Physical Chemistry

Anton Soria-López, María García-Martí and Juan C. Mejuto*

Ionic surfactants critical micelle concentration modelling in water/organic solvent mixtures using random forest and support vector machine algorithms

https://doi.org/10.1515/tsd-2024-2636
Received November 1, 2024; accepted November 22, 2024;
published online December 11, 2024

Abstract: The physicochemical property of surfactants that is widely used to study their behavior is the critical micellar concentration (CMC). The value of this property is specific to each surfactant as it depends on a number of external factors and the chemical composition of the surfactant. This research focused on using two new machine learning approaches, Random Forest (RF) and Support Vector Machine (SVM), to predict the logarithmic CMC value of 10 ionic surfactants. The same database from the previous study (a total of 258 experimental cases) was used with the same input variables – those defining the mixture of the organic solvent-water: T, molecular weight, molar fraction and log P; and the chemical composition of the surfactant: number of atoms of each element of the surfactant – to develop the predictive models. The best RF and SVM models were then compared with the best ANN model developed in the previous study. According to the results, the normalized models were those that presented the lowest RMSE values in the validation phase. Finally, the two approaches proposed in this research are suitable tools, together with the ANN, for the prediction of CMC and as possible alternative methods to replace expensive experimental laboratory measurements.

Keywords: CMC; ionic surfactants; machine learning; prediction; random forest; support vector machine

***Corresponding author: Juan C. Mejuto**, Departamento de Química Física, Facultade de Ciencias, Universidade de Vigo, 32004 Ourense, Spain, E-mail: xmejuto@uvigo.es
Anton Soria-López, Departamento de Química Física, Facultade de Ciencias, Universidade de Vigo, 32004 Ourense, Spain, E-mail: anton.soria@uvigo.es
María García-Martí, Departamento de Química Analítica e Alimentaria, Facultade de Ciencias, Universidade de Vigo, 32004 Ourense, Spain, E-mail: maria.garcia.marti@uvigo.gal

1 Introduction

Surfactants are amphiphilic organic compounds that have hydrophobic groups on the tail and hydrophilic groups on the head.[1] These compounds are capable of reducing the surface tension between two immiscible phases, especially aqueous and oily substances.[2] As a consequence of the double affinity of these compounds, their stability is not maintained in either polar solvents or organic solvents. Therefore, for both affinities to work correctly, the polar solvent must surround the hydrophilic part of the compound, while the organic solvent must be in contact with its hydrophobic part. These conditions only occur between two immiscible phases.[2]

Ionic surfactants are a class of surfactants that includes anionic and cationic. Anionic surfactants ionize in water by acquiring a negative charge that allows them to bind to positively charged particles.[3,4] This type of surfactants is the most widely used compared to cationic and other types due to its ease of production and low cost of manufacture. Additionally, they are effective in removing clay, dirt, and soil stains.[4] Examples of these surfactants include sodium dodecysulfate (SDS) or sodium-N-lauroylsarcosinate (SDDS). Regarding cationic surfactants, they contain a positively charged head group.[5] They are mainly used as antistatic agents for hair conditioners.[4] However, they are rarely used in detergents as they tend to absorb in a high rate into soil without being released, limiting their effectiveness.[4,6] Examples of these cationic surfactants include cetylpyridinium chrolide (CPyCl) or tetradecyltrimethyl ammonium bromide (TTAB).

At a certain concentration of surfactant, known as the critical micelle concentration (CMC), micelles are formed.[7] In other words, when the surface boundary in an aqueous solution is saturated, surfactants promote a molecular organization (micelles) to stabilize the system. Micelles consist of aggregates whose interior is hydrophobic while the exterior is hydrophilic.[8] Each surfactant, at specific

temperature and electrolyte concentration, has a characteristic CMC value.[9] According to Perinelli et al.[10] CMC is influenced by the hydrophobicity of the amphiphilic, hydrophobic tail length, as well as by the properties of the solutions. Due theses particularities, the size and shape of the micelles can be adjusted by modifying the concentration and structure of the surfactant, the temperature, the properties of the solvent, and other factors.[9]

Micellization kinetics plays a significant role in several technological applications.[11] Indeed, there are notable changes in different physical and chemical properties of the solution such as viscosity, surface tension and reactivity, occur when the CMC is reached.[11] Therefore, the CMC can be measured and analysing using different experimental methods including the surface tension method, the conductivity method, and fluorescence spectrophotometry.[12] However, these methods have several drawbacks as high expensive cost and time-consuming.[13] In this sense, the use of mathematical methods and predictive models can be excellent alternatives in this field, since these approaches have proven to be good tools for optimize research in this area. In addition, these methods reduce the costs and the time required for experimental measurements.[13]

Machine learning (ML) is an artificial intelligence tool that employs algorithms to enable computers to learn complex relationships, both linear and non-linear, within large and diverse datasets to generate predictive models.[12,14] ML has been described as a significant promise to address complex data patterns due to its strong fitting capabilities.[15] Furthermore, empirical modelling can be a good alternative to traditional experimental measurements in laboratories,[13] so numerous studies can be found in the literature have employed machine learning methods to model the properties of surfactant properties.[13,16,17]

Recently, our research group analyzed the efficiency of machine learning methodologies for artificial neural network (ANN) models to predict CMC values of surfactants in solvent organics-in-water systems.[18] Our findings demonstrate that these predictive models are an excellent alternative to the traditional experimental approach. Furthermore, these tools allow for a deeper understanding of surfactant properties, such as critical micelle concentration, since by simulating the different conditions to predict their modelling, more complex relationships can be uncovered that are not evident when studied through traditional experimental techniques. Despite that, further research is needed that addresses these studies using different ML methodologies to compare the results. Accordingly, there are other algorithms such as Random Forest (RF) – developed from a combination of multiple decision trees and the where the final prediction is determined by majority vote (for classification) or by the average of the results of each individual tree (for regression) – and Support Vector Machine (SVM) – model that aims to find an optimal hyperplane that maximizes the separation between two classes (for classification) or the hyperplane that maximizes the epsilon distance (for regression) – that are widely known.[19–21] In fact, these two algorithms offer several advantages over other machine learning models such an ANNs. For example, RF models avoid overfitting, while SVM models are able to adequately handle binary, categorical and numerical targets, are flexible, are robust with small datasets and provide a unique solution.[22]

Therefore, this research aims to develop two additional ML tools, Random Forest (RF) and Support Vector Machine (SVM) algorithms, to model the logarithmic value of the CMCs for the 10 ionic surfactants mentioned in our previous research.[18] This study focuses on analysing the performance of these two machine learning-based approximation approaches and compares them with the best ANN model developed in the previous study. Finally, it is discussed whether these alternatives are also appropriate tools to solve this type of problem.

2 Materials and methods

2.1 Experimental dataset

The database used in this research was extracted from our previous study.[18] It consist of total of 258 experimental cases, with 12 input variables and one additional output variable to predict. The input variables are: molar fraction of organic solvent in water (dimensionless), molecular weight of organic solvent (in g mol^{-1}), octanol-water partition coefficient of solvent defined as log P or log K_{WO} (dimensionless), number of C, H, Br, Cl, N, Na, O and S atoms, and solvent temperature in water when measuring CMC defined as T (in K). The output variable is the logarithm value of CMC (in mol L^{-1} or M).

