**Renewable Energy Usage Clustering Using K-Means**

**Maesa Ken Neobi Arief1, Matthew William Siregar1**

1 Computer Science Department, School of Computer Science, Bina Nusantara University, Tangerang, Indonesia

**Abstract**

Renewable Energy is a method to shrink reliability on the use of fossil fuels and to prevent the problems that arise with its dependency. This study investigates the case of countries with varieties of Renewable Energy usage using the application of K-Means clustering. The study highlights the potential implementation of unsupervised machine learning in future research and studies., supporting the act on becoming less dependant to fossil fuel energy.

***Keywords.*** *Clustering, Renewable Energy, K-Means, Machine Learning.*

1. **Introduction**

Renewable Energy (RE) is an energy source that replenishes naturally without being depleted from the earth. This has been a constantly discussed topic among researchers and has led to some innovative solutions on how to preserve and create new sources of renewable energy. In the earth’s ever-growing population, the need for fossil fuel-based energy resources has skyrocketed. As of which, led to multiple problems and challenges such as depletion of fossil fuel reserves, environmental concerns, geopolitical conflicts, and fuel price inflations [1].

In recent times, RE has been a hope to counter those problems that are caused by fossil fuel-based energy. RE has started to address the common issues such as fuel depletion and environmental concerns caused by global warming. The use of RE has been implemented in various sectors. Solar energy has been used for powering road lights by cities around the world. RE has also started to become useful in the domestic context. In several countries, solar energy has been used for heating water supplies in the houses’ water heaters [2].

To further implement RE, researchers have begun further study into the potential implementations of RE, one of which is the implementation of machine learning. A research conducted in 2020 by Shapi et al., implemented algorithms such as K-Nearest Neighbour (KNN), Support Vector Machines (SVM), and Artificial Neural Networks (ANN). The research used Microsoft Azure Machine Learning Studio (AzureML) using R programming language [3]. However, the limitation to this study is the compartment used in the modeling stage.

Another research conducted in 2023 by Benti et al. forecasted an increase of machine learning and deep learning innovations implementing generation of RE. The research showed current advances and what to expect in the next decade. The algorithms reviewed in the research were mostly used to forecast RE and some were also used to give predictions about the amount of RE used in a dataset [4].

The aim of this research is to build an unsupervised learning model that creates clusters of countries who are similar in RE generation. Once the clusters have been found, more research deeper into the topic can be conducted. This research uses the K-Means algorithm to create clustering in the midst of the dataset that we gained from kaggle.

1. **Literature Review**
2. **Machine learning-based energy consumption clustering and forecasting**

A similar research paper in 2020 by Culaba et al., created a clustering model using K-Means to characterize energy systems in the building sector that accounts for a significant portion in the consumption of global energy. The research gained new findings such as the descriptive understanding of the energy consumption data. The study addresses the clustering method in the context of energy consumption using Machine Learning (ML) algorithms. However this study has some limitations. The research can only be used for datasets in buildings within the same climate and time. Other than that, it is a small dataset with only 30 rows of data of buildings from Miami, Florida. And as a result, the model lacked applicability in other contexts of climates [5].

1. **Renewable Energy Sources Integration via Machine Learning modelling: a Systematic Literature Review**

A Systematic Literature Review (SLR) paper in 2024 by Alazemi et al., reviewed and highlighted different ML models implemented to research deeper about renewable energy. The research noted that solar and wind are the most implemented items of data for the ML models across the distribution grid. The results of the SLR concluded that different ML approaches such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), and other models have successfully been implemented to predict PV and wind power outputs. Even ensemble learning has successfully been used to enhance forecasting conditions within the high varying weather conditions. Simpler models like SVM have also proven to be an effective generalization algorithm even in conditions where the data is small. However, the scope of the SLR has not discovered ML implementation in forecasting energy price, demand, and load. A deeper analysis and discovery is needed to portray the true nature of ML models in RES [6].

