Brazilian Journal of Analytical Chemistry
2023, Volume 10, Issue 39, pp 52-69
doi: 10.30744/brjac.2179-3425.AR-30-2022

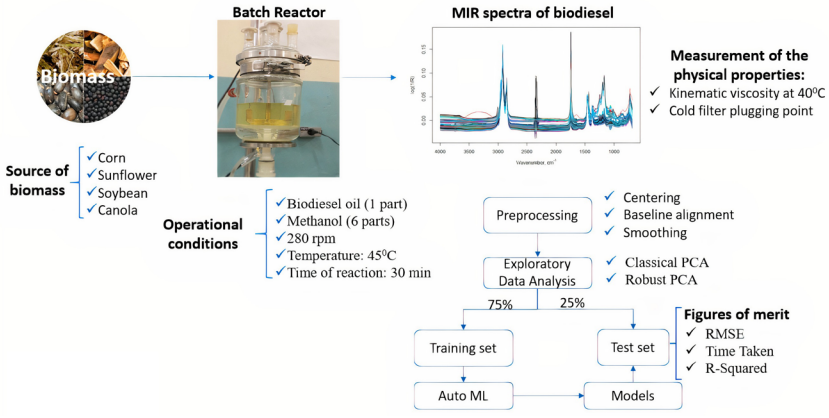
BrJAC

ARTICLE

Employing Auto-Machine Learning Algorithms for Predicting the Cold Filter Plugging and Kinematic Viscosity at 40 ºC in Biodiesel Blends using Vibrational Spectroscopy Data

Aderval Severino Luna*[1], Alexandre Rodrigues Torres[2], Camilla Lima Cunha[2], Igor C. A. Lima[1], Luis G. Nonato[3]

[1]*Universidade do Estado do Rio de Janeiro, Programa de Pós-graduação em Engenharia Química, Rua São Francisco Xavier, 524, Maracanã, 20550-900, Rio de Janeiro, RJ, Brazil*
[2]*Universidade do Estado do Rio de Janeiro, Faculdade de Tecnologia, Rodovia Presidente Dutra, km 298, Polo Industrial, 27537-000, Resende, RJ, Brazil*
[3]*Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação, Avenida Trabalhador São-Carlense, 400, 13560-970, São Carlos, SP, Brazil*



This work aims to develop an auto-machine learning method using Mid-Infrared (MIR) spectroscopy data to determine the cold filter plugging point (CFPP) and kinematic viscosity at 40 ºC of biodiesel, diesel, and mixtures samples. The biodiesel was obtained by the transesterification reaction and later purified. The first dataset was composed of 108 blends (biodiesel obtained from different biomass such as soy, corn, sunflower, and canola) with binary, ternary and quaternary mixtures. The second dataset was composed of 227 blends of diesel-biodiesel and diesel-biodiesel-ethanol, respectively. The physical properties of the samples were obtained according to ABNT NBR 14747 and ABNT NBR 10441, respectively. The MIR Spectroscopy data were acquired from 7,800 to 450 cm^{-1}, with a 4 cm^{-1} resolution and 20 scans. The spectra' baseline alignment was carried out using the asymmetric least squares method. A Savitzky–Golay filter was applied to a set of digital data points to smooth the data. This work used a first-order polynomial and a zero derivative function to smooth the spectra. The dataset was split into training and test sets using the function CreateDataPartition from the caret package. It was adopted 70% for training and 30% for test sets. In this work, the model training process was carried out using the open-source Python library LazyPredict. The LazyPredict returns the

Cite: Luna, A. S.; Torres, A. R.; Cunha, C. L.; Lima, I. C. A.; Nonato, L. G. Employing Auto-Machine Learning Algorithms for Predicting the Cold Filter Plugging and Kinematic Viscosity at 40 ºC in Biodiesel Blends using Vibrational Spectroscopy Data. *Braz. J. Anal. Chem.*, 2023, *10* (39), pp 52-69. http://dx.doi.org/10.30744/brjac.2179-3425.AR-30-2022

Submitted 07 May 2022, Resubmitted 26 June 2022, 2nd time Resubmitted 03 August 2022, Accepted 08 August 2022, Available online 31 August 2022.

trained models and their performance metrics. The kinematic viscosity at 40 ºC of the biodiesel samples and their blends could be modeled using the MIR Spectroscopy dataset using different auto-machine learning algorithms. The RMSEP (Root Mean Square Error of Prediction) (≤ 0.02 mm^2 s^{-1}) was similar to the experimental error obtained after log transformation. The CFPP of the biodiesel samples and their blends could be modeled using the MIR Spectroscopy dataset by different auto-machine learning algorithms with an RMSEP (≤ 1.6 ºC) similar to the experimental error obtained by traditional methodology. Based on the lower computational time and the same performance observed by the RMSEP and R^2 (coefficient of determination) values from different algorithms, it is recommended to use Ridge or Ridge Cross-Validation Regression models for both physical properties using MIR Spectroscopy data.

Keywords: auto-machine learning algorithms, biofuels, cold filter plugging point, ridge regression, kinematic viscosity at 40 ºC

INTRODUCTION

Due to population growth and global economic development, the world faces energy demand problems. Therefore, global energy demand triggers the excess consumption of fossil fuels. This phenomenon produces the following main issues: (1) excess global pollution accentuated by the increase in the greenhouse effect, (2) global warming and climate change, and (3) depletion of fossil fuels. Renewable energy emerges as an alternative for the energy supply faced with these severe problems. Biodiesel has been recognized to produce power with lower environmental impacts than fossil fuels.[1]

Furthermore, with the uncertainty of fossil fuels regarding their future availability and the need for green fuels, there is increasing attention to using biodiesel as an alternative fuel. As biodiesel can be produced by different biomasses, such as agricultural residues, these can also serve as raw materials for biofuel production, according to Bobadilla et al. (2018).[2]

Biodiesel is considered the future fuel, especially as the trend is for oil to become so scarce that its price becomes unaffordable. Furthermore, the world is looking for less polluting solutions to maintain the planet's sustainability. The good news is that Brazil is one of the largest biodiesel producers globally and has superior technology, generating a higher quality product. Biodiesel production also reduces the need to import oil, balancing the country's economy.[3-4]

Several parameters are recommended by the ANP (National Petroleum, Natural Gas, and Biocombustible Agency) to verify the quality of biodiesel and its blends. The CFPP and the kinematic viscosity at 40 ºC were chosen in this work.[5] Concerning the CFPP at low temperatures, biodiesel partially solidifies or loses its fluidity, leading to fuel flow interruption and clogging of the filtration system, causing problems in engine starting.[6] Meanwhile, the kinematic viscosity at 40 ºC increases in value with the length of the carbon chain and the degree of saturation and influences the burning process in the engine's combustion chamber.[6] However, the traditional methodologies were slow and made it difficult to make a quick decision about the quality of these products. Therefore, using other analytical techniques faster and more reliable is recommended than traditional ones. In this context, infrared spectroscopy meets these requirements and is a well-established technique in analytical chemistry.