According to Soria-López et al.[18] the organic solvents dissolved in water used to measure the CMC were four alcohols (ethanol, ethylene glycol, isopropanol and methanol) and acetone. The temperature range of the mixture of organic solvent in water oscillated between 298.15 K and 323.15 K. The database includes experimental data of all the input variables mentioned above, as well as the logarithmic value of CMC for a total of 10 ionic surfactants, 4 of which are anionic and 6 cationic. The anionic ones are sodium deoxycholate (SDC), sodium dodecylbenzenesulphonate (SDBS), sodium dodecylsulfate (SDS) and sodium *N*-lauroylsarcosinate (SDDS), while the cationic ones are

benzyldodecyldimethyl ammonium bromide (BDAB), cetylpyridinium chrolide (CPyCl), cetyltrimethylammonium bromide (CTAB), docecylpyridinium chloride (DPC), dodecyltrimethyl ammonium bromide (DTAB) and tetradecyltrimethyl ammonium bromide (TTAB).

2.2 Data division

The previous step to the implementation of the algorithms to create different predictive models of the CMC values is the data division. According to Ishola et al.[23] the predictive models developed using machine learning have to be validated. This validation is used for the selection of an optimal internal validation model and for the evaluation of its generalized predictive performance or external validation.[23,24] In this research, the internal validation consists of a training group (T) and a validation group (V), while the test group (Z), in this case, is the external validation. Taking this into account, the database was divided into three large groups randomly. The first group, training (T), constitutes 70 % of the database to develop different models. The second group, validation (V), corresponds to 20 % of the database, and is used to find the best model from all the models developed in the training group. Finally, the third group, testing (Z), represents 10 % of the database and has the function of evaluating the model's performance with data that has not been used in the training group.

2.3 Machine learning models

As has been previously mentioned, two machine learning based algorithms were used to predict the CMC values of ionic surfactants: Random Forest (RF) and Support Vector Machine (SVM).

2.3.1 Random forest

The Random Forest (RF) is a class of ensemble learning algorithm first announced by Breiman[19] in 2001.[25] This algorithm has been widely used to deal with classification and regression problems.[26] This algorithm uses bootstrap aggregation, commonly known as bagging, to produce decision trees.[27]

According to Iranzad and Liu,[28] two important techniques are integrated in RF: bagging and random node splitting. Bagging consists of repeatedly choosing a random sample with replacement from the training data and fitting trees to these samples. This leads to trees that grow from different samples and are quite distinct from one another. Regarding to random node splitting, this technique allows for the consideration of only a random subset of features in each node split of the tree. This leads to uncorrelated trees by preventing features that are very strong predictors of response from being selected by many trees.[28] Accordingly, this algorithm overcome certain limitations of decision trees such as overfitting problems and their inability to store non-linear and non-balanced data.[29,30] The final prediction of the RF observation is executed by majority voting in classification problems or by averaging the results of all trees in regression problems[31] (Figure 1).

In this research, all RF models generated employed the following three hyper-parameters combinations: number of trees (from 1 to 200 in 199 steps, using a linear scale), maximum depth (from 1 to 200, using a linear scale) and pre-pruning (true and false). It is also worth mentioning that two RF models were normalized to a certain range to avoid any of the variables having a greater influence on other variables causing a disproportionate effect on model training. In this study, two normalization methods were applied in linear scale to both input and output variables: range transformation (denoted by the subscript R) which normalizes the data on a scale between –1 and 1, and Z-transformation (denoted by the subscript Z). The normalization process was first applied to the training and validation data and then to the test data. All results were then de-normalized for comparison with other models. Therefore, in this research, three approaches were developed for the RF models (RF, RF_R and RF_Z).

2.3.2 Support vector machine

Support Vector Machine (SVM) is supervised machine learning algorithm first introduced by Cortes and Vapnik[20] in 1995.[29] This algorithm is based on the principle of structural risk minimization which can lead to an improvement of the generalization capability and a decrease in the upper limit of the generalization error.[32] In addition, according to Gaye, Zhang, and Wulamu,[33] SVM work well with high-dimensional data. This algorithm is a binary classifier in which the main aim is to find an optimal hyperplane that shows a maximum margin between the feature vectors of all the data of the different classes (Figure 2).[34,35]

This algorithm is used for both linear and nonlinear regression and classification problems.[36] SVM uses the kernel technique to solve in the case of nonlinear problems.[37] According to Boualem et al.,[38] kernel functions are used to translate the input data into a higher dimensional feature space. The most studied kernel functions are the linear, the polynomial of degree d and the radial basis function (RBF).[38]

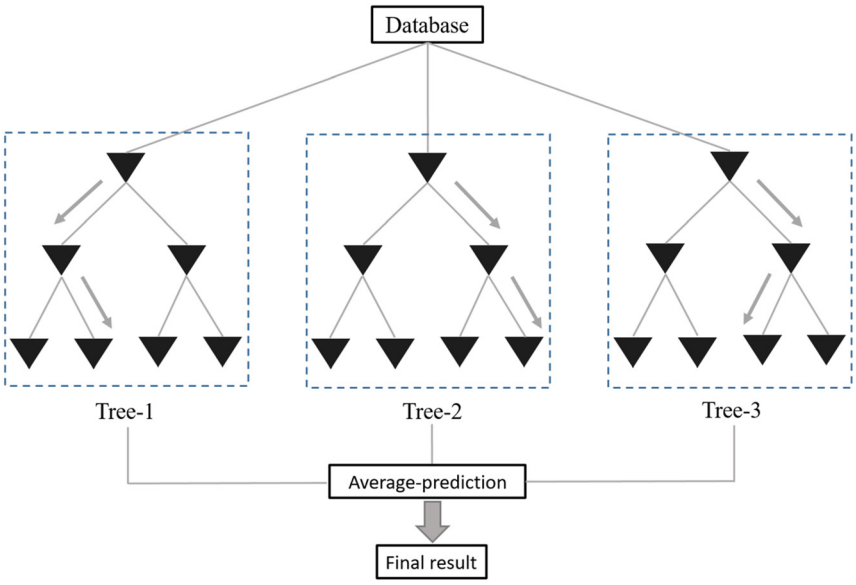

Figure 1: Set of individual trees that are part of the random forest for regression tasks. Inspired in Andrade Cruz et al.[30]

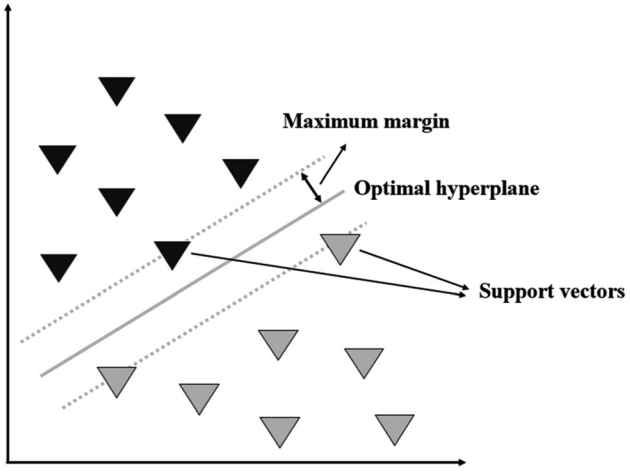

Figure 2: Support vector machine as a binary classifier. Inspired in Andrade Cruz et al.[30]

