A Machine Learning Approach for the prediction of solid fuels consumption in Turkey

Çelik N.[1], Konyalioglu A.K.[2]

[1] Gebze Technical University, Department of Mathematics,41400, Kocaeli-Turkey

[2]University of Strathclyde, Strathclyde Business School, 199 Cathedral St, Glasgow G4 0QU, The United Kingdom

*corresponding author:
e-mail: aziz.konyalioglu@strath.ac.uk

Abstract Solid fuels are very crucial energy sources as most of industries use them for obtaining heat, electricity and light. Furthermore, since solid fuels are scarce sources in Turkey, it is very important to forecast the consumption in order to effectively manage the energy policies and to conduct an effective planning for industries. In this study, it is aimed to forecast and to model the produce of solid fuels like lignite and coal in Turkey. In statistical analysis, machine learning techniques are applied for forecasting. There are several types of different machine learning algorithms such as supervised and unsupervised learning, reinforced learning, self-learning, feature learning etc. The methods we used are categorized as supervised learning since they build a mathematical model of a set of data that contains both the inputs and the desired outputs.

Keywords: Solid Fuels, Environmental Data Analysis, Machine Learning

1. Introduction

Solid fuels, including lignite and coals, have a major role to meet the energy demand (Adams and Shachmuvore, 2008). In Turkey, solid fuels are used to generate electric, heating and meet industrial purposes (Akbostancı et al., 2018). Considering these facts, forecasting solid fuels in Turkey is very crucial for policy makers, government and municipalities (Sözen et al., 2007). This forecasting also affects on economic policies related to energy resources and decision makings based on imports and exports (Ediger et al., 2006).

Furthermore, to forecast solid fuels consumption and demand, it is very essential to consider a wide range of factors that influence consumption patterns such as geological, economic, and environmental considerations. These factors may include economic growth, industrial activities, population dynamics, energy policies, technological advancements, environmental regulations, and international energy market dynamics (Melikoğlu, 2017; Aydin, 2015).

In this study, it is aimed to choose the best appropriate model in machine learning and statistical models in forecasting solid fuels including lignite and coal in Turkey. Furthermore, to forecast solid fuels in Turkey, different algorithms will be compared. This comparison will provide an effective algorithm to have an accurate forecasting. Furthermore, the insights gained from accurate forecasts are instrumental in developing strategies that optimize solid fuels utilization, reduce emissions, and foster a sustainable and resilient energy sector in Turkey.

2. Methodology

Machine learning field focuses on using data and algorithms to simulate how humans learn, gradually improving the accuracy of its predictions. In other words, machine learning is a branch of artificial intelligence where the number of rules evolves dynamically with the amount of data, unlike traditional programming techniques, which have a fixed number of rules. In many fields, including autonomous systems, applications for natural language processing, stock market, and financial transactions, the recognition of spam e-mail, medical image processing, voice and speech processing, bioinformatics, and the detection and prediction of natural calamities, machine learning is widely used. A crucial area of study for mathematicians is machine learning, which also necessitates proficiency in programming, calculus, linear algebra, statistics, and probability.

Machine learning is defined as “the study of computer algorithms that improve automatically through experience (Mitchell, 1997). Two terms “artificial intelligence” and “machine learning” are often used mistaken. Machine learning is actually a subset of artificial intelligence which is defined by computer scientists as any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goals (Poole, 1198). At the present time machine learning algorithms are widely used in many areas such as health, business, finance etc. In the theory of machine learning mathematics is used excessively especially the fields statistics and probability theory. Even though statistics and machine learning are related closely their main goal differs. Statistics draws population inferences from a sample, while machine learning finds generalizable predictive patterns (Bzdok et. al, 2018). There are several types of different machine learning algorithms such as supervised and unsupervised learning, reinforced learning, self-learning, feature learning etc. The methods we used are categorized as supervised learning since they build a mathematical model of a set of data that contains both the inputs and the desired outputs.

Regression is a statistical technique used to model and analyze the relationship between a dependent variable and one or more independent variables. It is used to predict the value of the dependent variable based on the values of the independent variables.

k-Nearest neighbor algorithm is an example of instance-based learning sometimes referred as memory-based learning. The main idea of this algorithm is "majority of votes". For example, for each point x to be classified, k-NN asks for class labels to the nearest k data point of x. When k is 1, the label of the nearest neighbor will be assigned to any data point to be classified. In this case the method will be called "nearest neighbor algorithm. There are several different distance metrics used in the K-NN algorithm, and these are Euclidean, Manhattan, Minkowski and Hamming metrics. Minkowski distance between two points X and Y as defined as follows where

$$\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_n}) \in \mathbb{R}$$
$$\mathbf{Y} = (\mathbf{y_1}, \mathbf{y_2}, \dots, \mathbf{y_n}) \in \mathbb{R}$$

$$\mathbf{D}(\mathbf{X}, \mathbf{Y}) = \left(\sum_{\mathbf{i=1}}^{\mathbf{n}} |\mathbf{x_i} - \mathbf{y_i}|^{\mathbf{p}} \right)^{\frac{\mathbf{1}}{\mathbf{p}}} \qquad (\mathbf{1})$$

If $\boldsymbol{p} = \mathbf{2}$, (1) becomes $\boldsymbol{D}(\boldsymbol{X}, \boldsymbol{Y}) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$ which is the Euclidean distance.

