REPUBLIC OF TÜRKİYE

YILDIZ TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

ESTIMATION OF BIOCRUDE OIL YIELD BY MACHINE LEARNING

Şerafettin GÜNGÖR

MASTER OF SCIENCE THESIS

Department of Chemical Engineering

Program of Chemical Engineering

Supervisor

Prof. Dr. Hasan SADIKOĞLU

July, 2024

REPUBLIC OF TÜRKİYE

YILDIZ TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

ESTIMATION OF BIOCRUDE OIL YIELD BY MACHINE LEARNING

A thesis submitted by Şerafettin GÜNGÖR in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE** is approved by the committee on 16.07.2024 in Department of Chemical Engineering, MSc Program.

Prof. Dr. Hasan SADIKOĞLU
Yıldız Technical University
Supervisor

Approved By the Examining Committee

Prof. Dr. Hasan SADIKOĞLU, Supervisor
Yildiz Technical University _____

Asst. Prof. Dr. Özgün YÜCEL, Member
Gebze Technical University _____

Assoc. Prof. Dr. Nil ACARALI, Member
Yildiz Technical University _____

I hereby declare that I have obtained the required legal permissions during data collection and exploitation procedures, that I have made the in-text citations and cited the references properly, that I haven't falsified and/or fabricated research data and results of the study and that I have abided by the principles of the scientific research and ethics during my Thesis Study under the title of "Estimation of Biocrude Oil Yield by Machine Learning" supervised by my supervisor, Prof. Dr. Hasan SADIKOĞLU. In the case of a discovery of false statement, I am to acknowledge any legal consequence.

Şerafettin GÜNGÖR

Signature

Dedicated to my family

and my best friend

ACKNOWLEDGEMENTS

I would like to thank my thesis advisors Prof. Dr. Hasan SADIKOĞLU and Res. Asst. Mert Akın INSEL for his invaluable guidance, support, and patience throughout the research and writing of the thesis. Their suggestions and constructive criticism are important in determining this project.

I would also like to thank Yıldız Technical University Chemical Engineering Department for providing the necessary resources and facilities for this study. Special thanks to the faculty and staff who contributed their expertise and helped when needed.

I would like to thank my family for their courage, understanding, and love. Their continued belief in my abilities has been a source of strength and motivation during difficult times.

I am grateful to my friends and colleagues for their support, good discussions, and honest support throughout this process. Their camaraderie makes the pursuit of knowledge truly rewarding.

Finally, I would like to thank the participants of this study, whose willingness to share their experiences and insights was important to the completion of this thesis.

Without the support and encouragement of all those mentioned above, this thesis would not have been possible.

Şerafettin GÜNGÖR

TABLE OF CONTENTS

LIST OF SYMBOLS vi
LIST OF ABBREVIATIONS vii
LIST OF FIGURES viii
LIST OF TABLES x
ABSTRACT xi
ÖZET xiii
1 INTRODUCTION 1
2 BACKGROUND 3
3 LITERATURE REVİEW 5
4 METHODOLOGY 6
4.1 Data Collection 6
4.2 Feature Selection 6
4.2.1 F-Test 6
4.3 Machine Learning 8
4.3.1 Decision Tree 8
4.3.2 K-Nearest Neighborhood 9
4.3.3 XGBoost 10
4.3.4 Gaussian Process Regressor 10
4.3.5 Artificial Neural Network 11
4.4 Validation Method 12
4.4.1 Cross Validation 13
4.5 Hyperparameter Optimization 14
4.6 Performance Metrics 15
4.6.1 R^2 Score 16
4.6.2 Root Mean Square Error (RMSE) 17
5 RESULTS AND DISCUSSION 18
5.1 Feature Selection 18
5.2 Estimation of Biocrude Oil Yield by Machine Learning 19
5.3 Feature Selection 26
6 CONCLUSION 34
REFERENCES 35
PUBLICATIONS FROM THE THESIS 39

LIST OF SYMBOLS

C	Carbon
CO_2	Carbon Dioxide
H	Hydrogen
N	Nitrogen
O	Oxygen
S	Sulphur

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
GPR	Gaussian Process Regressor
HHV	Higher Heating Value
HTL	Hydrothermal liquefaction
KNN	K-Nearest Neighbourhood
ML	Machine Learning
MSE	Mean Squared Error
R^2	Coefficient of Determination
RF	Random Forest
RMSE	Root Mean Squared Error
SAGE	Shapley Additive Global Importance
SHAP	Shapley Additive Explanations
SVM	Support Vector Machine

LIST OF FIGURES

Figure 4.1 Train, test and validation split.[31].....12

Figure 4.2 Cross-validation splits [31].....13

Figure 5.3 Van Krevelen Diagram. The colour scale shows biocrude oil yield. Darker means higher value and vice versa. 18

Figure 5.4 Box plot for each feature. The box plot was created with scaled data. 19

Figure 5.5 Subfigure (a) illustrates the comparison between the predicted values and the actual observed values for the test dataset using the Xgboost algorithm. Subfigure (b) shows the same comparison for the training dataset. 21

Figure 5.6 Subfigure (a) illustrates the comparison between the predicted values and the actual observed values for the test dataset using the K-Nearest Neighbors (KNN) algorithm. Subfigure (b) shows the same comparison for the training dataset. 22

Figure 5.7 Subfigure (a) illustrates the comparison between the predicted values and the actual observed values for the test dataset using the Artificial Neural Network (ANN) algorithm. Subfigure (b) shows the same comparison for the training dataset. 22

Figure 5.8 Subfigure (a) illustrates the comparison between the predicted values and the actual observed values for the test dataset using the Gaussian Process Regressor (GPR) algorithm. Subfigure (b) shows the same comparison for the training dataset. 23

Figure 5.9 Subfigure (a) illustrates the comparison between the predicted values and the actual observed values for the test dataset using the Random Forest (RF) algorithm. Subfigure (b) shows the same comparison for the training dataset. 23

Figure 5.10 Subfigure (a) illustrates the comparison between the predicted values and the actual observed values for the test dataset using the Gaussian Process Regressor (GPR) algorithm. Subfigure (b) shows the same comparison for the training dataset. 24

Figure 5.11 F-test result 26

Figure 5.12 Correlation matrix 27

Figure 5.13 This SHAP graph for the XGBoost model shows each feature's impact on predictions. Higher SHAP values indicate greater influence, with red representing high feature values and blue representing low values. 28

Figure 5.14 This SAGE graph for the XGBoost model illustrates the global importance of each feature. Higher values indicate greater overall influence on the model's predictions, providing insight into the contribution of each feature to the model's performance. 28

Figure 5.15 This SHAP graph for the K-Nearest Neighbor (KNN) model shows each feature's impact on predictions. Higher SHAP values indicate greater influence, with red representing high feature values and blue representing low values.....29

Figure 5.16 This SAGE graph for the K-Nearest Neighbor (KNN) model illustrates the global importance of each feature. Higher values indicate greater overall influence on the model's predictions, providing insight into the contribution of each feature to the model's performance.....29

Figure 5.17 This SHAP graph for the Artificial Neural Network (ANN) model shows each feature's impact on predictions. Higher SHAP values indicate greater influence, with red representing high feature values and blue representing low values.....30

Figure 5.18 This SAGE graph for the Artificial Neural Network (ANN) model illustrates the global importance of each feature. Higher values indicate greater overall influence on the model's predictions, providing insight into the contribution of each feature to the model's performance.....30

Figure 5.19 This SHAP graph for the Gaussian Process Regressor (GPR) model shows each feature's impact on predictions. Higher SHAP values indicate greater influence, with red representing high feature values and blue representing low values.....31

Figure 5.20 This SAGE graph for the Gaussian Process Regressor (GPR) model illustrates the global importance of each feature. Higher values indicate greater overall influence on the model's predictions, providing insight into the contribution of each feature to the model's performance.....31

Figure 5.21 This SHAP graph for the Random Forest (RF) model shows each feature's impact on predictions. Higher SHAP values indicate greater influence, with red representing high feature values and blue representing low values.....32

Figure 5.22 This SAGE graph for the Random Forest (RF) model illustrates the global importance of each feature. Higher values indicate greater overall influence on the model's predictions, providing insight into the contribution of each feature to the model's performance.....32

LIST OF TABLES

Table 4.1 Hyperparameter search range.....14

Table 5.2 Model Metrics.....20

Table 5.3 Models and the model's results. NP*: Not provided, it shows the study does not provide the information. 25

ABSTRACT

Estimation of Biocrude Oil Yield by Machine Learning

Şerafettin GÜNGÖR

Department of Chemical Engineering

Master of Science Thesis

Supervisor: Prof. Dr. Hasan SADIKOĞLU

Biofuel production requires bio-crude oil yield estimation for sustainable resource management, economic feasibility, and regulatory compliance. Accurate projections help investors, optimize processes, support energy security and carbon reduction, and influence technology development, market competitiveness, and environmental impact assessments. This work uses machine learning to forecast bio-crude oil yield to address sustainable energy supply. Extensive datasets are utilized to train various models, including regression, support vector machines, random forests, and neural networks. The results demonstrate the potential for accurate yield forecasting, offering insights for optimizing bio-crude oil production and advancing sustainable energy solutions. This study bridges bioenergy science and practical applications for a more sustainable energy future. This study evaluates XGBoost, Random Forest, SVM, and ANN models for bio-crude oil yield estimation using R^2, MSE, and RMSE. Rigorous preprocessing and training show high R^2 scores, indicating proficiency in capturing complex relationships. The R^2 and RMSE values of 0.86 and 0.0516 demonstrate predictive accuracy. Implications for the bioenergy sector highlight optimization potential and unique model strengths are outlined for tailored applications. Integrating XGBoost, Random Forest, SVM, and ANN, this work informs bioenergy decision-making, offering insights into model performance. This study presents a structured framework for applying machine learning to accurately estimate the yield of bio-crude oil. This

framework has the potential to enhance bioenergy production and have a significant impact on future research and industry in this field.

Keywords: Biocrude oil, machine learning, yield estimation.

ÖZET

Makine Öğrenmesi ile Biyo Kaynaklı Yakıt Verimi Tahmini

Şerafettin GÜNGÖR

Kimya Mühendisliği Anabilim Dalı

Yüksek Lisan Tezi

Danışman: Prof. Dr. Hasan SADIKOĞLU

Biyoyakıt üretimi, sürdürülebilir kaynak yönetimi, ekonomik fizibilite ve mevzuata uygunluk için biyo-ham petrol verimi tahminini gerektirir. Doğru tahminler yatırımcılara yardımcı olur, süreçleri optimize eder, enerji güvenliğini ve karbon azaltımını destekler ve teknoloji gelişimini, pazar rekabet gücünü ve çevresel etki değerlendirmelerini etkiler. Bu çalışma, sürdürülebilir enerji arzını ele almak amacıyla biyo-ham petrol verimini tahmin etmek için makine öğrenimini kullanıyor. Regresyon, destek vektör makineleri, rastgele ormanlar ve sinir ağları dahil olmak üzere çeşitli modelleri eğitmek için kapsamlı veri kümeleri kullanılır. Sonuçlar, biyo-ham petrol üretimini optimize etmeye ve sürdürülebilir enerji çözümlerini geliştirmeye yönelik öngörüler sunan doğru verim tahmini potansiyelini ortaya koyuyor. Bu çalışma, daha sürdürülebilir enerji geleceği için biyoenerji bilimi ile pratik uygulamalar arasında köprü kurmaktadır. Bu çalışma, R^2, MSE ve RMSE kullanılarak biyo-ham petrol verimi tahmini için XGBoost, Random Forest, SVM ve ANN modellerini değerlendirmektedir. Titiz ön işleme ve eğitim, yüksek R^2 puanlarını gösterir, bu da karmaşık ilişkileri yakalamadaki yeterliliği gösterir. 0,86 ve 0,0516'lık R^2 ve RMSE değerleri tahmin doğruluğunu gösterir. Biyoenerji sektörü için çıkarımlar, optimizasyon potansiyelini vurguluyor ve özel uygulamalar için benzersiz modelin güçlü yönleri özetleniyor. XGBoost, Random Forest, SVM ve ANN'yi entegre eden bu çalışma, biyoenerji karar alma

sürecine bilgi vererek model performansına ilişkin içgörüler sunuyor. Bu çalışma, biyo-ham petrol verimini doğru bir şekilde tahmin etmek için makine öğreniminin uygulanmasına yönelik yapılandırılmış bir çerçeve sunmaktadır. Bu çerçeve, biyoenerji üretimini geliştirme potansiyeline sahiptir ve bu alanda gelecekteki araştırmalar ve endüstri üzerinde önemli bir etkiye sahiptir.

Anahtar Kelimeler: Bio kaynaklı yakıt, makine öğrenmesi, verim tahmini.

1 INTRODUCTION

In today's world, where concerns about climate change increase and fossil fuel resources decrease, the search for sustainable energy sources has become more important than usual[1]. Among renewable options, biofuels produced from biomass stand out as the most promising option for reducing carbon monoxide emissions and reducing our dependence on limited resources.

One of the interesting areas in this field is the conversion of biomass into crude oil through processes such as hydrothermal liquefaction (HTL)[2]. However, accurately estimating how much biofuel we can get from different biomass sources is important to improve production and make biofuel production more economical.

Machine learning has revolutionized many industries by providing predictive insights and improving decision-making processes[3].

Integration of machine learning techniques provides a new way to overcome problems related to accurate prediction of bio-crude oil production. By analyzing large datasets containing yield data from biomass, processing, and bio-crude products, learning models can be learned to predict yield more effectively[2].

