**Operational condition optimization of *Schinus terebinthifolius* supercritical extraction using machine learning models.**

Ana L. N. Santos,a,c Leonardo O. S. Santana,a Victor L. S. Dias,a  Ana L. B. Souza,a,b Fernando L. P. Pessoaa,b

a SENAI CIMATEC University Center, Avenida Orlando Gomes, Salvador 41650-010, Brazil

b Federal University of Rio de Janeiro (UFRJ), Av. Athos da Silveira Ramos, Rio de Janeiro 21941-909, Brazil.

c Federal University of Bahia (UFBA), R. Prof. Aristídes Novis, Salvador 40210-630, Brazil

laise.nascimento.sant@gmail.com

Abstract

*Schinus terebinthifolius*, or pink pepper, a native Brazilian species, is globally significant for its phytotherapeutic, value in essential oil. This oil, rich in bioactive compounds, offers nutraceutical benefits and has diverse industrial applications, including in food, pharmaceuticals, and pesticides. Supercritical fluid extraction (SFE) using carbon dioxide is a promising alternative to traditional solvent extraction, producing residue-free products that comply with health regulations. Optimizing extraction conditions is crucial for maximizing yield. However, traditional experimentation for optimization is costly and extensive. This study employs machine learning to predict the yield of pink pepper essential oil from supercritical extraction. Data was generated through experimental methods, using a 2³ factorial design with central point in triplicate, further analysed by bootstrap method (200 resampling’s) and Monte Carlo simulation (1000 random data points). Six Machine Learning Regression Methods (Support Vector, K-Nearest Neighbour, Random Forest, Ridge, Lars, and Lasso regression) were employed. The Support Vector Regression with Standard Scaler demonstrated superior accuracy (MAE = 0.1668, MSE = 0.0653, RMSE = 0.2555, NMSE = 0.0374, Pearson = 0.9812), closely mirroring experimental results. This approach enables precise optimization for extracting high-value compounds from pink pepper essential oil in the best extraction conditions to maximize the yield of some compounds of interest. The yield is maximized with lower temperature and higher-pressure conditions, reaching higher yield values when the temperature is around 40 °C and the pressure is around 200 bar.

**Keywords**: *Schinus terebinthifolius*, Machine Learning, Super Critical Fluid Extraction, Support Vector Regression

* 1. Introduction

*Schinus terebinthifolius*, commonly referred to as pink pepper, is a plant of significant medicinal and industrial value, renowned for its bioactive compounds with antioxidant and antimicrobial properties. The extraction of these compounds, particularly through supercritical CO2 extraction methods, has become increasingly relevant due to its environmentally friendly nature and ability to yield high-quality extracts​​. (Araújo dos Santos et al., 2020)

Despite its advantages, the supercritical extraction process presents several experimental challenges. Firstly, the complexity of the extraction process, influenced by various factors such as pressure, temperature, and the nature of the plant material, makes the standardization of extraction conditions difficult​​. High pressures and temperatures, essential for achieving supercritical states, pose operational challenges and safety concerns, necessitating specialized equipment and expertise. Moreover, Brazil's rich biodiversity, including plants like *Schinus terebinthifolius*, adds another layer of complexity due to the varied chemical compositions of these plants, which can significantly influence extraction outcomes​​​​. (De Souza et al.,2023)

Considering these challenges, there is a growing interest in applying machine learning models to predict and optimize extraction yields. Machine learning offers a promising solution to overcome the limitations of traditional experimental approaches. By analyzing complex datasets and identifying patterns, machine learning models can predict the yields of essential oils under various conditions, thereby reducing the need for extensive and costly experimentation. This approach is particularly advantageous given the difficulty in obtaining experimental data and the complexity of controlling and replicating extraction conditions in a laboratory setting.

The training of machine learning models requires a large amount of data. Therefore, statistical methods can be used when the amount of experimental data is small (less than 10 experimental points). Thus, data resampling using the bootstrap method appears as an alternative to increase the amount of data. Furthermore, increasing the amount of data with the bootstrap method can estimate the uncertainty and probability distribution of the data. With the distribution of probability and uncertainty, it is possible to generate an even greater amount of data using Monte Carlo simulation. (Thebelt et al., 2022)

Lastly, this study aims to employ machine learning techniques to predict the yield of *Schinus terebinthifolius* extract from supercritical CO2 extraction and determine the better extraction conditions. By integrating advanced data analysis tools with experimental insights, this research endeavors to enhance the efficiency and effectiveness of the extraction process, contributing to the sustainable and economic production of essential oils.

