Anomaly Detection on World Energy Resources Consumption

Kenzie Harsanto
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
kenzie.harsanto@binus.ac.id

Kevin Tjahjadi
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
kevin.tjahjadi002@binus.ac.id

Nathaniel Emmanuel Jiaw
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
nathaniel.jiaw@binus.ac.id

Henry Lucky
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
henry.lucky@binus.ac.id

***Abstract*—Global energy consumption is a very important factor in the world, as it directly impacts the economy of the world. Due to this importance, it is essential to monitor the energy consumption in countries from different aspects. This report explores the energy consumption in each country by doing anomaly detection using data mining techniques and three different methods, which includes K-means, Isolation Forest, and Mahalanobis Distance.**

Keywords — Anomaly, Outlier, Energy, K-means, Isolation Forest, Mahalanobis Distance

I. Introduction

Understanding energy resource usage and production patterns is important for addressing global challenges such as energy security, sustainability, and climate change. While global energy consumption data provides invaluable insights into regional and national trends, it also reveals deviations—countries that use or produce very deviated energy. Identifying these anomalies or outliers is not only essential for recognizing vulnerabilities but also for uncovering strategies or circumstances that may inform broader policy decisions.

In recent years, the availability of large datasets covering various aspects of energy consumption has increased. This provides opportunities for various methods and other techniques to be implemented to help optimize energy management systems. Moreover, these algorithms and techniques can also help monitor and detect anomalies inside these large datasets, providing valuable insights, helping reduce economic loss, and early detection of inefficiency in the consumption of energy [1] [2].

In this report, we analyze global energy consumption data taken from Kaggle, titled *CO2-Emission by Country Growth and Population* [3] dataset, and the *Energy Data by Country* [4] dataset, to detect and characterize countries with anomalous energy resource patterns. Several methods are used to help detect anomalies inside the datasets, which include K-means algorithms, Isolation Forest, and Mahalanobis Distance. Data mining techniques are also used to extract patterns of energy consumption to detect anomalies. By applying these methods and data mining techniques, we aim to identify anomalies in both energy usage and production metrics.

II. Theoretical Basis

A. Anomaly

Anomalies are occurrences in a dataset that are in some way unusual and do not fit the general pattern that exists. These irregularities arise from rare events like data entry errors, measurement issues, or unexpected variations in system behavior

B. Knowledge Discovery in Databases (KDD)

KDD is a process that extracts useful information from large datasets and uses it to make decisions or predictions. There are several steps in the KDD process:

1. Data Cleaning
   Handling the missing and noisy data, removing the outliers, and correcting inconsistent data.
2. Data Integration
   Combining multiple sources of data into a coherent dataset.
3. Data Transformation
   Transforming data into the appropriate form.
4. Data Selection
   Selecting relevant data to be analyzed.
5. Data Mining
   Applying techniques that are used to extract patterns to find useful information.
6. Pattern Evaluation
   Identifying patterns that represent knowledge based on given data.
7. Knowledge Presentation
   Presenting the results in a meaningful way can be used to make decisions.

KDD is an iterative process in which evaluation metrics can be improved, data mining techniques can be adjusted, and new data can be incorporated and transformed to achieve more suitable and varied outcomes.

C. K-means Clustering

K-means clustering is an unsupervised machine learning algorithm that is tasked to separate unlabeled data points in

a dataset into *k* clusters based on their distance to the centroids, which is the central point of each cluster. Firstly, centroids are randomly created in the space where the data points are located. Each data point is then assigned to a cluster with the closest centroid. After all data points are assigned to a cluster, the algorithm will create new centroids based on the mean value of all the data points in each cluster. The data points will then be reassigned to a cluster based on the new centroids. This process will keep on repeating until the centroids are no longer changing in which the algorithm will stop.

K-means clustering is used in anomaly detection to find the distance between each point to the cluster centroid [5]. If the distance between a data point and the nearest cluster centroid is really large, then the point can be identified as an anomaly or an outlier in the dataset. These data points can be determined as an anomaly because they deviate from the norm.

