Jurnal Rekayasa Mesin
p-ISSN: 1411-6863, e-ISSN: 2540-7678
Vol.20, No.2, Agustus **2025**, hal. 279-288

Clustering-Based Analysis of Fuel Efficiency and Emissions in Automotive Data Using PCA and K-Means

Sunardi*, Ananda Nur Daffa Zain
Departement of Mechanical Engineering, Universitas Muhammadiyah Yogyakarta, JL. Brawijaya, Kasihan, Bantul, Yogyakarta 55183, Indonesia

*E-mail: sunardi@umy.ac.id.

Submitted: 21-06-2025; Accepted: 23-08-2025; Published: 31-08-2025

Abstract

Growing concerns regarding greenhouse gas emissions and fuel consumption have placed considerable demands on the automotive sector. To address these issues, this research applies unsupervised learning approaches namely Principal Component Analysis (PCA) and K-Means Clustering to categorize vehicles based on attributes associated with energy efficiency and environmental impact. Using a publicly available vehicle dataset, PCA was used to simplify the data by reducing dimensionality while preserving significant patterns. Subsequently, K-Means was employed to segment the data into three distinct clusters according to shared features like engine size, fuel usage, and CO_2 output. The resulting groupings effectively identified categories such as fuel-efficient, moderately consuming, and high-consumption vehicles. Visual representation in two-dimensional space further confirmed meaningful distinctions among the clusters, offering practical insights for both manufacturers and consumers.

Keywords: *clustering; CO_2 emissions; data analysis; fuel consumption; machine learning*

Abstrak

Tingginya kekhawatiran terhadap emisi gas rumah kaca dan konsumsi bahan bakar menuntut sektor otomotif untuk menemukan pendekatan analitis yang lebih efektif. Studi ini menerapkan metode pembelajaran mesin tak terawasi yaitu *Principal Component Analysis* (PCA) dan *K-Means Clustering* untuk mengelompokkan kendaraan berdasarkan atribut-atribut terkait efisiensi bahan bakar dan dampak lingkungan. Dataset kendaraan yang tersedia secara publik digunakan untuk mereduksi dimensi data menggunakan PCA tanpa kehilangan pola penting. Setelah itu, algoritma *K-Means* digunakan untuk membagi data ke dalam tiga klaster berdasarkan kesamaan karakteristik seperti kapasitas mesin, konsumsi bahan bakar, dan emisi CO_2. Hasil klastering ini mengungkap tiga kelompok kendaraan dengan kategori efisiensi tinggi, konsumsi sedang, dan konsumsi tinggi. Visualisasi dua dimensi menunjukkan pemisahan klaster yang signifikan, memberikan wawasan praktis bagi produsen dan konsumen.

Kata kunci: clustering; emisi CO_2; analisis data; konsumsi bahan bakar; machine learning

1. Introduction

As global environmental concerns intensify, the automotive sector is undergoing significant transformation [1]. Modern vehicles are now expected not only to perform efficiently but also to meet strict environmental regulations [2]. Increasing carbon emissions from internal combustion engines have become a major contributor to air pollution and climate change [3, 4]. Consequently, governments and environmental organizations across the world are pushing for improvements in vehicle fuel efficiency and reductions in greenhouse gas emissions [5, 6]. Consumers are likewise more inclined to choose vehicles that are both fuel-efficient and environmentally sustainable [7-9].

In response to this shift, leveraging advanced data analysis techniques has become a critical tool in understanding and improving vehicle performance [10]. Among these approaches, machine learning has emerged as an effective solution for discovering insights from complex datasets [11]. Specifically, unsupervised learning a category of machine learning that identifies patterns in data without predefined labels offers valuable potential for clustering vehicles based on operational and environmental metrics [12].

Two widely adopted techniques in unsupervised learning are Principal Component Analysis (PCA) and K-Means Clustering [13]. PCA is used to simplify high-dimensional data by projecting it onto a lower-dimensional space while preserving the most significant variance. This not only improves visualization but also enhances computational efficiency and reduces redundancy [14, 15]. K-Means, meanwhile, is a clustering algorithm that organizes data into groups based on similarity, typically using Euclidean distance as a metric [16]. The synergy of PCA and K-Means has proven effective in applications involving classification, segmentation, and trend detection in various domains [17].