* 1. Methodology
     1. Dataset

The experimental data was obtained from the experiment done in the work of De Souza et al., (2023). An extensive range of data was generated from this experimental data using a combination of the bootstrap method and Monte Carlo simulation in MS Excel, using Microsoft Excel’s programming language VBA, and the generation of random data tool. The original experimental data was obtained using a 2³ factorial design with the central point (50 °C, 145 bar) in triplicate, and then the bootstrap method was performed (with 200 resamples) using this central point data. From these resampling data, it was assumed that the uncertainty was the same for the other P-T conditions, displayed in Table (1) below.

Table 1. Experimental P-T conditions.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Temperature (°C) | 36 | 40 | 40 | 50 | 50 | 50 | 60 | 60 | 64 |
| Pressure (bar) | 145 | 90 | 200 | 67 | 145 | 223 | 90 | 200 | 145 |

To achieve this, pseudo-experimental data were generated from the bootstrap using a sample of 200 data for each P-T condition. A normal distribution was obtained, and the uncertainty of the central point was generated for the other P-T conditions. then, a Monte Carlo simulation was performed using the standard deviation and mean obtained in the bootstrap, with 1000 randomly generated data.

For the data preprocessing the dataset generated underwent an exploratory analysis, correlation analysis, and distribution assessment. The data was divided into 75% for training and the remaining 25% as test data, never seen by the models during the training. Within the training set, a split of 95% of the data for training and 5% for validation during the training process was made.

* + 1. Implementation of regression Machine Learning models

In this study, regression methods such as support vector regression (SVR) (Drucker, H., et al., 1997), k-nearest neighbors regressor (KNN) (Cover, T., and Hart, P., 1967), random forest regressor (RFR) (Breiman, L., 2001), ridge regression (RDG) (Hoerl, A. E. and Kennard, R. W., 1970), least angle regression (LARS) (Efron, B., et al., 2004), and lasso regression (LASSO) (Tibshirani, R., 1996) were used. Machine learning models typically perform better with standardized data (BISHOP, Christopher M., 2006), hence the implementation of the MinMaxScaler, StandardScaler, and RobustScaler methods as well.

These regression and standardization methods were developed using the scikit-learn library version 1.3.2, in the Python programming language. A total of 24 models were developed, combining the six models with four methods of normalization.

The models underwent an exploratory search to identify the best set of hyperparameters using GridSearchCV, available in the Keras Tuner library. Cross-validation (CV) with 20 “folds” was conducted to compare performance and enhance the models' generalization capability.

* + 1. Evaluation metrics

Initially, the Mean Squared Error (MSE), Eq. (1) below, was used to evaluate the performance of all possible configurations during GridSearchCV.

|  |  |
| --- | --- |
|  | (1 (1) |

In this metric, *n* represents the number of samples, *y* and *y’* are the actual and predicted data, respectively, and *i* indicates the index of the sample.

To assess the best models from each configuration, specifically those with the lowest MSE, additional evaluations were conducted using the Mean Absolute Error (MAE), Eq. (2), Root Mean Squared Error (RMSE), Eq. (3), Normalized Mean Square Error (NMSE), Eq. (4), and Pearson Correlation Coefficient (ρ), Eq. (5).

|  |  |
| --- | --- |
|  | (2) |
|  | (3) |
|  | (4) |
|  | (5) |

For the above statistical metrics, values close to 0 are more suitable for MAE, MSE, RMSE, and NMSE, while values close to 1 are preferred for ρ.

* 1. Results and Discussion
     1. Data Analysis and generation

The dataset generated from the bootstrap and Monte Carlo simulation had a total of 9011 datapoints, for better physical suitability the negative values were removed from the dataset resulting in a total of 8570 datapoints. To analyze the correlation of the target variable which is the extraction yield with the pressure and temperature conditions Pearson and Spearman correlation analysis, that said, for the Pearson correlation the pressure variable showed a positive correlation of 0.8 and for the temperature variable a negative correlation of -0.4. For the Spearman correlation the pressure also presented a positive correlation slightly higher of 0.9 and for the temperature also a negative correlation was observed of -0.3. That means in practical terms that the extraction yield is increased with higher pressures and decreased with higher temperatures.

