REPUBLIC OF TURKEY

YILDIZ TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

DEVELOPMENT OF A CORRELATION FOR ESTIMATION OF BIOMASS HIGHER HEATING VALUE WITH STATISTICAL AND MACHINE LEARNING BASED APPROACHES

Havva YAKA

MASTER OF SCIENCE THESIS

Department of Chemical Engineering

Chemical Engineering Program

Supervisor

Prof. Dr. Hasan SADIKOGLU

Co-Supervisor

Assoc. Prof. Dr. Ozgün YUCEL

July, 2022

REPUBLIC OF TURKEY

YILDIZ TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

BİYOKÜTLE YÜKSEK ISIL DEĞERİ TAHMİNİ İÇİN İSTATİSTİK VE MAKİNE ÖĞRENİMİNE DAYALI YAKLAŞIMLARLA KORELASYON GELİŞTİRİLMESİ

A thesis submitted by Havva YAKA in partial fulfillment of the requirements for the degree of Master is approved by the committee on 01.07.2022 in Department of Chemical Engineering, Chemical Engineering Program.

Prof. Dr. Hasan SADIKOGLU Yıldız Technical University Supervisor	Assoc. Prof. Dr. Ozgun YUCEL Yıldız Technical University Co- supervisor

Approved By the Examining Committee

Prof. Dr. Hasan SADIKOGLU, Supervisor

Yildiz Technical University _____

Prof. Dr. Murat OZDEMIR, Member

Gebze Technical University _____

Assoc. Prof. Dr. Nil Acaralı , 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 Development of a Correlation for Estimation of Biomass Higher Heating Value with Statistical and Machine Learning Based Approaches supervised by my supervisor, Prof. Dr. Hasan Sadikoglu and my co-supervisor Assoc. Prof. Dr. Ozgun Yucel. In the case of a discovery of false statement, I am to acknowledge any legal consequence.

Havva YAKA

Signature

Dedicated to my family

and my best friend

ACKNOWLEDGEMENTS

To my advisors Prof. Dr. Hasan SADIKOGLU and Assist. Prof. Dr. Ozgun YUCEL, who shared their valuable knowledge and experience with me within the scope of my thesis work, for their support and guidance,

To my dear friends Meryem MUSLU and Ecem KOSURGELI, who stood by me throughout the thesis, encouraged me whenever I had difficulties, and supported me,

To my most precious people, my family, whose moral support I have always felt with me and whose rights I can never repay; to my father, my mother, my sister and my brother,

My most heartfelt thanks.

Havva YAKA

TABLE OF CONTENTS

LIST OF ABBREVIATIONS vi

LIST OF FIGURES vii

LIST OF TABLES viii

ABSTRACT ix

ÖZET xi

1 INTRODUCTION 1

1.1 Literature Review 1

1.2 Objective of the Thesis 7

1.3 Hypothesis 8

2 METHODOLOGY 9

2.1 Collection of Data 9

2.2 Lineer Regression 11

2.3 Polynomial Regression 11

2.4 Support Vector Machine 12

2.5 Decision Tree Regression 13

2.6 Gaussian Process Regression 14

2.7 Random Forest Regression 14

2.8 Validation Methods: Hold-out Validation & k-fold Cross Validation 15

3 CONCLUSION 17

3.1 HHV Estimation by Machine Learning Methods 17

3.2 Classification of Fuel 22

3.2.1 Investigation of Metadata of Ultimate Analysis Dataset 22

3.2.2 Van Krevelen Diagram 23

4 RESULTS AND DISCUSSIONS 25

REFERENCES 26

PUBLICATIONS FROM THE THESIS 31

LIST OF ABBREVIATIONS

AAE	Average absolute error
ABE	Average bias error
ANN	Artificial neural network
C	Carbon
CC	Correlation coefficient
DAF	Dry, Ash Free
DTR	Decision Tree Regression
FC	Fixed carbon
GP	Genetic Programming
GPR	Gaussian Process Regression
H	Hydrogen
HHV	Higher heating value
LHV	Lower heating value
MAE	Mean absolute error
N	Nitrogen
N-RMSE	Normalized root mean square error
O	Oxygen
RFR	Random Forest Regression
RMSE	Root mean square error
S	Sulfur
SPR	Support Vector Regression

LIST OF FIGURES

Figure 2. 1 Histogram of the collected data 12
Figure 2. 2 Kernel function 13
Figure 2. 3 The principles of hold-out validation (a) and k-fold cross validation methods(b) 16
Figure 3. 1 The correlation heatmap of the input data.....17
Figure 3. 2 HHV prediction with proposed methods 19
Figure 3. 3 HHV prediction with proposed linear and polynomial models 20
Figure 3. 4 Atomic O/C ratio x 10 vs Atomic H/C ratio graph for fuel samples..24

LIST OF TABLES

Tablo 1. 1 Summary of critical studies in the subject of HHV prediction.....4

Table 2. 1 Statistical information of ultimate analysis data of fuels obtained from Phyllis 2 database 9

Table 2.2 Class, content and number of fuel samples..... 10

Table 3. 1 Developed HHV correlations for multilinear and polynomial regressions.....18

Table 3. 2 Performance evaluation of the proposed machine learning algorithms 18

Table 3. 3 Suggested fuel classes for application of each method. 21

ABSTRACT

DEVELOPMENT OF A CORRELATION FOR ESTIMATION OF BIOMASS HIGHER HEATING VALUE WITH STATISTICAL AND MACHINE LEARNING BASED APPROACHES

Havva YAKA

Department of Chemical Engineering

Master of Science Thesis

Supervisor: Prof. Dr. Hasan SADIKOGLU

Co- supervisor: Assoc. Prof. Dr. Ozgun YUCEL

Higher heating value (HHV) is one of the most important parameters to consider while obtaining energy efficiently from fuels. It provides means to estimate the quality of the fuel. However, measuring the HHV of fuels requires advanced devices, is expensive and time consuming. Thus, sufficient estimation of the HHV is crucial for development of fuel technologies and numerous studies have been performed about this subject. In this study, several machine learning algorithms (DTR, SVR, GPR, RFR, Multi-Linear Regressions, and Polynomial Regression) were utilized to construct models for estimation of the HHV from the largest open dataset of ultimate analysis of fuels. Many different solid, liquid, and gaseous fuels such as biomass, biochar, msw, kerosene, gasoline, fuel oil, algae, natural gas were analyzed. 10-fold cross validation method was used to assess the validity of the

constructed models in an unbiased manner. The R^2, adjusted R^2, RMSE, N-RMSE, and AAE values of each model were computed for performance evaluations. In addition, the metadata of the ultimate analysis is investigated for each class of fuel and an extensive Krevelen diagram was produced from the largest dataset to date, which illustrated the relationships between the atomic O/C ratio and the atomic H/C ratio of the fuels. Finally, the results were compared with the literature, and advantages and disadvantages of each method were discussed in terms of both computational complexity and prediction accuracy.

Keywords: Higher heating value, machine learning, biomass, regression

ÖZET

BİYOKÜTLE YÜKSEK ISIL DEĞERİ TAHMİNİ İÇİN İSTATİSTİK VE MAKİNE ÖĞRENİMİNE DAYALI YAKLAŞIMLARLA KORELASYON GELİŞTİRİLMESİ

Havva YAKA

Kimya Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Danışman: Prof. Dr. Hasan SADIKOGLU

Eş-Danışman: Doç. Dr. Ozgun YUCEL

Yüksek ısıl değer (HHV), yakıtlardan verimli enerji elde edilirken dikkate alınması gereken en önemli parametrelerden biridir. Yakıtın kalitesini tahmin etmek için araçlar sağlar. Ancak yakıtların HHV'sini ölçmek gelişmiş cihazlar gerektirir, pahalıdır ve zaman alıcıdır. Bu nedenle, yakıt teknolojilerinin geliştirilmesi için HHV'nin yeterli tahmini çok önemlidir ve bu konuda çok sayıda çalışma yapılmıştır. Bu çalışmada, yakıtların nihai analizinin en büyük açık veri setinden HHV tahmini için modeller oluşturmak için birkaç makine öğrenme algoritması (DTR, SVR, GPR, RFR, Çok Doğrusal Regresyon ve Polinom Regresyon) kullanıldı. Biyokütle, biochar, msw, kerosen, benzin, fuel oil, alg, doğal gaz gibi birçok farklı katı, sıvı ve gaz yakıt analiz edilmiştir. Oluşturulan modellerin geçerliliğini tarafsız bir şekilde değerlendirmek için 10 kat çapraz doğrulama yöntemi kullanılmıştır. Performans değerlendirmeleri için her modelin R^2, düzeltilmiş R^2, RMSE, N-RMSE

ve AAE değerleri hesaplanmıştır. Ek olarak, nihai analizin meta verileri her bir yakıt sınıfı için araştırılmış ve bugüne kadarki en büyük veri setinden kapsamlı bir Krevelen diyagramı üretilmiştir; yakıtlar. Son olarak, sonuçlar literatür ile karşılaştırılmış ve her bir yöntemin hem hesaplama karmaşıklığı hem de tahmin doğruluğu açısından avantaj ve dezavantajları tartışılmıştır.