Several methods can be used to determine the optimal number of clusters for the K-means algorithm on a dataset. One of the most popular methods is the elbow method [6]. The elbow method plots a graph that shows the relationship between the number of clusters on the x-axis and the Within-Cluster Sum of Squares (WCSS) on the y-axis. WCSS measures the compactness of the data points in the clusters, a lower WCSS indicates a more compact cluster, which means that the data points are closer to their centroid, while a higher WCSS indicates a less compact cluster. After plotting the graph, the optimal number of clusters is determined by the "elbow" point, which is a point in the graph that resembles an elbow and the decrease in WCSS starts to slow down.

D. Isolation Forest

Isolation Forest is an algorithm for data anomaly detection using binary trees. It is suitable for unsupervised learning. Isolation doesn't actually classify data into classes, but it only calculates how easy it is to isolate data during splitting [7]. The algorithm constructs multiple random decision trees and measures the path length needed to isolate each point. It measures how easily a data point can be separated from the rest of the data. Each data that is separated is assigned an anomaly score.

Anomaly scores are calculated based on how fast the data can be isolated from the rest of the data, if they require fewer random splits to reach the leaf node it is likely that they are considered an anomaly since they fall outside of the typical range of values of the feature. Anomaly detection is calculated by averaging the path lengths across all of the isolation trees in the forest. Anomaly Scores have a threshold of 0.5, with higher scores falling in the category of an outlier.

E. Mahalanobis Distance

Mahalanobis distance is a statistical measure used to determine the distance between a point and a distribution. It accounts for the correlations between variables and the spread of the data, making it very good for multivariate data [8]. Mahalanobis distance is particularly effective for anomaly detection because it identifies data points that derive significantly from the normal distribution of data. Anomalies often fall from the central cluster of points in a dataset, and Mahalanobis distance will capture the deviation by considering both the mean and variance structure of the data. It works first by calculating the mean and covariance matrix of the referenced data, the new data points are then compared to the referenced distribution using the formula.

$$Dm = \sqrt{(x - \mu)^{T\Sigma^{-1}}(x - \mu)}$$

If the distance exceeds a certain threshold, the point is considered an anomaly.

F. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a statistical algorithm used to reduce a dataset's dimensionality while retaining important patterns and information or variance. PCA is widely used in datasets that have a large number of features or attributes that can lead to issues such as overfitting. The features in the dataset are combined to create new axes called the Principal Components (PC), with the first PC having the most variance in the data, while the second PC has the second most variance in the data while being orthogonal to the first PC. Each subsequent PC captures the variance that is orthogonal to the previous PC and has less variance than the previous PC. Furthermore, PCA uses covariance matrix, eigenvalues, and eigenvectors to determine the direction and the magnitude or the amount of variance of each PC [9].

Elbow Method can also be used to determine the optimal number of Principal Components in PCA for a dataset. The graph that is created to find the "elbow" point is known as the Scree Plot, which shows the relationship between the number of components on the x-axis and the eigenvalue, which defines the amount of variance in each PC, on the y-axis. Similar to K-means clustering, the "elbow" point that determines the optimal number of PC is located where the curve begins to flatten and an increase in the number of PC does not significantly decrease the eigenvalue.

G. MinMax Scaling

MinMax scaling is a normalization technique used in data preprocessing to rescale the features of a dataset to a specific range. The scale of the features in MinMax scaling is adjustable, typically from 0 to 1 (but are able to be adjusted so it can handle negative values). This makes it useful for algorithms that are sensitive to the size of the data, such as gradient-based methods (e.g. logistic regression, neural networks) or distance-based models (e.g. k-NN, k-Means Clustering) [10].

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Figure 2.1. MinMax scaling formula

H. Silhouette Score

Silhouette score is a metric that is used to evaluate the similarity between data points in each cluster and the clustering quality. It shows how well the data points are clustered in each cluster. Silhouette scores that are above 0.7 are considered very good, while anything under 0.25 is considered severely bad.