* + 1. Performance evaluation of the proposed machine learning models

Twenty-four prediction models were created and trained, the metrics of these configurations (model + standardization method) for the yield prediction are displayed in figure 1 bellow.

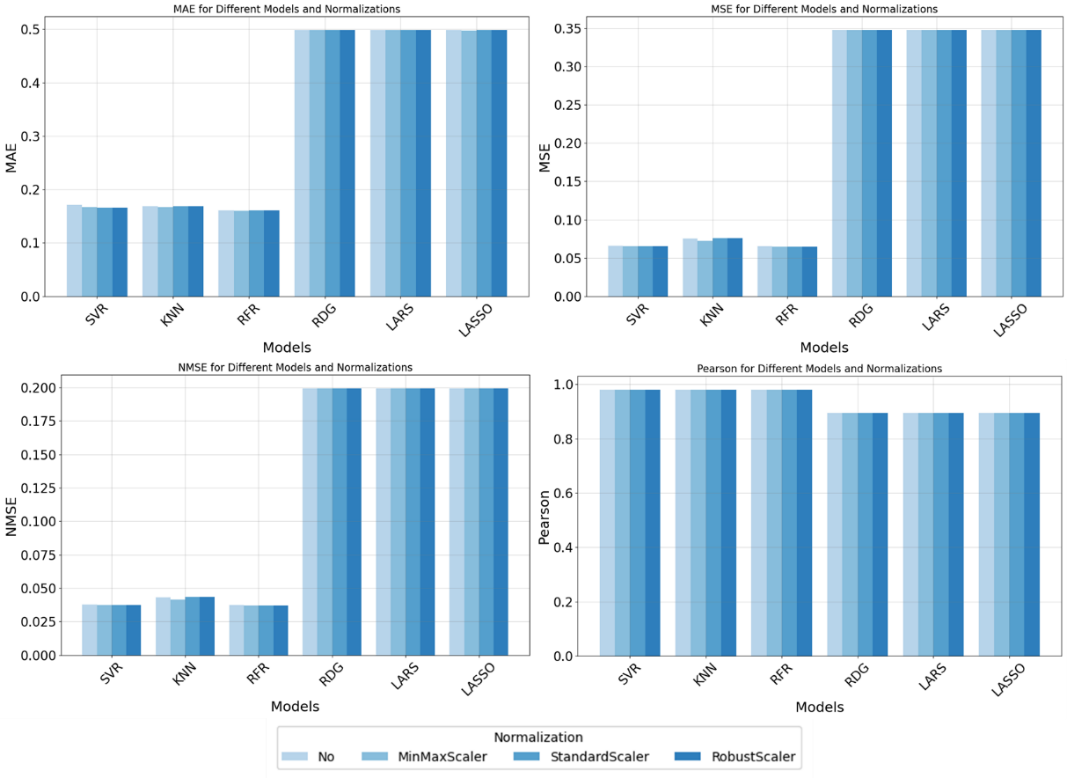


Figure 1. Performance metrics for every model trained.

Figure 1 shows that in the concern of the performance metrics and Pearson correlation the different methods of normalization did not affect the model performance resulting in little variation of the performance. Regarding the different models evaluated SVR + standard scaler normalization showed better overall performance with the lowest errors and higher correlation (MAE = 0.1668, MSE = 0.0653, RMSE = 0.2555, NMSE = 0.0374, and Pearson = 0.9812), this can be explained by the nature of the SVR model that has an advantage predicting non-linear values because of the introduction of a kernel function that project the input data into high dimensional linearly separable space (Ji et al., 2022)

To visualize the prediction results a surface graph was plotted as it can be seen in figure 2 bellow.