1. **Renewable Energy Prediction through Machine Learning Algorithms**

A research paper in 2020 by Alvarez et al., implemented efficient RE selection and chose a specific geographic location of Aguascalientes, Mexico using different types of ML algorithms. The algorithms used were Multi-Layer Perceptron (MLP), SVM, Random Forest (RF), and Linear Regression (LR). The research compared four different classification algorithms using the same geographic location and dataset and attempted to forecast solar iridescence and relative wind energy to predict solar and wind production respectively. The study is reliable, but it is limited to one specific geographical location. Further research and study using the same method is needed in a broader and more global context. In addition, combining different datasets from different regions all over the world can improve its applicability in multiple different situations. The flexibility of the method can be better from that [7].

1. **Method**

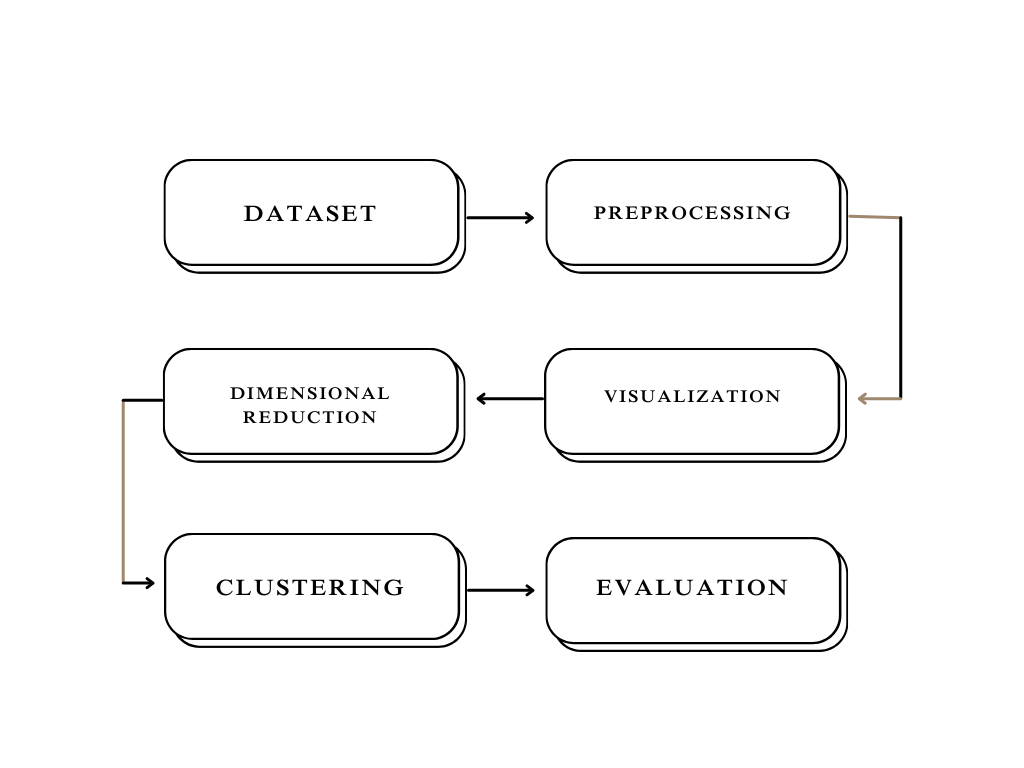
****

Fig. 1 Method Flowchart

1. **Dataset**

The dataset utilized in this research paper is the Global Data on Sustainable Energy (2000-2020) dataset from Kaggle. It is a tabular data composed of 21 features such as Entity, Year, Access to electricity, Access to clean fuels for cooking, Renewable energy capacity per person, Financial flows, Renewable energy share, Electricity from fossil fuels, Energy from nuclear, Energy from renewables, low-carbon electricity value, Primary energy consumption, Energy intensity level, carbon dioxide emissions, equivalent primary energy, GDP growth, GDP per capita, density, land area, latitude, longitude. This dataset is the first step in creating a clustering model for the research and is important for the latter stages in the study [8].

1. **Preprocessing**

The preprocessing phase is an important step because it is to ensure that the data is clean. First we check for any missing values in the dataset, then if there are any missing values, it is filled out with the mean of the column.

