Predicting Coal Quality Using Decision Tree Algorithm

Reza Naquib Faishal[1], Kiagus Muhammad Arsyad[2], Ariana Yunita[3]

{sholsholfaishol@gmail.com[1] , muhammadarsyadv1@gmail.com[2],
ariana.yunita@universitaspertamina.ac.id[3]}

Department of Geophysical Engineering, Universitas Pertamina, Jakarta, Indonesia[1]
Department of Computer Science, Universitas Pertamina, Jakarta, Indonesia[2,3]

Abstract. Predictions are often criticized for the lack of interpretability, which is often in many real-world applications. High-quality coal is required to meet industrial demands, increase national energy security, and export. This research aims to show that the Decision Tree algorithm can classify coal quality based on volatile matter, fixed carbon, and heating values. The dataset used in this study is synthetic data generated based on the ASTM (America Society for Testing and Materials) rankings. The model's accuracy for predicting coal quality is 96 percent, and the tree has a depth of 5. This study demonstrates how decision tree algorithms produce reasonable predictions.

Keywords: Coal, Classification, Decision tree, Machine learning, Prediction

1 Introduction

Predictions are often criticized for their need for more interpretability in many real-world applications. For example, deep learning, an algorithm that gives high accuracy, is such a black box algorithm that it is hard to interpret [1], [2]. However, several predictions urge us to understand the interpretation, such as in health, for predicting heart disease [3], and in education, for assessing scholarship [4].

Coal is a fossil fuel or sedimentary rock that can be burned, formed from organic deposits, primarily plant remains, and created through coalification [5]. Coal plays an important role in national development. Coal contributes significantly to satisfying energy needs, increases state revenue from exports, and coal-derived products form new business development and job creation. The demand for high-quality coal is rising in parallel with the expansion of the domestic industry. Therefore, classifying coal quality is an important issue for nations.

Several factors influence coal quality, including moisture, ash, volatile matter, and fixed carbon. A proximate analysis is required to classify the quality of coal. Proximate analysis is an initial method for determining coal quality, including moisture, ash, volatile matter, and fixed carbon content [6]. The lower the calorific value of coal, the higher the moisture and ash content. Meanwhile, the higher the volatile matter, the more coal can spontaneously burn [5], [6].

According to [7], coal is classified into four types based on its quality. Lignite is the lowest-quality type of coal. This is because lignite is young coal. There is a Subbituminous layer above

ICO-SEID 2022, November 23-24, Jakarta, Indonesia

DOI 10.4108/eai.23-11-2022.2338880

lignite. Bituminous exists above the Subbituminous. Anthracite coal is the highest quality coal. Anthracite is a type of hard coal that is shiny, solid black and has the highest carbon content, producing significantly more energy than other types of coal.

The quality of coal must be analyzed with precision so that the identification of the type of coal can be more accurate. Because of the flexibility of the use of coal and because each type of coal has its benefits, precise identification of coal types is necessary so that coal can be used appropriately. Not only accuracy in determining the kind of coal quality but also speed is needed so that the mining process is faster, more productive, and more profitable for coal mining companies.

The decision tree algorithm, one of the simplest classification methods in machine learning, can be used to optimize the classification of coal quality types. The decision tree is a powerful and well-known classification and prediction method. The decision tree method converts much information into a decision tree representing the rules [8].

While another study attempts to determine coal quality using the K-Means algorithm for clustering [9], this study aims to classify coal quality using the Decision Tree algorithm. The paper is organized as follows. Section two will discuss the theoretical foundations. Section three will explain the methods. Section four will show the results and discussion. The final section will present the conclusion and future research.

2 Theoretical Foundations

This section shows the theoretical foundations of the Decision Tree algorithm, the evaluation, and the Proximate Analysis of Coal Quality.

2.1 Decision Tree

The Decision Tree algorithm is one of the tree-based algorithms often used for the classification problem, but it may also apply to regression tasks. Many studies employ this algorithm for classifying and predicting. Several previous works use the Decision Tree algorithm, such as Prakoso et al. [8] employ the Decision Tree algorithm for gold content calculation. They conducted using the RapidMiner tool, which resulted in high accuracy. Another study employs the Decision Tree algorithm, but they use Weka to predict heart disease [3]. The accuracy is not as high as Prakoso et al., and they only result in 76,66 accuracy. Wang et al. [4] evaluate scholarship using C45 Decision Tree and result with 91% accuracy. Because of the tendency of resulting high accuracy from the three related studies, we employ the Decision Tree algorithm for classifying quality.