In this study, all SVM models were generated using the following four hyper-parameters combinations: SVM type, kernel type, *gamma* and *C*. The guide of Hsu et al.,[39] was used as a reference study to select the values of the *gamma* and *C* hyper-parameters. In addition, LibSVM, an SVM library proposed by Chang and Lin[40] was used.[41] The SVM models were generated using SVM type (*epsilon*-SVR and *nu*-SVR, both work on regression problems), kernel type (radial basis function, RBF), *gamma* (values between $9.5{\cdot}10^{-7}$ and 256 in 28 steps, in linear or logarithm scale) and *C* (values between $9.8{\cdot}10^{-4}$ and 1,048,576 in 30 steps, in linear or logarithm scale). The two normalization methods mentioned above were also applied in some SVM models on both in linear and logarithmic scales, described with the subscript L in this case, to both input and output variables: range transformation (between –1 and 1), and Z-transformation. The normalized results were then de-normalized in the same way as already mentioned in the RF models (RF_R and RF_Z). Therefore, six approaches were developed for the SVM models (SVM, SVM_L, SVM_R, $SVM_{R\text{-}L}$, SVM_Z and $SVM_{Z\text{-}L}$).

2.4 Metrics

The evaluation of the modelling fit and the prediction performance of the different algorithms were measured using the following three statistical parameters: root mean squared error (RMSE) (Eq. (1)), mean absolute percentage error (MAPE) (Eq. (2)) and the linear squared correlation coefficient (R^2) (Eq. (3)). All these parameters were calculated for all phases.

RMSE measures the error between two data. Its values range from 0 to ∞ and closer the value is to 0, the less error there is between two data. MAPE measures the percentage of absolute error. Its values range is also from 0 to ∞, being 0 % error when MAPE is equal to 0. Finally, the range of R^2 values is from –∞ to 1, with R^2 = 1 representing a perfect fit.[42] Therefore, at lower values of RMSE and MAPE and higher values of R^2, the predictive model shows a better fit.[43]

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}} \tag{1}$$

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{n}\left|\frac{x_i - y_i}{y_i}\right| \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(x_i - y_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{3}$$

where y and x represent the actual and predicted values, respectively. Furthermore, $\bar{y}$ is the mean of data y and n is the total number of the data.

2.5 Computational resource and software

The different RF and SVM models were developed using the RapidMiner Studio Educational 10.2.000 version (RapidMiner GmbH.). The computational equipment used were an Intel® Core™ i9-10900 K at 3.70 GHz with 64 GB RAM with Windows 11 Pro. The obtained data (258 experimental cases) were collected using Microsoft Excel 2013 (Microsoft). Figures 1 and 2 were made with Microsoft PowerPoint 2016 (Microsoft). The graphical representations (Figures 3, 5 and 7) and the scatter plots (Figure 4 and 6) were made with SigmaPlot 14.0 (Systat Software Inc.).

3 Results and discussion

In this section, the performance of the RF and SVM models in predicting the CMC values of several ionic surfactants is statistically described. Then, these models are compared with the best ANN model developed in the previous study to verify whether these two machine learning-based algorithms used in this research are suitable for CMC modelling.

The statistical parameters RMSE, MAPE and R^2 are used to evaluate the model fit and prediction performance. Their values provide statistical information about the behavior of the models. Finally, the RMSE values are taken into account for the selection of the best model for each algorithm studied in this research.

3.1 Random forest models

In the case of RF models, three different predictive models have been developed using the RF algorithm. Figure 3 presents the values of statistical parameters of RMSE, while Table 1 shows the values of the statistical parameters for MAPE and R^2 for each of the predictive models for the training, validation and testing phases.

According to the results, minimal differences were observed in the values of the three statistical parameters across all phases for the three RF models developed. For the RMSE values, the range oscillates between 0.069 M and 0.095 M (Figure 3). In contrast, the MAPE values ranged from 4.6 % to 6.5 %, while the R^2 values varied between 0.954 and 0.980 (Table 1).

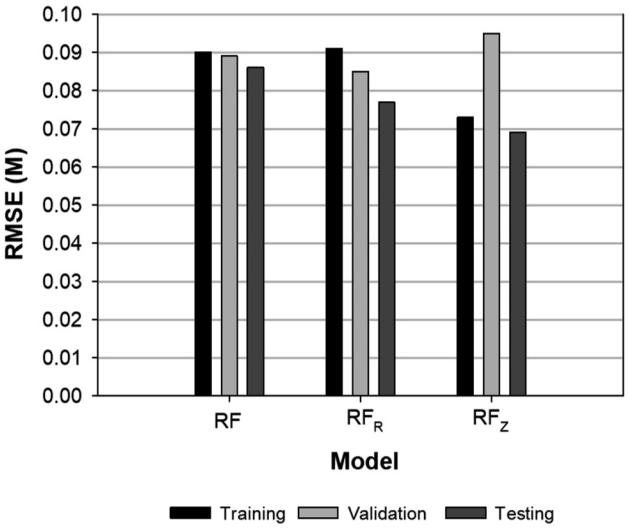

Figure 3: Graphical representation of the RMSE values for training, validation and testing for the RF models. RMSE is root mean square error (M).

Table 1: Values of the statistical parameters MAPE and R^2 obtained by each of the three RF models developed for all phases (training, validation and testing). MAPE is mean absolute percentage error (%) and R^2 is the linear squared correlation coefficient.

	Training		Validation		Testing	
Model	**MAPE**	**R^2**	**MAPE**	**R^2**	**MAPE**	**R^2**
RF	5.5	0.967	5.9	0.977	5.5	0.954
RF_R	6.1	0.966	6.2	0.980	5.6	0.964
RF_Z	5.4	0.979	6.5	0.979	4.6	0.970

Regarding the statistical performance of the models for the training phase, the RF_Z model demonstrated the lowest RMSE value (0.073 M) corresponding to a MAPE value of 5.4 %. On the other hand, the RF_R model showed the highest RMSE and MAPE value (0.091 M and 6.1 %). In the validation phase, the RF_R model achieved the lowest RMSE value (0.085 M), followed by the RF model (0.089 M) and the RF_Z model (0.095 M), corresponding to MAPE values of 6.2 %, 5.9 % and 6.5 %, respectively (Figure 3 and Table 1, respectively). The R^2 values were similar in the three models during both training and validation phases (Table 1). Finally, during testing phases, the RF_Z model showed the best performance, with the lowest RMSE and MAPE values (0.069 M and 4.6 %, respectively) and highest R^2 value (0.970) for the testing phase. The other two models were worse in terms of the three statistical parameters (Figure 3 and Table 1).