Anahtar Kelimeler: Yüksek ısıl değeri, makine öğrenmesi, biyokütle, regresyon

1
INTRODUCTION

1.1 Literature Review

The increase in global energy demand and the rapid depletion of fossil resources have caused an increase in the use of renewable energy resources [1]. Due to the ever-growing population and the industrialization process, biomass is one of the most important resources that will ensure that the energy need is not harmful to the environment and that it is sustainable [2].

The damage to the environment is significantly lower because biomass does not contain carcinogenic substances and has low sulfur compared to fossil fuels. There are several advantages of biomass compared to other energy sources; it is abundant, inexpensive and easily storable [2,3]. It does not create greenhouse effect like fossil fuels and does not cause acid rain. With such properties, the utilization of biomass has been increasing gradually around the world.

Biomass is transformed into gas or liquid fuels by using many different conversion methods as a solution to the energy needs with energy demand [4]. High heating value (HHV) is a major property in obtaining biofuels using biomass and thereby the HHV of a certain number of burning materials determines the quantitative energy content of fuels [5]. HHV is known as one of the most important properties of fuels since it yields the high energy content, which determines the efficient use of the energy resources.

HHV is obtained from the numerical simulations or design calculations of thermal conversion systems for fuels. Since the HHV of fuels results the quantitative energy content of fuels, it is of major importance in comprehensive studies [6]. Generally, bomb calorimeter is utilized to specify HHV of fuels, but this method is relatively expensive and time-consuming. The use of ultimate and proximate fuel analysis to estimate the HHV of the fuel in previous studies shows that this method is widely used as opposed to using bomb calorimeter [7].

While ultimate analysis requires cost-intensive and advanced equipment compared to proximate analysis, ultimate analysis gives more comprehensive results than proximate analysis [8].

During the past decades, several researchers have attempted at various correlations for predicting HHV from proximate and ultimate analysis of fuel [9]. Xing et al. [10] published an exhaustive study on estimation of HHV of fuels from ultimate and proximate analysis using machine learning methods.

In the literature, at most 737 data were collected to conduct estimations from proximate or ultimate analysis (see Table 1). The literature reviews also showed that machine learning approaches result better predictions, with R^2 values bigger than 0.94. With new approaches, studies on the use of many methods for HHV prediction has become more common. For example, Nhuchhen and Salam [11] have developed a model for estimation of higher heating value from proximate analysis of fuels with a new approach.

The results show that the proposed correlation yields better results compared to other correlations. In this study, 250 data were used. When the results were examined, it was seen that the new developed correlation gave better accuracy. In this study, the average absolute error was found as 9.43% and 5.88%.

According to another study, Hosseinpour et. al. [12], authors estimated the HHV of fuels by using proximate analysis using the improved PCA-INFLPS model. This developed model has been compared with the other models such as ANFIS, NFPLS, PLS, and INFPL. As a result, the developed model (PCA-INFPLS) predicted HHV more accurately compared to other models. In this work, despite the 350 data collected, the developed model (PCA-INFPLS) yielded sufficient results for the estimation of the high calorific value. For this study, R^2 value was bigger than 0.96.

In a previous work by Parikh et. al. [13], the authors predicted a correlation for higher heating value from proximate analysis of solid fuels. In this study, 450 data points have been used for assuming correlation. 100 data were shared in the chart. In this study, an average absolute error was found as 3.74%.

Two years later, the authors derived a new correlation [14], calculated the fundamental composition from proximate analysis of the fuels. In this study, 250 data points were collected. The results show that average absolute error 3.21%, and percentage of carbon, hydrogen and oxygen are 3,21%C, 4.79%H and 3.4%, O.

According to another comprehensive study, authors carried out genetic programming-based correlations for the prediction of the HHV of coals from diverse geographies and of diverse ranks [15]. The resulted correlation coefficient of the GP I is 0,99 also correlation coefficient of the GP II is the same as GP I, and the correlation coefficient of the GP III is 0,91. Considering the results of the GP models, it is observed that the correlation coefficients of these models are more than 0.90. These high results are important for the usability of these algorithms.

When the literature is examined, it is observed that there are many studies which determined the HHV value by deriving new correlations or using more comprehensive models. It is also seen that utilization of different methods may result in differences in the accuracy and reliability of the results of published correlations.

In the developed correlation studies, many definitions are used for evaluating the reliability and accuracy of the results and the performance estimation of the methods proposed. These definitions give reliable results when determining the usability of the proposed method. The correlation coefficient (R^2) describes the direction and strength of the relationship between variables. Absolute average error (AAE), on the other hand, provides information about accuracy in performance estimation similar to other definitions [16]. Adjusted R^2 shows how well terms fit a line but adjusts the number of terms in a pattern [17]. Root mean square error (RMSE) is commonly used to evaluate the accuracy of a proposed system [18]. Normalized root means square error (NRMSE) is a standard statistical measure and this definition is used to measure model performance [19].

$$\boldsymbol{R^2 = 1 - \left[\frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y}_i)^2}\right]} \quad (1.1)$$

$$\boldsymbol{Adj.R^2 = 1 - \left[\frac{(1-R^2)*(m-1)}{m-n-1}\right]} \quad (1.2)$$

$$\boldsymbol{RMSE = \sqrt{\frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{m}}} \quad (1.3)$$

$$NRMSE = \frac{\sqrt{\frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\hat{y}_i}}}{m} \quad (1.4)$$

$$AAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - x| \quad (1.5)$$

As a summary of literature, Table 1.1 presents the methods used in HHV prediction studies, number of data, validation methods and the values of R^2 / AAE/ MAE metrics.

Tablo 1. 1 Summary of critical studies in the subject of HHV prediction

Data Size	Method	Validation method (Test/Train)	R^2 / AAE/ MAE	Ref. no
275	Multiple linear regression	Hold-out validation 50/225	AAE= 1.45%	[9]
737	Multiple non-linear regression and Machine learning approaches: (ANN, SVM and RF)	10-fold cross validation	machine learning approaches: $R^2 >$ 0.90 linear and nonlinear correlations: $R^2 <$ 0.70	[10]
250	Linear and nonlinear regression	Hold-out validation 10/240	AAE = 9.43% and 5.88%	[11]
350	Iterative network-based fuzzy partial least squares coupled with principle component analysis (PCA-INFPLS)	Cross-validation 15%/15%/70%	$R^2 > 0.96$	[12]

Table 1.1 Summary of critical studies in the subject of HHV prediction (continued)

550	**Multiple linear and non-linear regression**	**Hold-out validation 100/450**	**AAE = 3.74%**	**[13]**
250	Multiple linear and non-linear regression	Hold-out validation 50/200	AAE = 3.21% C, 4.79% H 3.4% O	[14]
250	Multiple non-linear regression	Hold-out validation 86/164	$R^2 = 0.98$	[7]
340	Multiple non-linear regression	Hold-out validation 40/300	$R^2 = 0.93$, 0.83, and 0.67 of C, H, and O in high-rank bituminous, $R^2 =$ 0.86, 0.61, and 0.71 of C, H, and O in subbituminous, $R^2 =$ 0.92, 0.67, and 0.66 of C, H, and O in lignite.	[20]
16	Multiple linear regression	-	$R^2 = 0.9997$	[21]
25	GP models	Hold-out training 75%/15%	$R^2 =$ 0.99 GP I $R^2 = 0.99$ GP II $R^2 = 0.91$ GP III	[15]

Table 1.1 Summary of critical studies in the subject of HHV prediction (continued)

382	Machine learning approaches: GA-radial basis function (GA-RBF) and multivariate polynomial regression (MPR)	Hold-out 30 % /70 %	GA-RBF R^2 = 0.9591 MPR R^2 = 0.9597	[22]
29	Genetic Programming (GP I,II)	Hold-out Validation 15/14	R^2 = 0.995 GP I R^2 = 0.998 GP II	[23]
171	Multiple Regression and Genetic Programming	Hold-out validation 75% / 25%	R^2 = 0.2037 - 0.5429 R^2 = 0.0887 - 0.7547 R^2 = 0.7611- 0.8804	[24]
44	Linear Regression	No validation 0%/100%	R^2 = 0.99	[25]
1566	**Multiple Non-linear regression and Machine Learning Approaches**: (DTR, SVR, GPR, RFR, Multi-Linear Regressions, and Polynomial Regression)	**10-fold cross validation**	R^2 = 0.88 - 0.93 Adj. R^2 = 0.88 - 0.93 RMSE = 1.3 – 1.8 N-RMSE = 0.066 – 0.094 AAE = 0.040 – 0.058	**This Study**

In this study, we utilize machine learning algorithms such as Decision Tree Regression (DTR), Support Vector Regression (SVR), Gaussian Process Regression (GPR), Random Forest Regression (RFR), Linear Regression (CH), Multi-Linear Regressions (CHO and CHONS), and Polynomial Regression (C1.4H) to estimate HHV value of fuels from ultimate analysis data.