Studies in the literature used machine learning algorithms to predict the properties of biodiesel and diesel-biodiesel blends. Several articles published in the literature concerning this application were highlighted. Pimentel et al. (2006)[9] evaluated the application of calibration models multivariate by PLS (Partial Least-Squares) based on MIR (Mid-Infrared) Spectroscopy (4000 to 650 cm^{-1}; ATR: Attenuated Total Reflectance) and NIR (Near-Infrared) Spectroscopy (12000 to 4000 cm^{-1}; optical path = 1.0 cm) spectra to predict the biodiesel content in diesel oil blends considering the presence of vegetable oils. According to the authors, the F test (with a 95% confidence level) showed no statistically significant difference between the models built using these techniques. However, the F test should not be used compared to the models since the RMSEP (Root Mean Square Error of Prediction) values do not follow the F distribution or chi-square distribution.[7] Baptista et al. (2008)[10] evaluated NIR spectroscopy for predicting biodiesel properties,

such as the iodine index, the CFPP, density at 15 ºC, and viscosity kinematics at a temperature of 40 ºC. The CFPP biodiesel prediction model using PLS produced an R^2 (coefficient of determination) of 0.951 and RMSEP 1.0 ºC. The best prediction model for density at 15 ºC was obtained by PLS with R^2 of 0.999 and RMSEP of 0.9 kg m^{-3}. However, it is noticed that this study was carried out with only 71 samples (49 samples for validation test and 22 samples for test samples) to predict the CFPP.[8] Lira et al. (2010)[11] prepared blends using soy methyl esters, castor oil, cottonseed oil, canola oil, sunflower oil, and diesel samples from different regions of Brazil. For the density prediction model using NIR spectroscopy data, the PLS model showed R^2 values of 0.99 and RMSEP of 0.56 kg m^{-3}. This prediction was adequate and showed an error with similar performance to those obtained for traditional techniques.[9] Balabin and Safieva (2011)[12] developed an artificial neural network (ANN) model to predict fuel properties, including the CFPP, based on a NIR Spectroscopy dataset. The ANN model showed a better performance when compared to the other multivariate linear regressions. It was a surprise that the authors used a multiple linear regression (MLR) because it is well known that near-infrared spectra were highly correlated. The multicollinearity must be considered, which these authors cannot ignore.[10] Filgueiras et al. (2014)[13] compared the PLS and SVM (Support Vector Machine) models' performance for predicting API gravity, kinematic viscosity, and water content in petroleum using Fourier Transform Infrared Spectroscopy with Attenuated Total Reflectance (FTIR-ATR). The authors only used 68 samples to carry out this work, and the performance of the SVM model was better than the PLS model for predicting kinematic viscosity. It was noticed that the RMSEP reported for both models was higher than the actual values of this property. Probably, the authors committed a mistake.[11]

Cunha et al. (2017)[15] used PLS and SVM to model the relationship between the FT-IR Spectroscopy data and density, refraction, and CFPP of pure biodiesel samples and their mixtures.[12] According to the authors, it is not recommended to use the F-test, and the SWTP (Sum of Wilcoxon Test Probability) is the better choice.[13] The best CFPP prediction was obtained using SVM regression, which showed an equal RMSEP at 0.6 ºC. The PLS model resulted in the best density and refractive index prediction with RMSEP values equal to 0.2 kg m^{-3} and 0.0001, respectively. The predicted values for physical properties were similar to those obtained by conventional techniques, demonstrating the feasibility of using machine learning techniques when coupled with NIR spectroscopy.[12]

Through the bibliographic research carried out, it was observed that the most used spectroscopic techniques involving machine learning algorithms were NIR and MIR spectroscopies. On the other hand, the main characteristics of this work that differ from similar results reported in the literature were the high number of samples, the different sources of raw material used in the production of biodiesel, and the wide range of physical properties measured in the biodiesel samples and their mixtures.

Researchers spend enormous time searching for the most suitable algorithm and preprocessing methods to solve a predictive task. However, automatic machine learning is a viable alternative for solving numerous regression and classification problems. In this sense, Auto Machine Learning (AutoML) is a modern approach to automated model retrieval, training procedures, and hyperparameter optimization for specific problems. One of the most famous AutoML methods is meta-learning (MTL), which proposed to develop models that offer algorithm recommendations and parameter values to be adopted for each new problem. This segment has gained notoriety due to its excellent ability to generate suitable models without human intervention.[14] Its main objective is to reduce the number of tested algorithms to optimize experimentation time with minimal loss in the quality of results.[15-16]

Some technologies facilitate the steps of an AutoML project, such as Auto-Keras,[17] Auto Sklearn,[18] Cloud AutoML,[19] H2O,[20] MLBox,[21] Lazy Predict[22]. All these tools have specific configurations. For the present work, the Lazy Predict framework is more appropriate, considering that it has an open-source code, comprises steps deemed necessary for evaluating MTL, and can potentially train many models. In addition, it is applicable in the Python computational program,[23] which presents a simple, effective, and versatile language.

Additionally, this is the first work in this area to employ automatic machine learning (Auto ML) methods to predict biodiesel's physical properties using MIR spectroscopy data. Therefore, this work aims to develop an auto-machine learning method to determine the following physical properties: CFPP and kinematic viscosity at 40 ºC of biodiesel, diesel, and mixtures samples using MIR Spectroscopy data.

MATERIALS AND METHODS

The methodology of producing biodiesel and its blends can be found in the literature.[1]

Dataset 1: Mixtures of Biodiesel

The pure samples were composed of pure biodiesel from soybeans, canola, sunflower, corn, and biodiesel provided by a distributor in southern Brazil, totaling 41 samples. Mixtures containing the four different sources of biodiesel (soybean, canola, sunflower, and corn) were prepared in triplicate and volumetric base. Thus, as 36 distinct mixtures were established, it was necessary to prepare 108 blends.

Tables S1 – S3 (Supplementary Material) show the compositions of the different mixtures of biodiesel samples in percentage volumetric. A distributor provided 2 liters of biodiesel in the southern region to compose the blends. Canola biodiesel has the value of the most distinct cold filter plugging point among the types of biodiesels. In this study, it was chosen to compose the binary mixture with biodiesel from the southern region. In summary, Dataset 1 consists of 40 pure samples, 61 binary, 27 ternary, and 21 quaternary samples, 149 samples.

Dataset 2: Diesel-biodiesel and diesel-biodiesel-ethanol blends

For Dataset 2 preparation, diesel-biodiesel and diesel-biodiesel-ethanol blends were prepared in triplicate and volumetric base. Pure samples used in the preparation of the mixtures of this step were also included in the dataset, totaling 33 pure samples (Diesel S-10, Diesel S-500, Standard pure diesel, rapeseed biodiesel, soy biodiesel, sunflower biodiesel, corn biodiesel, biodiesel from the southern region and biodiesel from a distributor in Rio de Janeiro State). 2 liters of soy biodiesel, 1 liter of sunflower biodiesel, 1 liter of canola biodiesel, 1 liter of corn biodiesel, 1 liter of biodiesel from the south region, 0.5 liter of biodiesel from R.J., 12 liters of S-10 diesel, 5 liters of S-500 diesel, 1.5 liters of pure standard diesel (no biodiesel added) and 0.5 liter of anhydrous ethanol to compose the mixtures.