According to the previously described, it can be concluded that the three models performed adequately with

internal data (training and validation) and showed excellent generalized predictive performances with external data (testing), indicating that these predictive models did not present overfitting problems.

The best RF model was selected to further study its performance in CMC modelling of ionic surfactants. The selection criterion used for selection was the lowest RMSE value during the validation phase. Based on this criterion, the normalized model in the range –1 to 1 (RF_R) was selected, since it showed the lowest RMSE (0.085 M), although it did not show the lowest RMSE value for the testing phase. Therefore, the RF_R model was selected for further performance analysis.

Figure 4 illustrates the dispersion of the real CMC values versus those predicted by the RF_R model for the three phases. The dispersion of the CMC values for the internal data (training and validation cases) is shown in Figure 4A. According to this figure, it can be observed that, in general, the points are close to the red dashed line with the slope of 1. This indicates that the most cases exhibited low fitting errors, i.e., minimal dispersion between the real and predicted values. However, there are some cases that are noticeably far from the straight line, showing significant dispersion errors between real and predicted values. In relation to these special cases, it is worth mentioning three training cases with the highest fitting errors: (–0.50, –1.07), (–0.77, –1.28) and (–0.23, –0.61).

Regarding the first case (–0.50, –1.07), the model predicted a value of –1.07 M, while the actual value was –0.50 M, (model overestimation). This case showed the highest absolute error (0.572 M), since it is the point that is farthest from the slope line 1, and a significant absolute percentage error value (115.2 %). The second case (–0.77, –1.28) also presented an overestimation of the model (–1.28 M vs –0.77 M). In addition, it showed the second highest value of absolute error (0.511 M) and the third highest value of percentage error value (66.2 %). Finally, the third case (–0.23, –0.61), the model predicted a log CMC value of –0.61 M versus actual value (–0.23 M) leading a model overestimation of the model. Furthermore, this case showed an absolute error of 0.377 M and the highest value of absolute percentage error (160.7 %). In fact, if these three training cases are removed, the values of the three statistical parameters are significantly improved (RMSE = 0.070 M, MAPE = 4.6 %, and R^2 = 0.980).

Figure 4B illustrates the dispersion of CMC values for the external data in testing phase. According to this figure, it is observed that most of the cases are close to the slope line 1 (red dashed). However, two test cases with the highest error values are highlighted. In this sense, the first case (–1.36, –1.52), in which the model predicted a value of –1.52 M versus –1.36 M resulting in a model overestimation of 11.7 %, had the highest absolute error value (0.159 M). On the other hand, the second case (–0.84, –0.73) presented the highest absolute percentage error value (12.9 %) and the second highest absolute error value (0.108 M).

3.2 Support vector machine models

Regarding SVM models, six different predictive models have been developed using SVM algorithm. Figure 5 presents the values of statistical parameters of RMSE. On the other hand, statistical parameters for MAPE and R^2 are represented (Table 2).

Considerable differences were observed in the values of the statistical parameters RMSE (Figure 5) and MAPE (Table 2) for all phases between the SVM models developed using a linear scale and those using a logarithmic scale. In

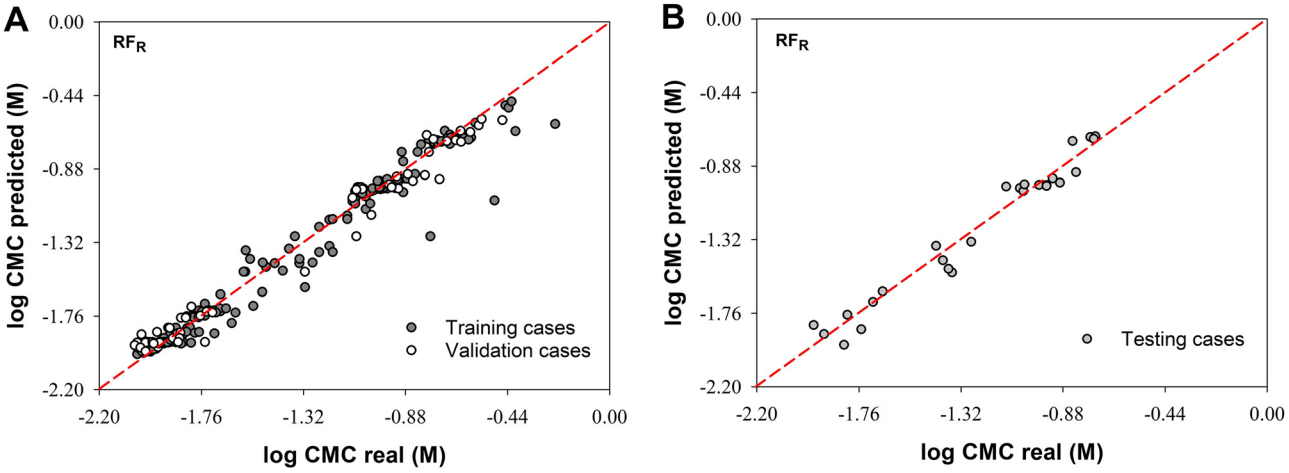

Figure 4: Scatter plots presenting actual and predicted values of log CMC real (*x*-axis) and log CMC predicted (*y*-axis) for the RF_R model in the training and validation phases (A) and testing phase (B). The red dashed line corresponds to the line with slope one.

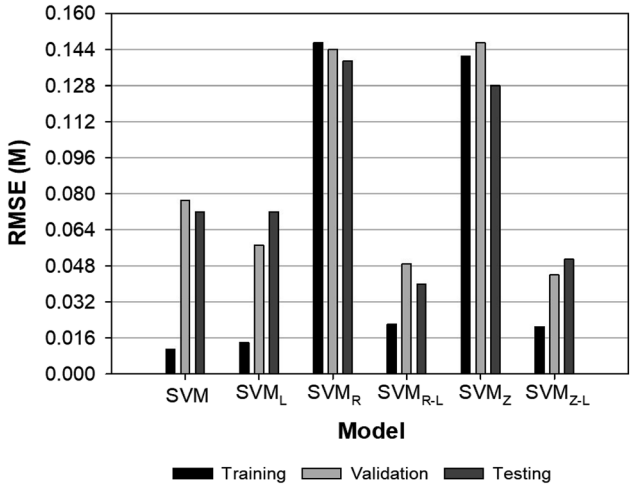

Figure 5: Graphical representation of the RMSE values for training, validation and testing for the SVM models. RMSE is root mean square error (M).

Table 2: Values of the statistical parameters MAPE and R^2 obtained by each of the six SVM models developed for all phases (training, validation and testing). MAPE is mean absolute percentage error (%) and R^2 is the linear squared correlation coefficient.

	Training		Validation		Testing	
Model	**MAPE**	**R^2**	**MAPE**	**R^2**	**MAPE**	**R^2**
SVM	0.2	0.999	5.1	0.983	3.8	0.970
SVM$_L$	0.4	0.999	3.1	0.991	3.4	0.969
SVM$_R$	9.5	0.913	8.7	0.933	7.9	0.922
SVM$_{R-L}$	0.6	0.998	2.9	0.993	2.1	0.991
SVM$_Z$	8.7	0.916	8.8	0.932	7.0	0.920
SVM$_{Z-L}$	0.6	0.998	2.6	0.994	2.6	0.985

this sense, the SVM$_R$ and SVM$_Z$ models showed the highest values of these two statistical parameters compared to the other SVM models.