This article focuses on building predictive models using machine learning algorithms such as regression and decision trees[1]. We aim to improve the conversion process and improve bio-crude oil yield prediction by understanding how factors such as biomass composition, activity, and catalyst usage affect bio-crude oil yield.

As the world moves towards renewable energy, the findings of this article can contribute to the sustainable development of biofuel production technology. By harnessing the power of machine learning, we can pave the way for a greener, more sustainable energy future[3].

With increasing concerns about the environmental impact of depletion and combustion of natural gas, research on alternative fuel sources has gained importance. Among the renewable energy options, biomass energy stands out as a

successful candidate for fuel production that can replace gasoline. The main goal is to increase conversion efficiency, with biochemical and thermochemical processes gradually becoming the main means of converting biomass into energy, fuels, and bioproducts[4].

Hydrothermal liquefaction (HTL) has made significant progress as a technology operating in subcritical/supercritical water to efficiently convert biomass into oil. Bio-oil derived from HTL in particular has high energy density, making it a viable alternative to petroleum. However, these challenges still exist due to the negative properties of bio-oils, such as high viscosity, acidity, and nitrogen content, which cause environmental concerns[4].

Efforts to improve biofuel production from HTL include understanding and controlling key factors such as biomass composition and performance such as temperature and residence time. Although experimental research is useful, it is time-consuming and expensive. The emergence of artificial intelligence, especially machine learning (ML), has a chance to overcome these problems[4].

ML methods are more commonly used to predict biomass conversion behaviour, including HTL. Various machine learning algorithms such as Random Forest (RF) and Deep Neural Networks show promise in predicting bio-oil yields and products with higher accuracy compared to traditional models. Despite this progress, gaps remain in understanding the interaction between biomass composition, performance, and biofuel[5], [6].

Recent research has focused on closing these gaps by developing machine-learning models to predict biomass. - Oil production and products based on different materials. However, post-test guidance is needed to make biofuel products. A new learning model-based optimization has been proposed to guide experimental studies, which provides a good way to achieve energy recovery and low nitrogen content in bio-oil[4].

Integration of machine learning and optimization algorithms is effective in promoting the potential of bio-oil production from biomass HTL. Using data-based insights, scientists aim to improve experimental processes and improve biofuel production, paving the way for a strong and useful future.

2
BACKGROUND

Microalgae have the potential to be a sustainable source of oil production, but many challenges have been encountered on the path to commercial value. The biggest challenge lies in the efficient use of resources such as carbon dioxide and fertilizer[7]. Cultivation systems should make good use of renewable resources to reduce dependence on external resources[8]. Introducing cyanobacteria or artificial microalgae with nitrogen-fixing properties can replace the food supply and reduce the need for purchased fertilizer[8].

The huge demand for nitrogen and phosphate fertilizers has caused a major impact on microalgae biomass production [8]. The environmental impact of fertilizer production, including energy consumption and carbon dioxide emissions, makes the sustainability of microalgae-based biofuel production difficult[8]. Although recovering nutrients from wastewater and using anaerobic digestion for nutrient recovery is promising, proper management is essential to maintain ecosystem balance[8].

It is important to obtain more carbon dioxide in microalgae culture. While it is possible to capture carbon dioxide from industrial sources such as power plants, finding effective ways to use atmospheric carbon dioxide is important to reduce dependence on content[8]. Technologies such as genetic modification to improve carbon dioxide have the potential to increase microalgae productivity[8].

In addition, the choice of carbon source for microalgae culture can affect efficiency and environmental footprint[8]. Although heterotrophic growth provides higher efficiency, challenges such as limited waste materials and expensive bioreactors hinder efficiency[8]. The combination of autotrophic and heterotrophic growth modes leads to stress due to resistance to pollution and reduced productivity[8].

Hydrothermal liquefaction (HTL) has emerged as an effective method to convert microalgal biomass into crude oil[9]. HTL involves heating microalgae in the presence of water at high pressure, leading to the production of bio-crude

oil[9]. Subsequent reforming processes, such as catalytic cracking and hydrotreating, convert bio-crude oil into even more valuable petroleum products[9].

Machine learning provides an effective tool to predict bio-crude oil from microalgae biomass. By analyzing various parameters such as algae type, growth, and HTL process, the learning model can provide better insights into bio-crude fuel efficiency[9]. Using data from research experiments and simulations, these models can improve the accuracy of algorithms[9].

In summary, predicting bio-crude oil production from microalgae biomass through machine learning represents a successful way to make biofuel sustainable. Machine learning can solve problems related to food management, CO2 consumption, and optimization, thus facilitating the development of efficient and consistent technologies for biofuels[9].

3
LITERATURE REVİEW

In the field of biomass conversion, understanding and predicting biofuel yields plays an important role in evaluating the effectiveness and feasibility of processes such as hydrothermal liquefaction (HTL). This study covers process parameters and focusing on the integration of principal component analysis (PCA) and machine learning (ML). Additionally, literature studies that focus on biocrude oil yield were collected in Table 5.3.

PCA has become a simple method for estimating biological crude oil yield by converting complex data including biomass material, activity, and characteristics into meaningful terms. More importantly, PCA compressed more than 90% of the original data, providing a better understanding of the biomass HTL system[10].

Machine learning algorithms, including XGBoost (XGB), Random Forest (RF) and Support Vector Regression (SVR), have attracted attention for their high accuracy in predicting bio-crude oil. Among these, XGB stands out for its ability to simulate complex HTL processes by providing information on biological crude product yield (BY) and higher heating value (HHV). Temperature (T) has become an important factor affecting BY[11].

Recent advances have demonstrated ML-based ensemble optimization models such as Decision Regression Tree (DRT), Gradient Boosting Regression (GBR), and RF for multi-objective prediction and optimization of biological crude product yield. These models provide a good way to conduct experiments and optimize biological crude oil production processes[12]. These improvements pave the way for more efficient biomass conversion.

4
METHODOLOGY

4.1 Data Collection

The data set that is used to train the machine-learning model is from an article [13]. HHV values were obtained from an article using a machine-learning model[14]. The machine learning model's parameters were the dataset that was used to estimate biocrude oil yield. After that, the data was checked to make sure it was good and consistent. Then the data was put on the same scale so that it is easy to process the data. The data was scaled using the min-max scaler, as described in Equation 4.1. After scaling, the data was split into two parts: a training dataset and a test dataset. Both datasets were saved on a hard drive for later use. Consequently, all models used in this study were trained on the same training dataset and tested on the same test dataset. This data then was used to train our model to predict bio-crude oil production. Finally, the model's ability to make predictions based on what it has learned has been examined. This study's goal is to find a reliable way to predict bio-crude oil using machine learning.

$$x_{Scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{4.1}$$

4.2 Feature Selection

4.2.1 F-Test

The F-test stands as a pivotal tool in statistical analysis, especially within regression analysis. Its primary role is to scrutinize the hypothesis that all model parameters equate to zero. Ronald A. Fisher pioneered this test, hence the name F-test, and its foundation lies in the ratio of variances. Formally, the F-statistic is given by the ratio of the explained variance to the unexplained variance, as represented by the equation 4.2 [15].

$$F = \frac{(Explained\ Variance)}{(Unexplained\ variance)} \tag{4.2}$$

In regression analysis, a widely used rule of thumb is that if the F-value is greater than 2.5, it is reasonable to reject the null hypothesis. If the model gives a result that is not zero, then there must be something in the model that is not zero [15].

Interestingly, the F-test offers a means to convert R2, a measure of the proportion of the variance in the dependent variable that is predictable from the independent variables, into a test statistic. This conversion takes the form of a ratio between the average explained variance and the average residual variance. When the explanatory variables collectively fail to augment the fit of a baseline model beyond what might be due to chance, this statistic adheres to an F-distribution. It's worth noting that this reference is not derived from the raw observed data but from a summary measure thereof. Essentially, this comparison contrasts the estimated systematic model against a rudimentary model, typically the mean of the dependent variable Y [16].

However, the choice of this baseline model often stirs debate. Adopting the mean of Y as a model essentially assumes independence between all independent variables X and the outcome Y. Such an assumption, while sometimes useful, can be overly simplistic. In reality, it might be unrealistic to expect Y to be completely independent of all X variables. Thus, the F-test's effectiveness as a goodness-of-fit measure is somewhat tempered by its reliance on an arbitrary or situation-dependent baseline model. Unlike traditional goodness-of-fit tests that use observed data as their natural baseline, the F-test operates more as a model comparison tool. Therefore, while it serves as a model test, the F-test doesn't quite fit the mould of a pure goodness-of-fit test [16].

The F test was used to evaluate the relationship between bio-crude oil yield and individual parameters. It is worth mentioning that the significance of the parameters in terms of their impact on the bio-crude oil yield is more than 5. Therefore, these parameters were maintained, and its performance was verified. Figure 5.11 shows that there is a significant relationship between biocrude oil yield and the other variables.

4.3 Machine Learning

Machine learning has programs that can adjust its performance based on accurate data. Tests were performed using parametric models with adaptive regression. Machine learning is a branch of artificial intelligence (AI) that includes three main categories: supervised learning, unsupervised learning, and reinforcement learning[17].

The learning method consists of algorithms that learn by collecting examples and presenting each input. Supervised learning includes logistic regression, support vector machines, decision trees, and random forests. Unsupervised learning involves algorithms that learn from anonymous examples and evaluate data based on statistical, geometric, or similar patterns. For example, the word k means that combination and speed estimation can be discussed [17].

Reinforcement learning relies on feedback to improve problem-solving. In this case, criticism will show you the quality of your solution, but not how you can improve it. This research specifically focuses on evaluating machine learning in deployment. The first step as a data scientist is to define the problem you want to solve. Select the appropriate method based on the type of response desired [17].

4.3.1 Decision Tree

A decision tree is a method used to approximate discrete objective functions, representing the learned function in the form of a tree structure. In this technique, instances are classified by ordering them from root to specific leaf nodes based on their property values. At each node, a decision or test condition is made for an instance attribute, with each branch representing a possible value for that element. The classification process begins at the root node, known as the decision node, where the tree traverses down along the edge corresponding to the output value of the feature test. This traversal continues in the subtree, headed by a new node at the end of the previous edge until a leaf node is reached, which denotes the final decision or classification category [18].

The key goal when using a decision tree is to determine the best classifier attribute at the level of each node. This decision is often based on statistical measures such as information gain, Gini index, chi-square, and entropy, which are calculated for each node to assess its fitness. Various algorithms are used to implement decision

trees, including classification and regression tree (CART), iterative dichotomizer 3 (ID3), automatic interaction detection (CHAID), chi-square C4.5 and C5.0, and M5. These algorithms help construct decision trees that accurately represent the underlying relationships in the data, making them effective tools for classification and decision-making tasks in machine learning and data analysis [18].

4.3.1.1 Random Forest

In a regression and classification model, the use of random forests involves training multiple tree predictors for particular data sets. By averaging the results of all trees, predictions are made without excessive fitting [8]. The RF regression method uses bootstrapped samples and multiple regression trees to be split. Each tree is constructed from a random subset of input variables. Maximizing RMSE on the training set helps to determine how parameters should be tuned for accurate prediction, including the number of trees and input variables per node [9].

Bagging is used to generate training data in RF regression, which involves growing trees with randomized variable selection at each node. Trees can be pruned to their maximum height without harm, as they are guided by the Gini Index's guideline on variable selection. Numbers are generated by using user-defined parameters, including the number of variables and trees. Reliability and accuracy are ensured by the averaged generalization error in regression tasks [10].

4.3.2 K-Nearest Neighborhood

Classification and regression problems in data science and machine learning are frequently tackled using the KNN algorithm. It operates by identifying the data points that are nearest to a given point and then categorizing or forecasting using the majority class or average value among its closest neighbors. The non-parametric nature of KNN, along with its simplicity and effectiveness, makes it a good fit for tasks that require little knowledge of the data structure. The selection of the optimal number of neighbors (K) and distance measure can have a significant impact on its performance. Additionally, KNN's drawbacks involve computational time, memory requirements, and the difficulty of determining the optimal distance function, which has been tackled by techniques such as ensemble learning and dataset reduction [19], [20].

The k Nearest Neighbor algorithm (kNN) is also used for classification and regression, which works by finding the nearest neighbors of a given data point. It uses the values of its nearest neighbors to predict continuous values in kNN regression. The algorithm's need to store all training data for testing can result in slow execution and excessive memory usage, especially when dealing with large datasets. The main area of research is kNN classification, with little attention given to studying neoclassical data or applying it in a regression model [21].

4.3.3 XGBoost

The XGBoost ensemble learning algorithm's flexibility, predictability, scalability and efficiency have made it a topic of interest in machine learning and data mining [22].

To address data imbalances, researchers have been exploring ways to use XGBoost in conjunction with techniques such as SMOTE and EasyEnsemble to optimize detection and classification tasks. Due to the diverse data distributions, XGBoost often fails to provide optimal results when mixed with over-sampling and under-sampling methods [22].

XGBoost's success is due to its ability to scale and accelerate. In contrast to other popular solutions, it operates at a much faster rate and can process billions of examples even in memory-limited or distributed environments. The scalability is made possible by various optimizations, such as a novel tree learning algorithm for sparse data, utilizing a 'weighted' quantile sketch procedure to handle instance weights and parallel/distributed computing methods [23].