Despite the widespread use of PCA and K-Means in fields such as marketing, medical diagnostics, and industrial quality control [18-21], their implementation in automotive performance analysis especially for fuel economy and CO_2 emissions is still relatively underutilized. While prior research often focuses on predictive models for fuel consumption [22] or supervised classification based on vehicle types [23], the potential for unsupervised techniques to reveal hidden clusters among vehicle characteristics remains largely untapped.

This study addresses that gap by applying PCA and K-Means to a publicly available vehicle dataset. The dataset includes six primary features: fuel consumption in urban and highway conditions, engine displacement, cylinder count, fuel cost per year, and CO_2 emissions from the tailpipe. The use of PCA allows these variables to be compressed into two principal components, enabling easier interpretation and visual representation. K-Means clustering is then used to group vehicles into distinct categories based on performance and environmental efficiency.

The goal is to identify clusters of vehicles with similar behavior—such as low-emission, fuel-efficient models versus high-emission, fuel-intensive models—without requiring labeled data. This approach offers valuable insights that can assist manufacturers in benchmarking their products, policymakers in designing environmental strategies, and consumers in making informed choices.

The research objectives are as follows:To reduce data complexity using PCA while maintaining interpretability; To cluster vehicles using K-Means based on relevant performance metrics; To visualize the clusters in two-dimensional space for clearer understanding; To explore the implications of these clusters on energy efficiency and emission reduction strategies. Through this method, the study contributes to the growing interest in data-driven sustainable engineering, demonstrating how unsupervised learning can support environmental assessment in the automotive field.

2. Material and Method

This study employs a publicly available dataset containing detailed specifications of various vehicles, which was sourced from GitHub (https://raw.githubusercontent.com/hadley/fueleconomy/master/data-raw/vehicles.csv). The dataset includes variables such as fuel consumption in city and highway driving conditions (miles per gallon), engine displacement (in liters), number of cylinders, estimated annual fuel cost (in US dollars), and tailpipe CO_2 emissions (grams per mile). These features were selected because they offer measurable indicators of a vehicle's environmental and operational performance. **Table 1** shows the number of data and the percentage of data distribution.

Table 1. Number of data and percentage of data distribution

Category	Number	Percentage (%)
Total of Data	34565	100
Train Data	24195	70
Test Data	6913	20
Validation Data	3456	10
Sample Data	15	0.04

2.1. Data Cleaning and Standardization Process

The dataset was initially examined to assess completeness. Rows containing missing or non-numeric entries were excluded to maintain the integrity and accuracy of the analysis. Following data cleaning, a standardization process was conducted using the StandardScaler function from the Scikit-learn library. This step ensures that all numerical features are transformed to a common scale with zero mean and unit variance, preventing any single feature from dominating the clustering results due to differences in magnitude.

2.2. Principal Componen Analysis (PCA)

To reduce data dimensionality and enhance visualization, Principal Component Analysis (PCA) was applied. This technique transforms the original six-dimensional data into two new uncorrelated components, which capture the highest variance in the dataset. The two principal components were then used as the input for the clustering phase. PCA not only improves computational efficiency but also allows the data to be visualized in a two-dimensional plot, making it easier to interpret the relationships among vehicles. The transformation can be expressed using the following formula [1]:

$$\boldsymbol{Z} = \boldsymbol{X} \cdot \boldsymbol{W} \tag{1}$$

In this study, $\boldsymbol{Z}$ denotes the projected data in the reduced-dimensional space obtained after transformation, while $\boldsymbol{X}$ represents the standardized data matrix, where rows correspond to the samples and columns correspond to the features. The transformation is achieved using $\boldsymbol{W}$, which is the matrix of eigenvectors extracted from the covariance matrix of $\boldsymbol{X}$. Through this projection, the high-dimensional data $\boldsymbol{X}$ is mapped into a lower-dimensional representation $\boldsymbol{Z}$, thereby preserving the most significant variance in the dataset while reducing redundancy and computational complexity.

