Int. J. Adv. Eng. Pure Sci. 2020, 32(2): 145-151
DOI: 10.7240/jeps.558378

**RESEARCH ARTICLE** / ARAŞTIRMA MAKALESİ

# Machine Learning Based Approach for Predicting of Higher Heating Values of Solid Fuels Using Proximity and Ultimate Analysis

## Makine Öğrenmesi ile Kısa ve Elemental Analiz Kullanarak Katı Yakıtların Üst Isı Değerinin Tahmin Edilmesi

**Furkan ELMAZ[1], Özgün YÜCEL*[2], Ali Yener MUTLU[1]**

*[1]İzmir Kâtip Çelebi University, Electrical and Electronics Engineering, Çiğli, İzmir, Turkey*
*[2]İzmir Kâtip Çelebi University, Environmental Engineering, Çiğli, İzmir, Turkey*

**Abstract**
Prediction of higher heating value (HHV) using proximity and ultimate analysis is an important procedure for understanding the characteristic attribute of a fuel. Researches put effort to model the relationship between the HHV value and those analyses. But conducted methods usually included only simple statistical analysis. In this paper we approach this prediction problem from the machine learning perspective, we employ four machine learning methods, i.e. linear regression, polynomial regression, decision tree regression and support vector regression to predict HHV using proximity and ultimate analysis of different type of materials. Data set used is collected from literature and is categorized, where the resulting categories are used as features to be fed to the machine learning models to create prediction models as accurate as possible. Performances of the proposed methods are evaluated with k-fold cross-validation technique and each method's pros and cons are discussed for both prediction accuracy and computational complexity. Polynomial regression proved itself as the most optimal choice among others from these perspectives.
**Keywords:** Biomass, coal, higher heating value, decision tree, support vector machines, cross-validation.

**Öz**
Kısa ve elemental analiz kullanılarak üst ısıl değerinin (ÜID) öngörülmesi, bir yakıtın karakteristik niteliğini anlamak için önemli bir prosedürdür. Araştırmalar, ÜID değeri ile bu analizler arasındaki ilişkiyi açıklamak için modelleme çalışmaları yapmışlardır. Ancak uygulanan yöntemler genellikle sadece basit istatistiksel analizleri içermektedir. Bu makalede, bu tahmin sorununa makine öğrenme perspektifinden yaklaşılmaktadır, farklı türdeki malzemelerin kısa ve elemental analizini kullanarak UID'yi tahmin etmek için dört makine öğrenme yöntemi, yani doğrusal regresyon, polinom regresyonu, karar ağacı regresyonu ve destek vektör regresyonunu kullanılmıştır. Kullanılan veri seti literatürdeki farklı kaynaklardan temin edilerek, kategorilere ayrılmış; sonuçta elde edilen kategoriler, mümkün olduğunca doğru tahmin modelleri oluşturmak için makine öğrenme modellerine beslenecek girdiler olarak kullanılmıştır. Önerilen yöntemlerin performansları k-katlı çapraz doğrulama tekniğiyle değerlendirilerek, her yöntemin performans değerleri hem tahmin doğruluğu hem de hesaplama karmaşıklığı açısından tartışılmıştır.
**Anahtar Kelimeler:** Biyokütle, çapraz doğrulama, destek vektör makinaları, karar ağacı, kömür, üst ısı değeri.

## I. INTRODUCTION

HHV is an important characteristic of a fuel, defined as the amount of heat released during the combustion of a specified amount of the material. Usage of ultimate analysis (elemental composition) of the fuel is a widely used methodology for researchers to predict HHV of a fuel. Ultimate analysis gives the elemental composition of a fuel. Its determination is relatively difficult and expensive compared to proximate analysis [2]. Various types of correlation equations between ultimate analysis (C%, H%, O%, N%, S%) and HHV (MJ/kg) have been proposed in previous studies which try to make a most accurate estimation. In Selvig et. al., Strache and Lant, D'Huart, Gumz et.al., authors attempted to estimate HHV of few types of coal with experimental data they acquired [1-4]. With little to no difference between obtained correlation equations, their methodology was to fit a linear line to the data, i.e. linear regression which is a simple statistical analysis. Even though it is known that there is a linear