III. METHODOLOGY

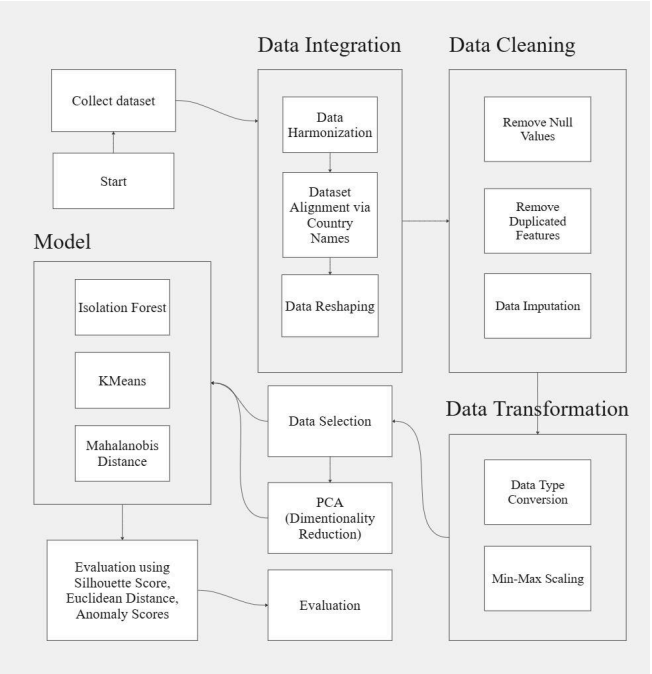

Fig 2. Proposed Method

A. Dataset

This paper uses 2 (two) different datasets that are then integrated and merged into 1 (one) dataset. The datasets taken are *the CO2-Emission by Country Growth and Population* [3] dataset and the *Energy Data by Country* [4] dataset which are both taken from Kaggle. Both datasets contain different features and rows that describe each country's energy consumption on each dataset.

1) Dataset Overview

- *CO2-Emission by Country Growth and Population [2]*
  - The dataset contains 10 attributes, with 2 of the features categorized as categorical while the rest of the features are numerical
  - Summary of the data
    - Energy Types consist of:
      a. All Energy Types
      b. Coal
      c. Natural Gas
      d. Petroleum
      e. Nuclear
      f. Renewables
    - GDP measures *gross domestic product*, Population, Country names, energy intensity per capita, energy intensity by GDP, and CO2 emission
  - The total country for the dataset is 231
- *Energy Data by Country [3]*
  - The dataset contains 44 features or attributes with 12 of the features categorized as categorical while the rest of the features are numerical.
  - Summary of the data
    - There are 5 types of energy:
      a. Oil
      b. Gas
      c. Coal
      d. Nuclear
    - Country name, energy consumption, and world shares
  - The detail for every energy is as follows:
    a. Reserves
    b. Consumption
    c. Production
    d. Net Imports
    e. Year
    f. Units
    g. Net Exports
    h. Deficit
  - The total country for the dataset is 212, there are 41 features
- *Both Dataset Overview*

  The first dataset complements the second dataset as the second dataset does not have several data such as GDP, energy intensity per capita, and energy intensity per GDP. The first dataset also contains energy types that do not exist in the first dataset

B. Data Integration

a. Data Harmonization

The first dataset contains data from the year 2016 and the second dataset contains data from the years 1980 - 2016. To ensure the integrity of the data, we take data from 2016

b. Dataset Alignment via Country Names

The first dataset comprises data from 212 countries, while the second dataset includes data from 231 countries. To address this issue, countries that do not appear in both datasets were excluded

c. Data Reshaping

Country	Energy_type	Year
World	all_energy_types	1980
World	coal	1980
World	natural_gas	1980
World	petroleum_n_other_liquids	1980
World	nuclear	1980
World	renewables_n_other	1980

Fig 3. Vertical Data Structure in First Dataset

Figure 3 shows the structure of the first dataset, here the energy types are distributed vertically. To integrate the first and second datasets, making the data horizontal is needed. To address this issue, we are using a method called pivoting or data transposition which transforms or flips the rows into columns and vice versa

d. Data Integration

Once the problems are all sorted, the first and second datasets will be merged into one by using the country name

C. Data Cleaning

Upon merging both datasets, the data cleaning process was conducted using a data summary analysis.