Other studies that related with coal are Zeng et al. [10] and Pekel et al. [11]. Zeng et al. employ decision tree for analyzing sustainability of Chinese coal cities. They visualize the tree and explain the rule. Pekel et al. predict low-rank coal moisture using Decision Tree Regressor. This study visualize the tree rule for their study case. Another reason is that Decision Tree is easy to interpret by humans so this study might reveal the tree visualization of coal quality.

The decision tree method uses a flowchart structure that resembles a tree structure. The selection of attributes in the decision tree is based on the highest gain value of all gain values in each attribute. To calculate the gain, the following formula is used [8].

$$\text{Gain (S, A)} = \text{Entropy (S)} - \sum_{i=1}^{n} |S_i|/S \times \text{Entropy } (S_i) \quad (1)$$

S : Case set

A : Attribute

n : Partitions number of attribute A

$|S_i|$: Number of cases on the i-th partition

S : Number of cases in S

To calculate the gain value, the entropy value must first be found out. The following formula is used to calculate the entropy value [8].

$$\text{Entropy (S)} = \sum_{i=1}^{n} - p_i \, log_2 \, p_i \quad (2)$$

S : Case set

n : Number of S partitions

p_i : Proportion of S_i to S

2.2 Evaluation of Decision Tree Algorithm

The confusion matrix is a tool for predictive analysis in machine learning including a decision tree algorithm and is used to assess the performance of a classification model. According to Karimi [12], a confusion matrix is a table that contains four different combinations of predicted and actual values. In the confusion matrix, four terms represent the classification process results: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

Karimi et al. [12] explained about TP, TN, FP, and FN, so that accuracy, precision, recall, and F-1 score could be calculated. Accuracy measures how many correct predictions our model made across the entire test dataset using the formula:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (3)$$

Precision informs us that the number of correctly predicted cases was positive. This would decide whether our model can be trusted or not. The precision formula is:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{TN}} \quad (4)$$

Recall indicates how many of the actual positive cases we could correctly predict with our model. A higher recall indicates that most positive cases (TP + FN) will be labeled as positive (TP). This will result in a more significant number of FP measurements and a lower overall accuracy. The recall formula is as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

F1-Score is a harmonic mean of precision and recall, providing a comprehensive picture of these two metrics. It is greatest when precision equals recall. The precision formula is as follows:

$$\text{F1-Score} = \frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} \tag{6}$$

2.3 Proximate Analysis of Coal Quality

One of the first analyses performed on coal after mining is proximate analysis. Coal quality in this study is based on ASTM (America Society for Testing and Materials) rankings. This classification is based on the degree of metamorphism or alteration that occurred during the coalification process (ranging from Lignite to Anthracite) [7]. The data required in this classification are fixed carbon, volatile matter, and heating values.

According to Nur et al.[7], fixed carbon is the carbon that remains after the determination of flying substances. The value of fixed carbon is calculated using the following formula:

$$\text{FC} = 100 - \text{M} - \text{A} - \text{VM} \tag{7}$$

FC : Weight % Fixed carbon

M : Weiht % Moisture

A : Weight % Ash

VM : Weight % Volatile matter

Nur et al. [7] also explained that volatile matter is a value that represents the percentage of flying substances found in coal. The formula used to calculate the value of volatile matter is:

$$\text{VM} = (\text{BM} + \text{EM}) - \text{M} \tag{8}$$

M : Moisture (%)

VM : Volatile matter (%)

BM : Burned Materials (%)

EM : Evaporated Materials (%)

3 Methods

This study aims to implement the Decision Tree algorithm for the coal quality dataset to show the algorithm's interpretability. General machine learning workflow [13] was applied for this study. We use a coal classification dataset based on ASTM. Features of the dataset are fixed carbon, volatile matter, and heating values to determine coal ranking. This classification attempts to separate coal into four types: lignite, subbituminous, bituminous, and anthracite. Higher-ranking coal is generally more complex, contains more carbon, has lower humidity levels, and produces more energy with high heating values. Anthracite is the highest rank and best coal quality type because its formation time is the longest and requires extremely high temperatures and pressure to form.

