Coal Characterization of South Sumatera Basin using the Unsupervised Machine Learning Method

E Ayustyana¹, S A Wibisono², F M H Sihombing¹

¹Geology Study Program, Faculty of Mathematics and Natural Science (MIPA), Depok, West Java, 16424, Indonesia
²Pusat Sumber Daya Mineral Batu Bara dan Panas Bumi (PSDMBP), Soekarno-Hatta, Bandung, West Java, 40254, Indonesia

*Corresponding Author : felix.sihombing@ui.ac.id

Abstract. South Sumatra Basin is one of the sedimentary basins in Indonesia which has coal bearing formations. Numerous coal proximate data (free moisture, total moisture, moisture, inherent moisture, ash content, fixed carbon, sulfur content) and ultimate data (carbon (C), hydrogen (H), oxygen (O), nitrogen (N)) have been collected in this basin. This study aims to determine the characterization of coal in the South Sumatra Basin based on the proximate data and ultimate data using the unsupervised machine learning methods. The machine learning method has several basic concepts, namely being able to predict data by studying several patterns and factors that have been trained in a short amount of time. The study able to cluster coal in the basin into two cluster of coals with striking difference. The distribution of the two coal clusters in the South Sumatra Basin possibly influenced by the age of the formation in the South Sumatra Basin. In the first cluster, it is distributed in the older Airbenakat Formation and Muaraenim Formation, while in the second cluster it is scattered in the younger Muaraenim Formation and the Kasai Formation. The formation ages of the youngest are the Kasai Formation, the Muaraenim Formation, and the Airbenakat Formation.

Keywords : coal, South Sumatera Basin, Machine Learning, Muaraenim Formation, Airbenakat Formation

1. Introduction

Coal is a source of energy that comes from organic sediment. Organic sediment in the formation of coal comes from plant remains which undergo a process of decay. Most of Indonesia's coal is generally used as a source of energy for electricity generation. Based on data from the Center for Mineral, Coal and Geothermal Resources in 2016, the formation of Indonesian coal carriers are mostly located on Sumatra and Kalimantan island.

Coal bearing formations are rock formations that have coal layers, to be able to find these formations requires geological investigation and mapping activities in the field. Coal-bearing formations are generally located in sedimentary basins. The South Sumatra Basin is one of the sedimentary basins in Indonesia which has coal-bearing formations. The exploration and production activities of coal in the South Sumatra Basin have been widely carried out. However, it is still not well categorized. Therefore, coal characterizations of South Sumatra Basin are grouped using machine learning methods.

Machine learning is part computer science, which provides something without the need to be explicitly programmed. This machine learning requires only a few definitive parameters with one or more variable features. This data or variable is needed as material for further analysis so that the machine can produce results (Samuel,
The machine learning method has several basic concepts, namely being able to predict data by studying several patterns and factors that have been trained in predicting data in a short amount of time.

2. Geological Overview

The South Sumatra Basin was formed due to the submergence of the Indo-Australian Plate moving from north to the northeast towards the immobile Eurasian plate. In the South Sumatra Basin, there is a plate subduction zone covering the western area of Sumatra Island and southern Java Island. As a result of the submergence of the Indo-Australian plate, it may affect the condition of rocks, morphology, tectonics, and structures in South Sumatra. There are several boundaries around the South Sumatra Basin, namely the southeast are bounded by the Sunda Basin and the Twelve Mountains, the southwest is bordered by the Bengkulu Basin, Bukit Barisan, and the Semangko Fault, the north is bordered by the Thirty Mountains and the Central Sumatra Basin, and the east is bordered by the Sunda Shelf.

The stratigraphy of the South Sumatra Basin is referred to as a large cycle consisting of transgression and regression. The formation in this basin is divided into two large formation groups, namely the Telisa group, it is when the transgression phase occurs and the Palembang group, namely when the regression phase occurs, which includes the Telisa group namely the Talang Akar Formation, the Baturaja Formation, and the Gumai Formation. Meanwhile, those included in the Palembang group are the Airbenakat Formation, the Muaraenim Formation, and the Kasai Formation. The stratigraphy of the South Sumatra Basin can be seen in Figure 1.

![Stratigraphy of the South Sumatra Basin](image)

Figure 1. Stratigraphy of the South Sumatra Basin (De Coster, 1974)

One of the main coal-bearing formations in the South Sumatra Basin is the Muaraenim Formation (Fatimah, 2015). The Muaraenim Formation is formed in a regression phase, from shallow seas (Airbenakat Formation) to terrestrial environments (Kasai Formation).