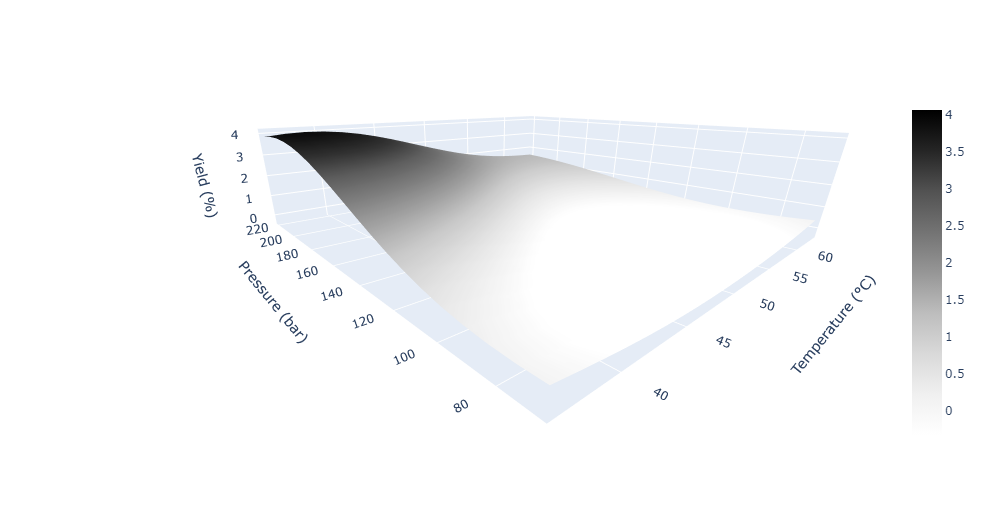


Figure 2. Surface response graph for the SVR model

The yield is maximized with lower temperature and higher-pressure conditions, which corresponds to the Pearson correlation values reaching higher yields values when the temperature is around 40°C and the pressure is around 200 bar.

* 1. Conclusion

The machine learning models demonstrated varying degrees of efficacy in optimizing the yield of essential oils. Among them, certain models, particularly SVR, showed high accuracy across the range of P-T conditions, showing higher Pearson coefficient e higher performance metrics, it was possible to conclude that the temperature hurts the oil yield, this can occur because of the degradation of the material compromising the final result, on the other hand, the increase of pressure has a positive effect on the oil extraction. This work was efficient in showing the best extraction conditions for supercritical extraction, providing a relevant result with three base experiments. This indicates the potential of machine learning models in precisely optimizing extraction conditions to maximize yield. The use of different normalization techniques in the models highlighted the stability and effectiveness of these methods under varying data scaling scenarios.

References

A. C. Araújo dos Santos, G.A.Batista Nascimento, N.Barbosa da Silva,V. Laurent, Y. Valdez,A. Barbosa, E. Calixto, F. Pessoa,2020, Mathematical modeling of the Extraction Process essential oils *Schinus terebinthifolius* Raddi Using Supercritical Fluids, Journal of bioengineering and technology applied to health,2,130-135

A. de Souza, G. Santos, A. Bispo, K. Hodel, B. Machado, D.Chaves, M.Mendes, F. Pessoa, 2023,Technical and Economic Evaluation of Bioactive Compounds from *Schinus terebinthifolius* Using Supercritical Carbon Dioxide, Applied Sciences, 13,21,11897

A. E. Hoerl, R. W. Kennard, 1970, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics, 12, 1, 55-67.

A.Thebelt, J.Wiebe, J.Kronqvist, C.Tsay, R. Misener,2022, Maximizing information from chemical engineering data sets: Applications to machine learning, Chemical Engineering Science,252,117469

B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, 2004, Least angle regression, The Annals of Statistics, 32, 2, 407-499.

C. Ji, F. Ma, J. Wang, W. Sun,2022, Early Identification of Abnormal Deviations in Nonstationary Processes by Removing Non-Stationarity, Computer Aided Chemical Engineering, 49, 1393-1398

C. M. Bishop, 2006, Pattern Recognition and Machine Learning, Springer.

H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, V. Vapnik, 1997, Support Vector Regression Machines, In Advances in Neural Information Processing Systems, 9, pp. 155-161, MIT Press.

L. Breiman, 2001, Random forests, Machine Learning, 45, 1, 5-32.

R. Tibshirani, 1996, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Methodological), 58, 1, 267-288.

T. Cover, P. Hart, 1967, Nearest neighbor pattern classification, IEEE Transactions on Information Theory, 13, 1, 21-27.