Lignite, subbituminous, bituminous, and anthracite contain fixed carbon, volatile matter, and heating values with different amounts. Anthracite contains around 86–98 percent fixed carbon levels, volatile matter 2–15 percent, and 7740–8300 Kcal/kg. Bituminous contains around 54–86 percent fixed carbon levels, 14–54 percent volatile matter, and 6765–8741 Kcal/kg. Subbituminous contains fixed carbon levels of about 53 – 55 percent, volatile matter of 53 – 55 percent, and 5990 – 7540 Kcal / kg. Lignite contains fixed carbon levels of around <53 percent, volatile matter <53 percent, and 5250 – 6360 Kcal/kg [7]. The synthetic data is based on all ranges of those values.

Table 1. ASTM Specification for Coal Quality Classification. Modified from [7]

Class	Fixed Carbon (%)	Volatile Matter (%)	Heating Values (Kcal/kg)
Anthracite	86-98	2-15	7740-8300
Bituminous	54-86	14-54	6765-8741
Subbituminous	53-55	53-55	5990-7540
Lignite	<53	<53	5250-6360

We used synthetic data related to coal quality which comprises three features, namely Fixed Carbon (percent), Volatile Matter (percent), and Heating Values (Kcal/kg). The selection of such features is based on the production of coal ranking following ASTM standards. All data in the decision tree must be numerical. As a result, we convert columns containing non-numeric data, such as Class, into numeric data. Number 4 was converted from anthracite, Bituminous was altered to 3, Subbituminous was converted to 2, and Lignite was converted to 1.

	Fixed Carbon (%)	Volatile Matter (%)	Heating Values (Kcal/kg)	Class
0	60	30	8200	Bituminous
1	86	7	7900	Anthracite
2	54	21	8100	Bituminous
3	56	55	7000	Subbituminous
4	52	52	5250	Lignite
...	...	...	...	...
95	54	53	6700	Subbituminous
96	56	55	6820	Subbituminous
97	88	25	8260	Bituminous
98	90	10	8300	Anthracite
99	88	18	8600	Bituminous

100 rows × 4 columns

Fig. 1. Snapshot of Dataset

The dataset contains four columns, including the feature and target columns. Fixed Carbon (percent), Volatile Matter (percent), and Heating Values (Kcal/kg) are all included in the feature column. Meanwhile, Class is included in the target column. This snapshot of dataset can be seen in **Figure 1**.

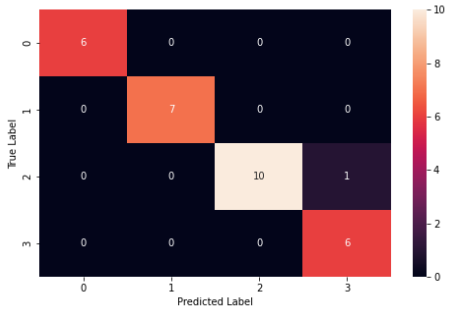
This study uses Python as a programming language, and Sci-kit Learn as a library. For training and testing, the data is then divided into training and testing sets using 'train test split' algorithm in scikit-learn. This study divides training and testing data into 70% and 30%.

4 Results and Analysis

In general, this study results in the accuracy of the predictive model and tree visualization.

To check the performance of the classification, the confusion matrix was used. By using the confusion matrix, we can find the number of correct and incorrect predictions so that it can be evaluation material for us. The confusion matrix can be seen in **Figure 2a**. It can be seen that the comparability between True Label on the Y axis and Predicted Label on the X axis is very high. Number of True Positive for Class 0, Class 1, Class 2 and Class 3 are 6, 7, 10, and 6 respectively. Furthermore, number of false negative for class 2 is 1, which means the true label is Class 2, but it is predicted as Class 3.

We calculate the accuracy, precision, recall, and F1-Score to evaluate the results of the confusion matrix. **Figure 2b**. depicts that the accuracy is 97%.



(a)

	precision	recall	f1-score	support
1	1.00	1.00	1.00	6
2	1.00	1.00	1.00	7
3	1.00	0.91	0.95	11
4	0.86	1.00	0.92	6
accuracy			0.97	30
macro avg	0.96	0.98	0.97	30
weighted avg	0.97	0.97	0.97	30

(b)

Fig. 2. (a) Confusion Matrix (b) Confusion matrix evaluation

We then perform a visualization with such high precision. The purpose of this visualization is to determine the relationship between the three feature columns (Fixed Carbon (percent), Volatile Matter (percent), and Heating Values (Kcal / kg)) and the target column, Class. Based on the tree visualization, we can decide or identify coal types, whether Lignite, Subbituminous, Bituminous, or Anthracite.