The novel application of these algorithms when estimating the HHV of the fuels is crucial for the development of fuel technologies. We used a large dataset (1566 data points) while constructing our models, which is considerably larger than other studies found in the literature. We have conducted 10-fold cross validation to assess the validity of the constructed models. The R^2, adjusted R^2, RMSE, N-RMSE, and AAE values obtained for each model to evaluate the performances. The data validation and performance evaluation in this study is more extensive and accurate when compared to the other studies in the literature, since the dataset is considerably larger. In addition, by investigating the metadata of the large dataset, we provide crucial insights about the classes of fuels for researchers.

We also presented the extended version of Krevelen diagram of the classified fuels to illustrate the relations between atomic O/C ratio and atomic H/C ratio by using the dataset to characterize the maturity and source of organic matters. Finally, we compare our results both with the literature and with themselves.

The advantages and disadvantages of each proposed model is discussed detailly in the following chapters. As a result, we observe that while the RFR method yields the best fit for the data, the GPR is more accurate when estimating the HHV in general. However, the computation of HHV with the proposed CHONS method is more straightforward and may be preferable to the other methods.

1.2 Objective of the Thesis

By using machine learning methods, HHV estimation is made with the help of ultimate analysis results and the results of the methods used are compared and advantages and disadvantages are determined.

1.3 Hypothesis

By using machine learning methods, HHV estimation is more applicable with the help of ultimate analysis results.

2
METHODOLOGY

2.1 Collection of Data

Ultimate analysis of 1566 fuel samples are used to construct the models. The data is obtained from Phyllis 2 database [26]. The statistical information regarding the data is presented in Table 2.1. While the Phyllis 2 database contains more samples, utilized data in this study is reduced due to the overlaps present in the database. The data consists of ultimate analysis of the fuels, i.e. C (%), H (%), O (%), N (%), S (%), LHV (MJ/kg) and HHV (MJ/kg). Hence, in this study, we have six features and one output (HHV).

Table 2. 1 Statistical information of ultimate analysis data of fuels obtained from Phyllis 2 database [26]

Ultimate analysis	Unit (daf)	Min.	Max.	Median	Mean	Std dev		Samples
Carbon	wt%	0.00	98.56	50.45	53.42	10.74	20%	1820
Hydrogen	wt%	0.00	90.66	6.12	6.33	2.53	40%	1813
Oxygen	wt%	0.00	77.98	42.27	38.39	11.11	29%	1668
Nitrogen	wt%	0.00	21.56	0.65	1.45	2.17	150%	1841
LHV	MJ/kg	10.30	50.06	18.90	20.47	4.92	24%	1610
HHV	MJ/kg	10.35	55.55	20.23	21.88	5.24	24%	1689

While majority of the data consist of biomass, other kind of fuels are also included in our analysis (see Table 2.2). In addition, data of the major materials such as wood waste, sewage sludge, olive residue, switchgrass, rice straw, etc. are also collected for the present study [14]. While Phyllis 2 considers 16 classes of fuel [26], we have merged some of the classes as presented in Table 2.2, resulting 8 classes of fuel.

Table 2.2 Class, content and number of fuel samples

Class	Content	Data Size
Torrefied materials	Mixed wood, straw, organic residue, hard wood and soft wood	28
Untreated woods	Bark, beech, birch, pine, leaves, hard wood and soft wood	413
Treated woods	Composite, demolition, impregnated woods, particle boards	82
Wastes	Organic, non-organic residues, RDF, MSW	207
Manure & Sludge	Manure, sludge, pulp, sewage	77
Agricultural wastes	Grass, plant, husk, shell, pit, straw	454
Char & Fossil fuels	Biochar, char, fossil fuels, oils (kerosane, gasoline, fuel oil, diesel), natural gas	264
Others	Non-fossil fuels, marine biomass (micro and macro algea)	41

2.2 Lineer Regression

Linear regression is a statistical method and is used to calculate the value of a dependent variable from the independent variable. It is the most widely used method among all statistical techniques [7]. When machine learning methods are examined, it is clearly observed that there are many studies using linear regression method for HHV prediction. This method is a variable method and displays the output as a combination of properties with a simple value. Linear regression is generally expressed as in equation bellow [5].

$$\hat{y} = \beta_0 + x_1 * \beta_1 + x_2 * \beta_2 + \ldots x_n * \beta_n \quad (2.1)$$

where, x1, x2, … , xn are the features of the data set starting from 1 up to n, yˆ is the predicted output and β0, β1, … , βn are the coefficient determined in the training phase

2.3 Polynomial Regression

When the data is approximated by a polynomial function, the regression is termed as polynomial regression [27]. It contains high order terms of variables used in estimating polynomial regression. In this way, it makes the response surface curved [28]. Since there is no uniform polynomial equation, the equation should be decided according to the problem encountered [5]. General expression of a polynomial is a function that takes the form:

When the data is approximated by a polynomial function, the regression is termed as polynomial regression [27]. It contains high order terms of variables used in estimating polynomial regression. In this way, it makes the response surface curved [28]. Since there is no uniform polynomial equation, the equation should be decided according to the problem encountered [5]. General expression of a polynomial is a function that takes the form:

$$f(x) = c_0 + c_1 x + c_2 x^2 \ldots c_n x^n \quad (2.2)$$

where n is the degree of the polynomial and c is a set of coefficients.

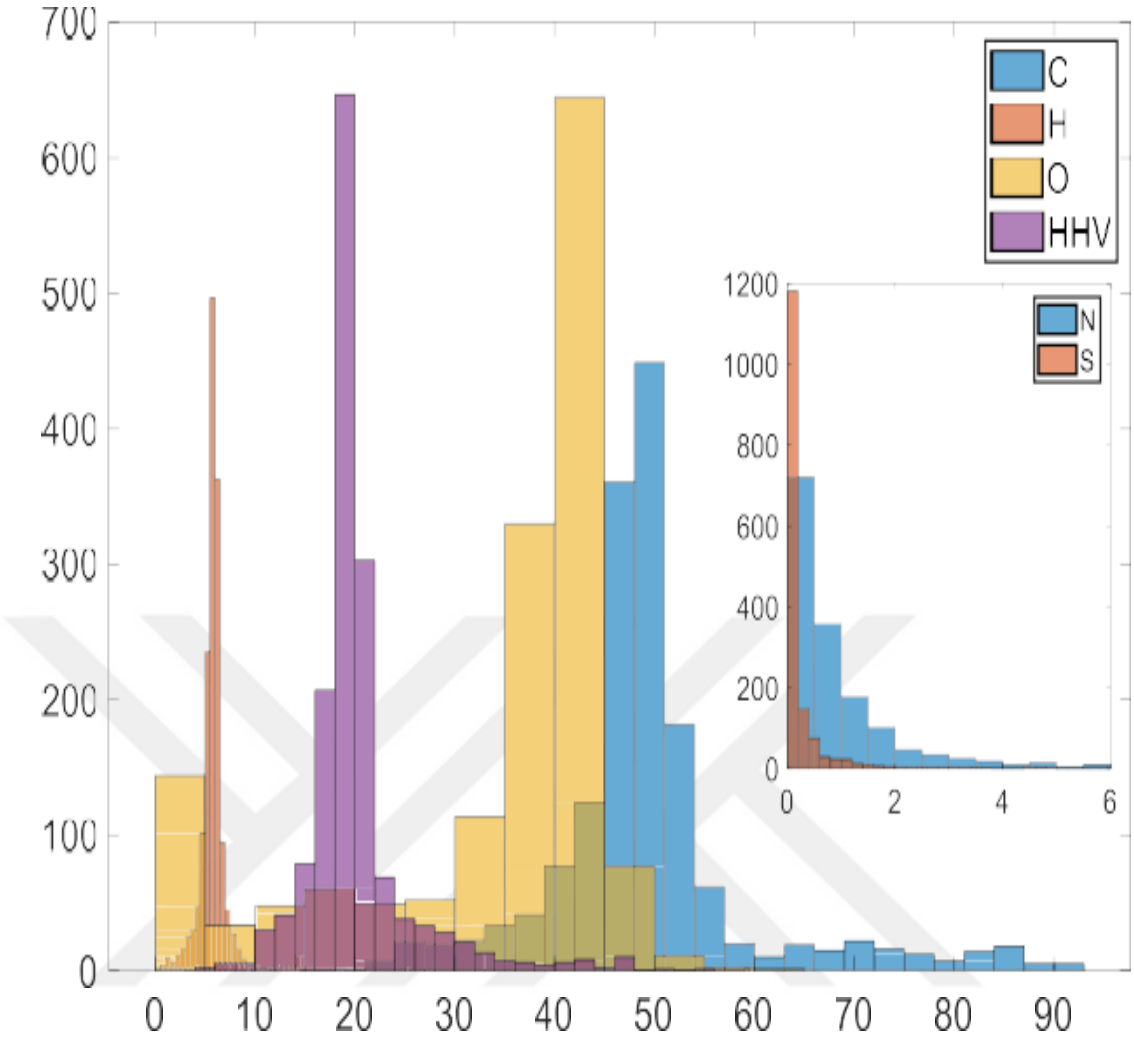


Figure 2. 1 Histogram of the collected data

2.4 Support Vector Machine

Support vector regression (SVR) has emerged as a successful method with studies conducted in the last decade. This method is based on a statistical learning theory. In this way, it solves learning problems with fewer examples [29,30]. Furthermore, support vector regression is used to separate the data belonging to different classes by creating hyperplanes.