As a representative of diesel in ternary blends, the Diesel S-10 was chosen for having a higher cetane number (48) than the S-500 (42) and for offering to any diesel vehicle, even those manufactured before 2012, better engine conservation and reduced maintenance costs. Soybean biodiesel was elected as the representative of biodiesel in the blends for being a widespread oleaginous source in producing this biofuel to compose the ternary mixtures. Table S4 (Supplementary Material) shows the compositions (% v/v) of the 16 ternary mixtures between the Diesel S-10, soybean biodiesel, and anhydrous ethanol.

Diesel S-10 (10 ppm sulfur), S-500 (500 ppm sulfur), and standard diesel (pure) were used in the composition of the mixtures. Table S5 (Supplementary Material) shows the 34 binary diesel-biodiesel mixtures prepared in triplicate, resulting in 102 samples.

CFPP

The CFPP is the temperature, in ºC, at which a specific sample volume does not pass through a metal filter standard in a specified period when cooled under certain conditions. The method is based on cooling, with a rate of 1°C/min, a volume of 45 mL sample, which is sucked into a pipette through a standardized metal mesh filter under a controlled vacuum. The procedure is repeated as many times as possible until the amount of crystals that separate from the solution is sufficient to interrupt or reduce the circulation of the sample through the filter. Alternatively, this procedure is repeated if the time required to fill the pipette exceeds 60 seconds or if the sample fails to return entirely to the test container before being cooled another 1 ºC. The temperature at which the final filtration was started is the cold filter plugging point. The CFPP tests were performed on TANAKA brand equipment, model AFP-102. The measurements were

obtained according to ABNT NBR 14747[24] and specified by the ANP Resolution 45/2014[5]. The repeatability of the measurement method was 1.8 ºC, and the reproducibility was 2.0 ºC for CFPP (ºC), with a 95% confidence level. The range of observed variation for the dataset under consideration was -26.0 ºC to 7.0 ºC, with a measurement error equal to 1.6 ºC.

Kinematic viscosity at 40 ºC

Kinematic viscosity is a property that measures resistance to flow under the gravity of a certain mass of fluid concerning its volume, which can be understood as the ratio between the dynamic viscosity and the specific mass of the liquid. The sample was homogenized in the original bottle during viscosity measurements. An aliquot of 10 mL was filtered and transferred via filter-syringe (PTFE 0.2 µm and 25 mm) for the Cannon-Fenske capillary viscometer. The sample was sucked above the upper line, and the time taken for the upper meniscus to pass successively through the two calibration marks was noted. Ten measurements of this time were carried out. The measurements were obtained according to ABNT NBR 10441[25] and specified by the ANP Resolution 45/2014[5]. The repeatability of the kinematic viscosity measurement method was 0.0155 $mm^2\ s^{-1}$, for a confidence level of 95%, and the reproducibility of 0.0279 $mm^2\ s^{-1}$. The observed range of variation for the dataset under consideration was 2.6629 to 4.8524 $mm^2\ s^{-1}$ with a measurement error equal to 0.0267 $mm^2\ s^{-1}$.

MIR Spectroscopy

The MIR Spectroscopy data were acquired in the range of 7800 to 450 cm^{-1}, with a resolution of 4 cm^{-1}, data intervals of 1 cm^{-1}, and 20 scans. The Horizontal ATR with zinc selenide (ZnSe) crystal from PIKE Technologies coupled to the Perkin Elmer infrared spectrophotometer model FT-MIR/NIR Frontier. The measurement of reflectance was transformed as follows: $log\left(\frac{1}{R}\right)$. Samples of mixtures of different biodiesel oils were analyzed placing 0.2 mL of each combination in the sample holder (crystal) – the mid-infrared spectra of blends and pure biodiesel. Tissue papers moistened in distilled water followed by tissue moistened with ethyl alcohol were used to clean the surface of the crystal. After evaporation of ethyl alcohol, the blank spectrum was acquired to verify the crystal's absence of residues and contaminants.

Softwares

The Python version 3.9.5,[23] R version 4.0.2,[26] RStudio version 1.4.1717,[27] and Visual Studio Code version 1.58.2 for Windows 64 bits were used in this work. LazyPredict Module builds many basic models without much code and helps understand which models work better without parameter tuning. The documentation can be found on the site: https://lazypredict.readthedocs.io. The computational time to build 43 regression models was inferior to 5 minutes using this dataset for each property.

Hardware

The statistical analysis was carried out on Notebook, Core i7, 8th Generation, 16 GB RAM, 256 GB SSD, and Video Nvidia GEFORCE GTX 1650.

RESULTS AND DISCUSSION

MIR Spectroscopy data of the biodiesel samples

MIR Spectroscopy of pure biodiesel from five different sources and their blends (binary, ternary and quaternary) were obtained in the spectral region from 4000 to 680 cm^{-1}. Figure 1 (left) shows the raw spectra of the biodiesel samples and their blends. As shown in Figure 1 (left), the sample (Am158: indicated by the red line) has a different behavior from the similar examples that make up the triplicate samples. This distinct behavior was attributed to water absorption, probably due to an error in cleaning the equipment between sample measurements. For this reason, it was necessary to remove it from the dataset.

In addition, the signal observed in the spectral region between 2400 to 2300 cm^{-1}, also observed in Figure 1 (left), was identified as the crystal response signal of zinc selenide (ZnSe) of the ATR accessory

and not of the spectral band of biodiesel samples. For this reason, it was decided to replace this region using the mean of 50 neighbor variables above and below this part of the spectra.