1. **Visualization**

Data visualization is aimed to give a visual representation of the dataset that is being utilized. The visualization takes place using ‘seaborn’ and ‘matplotlib’. The data is visualized through a series of charts and graphs to ensure that the study has an illustration of what the data is about.

1. **Dimensional Reduction**

After visualizing the data, the dataset’s dimensions are reduced by using Principal Component Analysis (PCA). PCA works by extracting the variance structure from high dimensional datasets. Because there are 21 features, dimensional reduction is needed to summarize the dataset with smaller numbers of representative variables that explain most of the variability in the data.

1. **Clustering**

K-means is the clustering method used in this research. The method is to classify the given data points into k different clusters through the iteration while converging to a local minimum. The algorithm consists of two separate phases. First, the algorithm selects k centers randomly, where the value k is already determined. Then, the euclidean distance between the data points and the cluster center is calculated. When all data points are determined to be put in some clusters, that is where the first step is done. Suppose that the target object is x, then the equation is as follows:

(1)

E is the sum of the squared error of all data points. The distance of criterion function is Euclidean distance, which is used to determine the nearest distance between a data point and a cluster center. The k-means clustering algorithm always converges to the local minimum. Then when the data points have clusters, the iterations then are executed for a number of times and recalculates the cluster center until there are no changes in the center of clusters [9].

1. **Evaluation**

For this research, the evaluation method used is the silhouette method. The silhouette score is a measure of how similar a data point is within a cluster compared to other clusters. The equation for calculating the silhouette coefficient is displayed below:

(2)

In equation (2), is the silhouette coefficient of the data point i. is the average distance between i and all the other data points in the cluster to which i belongs. And is the average distance from i to all clusters to which i does not belong. After that, the average silhouette for every k is calculated with this equation:

(3)

And lastly, the silhouette score is used to calculate the similarity of clusters. Scikit-learn’s silhouette score computes the mean silhouette coefficient of all samples. The determinator for evaluation is listed as:

* A silhouette score with a value near +1 means the data point is in the correct cluster.
* A silhouette score with a value close to 0 means the data point might belong in some other cluster
* A silhouette score with a value close to -1 means that the data point is in the wrong cluster.

To conclude, the higher the silhouette score is, the better the clustering model [10].