Regarding the RMSE values, their range oscillated between 0.011 M and 0.147 M, and between 0.2 % and 9.5 % in the case of MAPE (Figure 5 and Table 2 respectively). In addition, significant variations were observed for R^2 values between different SVM models (between 0.913 and 0.999).

In terms of the statistical performance of the models for the training phase, the SVM$_R$ model presented the highest RMSE and MAPE values (0.147 M and 9.5 %, respectively), and the lowest R^2 value (0.913), while the SVM model presented the lowest RMSE and MAPE values (0.011 M and 0.2 %). The fits of the other SVM models (SVM$_L$, SVM$_{Z-L}$ and SVM$_{R-L}$) were very close to those of the SVM model (Figure 5 and Table 2, respectively). During the validation phase, the SVM$_{Z-L}$ model demonstrated the lowest RMSE value (0.044 M) followed closely by the SVM$_{R-L}$ model (0.049 M). Moreover, the R^2 values for these twoo models were very similar (Table 2). In contrast, the SVM$_Z$ model exhibited the highest value (0.147 M) and a MAPE of 8.8 %, followed by the SVM$_R$ model (with RMSE = 0.144 M and MAPE = 8.7 %). Finally, the SVM$_{R-L}$ model showed the lowest RMSE and MAPE values (0.040 M and 2.1 %, respectively) for the testing phase. The SVM$_{Z-L}$ model was not far behind the SVM$_{R-L}$ model in terms of the three statistical parameters. In contrast, the SVM$_R$ and SVM$_Z$ models showed the poorest performance. These results are illustrated in Figure 5 and Table 2.

According to these results, SVM models using a logarithmic scale achieved better results with the internal data compared to SVM models using a linear scale. Moreover, these predictive models showed the best generalized predictive performances with the external data, indicating that they are more suitable for modelling the logarithmic CMC values.

As mentioned in the previous section, the best SVM model was chosen to further evaluate its performance. Considering the criterion previously described, the model (SVM$_{Z-L}$), with a RMSE value of 0.044 M, was selected for the validation phase, although it was not the one with the lowest value for the test phase (Table 2 and Figure 5).

Figure 6 shows the deviation of the actual CMC values from those predicted by the SVM$_{Z-L}$ model for the three phases. Figure 6A shows for the training and validation cases. According to this figure, the case (–1.31, –1.49) is noteworthy, as it contains high fitting errors and is in a position away from the red dashed line with respect to the other cases. In fact, this case has the highest absolute error (0.172 M). Furthermore, in this point, the model predicted a value of –1.49 M versus –1.31 M (real value), leading to an overestimation of the model (absolute percentage error of 13.1 %).

Figure 6B shows that most of the test cases are close to the slope line 1. However, one case (–1.36, –1.55) is worth mentioning, which is very attractive to the human eye and is located in the middle zone of the graph. For this test case, the model predicted a value of –1.55 M versus –1.36 M (overestimation of model). The absolute error and the absolute percentage error were the highest of the all the test cases (0.194 M and 14.3 %, respectively). In fact, when this point is removed, the RMSE and MAPE values drop to 0.034 M and 0.020 M, and R^2 increases to 0.993.

3.3 Comparison with the ANN model obtained from the previous study

According to the previous sections, the best models in this research are: RF$_R$ for the RF algorithm and SVM$_{Z-L}$ for the

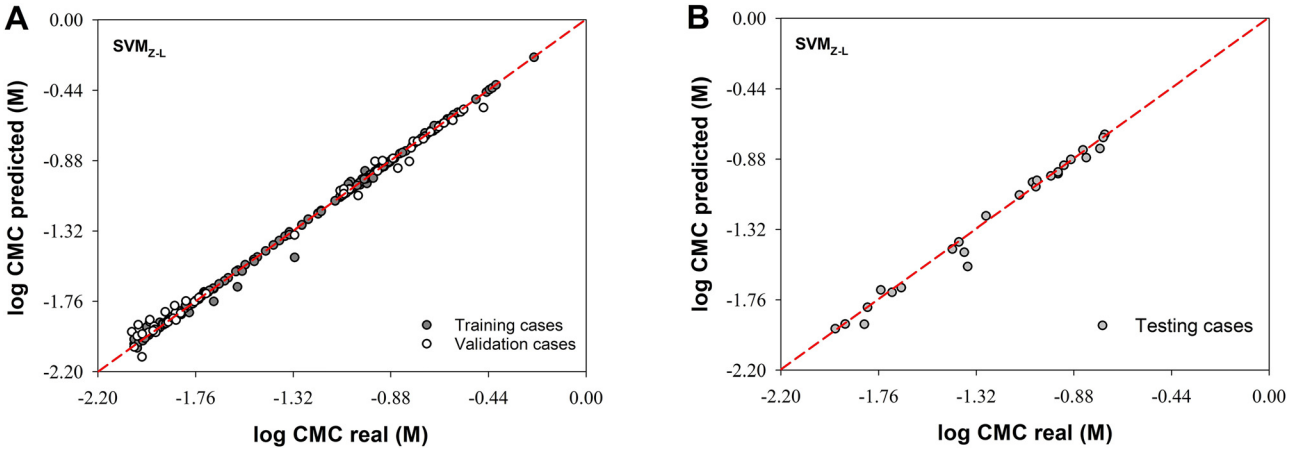

Figure 6: Scatter plots presenting actual and predicted values of log CMC real (*x*-axis) and log CMC predicted (*y*-axis) for the SVM_{Z-L} model in the training and validation phases (A) and testing phase (B). The red dashed line corresponds to the line with slope one.

SVM algorithm. In this section, a comparison of these models is carried out with the best ANN model shown in the previous study[18] (Figure 7).

According to the results, it is stated that the best models for the three algorithms (RF, SVM and ANN) are the normalized ones. This corroborates that normalization usually provides excellent fits because this method prevents any variable from having a greater effect on others during model training. In fact, the previous study by Soria-Lopez et al.[18] demonstrated that all the normalized ANN models were the ones that showed the best fits with respect to the non-normalized ANN models, with the lowest RMSE values for the validation phase. However, the models developed in this research did not follow that pattern. Regarding the RF and SVM models, no clear differences, no defined pattern, were observed between the non-normalized and normalized ones. The three RF models developed were similar in terms of fits both for the validation phase and for the rest of the phases, with the RF_Z model showing the lowest RMSE value in the validation phase (Figure 3). As for the SVM models, the following pattern was analysed: linear scale models showed the worst fits for all statistical parameters compared to those developed from a logarithmic scale for all phases. Among the SVM models developed using a logarithmic scale, the SVMZ-L model showed the lowest RMSE value in the validation (Figure 5).