The objective function of XGBoost is composed of an inherent loss function, a regularization term, and specialized complexity control to prevent overfitting. Each sample's value is then iteratively determined from the contribution of each tree and its prediction of another. By applying a Taylor expansion to the loss function in regressive time, one can express the objective function using gradient and Hessian values. To control model complexity and optimize the contribution of trees, these values are used during training to update the parameters and reduce loss [24].

4.3.4 Gaussian Process Regressor

Gaussian Processes (GP), a flexible nonparametric supervised learning technique utilized for regression and probabilistic classification tasks. GP has benefits such as

the ability to estimate values between known data points, provide predictions with associated probabilities, and adjust the parameters of the kernel function. However, its usage is constrained by the lack of sparsity and reduced efficiency in high-dimensional spaces, which restricts its usefulness in specific situations [25].

Gaussian Process Regressor (GPR) is a significant instrument in scientific research, relying on mathematical principles that are fundamental for modelling functions in disciplines such as chemistry and physics. Scientists utilize interpolation and regression methods to connect data points and create continuous function representations, enabling them to comprehend intricate quantum phenomena and gas-particle interactions [26].

The Gaussian Process Regressor is a crucial tool for applying Gaussian processes in regression tasks. GP, or Gaussian Process, uses prior knowledge and covariance functions to find the best parameters and maximize the likelihood of the data. This allows it to make predictions with both the mean and the standard deviation of the data. the model is designed to handle noise level requirements and surpasses standard estimators by allowing prediction without prior fitting and providing techniques for selecting hyperparameters [25].

4.3.5 Artificial Neural Network

Artificial Neural Networks (ANNs) replicate the human brain's information processing, learning from past experiences to identify patterns within data [27]. Chemists are increasingly utilizing ANNs to investigate intricate relationships among multiple variables, capitalizing on their capacity to handle nonlinear dependencies proficiently [28].

Although ANNs can accurately depict linear relationships, traditional statistical methods may often outperform them [28]. Therefore, it is recommended to combine ANNs with conventional statistical methods, particularly when the relationships between variables are not clearly defined [28].

ANNs are composed of interconnected neurons that process input data through layers, ultimately producing output estimations [29]. These networks possess the ability to perform multiple tasks simultaneously, which enhances the efficiency of data analysis [29].

One practical benefit of ANNs is their ability to adapt to hardware, which simplifies data processing [29].

To sum up, ANNs are a powerful tool for analyzing data in analytical chemistry, as they can effectively uncover intricate connections within the data.

4.4 Validation Method

To address the issue of overfitting, which frequently arises in predictive modelling, model validation experts utilize various techniques. The common method is to split the available data into three different groups: the training set, the validation set, and the testing set, which is also known as the holdout method [30]. Figure 4.2 represents how the data split, trained and validated.

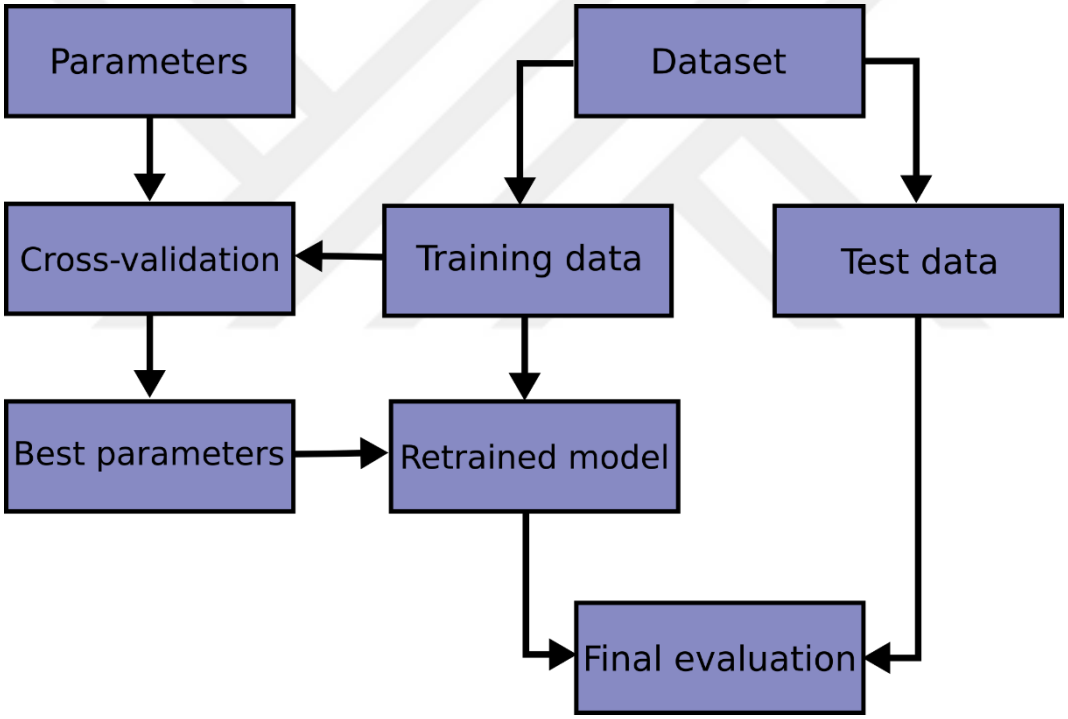


Figure 4.1 Train, test and validation split.[31]

The training set is the starting point for creating diverse models, each with its structure and intricacy. After creating the model, the validation set is vital in choosing the model with the least prediction error, which is commonly referred to as the training error [30].

The selected model is subjected to thorough assessment using the testing set, which is not used during the model development phase. This self-evaluation identifies the

model's ability to generalize or its error rate on data it has not seen before, offering valuable information about its predictive capabilities [30].

4.4.1 Cross Validation

Cross-validation enables us to evaluate the model's performance without relying on the same data that was used during training. The avoidance of overfitting ensures that a model can accurately recall training data without any issues when it needs to be updated [31].

Using the "folds," the dataset is divided into various groups for testing and training, which are then further divided by k-fold cross-validation. Repeating this process multiple times allows us to estimate the model's performance. LOOCV (Leave-One-Out cross-validation) is a testing technique that requires the use of every data point once, but it is known to be both unbiased and computationally expensive [32].

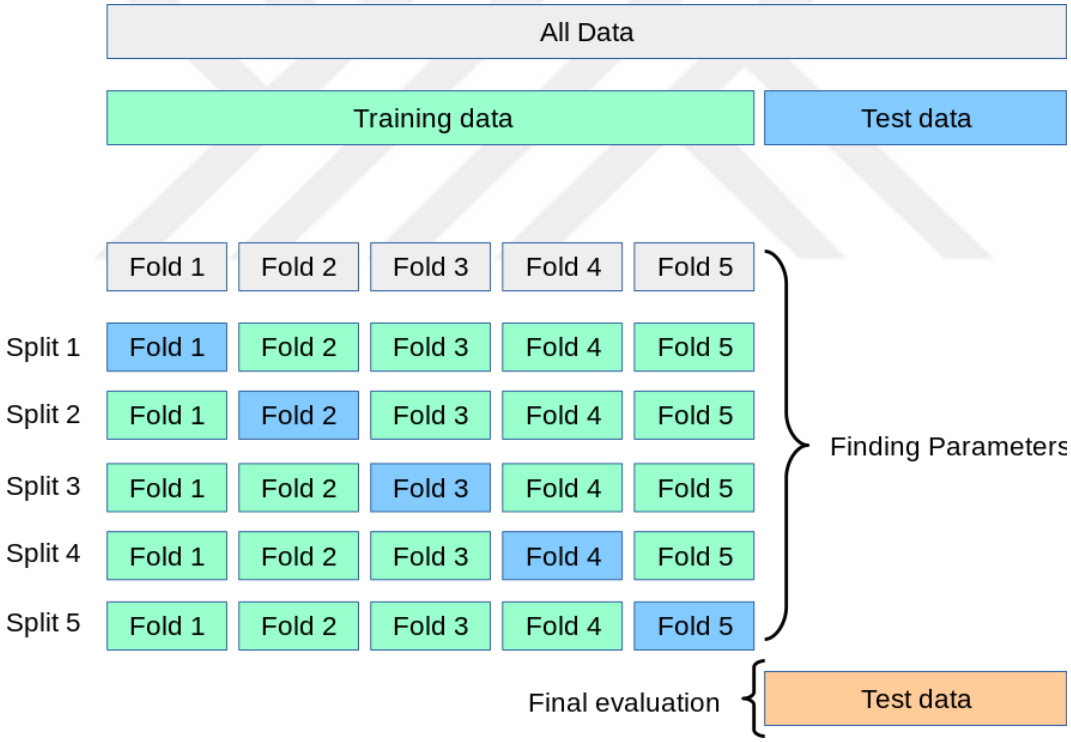


Figure 4.2 Cross-validation splits [31]

Choosing how to measure a model is important. There exist methods that can be misleading in their effectiveness. A single test may not accurately reflect the results of others. Comparing different models can be more complex by using all the predicted values at once when calculating performance metrics [32].

Although cross-validation was first utilized for evaluating linear regression models, it has since become a common practice for testing various model types. The primary aim is to minimize overfitting and measure the applicability of a model when new data is provided. Without requiring new data validation, it is possible to estimate the model's performance through cross-validation [33]. Figure 4.2 shows how cross validation works with 5 folds. This study was used 10-fold cross validation to train all models.

4.5 Hyperparameter Optimization

BayesSearchCV conducts Bayesian optimization on hyperparameters. This tool provides several methods such as "fit" and "score". If the chosen estimator supports them, it also offers additional methods [34].

The hyperparameters for the estimator are optimized through a cross-validated search of parameter settings. A fixed number of parameter settings are sampled from specified distributions, and the number of parameter settings tested is determined by n_iter [34].

Table 4.1 Hyperparameter search rangeindicates hyperparameters search range. Bayes search cv was used to search the best hyperparameters during the training. 200 iterations had done for each model.

Table 4.1 Hyperparameter search range

Model Name	Hyper Parameter Search Range
XGBOOST	• 'n_estimators':from 1 to 150, • 'max_depth': from 5 to 10, • 'gamma': from 1 to 12, • 'reg_alpha' : from 40 to 180, • 'reg_lambda' : from 0 to 10, • 'colsample_bytree' : from 0,5 to 1,0, • 'min_child_weight' : from 0 to 10, • 'seed': from 5 to 15, • 'eta' : [0.3,0.2,0.1,0.05,0.01,0.0001,None], • 'subsample' : from 5 to 10,

Table 4.1 Hyperparameter search range (Continued…)

KNN	• 'n_neighbors': from 1 to 100, • 'weights':['uniform','distance'], • 'algorithm':['auto', 'ball_tree', 'kd_tree', 'brute'], • 'leaf_size': from 1 to 100, • 'p':[1,2], • 'metric':['minkowski','cityblock','euclidean','l1','l2','manhattan'],
ANN	• Dense(units= Increasing by five from 100 to 300., activation= ["relu","leaky_relu","tanh"]), • Dropout(Increasing by 0,1 from 0 to 0,5), • Dense(units= Increasing by five from 0 to 270, activation= ["relu","sigmoid","leaky_relu","exponential","sigmoid","tanh"]), • Dense(activation= ["relu","leaky_relu","sigmoid","tanh"])) • Compile(optimizer=["RMSprop","Nadam"]))
Rando m Forest	• 'n_estimators': from 1 to 150, • 'criterion':['squared_error', 'absolute_error', 'friedman_mse', 'poisson'], • 'max_depth':from 1 to 20 and None, • 'min_samples_split': from 2 to 10,
GPR	• 'kernel_type': ['RBF', 'Matern', 'RationalQuadratic', 'DotProduct_WhiteKernel_Exp']), • 'constant': from 1e-2 to 1e+2), • 'length_scale': from 1e-2 to 1e+2), • 'alpha': from 1e-11 to 1e+2), • 'exponent': from 1 to 5, • 'n_restarts_optimizer': from 0 to 10

4.6 Performance Metrics

When assessing ML regression models, performance metrics are crucial as they compare the model predictions with the actual test data. this sentence: Metrics like

mean squared error, mean absolute error, Pearson correlation coefficient, and coefficient of determination are commonly applied for this objective [35]. While loss functions are used to train models, performance metrics are used to assess the accuracy of the trained models [35].

Regression performance can be measured using functions such as mean_squared_error, r2_score, and explained_variance_score from the sklearn.metrics module [35]. These methods provide adaptability in managing multioutput scenarios and permit diverse averaging techniques. It is worth noting that r2_score and explained_variance_score can be given more importance by weighting them based on their variance, which can highlight the significance of variables with higher variance variables [36].

To accurately assess machine learning models, it is crucial to select appropriate performance metrics. The choice of metric varies depending on the goal of the task, and each metric has a specific function. this sentence: Combined metrics like SAR can offer a holistic view of algorithm performance, while RMSE remains a reliable choice for regression tasks [37],[35].

4.6.1 R^2 Score

R2 is a term that indicates how well a regression model matches the data. It shows how close the regression line is to the real data points. The value of R2 is essential in forecasting future results or validating hypotheses in statistical models [35],[38].

R2 shows how much of the variation in the outcome (y) can be attributed to the factors that are included in the model. This ratio offers an understanding of how well the model fits the data and how well it can predict new, unseen data [35].

Nevertheless, R2 is not a universal measure, as it can give different results depending on the data used. A perfect R2 score means that the model explains all the variation in the y variable, while a negative score means that the model predicts lower y values than a simple model that always predicts the mean y value.