As seen in **Figure 2**, Volatile Matter (percentage) = 14.5 indicates that any coal with a Volatile Matter of 14.5 percent or less will follow the left arrow (true), while the rest will follow the right arrow (false). Gini is a method that is used in decision tree algorithms. Gini is the simplest method for performing binary splitting in a decision tree. Samples = 100 indicates that the total data is 100, while Value = [16,23,40,21] demonstrates that 16 samples are Lignite, 23 are Subbituminous, 40 are Bituminous, and 21 are Anthracite.

At depth 1, parent nodes generate two branch nodes. The left branch nodes have gini = 0, indicating that all samples produce the same result. Samples = 20 indicates that there are 20 types of coal left in this branch node. Value = [0,0,0,20] demonstrates that the existing 20 samples are of the Anthracite type. The right branch node's fixed carbon (percent) is 56.5, which means that any coal with a fixed carbon of 56.5 percent or less will follow the left arrow, and the rest will follow the right arrow. With a Gini impurity worth 0.627 and a total of 80 samples, the value obtained is 16 samples of Lignite, 23 samples of Subbituminous, 40 samples of Bituminous, and 1 sample is Anthracite.

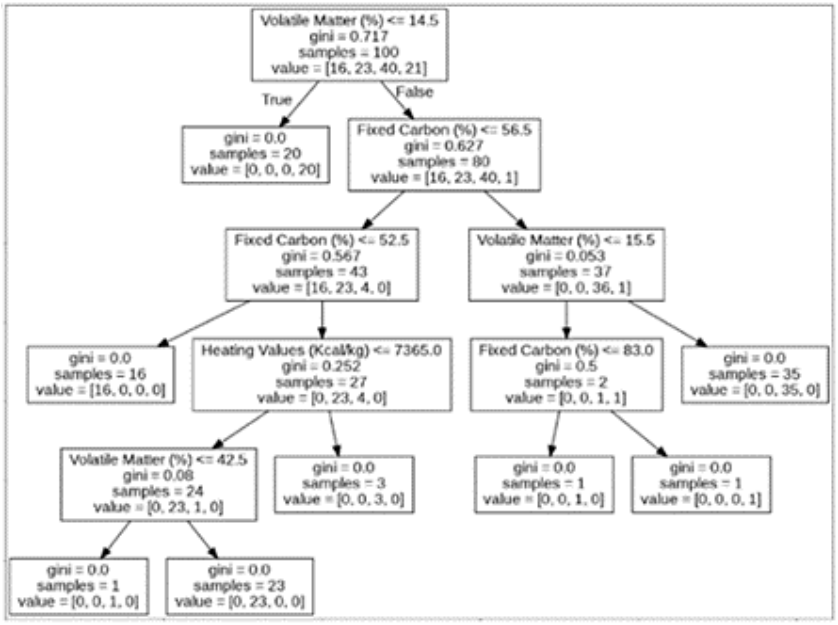


Fig. 3. Decision Tree for Classifying Coal Quality

The predictive model for this study is a decision tree with a depth of 5. **Figure 3**. shows that if data has volatile matter more or equal to 14.5, then it was classified as class 4 (Anthracite). It can be seen that Class 4 was correctly classified for all 20 points. Furthermore, if the tree was traversed to another leaf node, Volatile Matter less than 14.5, and fixed carbon less than or equal to 52.5, then it was classified as Class 1 (Lignite). It is important to note that the Decision Tree algorithm tends to overfit. Hence another tree-based approach is taken into consideration for application, such as XGBoost, and Random Forest.

With such information, we hope to obtain decision results in the form of more detailed and accurate identification of coal types. This is also what motivates us to use decision tree algorithms. We value the accuracy of coal-type data because the kind of coal determines the actual quality of coal. The quality of coal will determine the price of coal, which will significantly impact the company's profits. Furthermore, the speed of coal quality identification in decision trees can be obtained, allowing the process of exploration and mining exploitation to be accelerated, allowing the company to achieve production and profit targets.

5 Conclusion

A number of conclusions can be drawn from the study we have conducted. First, the decision tree algorithm can accurately classify coal quality types. The accuracy rate we obtained in our study, which used a dataset with 100 rows and four columns, was 97 percent. Second, the decision tree algorithm can find random and complex number combinations from columns (Fixed Carbon (percent), Volatile Matter (percent), and Heating Values (Kcal/kg). The decision tree is so easy to interpret to show how the coal quality is formed, including Lignite, Subbituminous, Bituminous, and Anthracite, based on the combination of these numbers.