With its nonparametric structure and the use of kernel functions, support vector machine regression is one of the methods widely used among machine learning methods. In this way, the difference occurs with the polynomial regression. This case shows that the support vector regression method is a controlled learning technique and minimizing the error.

The difference of support vector regression from other methods is that it optimizes the quadratic function of variables using inequality-type constraints [15]. Looking at some of its disadvantages, the support vector regression methodology is not proper for large data sets. In addition, SVR performs poorly when the number of features for each data point exceeds the training data sample.

Properties are first mapped to a high-dimensional space (F) with a kernel function and associated linearly with the outputs, thus creating the Support Vector Machine model for a regression problem [22]. The general function of Support Vector Machine Regression is given in Equation (8).

$$f(x) = \sum_{i=1}^{n} a_i * K(x_i, x_j) + b \tag{2.3}$$

where, f(x) is the kernel function, b is the bias term and (i = 1, 2, 3 …, n) are constant parameters.

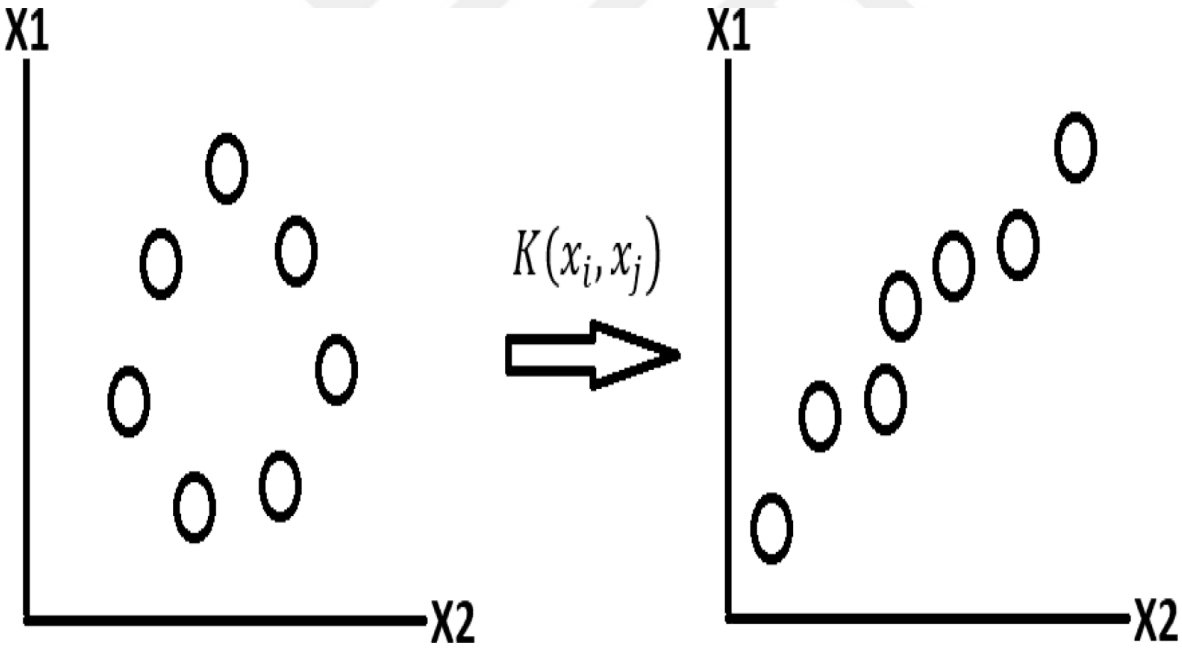


Figure 2. 2 Kernel function

2.5 Decision Tree Regression

A decision tree is a machine learning method commonly utilized to construct predictive models based on basic logical expressions [23]. The basic decision tree, which is directional, consists of a root node, internal nodes, and terminal nodes [31,32].

Decision tree regression (DTR) is used for the prediction of continuous type variables instead of categorical variables.

DTR divides the dataset successively into smaller clusters so that it distributes more homogeneously. Then, it tries to minimize the guessing error in each section. Therefore, the prediction of the continuous output variable is obtained using the logical rule.

The biggest advantage of DTR over other modeling techniques is its ability to construct a model that can be represented as a set of logical and rule expressions, and it can handle multi-output problems. In addition, it is easy to understand and interpret. The tree structures used can be visualized [5]. Despite the remarkable features mentioned, decision trees has disadvantages such as not enabling outputs with more than one feature, yielding partially variable results, being highly sensitive to even slight changes in the test data, and creating highly complex tree structures for numerical data sets [31].

It progresses through the use of decision trees in classification and estimation, constructing a decision tree model based on the training data, evaluating this model using test data using appropriate testing criteria, and using the relevant model in predicting future values [32].

2.6 Gaussian Process Regression

Gaussian process regression (GPR) is one of the machine learning algorithms. This machine-learning algorithm is kernel-based, nonparametric, and based on probability. Regression models based on Gaussian processes (GP) are completely probabilistic, flexible, and simple to apply, and therefore their use is widespread in many application areas. When you look at the disadvantage of this method, it is effectively limited by the problems of a few thousand cases at most. This is because computational demands and memory requirements increase as the number of training cases in cubic and squared, respectively [33].

2.7 Random Forest Regression

Random forest is an emerging learning method integrated with many decision trees [34]. This method can be utilized for regression (RFR) and classification (RFC). While training, by using bootstrap sampling, subsamples of input training samples are

collected, and the forest grows more than one tree to match these subsamples. For estimation process, the model takes average of the results of generated trees to obtain the decisive estimate. Number of features to consider when growing trees and number of trees in the forest are the two crucial parameters in the RFR model [35].

2.8 Validation Methods: Hold-out Validation & k-fold Cross Validation

In general, three types of evaluation methods are used in machine learning approaches. The most common of these are k-fold and Hold-out evaluation methods. The Hold-out evaluation method divides the entire data set into 3 different sets. These sets are validation set, training set, and holdout/test set. It trains the model through the training set, then assesses the validity over the validation set, and finds out the best parameters for the method. The holdout/test sets are used to evaluate the modelling performance. A classification algorithm is used to test the model. It is a method of developing the hold-out method [36].

The k-fold cross-validation is a method that is commonly used in applications, in which the data is separated into k subsets [37]. The evaluation method of dividing a data set into k multiples of equal size randomly is the k-fold cross-validation method. In k-fold cross-validation, plies are utilized for construction of model, and the left-out ply is allocated for pattern validation. This shows that model construction is emphasized rather than the validation procedure. However, some studies have affirmed that more attention on the validation procedure may lead to an improvement in model selection [34,38].

This evaluation model presents what kind of result will be obtained in the data set of the statistical analysis. Cross-validation provides an unbiased estimation of prediction error without splitting the data into train and test sets. The principles of these validation methods are illustrated in Figure 2.4.