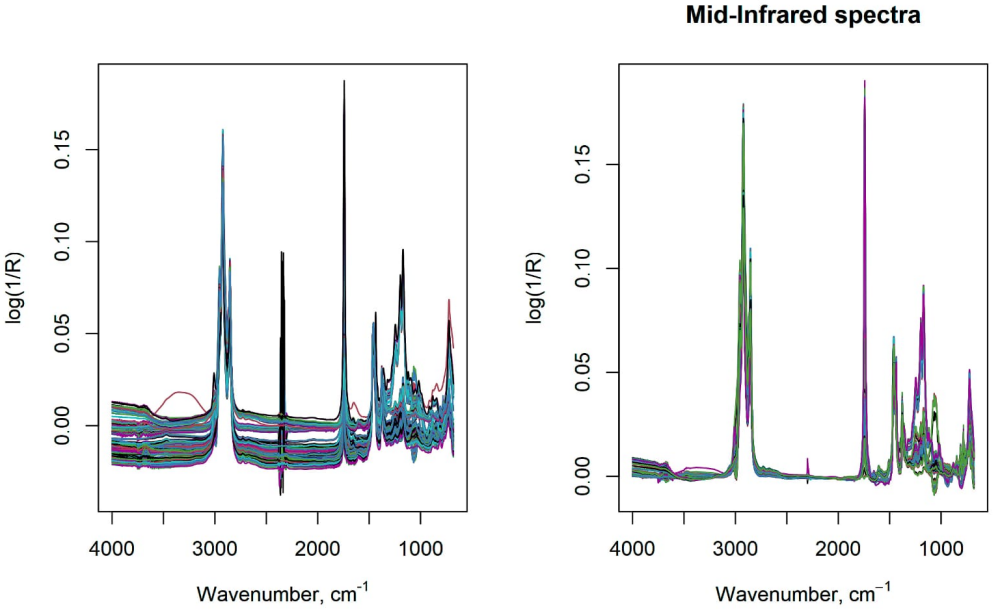


Figure 1. Raw mid-infrared spectra of the biodiesel samples (left) and after sample removal (Am158), followed by the replacement of the peak region assigned to the ZnSe crystal by the average of 50 variables above and below the region (right).

Interpretation of MIR Spectroscopy data of the biodiesel samples

The MIR Spectroscopy data region between 3000 and 2840 cm^{-1} showed C- H's fundamental stretching bands. The range between 1750 and 1730 cm^{-1} displays the carbonyl stretching band from the mid-range infrared spectra. The spectral region between 1500 and 1400 cm^{-1} displays bands out of plane bending vibration of the mid-infrared spectra. The spectral region between 1300 – 1000 cm^{-1} is where the fundamental stretching bands of C–O appear. The part between 750 – 700 cm^{-1} (methylene rocking vibration) indicates a long-chain linear aliphatic structure. Table I summarizes the interpretation of the MIR Spectroscopy data of the biodiesel samples.

Table I. Summary of MIR spectroscopy interpretation of biodiesel samples

Region	Wavenumber, cm^{-1}	Probably group	Class of compound
1	3000 – 2840	v C – H	Alkanes
2	1750 – 1730	v C = O	Carbonyl compounds
3	1500 – 1400	δ C – H	Alkanes
4	1300 – 1000	v C – O	Carboxylic acids, Esters
5	750 – 700	δ C – H	$(CH_2)_n$; n ≥ 4

v = stretching vibration; δ = out of plane bending.[28]

Baseline alignment and smoothing of the infrared spectra by the Savitsky-Golay method

All spectra' baseline alignment was carried out using the asymmetric least squares method.[29] Lately, a first-order polynomial and a zero derivative function have been used to smooth the spectra.[30] Figure 1 (right) shows the infrared spectra after applying baseline alignment and smoothing.

Exploratory data analysis using principal component analysis (PCA)

If the PCA model is made for a dataset from the same population, the orthogonal and score distances can find outliers and extreme objects. One of the ways to do this is to compute critical limits for the distances assuming that they follow a specific theoretical distribution. The residual/distance for this PCA model showed two critical limits: the dashed line is a limit for extreme objects, and the dotted line is a limit for outliers.

Principal component analysis (PCA) applied to MIR Spectroscopy data

Table II summarizes the PCA analysis for the MIR Spectroscopy dataset. The explained variance (%) determined the optimal number of principal components of the PCA model. The last number of components was used when the explained variance became less than 1%. In this case, six PCs were used for the PCA model.

Table II. Summary of the PCA (class PCA) for the MIR Spectroscopy dataset. Type of limits: (ddmoments) calculates critical limits for distance values using data-driven moments approach, alpha = 0.05, and gamma = 0.01. Alpha is the significance level for detecting extreme objects, and gamma is for detecting outliers.

PC	Eigenvalues	Explained Variance	Cumulative explained variance	N_q	N_h
1	1537.444	46.30	46.30	1	19
2	1021.223	30.75	77.05	1	1
3	336.135	10.12	87.17	1	1
4	157.657	4.75	91.92	1	1
5	91.634	2.76	94.68	1	1
6	57.298	1.73	96.40	2	1
7	27.859	0.84	97.24	2	1
8	18.264	0.55	97.79	2	1
9	17.000	0.51	98.30	2	1
10	12.434	0.37	98.68	3	1
11	8.956	0.27	98.95	2	1
12	6.268	0.19	99.13	2	1

N_q and N_h are the numbers of the degree of freedom (DoF) associated with h_0 and q_o, scaling factors, respectively.

However, when the dataset contains outliers using the classic estimators, the classical PCA is inappropriate based on the conventional mean and variance values. In these cases, the mean and variance of the corresponding distance are replaced with their robust analogs, namely median and interquartile range statistics.[31] Figure 2 shows the graph of distances obtained by the classical PCA and Robust PCA.