It is the data point to be classified as shown in the green dot below. If K is selected as 3, it will be classified as a triangle since the three closest neighbors are two triangles and a square. If K is selected as 5, it is classified as a square since the majority of its five closest neighbors are squares.

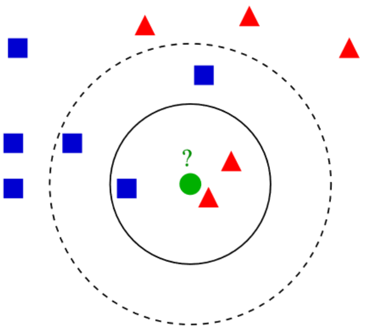

Figure 1. An example of k-NN classification

Support vector machines are another powerful machine learning model used to solve classification and regression problems. It is commonly known as a supervised model but can also work as an unsupervised model in the case of unlabeled data. SVM's motivation is to find the maximum margin hyperplane among the infinite number of hyperplanes of data points. Geometrically, a hyperplane is defined as an (n-1)-dimensional subspace of an ambient space of dimension n. The algorithm starts with selecting a random subset of the original dataset, with replacement (known as bootstrap aggregating or bagging). This subset will be used to train the first decision tree. To make a prediction, the random forest algorithm takes the average of the predictions made by each decision tree.

It is commonly known as a supervised model but can also work as an unsupervised model in the case of unlabeled data. SVM's motivation is to find the maximum margin hyperplane among the infinite number of hyperplanes of data points. Geometrically, a hyperplane is defined as an (n-1)-dimensional subspace of an ambient space of dimension n. In machine learning, a hyperplane with a maximum distance to the closest data points is called the "maximum margin hyperplane" and such points "support vectors". The choice of maximum margin hyperplanes is directly dependent on support vectors.

Equation (2) and (3) are the equations for hyperplanes in 2-dimensions and in p-dimensions respectively.

$$\boldsymbol{\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0} \qquad (2)$$
$$\boldsymbol{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p = 0} \qquad (3)$$

Equation (4) represents a feature vector where $\boldsymbol{X_{1,}X_{2,\dots,}X_p}$ are points on hyperplane.

$$\boldsymbol{X = [X_{1,}X_{2,\dots,}X_p]^T} \qquad (4)$$

Equations (5) and (6) can be used to specify classes when performing a binary classification.

$$\boldsymbol{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p > 0} \qquad (5)$$
$$\boldsymbol{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p < 0} \qquad (6)$$

Figure 2 mainly shows the classification of data by VSM.

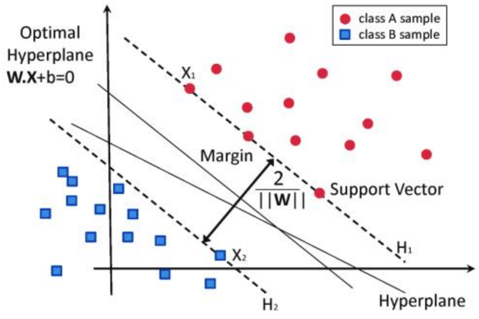

Figure 2. Classification of data by support vector machine (SVM).

3. Application

Lignite is a type of coal that is primarily used for electric power generation. Coal, on the other hand, is a fossil fuel that is widely used for energy production (Ozturk and Ozturk, 2018;Ediger and Akar, 2007). The production of lignite and coal is influenced by a variety of factors, including geological, economic, and environmental considerations. For this reason, in application part, we use five different features in order to obtain the model for the production of coal and lignite. The first feature is energy demand for the households and the firms in Turkey, (X_1 and X_2). The other independent variables for modelling the production of the lignite and the coal are the price of oil because of the transportation cost (X_3), as the price of electricity (X_4) and the salary for the workers (X_5). The data to forecast lignite and coal consumption has been taken from Turkish Statistical Institute (2022).