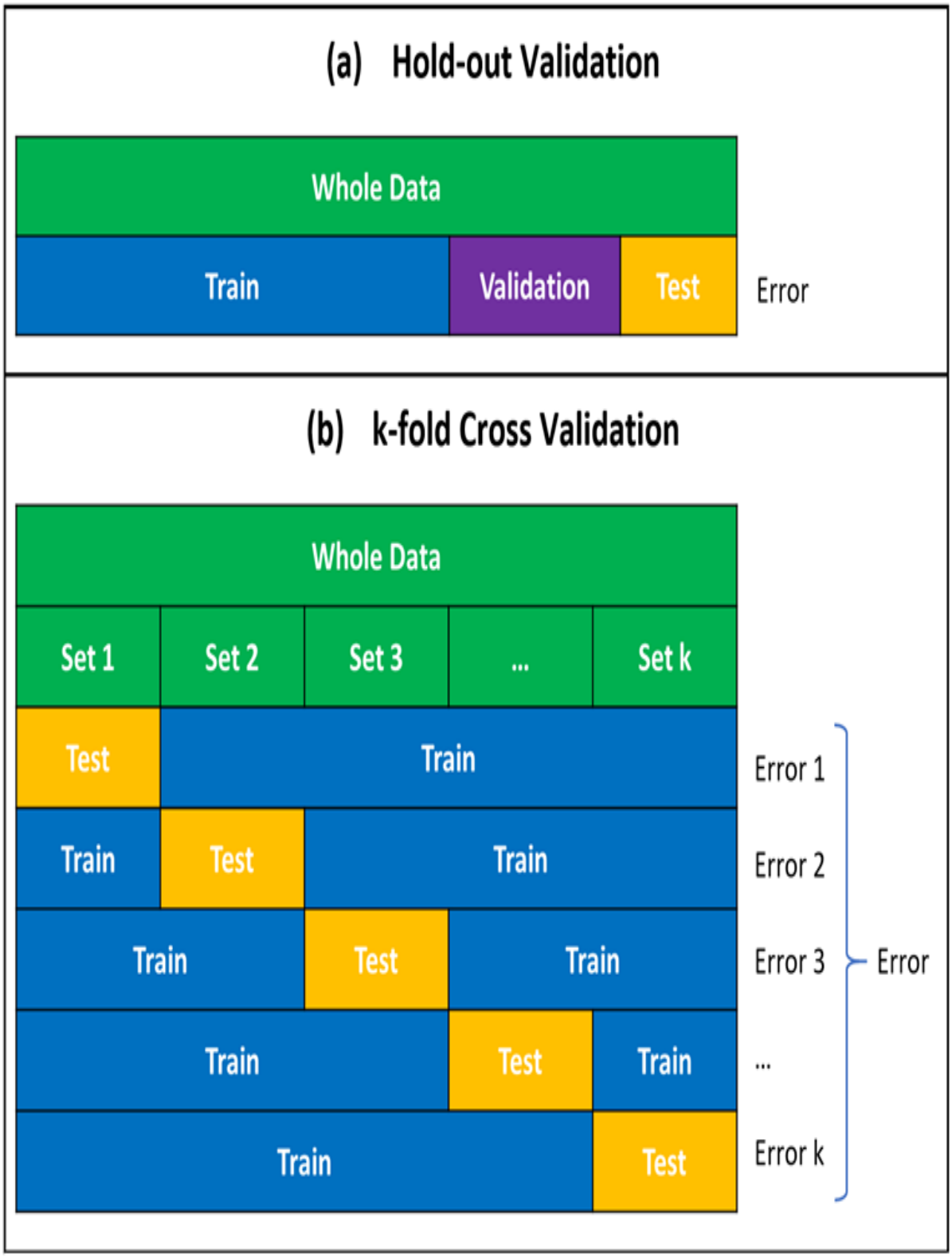


Figure 2. 3 The principles of hold-out validation (a) and k-fold cross validation methods(b)

3
CONCLUSION

3.1 HHV Estimation by Machine Learning Methods

In this study, six features (C (%), H (%), O (%), N (%), S (%), LHV (MJ/kg)) are used to predict the HHV (MJ/kg). Ultimate analysis of fuels is used as the dataset. Recommended machine learning methods are carried out in the MATLAB 2017a environment. The correlations obtained by the multi-linear and polynomial regressions is given in Table 3.1. The correlation heatmap of the inputs are presented in Fig.3.1, and it can be observed that the input features are independent.

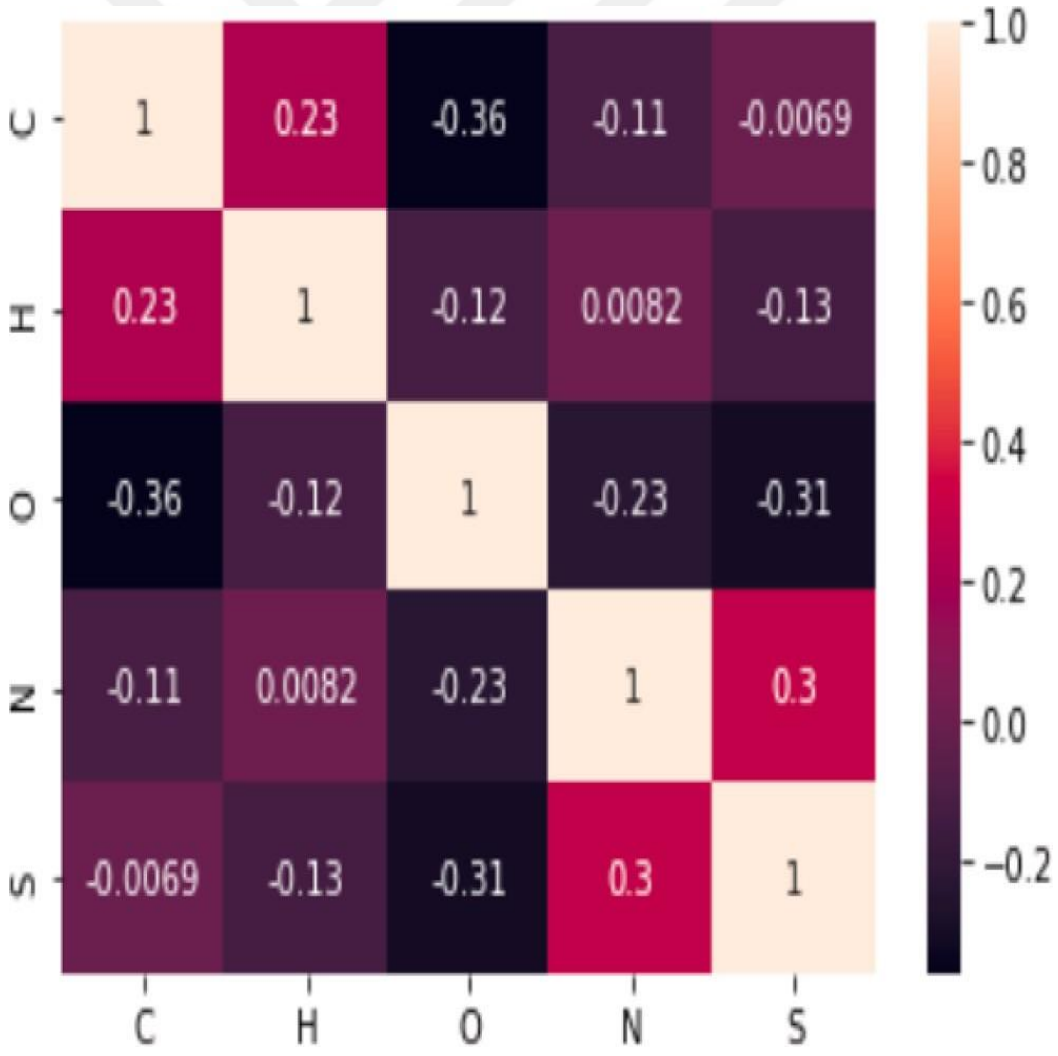


Figure 3. 1 The correlation heatmap of the input data

Table 3. 1 Developed HHV correlations for multilinear and polynomial regressions

Label	Data Size	Correlation
CHONS	1566	$HHV = 0.3988\ C + 0.4422\ H - 0.03281\ O + 0.14161\ N + 0.16449\ S - 1{,}2946$
CHO	1566	$HHV = 0.40339\ C + 0.64814\ H - 0.046934\ O - 1.9085$
CH	1566	$HHV = 0.42535\ C + 0.65981\ H - 4.6428$
$C^{1.4}H$	1566	$HHV = 0.062488\ C^{1.4} + 0.6152\ H + 1.7542$

10-fold cross-validation technique is used in the validation of methods. Average absolute errors (AAE), root-mean square-error (RMSE), root-mean-square-logarithmic error (NRMSE), correlation coefficient (R^2) and adj. R^2 metrics are computed to evaluate the performance of the proposed methods (see Table 3.2). The correlation diagrams of each proposed method is presented in Figure 3.1.

Table 3. 2 Performance evaluation of the proposed machine learning algorithms

	R^2	Adj. R^2	RMSE	N-RMSE	AAE
DTR	0.9019	0.9017	1.6360	0.0873	0.0496
SVR	0.8926	0.8924	1.7121	0.0819	0.0492
GPR	0.9174	0.9173	1.5192	0.0660	0.0408
RFR	0.9244	0.9243	1.4359	0.0805	0.0440
CHONS	0.8916	0.8913	1.7195	0.0848	0.0501
CHO	0.8894	0.8892	1.7374	0.0849	0.0516
CH	0.8849	0.8848	1.7932	0.0935	0.0580
$C^{1.4}H$	0.8903	0.8902	1.7506	0.0845	0.0505