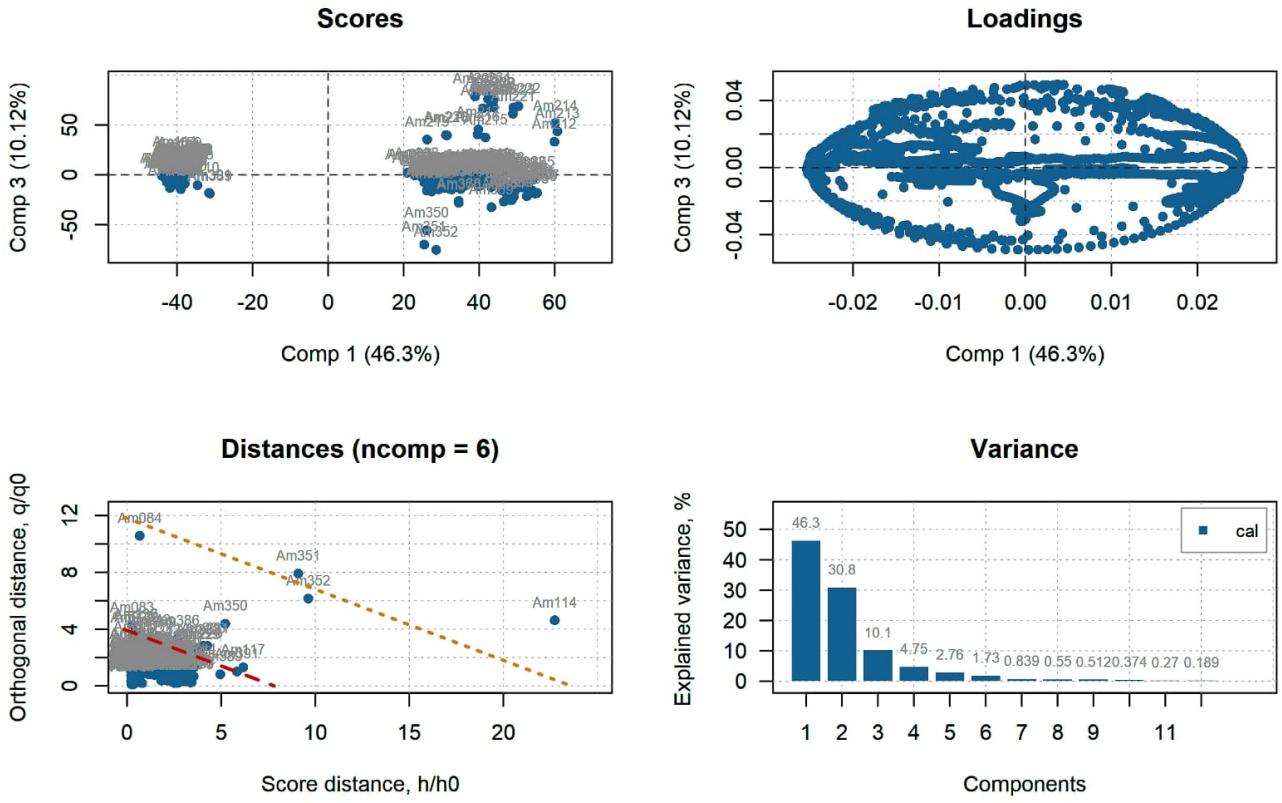


Figure 2. Distances (top left): Orthogonal distance, q versus Score distance, h; Distances (top right): Orthogonal distance, (q/q_0) versus Score distance, (h/h_0); Distances (down left): Orthogonal distance, $\log(1+(q/q_0))$ versus Score distance, $\log(1+(h/h_0))$; Distances (down right): Orthogonal distance, $\log(1+(q/q_0))$ versus Score distance, $\log(1+(h/h_0))$.

The distance plot is well-known and intensively employed. If the distances are spread, applying a simple log transformation of the axes can improve the plot visibility, as shown in Figure 2 (down left: Classical PCA, downright: Robust PCA). Figure 2 shows the distance plot for a classical PCA (top) and robust PCA (top) when the distance values were normalized. The number of outliers remained equal to two regardless of the PCA models used: classic or robust (Am 114 and Am 351). As the outlier number was minimal, it was decided to keep them in the MIR Spectroscopy dataset.

Splitting the datasets into training and test sets

The datasets were split into training and test sets using the function CreateDataPartition from the caret package.[32] The splitting is based on the outcome. For numeric **y**, the sample is split into sections based on percentiles, and sampling is done within these subgroups. The number of percentiles is set via the groups' argument for this function. It was adopted 70% for training and 30% for test sets.

Selecting the prediction model using auto-machine learning

Automated machine learning (Auto-ML) was used to reduce the workload of data training, hyperparameter tuning, and others. The model training process was carried out using the open-source Python library LazyPredict.[22] This library automates the model training pipeline and speeds up the workflow. This Python library contains 43 algorithms for multivariate regression tasks. The Python code line inserts the training and test sets for x (features) and y (response), respectively. The LazyPredict returns the trained models and their performance metrics; it is an advantage. Furthermore, one can compare the performance metrics of each model, and it is possible to tune the best model to improve the performance if desired.

Predicting the CFPP of biodiesel samples using MIR Spectroscopy data

Table III summarizes the results obtained for the predicted models using R^2 and RMSEP as the figures of merit for the CFPP of biodiesel samples. In this table, it was only reported predict models with $R^2 \geq 0.70$.

Table III. Figures of merit from predicting models of CFPP of biodiesel samples using the MIR Spectroscopy data

Model	R^2	RMSEP, ^{0}C	Time Taken, s
Bayesian Ridge	0.95	1.38	0.25
Lasso Lars CV	0.95	1.39	8.73
Ridge	0.94	1.45	0.09
Ridge CV	0.94	1.45	0.10
AdaBoosting Regressor	0.94	1.46	4.17
Orthogonal Matching Pursuit CV	0.94	1.49	0.48
Extra Trees Regressor	0.93	1.60	6.95
Huber Regression	0.93	1.64	0.78
MLP Regressor	0.92	1.69	3.22
Linear SVR	0.91	1.76	1.66
Passive Aggressive Regressor	0.86	2.25	0.23
Tweedie Regressor	0.86	2.26	0.21
Transformed Target Regressor	0.85	2.32	0.22
Linear Regressor	0.85	2.32	0.15
Lasso Lars IC	0.83	2.46	0.35
Decision Tree Regressor	0.81	2.60	0.43
K Neighbors Regression	0.77	2.86	0.08
Random Forest Regressor	0.74	3.06	21.93
Kernel Ridge	0.70	3.31	0.08
Bagging Regressor	0.70	3.31	2.13

Concerning the algorithm's performance based on the highest R^2 and lowest RMSEP, the Bayesian Ridge, Lasso Lars CV, Ridge, Ridge CV, AdaBoosting, Orthogonal Matching Pursuit CV, and Extra Trees Regressor models could be used. They all showed an RMSEP equal or inferior to the experimental error of 1.6 ºC. However, the Ridge and Ridge CV Regressor models had a processing time inferior to or equal to 0.10 s, so they were recommended in this work.

Predicting the kinematic viscosity of biodiesel using MIR Spectroscopy data

Although all the predicted models showed a higher R^2, the RMSEP values were higher than the experimental error (0.0267 $mm^2\ s^{-1}$). Given the results obtained, it is necessary to understand better the behavior of kinematic viscosity to choose how to model using MIR Spectroscopy data.

Concerning the estimation of low-temperature liquid viscosity

Orrick and Erbar[33] and Sastri-Rao[34] are estimation methods for liquid viscosity at low temperatures based on logarithm, which employ structural-sensitive parameters valid only for specific homologous series or are from group contributions. Both methods are widely used and are limited to reduced temperatures, but none of these methods considered are particularly reliable. Neither is reliable for highly branched structures. In using viscosity in engineering calculations, one is often interested not in the dynamic viscosity but in the ratio of the dynamic viscosity to the density. This quantity, called kinematic viscosity, would generally be expressed in $m^2\ s^{-1}$ or stokes. Pure liquid viscosities at high reduced temperatures are usually correlated with variations of the law of corresponding states. At low temperatures, most methods are empirical and involve a group contribution approach. Current liquid mixtures correlations essentially mix rules relating pure component viscosities to composition. The log transformation was applied to the kinematic viscosity values based on these studies. Therefore, it was challenging to model a viscosity of a mixture of fluids using infrared spectral data, particularly with the biodiesel blend samples.

Table IV summarizes the results obtained for the predicted models using R^2 and RMSEP as the figures of merit for the kinematic viscosity of biodiesel samples. In this table, it was only reported predict models with $R^2 \geq 0.84$.