1. **Result**
2. **Result Visualization**

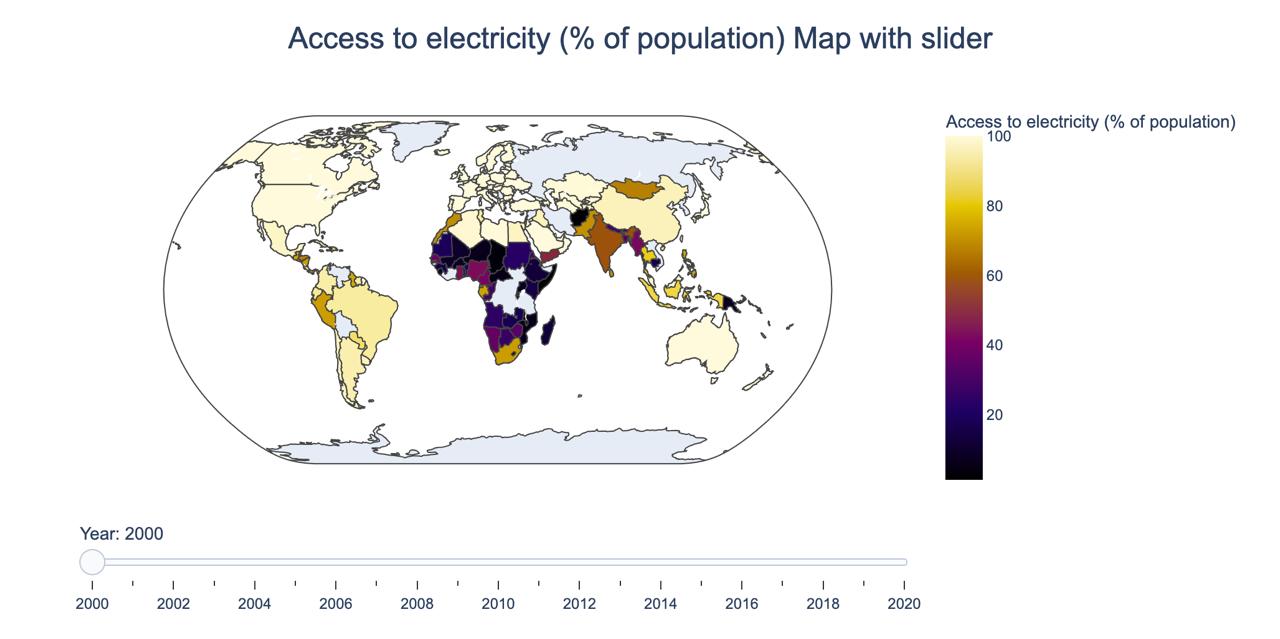
****

Fig 2. World Map of Access to Electricity with Slider

Fig. 2 illustrates the access to electricity in almost every country in the world from 2000 - 2020. Access to electricity in North America and Europe is much easier compared to Asia and Africa. In the code, Fig. 2 also has a slider that can be used to see the different segmentations of access to electricity in every country per year. Some data is not available due to data limitations from the dataset itself.

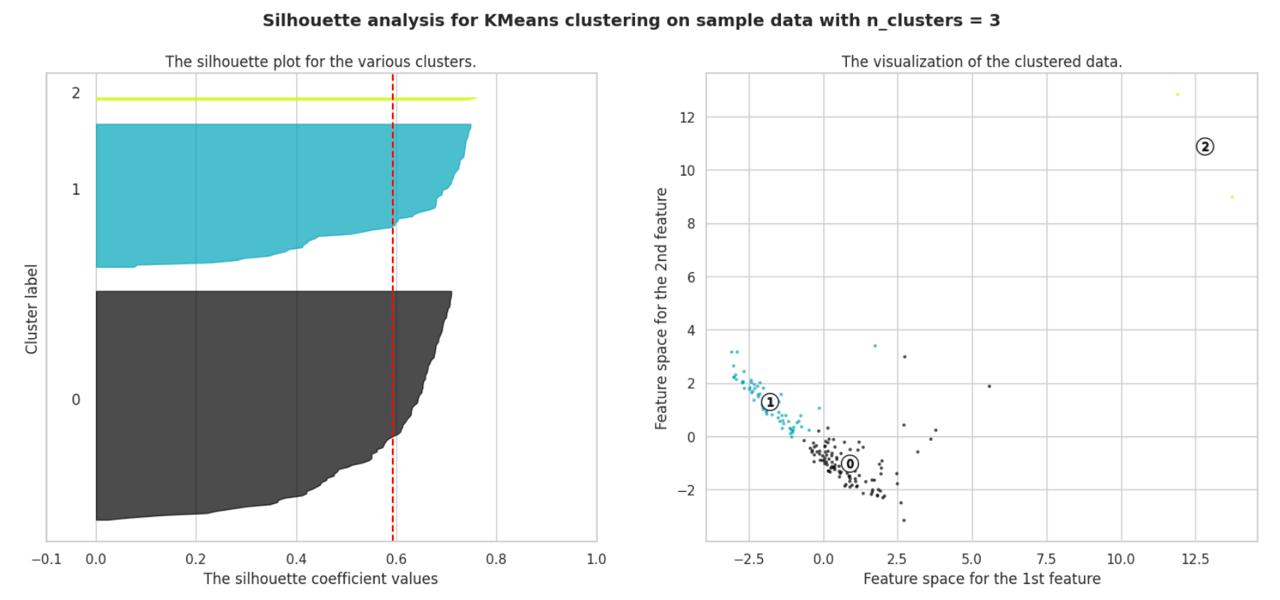


Fig. 3 Silhouette analysis for KMeans clustering with k = 3

Fig. 3 is describing the analysis based on the silhouette coefficient after using the silhouette method. In k = 3, the x axis in the left subplot shows the silhouette values, whereas the y axis shows the number of data points in the same color as the cluster. The red line shows the average silhouette value for all the clusters, in this case it is 0.59242. In the right subplot, it shows the black and blue subplot is higher than all of the green ones. Which explains the worse silhouette value on the left subplot. However, the three subplots look well-separated.