**Corresponding Author:** Özgün YÜCEL, Tel: +902323293535 / 3807, e-posta: ozgun.yucel@ikc.edu.tr
**Submitted**: 26.04.2019, **Revised**: 23.10.2019, **Accepted**: 18.02.2020

relationship between ultimate analysis and HHV, expanding the hypothesis equation with different interactions of the terms could help to improve the accuracy and validity of the proposed correlations. Although the proposed correlations are satisfactory for predicting the HHV from the ultimate analysis, researchers encounter with a different problem, requirement of advanced laboratory equipment for ultimate analysis. This requirement makes the analysis of a material for its HHV harder and expensive. For this purpose, researchers resort to estimate HHV by proximite analysis. Proximate analysis of a fuel provides the percentage of the material that burns in a gaseous state (volatile matter), in the solid state (fixed carbon), and the percentage of inorganic waste material (ash), and is therefore of fundamental importance for biomass energy use [4]. This analyzing technique requires nothing more than commonly available laboratory equipment. But, as a consequence, prediction of HHV by proximity analysis resulted in worse accuracy compared to the ultimate analysis. In Matin et. al [5], authors used multiple linear regression and random forest methods to predict HHV by using proximity analysis of coals obtained from different states of USA. Even though results are extremely good, the overfitting problem and scalability of the methods for other materials are not discussed, therefore the reliability and stability of the prediction models are not clearly justified. In Parikh et.al [6], authors used a large number of data sets with different types of materials, i.e., different types of coals and biomass. They employed linear regression method to find a correlation. Obtained results were somewhat promising because authors did not split the data set into training and test folds, the true prediction capabilities of the models weren't investigated. Also, the lack of application of different modeling techniques resulted in incapability for explaining the phenomenon in-depth.

When the importance of HHV for thermo-chemical applications is concerned, development of prediction models in order to predict HHV of material rather using the proximity analysis or ultimate analysis is a crucially important step. Therefore, in this paper, we approach the prediction of HHV from the proximity analysis and ultimate analysis as a supervised machine learning problem. We merged a data set from the literature with samples consisting of proximity analysis, ultimate analysis and HHV of different types of materials. Furthermore, the samples are categorized according to the type of material and these categories are regarded as features of the HHV, which are used during the development of the models. Linear regression, decision tree regression methods which are used previously in similar studies as well as polynomial regression and support vector regression regression methods which have not been used in the literature are employed to develop prediction models. Moreover, k-fold cross-validation technique is used to show unbiased prediction accuracy and reliability of each method.

# II. METHODS

## 2.1. Data Collection and Preprocessing

Data set used in this paper is collected from previous studies in the literature, the data set consists of material type; material type, proximity analysis of the material, i.e., FC (%), VM (%) and ash (%), ultimate analysis of the material, i.e. C (%), H (%), O (%), N (%), S (%) and HHV (MJ/kg) [5,6]. Moreover, we categorized the material types of each data set into 14 categories, i.e. coals/coke, manufactured fuel/wood, pit/shells/seeds/cobs, wood/energy crops, barks/prunings, straws, stalks, fibrous material/leaves/grass, hull/husk/dust, biomass waste material, milling industry waste, refuse/MSW, biomass chars and other biomass/misc. The created categories are also used as features during the development of the machine learning models. Therefore, we create a data set consisting of 185 samples, 21 features and one output.

In order to prepare data set to be used in machine learning methods, preprocessing of the set must be handled. Data preprocessing consists of several steps which depend on the quality of the data set. Data set we merged consists of missing values and categorical variables due to categorization. Therefore, missing values are filled with the median of the corresponding column and categorical variables are one-hot encoded into binary digits [5]. Moreover, numeric columns in the data set is linear scaled to the unit range (Eq. 1) to ensure none of the larger valued features have excessive dominance to the smaller ones and to have better convergence speed during the training phase of the models [5]:

$$\tilde{x}_i[n] = \frac{x_i[n] - \min(x_i)}{\max(x_i) - \min(x_i)} \quad (1)$$

where, $x_i[n]$ is the n$^{th}$ sample of the i$^{th}$ feature of the data set, $\max(x_i)$ and $\min(x_i)$ are the maximum and minimum values of the i$^{th}$ feature, and $\tilde{x}_i[n]$ is the scaled n$^{th}$ observation of the i$^{th}$ feature.