It's important to remember that if the residuals from a prediction model have a mean of zero, the R2 score is the same as the Explained Variance score [35].

In summary, R2 indicates the extent to which a regression model captures the variation in the data. Although it provides valuable information about model

performance, comparing results across various datasets may not always be valid due to its dependence on the specific dataset used [35].

4.6.2 Root Mean Square Error (RMSE)

RMSE, which stands for Root Mean Square Error, is a fundamental measurement utilized in various fields to evaluate the effectiveness of models (5). It is derived from the Mean Squared Error (MSE) by taking the square root, resulting in a unit of measurement equivalent to the predicted values. This characteristic makes RMSE a suitable metric for evaluating model accuracy in cases where the errors are normally distributed [39].

Due to its squared nature, RMSE penalizes larger errors and reflects their impact on the overall error. Therefore, the value of RMSE can vary depending on the distribution of error magnitudes and the size of the sample (5). Despite being sensitive to error variance, RMSE remains a widely used metric for model evaluation due to its ability to provide insights into the overall error magnitude and distribution [39].

It is important to note that RMSE is a standalone metric with its unique advantages and interpretation. As a distance metric, RMSE also satisfies the triangle inequality requirement, further validating its role as a robust evaluation tool [39].

In conclusion, RMSE plays a crucial role in assessing model accuracy, particularly effective in cases where the errors follow a normal distribution. Its calculation and interpretation offer valuable insights into the performance of the model [39].

R2 and RMSE metrics used to measure the model's success are shown in equations 4.3 and 4.4 [12].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_{Pred,i} - y_{Exp,i}\right)^2}{\sum_{i=1}^{n}\left(y_{Exp,i} - y_{Exp,average}\right)^2} \tag{4.3}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{pred,i} - y_{exp,i}\right)^2}{n}} \tag{4.4}$$

5
RESULTS AND DISCUSSION

5.1 Feature Selection

The dataset that it has used to train the models includes HHV values. It makes this study unique. The HHV values were obtained synthetically from an article.[14]Before the training, the F-Test was used to see the relationship between train values and biocrude oil yield. Figure 5.4 is a box plot obtained from the dataset after the dataset scaled.

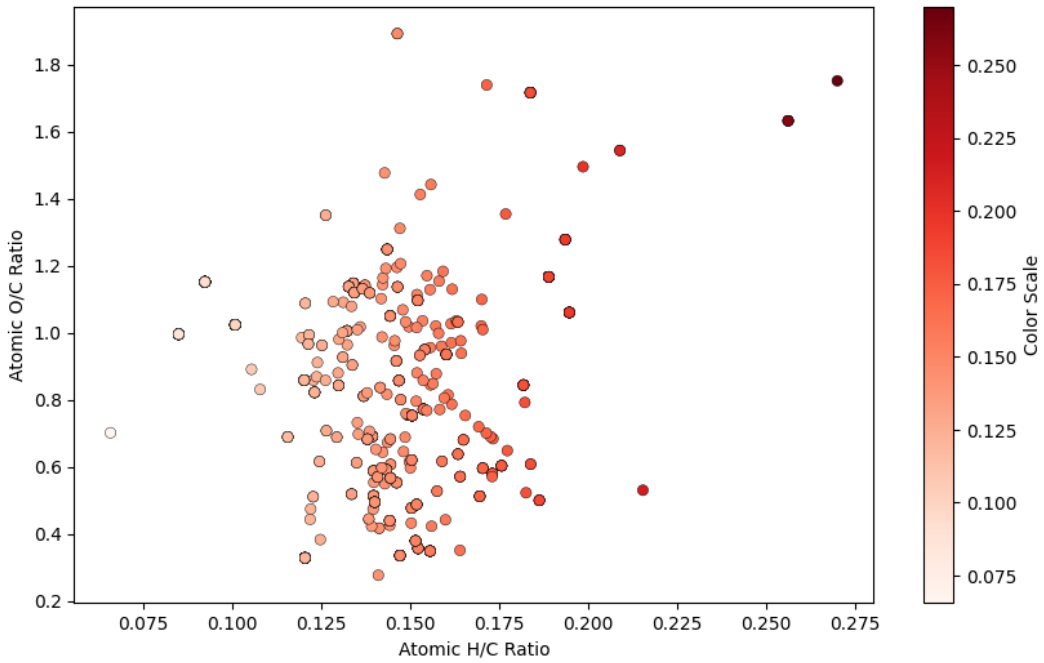


Figure 5.3 Van Krevelen Diagram. The colour scale shows biocrude oil yield. Darker means higher value and vice versa.

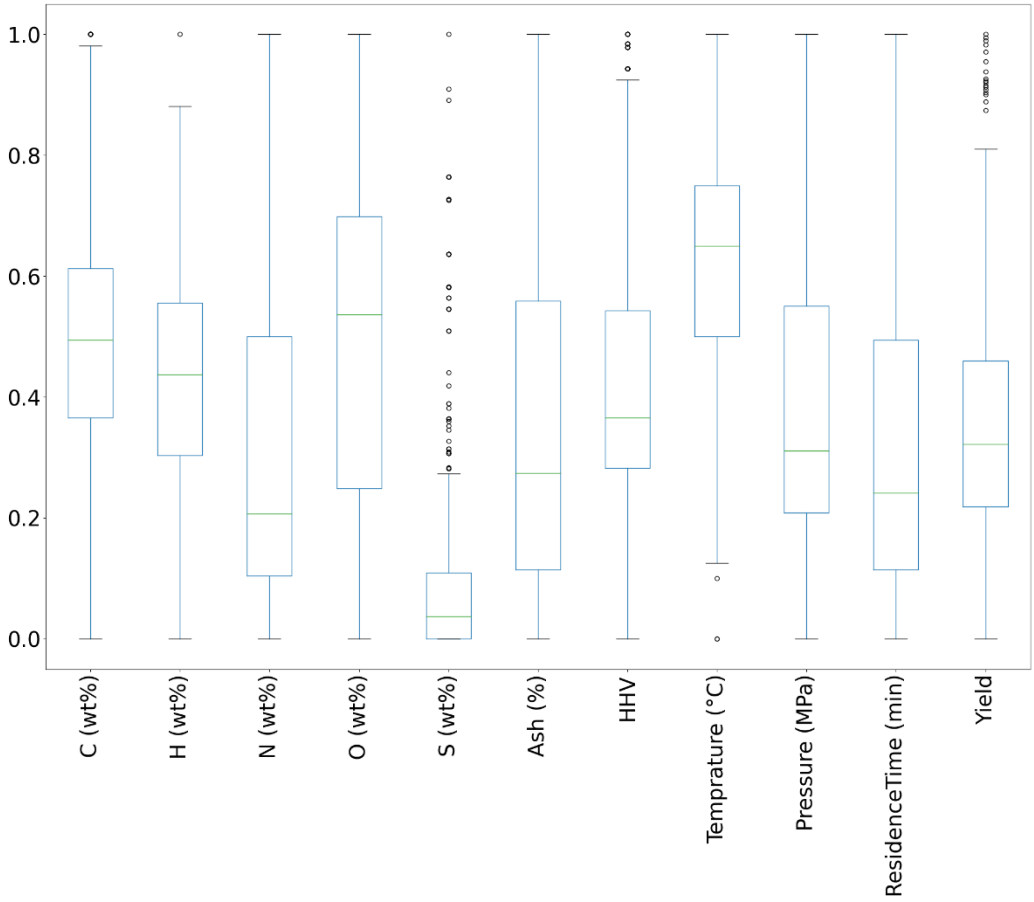


Figure 5.4 Box plot for each feature. The box plot was created with scaled data.

5.2 Estimation of Biocrude Oil Yield by Machine Learning

Our research focused on using machine learning techniques to predict bio-crude oil, and we discovered some important findings. Among the multivariate models used, the xgboost and random forest showed good results with an index R^2 value of 0.86 and an R^2 value of 0.99 for the entire dataset. This shows that it is effective at capturing local data patterns, making it a good choice for predictive work in this area. Additionally, xgboost model achieved the lowest root mean square error (RMSE) of 0.067, further highlighting its accuracy in predicting bio-crude oil. Table 5.2 indicates the model's metrics.

Table 5.2 Model Metrics

Model Name	HyperParameter Optimization	Test R^2	Test RMSE	Train R^2	Train RMSE	Entire Dateset R2	Entire Dataset RMSE
K-Nearest Neighborhood	BayesSearchCV; iterations=200; CV=10	0,85	0,066	0,99	0,019	0,97	0,033
Random Forest	BayesSearchCV; iterations=200; CV=10	0,85	0,069	0,96	0,037	0,93	0,045
XGBoost	BayesSearchCV; iterations=200; CV=10	0,86	0,067	0,99	0,019	0,97	0,034
GPR	BayesSearchCV; iterations=200; CV=10	0,84	0,071	0,99	0,022	0,96	0,037
ANN	BayesSearchCV; iterations=200; CV=10	0,81	0,077	0,95	0,042	0,92	0,051

The random forest model also performed well, with the test R^2 value of 0.85 and an R^2 value of the entire dataset of 0.96 In addition, it is similar xgboost model in terms of performance, its ability to learn data is obvious. However, it is worth noting that the random forest model reports a higher RMSE of 0.069 compared to the xgboost model

Similarly, the XGBoost model, given the results, offers a slightly lower R^2 value of 0.86. Its RMSE is 0.067, still one of the best ones in the comparison.

The GPR algorithm's performance is in between the other models. 0.84 R^2 value and 0.071 RMSE value make the model worthwhile as much as the other machine learning algorithms in this study.

ANN model shows relatively low performance against the other models. The R^2 and RMSE values are 0.81 and 0.077, respectively. Machine learning algorithms may be more beneficial for biocrude oil yield estimation. For the reason that, training machine learning algorithms consume fewer sources than neural networks. [40] Moreover, machine learning algorithms run faster than neural networks [40].

These findings highlight the importance of using multiple methods to understand bio-crude oil prediction. By knowing the advantages and disadvantages of each model, we can improve the accuracy of prediction and lay a foundation for future research in prediction models and data analysis in the field of bio-crude oil prediction. Overall, our study provides useful information that can inform practical applications and further research in this field.

A van Krevelen diagram is a graphical representation of organic matter composition based on the ratios of hydrogen-to-carbon (H/C) and oxygen-to-carbon (O/C). It allows for visualization of the elemental composition of organic compounds, aiding in the characterization of natural organic material. Researchers use van Krevelen diagrams to understand natural organic material diversity, identify chemical classes, and compare samples across different studies. Its utility lies in providing insights into natural organic material composition and facilitating the interpretation of complex data in characterization studies [41]. The dataset that is used to train in this study is characterized in Figure 5.3.

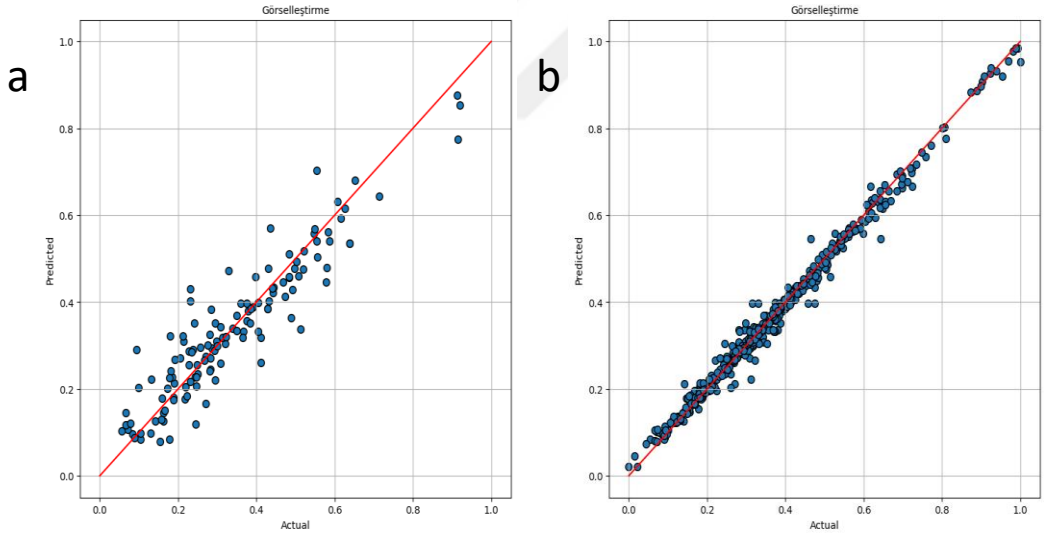


Figure 5.5 Subfigure (a) illustrates the comparison between the predicted values and the actual observed values for the test dataset using the Xgboost algorithm. Subfigure (b) shows the same comparison for the training dataset.