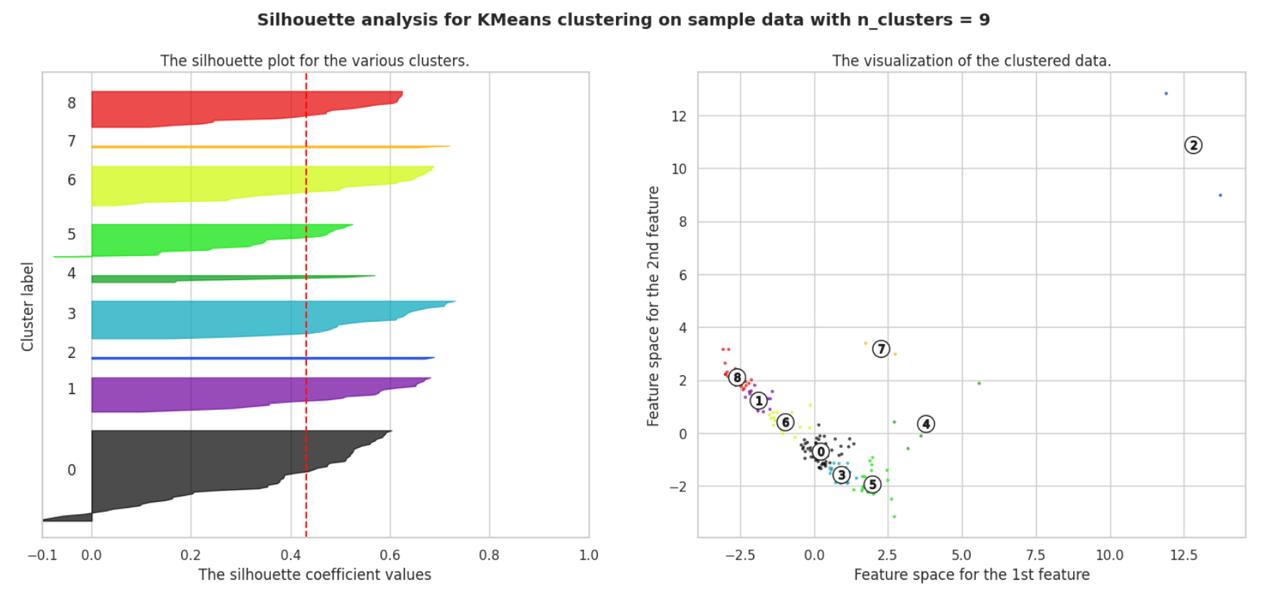


Fig 4. Silhouette analysis for KMeans clustering with k = 9

In comparison with fig 4., which is a silhouette analysis for clustering with k = 9. The left subplot shows that there are less data points in each cluster and the average silhouette value is much lower than the one in Fig 3. More specifically in cluster 2, 4, and 7. The coefficient value is also much lower than the one with k = 3. The right subplot displays that cluster 0, 1, 3, 5, 6, 8 are at a close proximity within each other and not really well-separated between each other. This makes a more challenging method to do clustering because of the low silhouette score of 0.43078.

1. **Result Evaluation**

| **k** | **Silhouette Score** |
| --- | --- |
| 2 | 0.85106 |
| 3 | 0.59242 |
| 4 | 0.46287 |
| 5 | 0.48404 |
| 6 | 0.49557 |
| 7 | 0.50019 |
| 8 | 0.44544 |
| 9 | 0.43078 |
| 10 | 0.43093 |

Table 1. Silhouette Scores for every K

Table 1. shows the evaluation method of clustering which is the silhouette score for every K. Silhouette scores measure the relations between cluster cohesion and cluster separation. The mean of these values represent the balance of the overall cohesion and separation in all the clusters. The cohesion and separation of clusters are better when the silhouette values are between 0.5 - 1. Considering the silhouette values, acceptable solutions are those with 2, 3, and 7 clusters. 2 clusters having the much higher silhouette score with 0.85106, followed by 3 clusters with 0.59242, then 7 clusters with 0.50019.