## 2.2 Linear Regression

Linear regression is a simple statistical analysis technique which tries to demonstrate the output as a linear combination of the features with a constant value (Eq. 2). It is widely used in different disciplines due to its easy implementation and high prediction accuracy when the features and output have close to linear correlation [6]. As discussed in the introduction part, most of the studies focused to predict HHVs with using linear regression analysis. Therefore, we employ the linear regression method to create a reference to the results of other methods to show the advantages and disadvantages of each method. Least-Square approximation, i.e. normal equation is used to train the

model. Even though this approach is quite computationally expensive with larger data sets with an excessive number of samples and features, for the data set we collected, we encountered no such problem.

$$\hat{y} = \beta_0 + x_1 * \beta_1 + x_2 * \beta_2 + \cdots x_n * \beta_n \quad (2)$$

where, $\hat{y}$ is the predicted output, $x_1, x_2, \ldots x_n$ are the features of the data set starting from 1 up to n, $\beta_0, \beta_1, \ldots \beta_n$ are the coefficient determined in the training phase of the model.

**2.3 Polynomial Regression**

Polynomial regression expands the idea of linear regression into polynomial equations. Polynomial regression aims to present a relationship between output and the features with a polynomial equation. Because there is not a single type of polynomial equation, one must decide the equation according to the problem encountered. In this paper, several polynomial equations have experimented and quadratic polynomial equation selected as the best performing equation for predicting the HHVs by using both proximity and ultimate analysis (Eq. 3). One of the superior attributes of the polynomial regression compared to the linear regression is the usability of different interactions between features such as the square of a feature, or any interaction term between features can be used during the selection of the polynomial model. But, this attribute results with a possible overfitting problem where models learn the training data very well, which results in reduced prediction performance when the model tries to estimate data it has never seen before referred to as testing data [7]. This situation must be avoided to make successful prediction models with high generalization ability. In this paper, we add regularization term to the cost function (Eq. 4) to prevent the prediction model from overfitting:

$$\hat{y} = \sum_{i=1}^{n}\sum_{j=i}^{n} \beta_{ij} * x_i * x_j + \sum_{k=1}^{n} \alpha_k * x_k \quad (3)$$

$$J = \sum_{t=1}^{m} (y_i - \hat{y}_i)^2 + \lambda \left( \sum_{i=1}^{n}\sum_{j=i}^{n} {\beta_{ij}}^2 + \sum_{k=1}^{n} {\alpha_k}^2 \right) \quad (4)$$

where, J is the cost function, y is the actual output vector, m is the number of samples in the data set, $\beta_{ij}$ and $\alpha_k$ are the parameters determined in the training phase, and $\lambda$ is the regularization parameter.

**2.4 Decision Tree Regression**

Decision Tree Regression (DTR) is a type of decision tree algorithms. Unlike the basic decision tree which is used for classification problems, DTR is used for predicting continuous type variables rather than the categorical ones. The DTR uses binary tree to recursively divide the output space into sub-sets where the distribution of output is more homogeneous in succession [8]. The algorithm then tries to minimize the prediction error in the separate parts. The same process is then applied to each new branch. Therefore, the prediction of continuous output variable is achieved by using a set of logical rules. Like many of the complex machine learning algorithms, DTR may fall into overfitting problem. In order to prevent overfitting, pruning process is implemented in the training phase of the model [9]. The greatest advantage of DTR compared to other modeling techniques is the capability of producing a model that can be represented as set of rules and logical statements. In addition, results of the decision tree model provide easy to understand information, for example importance of each feature can be understood by just looking at the presence of the interested feature in hierarchical structure of the DTR model. Also, the basic logical structure of tree is compatible to use with hardware which have low computational power.