2.3. K-Means Clustering

For the clustering process, K-Means Clustering was implemented. This algorithm partitions the data into k distinct groups by minimizing the variance within each cluster. Based on exploratory analysis and domain knowledge, the number of clusters (k) was set to three. These clusters are intended to represent three general categories of vehicle efficiency: low, moderate, and high fuel consumption/emission groups. The K-Means model was initialized with a fixed random state to ensure reproducibility. The objective function of K-Means is defined as [2]:

$$\boldsymbol{J} = \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{k}} \sum_{\boldsymbol{x_j} \in \boldsymbol{C_i}} \left\| \boldsymbol{x_j} - \boldsymbol{\mu_i} \right\|^{\mathbf{2}} \tag{2}$$

In the context of K-Means clustering, the objective function aims to minimize the total within-cluster variation, which is expressed as $\boldsymbol{J}$. This value represents the total within-cluster sum of squared distances, serving as a measure of compactness for the clusters. Each $\boldsymbol{x_j}$ corresponds to the $\boldsymbol{j}$-th data point that belongs to cluster $\boldsymbol{C_i}$, while $\boldsymbol{\mu_i}$ denotes the centroid or the mean position of all data points within cluster $\boldsymbol{i}$. The parameter $\boldsymbol{k}$ indicates the number of clusters specified in the analysis. By iteratively updating the centroids and reassigning data points, the K-Means algorithm seeks to minimize $\boldsymbol{J}$, thereby ensuring that the data points within each cluster are as close as possible to their respective centroid, leading to well-defined and meaningful groupings.

The entire process data preprocessing, dimensionality reduction, and clustering was conducted using Google Colab, a cloud-based environment that supports Python and data science workflows. All code was executed using libraries such as Pandas, Matplotlib, Seaborn, and Scikit-learn.

A simplified version of the methodology is summarized in the following workflow:

1. Import vehicle dataset
2. Data cleaning (remove incomplete rows)

3. Standardize numerical features
4. Apply PCA to reduce to 2 components
5. Perform K-Means clustering with $k = 3$
6. Visualize clusters in 2D PCA space

This methodological framework ensures a robust, interpretable, and replicable analysis of vehicle performance and environmental categorization.

3. Results and Discussion

This section presents the outcome of the data preprocessing, dimensionality reduction using PCA, and clustering using K-Means. The results are analyzed both quantitatively and visually to uncover underlying patterns in the vehicle dataset related to fuel consumption and emissions.

3.1. Initial Data Analysis

The cleaned dataset consists of 15 vehicle samples selected from different brands, focusing on six critical numerical features: city fuel economy (city), highway fuel economy (highway), engine displacement (displ), number of engine cylinders (cylinders), estimated annual fuel cost (fuelCost), and tailpipe CO_2 emissions (CO_2TailpipeGpm). The data was standardized to ensure equal weighting in the clustering process. **Table 2** displays a snapshot of the standardized values for the 15 vehicles. The wide range in values confirms the variation in vehicle specifications and performance, which is necessary for meaningful clustering.

Although the complete dataset comprises more than 34,000 vehicle entries, only 15 representative samples are shown in **Table 2** to provide a clearer illustration of the preprocessing, standardization, and initial clustering process. These selected samples demonstrate the diversity of vehicle specifications across different brands and categories. The subsequent PCA transformation and K-Means clustering, however, were applied to the entire dataset, and the comprehensive results are presented in **Figures 3** and **4** as well as **Table 3**. Thus, the 15 vehicles serve only as illustrative examples, while the main basis of analysis relies on the full dataset.

Table 2. Sample of Standardized Data for 15 Vehicles

Car Brands	City	Highway	Cylinders	Displ	FuelCost	CO_2TilePigeGpm
TOYOTA	13	17	8	5.7	3650	592.47
HONDA	26	35	4	1.7	1850	296.23
FORD	31	39	4	2	1750	299.41
CHVROLET	13	17	8	6.5	4250	727.14
BMW	17	25	6	3	3000	444.35
MERCEDES	20	28	6	3	2600	445.00
AUDI	20	28	4	2	2600	393.00
NISSAN	12	18	8	5.6	2950	634.79
HYUNDAI	18	27	6	3.8	2600	423.19
SUBARU	22	30	4	2	2400	360.00
KIA	18	25	6	3.5	2600	423.19
MAZDA	25	34	4	1.6	1950	317.39
VOLKSWAGEN	21	26	4	1.8	2400	386.39
PORCHE	17	23	6	3.8	3150	467.74
LAND ROVER	11	14	8	4.6	5050	740.58

3.2. Dimensionality Reduction using PCA

To visualize the data more effectively, Principal Component Analysis (PCA) was applied to reduce the original six features into two principal components. These components explain the highest variance in the dataset, making it possible to represent multidimensional data in two dimensions without significant loss of information.