1. **Conclusions**

This study evaluated the implementation of clustering in a dataset based on renewable energy generation for each country. Through a series of different amounts of clusters, the model was fitted to make several clusters for each country. After reducing the dimensions for the features and choosing the best features, the model was able to get a strong silhouette score in some clusters, and make reasonable clusters for the others.

The limitations for this study include the lack of a middle ground between the clusters and that’s up for any future study to pick up on. This study concluded that clustering of countries within the parameters of RE generation is feasible and hopes to be a groundwork for future research and study. Potential variations of this study will arise in the future, and can be essential for preventing the use of fossil fuels and preventing the problems that rise with the earth that is continuously populating.

**References**

1. P. A. Owusu and S. Asumadu-Sarkodie, “A Review of Renewable Energy sources, Sustainability Issues and Climate Change Mitigation,” Cogent Engineering, vol. 3, no. 1, Apr. 2016, doi: <https://doi.org/10.1080/23311916.2016.1167990>.
2. A. Qazi et al., “Towards Sustainable Energy: a Systematic Review of Renewable Energy Sources, Technologies, and Public Opinions,” IEEE Access, vol. 7, no. 7, pp. 63837–63851, 2019, doi: <https://doi.org/10.1109/access.2019.2906402>.
3. M. K. M. Shapi, N. A. Ramli, and L. J. Awalin, “Energy Consumption Prediction by Using Machine Learning for Smart Building: Case Study in Malaysia,” Developments in the Built Environment, vol. 5, p. 100037, Dec. 2020, doi: <https://doi.org/10.1016/j.dibe.2020.100037>.
4. N. E. Benti, M. D. Chaka, and A. G. Semie, “Forecasting Renewable Energy Generation with Machine Learning and Deep Learning: Current Advances and Future Prospects,” Sustainability, vol. 15, no. 9, p. 7087, Jan. 2023, doi: <https://doi.org/10.3390/su15097087>.
5. A. B. Culaba, A. J. R. Del Rosario, A. T. Ubando, and J. Chang, “Machine Learning‐based Energy Consumption Clustering and Forecasting for Mixed‐use Buildings,” International Journal of Energy Research, vol. 44, no. 12, pp. 9659–9673, May 2020, doi: <https://doi.org/10.1002/er.5523>.
6. T. Alazemi, M. Darwish, and M. Radi, “Renewable Energy Sources Integration via Machine Learning modeling: a Systematic Literature Review,” Heliyon, vol. 10, no. 4, p. e26088, Feb. 2024, doi: <https://doi.org/10.1016/j.heliyon.2024.e26088>.
7. L. F. J. Alvarez, S. R. González, A. D. López, D. A. H. Delgado, R. Espinosa, and S. Gutiérrez, “Renewable Energy Prediction through Machine Learning Algorithms,” in IEEE Xplore, Oct. 2020, pp. 1–6. doi: <https://doi.org/10.1109/ANDESCON50619.2020.9272029>.
8. A. Tanwar, “Global Data on Sustainable Energy (2000-2020),” www.kaggle.com, 2023. <https://www.kaggle.com/datasets/anshtanwar/global-data-on-sustainable-energy/data>
9. S. Na, L. Xumin, and G. Yong, “Research on k-means Clustering Algorithm: an Improved k-means Clustering Algorithm,” in 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, Apr. 2010. doi: https://doi.org/10.1109/iitsi.2010.74.
10. K. R. Shahapure and C. Nicholas, “Cluster Quality Analysis Using Silhouette Score,” in 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), Oct. 2020. doi: https://doi.org/10.1109/dsaa49011.2020.00096