**2.5 Support Vector Regression**

Support vector machines (SVM) is a widely used machine learning algorithm used in various areas of applications [10]. The popularity of the SVM is due to its optimization objective, unlike the polynomial regression and decision tree regression, SVR aims to minimize generalization error bound rather than the sum of square errors between prediction and actual outputs [11]. Therefore, during the training of SVR models, the aim to develop a model which generalizes the problem which is the desired attribute expected from a machine learning model. Support vector regression (SVR) is a special case of SVM where the aim is to predict continuous variables. In order to predict a continuous variable, we employed regularized cost function with robust ε-insensitive loss function (Eq. 5) which (in the best case) all outputs would be predicted with the error up to ε value.

$$L_\varepsilon = \begin{cases} 0 & if\ |y - f(x)| \le \varepsilon \\ |y - f(x)| - \varepsilon & otherwise \end{cases} \quad (5)$$

$$J = \sum_{i=1}^{n} L_\varepsilon(f(x_i), y) + \frac{\lambda}{2} |w|^2 \quad (6)$$

where w is the weight matrix determined during the training phase. Another reason for excessive usage of SVM technique is due to its use of kernel functions. With the help of kernel functions, features are mapped into higher vector dimension and simple linear regression is performed with the mapped features to make a prediction (Eq. 7). In this paper, linear kernel function (Eq. 8) is observed to be the kernel function with the highest prediction success.

$$f(x) = \sum_{i=1}^{n} a_i * K(x_i, x_j) + b \quad (7)$$

$$K(x_i, x_j) = (x_i' x_j + 1) \quad (8)$$

where $K(x_i, x_j)$ is the selected kernel function, b is the constant bias term.

### 2.6 K-Fold Cross Validation

K-fold cross-validation method is a technique to evaluate the performance of a method and its implementation with unbiased manner [12]. In k-fold cross-validation the data is divided into k equal sized partitions, i.e. folds, then the candidate model trained using the k-1 folds and tested on the remaining one, this process repeated k times with a different selection of test folds to ensure all of the data is used for testing purpose. Therefore, all of the data is used as both training and testing sets, and the candidate method is judged by its testing performance. This approach not only shows the prediction accuracy and usability of the candidate method, it shows the possible performance variance resulted with random splitting of the data as the model is trained and tested with different folds, this attribute of the k-fold cross validation shows the true and unbiased prediction accuracy and generalization performance of a method which is not possible with traditional train – test splitting. In the present paper, we chose 10 as the 'k' number because it has been observed that 10-fold cross validation provides the most unbiased performance evaluation [13].

## III. RESULTS AND DISCUSSION

The proposed machine learning methods are implemented in MATLAB 2018a environment. As a goal of the present paper, ultimate analysis and proximity analysis features are used individually, and one-hot encoded variables are used mutually to predict the HHVs. Which results with, 18 features and 16 features are used for predicting HHVs by using proximity analysis and ultimate analysis data, respectively. Testing performance of k-fold cross-validation technique is concerned during the evaluation of the methods and correlation coefficient ($R^2$), Adj. $R^2$, root-mean-square-error (RMSE), and root-mean-square-logarithmic-error (RMSLE) metrics are employed to evaluate the prediction performance of the proposed methods. Results are given in Table 1 and Table 2. Moreover, in order to investigate each method's advantages and disadvantages, actual output values versus predictions graphs are plotted in Figure 1 and Figure 2.

$$R^2 = 1 - \left[\frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2}\right] \quad (9)$$

$$Adj.R^2 = 1 - \left[\frac{(1 - R^2) * (m - 1)}{m - n - 1}\right] \quad (10)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{m}} \quad (11)$$

$$RMSLE = \sqrt{\frac{\sum_{i=1}^{m}[\log(y_i + 1) - \log(\hat{y}_i + 1)]^2}{m}} \quad (12)$$