As seen in **Figure 1**, PCA projects the samples onto a 2D space, revealing clusters of vehicles with similar characteristics. Vehicles with high efficiency tend to group together, while those with poor fuel economy and high emissions are separated along the principal axes.

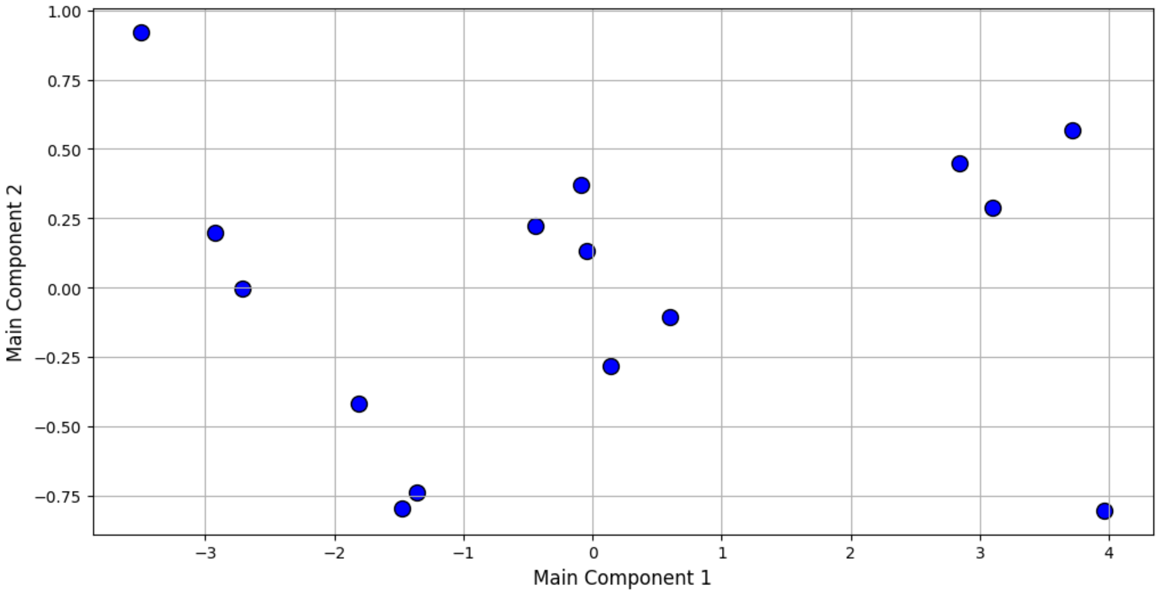

Figure 1. 2D Visualization of 15 Vehicle Samples using PCA

3.3. Clustering using K-Means

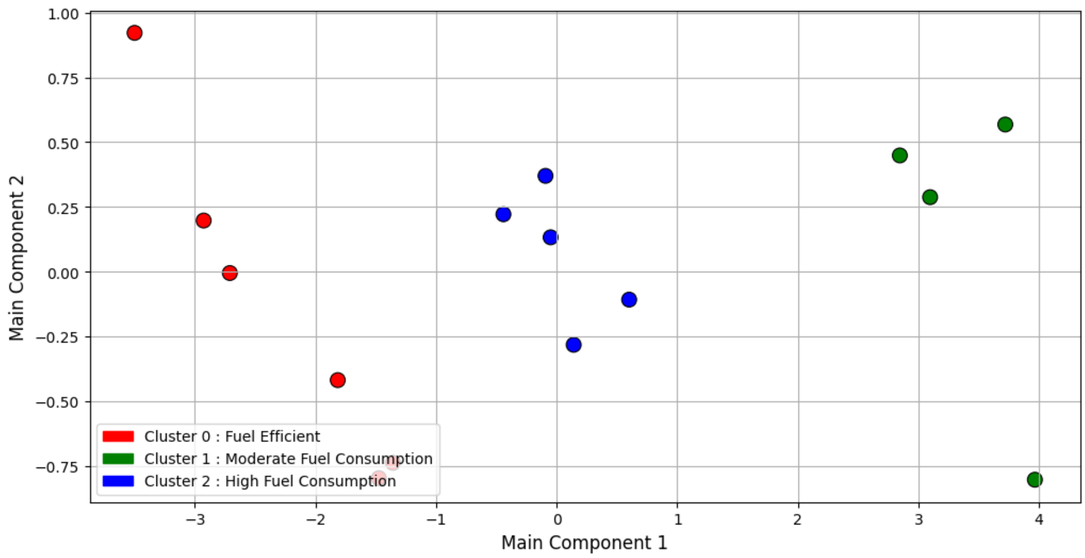

Figure 2. K-Means Clustering Output for 15 PCA-Transformed Vehicle Samples

Following PCA, K-Means Clustering was used to segment the vehicle samples into three clusters. The number of clusters ($k = 3$) was chosen based on domain knowledge and initial experimentation. Each cluster represents a distinct category of vehicle efficiency: Cluster 0: Vehicles with low fuel consumption and emissions (efficient group); Cluster 1: Moderately efficient vehicles; Cluster 2: Vehicles with high fuel consumption and CO_2 emissions