On the other hand, it can be observed that the RF_R model showed the worst fits for all phases in the RSME (Figure 7), MAPE and R^2 values (Table 3). On the other hand, ANN_Z model presented a lower RMSE value for the validation phase (0.040 M), followed very closely by the SVM_{Z-L} model (0.044 M). Moreover, the MAPE and R^2 values were practically similar. During the training phase, the SVM_{Z-L} model presented slightly better fits than the ANN_Z model. Furthermore, this model showed better fits for the external

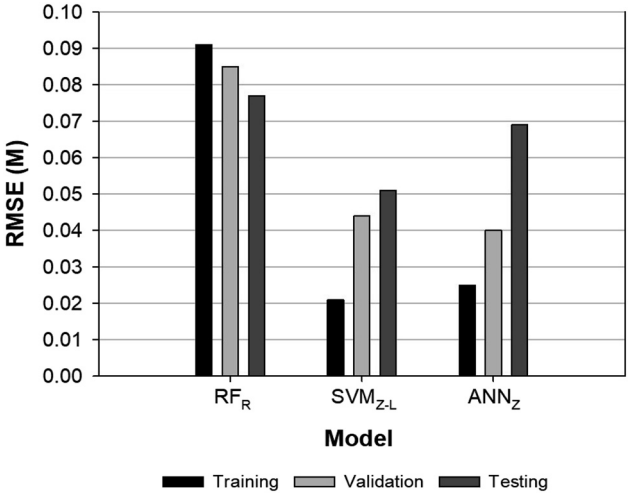

Figure 7: Graphical representation of the RMSE values (in M) for training, validation and testing for the three better models. RMSE is root mean square error (M).

Table 3: Values of the statistical parameters MAPE and R^2 obtained by each of the best RF and SVM models for all phases (training, validation and testing). Furthermore, the values of the statistical parameters of the best ANN model obtained in the previous study were also included. MAPE is the mean absolute percentage error (%) and R^2 is the linear squared correlation coefficient.

	Training		Validation		Testing	
Model	MAPE	R^2	MAPE	R^2	MAPE	R^2
RF_R	6.1	0.966	6.2	0.980	5.6	0.964
SVM_{Z-L}	0.6	0.998	2.6	0.994	2.6	0.985
ANN_Z	1.5	0.997	2.7	0.995	3.5	0.970

data (testing phase) with RMSE and MAPE values of 0.051 M and 2.6 % versus 0.069 M and 3.5 % for ANN_Z model (Table 3).

According to these results, it can be concluded that machine learning-based models are suitable for modelling the logarithmic CMC values of ionic surfactants. This research demonstrates that other machine learning approaches, such as the Random Forest and the Support Vector Machine, are promising tools to replace traditional laboratory measurements. These algorithms were successfully applied in this research. In this sense, RF and SVM together with ANN are three approaches that work adequately for this type of problems.

Other studies can be found in the literature that have used machine learning-based models to model the logarithmic CMC values. A complete comparison of the best models obtained in our study with other previous models in the literature is important to know the scope of our models. However, this comparison is difficult due to differences in data sets, hyper-parameter usage, and surfactant types used to predict in the models found in the literature. The study by Boukelkal et al.[44] showed that nonlinear machine learning-based methods (ANN, RFR, and SVR) provided better fits for the prediction of CMC of different surfactant classes. Among those models, the SVR model for the global phase was the best with an RMSE value of 0.205 M and an R^2 of 0.974 for a total of 593 experimental cases.[44] Our best SVM model (RMSE = 0.031 M and R^2 = 0.996) for a total of 258 experimental cases presented better statistical values (Figure 7 and Table 3). Another study by Rahal, Hadidi, and Hamadache[45] used regression and ANN methods to predict the CMCs of 50 anionic surfactants. The results showed that the ANN model with a 4-3-1 architecture presented the best results with an R^2 of around 0.940 for the training phase.[45] Our best ANN for this phase showed a higher R^2 value (0.997) (Table 3). Finally, Chen and colleagues[12] used tree-based ensemble algorithms including RF to predict the CMCs of 779 experimental cases of different surfactant classes. The results showed that the RF model exhibited an RMSE, MAPE, and R^2 of 0.200 M, 23.93 %, and 0.972, respectively, for the training phase, while 0.325 M, 10.32 %, and 0.927, respectively, for the testing phase.[12] In our study, the RF model showed better fits (Figure 7 and Table 3).

Based on these results, it can be stated that our models outperform all previously published models in all statistical metrics available for comparison, as they exhibit a lower root mean square error and a higher correlation coefficient. However, further research is still needed on these machine learning approaches in this area. New parameter combinations, data splitting, new normalization methods, among others, need to be studied to assess whether the models can be further improved.

4 Conclusions

The study of CMC is of great interest for industrial and academic applications within the field of surfactant. The use of predictive models using machine learning algorithms is a suitable tool and a possible alternative to expensive experimental measurements in the laboratory. In this research, two machine learning-based approaches, namely RF and SVM, were applied to model the logarithmic CMC values of 10 ionic surfactants using the same input variables as in the previous study.[18] The best RF and SVM models were normalized according to the criterion for the selection of the best model. Then, these models were compared with the best ANN model developed in the previous study. The ANN_Z model showed the best fit for the validation phase (RMSE = 0.040 M and with a MAPE = 2.7 %), while the $SVM_{Z\text{-}L}$ model showed better fits for the external data (RMSE = 0.051 M and with a MAPE = 2.6 %). Generally, the three predictive models developed presented adequate fits. Therefore, the implementation of the RF and SVM algorithms developed in this study and the ANN algorithm developed in the previous study for CMC prediction are effective tools and good substitutes for experimental laboratory measurements. Finally, further studies of these approaches in CMC prediction using new methods (other hyper-parameter combinations, new data splits, and more experimental cases, etc.) are needed to further improve these results.

Acknowledgments: This research was supported by an FPU grant from the Spanish Ministry of Science and Innovation (MCINN) to Anton Soria-López (FPU2020/06140). The authors would like to thank RapidMiner Inc. for the Educational and the free license of RapidMiner Studio software (version 10.2.000).

Research ethics: Not applicable.

Informed consent: Not applicable.

Author contributions: All authors have read and agreed to the published version of the manuscript. Conceptualization, J.C.M. and A.S-L. Methodology, A.S-L. Validation, A.S-L. and M.G-A. Formal analysis, A.S-L. and M.G-A. Investigation, A.S-L. Writing – original draft preparation, A.S-L. Writing – review and editing, M.G-A. and J.C.M. Visualization, A.S-L. Supervision, J.C.M. Project administration, J.C.M. Funding acquisition, J.C.M.

Use of Large Language Models, AI and Machine Learning Tools: None declared.

Conflict of interest: The authors declare no conflict of interest.

Research funding: This research was supported by an FPU grant from the Spanish Ministry of Science and Innovation (MCINN) to Anton Soria-López (FPU2020/06140).

Data availability: The data presented in this study are available on request from the corresponding authors. The data are not publicly available due to privacy restrictions.