**Table 1.** Performance evaluation for predicting HHVs by using proximity analysis and categorical variables

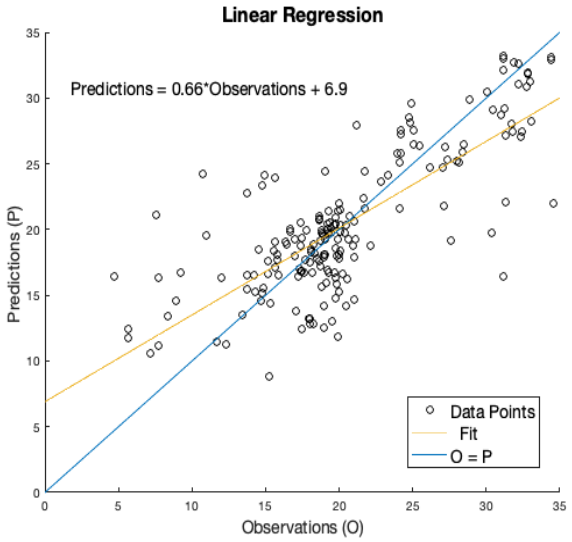
| | Linear Regression | Polynomial Regression | Decision Tree Regression | Support Vector Regression |
|---|---|---|---|---|
| **$R^2$** | 0.64 | 0.84 | 0.78 | 0.84 |
| **Adj.$R^2$** | 0.57 | 0.81 | 0.74 | 0.81 |
| **RMSE** | 3.465 | 2.275 | 2.701 | 2.276 |
| **RMSLE** | 0.167 | 0.115 | 0.133 | 0.123 |

**Table 2.** Performance evaluation for predicting HHVs by using ultimate analysis and categorical variables

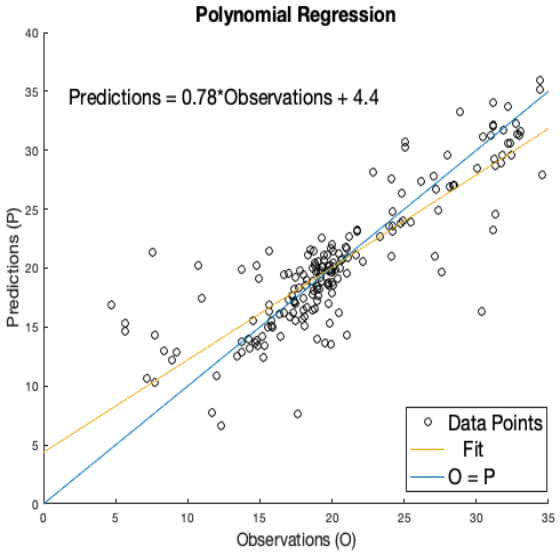
| | Linear Regression | Polynomial Regression | Decision Tree Regression | Support Vector Regression |
|---|---|---|---|---|
| **$R^2$** | 0.75 | 0.76 | 0.76 | 0.77 |
| **Adj.$R^2$** | 0.73 | 0.75 | 0.74 | 0.75 |
| **RMSE** | 3.137 | 2.957 | 3.065 | 2.896 |
| **RMSLE** | 0.209 | 0.197 | 0.213 | 0.188 |

When Table 1 is examined, it is obvious that linear regression was not able to make good predictions of HHVs using proximity analysis compared to other methods, even though it is excessively used in previous studies in the literature for the same problem. One of the primary reasons for this behavior is the limitations of a linear line where the linear regression tries to fit to the data. It can be concluded that only a linear line is not enough to make highly accurate predictions for such a problem. For the case of decision three regression method, although its prediction accuracy was higher than the linear regression method, it is still worse than the remaining two methods. Due to the binary splitting structure of the tree it is not a surprise that the decision tree can't perform as good as in regression as it does at classification problems [14]. On the other hand, polynomial regression and support vector regression performed very similar and better than the rest of the methods. For polynomial