3. Methods

This research uses data processing as the first stage method. At this stage, dataset preparations are carried out to prepare some of the data that will be needed on the next stage. The datas needed in this research are proximate data and ultimate data. The next stage is pre-processing data. In this stage, the method used is selecting features from existing data and normalizing the data. Feature selection is used to see how influential one data is with other
data. Data normalization has a function to equalize data with the same scale on all data, that is, the original data will be converted into a scale of 0 - 1. In the pre-processing stage, the PCA (Principal Component Analysis) method is also carried out. PCA has a function to transform data into smaller dimensions.

The next stage in pre-processing is to carry out the feature selection process. Features selection process is used to see the effect of data from one feature to another. The next step is to enter all the libraries that will be used in processing data. Next, put the dataset that will be used to carry out unsupervised machine learning activities. After entering the dataset on Jupyter Notebook, the next step is to normalize the data because each data has a different range of values. So with normalization, all data will have the same range, namely 0 - 1. After the data is normalized, the next step is to do principal component analysis (PCA) to reduce the dimensions of the proximate data and ultimate data.

The next stage is, determining the cluster. The determination of the best cluster is done by using the elbow and k-means methods. The elbow method is a method that sees the best number of clusters by looking at the percentage value of the comparison between the number of clusters by forming an elbow at a point. The K-Means method has a function to create clusters or groups of data that have been determined by the elbow graph. In determining this cluster, the first step is the Elbow method. The graph in this elbow method serves to determine the best number of clusters.

The last stage is to determine the number of K (cluster) in the clustering process using K-Means. At this stage, the authors tried several clusters as a consideration, namely two clusters, three clusters, and four clusters. The author's consideration is also carried out on the distribution of samples in the South Sumatra Basin. The author plots the map to make it easier to visualize the predefined clusters in machine learning. Map plotting also helps and makes geological interpretation easier. Map plotting using the Arc GIS application.

4. Results

The datas used in this research are coordinate, proximate, and ultimate data. The proximate data used are those that are closely related to moisture content. Data on free moisture, total moisture, moisture, inherent moisture, ash content, fixed carbon, and sulfur content. Ultimate data is closely related to the content of elements contained such as carbon (C), hydrogen (H), oxygen (O), and nitrogen (N). This ultimate data has a different range from one another. The data used are 656 data. Based on the results issued by machine learning, the results analyzed are two clusters, three clusters, and four clusters.

This two-cluster analysis is using descriptive statistics and geological conditions. The data used are total moisture, ash content, volatile matter (VM), fixed carbon, sulfur content, carbon (C), hydrogen (H), oxygen (O), and nitrogen (N). Figure 2 shows the results of a box and whisker diagram for a two-cluster analysis.
Apart from a descriptive statistical analysis, the analysis was also carried out with a spatial density approach. The authors plotted the clustering results into a map of the South Sumatra Basin using the Arc GIS application. Based on the results of the plot, it can be seen in Figure 3 which resulted the distribution of samples between cluster zero and cluster one. The results of the sample distribution are as follows.

Figure 3. Two clusters in the South Sumatra Basin
This three-cluster analysis is using descriptive statistics and geological conditions. The datas used are total moisture, ash content, volatile matter (VM), fixed carbon, sulfur content, carbon (C), hydrogen (H), oxygen (O), and nitrogen (N). Figure 4 shows the results of a box and whisker diagram for a three-cluster analysis.

Figure 4. Three cluster box and whisker diagram

Apart from the descriptive statistical analysis, the analysis was also carried out with a spatial density approach. The authors plotted the clustering results into a map of the South Sumatra Basin using the Arc GIS application. Based on the results of the plot, it can be seen in Figure 5 that the sample distribution between cluster zero, cluster one, and cluster two. The results of the sample distribution are as follows.