- Remove Null Values

  The data summary revealed there are missing values in certain features. To address this, features

with a majority of null values were excluded from the dataset to ensure data quality

- Remove Duplicated Features

  After integrating the dataset, there are features that have the same meaning as another feature such as "population_2016" and "Population". The duplicated feature is then removed to ensure the data quality

- Data Imputation

  For features containing a small proportion of missing values, the missing values were imputed using the mean value of the respective feature

D. Data Transformation

- Data Type Conversion
  Certain features contained percentage values stored as strings. To ensure numerical consistency, these values are converted into decimals. Transforming them into numerical data types
- Min-Max Scaling
  Min-Max Scaling is used to ensure the features in the dataset are scaled within a range while preserving the data distribution. This technique maintains negative values as negative and positive values as positive.

E. Data Selection

Following the integration and cleaning of the dataset, a correlation matrix was used to ensure the data selected for analysis was both relevant and meaningful. The correlation matrix measures the degree of association between features, enabling the identification of redundant features and those that may not contribute significant information to the analysis. It minimizes the risk of overfitting.

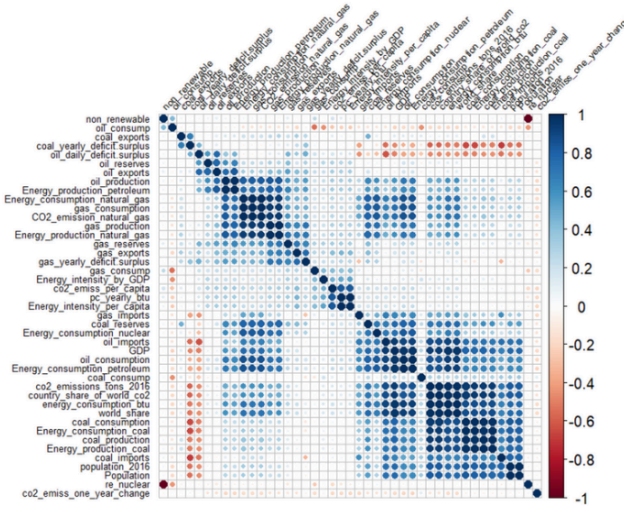

Fig 4. Correlation Matrix

Figure 4 shows the correlation among all features in the dataset. Weak correlations are represented by white cells, while strong correlations are depicted using blue or red hues, indicating the strength of the correlation. Redundant features are identified when multiple features have the same color pattern. Features that are redundant and have a weak correlation will be removed. This ensures the data are relevant and meaningful.

F. Principal Component Analysis

After completing the data selection process, our analysis will be conducted using both non-PCA (untransformed) and PCA-transformed features. To determine the optimal number of principal components to use, we will use a scree plot

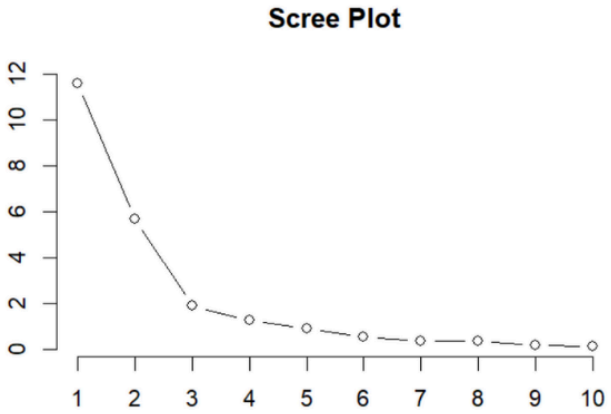

Fig 5. Scree Plot

Based on the result shown in Figure 5, the scree plot suggests that the optimal number of components is 3 as the explained variance reduces after the third component

G. Modelling

This study uses Isolation Forest, KMeans Clustering, and Mahalanobis Distance. Isolation Forest is an ensemble-based anomaly detection method, KMeans is a distance-based anomaly detection, and Mahalanobis Distance are statistical-based anomaly detection.
Isolation Forest is specifically built for detecting outliers. KMeans is primarily a clustering algorithm but can be adapted for anomaly detection by analyzing the distance between the data points and the cluster centroids by using metrics such as Euclidean Distance and Silhouette Score, the degree of deviation from the cluster can identify outliers. To determine how many clusters we are using for KMeans, the elbow method is used.