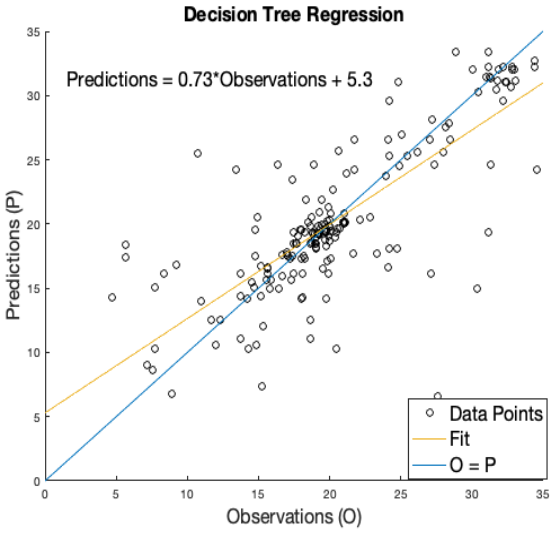
regression, one can see that the quadratic representation of the features is superior to the simple linear form and it can be used for predicting the HHVs using proximity analysis. SVR had almost identical performance to the polynomial regression, but the best performing and selected kernel was the linear kernel, this selection of kernel is not greatly optimal, because we get the most benefit from selection more complex kernel functions for to be used in SVR. Moreover, when the computational complexity of the SVR is concerned, polynomial regression may be a better choice for predicting HHVs from proximity analysis.



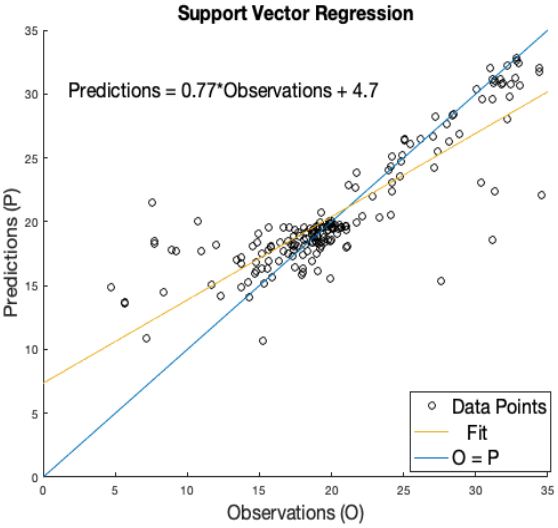

(a)




(b)




(c)



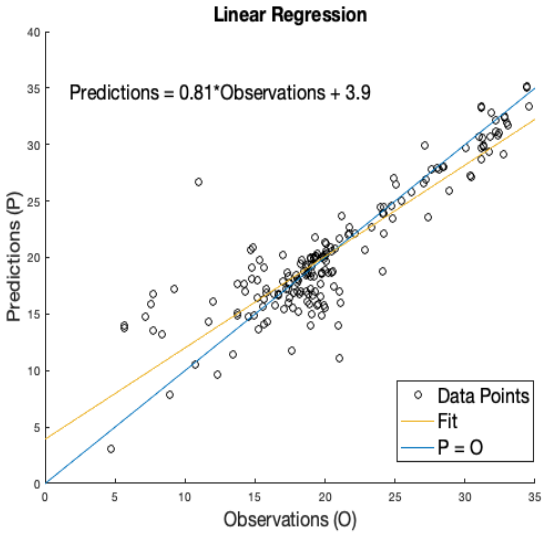

(d)

**Figure 1.** Predictions vs observations graph for all methods using proximity analysis

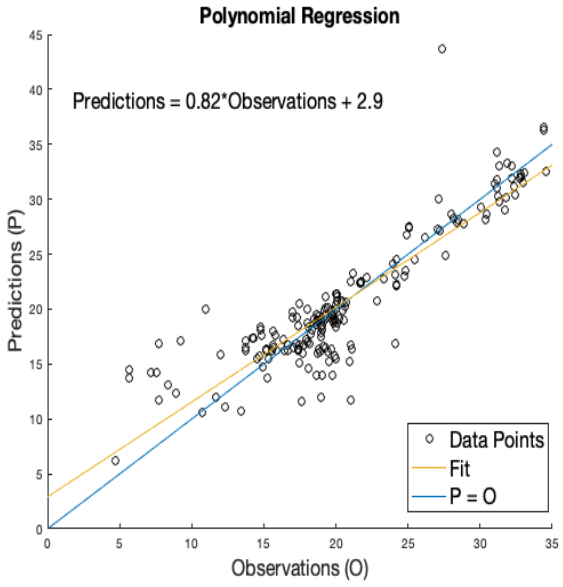
For the prediction of HHVs from ultimate analysis data, as one can see that from Table 2, all of the methods performed almost identical. The only reason for this behavior is ultimate analysis and HHVs have a linear correlation to some degree, which explains why the linear regression performed as same as the other methods. Even though ultimate analysis of material requires an advanced laboratory, if one was able to perform the analysis, it is logical to use linear regression to make accurate predictions for HHVs due to its low computational cost and easiness of the implementation. One can say that other and more advanced methods proposed in this paper is not required and can be a waste of computational resources for such a problem.