Figure 5. Three clusters in the South Sumatra Basin
This four-cluster analysis is using descriptive statistics and geological conditions. The datas used are total moisture, ash content, volatile matter (VM), fixed carbon, sulfur content, carbon (C), hydrogen (H), oxygen (O), and nitrogen (N). Figure 6 shows the results of a box and whisker diagram on a four-cluster analysis.

\[ \text{Figure 6. Four cluster box and whisker diagram} \]

Apart from the descriptive statistical analysis, the analysis was also carried out with a spatial density approach. The authors plotted the clustering results into a map of the South Sumatra Basin using the Arc GIS application. Based on the results, it can be seen in Figure 7 that the sample distribution between cluster zero, cluster one, cluster two, and cluster three. The results of the sample distribution are as follows.

\[ \text{Figure 7. Four clusters in the South Sumatra Basin} \]
5. Discussion

Cluster determination is determined from descriptive statistical data and geological conditions that have been analyzed. Based on this data, the best cluster is two clusters. The two cluster has a significant difference between cluster zero and cluster one. Based on the box and whisker diagrams, the total humidity value (TM) in cluster zero has a high enough moisture value compared to cluster one. At the ash content, the zero cluster has a fairly low ash value compared to cluster one. Based on the value of volatile matter and carbon tethered to cluster zero, it has a higher value than cluster one.

The ultimate analysis is a carbon (C), hydrogen (H), oxygen (O), and nitrogen (N). The values of C, H, O are influenced by the rank of coal. Based on the results of the two clusters in the zero cluster shows the high carbon, low hydrogen, and low oxygen values, so it can be concluded that cluster zero has an older age when compared to cluster one. Based on the distribution of points on the geological map, the two clusters have sufficient results according to the age of the formation. In the zero clusters, the distribution occurs in the Airbenakat Formation and the Mauarenim Formation, while the one cluster is in the Mauarenim Formation and the Kasai Formation.

The reason for not using three clusters is because, cluster zero and cluster two has almost the same range of values from ash content, volatile matter, tethered carbon, sulfur content, carbon elements, hydrogen elements, oxygen elements, and nitrogen elements. Based on the ultimate data analysis when juxtaposed with geological maps, the three-cluster analysis has a formation distribution that does not match the results of the box and whisker diagram plots on carbon (C), hydrogen (H), and oxygen (O) data.

The reason for not using four clusters is that the proximate data and ultimate data on the box and whisker diagram, and based on the distribution on the geological map, have a very random data distribution so that it is difficult to distinguish. Between clusters have a fairly wide distribution of formations and do not have a special pattern, so that it is difficult to distinguish between clusters from one another.

6. Conclusion

The character of coal in the South Sumatra Basin is based on the results issued by machine learning, namely having two clusters. The analysis was carried out in determining the two clusters, namely based on proximate data and ultimate data. Cluster zero has coal with a high total moisture content compared to cluster one, for ash content in cluster zero it is quite low compared to cluster one, for volatile matter and fixed carbon values in cluster zero are higher than with cluster one. Based on the ultimate data analysis, the results of the two clusters in the zero cluster have high carbon, low hydrogen, and low oxygen values, it can be concluded that cluster zero has an older age when compared to cluster one.

The distribution of the two coal clusters in the South Sumatra Basin is influenced by the age of the formation in the South Sumatra Basin. In cluster zero, it is scattered in the Airbenakat and Mauarenim Formations, while in the first cluster it is scattered in the Mauarenim Formation and the Kasai Formation. The formation ages of the youngest are the Kasai Formation, the Mauarenim Formation, and the Airbenakat Formation.

7. Acknowledgements

PUTI Grant No. NKB-969/UN2.RST/HKP.05.00/2020

References

[1] Barber, A., J., Crow M., J., dan Milsom, J., S. (2005). *Sumatra Geology Resources and Tectonics Evolution*, Geological Society Memoir. 31. The Geological Society, London.

[2] Bishop, M. G. (2001). *South Sumatra Basin Province, Indonesia: The Lahat/Talang Akar-Cenozoic Total Petroleum System*. Open File Report 99-50-S USGS. Colorado.
[3] De Coster, G.L. (1974). *The Geology of The Central and South Sumatra Basin*. Proceedings Indonesian Petroleum Association 3rd Annual Convention hlm. 70-110. Jakarta: IPA.

[4] Fatimah, Suryanah A., Suhada I. D., Hidayat R. (2015). *Deep Seated Coal Batubara Bawah Permukaan*. Pusat Sumber Daya Geologi – Badan Geologi. Bandung.

[5] Gafoer, S., Cobrie, T., Purnomo, T. (1986). *Geologi Lembar Lahat, Sumatera*, Puslitbang Geologi. Bandung.

[6] Samuel, Arthur., (1959). *Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed*. 