Table IV. Figures of merit from predicting models of kinematic viscosity at 40 ºC of biodiesel samples using MIR Spectroscopy data after log transformation

Model	R^2	RMSEP, $mm^2\ s^{-1}$	Time Taken, s
Extra Trees Regressor	0.94	0.01	11.60
Bayesian Ridge	0.93	0.02	0.28
Ridge	0.92	0.02	0.08
Ridge CV	0.92	0.02	0.09
Nu SVR	0.92	0.02	0.20
Linear Regression	0.92	0.02	1.65
Extra Tree Regressor	0.92	0.02	0.20
Huber Regression	0.91	0.02	1.02
Lasso CV	0.90	0.02	83.50
Orthogonal Matching Pursuit	0.89	0.02	0.45
Lasso Lars CV	0.89	0.02	8.66
Linear Regression	0.89	0.02	0.15
Transformed Target Regressor	0.89	0.02	0.15
Tweedie Regressor	0.88	0.02	0.12
K Neighbors Regressor	0.88	0.02	0.10
Gamma Regressor	0.87	0.02	0.24
LGBM Regressor	0.87	0.02	1.59
Poisson Regressor	0.86	0.02	0.13
Hist Gradient Boosting Regressor	0.86	0.02	14.84

(continues on the next page)

Table IV. Figures of merit from predicting models of kinematic viscosity at 40 ºC of biodiesel samples using MIR Spectroscopy data after log transformation (continuation)

Model	R^2	RMSEP, $mm^2\ s^{-1}$	Time Taken, s
Bagging Regression	0.86	0.02	2.00
Gradient Boosting Regressor	0.84	0.02	13.72

Twenty-one of the predicted models showed a higher R^2, and the RMSEP values were lower or equal to the experimental error (0.0267 $mm^2\ s^{-1}$). However, the Ridge and Ridge CV Regressor models had a processing time inferior to or similar to 0.09 s, which was recommended in this work.

CONCLUSIONS

Using the MIR Spectroscopy dataset, the physical properties of biodiesel and its blends, such as CFPP and kinematic viscosity at 40 ºC, could be modeled by auto-machine learning algorithms.

The CFPP and kinematic viscosity at 40 ºC of the biodiesel samples and their blends could be modeled using MIR Spectroscopy datasets by different auto-machine learning algorithms with an RMSEP similar to the experimental error obtained with a classical procedure in a short time. There is a great advantage because it can predict this property quickly compared to the traditional methodology.

The auto-machine learning algorithms for these modeling were selected based on the figures of merit expressed by the RMSEP (lower value) and R^2 (high value) and by computational time (lower value). This work recommends using the Ridge and Ridge Cross-Validation Regression methods for modeling these properties using the MIR Spectroscopy dataset.

Ridge regression shrinks the coefficients, and it helps to reduce the model complexity and multicollinearity. It was highlighted that the linear regression model with a penalization parameter (L2) could model these properties. However, the kinematic viscosity must be transformed using the log transformation before modeling to get good results. Therefore, it is essential to mention that the Ridge regression is adequate in scenarios where independent variables are highly correlated and occurred with MIR Spectroscopy data. On the other hand, it should be noted that this is the first work in the literature to use automatic machine learning algorithms to predict physical properties in biofuels and their blends using MIR Spectroscopy data.

Conflicts of interest

Regarding conflicts of interest and on behalf of all authors, I declare there are no financial conflicts of interest or lack thereof.

Acknowledgements

The authors are thankful to “Conselho Nacional de Desenvolvimento Científico e Tecnológico” (CNPq), “Fundação de Amparo à Pesquisa no Rio de Janeiro” (FAPERJ), “Coordenação de Aperfeiçoamento de Pessoal de Nível Superior” (CAPES), and “Universidade do Estado do Rio de Janeiro” (Programa Pró-Ciência) for their financial support. ASL has research scholarships from UERJ (Programa Pró-Ciência), FAPERJ (Programa Ciência do Nosso Estado), and CNPq (Bolsista de Produtividade 1D), respectively.

REFERENCES

(1) Cunha, C. L. Application of multivariate calibration methods to predict properties of fossil fuels and biofuels using mid-range infrared spectrum and near-infrared spectroscopy and Raman spectroscopy data. Ph.D. Thesis, Rio de Janeiro State University, Rio de Janeiro, Brazil, **2019**.

(2) Bobadilla, M. C.; Martinez, R. F.; Lorza, R.; L.; Gómez, F. S.; González, E. P. V. Optimizing Biodiesel Production from Waste Cooking Oil Using Genetic Algorithm-Based Support Vector Machines. *Energies* **2018**, *11*, 2995. https://doi.org/10.3390/en11112995

(3) Luque, R.; Lovett, J. C.; Datta, B.; Clancy, J.; Campelo, J. M.; Romero, A. A. Biodiesel as feasible petrol fuel replacement: a multidisciplinary overview. *Energ. Environ. Sci.* **2010**, *3*, 1706-1721. https://doi.org/10.1039/C0EE00085J
(4) FRAGMAQ: Entenda a importância da utilização do biodiesel para o Brasil, suas vantagens e desvantagens. Available at: https://www.fragmaq.com.br/blog/entenda-importancia-da-utilizacao-do-biodiesel-para-o-brasil-suas-vantagens-e-desvantagens/. [Accessed July 2021].
(5) Agência Nacional de Petróleo, Gás Natural e Biocombustíveis. Produção e fornecimento de biocombustíveis. Resolution ANP Number 45. **2014**. Available at: https://www.gov.br/anp/pt-br/assuntos/producao-e-fornecimento-de-biocombustiveis/biodiesel/especificacao-do-biodiesel. [Accessed Jan. 2022].
(6) Lôbo, I. P.; Ferreira, S. L. C.; Cruz, R. S. Biodiesel: parâmetros de qualidade e métodos analíticos. Quím. Nova **2009**, *32* (6), 1596 – 1608. https://doi.org/10.1590/S0100-40422009000600044
(7) Associação Brasileira de Normas Técnicas (ABNT). NBR 10441:2014. Petroleum products – Transparent and opaque liquids – Determination of kinematic viscosity and calculation of dynamic viscosity. São Paulo, **2014**.
(8) Associação Brasileira de Normas Técnicas (ABNT). NBR 14747: 2015. Óleo diesel – Determinação da temperatura de entupimento de filtro a frio. São Paulo, **2015***.*
(9) Pimentel, M. F.; Ribeiro, G. M. G. S.; Cruz, R. S.; Stragevitch, L.; Pacheco-Filho, J. G. A.; Teixeira, L. S. G. Determination of biodiesel content when blended with mineral diesel fuel using infrared spectroscopy and multivariate calibration. *Microchem. J*. **2006**, *82*, 201-206. https://doi.org/10.1016/j.microc.2006.01.019
(10) Baptista, P.; Felizardo, P.; Menezes, J. C.; Correia, M. J. N. Multivariate near-infrared spectroscopy models for predicting the iodine value, CFPP, kinematic viscosity at 40^{0}C, and density at 15 ºC of biodiesel. *Talanta* **2008**, *77*, 144 – 151. https://doi.org/10.1016/j.talanta.2008.06.001
(11) Lira, L. F. B.; Vasconcelos, F. V. C.; Pereira, C. F.; Paim, A. P. S.; Stragevitch, L.; Pimentel, M. F. Prediction of properties of diesel/ biodiesel blends by infrared spectroscopy and multivariate calibration. *Fuel* **2010**, *89* (2), 405- 409. https://doi.org/10.1016/j.fuel.2009.05.028
(12) Balabin, R. M.; Safieva, R. Z. Near-Infrared (NIR) Spectroscopy for Biodiesel Analysis: Fractional Composition, Iodine Value, and Cold Filter Plugging Point from One Vibrational Spectrum. *Energ. Fuels* **2011**, *25* (5), 2373 – 2382. https://doi.org/10.1021/ef200356h
(13) Filgueiras, P. R.; Sad, C. M. S.; Loureiro, A. R.; Santos, M. F. P.; Castro, E. V. R.; Dias, J. C. M.; Poppi, R. J. Determination of API gravity, kinematic viscosity, and water content in petroleum by ATR-FTIR spectroscopy and multivariate calibration. *Fuel* **2014**, *116*, 123 – 130. https://doi.org/10.1016/j.fuel.2013.07.122
(14) Cunha, C. L.; Torres, A. R.; Luna, A. S. Multivariate regression models obtained from near-infrared spectroscopy data for prediction of the physical properties of biodiesel and its blends. *Fuel* **2020**, *261*, 116344. https://doi.org/10.1016/j.fuel.2019.116344
(15) Cunha, C. L.; Luna, A. S.; Oliveira, R. C. G.; Xavier, G. M.; Paredes, M. L. L.; Torres, A. R. Predicting the properties of biodiesel and its blends using mid-FT-IR spectroscopy and first-order multivariate calibration. *Fuel* **2017**, *204*, 185-194. https://doi.org/10.1016/j.fuel.2017.05.057
(16) Neto, H. A. Metodologia de aprendizado AutoML baseado em informações de complexidade de instâncias. Ph.D. Thesis, Minas Gerais Federal University, Brazil, 2017. Available at: http://hdl.handle.net/1843/35575 [Accessed Jan. 2022].
(17) Balaji, A.; Allen, A. Benchmarking Automatic Machine Learning Frameworks. arXiv:1808.06492 [cs.LG] https://doi.org/10.48550/arXiv.1808.06492
(18) Chen, B.; Wu, H.; Mo, W.; Chattopadhyay, I.; Lipson, H. Autostacker: A compositional evolutionary learning system. Proceedings of the 2018 Genetic and Evolutionary Computation Conference (GECCO 2018). Anais, Kyoto, Japan, **2018**.