Furthermore, one can analyze Figure 1 to have a deeper understanding of each model's ability to predict HHVs using proximity analysis. Linear regression was able to

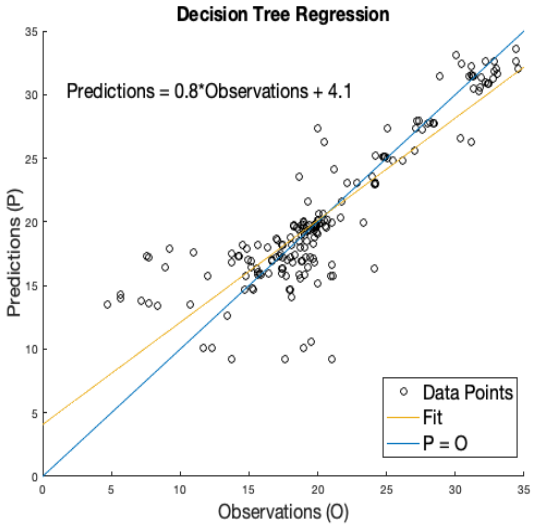
predict just a few of the data points correctly which justifies the unsatisfied results given in Table 1. For the case of decision tree regression, some of the data points were predicted with almost perfect accuracy, but other points were extremely irrelevant. This is due to the nature of the decision tree regression as it uses a strict set of rules, but one must note that with a higher number of samples decision tree regression may perform better. For SVR and polynomial regression, the results are similar and with the exception of a few outlier data points, predictions were close to the observations and have a reasonable distribution.



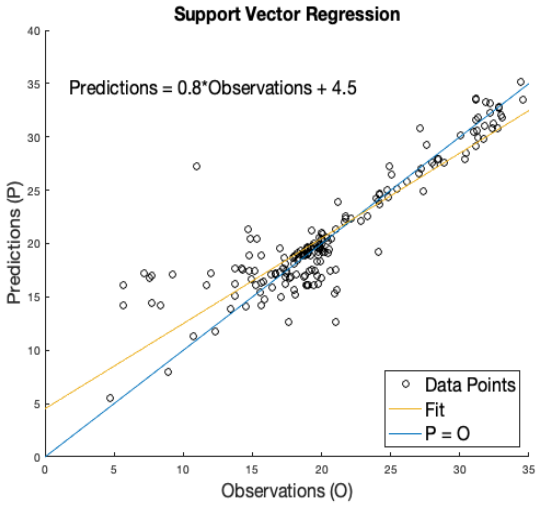

(a)




(b)




(c)




(d)

**Figure 2**. Predictions vs observations graph for all methods using ultimate analysis

In Figure 2, it is obvious that almost all of the methods made high accuracy predictions with exception of few outlier data, and some of the same outlier data points were predicted with large error margin with all methods, one can conclude that those points don't follow the same pattern, don't contain true information for the dynamics of the system and can be ignored, one must note that these outliers may occur due to the handling of the missing data discussed in section 2.1. After all, as a simple statistical analysis technique, linear regression can perform as well as the other methods and it is enough to show the relationship between ultimate analysis and HHVs.

Also, one must note that among each proposed method, SVR is the most computationally complex algorithm due to its hyperplane solution to given regression problems as explained in Section 2.5. On the other hand, linear regression and polynomial regression tends to have similar computational complexity

because they solve single optimization problem throughout their training. Decision tree lays between these methods from the computational requirement perspective and it predict using entropy-based structure as explained in Section 2.4.