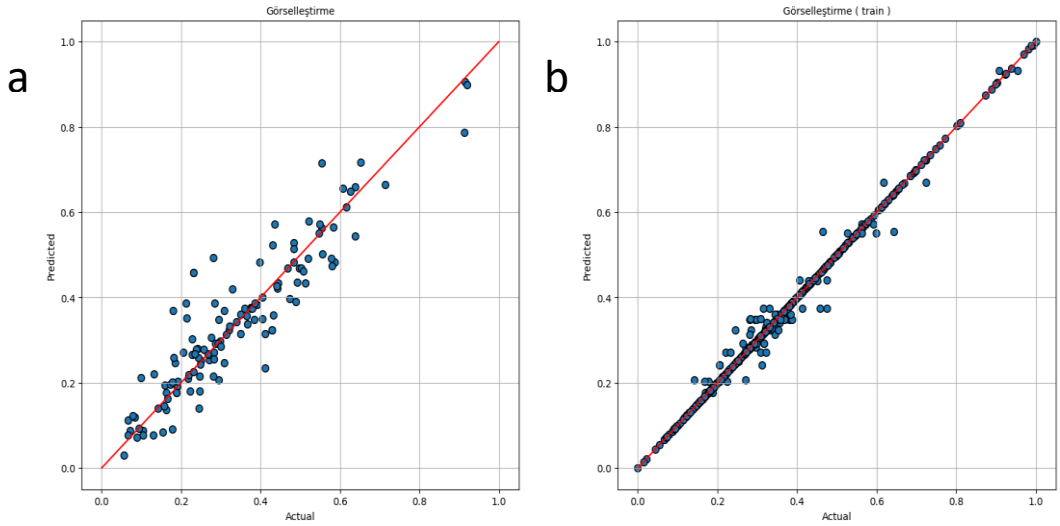


Figure 5.6 Subfigure (a) illustrates the comparison between the predicted values and the actual observed values for the test dataset using the K-Nearest Neighbors (KNN) algorithm. Subfigure (b) shows the same comparison for the training dataset.

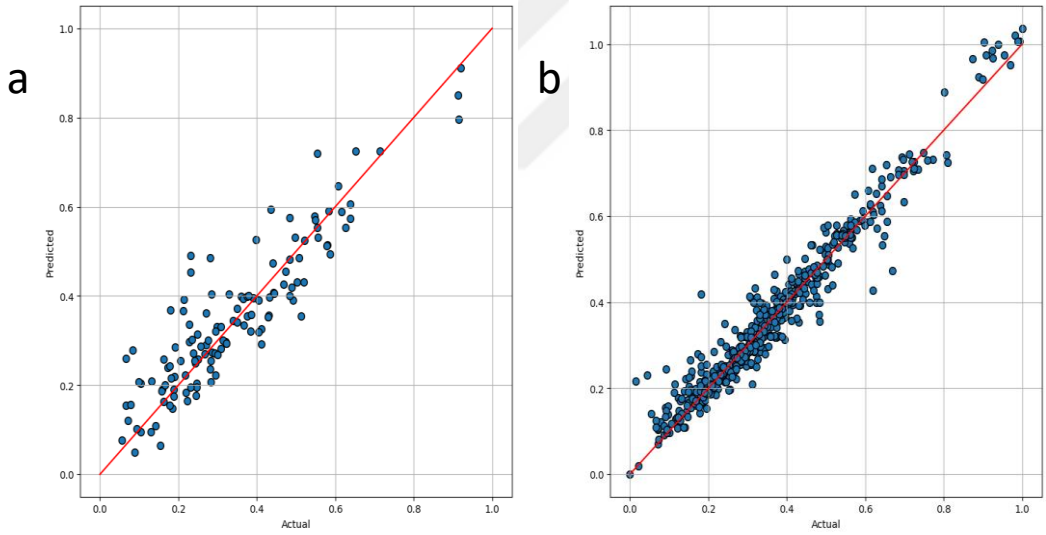


Figure 5.7 Subfigure (a) illustrates the comparison between the predicted values and the actual observed values for the test dataset using the Artificial Neural Network (ANN) algorithm. Subfigure (b) shows the same comparison for the training dataset.

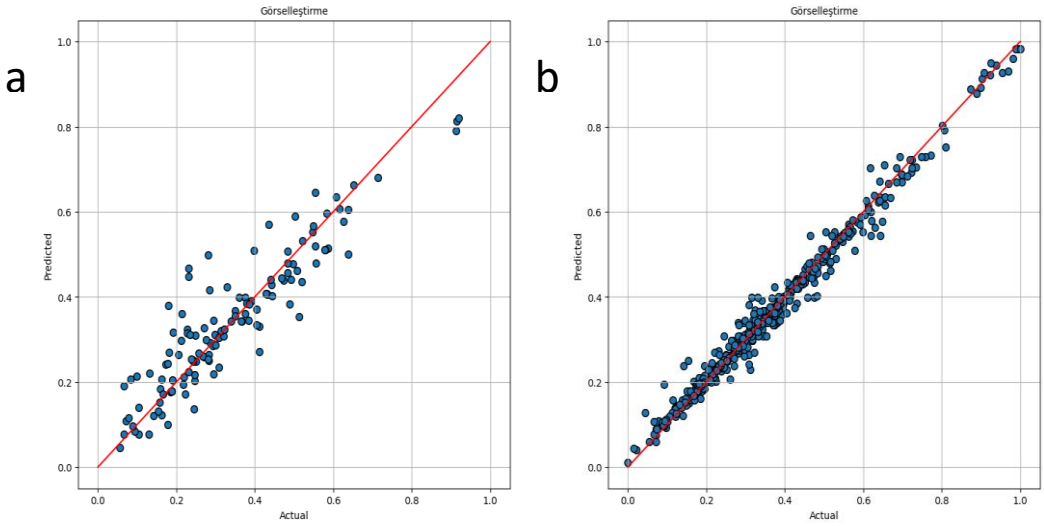


Figure 5.8 Subfigure (a) illustrates the comparison between the predicted values and the actual observed values for the test dataset using the Gaussian Process Regressor (GPR) algorithm. Subfigure (b) shows the same comparison for the training dataset.

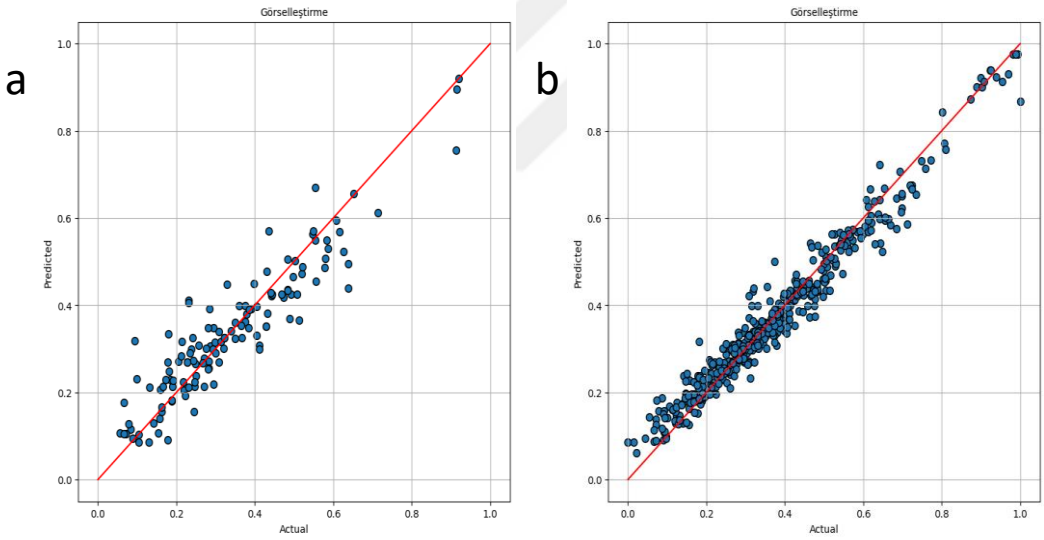


Figure 5.9 Subfigure (a) illustrates the comparison between the predicted values and the actual observed values for the test dataset using the Random Forest (RF) algorithm. Subfigure (b) shows the same comparison for the training dataset.

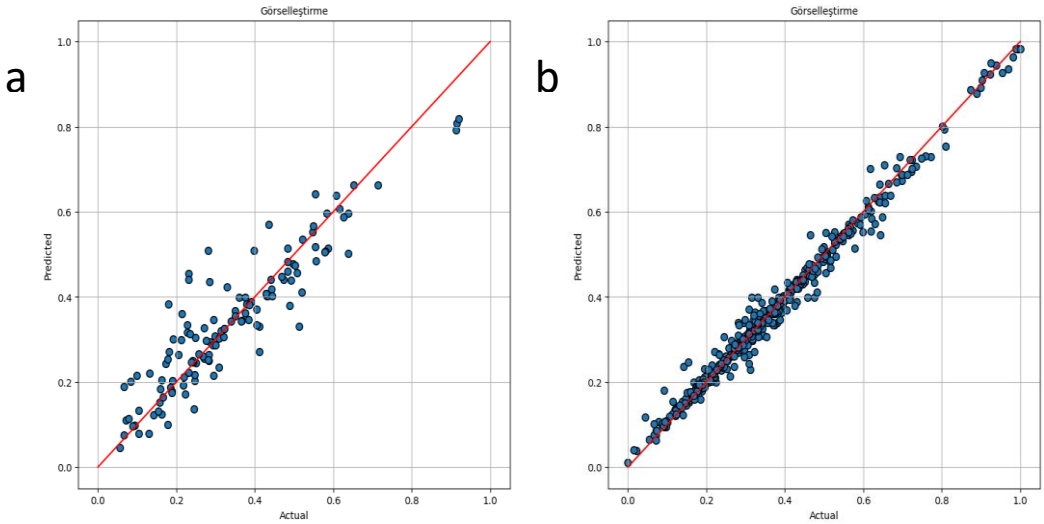


Figure 5.10 Subfigure (a) illustrates the comparison between the predicted values and the actual observed values for the test dataset using the Gaussian Process Regressor (GPR) algorithm. Subfigure (b) shows the same comparison for the training dataset.

When discussing the prediction of bio-crude oil production through machine learning, we delve into the complex processes based on biomarkers and biomass properties. Factors such as final analysis and process failure play an important role in accurately predicting results. Machine learning techniques such as random forests and decision trees help show relationships between these factors, improving prediction accuracy.

We use large amounts of training data and powerful validation tools to ensure the reliability of the machine-learning model. Despite these efforts, challenges remain, including limitations related to data availability and translation issues. However, machine learning models consistently outperform traditional methods regarding accuracy and adaptability.

The discussion highlighted the importance of accurately predicting biofuel production to optimize biomass conversion and achieve usage targets. Electricity continues. It highlights the need for collaborative collaboration and the use of advances in machine learning to solve problems and promote effective solutions. By addressing these problems, we can pave the way for more efficient and green energy.

Table 5.3 Models and the model's results. NP*: Not provided, it shows the study does not provide the information.

Data Size	Dataset	Method	Fuel Class	Validation Method	Performance Metric	Year	Ref.
250	UP	PCA+ML	Biomass	10-CV	R2=0.8 RMSE=0.005	2023	[10]
325	UP	XGB	Biomass	10-CV	R2=0.87 RMSE=0.0446	2022	[11]
325	UP	Random Forest	Biomass	10-CV	R2=0,84 RMSE=0,0487	2022	[11]
325	UP	Kernel Ridge Regression	Biomass	10-CV	R2=0,82 RMSE=0,0878	2022	[11]
325	UP	Support Vector Machine	Biomass	10-CV	R2=0,85 RMSE=0,0474	2022	[11]
448	UP	GBR	Biomass	5-CV	R2=0.9 RMSE=0.00614	2021	[12]
36	P	Multi-Component Linear Additivity Model	Biomass	NP*	R2=95.62	2018	[42]
652	UP	GPR	Biomass	5-CV	R2=92.6 MAE<0.031	2022	[13]
NP*	UP	Nonlinear Regression	Biomass	NP*	R2=94.3	2019	[43]

Table 5.3 Models and the model's results. NP*: Not provided, it shows the study does not provide the information. (Continued…)

NP*	U	Nonlinear Regression	Biomass	NP*	R2=98.1	2018	[44]
652	UP	KNN	Biomass	10-CV	R2=97.0 RMSE=0.034	2024	This Study
652	UP	Random Forest	Biomass	10-CV	R2=0.93 RMSE=0.045	2024	This Study
652	UP	Xgboost	Biomass	10-CV	R2=0.97 RMSE=0.034	2024	This Study
652	UP	GPR	Biomass	10-CV	R2=0.96 RMSE=0.037	2024	This Study
652	UP	ANN	Biomass	10-CV	R2=0.92 RMSE0.051	2024	This Study

5.3 Feature Selection

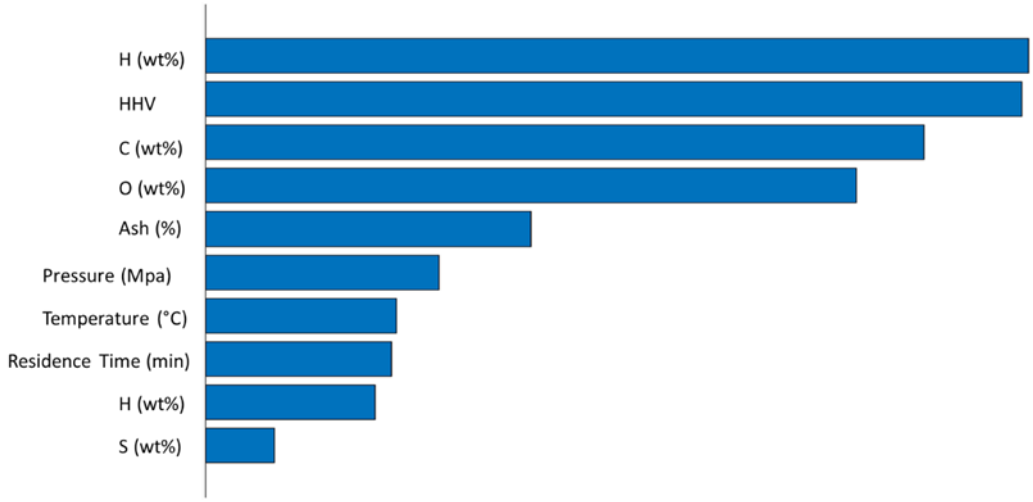


Figure 5.11 F-test result

Figure 5.12 represents a correlation between variables. The scale ranges from 0.0 to 1.0. High value means high correlation and vice versa. It helps to determine whether or not to keep variables in the model.