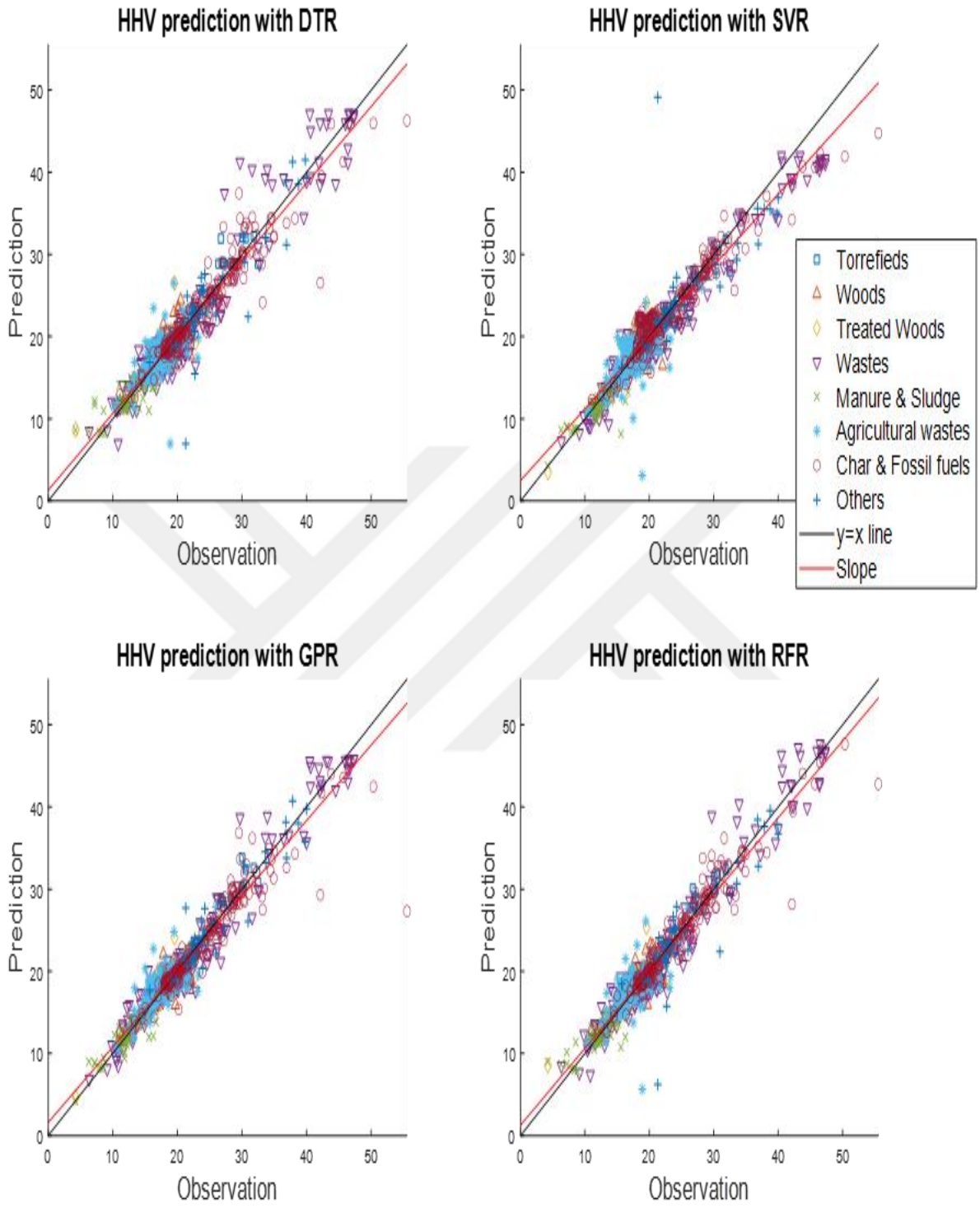


Figure 3. 2 HHV prediction with proposed methods

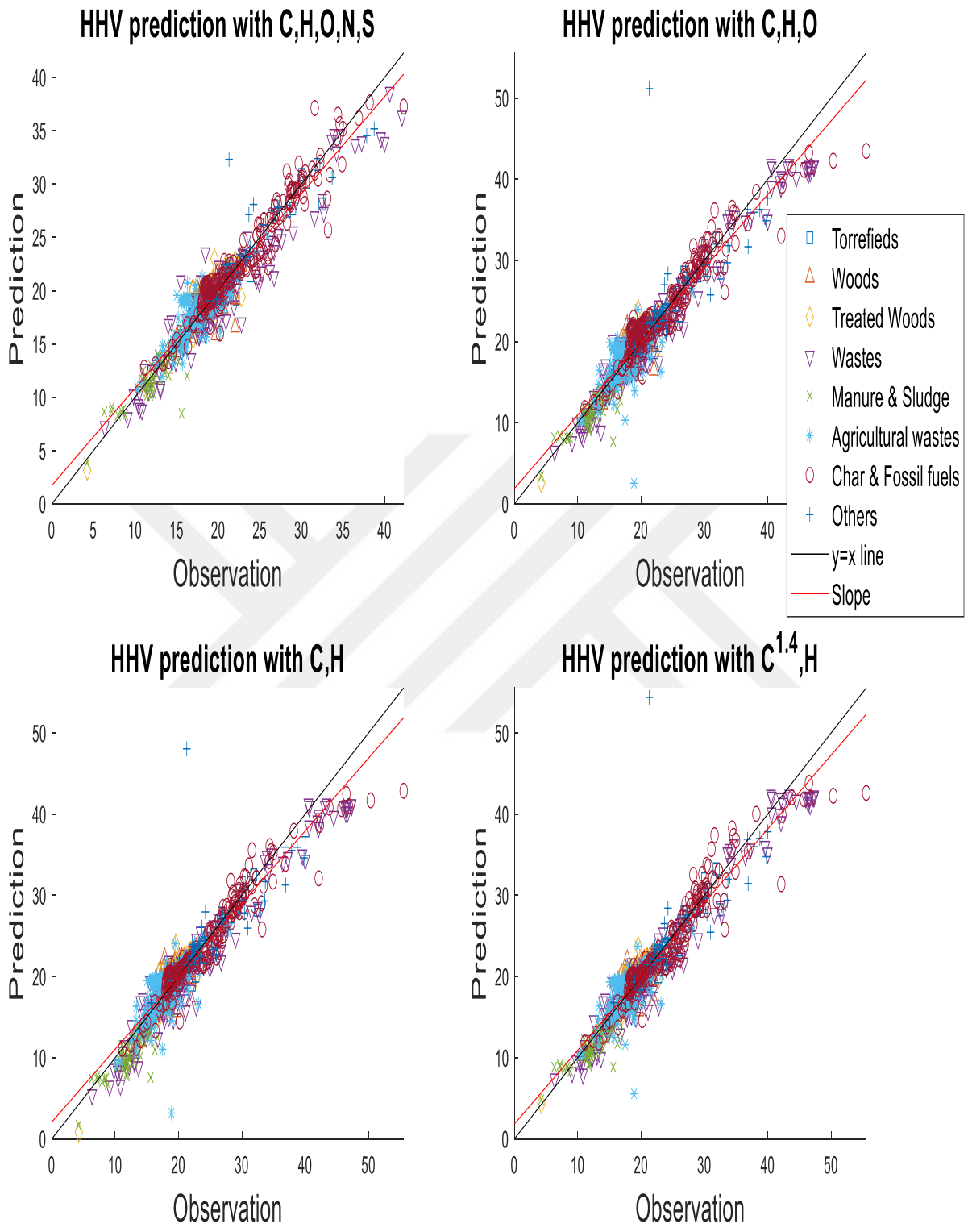


Figure 3. 3 HHV prediction with proposed linear and polynomial models

From Table 3.2, it is seen that the highest R^2 and adjusted R^2 values obtained by RFR, followed by GPR and CHONS. The RMSE, N-RMSE, and AAE values were also the lowest for these methods. It is also observed from Figure 3.2 that all proposed methods perform sufficiently well at prediction of HHVs. However, each method was more preferable for some fuel classes and undesired for others.

The DTR correctly estimated the HHV of torrefields, woods, treated woods, and manure & sludge, but slightly underestimated the HHV of char & fossil fuels, and overestimated the HHV of wastes. The SVR performed considerably better than DTR in estimation of HHV of wastes, but underestimation of HHV of char & fossil fuels was still present. In fact, only the CHONS method was successful at estimating the HHV of char & fossil fuels among the proposed methods.

GPR performed exceptionally well overall and may be preferable to the other methods except CHONS. The RFR, even though it obtained highest R^2 values, resulted noticeable underestimations while predicting the HHV of agricultural wastes, char & fossil fuels, and others. The multi-linear regression containing 5 variables, CHONS, performed the best and sufficiently represented all classes of fuels. CHO, CH, $C^{1.4}H$ methods were sufficient at prediction of HHV of torrefields, woods, treated woods, wastes, and Manure & Sludge.

Since the utilization of linear and polynomial regression is relatively straightforward, these methods may be more preferable to the DTR, SVR, and RFR methods. Moreover, with the hyper-parameter optimization, it would be possible to obtain greater prediction performance from these machine learning algorithms. The suggested fuel classes for application of each method are summarized in the Table 3.3.

Table 3.3 Suggested fuel classes for application of each method

Method	Suggested Biomaterial Classes for Application
DTR	Torrefields, Woods, Treated Woods, Manure & Sludge
SVR	Torrefields, Woods, Treated Woods, Manure & Sludge, Wastes
GPR	Torrefields, Woods, Treated Woods, Manure & Sludge, Agricultural Wastes, Others

Table 3.3 Suggested fuel classes for application of each method (continued)

RFR	Torrefields, Woods, Treated Woods, Manure & Sludge
CHONS	All classes
CHO	Torrefields, Woods, Treated Woods, Wastes, Manure & Sludge
CH	Torrefields, Woods, Treated Woods, Wastes, Manure & Sludge
$C^{1.4}H$	Torrefields, Woods, Treated Woods, Wastes, Manure & Sludge

When these results are examined, it is observed that GPR and CHONS give better results compared to other proposed methods. However, as previously discussed, each method has its own strengths, and for specific classes of biomaterial, the other methods may preferably be utilized as well.

3.2 Classification of Fuel

3.2.1 Investigation of Metadata of Ultimate Analysis Dataset

A large amount of technological options available for generation of power from fuels. Hence, the type of fuel is of great importance in energy applications in order to choose best procedures and processing methods to obtain optimum results.

While the ultimate analysis data itself may be utilized for constructing sufficient correlations and algorithms to classify fuels, as discussed in a previous study [34], we have observed that the metadata also provides crucial insight about the classes of fuels. Hence, the minimum, maximum and average values of ultimate analysis data for each class of fuel is presented in related table.