The results of clustering are shown in **Figure 2** where each color denotes a different cluster. Vehicles are clearly grouped, supporting the effectiveness of PCA in simplifying the data for clustering purposes.

Table 3. Cluster Assignments after PCA and K-Means Clustering

Car Brands	City	Highway	Cylinders	Displ	FuelCost	CO_2TilePigeGpm	Cluster
TOYOTA	13	17	8	5.7	3650	592.47	1
HONDA	26	35	4	1.7	1850	296.23	0
FORD	31	39	4	2	1750	299.41	0
CHVROLET	13	17	8	6.5	4250	727.14	1
BMW	17	25	6	3	3000	444.35	2
MERCEDES	20	28	6	3	2600	445.00	2
AUDI	20	28	4	2	2600	393.00	0
NISSAN	12	18	8	5.6	2950	634.79	1
HYUNDAI	18	27	6	3.8	2600	423.19	2
SUBARU	22	30	4	2	2400	360.00	0
KIA	18	25	6	3.5	2600	423.19	2
MAZDA	25	34	4	1.6	1950	317.39	0
VOLKSWAGEN	21	26	4	1.8	2400	386.39	0
PORCHE	17	23	6	3.8	3150	467.74	2
LAND ROVER	11	14	8	4.6	5050	740.58	1

Table 3 summarizes the cluster assignment for each of the 15 samples, including their original performance metrics. This grouping makes it easier to identify which vehicles fall into desirable categories for environmental and economic considerations.