For subsequent studies, we suggest using data from proximate analysis and adding data from ultimate analysis. The ultimate analysis is carried out to determine the content of chemical elements in coal, such as carbon, hydrogen, oxygen, nitrogen, sulphur, auxiliary elements, and rare earth elements. With the combination of data from proximate analysis and ultimate analysis, it is hoped that the results of predictions and classifications of coal quality types can be more accurate using a decision tree.

References

[1] L. Deng and D. Yu, "Deep learning: Methods and applications," *Foundations and Trends in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2013, doi: 10.1561/2000000039.

[2] A. Yunita, H. B. Santoso, and Z. A. Hasibuan, "Deep learning for predicting students' academic performance," in *Proceedings of 2019 4th International Conference on Informatics and Computing, ICIC 2019*, 2019, pp. 1–6. doi: 10.1109/ICIC47613.2019.8985721.

[3] S. Maji and S. Arora, "Decision Tree Algorithms for Prediction of Heart Disease," in *Information and Communication Technology for Competitive Strategies*, S. Fong, S. Akashe, and P. N. Mahalle, Eds., Singapore: Springer Singapore, 2019, pp. 447–454.

[4] X. Wang, C. Zhou, and X. Xu, "Application of C4.5 decision tree for scholarship evaluations," *Procedia Computer Science*, vol. 151, pp. 179–184, 2019, doi: https://doi.org/10.1016/j.procs.2019.04.027.

[5] S. F. Greb, C. F. Eble, and J. C. Hower, "Coal," in *Encyclopedia of Geochemistry: A Comprehensive Reference Source on the Chemistry of the Earth*, W. M. White, Ed., Cham: Springer International Publishing, 2017, pp. 1–16. doi: 10.1007/978-3-319-39193-9_153-1.

[6] T. A. Oratmangun, S. H. Yuwanto, and L. Utamakno, "ANALISIS PROKSIMAT DALAM PENENTUAN KUALITAS DAN JENIS BATUBARA PADA PT. BUMI MERAPI ENERGI, KABUPATEN LAHAT, PROVINSI SUMATRA SELATAN," *Jurnal Sumberdaya Bumi Berkelanjutan (SEMITAN)*, vol. 3, no. 1, pp. 56–59, 2021.

[7] Z. Nur, M. Oktavia, and Desmawita, "Analisis Kualitas Batubara Di Pit Dan Stockpile Dengan Metoda Analisis Proksimat Di Pt. Surya Anugrah Sejahtera Kecamatan Rantau Pandan Kabupaten Bungo Provinsi Jambi," *Mine Magazine*, vol. 7, no. 1, p. 283, 2019.

[8] B. S. Prakoso and G. D. Sutanto, "PENERAPAN METODE DECISION TREE DAN NAÏVE BAYES UNTUK MENGHITUNG KADAR KARAT EMAS," *ikraith-informatika*, vol. 3, no. 2, pp. 27–32, 2019.

[9] M. Bakri, "PENERAPAN DATA MINING UNTUK CLUSTERING KUALITAS BATU BARA DALAM PROSES PEMBAKARAN DI PLTU SEBALANG MENGGUNAKAN METODE K-MEANS," *Jurnal TEKNOINFO*, vol. 11, no. 1, p. 10, 2017.

[10] L. Zeng, J. Guo, B. Wang, J. Lv, and Q. Wang, "Analyzing sustainability of Chinese coal cities using a decision tree modeling approach," *Resources Policy*, vol. 64, p. 101501, Dec. 2019, doi: 10.1016/j.resourpol.2019.101501.

[11] E. Pekel, M. C. Akkoyunlu, M. T. Akkoyunlu, and S. Pusat, "Decision tree regression model to predict low-rank coal moisture content during convective drying process," *International Journal of Coal Preparation and Utilization*, vol. 40, no. 8, pp. 505–512, Aug. 2020, doi: 10.1080/19392699.2020.1737527.

[12] F. Karimi, S. Sultana, A. S. Babakan, and S. Suthaharan, "An enhanced support vector machine model for urban expansion prediction," *Computers, Environment and Urban Systems*, vol. 75, pp. 61–75, 2019.

[13] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc.," 2022.