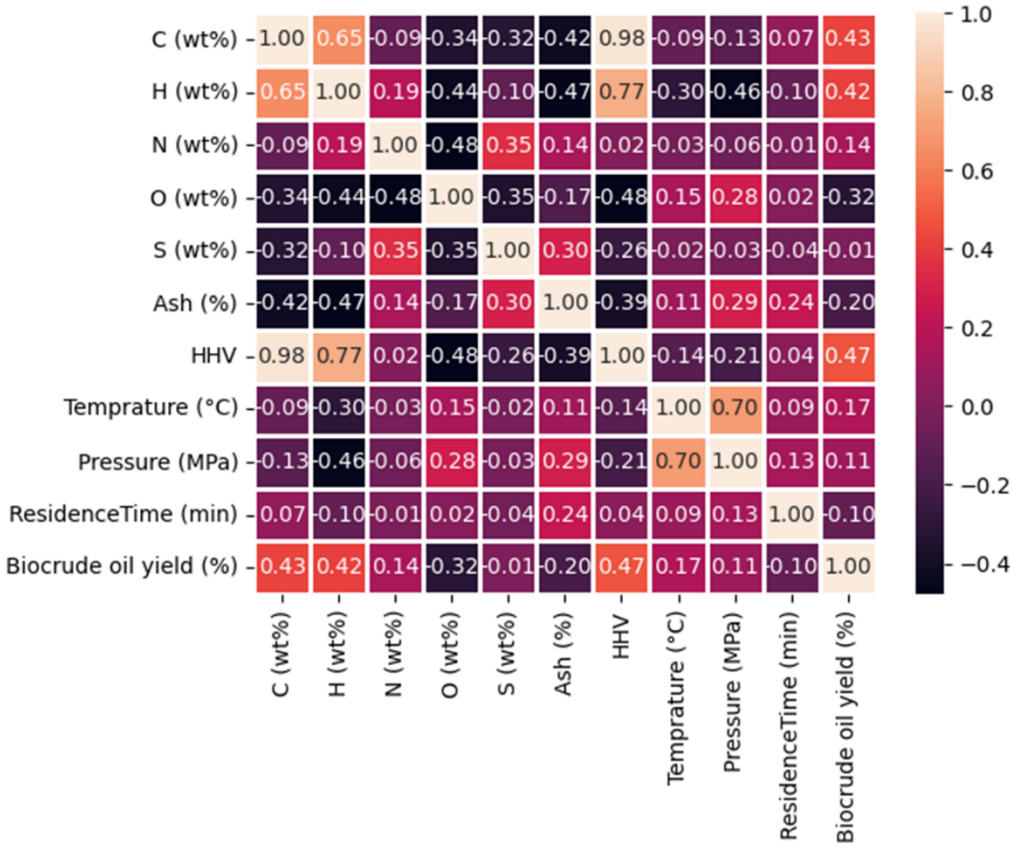


Figure 5.12 Correlation matrix

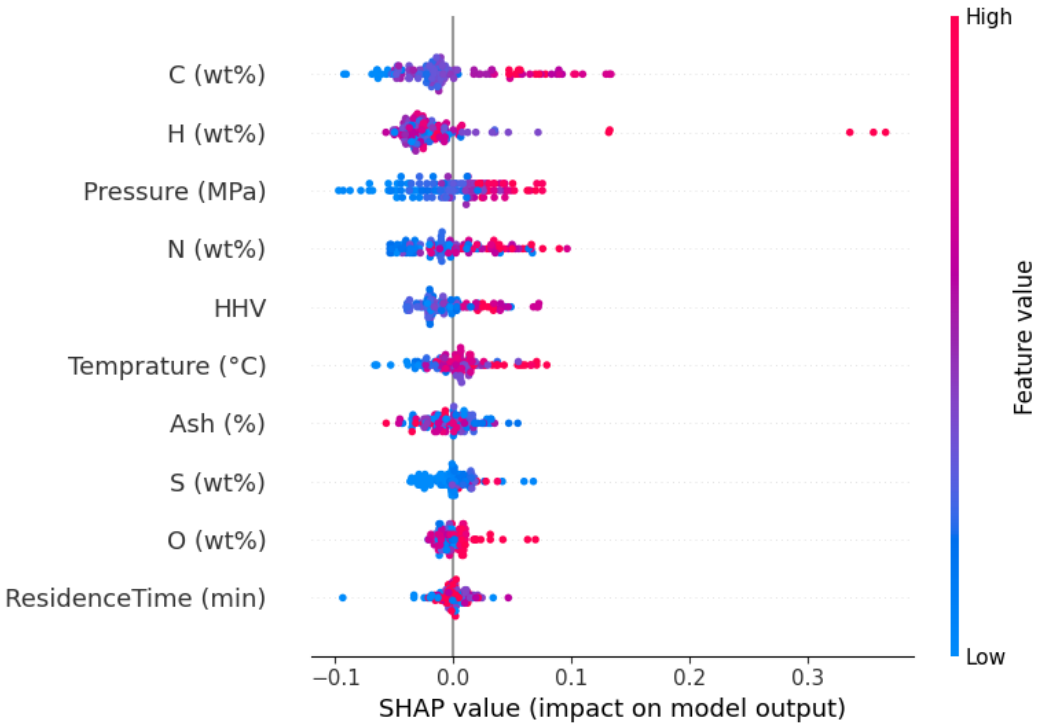


Figure 5.13 This SHAP graph for the XGBoost model shows each feature's impact on predictions. Higher SHAP values indicate greater influence, with red representing high feature values and blue representing low values.

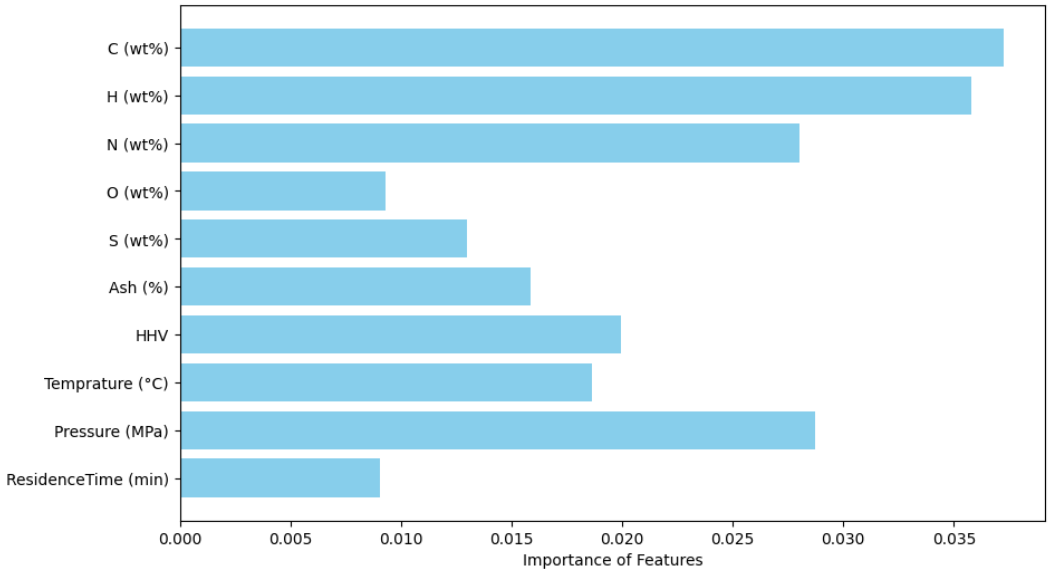


Figure 5.14 This SAGE graph for the XGBoost model illustrates the global importance of each feature. Higher values indicate greater overall influence on the model's predictions, providing insight into the contribution of each feature to the model's performance.

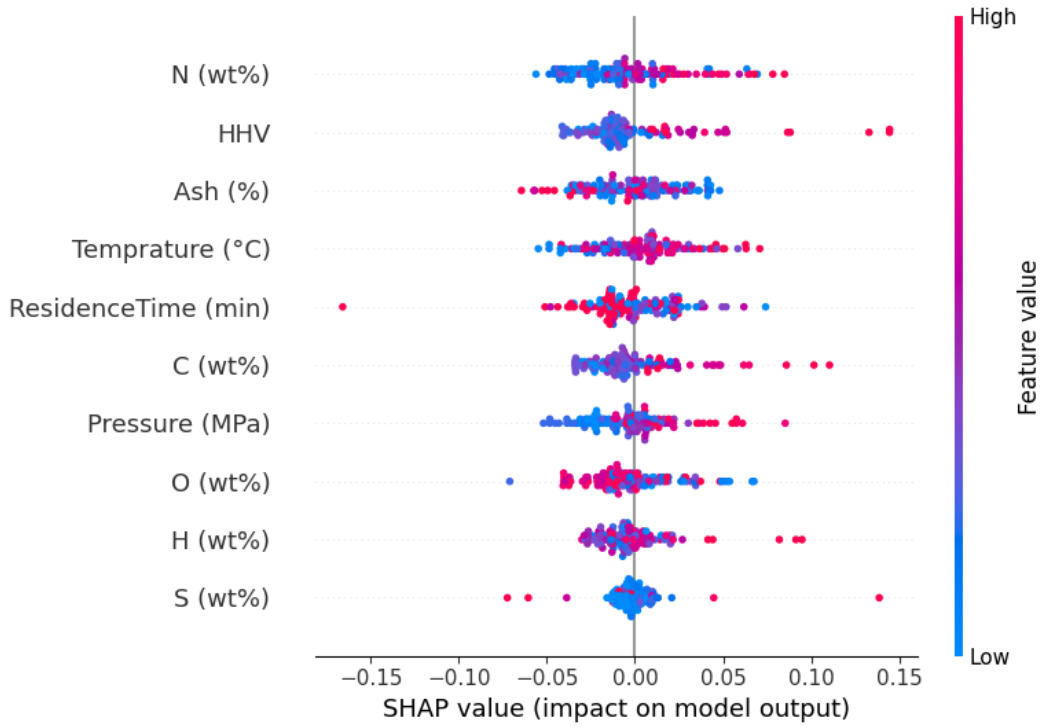


Figure 5.15 This SHAP graph for the K-Nearest Neighbor (KNN) model shows each feature's impact on predictions. Higher SHAP values indicate greater influence, with red representing high feature values and blue representing low values.

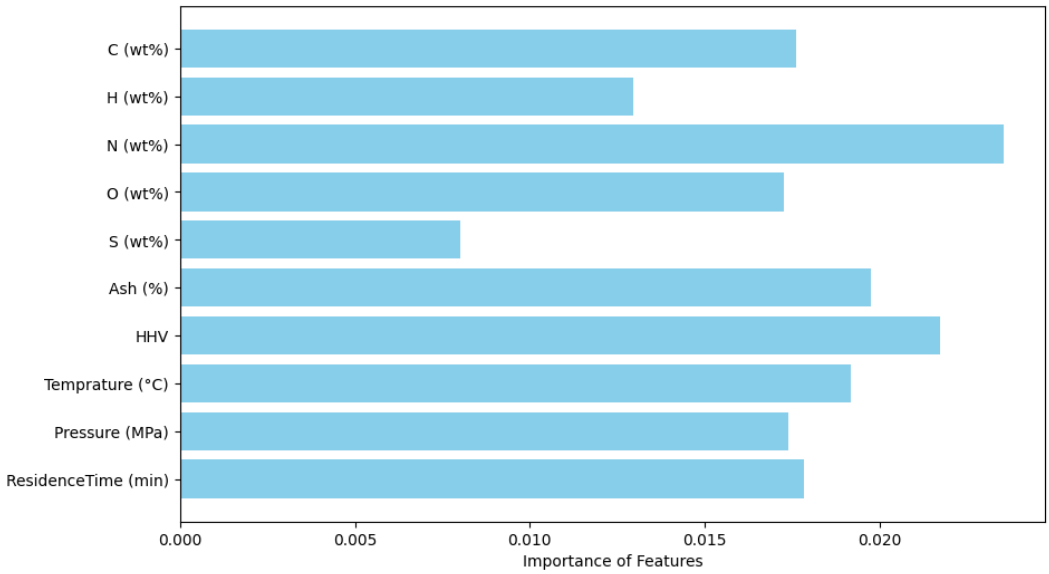


Figure 5.16 This SAGE graph for the K-Nearest Neighbor (KNN) model illustrates the global importance of each feature. Higher values indicate greater overall influence on the model's predictions, providing insight into the contribution of each feature to the model's performance.

