach Learn 1996;24:123–40. https://doi.org/10.1007/BF00058655.

[46] Breiman L. Random forest. Mach Learn 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.

[47] Schapire RE. A brief introduction to boosting. Proceedings of the 16th international joint conference on Artificial intelligenceume 2. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1999. p. 1401–6.

[48] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco California USA: ACM; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.

[49] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97. https://doi.org/10.1007/BF00994018.

[50] Shawe-Taylor J, Cristianini N. Kernel methods for pattern analysis. Cambridge: Cambridge University Press; 2004. https://doi.org/10.1017/CBO9780511809682.

[51] Hertz J, Krogh A, Palmer RG. Introduction to the theory of neural computation. Redwood City, Calif: Addison-Wesley Pub. Co; 1991.

[52] Lawrence S, Giles CL. Overfitting and neural networks: conjugate gradient and backpropagation. Proceedings of the IEEE-INNS-ENNS international Joint conference on neural networks. IJCNN 2000. Neural Comput: New Challenges and Perspectives for the New Millennium 2000;1:114–9. https://doi.org/10.1109/IJCNN.2000.857823. vol.1.

[53] Blum A, Kalai A, Langford J. Beating the hold-out: bounds for K-fold and progressive cross-validation. Proceedings of the twelfth annual conference on Computational learning theory. Santa Cruz California USA: ACM; 1999. p. 203. https://doi.org/10.1145/307400.307439. 8.

[54] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection (Montreal) Canada 1995;2:1137–43.

[55] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning data mining, interference and prediction. 2009. New York: Springter.

[56] Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: theory and practice. Neurocomputing 2020;415:295–316. https://doi.org/10.1016/j.neucom.2020.07.061.

[57] Feurer M, Hutter F. Hyperparameter optimization. In: Hutter F, Kotthoff L, Vanschoren J, editors. Automated machine learning: methods, systems, challenges. Cham: Springer International Publishing; 2019. p. 3–33. https://doi.org/10.1007/978-3-030-05318-5_1.

[58] Kreyszig E. Advanced engineering mathematics. tenth ed. New York: Wiley; 1979. p. p880.

[59] Carpenter RG. Principles and procedures of statistics, with special reference to the biological sciences. Eugen Rev 1960;52:172–3.

[60] Kononenko I, Hong SJ. Attribute selection for modelling. Future Generat Comput Syst 1997;13:181–95. https://doi.org/10.1016/S0167-739X(97)81974-7.

[61] Knothe G. Structure indices in FA chemistry. How relevant is the iodine value? J Americ Oil Chem Soc 2002;79:847–54. https://doi.org/10.1007/s11746-002-0569-4.

[62] Zendehboudi S, Rezaei N, Lohi A. Applications of hybrid models in chemical, petroleum, and energy systems: a systematic review. Appl Energy 2018;228: 2539–66. https://doi.org/10.1016/j.apenergy.2018.06.051.