3.4. Full Dataset Visualization and Cluster Distribution

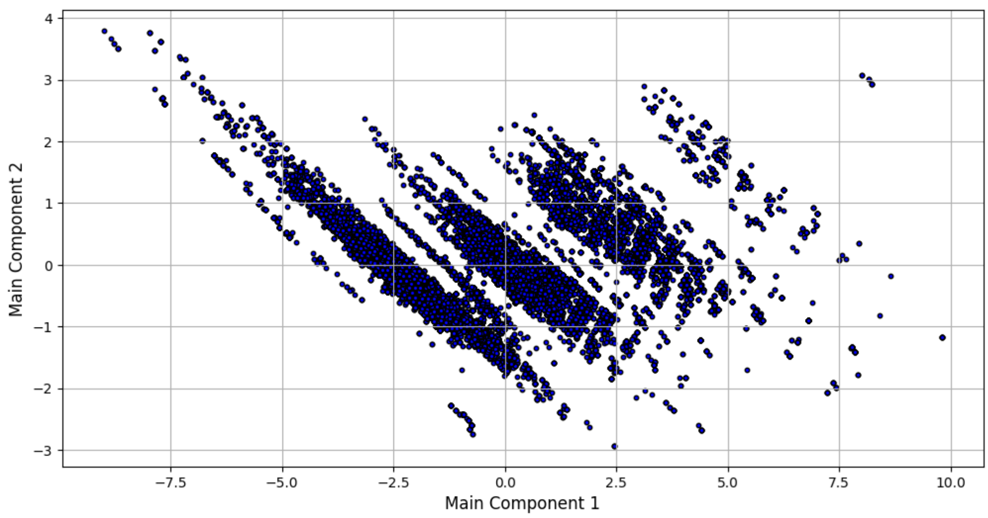

Figure 3. Clustering of Entire Vehicle Dataset (PCA 2D Projection)

Beyond the sample of 15 vehicles, the clustering model was applied to the entire dataset of vehicles from the original source. PCA transformation was again performed, and the K-Means algorithm classified all data points into the same three clusters. Vehicles were plotted in PCA space (**Figure 3**) and color-coded by cluster (**Figure 4**). The majority of vehicles fall into the middle-efficiency group (Cluster 1), while Clusters 0 and 2 contain vehicles that are clearly more or less efficient based on fuel consumption and CO_2 output.

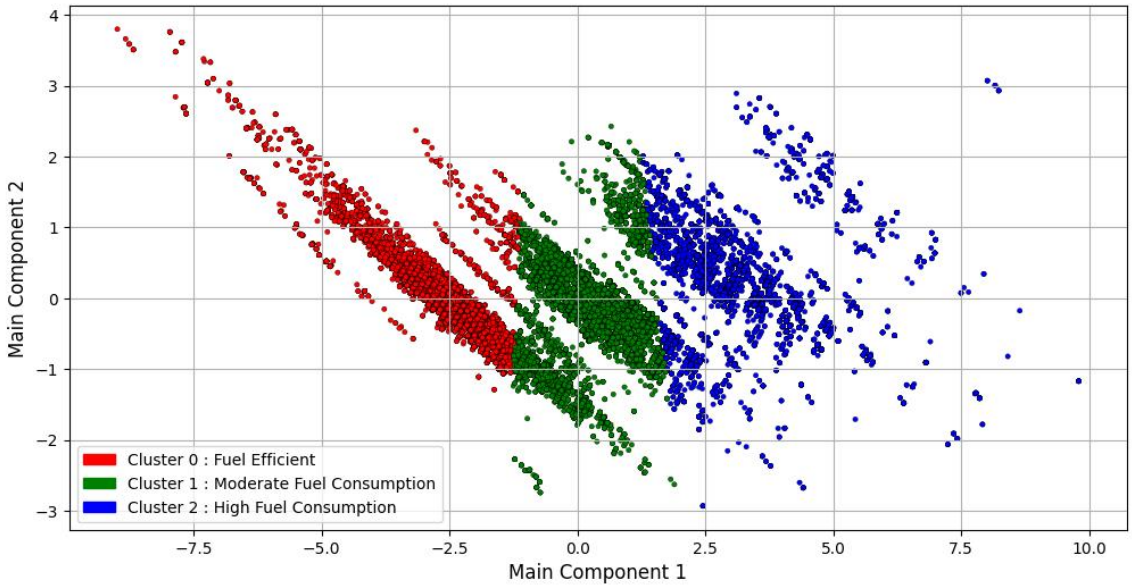

Figure 4. Visualization of K-Means Clustering with PCA of the Entire Data

Table 3 provides a breakdown of the number of vehicles in each cluster. This quantitative insight reveals that although the dataset contains many vehicles with moderate efficiency, a significant number are still categorized as either highly efficient or highly consuming.

Table 3. Number of Vehicles per Cluster (k=3)

Cluster	Total of data	Precentage (%)
0	10773	17
1	15323	35
2	8469	14

3.5. Interpretation and Discussion

The combination of PCA and K-Means has proven to be a robust approach in analyzing high-dimensional vehicle data. PCA effectively reduced data complexity, enabling clearer interpretation and visualization, while K-Means produced meaningful groupings that reflect substantial differences in vehicle performance and environmental impact.