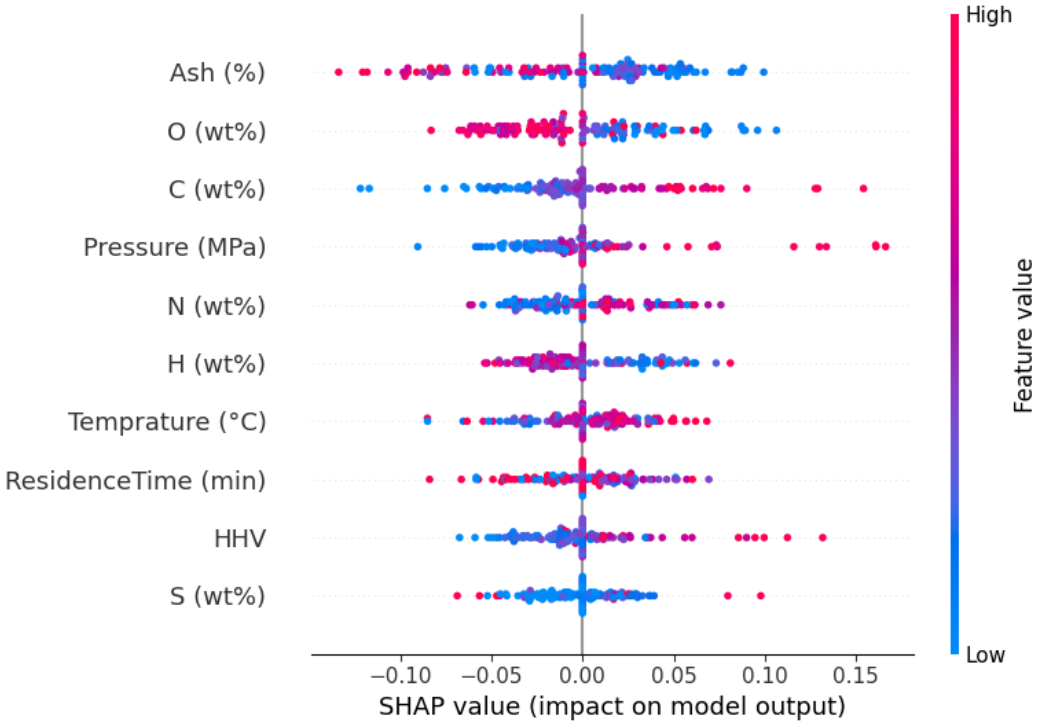


Figure 5.17 This SHAP graph for the Artificial Neural Network (ANN) model shows each feature's impact on predictions. Higher SHAP values indicate greater influence, with red representing high feature values and blue representing low values.

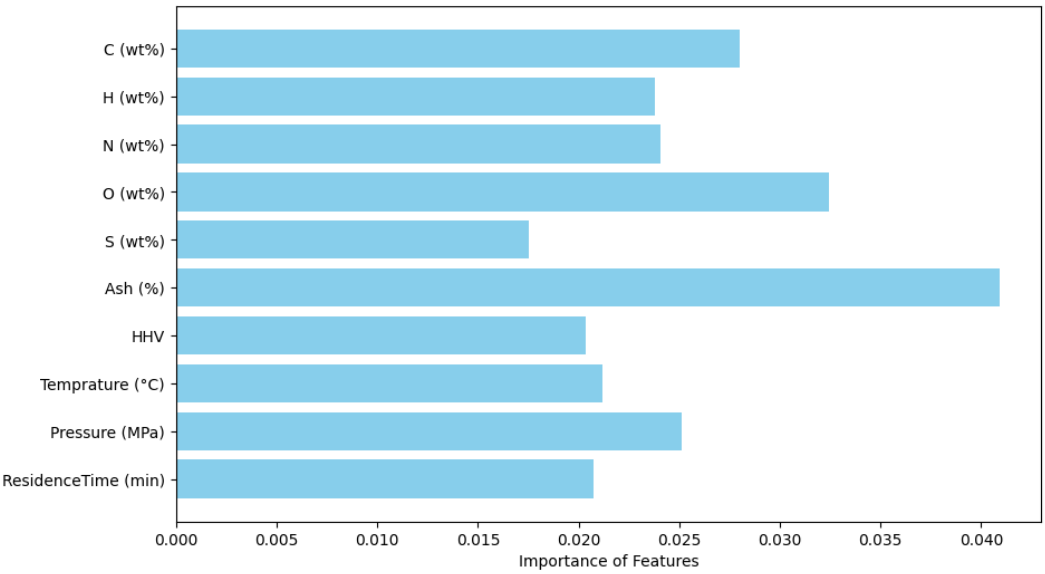


Figure 5.18 This SAGE graph for the Artificial Neural Network (ANN) model illustrates the global importance of each feature. Higher values indicate greater overall influence on the model's predictions, providing insight into the contribution of each feature to the model's performance.

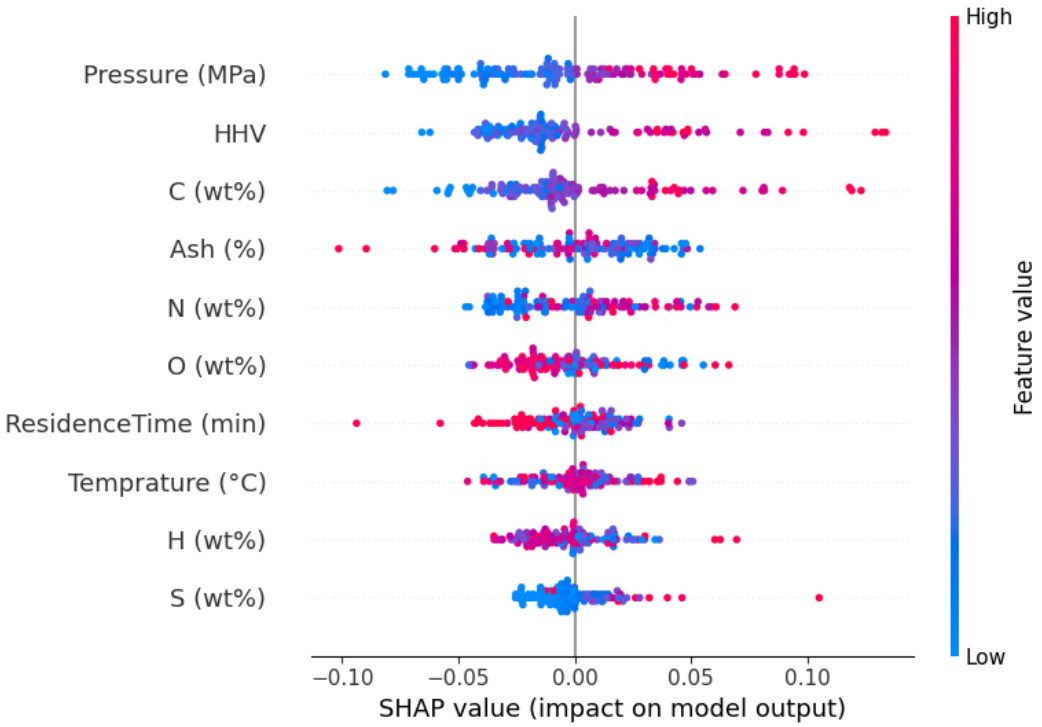


Figure 5.19 This SHAP graph for the Gaussian Process Regressor (GPR) model shows each feature's impact on predictions. Higher SHAP values indicate greater influence, with red representing high feature values and blue representing low values.

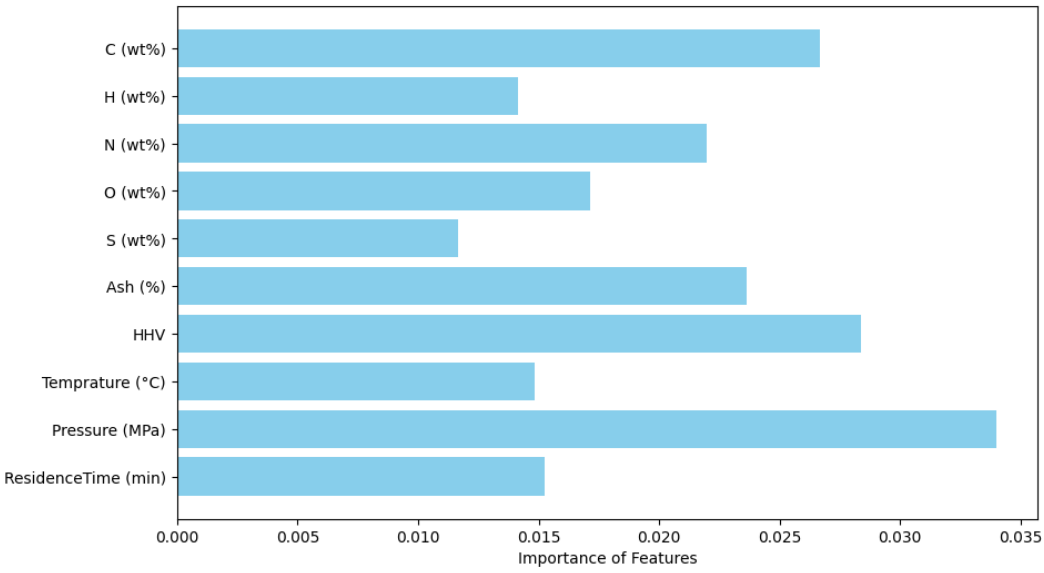


Figure 5.20 This SAGE graph for the Gaussian Process Regressor (GPR) model illustrates the global importance of each feature. Higher values indicate greater overall influence on the model's predictions, providing insight into the contribution of each feature to the model's performance.

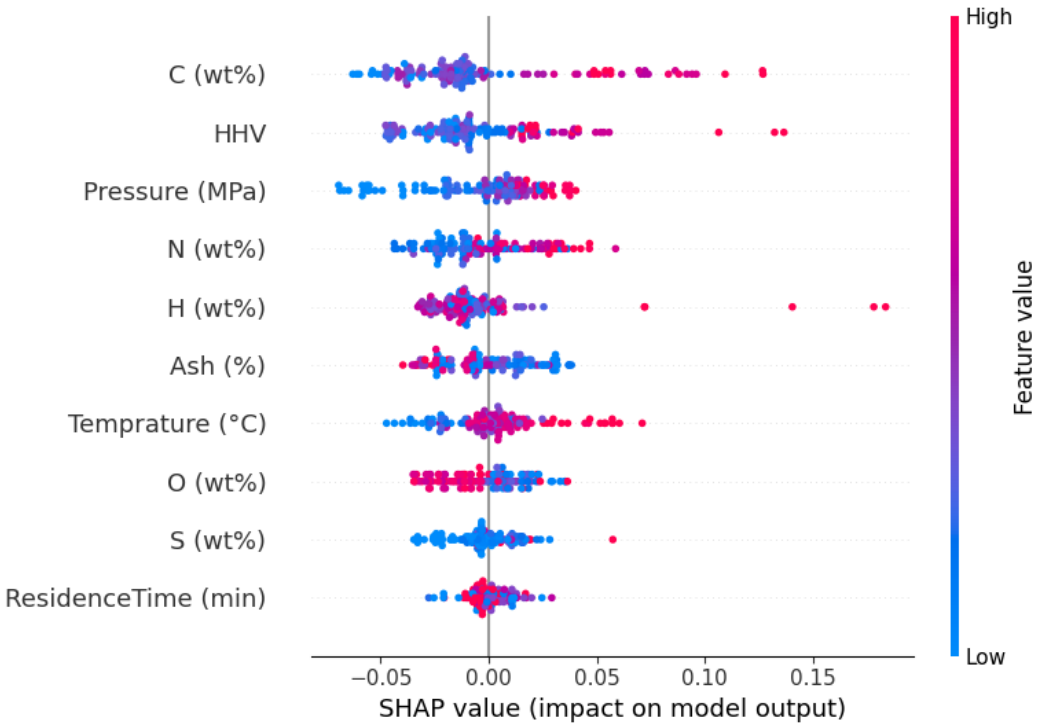


Figure 5.21 This SHAP graph for the Random Forest (RF) model shows each feature's impact on predictions. Higher SHAP values indicate greater influence, with red representing high feature values and blue representing low values.

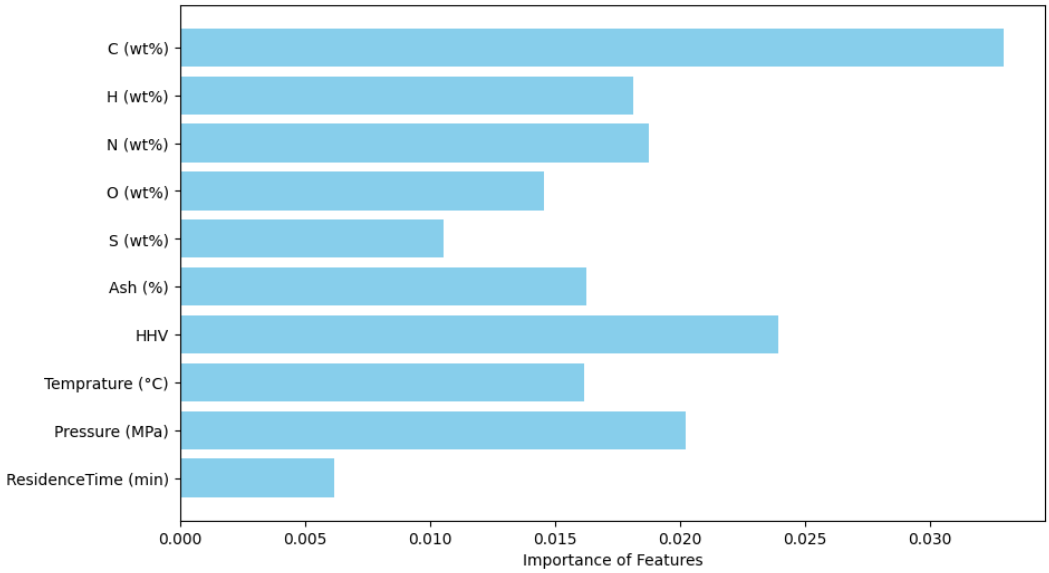


Figure 5.22 This SAGE graph for the Random Forest (RF) model illustrates the global importance of each feature. Higher values indicate greater overall influence on the model's predictions, providing insight into the contribution of each feature to the model's performance.