(19) Data Lab. AutoKeras. Available at: https://autokeras.com/. [Accessed Jan. 2022].
(20) Feurer, M.; Klein, A.; Eggensperger, K.; Springerberg, J. T.; Blum, M; Hutter, F. Auto-sklearn: Efficient and robust automated machine learning. In: Hutter, F.; Kotthoff, L.; Vanschoren, J. (Eds.). *Automated Machine Learning*, pp 113–134, Springer, Germany, **2019**.
(21) Google AutoML. Available at: https://cloud.google.com/automl/docs. [Accessed Jan. 2022].
(22) H2O.AI. H2O. Available at: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html. [Accessed Jan. 2022].
(23) Romblay, A. A. MLBox. Available at: https://mlbox.readthedocs.io/en/latest/. [Accessed Jan. 2022].
(24) Pandala, S. R. Lazy Predict Documentation, release 0.2.9. **2021**. Available at: https://github.com/shankarpandala/lazypredict/issues [Accessed Jan. 2022].
(25) Python Software Foundation. Python Language Reference, version 3.9.5. **2021**. Available at: http://www.python.org [Accessed Jan. 2022].
(26) R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, **2020**. Available at: https://www.R-project.org/ [Accessed Jan. 2022].
(27) RStudio Team. RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. **2020**. Available at http://www.rstudio.com/ [Accessed Jan. 2022].
(28) Silverstein, R. M.; Webster, F. X. *Spectrometric Identification of Organic Compounds*. 6th ed., John Wiley & Sons, Inc., New York, USA, **2005**.
(29) Lilan, K. H.; Almoy, T.; Mevik, B-H. Optimal Choice of Baseline Correction for Multivariate Calibration of Spectra. *Appl. Spectrosc.* **2010**, *64* (9), 1007-1016. https://doi.org/10.1366/000370210792434350
(30) Savitzky, A.; Golay, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* **1964**, *36* (8), 1627–1639. https://doi.org/10.1021/ac60214a047
(31) Rodionova, O. Y.; Kucheryavskiy, S.; Pomerantsev, A. L. Efficient tools for principal component analysis of complex data – a tutorial. *Chemometr. Intell. Lab.* **2013**, *213*, 104304. https://doi.org/10.1016/j.chemolab.2021.104304
(32) Kuhn, M. *Caret: Classification and Regression Training*. R package version 6.0-88. **2021**. Available at: https://CRAN.R-project.org/package=caret [Accessed Jan. 2022].
(33) Orrick, C.; Erbar, J. H. Private Communication to Reid. 1974. In: Poling, B. E.; Praunitz, J. M.; O'Connell, J. P. *The Properties of Gases and Liquids*. 5th ed., McGraw-Hill, Inc., New York, USA, **2001**.
(34) Sastri, S. R. S.; Rao, K. K. A new group contribution method for predicting viscosity of organic liquids. *Chem. Eng. J. Biochem. Eng.* **1992**, *50* (1), 9 – 25. http://ore.immt.res.in/handle/2018/359

SUPPLEMENTARY MATERIAL

Employing auto-machine learning algorithms for predicting the cold filter plugging point and kinematic viscosity at 40 ºC in biodiesel blends using vibrational spectroscopy data.

Table S1. Composition of binary biodiesel blends

Biodiesel 1	% v/v	Biodiesel 2	% v/v
Soybean	10	Corn	90
Soybean	30	Corn	70
Soybean	50	Corn	50
Soybean	70	Corn	30
Soybean	90	Corn	10
Soybean	10	Sunflower	90

(continues on the next page)

Table S1. Composition of binary biodiesel blends (continuation)

Biodiesel 1	% v/v	Biodiesel 2	% v/v
Soybean	30	Sunflower	70
Soybean	50	Sunflower	50
Soybean	70	Sunflower	30
Soybean	90	Sunflower	10
Soybean	10	Canola	90
Soybean	30	Canola	70
Soybean	50	Canola	50
Soybean	70	Canola	30
Soybean	90	Canola	10
Canola	10	South region	90
Canola	30	South region	70
Canola	50	South region	50
Canola	70	South region	30
Canola	90	South region	10

Table S2. Composition of ternary biodiesel blends

Biodiesel 1	% v/v	Biodiesel 2	% v/v	Biodiesel 3	% v/v
Soybean	50	Corn	40	Canola	10
Soybean	40	Corn	30	Canola	30
Soybean	50	Corn	10	Canola	40
Soybean	50	Corn	40	Sunflower	10
Soybean	40	Corn	30	Sunflower	30
Soybean	50	Corn	10	Sunflower	40
Soybean	50	Canola	40	Sunflower	10
Soybean	40	Canola	30	Sunflower	30
Soybean	50	Canola	10	Sunflower	40