References

1. Veeramanoharan, A.; Kim, S.-C. A Comprehensive Review on Sustainable Surfactants from CNSL: Chemistry, Key Applications and Research Perspectives. *RSC Adv*. **2024**, *14*, 25429–25471. https://doi.org/10.1039/D4RA04684F.
2. Aguirre-Ramírez, M.; Silva-Jiménez, H.; Banat, I. M.; Díaz De Rienzo, M. A. Surfactants: Physicochemical Interactions with Biological Macromolecules. *Biotechnol. Lett.* **2021**, *43*, 523–535. https://doi.org/10.1007/s10529-020-03054-1.
3. Williams, J. Formulation of Carpet Cleaners. In *Handbook for Cleaning / Descontamination of Surfaces*; Johansson, I.; Somasundaran, P., Eds.; Elsevier B.V.: Amsterdam, 2007; pp. 103–123.
4. Cheng, K. C.; Khoo, Z. S.; Lo, N. W.; Tan, W. J.; Chemmangattuvalappil, N. G. Design and Performance Optimisation of Detergent Product Containing Binary Mixture of Anionic-Nonionic Surfactants. *Heliyon* **2020**, *6*, e03861. https://doi.org/10.1016/j.heliyon.2020.e03861.
5. Rapp, B. E. Chapter 20 – Surface Tension. In *Microfluidics: Modelling, Mechanics and Mathematics*; Rapp, B. E., Ed.; Micro and Nano Technologies; Elsevier: Oxford, 2017; pp. 421–444.
6. St. Laurent, J. B.; de Buzzaccarini, F.; de Clerck, K.; Demeyere, H.; Labeque, R.; Lodewich, R.; van Langenhove, L. Laundry Cleaning of Textiles. In *Handbook for Cleaning / Descontamination of Surfaces*; Johansson, I.; Somasundaran, P., Eds.; Elsevier B.V.: Amsterdam, 2007; pp. 57–102.
7. Dini, S.; Bekhit, A. E.-D. A.; Roohinejad, S.; Vale, J. M.; Agyei, D. The Physicochemical and Functional Properties of Biosurfactants: A Review. *Molecules* **2024**, *29*, 2544. https://doi.org/10.3390/molecules29112544.
8. Poša, M. The Gibbs-Helmholtz Equation and the Enthalpy–Entropy Compensation (EEC) Phenomenon in the Formation of Micelles in an Aqueous Solution of Surfactants and the Cloud Point Effect. *J. Mol. Liq.* **2024**, *396*, 124109. https://doi.org/10.1016/j.molliq.2024.124109.
9. El-Dossoki, F. I.; Gomaa, E. A.; Hamza, O. K. Solvation Thermodynamic Parameters for Sodium Dodecyl Sulfate (SDS) and Sodium Lauryl Ether Sulfate (SLES) Surfactants in Aqueous and Alcoholic-Aqueous Solvents. *SN Appl. Sci.* **2019**, *1*, 933. https://doi.org/10.1007/s42452-019-0974-6.
10. Perinelli, D. R.; Cespi, M.; Lorusso, N.; Palmieri, G. F.; Bonacucina, G.; Blasi, P. Surfactant Self-Assembling and Critical Micelle Concentration: One Approach Fits All? *Langmuir* **2020**, *36*, 5745–5753. https://doi.org/10.1021/acs.langmuir.0c00420.
11. Astray, G.; Iglesias-Otero, M. A.; Moldes, O. A.; Mejuto, J. C. Predicting Critical Micelle Concentration Values of Non-ionic Surfactants by Using Artificial Neural Networks. *Tenside Surfactants Deterg.* **2013**, *50* (2), 118–124. https://doi.org/10.3139/113.110242.
12. Chen, J.; Hou, L.; Nan, J.; Ni, B.; Dai, W.; Ge, X. Prediction of Critical Micelle Concentration (CMC) of Surfactants Based on Structural Differentiation Using Machine Learning. *Colloids Surf. A Physicochem. Eng. Asp.* **2024**, *703*, 135276. https://doi.org/10.1016/j.colsurfa.2024.135276.
13. Abooali, D.; Soleimani, R. Structure-Based Modeling of Critical Micelle Concentration (CMC) of Anionic Surfactants in Brine Using Intelligent Methods. *Sci. Rep.* **2023**, *13*, 13361. https://doi.org/10.1038/s41598-023-40466-1.
14. Liao, Z.; Lu, J.; Xie, K.; Wang, Y.; Yuan, Y. Prediction of Photochemical Properties of Dissolved Organic Matter Using Machine Learning. *Environ. Sci. Technol.* **2023**, *57*, 17971–17980. https://doi.org/10.1021/acs.est.2c07545.
15. Zhong, S.; Zhang, K.; Bagheri, M.; Burken, J. G.; Gu, A.; Li, B.; Ma, X.; Marrone, B. L.; Ren, Z. J.; Schrier, J.; Shi, W.; Tan, H.; Wang, T.; Wang, X.; Wong, B. M.; Xiao, X.; Yu, X.; Zhu, J.-J.; Zhang, H. Machine Learning: New Ideas and Tools in Environmental Science and Engineering. *Environ. Sci. Technol.* **2021**, *55*, 12741–12754. https://doi.org/10.1021/acs.est.1c01339.
16. Qin, S.; Jin, T.; Van Lehn, R. C.; Zavala, V. M. Predicting Critical Micelle Concentrations for Surfactants Using Graph Convolutional Neural Networks. *J. Phys. Chem. B* **2021**, *125*, 10610–10620. https://doi.org/10.1021/acs.jpcb.1c05264.
17. Moriarty, A.; Kobayashi, T.; Salvalaglio, M.; Angeli, P.; Striolo, A.; McRobbie, I. Analyzing the Accuracy of Critical Micelle Concentration Predictions Using Deep Learning. *J. Chem. Theory Comput.* **2023**, *19*, 7371–7386. https://doi.org/10.1021/acs.jctc.3c00868.
18. Soria-Lopez, A.; García-Martí, M.; Barreiro, E.; Mejuto, J. C. Ionic Surfactants Critical Micelle Concentration Prediction in Water/Organic Solvent Mixtures by Artificial Neural Network. *Tenside Surfactants Deterg.* **2024**, *61* (6), 519–529. https://doi.org/10.1515/tsd-2024-2623.
19. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. https://doi.org/10.1023/A:1010933404324.
20. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20* (3), 273. https://doi.org/10.1007/BF00994018.
21. Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer-Verlag: New York, 1995.
22. AlKheder, S.; AlRukaibi, F.; Aiash, A. Support Vector Machine (SVM), Random Forest (RF), Artificial Neural Network (ANN) and Bayesian Network for Prediction and Analysis of GCC Traffic Accidents. *J. Ambient Intell. Humaniz. Comput.* **2023**, *14* (6), 7331–7339. https://doi.org/10.1007/s12652-022-04441-4.
23. Ishola, N. B.; Epelle, E. I.; Betiku, E. Machine Learning Approaches to Modeling and Optimization of Biodiesel Production Systems: State of Art and Future Outlook. *Energy Convers. Manag. X* **2024**, *23*, 100669. https://doi.org/10.1016/j.ecmx.2024.100669.
24. Mathai, N.; Chen, Y.; Kirchmair, J. Validation Strategies for Target Prediction Methods. *Brief. Bioinform.* **2019**, *21* (3), 791–802. https://doi.org/10.1093/bib/bbz026.
25. Lei, L.; Shao, S.; Liang, L. An Evolutionary Deep Learning Model Based on EWKM, Random Forest Algorithm, SSA and BiLSTM for Building Energy Consumption Prediction. *Energy* **2024**, *288*, 129795. https://doi.org/10.1016/j.energy.2023.129795.
26. Antoniadis, A.; Lambert-Lacroix, S.; Poggi, J.-M. Random Forests for Global Sensitivity Analysis: A Selective Review. *Reliab. Eng. Syst. Saf.* **2021**, *206*, 107312. https://doi.org/10.1016/j.ress.2020.107312.
27. Bagherzadeh, F.; Mehrani, M.-J.; Basirifard, M.; Roostaei, J. Comparative Study on Total Nitrogen Prediction in Wastewater Treatment Plant and Effect of Various Feature Selection Methods on Machine Learning Algorithms Performance. *J. Water Process Eng.* **2021**, *41*, 102033. https://doi.org/10.1016/j.jwpe.2021.102033.
28. Iranzad, R.; Liu, X. A Review of Random Forest-Based Feature Selection Methods for Data Science Education and Applications. *Int. J. Data Sci. Anal*. **2024**. https://doi.org/10.1007/s41060-024-00509-w.
29. S, K.; Ravi, Y. K.; Kumar, G.; Nandabalan, Y. K.; J, R. B. Microalgal Biorefineries: Advancement in Machine Learning Tools for Sustainable Biofuel Production and Value-Added Products Recovery. *J. Environ.*