SHAP (SHapley Additive exPlanations) and SAGE (Shapley Additive Global importancE) are techniques rooted in game theory that aid in interpreting machine learning models. SHAP provides insights by merging optimal credit distribution with localized explanations using foundational Shapley values and their extensions. Similarly, SAGE measures each feature's importance by evaluating its contribution to predictive power and accounts for complex feature interactions with Shapley values. Both methods offer a comprehensive understanding of model behavior through a game-theoretic lens [45],[46].

The SHAP value graph for the XGBoost model (Figure 5.13) indicates that HHV is the fifth most important feature. In Figure 5.15, it is the second most important feature. Figure 5.21 and Figure 5.19 display the same results. The SHAP graphs reveal that HHV values are as significant for the models as other features. However, Figure 5.17 is an exception, showing lower importance for HHV.

Figure 5.14 also serves as the SAGE graph for the XGBoost model, confirming that HHV values are as important as other features. Figure 5.22, Figure 5.20, and Figure 5.16 follow the same pattern. In contrast, Figure 5.18 shows relatively lower importance for HHV. According to the Table 5.3 Models and the model's results., the ANN model's performance is slightly lower than that of the other models. This demonstrates that HHV values positively impact the models.

6
CONCLUSION

In conclusion, our research on bio-crude oil prediction using machine learning shows some important findings. We find that different modelling approaches generally perform similarly. However, the random forest model stands out as the best-performing model by showing the highest Test R2 value among all models. This shows that its strong ability to understand social data is the best choice for predictive work in this context.

On the other hand, the K-Nearest Neighbor (KNN) model is also effective in creating good models by capturing local information reflected in the highest R2 values. These results highlight the importance of using different models to gain a better understanding and make informed decisions.

By knowing the advantages and disadvantages of each model, we can improve the accuracy of predictions and provide a basis for future modelling and research. Our results contribute to the advancement of technology in the estimation of bio-crude oil, providing practical use and insight for further research in this field.

REFERENCES

[1] J. P. Lange, “Lignocellulose Liquefaction to Biocrude: A Tutorial Review,” Mar. 22, 2018, *Wiley-VCH Verlag*. doi: 10.1002/cssc.201702362.

[2] A. Dimitriadis and S. Bezergianni, “Hydrothermal liquefaction of various biomass and waste feedstocks for biocrude production: A state of the art review,” Feb. 01, 2017, *Elsevier Ltd*. doi: 10.1016/j.rser.2016.09.120.

[3] J. A. Ramirez, R. J. Brown, and T. J. Rainey, “A review of hydrothermal liquefaction bio-crude properties and prospects for upgrading to transportation fuels,” 2015, *MDPI AG*. doi: 10.3390/en8076765.

[4] J. Li *et al.*, “Machine learning aided bio-oil production with high energy recovery and low nitrogen content from hydrothermal liquefaction of biomass with experiment verification,” *Chemical Engineering Journal*, vol. 425, Dec. 2021, doi: 10.1016/j.cej.2021.130649.

[5] T. Katongtung, T. Onsree, and N. Tippayawong, “Machine learning prediction of biocrude yields and higher heating values from hydrothermal liquefaction of wet biomass and wastes,” *Bioresour Technol*, vol. 344, Jan. 2022, doi: 10.1016/j.biortech.2021.126278.

[6] T. Katongtung, T. Onsree, K. Y. Tippayawong, and N. Tippayawong, “Prediction of biocrude oil yields from hydrothermal liquefaction using a gradient tree boosting machine approach with principal component analysis,” *Energy Reports*, vol. 9, pp. 215–222, Oct. 2023, doi: 10.1016/j.egyr.2023.08.079.

[7] F. Cheng *et al.*, “Hydrothermal liquefaction of high- and low-lipid algae: Bio-crude oil chemistry,” *Appl Energy*, vol. 206, pp. 278–292, Nov. 2017, doi: 10.1016/j.apenergy.2017.08.105.

[8] H. Kazemi Shariat Panahi, M. Tabatabaei, M. Aghbashlo, M. Dehhaghi, M. Rehan, and A. S. Nizami, “Recent updates on the production and upgrading of bio-crude oil from microalgae,” *Bioresour Technol Rep*, vol. 7, Sep. 2019, doi: 10.1016/j.biteb.2019.100216.

[9] R. Ren *et al.*, “High yield bio-oil production by hydrothermal liquefaction of a hydrocarbon-rich microalgae and biocrude upgrading,” *Carbon Resources Conversion*, vol. 1, no. 2, pp. 153–159, Aug. 2018, doi: 10.1016/j.crcon.2018.07.008.

[10] T. Katongtung, T. Onsree, K. Y. Tippayawong, and N. Tippayawong, “Prediction of biocrude oil yields from hydrothermal liquefaction using a gradient tree boosting machine approach with principal component analysis,” *Energy Reports*, vol. 9, pp. 215–222, Oct. 2023, doi: 10.1016/j.egyr.2023.08.079.

[11] T. Katongtung, T. Onsree, and N. Tippayawong, “Machine learning prediction of biocrude yields and higher heating values from hydrothermal liquefaction of wet biomass and wastes,” *Bioresour Technol*, vol. 344, Jan. 2022, doi: 10.1016/j.biortech.2021.126278.

[12] J. Li *et al.*, “Machine learning aided bio-oil production with high energy recovery and low nitrogen content from hydrothermal liquefaction of biomass with experiment verification,” *Chemical Engineering Journal*, vol. 425, Dec. 2021, doi: 10.1016/j.cej.2021.130649.

[13] A. Shafizadeh *et al.*, “Machine learning predicts and optimizes hydrothermal liquefaction of biomass,” *Chemical Engineering Journal*, vol. 445, Oct. 2022, doi: 10.1016/j.cej.2022.136579.

[14] H. Yaka, M. A. Insel, O. Yucel, and H. Sadikoglu, “A comparison of machine learning algorithms for estimation of higher heating values of biomass and fossil fuels from ultimate analysis,” *Fuel*, vol. 320, Jul. 2022, doi: 10.1016/j.fuel.2022.123971.

[15] R. Kissell and J. Poserina, “Chapter 2 - Regression Models,” in *Optimal Sports Math, Statistics, and Fantasy*, R. Kissell and J. Poserina, Eds., Academic Press, 2017, pp. 39–67. doi: https://doi.org/10.1016/B978-0-12-805163-4.00002-5.

[16] C. Hagquist and M. Stenbeck, “Goodness of Fit in Regression Analysis – R2 and G2 Reconsidered,” *Qual Quant*, vol. 32, no. 3, pp. 229–245, 1998, doi: 10.1023/A:1004328601205.

[17] L. Igual and S. Seguí, “Undergraduate Topics in Computer Science Introduction to Data Science.” [Online]. Available: http://www.springer.com/series/7592

[18] J. Alzubi, A. Nayyar, and A. Kumar, “Machine Learning from Theory to Algorithms: An Overview,” in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Nov. 2018. doi: 10.1088/1742-6596/1142/1/012012.

[19] V. B. S. Prasath, H. A. A. Alfeilat, O. Lasassmeh, and A. B. A. Hassanat, “Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier - A Review,” *CoRR*, vol. abs/1708.04321, 2017, [Online]. Available: http://arxiv.org/abs/1708.04321

[20] K. Taunk, S. De, S. Verma, and A. Swetapadma, “A Brief Review of Nearest Neighbor Algorithm for Learning and Classification,” in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, 2019, pp. 1255–1260. doi: 10.1109/ICCS45141.2019.9065747.

[21] Y. Song, J. Liang, J. Lu, and X. Zhao, “An efficient instance selection algorithm for k nearest neighbor regression,” *Neurocomputing*, vol. 251, pp. 26–34, 2017, doi: https://doi.org/10.1016/j.neucom.2017.04.018.

[22] P. Zhang, Y. Jia, and Y. Shang, “Research and application of XGBoost in imbalanced data,” *Int J Distrib Sens Netw*, vol. 18, no. 6, p. 15501329221106936, 2022, doi: 10.1177/15501329221106935.

[23] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD ’16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[24] C. Sheng and H. Yu, “An optimized prediction algorithm based on XGBoost,” in *2022 International Conference on Networking and Network Applications (NaNA)*, 2022, pp. 1–6. doi: 10.1109/NaNA56854.2022.00082.

[25] “Gaussian Processes,” https://scikit-learn.org/stable/modules/gaussian_process.html.

[26] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, “Gaussian Process Regression for Materials and Molecules,” *Chem Rev*, vol. 121, no. 16, pp. 10073–10141, Aug. 2021, doi: 10.1021/acs.chemrev.1c00022.

[27] S. Agatonovic-Kustrin and R. Beresford, “Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research,” *J Pharm Biomed Anal*, vol. 22, no. 5, pp. 717–727, 2000, doi: https://doi.org/10.1016/S0731-7085(99)00272-1.

[28] J. Zupan, “Introduction to artificial neural network (ANN) methods: what they are and how to use them,” *Acta Chim Slov*, vol. 41, no. 3, p. 327, 1994.

[29] D. F. Specht, “A general regression neural network,” *IEEE Trans Neural Netw*, vol. 2, no. 6, pp. 568–576, 1991.

[30] T. Burzykowski, M. Geubbelmans, A.-J. Rousseau, and D. Valkenborg, “Validation of machine learning algorithms,” *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 164, no. 2, pp. 295–297, Aug. 2023, doi: 10.1016/j.ajodo.2023.05.007.

[31] “3.1. Cross-validation: evaluating estimator performance — scikit-learn 1.4.1 documentation.” Accessed: Mar. 21, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/cross_validation.html

[32] C. Tantithamthavorn, S. McIntosh, A. E. Hassan, and K. Matsumoto, “An Empirical Comparison of Model Validation Techniques for Defect Prediction Models,” *IEEE Transactions on Software Engineering*, vol. 43, no. 1, pp. 1–18, 2017, doi: 10.1109/TSE.2016.2584050.

[33] M. W. Browne, “Cross-Validation Methods,” *J Math Psychol*, vol. 44, no. 1, pp. 108–132, 2000, doi: https://doi.org/10.1006/jmps.1999.1279.

[34] scikit-optimize, “skopt.BayesSearchCV — scikit-optimize 0.8.1 documentation,” https://scikit-optimize.github.io/stable/modules/generated/skopt.BayesSearchCV.html.

[35] “3.3. Metrics and scoring: quantifying the quality of predictions,” https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics. Accessed: Apr. 15, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics

[36] V. Plevris, G. Solorzano, N. Bakas, and M. Ben Seghier, “Investigation of performance metrics in regression analysis and machine learning-based prediction models,” Apr. 2022. doi: 10.23967/eccomas.2022.155.

[37] T. Doan and J. Kalita, “Selecting Machine Learning Algorithms Using Regression Models,” in *2015 IEEE International Conference on Data*

Mining Workshop (ICDMW), 2015, pp. 1498–1505. doi: 10.1109/ICDMW.2015.43.

[38] "Coefficient of Determination, R-squared," https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html.

[39] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," *Geosci Model Dev*, vol. 15, no. 14, pp. 5481–5487, 2022, doi: 10.5194/gmd-15-5481-2022.

[40] I. Nitze, U. Schulthess, and H. Asche, "Comparison of machine learning algorithms random forest, artificial neural network and support vector machine to maximum likelihood for supervised crop type classification," *Proceedings of the 4th GEOBIA, Rio de Janeiro, Brazil*, vol. 79, p. 3540, 2012.

[41] J. R. Laszakovits and A. A. MacKay, "Data-Based Chemical Class Regions for Van Krevelen Diagrams," *J Am Soc Mass Spectrom*, vol. 33, no. 1, pp. 198–202, Jan. 2022, doi: 10.1021/jasms.1c00230.

[42] A. Dimitriadis and S. Bezergianni, "Hydrothermal liquefaction of various biomass and waste feedstocks for biocrude production: A state of the art review," Feb. 01, 2017, *Elsevier Ltd*. doi: 10.1016/j.rser.2016.09.120.

[43] A. Aierzhati *et al.*, "Experimental and model enhancement of food waste hydrothermal liquefaction with combined effects of biochemical composition and reaction conditions," *Bioresour Technol*, vol. 284, pp. 139–147, 2019, doi: https://doi.org/10.1016/j.biortech.2019.03.076.

[44] L. Sheng, X. Wang, and X. Yang, "Prediction model of biocrude yield and nitrogen heterocyclic compounds analysis by hydrothermal liquefaction of microalgae with model compounds," *Bioresour Technol*, vol. 247, pp. 14–20, 2018, doi: https://doi.org/10.1016/j.biortech.2017.08.011.

[45] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[46] I. Covert, S. Lundberg, and S.-I. Lee, "Understanding Global Feature Contributions With Additive Importance Measures," Apr. 2020, [Online]. Available: http://arxiv.org/abs/2004.00668

PUBLICATIONS FROM THE THESIS

Conference Papers

1. Ş. Güngör, M. A. İnsel, and H. Sadıkoğlu, "Utilization of Machine Learning Algorithms for Enhanced Bio-Crude Oil Yield Estimation," in Proc. 7th International Conference on Engineering Sciences (ICES 2024), Ankara, Turkey, 2024, pp. 22.