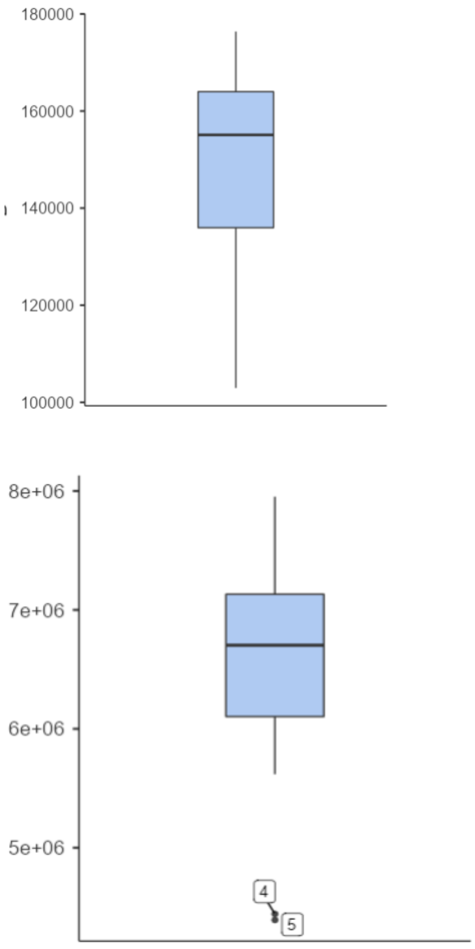

Figure 3. The Box plots of Coal and the Lignite data

Table 1 shows the results of the model accuracies based on R^2 values.

Table 1. The results of R^2 based on different algorithms

Method	Coal	Lignite
Regression	0.86	0.82
k-NN Algorithm	0.88	0.81
Support Vector Machines	0.91	0.84
Random Forest	0.93	0.89

Therefore, the Random Forest Algorithm is the best choice for modelling the production of the coal and the lignite based on R^2 values.

4. Conclusion

Solid fuels are still very important to provide energy for households, industries and municipalities in every country. In turkey, solid fuels still provide energy to maintain daily operations in facilities, household energy to be heated. Thus, an accurate forecasting is essential to plan demand and supply in solid fuels. In this study, we aimed to compare four different methods to forecast coal and lignite consumption in Turkey. We found that random forest has the highest R^2 value for coal and lignite. It means that the most accurate method can be implied as random forest between the used methods.

References

Adams, F. G., & Shachmurove, Y. (2008). Modeling and forecasting energy consumption in China: Implications for Chinese energy demand and imports in 2020. Energy economics, 30(3), 1263-1278.

Akbostancı, E., Tunç, G. İ., & Türüt-Aşık, S. (2018). Drivers of fuel based carbon dioxide emissions: The case of Turkey. Renewable and Sustainable Energy Reviews, 81, 2599-2608.

Aydin, G. (2015). The application of trend analysis for coal demand modeling. Energy Sources, Part B: Economics, Planning, and Policy, 10(2), 183-191.

Bulut Y. and Tez Z. (2007), Adsorption studies on ground shells of hazelnut and almond, *Journal of Hazardous Materials*, **149**, 35-41.

Bzdok D., N. Altman and M. Krzywinski, (2018), "Statistics Versus Machine Learning," Nature Methods.

Ediger, V. Ş., & Akar, S. (2007). ARIMA forecasting of primary energy demand by fuel in Turkey. Energy policy, 35(3), 1701-1708.

Ediger, V. Ş., Akar, S., & Uğurlu, B. (2006). Forecasting production of fossil fuel sources in Turkey using a comparative regression and ARIMA model. Energy Policy, 34(18), 3836-3846.

Lewis-Beck, C., & Lewis-Beck, M. (2015). Applied regression: An introduction (Vol. 22). Sage publications.

Melikoglu, M. (2017). Modelling and forecasting the demand for jet fuel and bio-based jet fuel in Turkey till 2023. Sustainable Energy Technologies and Assessments, 19, 17-23.

Ozturk, S., & Ozturk, F. (2018). Forecasting energy consumption of Turkey by Arima model. Journal of Asian Scientific Research, 8(2), 52-60.

Poole D., A. Mackworth and R. Goebel, (1998), Computational Intelligence: A Logical Approach, New York: Oxford University Press.

Sözen, A., Gülseven, Z., & Arcaklioğlu, E. (2007). Forecasting based on sectoral energy consumption of GHGs in Turkey and mitigation policies. Energy policy, 35(12), 6491-6505.

T. Mitchell, (1997), Machine Learning, McGraw Hill.

Yildiz, B., Bilbao, J. I., & Sproul, A. B. (2017). A review and analysis of regression and machine learning models on commercial building electricity load forecasting. Renewable and Sustainable Energy Reviews, 73, 1104-1122.

Zouhal, L. M., & Denoeux, T. (1998). An evidence-theoretic k-NN rule with parameter optimization. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 28(2), 263-271.