When the table is examined, it is observed that the torrefied materials has relatively high carbon and low sulfur content on average, with high values of LHV and HHV. The wastes have 6 times more sulfur content than torrefied materials and treated woods on average, and it also contains a noticeable amount of nitrogen. It has the highest hydrogen content, and moderate carbon content. The manure and sludge class fuels contain the highest amount of sulfur and nitrogen and should be treated accordingly. It is important that examine the biomass samples according to the features.

Agricultural wastes have the highest oxygen content on average which makes the fuels of this class easily ignitable, but it has the lowest carbon content, resulting lowest values of LHV and HHV among the 8 classes.

Char and fossil fuels, as expected, has the highest LHV and HHV values due to the high carbon content of the fuels. It has considerable amount of sulfur as well, but the high energy output of this type of fuels generally compensates for the complications of the additional treatments. The difference in the minimum and maximum values show that this class is more diverse than others and properties of fuels in this class may differ from one another.

3.2.2 Van Krevelen Diagram

In this study, Van Krevelen diagram (Figure 3.3) was plotted using the data available from Phyllis 2 database. This diagram shows cross-plots the O/C atomic ratio versus the H/C atomic ratio of carbon compounds. Van Krevelen method has been utilized for fuel assessment and interpretation of matter distributions in environmental samples [40]. Plotting the Van Krevelen diagram indicated that the classes of fuels have their distinct and individual region in the diagram.

One of the parameters that increase the higher heating value is the H/C ratio. According to the diagram, the H/C ratio of the samples specified as waste, coal & fossil fuel and others is higher than the other samples, while the H/C ratio of the other samples remains lower. This situation is expected since the C, H, O, N, ash, char contents of each sample were different from each other. The H/C ratio was in the range of 0.8–1.0 and the O/C ratio was in the range of 1.0–1.5 for agricultural wastes.

Almost all of the agricultural waste is taken at the center part of the diagram in, since the lignin ratio of these wastes are higher than the other fuel samples. Due to the same reason, we see that most of the treated woods data are located at the center of the diagram. Besides, wood fuel is in a similar relationship to the positions of processed waste, waste, char & fossil fuels. The O/C ratio was in the range of 0–0.6 and the H/C ratio was in the range of 0–2.0 for these fuel samples. When the diagram is examined in detail, we observe that some data in the waste group are at outliers. The most important reason for this is that the contents of the wastes vary, and include materials such as electronic devices, car wastes, PVC, carpets in non-organic samples. In the diagram, it is observed that some data are distributed on the y-axis.

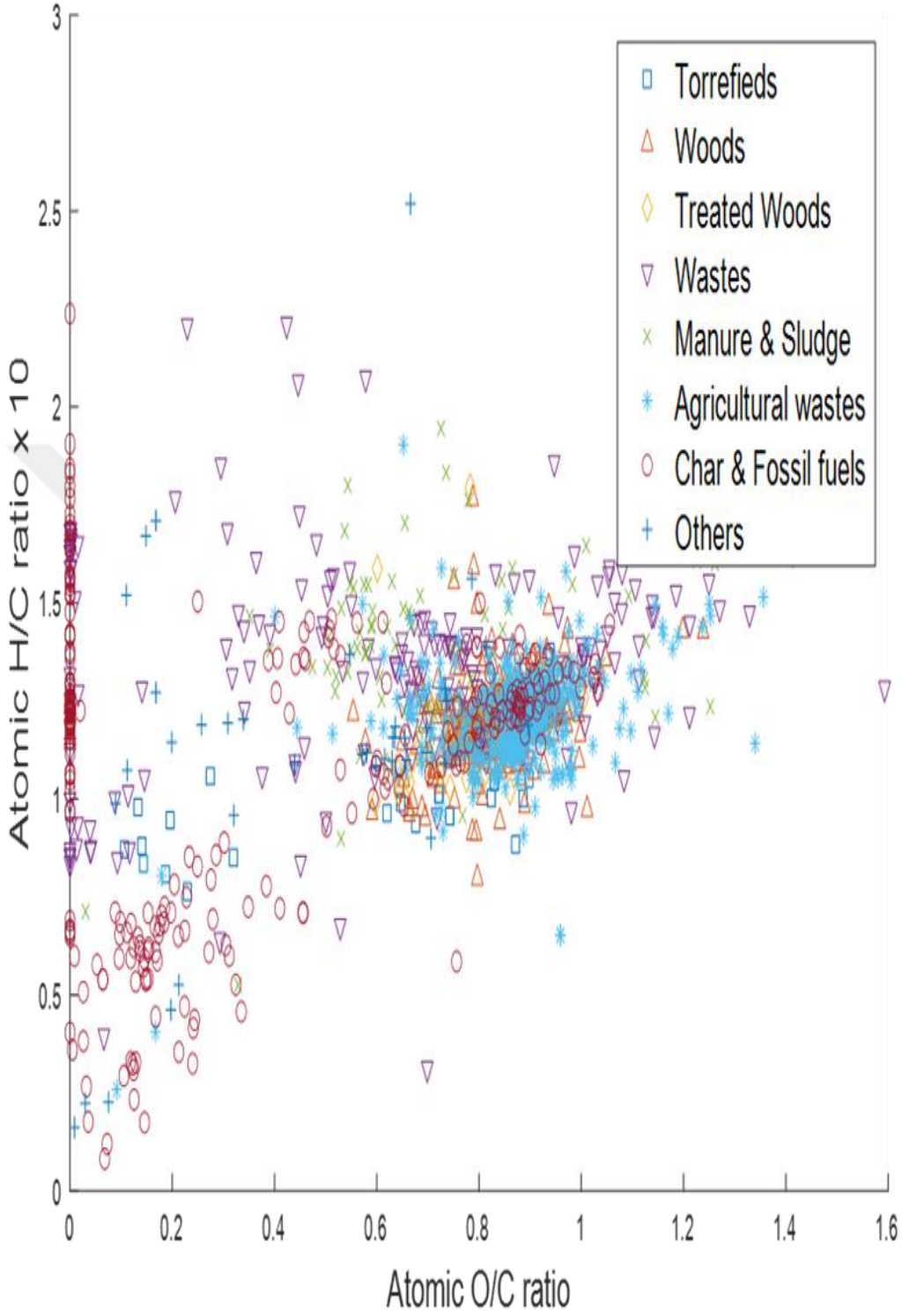


Figure 3. 4 Atomic O/C ratio x 10 vs Atomic H/C ratio graph for fuel samples

4
RESULTS AND DISCUSSIONS

In this study, different machine learning methods utilized to predict the HHVs of fuels from the ultimate analysis. 1514 datapoints were used to train and validate the proposed models, which is the largest data size used in the literature. All the methods were validated by using 10-fold cross validation technique. It was observed that the CHONS is sufficient for prediction of HHV of all kinds of fuel. The GPR method also performed remarkably well, only slightly underestimating HHV of some of the char & fossil fuels. The other methods proposed in this study may also be preferably used by researchers for specific fuel, as discussed in the previous chapter. Additionally, the metadata of the ultimate analysis is investigated in detail, and novel insights regarding the classes of fuels are presented, which may be useful for accurate determination of type of the fuels. Lastly, the Krevelen diagram was illustrated to observe the relationship between H/C and O/C ratios for different kinds of fuels. Among the studies done so far in the literature, the presented Krevelen diagram is the most extensive diagram in terms of considered data size.

This study provides an example of prediction of HHVs of fuels from a dataset of ultimate analysis by using machine learning algorithms. The proposed methods may be utilized by researchers to predict the HHV from ultimate analysis. The presented Krevelen diagram may be used to predict the classes of fuels with unknown origin. As a future direction, the performance of other machine learning approaches may be evaluated and compared, and the possibility of HHV estimation from other sources, such as proximate analysis, may be investigated further.

REFERENCES

[1] H. Şenol, E.A. Elibol, Ü. Açıkel, M. Şenol, Türkiye'de Biyogaz Üretimi İçin Başlıca Biyokütle Kaynakları, Bitlis Eren Üniversitesi Fen Bilim. Derg. 6 (2017) 81–92. https://doi.org/10.17798/bitlisfen.315118.

[2] E. Toklu, Biomass energy potential and utilization in Turkey, Renew. Energy. 107 (2017) 235–244. https://doi.org/10.1016/j.renene.2017.02.008.

[3] A. Anca-couce, C. Hochenauer, R. Scharler, Bioenergy technologies , uses , market and future trends with Austria as a case study, Renew. Sustain. Energy Rev. 135 (2021) 110237. https://doi.org/10.1016/j.rser.2020.110237.

[4] S.S. Toor, L.A. Rosendahl, J. Hoffmann, T.H. Pedersen, R.P. Nielsen, E.G. Søgaard, Application of Hydrothermal Reactions to Biomass Conversion, Hydrothermal Liq. Biomass. (2014) 189–217. https://doi.org/10.1007/978-3-642-54458-3.

[5] F. Elmaz, Ö. Yücel, A.Y. Mutlu, Makine Öğrenmesi ile Kısa ve Elemental Analiz Kullanarak Katı Yakıtların Üst Isı Değerinin Tahmin Edilmesi, Int. J. Adv. Eng. Pure Sci. 32 (2020) 145–151. https://doi.org/10.7240/jeps.558378.