Table S3. Composition of quaternary biodiesel blends

Biodiesel 1	% v/v	Biodiesel 2	% v/v	Biodiesel 3	% v/v	Biodiesel 4	% v/v
Soybean	25	Corn	25	Canola	25	Sunflower	25
Soybean	40	Corn	40	Canola	10	Sunflower	10
Soybean	40	Corn	10	Canola	40	Sunflower	10
Soybean	40	Corn	10	Canola	10	Sunflower	40
Soybean	10	Corn	40	Canola	40	Sunflower	10
Soybean	10	Corn	40	Canola	10	Sunflower	40
Soybean	10	Corn	10	Canola	40	Sunflower	40

Table S4. Composition of ternary diesel-biodiesel-anhydrous ethanol blends

Diesel S-10, % v/v	Soybean Biodiesel, % v/v	Anhydrous Ethanol
88.0	10.0	2.0
85.0	10.0	5.0
83.0	10.0	7.0
80.0	10.0	10.0
83.0	15.0	2.0
80.0	15.0	5.0
78.0	15.0	7.0
75.0	15.0	10.0
78.0	20.0	2.0
75.0	20.0	5.0
73.0	20.0	7.0
70.0	20.0	10.0
73.0	25.0	2.0
70.0	25.0	5.0
68.0	25.0	7.0
65.0	25.0	10.0

Table S5. Composition of binary diesel-biodiesel mixtures

Diesel, % v/v			Biodiesel, % v/v					
S-10	**S-500**	**Standard**	**Soybean**	**Canola**	**Sunflower**	**Corn**	**South region**	**RJ**
90.0			10.0					
85.0			15.0					
80.0			20.0					
75.0			25.0					
90.0				10.0				
85.0				15.0				
80.0				20.0				
75.0				25.0				
90.0					10.0			
85.0					15.0			
80.0					20.0			
75.0					25.0			
90.0						10.0		
85.0						15.0		
80.0						20.0		
75.0						25.0		
90.0							10.0	
85.0							15.0	
80.0							20.0	
75.0							25.0	
	90.0		10.0					
	85.0		15.0					
	90.0			10.0				
	85.0			15.0				
	90.0				10.0			
	85.0				15.0			
	90.0					10.0		
	85.0					15.0		
	90.0						10.0	
	85.0						15.0	

(continues on the next page)

Table S5. Composition of binary diesel-biodiesel mixtures (continuation)

Diesel, % v/v			Biodiesel, % v/v					
S-10	**S-500**	**Standard**	**Soybean**	**Canola**	**Sunflower**	**Corn**	**South region**	**RJ**
		90.0	10.0					
		85.0						15.0
80.0								20.0
70.0								30.0

PYTHON CODE

Directory pkgdir = 'D:/CURSO MBA CIÊNCIA DOS DADOS/Disciplina Metodologia e Projeto para Ciências de Dados/TCC/'

```
#!pip install lazypredict
pip install lazypredict
from lazypredict.Supervised import LazyRegressor
import numpy as np
import pandas as pd

# CFPP model using MIR dataset
train1_mir_cfpp = pd.read_csv(pkgdir+'train1_mir_cfpp.csv', sep=';', decimal=',')
y_cfpp_train1 = pd.read_csv(pkgdir+'y_cfpp_train1.csv', sep=';', decimal=',')
test1_mir_cfpp = pd.read_csv(pkgdir+'test1_mir_cfpp.csv', sep=';', decimal=',')
y_cfpp_test1 = pd.read_csv(pkgdir+'y_cfpp_test1.csv', sep=';', decimal=',')
train1_mir_cfpp.shape, y_cfpp_train1.shape
test1_mir_cfpp.shape, y_cfpp_test1.shape
train1_mir_cfpp = train1_mir_cfpp.to_numpy()
y_cfpp_train1 = y_cfpp_train1.to_numpy()
test1_mir_cfpp = test1_mir_cfpp.to_numpy()
y_cfpp_test1 = y_cfpp_test1.to_numpy()

# Fit all models
reg = LazyRegressor(predictions=True)
model_mir_cfpp, predictions = reg.fit(train1_mir_cfpp, test1_mir_cfpp, y_cfpp_train1.reshape((-1,)), y_
cfpp_test1.reshape((-1,)))
print(model_mir_cfpp)

# Kinematic viscosity using MIR dataset
train2_mir_visc = pd.read_csv(pkgdir+'train2_mir_visc.csv', sep=';', decimal=',')
y_visc_train2 = pd.read_csv(pkgdir+'y_visc_train2.csv', sep=';', decimal=',')
test2_mir_visc = pd.read_csv(pkgdir+'test2_mir_visc.csv', sep=';', decimal=',')
y_visc_test2 = pd.read_csv(pkgdir+'y_visc_test2.csv', sep=';', decimal=',')
train2_mir_visc.shape, y_visc_train2.shape
test2_mir_visc.shape, y_visc_test2.shape
```

```
train2_mir_visc = train2_mir_visc.to_numpy()
y_visc_train2 = y_visc_train2.to_numpy()
test2_mir_visc = test2_mir_visc.to_numpy()
y_visc_test2 = y_visc_test2.to_numpy()

# Fit all models
reg = LazyRegressor(predictions=True)
model_mir_visc, predictions = reg.fit(train2_mir_visc, test2_mir_visc, y_visc_train2.reshape((-1,)), y_visc_
test2.reshape((-1,)))
print(model_mir_visc)

# Kinematic viscosity using MIR dataset after log transformation
train2_mir_visc = pd.read_csv(pkgdir+'train2_mir_visc.csv', 68 sep=';', decimal=',') log_y_visc_
train2 = pd.read_csv(pkgdir+'log_y_visc_train2.csv', sep=';', decimal=',') test2_mir_visc = pd.read_
csv(pkgdir+'test2_mir_visc.csv', sep=';', decimal=',') log_y_visc_test2 = pd.read_csv(pkgdir+'log_y_visc_
test2.csv', sep=';', decimal=',')
train2_mir_visc.shape, log_y_visc_train2.shape
test2_mir_visc.shape, log_y_visc_test2.shape
train2_mir_visc = train2_mir_visc.to_numpy()
log_y_visc_train2 = log_y_visc_train2.to_numpy()
test2_mir_visc = test2_mir_visc.to_numpy()
log_y_visc_test2 = log_y_visc_test2.to_numpy()

# Fit all models after log transformation
reg = LazyRegressor(predictions=True)
model_mir_log_visc, predictions = reg.fit(train2_mir_visc, test2_mir_visc, log_y_visc_train2.reshape((-1,)),
log_y_visc_test2.reshape((-1,)))
print(model_mir_log_visc)
```