Manage. **2024**, *353*, 120135. https://doi.org/10.1016/j.jenvman.2024.120135.
30. Andrade Cruz, I.; Chuenchart, W.; Long, F.; Surendra, K. C.; Renata Santos Andrade, L.; Bilal, M.; Liu, H.; Tavares Figueiredo, R.; Khanal, S. K.; Fernando Romanholo Ferreira, L. Application of Machine Learning in Anaerobic Digestion: Perspectives and Challenges. *Bioresour. Technol.* **2022**, *345*, 126433. https://doi.org/10.1016/j.biortech.2021.126433.
31. Hu, J.; Szymczak, S. A Review on Longitudinal Data Analysis with Random Forest. *Brief. Bioinform.* **2023**, *24* (2), 1–11. https://doi.org/10.1093/bib/bbad002.
32. Li, X.; Yu, J.; Jia, Z.; Song, J. Harmful Algal Blooms Prediction with Machine Learning Models in Tolo Harbour. In *2014 International Conference on Smart Computing*: Hong Kong, China: IEEE, 2014, pp. 245–250. https://doi.org/10.1109/SMARTCOMP.2014.7043865.
33. Gaye, B.; Zhang, D.; Wulamu, A. Improvement of Support Vector Machine Algorithm in Big Data Background. *Math. Probl. Eng.* **2021**, *2021*, 1–9. https://doi.org/10.1155/2021/5594899.
34. Soria-Lopez, A.; Sobrido-Pouso, C.; Mejuto, J. C.; Astray, G. Assessment of Different Machine Learning Methods for Reservoir Outflow Forecasting. *Water* **2023**, *15*, 3380. https://doi.org/10.3390/w15193380.
35. Fan, J.; Jing, F.; Fang, Z.; Tan, M. Automatic Recognition System of Welding Seam Type Based on SVM Method. *Int. J. Adv. Manuf. Technol.* **2017**, *92*, 989–999. https://doi.org/10.1007/s00170-017-0202-8.
36. Akter, T.; Bhattacharya, T.; Kim, J.-H.; Kim, M. S.; Baek, I.; Chan, D. E.; Cho, B.-K. A Comprehensive Review of External Quality Measurements of Fruits and Vegetables Using Nondestructive Sensing Technologies. *J. Agric. Food Res.* **2024**, *15*, 101068. https://doi.org/10.1016/j.jafr.2024.101068.
37. Nie, Z.; Bai, X.; Nie, L.; Wu, J. Optimization of the Economic and Trade Management Legal Model Based on the Support Vector Machine Algorithm and Logistic Regression Algorithm. *Math. Probl. Eng.* **2022**, *2022*. https://doi.org/10.1155/2022/4364295.
38. Boualem, A. D.; Argoub, K.; Benkouider, A. M.; Yahiaoui, A.; Toubal, K. Viscosity Prediction of Ionic Liquids Using NLR and SVM Approaches. *J. Mol. Liq.* **2022**, *368*, 120610. https://doi.org/10.1016/j.molliq.2022.120610.
39. Hsu, C.-W.; Chang, C.-C.; Lin, C.-J. A Practical Guide to Support Vector Classification, 2003. https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.
40. Chang, C.-C.; Lin, C.-J. LIBSVM: A Library for Support Vector Machine. *ACM Trans. Intell. Syst. Technol.* **2011**, *2* (27), 1–27. https://doi.org/10.1145/1961189.1961199.
41. RapidMiner. Support Vector Machine (LibSVM). https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/support_vector_machines/support_vector_machine_libsvm.html (accessed 2024-03-04).
42. Iida, T. Identifying Causes of Errors between Two Wave-Related Data Using Performance Metrics. *Appl. Ocean Res.* **2024**, *148*, 104024. https://doi.org/10.1016/j.apor.2024.104024.
43. Chicco, D.; Warrens, M. J.; Jurman, G. The Coefficient of Determination R-Squared Is More Informative Than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation. *PeerJ Comput. Sci.* **2021**, *7*, e623. https://doi.org/10.7717/peerj-cs.623.
44. Boukelkal, N.; Rahal, S.; Rebhi, R.; Hamadache, M. QSPR for the Prediction of Critical Micelle Concentration of Different Classes of Surfactants Using Machine Learning Algorithms. *J. Mol. Graph. Model.* **2024**, *129*, 108757. https://doi.org/10.1016/j.jmgm.2024.108757.
45. Rahal, S.; Hadidi, N.; Hamadache, M. In Silico Prediction of Critical Micelle Concentration (CMC) of Classic and Extended Anionic Surfactants from Their Molecular Structural Descriptors. *Arab. J. Sci. Eng.* **2020**, *45*, 7445–7454. https://doi.org/10.1007/s13369-020-04598-0.

Bionotes

Anton Soria-López
Anton Soria-López is dooing his PhD at the Agri-Environmental and Food Research Group at Ourense Campus (University of Vigo). His research interest is focused in the applications of Artificial Neural Networks to chemical and biological problems.

María García-Martí
María García-Martí is PosDoctoral Research at the Agri-Environmental and Food Research Group at Ourense Campus. His research interest is focused in food chemistry and environmental research.

Juan C. Mejuto
Juan C. Mejuto currently is Full Professor in the Physical Chemistry Department of University of Vigo. He is the head of the Agri-Environmental and Food Research Group at Ourense Campus. His research interest comprises (i) physical organic and physical inorganic chemistry, (ii) reactivity mechanisms in homogeneous and micro heterogeneous media, (iii) stability of self-assembly aggregates and (iv) supramolecular chemistry.