[6] S. Acar, A. Ayanoglu, A. Demirbas, Determination of Higher Heating Values (Hhvs) of Biomass Fuels, Uluslararası Yakıtlar Yanma Ve Yangın Derg. 119 (2016) 1–3. https://dergipark.org.tr/fce/issue/29250/313144.

[7] A.K. Majumder, R. Jain, P. Banerjee, J.P. Barnwal, Development of a new proximate analysis based correlation to predict calorific value of coal, Fuel. 87 (2008) 3077–3081. https://doi.org/10.1016/j.fuel.2008.04.008.

[8] T. Cordero, F. Marquez, J. Rodriguez-Mirasol, J. Rodriguez, Predicting heating values of lignocellulosics and carbonaceous materials from proximate analysis, Fuel. 80 (2001) 1567–1571. https://doi.org/10.1016/S0016-2361(01)00034-5.

[9] S.A. Channiwala, P.P. Parikh, A unified correlation for estimating HHV of

solid, liquid and gaseous fuels, Fuel. 81 (2002) 1051–1063. https://doi.org/10.1016/S0016-2361(01)00131-4.

[10] J. Xing, K. Luo, H. Wang, Z. Gao, J. Fan, A comprehensive study on estimating higher heating value of biomass from proximate and ultimate analysis with machine learning approaches, Energy. 188 (2019) 116077. https://doi.org/10.1016/j.energy.2019.116077.

[11] D.R. Nhuchhen, P. Abdul Salam, Estimation of higher heating value of biomass from proximate analysis: A new approach, Fuel. 99 (2012) 55–63. https://doi.org/10.1016/j.fuel.2012.04.015.

[12] S. Hosseinpour, M. Aghbashlo, M. Tabatabaei, Biomass higher heating value (HHV) modeling on the basis of proximate analysis using iterative network-based fuzzy partial least squares coupled with principle component analysis (PCA-INFPLS), Fuel. 222 (2018) 1–10. https://doi.org/10.1016/j.fuel.2018.02.126.

[13] J. Parikh, S.A. Channiwala, G.K. Ghosal, A correlation for calculating HHV from proximate analysis of solid fuels, Fuel. 84 (2005) 487–494. https://doi.org/10.1016/j.fuel.2004.10.010.

[14] J. Parikh, S.A. Channiwala, G.K. Ghosal, A correlation for calculating elemental composition from proximate analysis of biomass materials, Fuel. 86 (2007) 1710–1719. https://doi.org/10.1016/j.fuel.2006.12.029.

[15] S.B. Ghugare, S.S. Tambe, Genetic programming based high performing correlations for prediction of higher heating value of coals of different ranks and from diverse geographies, J. Energy Inst. 90 (2017) 476–484. https://doi.org/10.1016/j.joei.2016.03.002.

[16] H. Akoglu, User's guide to correlation coefficients, Turkish J. Emerg. Med. 18 (2018) 91–93. https://doi.org/10.1016/j.tjem.2018.08.001.

[17] L. Ricci, R. Martínez, Adjusted R2-type measures for Tweedie models, Comput. Stat. Data Anal. 52 (2008) 1650–1660. https://doi.org/10.1016/j.csda.2007.05.017.

[18] W. Wang, Y. Lu, Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model, IOP Conf. Ser. Mater. Sci. Eng. 324 (2018). https://doi.org/10.1088/1757-

899X/324/1/012049.

[19] E. Tan, Evaluation of NCEP/NCAR Reanalysis Precipitable Water Data Comparing to Radiosonde Observations for Turkey, Cumhur. Sci. J. 40 (2019) 527–535. https://doi.org/http://dx.doi.org/10.17776/csj.393237.

[20] L. Yi, J. Feng, Y.H. Qin, W.Y. Li, Prediction of elemental composition of coal using proximate analysis, Fuel. 193 (2017) 315–321. https://doi.org/10.1016/j.fuel.2016.12.044.

[21] A. Demirbas, Calculation of higher heating values of fatty acids, Energy Sources, Part A Recover. Util. Environ. Eff. 38 (2016) 2693–2697. https://doi.org/10.1080/15567036.2015.1115924.

[22] A. Dashti, A.S. Noushabadi, M. Raji, A. Razmi, S. Ceylan, A.H. Mohammadi, Estimation of biomass higher heating value (HHV) based on the proximate analysis: Smart modeling and correlation, Fuel. 257 (2019) 115931. https://doi.org/10.1016/j.fuel.2019.115931.

[23] K. Özkan, Ş. Işık, Z. Günkaya, A. Özkan, M. Banar, A heating value estimation of refuse derived fuel using the genetic programming model, Waste Manag. 100 (2019) 327–335. https://doi.org/10.1016/j.wasman.2019.09.035.

[24] I. Boumanchar, K. Charafeddine, Y. Chhiti, F.E. M'hamdi Alaoui, A. Sahibed-dine, F. Bentiss, C. Jama, M. Bensitel, Biomass higher heating value prediction from ultimate analysis using multiple regression and genetic programming, Biomass Convers. Biorefinery. 9 (2019) 499–509. https://doi.org/10.1007/s13399-019-00386-5.

[25] C.-Y. Yin, Prediction of higher heating values of biomass from proximate and ultimate analyses, Fuel. 90 (2010) 1128–1132. https://doi.org/10.1016/j.fuel.2010.11.031.

[26] Phyllis 2, Database for biomass and waste, (n.d.). ecn.nl/phyllis2.

[27] T. Zhang, Q. Zhang, Q. Wang, Model detection for functional polynomial regression, Comput. Stat. Data Anal. 70 (2014) 183–197. https://doi.org/10.1016/j.csda.2013.09.007.

[28] T.S. Gendy, T.M. El-Shiekh, A.S. Zakhary, A polynomial regression model for stabilized turbulent confined jet diffusion flames using bluff body

burners, Egypt. J. Pet. 24 (2015) 445–453. https://doi.org/10.1016/j.ejpe.2015.06.001.

[29] B. Venkoba Rao, S.J. Gopalakrishna, Hardgrove grindability index prediction using support vector regression, Int. J. Miner. Process. 91 (2009) 55–59. https://doi.org/10.1016/j.minpro.2008.12.003.

[30] C.J. Lu, T.S. Lee, C.C. Chiu, Financial time series forecasting using independent component analysis and support vector regression, Decis. Support Syst. 47 (2009) 115–125. https://doi.org/10.1016/j.dss.2009.02.001.

[31] L. Rokach, O. Maimon, Classification Trees, in: O. Maimon, L. Rokach (Eds.), Data Min. Knowl. Discov. Handb., Springer US, Boston, MA, 2010: pp. 149–174. https://doi.org/10.1007/978-0-387-09823-4_9.

[32] Y. Zhao, Y. Zhang, Comparison of decision tree methods for finding active objects, Adv. Sp. Res. 41 (2008) 1955–1959. https://doi.org/10.1016/j.asr.2007.07.020.

[33] J.Q. Candela, C. Rasmussen, A Unifying View of Sparse Approximate Gaussian Process Regression, J. Mach. Learn. Res. 6 (2005) 1939–1959.

[34] F. Elmaz, B. Büyükçakır, Ö. Yücel, A.Y. Mutlu, Classification of solid fuels with machine learning, Fuel. 266 (2020). https://doi.org/10.1016/j.fuel.2020.117066.

[35] S. Yang, L. Zhang, J. Fan, B. Sun, Experimental study on erosion behavior of fracturing pipeline involving tensile stress and erosion prediction using random forest regression, J. Nat. Gas Sci. Eng. 87 (2021) 103760. https://doi.org/10.1016/j.jngse.2020.103760.

[36] T.T. Wong, Parametric methods for comparing the performance of two classification algorithms evaluated by k-fold cross validation on multiple data sets, Pattern Recognit. 65 (2017) 97–107. https://doi.org/10.1016/j.patcog.2016.12.018.

[37] Z. Xiong, Y. Cui, Z. Liu, Y. Zhao, M. Hu, J. Hu, Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation, Comput. Mater. Sci. 171 (2020) 109203. https://doi.org/10.1016/j.commatsci.2019.109203.

[38] Y. Jung, Multiple predicting K-fold cross-validation for model selection, J. Nonparametr. Stat. 30 (2018) 197–215. https://doi.org/10.1080/10485252.2017.1404598.

[39] H. Song, K.S. Quinton, Z. Peng, H. Zhao, N. Ladommatos, Effects of Oxygen Content of Fuels on Combustion and Emissions of Diesel Engines, Energies. 9 (2016). https://doi.org/10.3390/en9010028.

[40] S. V. Vassilev, D. Baxter, L.K. Andersen, C.G. Vassileva, An overview of the chemical composition of biomass, Fuel. 89 (2010) 913–933. https://doi.org/10.1016/j.fuel.2009.10.022.

PUBLICATIONS FROM THE THESIS

Papers

1. H. Yaka, M.A. Insel, O. Yucel, Hasan Sadikoglu “A Comparison of Machine Learning Algorithms for Estimation of Higher Heating Values of Biomass and Fossil Fuels From Ultimate Analysis”, at Fuel, July 2022.