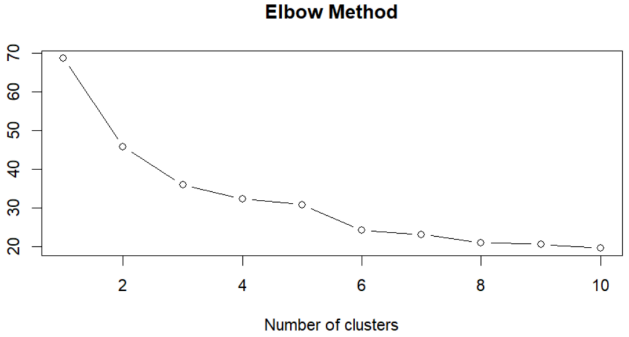

Fig 6. Elbow Method

Based on Figure 6. the most fitted number of clusters is 3. Mahalanobis measures the distance between a data point and the mean of the dataset using the correlations

between features. These 3 models excel in detecting outliers in multivariate datasets.

H. Metrics

- Isolation Forest
  - Anomaly Scores
    Anomaly scores are assigned to each data point based on how quickly the data can be isolated in a randomly constructed isolation tree. Data points that are easier to isolate have higher anomaly scores. The threshold for anomaly scores is usually 0.5, but adjustments can be made. The lower the anomaly score is, the better the data point
- K-Means Clustering
  - Euclidean Distance
    Euclidean distance is used to measure the distance between a data point and its cluster centroids.

$$d = \sqrt{\sum_{i=1}^{n} (xi - \mu i)^2}$$

    By measuring the distance, data points that are significantly farther from their clusters can be identified as anomalies.
  - Silhouette Score
    Silhouette Score measures how well a data point fits within its cluster. A low silhouette score can indicate a potential anomaly since the point doesn't belong clearly to its assigned clusters. A data point that has a Silhouette score under 0.2 is considered bad and can be identified as an anomaly.

$$S = \frac{b - a}{max(a, b)}$$

- Mahalanobis Distance
  - Chi-Square Distribution
    To identify anomalies, the Mahalanobis distance can be compared against a critical value derived from the 99% chi-square confidence interval. It sets a threshold based on the number of features

IV. Results and Discussion

Each of the methods is tested with and without PCA to see the difference and compare the effects of PCA on this dataset. Then, the result is calculated and the top outliers are recorded.

Isolation Forest	Anomaly Score Average
Without PCA	0.356
With PCA	0.363

Table 1. The average of anomaly scores of Isolation Forest without and with PCA.

K-Means Clustering	Silhouette Score
Without PCA	0.71
With PCA	0.804

Table 2. The silhouette scores of K-Means Clustering without and with PCA.

Table 1 presents the average anomaly scores calculated using the Isolation Forest method, both with and without PCA. Without applying PCA, the average anomaly score is 0.356. Applying PCA results in a slight increase, with the average anomaly score rising to 0.363. A threshold of 0.55 was chosen to identify the top outliers, as using the default threshold of 0.5 produced a huge amount of outliers. Without PCA, the countries identified as outliers include the United States, Russia, China, India, Saudi Arabia, Japan, Iran, Germany, and Canada. When PCA is applied, the same countries are identified as outliers, with the addition of four more: France, Iraq, Australia, and Qatar. This indicates that applying PCA slightly enhances the method's sensitivity to additional outliers.