## IV. CONCLUSION

In this paper, we collected and merged a data set contains proximity and ultimate analysis of various type of materials and their corresponding HHVs. Furthermore, 4 different machine learning methods are employed and used to predict HHVs using proximity and ultimate analysis separately as well as using categorical information of the materials. For the case of using proximity analysis to predict HHVs, the performance of the linear regression is inadequate as a contrary to its wide usage in literature. Polynomial regression and SVR had similar results and due to simpler implementation and computational requirement, one can prefer to use polynomial regression for such problem. Decision tree regression performed better than linear regression but fell short compared to SVR and polynomial regression. But this situation may differ if one can see a larger data set with more number of samples. For predicting the HHVs from the ultimate analysis, all of the proposed methods made predictions with high accuracy and in a similar fashion. Due to high performance in linear regression, one can say that ultimate analysis and HHV have a high linear correlation. Therefore, advanced machine learning methods are not required for such problem and one can employ simple statistical analysis techniques. As the best performing method, polynomial regression managed to predict HHV value from both proximity and ultimate analysis results among other methods by reaching $R^2$ =78 and $R^2$=82, respectively. For future work, we'll focus on applying a greater number of methods as well as using a larger data set, also other than predicting the HHVs, we aimed to predict ultimate analysis using the proximity analysis with various type of machine learning based modeling methods. We believe that machine learning can help researchers to make highly accurate models of HHV by using proximity and ultimate analysis of the materials.

## REFERENCES


**[1]** Selvig WA, G. I., "Calorific value of coal," *Chem. coal Util.*, vol. 1, p. 139, 1945.

**[2]** Strache H, L. R., "Kohlenchemie," *Akad. Verlagsgesellschaft*, p. 476, 1924.

**[3]** Hosokai, S., Matsuoka, K., Kuramoto, K., and Suzuki, Y., "Modification of Dulong's formula to estimate heating value of gas, liquid and solid fuels," *Fuel Process. Technol.*, vol. 152, pp. 399–405, 2016.

**[4]** Rd, M. and Md, H., "Mass-fraction of oxygen as a predictor of HHV of gaseous, liquid and solid fuels," *Energy Procedia*, vol. 142, pp. 4124–4130, 2017.

**[5]** Matin, S. S. and Chelgani, S. C., "Estimation of coal gross calorific value based on various analyses by random forest method," *Fuel*, vol. 177, pp. 274–278, 2016.

**[6]** Channiwala SA and Parikh PP., "A unified correlation for estimating HHV of solid, liquid and gaseous fuels.," *Fuel*, vol. 81, pp. 1051–63, 2002.

**[7]** Ng, A. Y., "Preventing Overfitting of Cross-Validation Data," in *ICML '97 Proceedings of the Fourteenth International Conference on Machine Learning*, 1997, pp. 245–253.

**[8]** Podgorelec, V. and Zorman, M., "Decision Tree Learning," *Encycl. Complex. Syst. Sci.*, pp. 1–28, 2015.

**[9]** Bertsimas, D. and Dunn, J., "Optimal classification trees," *Mach. Learn.*, vol. 106, no. 7, pp. 1039–1082, 2017.

**[10]** Chih-Wei Hsu, Chih-Chung Chang, and C.-J. L., "A Practical Guide to Support Vector Classification," *BJU Int.*, vol. 101, no. 1, pp. 1396–1400, 2008.

**[11]** Noble, W. S., "What is a support vector machine?," *Nat. Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, 2006.

**[12]** Elmaz, F., Yücel, Ö., and Mutlu, A. Y., "Evaluating the Effect of Blending Ratio on the Co-Gasification of High Ash Coal and Biomass in a Fluidized Bed Gasifier Using Machine Learning," *Mugla J. Sci. Technol.*, vol. 5, no. 1, pp. 1–15, Jun. 2019.

**[13]** Mutlu, A. Y. and Yucel, O., "An artificial intelligence based approach to predicting syngas composition for downdraft biomass gasification," *Energy*, vol. 165, pp. 895–901, Dec. 2018.

**[14]** Xu, M., Watanachaturaporn, P., Varshney, P. K., and Arora, M. K., "Decision tree regression for soft classification of remote sensing data," *Remote Sens. Environ.*, vol. 97, no. 3, pp. 322–336, 2005.