Table 2 shows the silhouette score results using the k-means clustering method with and without PCA using 3 clusters. Without PCA, the silhouette score is 0.71, whereas, after applying PCA, the silhouette score increases to 0.804. Anomalies also were identified using both Euclidean distance and silhouette scores. Before applying PCA, countries identified as anomalies based on the Euclidean distance metric include India, Japan, China, the United States, and Russia. Moreover, when using the silhouette score, countries such as China, Qatar, Saudi Arabia, Iran, Canada, and the United States are considered anomalies. After using PCA, countries that are considered anomalies based on the Euclidean distance metric are Saudi Arabia, India, the United States, Russia, and Japan. Furthermore, countries identified as anomalies using the silhouette score include China, Russia, the United States, and India.

The Mahalanobis Distance method is computed using the default threshold, which is the 99% chi-square confidence. Without applying PCA, the method identifies a large number of outliers, totaling 20 countries: Singapore, China, Australia, Qatar, Iraq, Algeria, Norway, Indonesia, United Kingdom, Saudi Arabia, Brazil, France, South Korea, Iran, Germany, Canada, Russia, United States, Japan, and India. However, when PCA is applied, the number of outliers decreases significantly to just 5 countries: the United States, China, Russia, Saudi Arabia, and India. This reduction indicates that the Mahalanobis Distance method becomes more efficient with PCA, as PCA reduces the number of features, simplifying the data and focusing on the most significant components.

The result of the three methods that are used gives a similar outcome with the top 5 countries that are considered

anomalies being China, the United States, Saudi Arabia, Russia, and India. The result can be divided into High-consumption energy outliers and Low-consumption energy outliers by looking at the population, GDP, and total energy consumption per btu (British Thermal Unit).

The outlier countries can take specific actions based on their energy use to address their issues. High-consumption countries should focus on setting rules to make industries use energy more efficiently and penalize those that waste energy. They can also offer subsidies or tax cuts to support renewable energy projects and adopt energy-saving technologies in factories, buildings, and transportation. For low-consumption countries, the focus can be on making energy more affordable by reducing costs for low-income households. They can also work with international organizations to get funding and support to improve their energy systems. These steps can help both types of countries use energy more wisely and support global efforts for sustainable energy use.

The code of this project is in this link: https://github.com/k3ntut/Data-Mining-Project/

V. Conclusion

The results from the three anomaly detection methods, Isolation Forest, K-means clustering, and Mahalanobis Distance, indicate consistent identification of anomaly countries. Both with and without PCA, the results are alike, with similar countries identified as outliers. However, the Mahalanobis Distance method showed a significant difference, identifying a large number of outliers without PCA but significantly fewer with PCA, highlighting its improved efficiency when the number of features is reduced.

The analysis also revealed that factors such as GDP, population, and average energy usage moderately influence the identification of outliers. Countries with high energy production and consumption are more often identified as outliers, given their significant contributions to global energy metrics. Small countries like Qatar also appear as outliers given their heavy production on energy resources.

The methods were able to find outliers in both big and small countries, showing that they can work well with different levels of energy production and consumption. This highlights the need for unique solutions to address the specific energy challenges of each country.

References

[1] https://ieeexplore.ieee.org/document/8621948
[2] https://www.researchgate.net/publication/353701611_On_the_nature_and_types_of_anomalies_a_review_of_deviations_in_data
[3] https://www.kaggle.com/datasets/lobosi/c02-emission-by-countrys-growth-and-population (Dataset 1)
[4] https://www.kaggle.com/datasets/geekraccoon/energy-data-by-country/data (Dataset 2)
[5] https://www.researchgate.net/publication/271891551_Network_Anomaly_Detection_by_Cascading_K-Means_Clustering_and_C45_Decision_Tree_algorithm (K-Means Clustering for Anomaly Detection Method)
[6] https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/ (Elbow Method)
[7] https://www.researchgate.net/publication/224384174_Isolation_Forest (Isolation Forest)
[8] https://arxiv.org/pdf/2003.00402 (Mahalanobis Distance)
[9] https://www.geeksforgeeks.org/principal-component-analysis-pca/ (PCA)
[10] https://www.atoti.io/articles/when-to-perform-a-feature-scaling/ (Min-Max Scaling)