From the clustering results, three distinct groups of vehicles were identified. Cluster 0 represents the most fuel-efficient vehicles with smaller engine displacements, fewer cylinders, and lower CO_2 emissions. These characteristics are commonly found in compact cars or hybrid models designed for urban usage. Such vehicles demonstrate not only economic advantages through reduced fuel costs but also environmental benefits by producing less greenhouse gas emissions. Cluster 1, which accounts for the majority of vehicles, consists of moderately efficient models. These typically represent mid-size sedans and crossovers, balancing performance and efficiency. Cluster 2 includes vehicles with larger engines, higher cylinder counts, and significantly higher CO_2 output. This cluster is dominated by SUVs, luxury vehicles, and performance cars where consumer preferences are oriented toward power rather than efficiency.

When comparing the results to prior studies, similarities and differences can be observed. For example, Salloum et al. (2024) applied PCA-KMeans for tweet classification and reported improved interpretability of clusters, which parallels the clarity gained in this automotive study [17]. Meanwhile, Zhao et al. (2023) focused on predictive models of fuel consumption, whereas our unsupervised approach provides exploratory insights without relying on labels [22]. Furthermore, Tyagi et al. (2022) demonstrated that PCA-KMeans synergy works well for edge computing data, and our findings confirm that this synergy is equally applicable to automotive energy efficiency. The consistency across these studies highlights the generalizability of PCA-KMeans for complex, multidimensional datasets [13].

From a practical perspective, the clusters generated in this study hold valuable implications. For automotive manufacturers, understanding that vehicles in Cluster 0 are favored in sustainability-driven markets can guide investment toward developing compact and hybrid technologies. Cluster 1 vehicles, representing the mainstream market, indicate the need for incremental efficiency improvements to comply with tightening regulations. Vehicles in Cluster 2, however, may face increasing challenges as carbon taxation policies expand worldwide. Manufacturers producing vehicles in this group may need to invest in lightweight materials, advanced combustion technologies, or electrification to remain competitive.

For policymakers, the clustering results provide empirical evidence that regulatory interventions can be targeted more effectively. For instance, subsidies or incentives can be directed toward models falling within Cluster 0, while stricter taxation and emission penalties can be applied to vehicles in Cluster 2. This differentiated approach ensures that regulations are based on actual performance rather than nominal categories such as engine type or vehicle size.

From the consumer standpoint, the clustering visualization offers a simplified decision-making tool. Prospective buyers can easily identify whether a vehicle model aligns with their environmental and economic preferences. For environmentally conscious consumers, Cluster 0 models become an obvious choice, while those prioritizing performance may recognize the trade-offs associated with selecting vehicles in Cluster 2.

Nevertheless, this study has limitations that should be acknowledged. First, the analysis is based on six primary variables, which although critical, do not fully capture the multifaceted nature of vehicle performance. Factors such as maintenance cost, fuel type (e.g., diesel, ethanol, hybrid, or electric), and lifecycle emissions were not included. Second, the clustering algorithm used (K-Means) assumes spherical cluster shapes and may not perfectly capture more complex distributions. Alternative clustering methods such as DBSCAN or hierarchical clustering may provide additional insights. Finally, the dataset primarily reflects vehicles from the North American market, which may limit generalizability to other regions where engine types, fuel standards, and driving behaviors differ.

Future research should expand the dataset to include electric and hybrid vehicles, enabling analysis of emerging trends in sustainable transportation. Incorporating additional attributes such as acceleration performance, safety ratings, or production costs could enhance the robustness of clustering results. Moreover, integrating multi-method approaches (e.g., PCA with fuzzy clustering or neural embeddings) may capture more nuanced patterns in automotive data.

In summary, the interpretation of results confirms that the proposed PCA-KMeans framework is effective in categorizing vehicles based on efficiency and emissions. The expanded discussion underscores its relevance for multiple stakeholders and positions the method as a promising tool for sustainable automotive design and policy-making.

4. Conclusion

This study applied Principal Component Analysis (PCA) and K-Means Clustering to analyze vehicle performance based on six features related to fuel efficiency and emissions. PCA effectively reduced the data's dimensionality, making it easier to visualize and interpret. K-Means grouped the vehicles into three distinct clusters, reflecting different efficiency levels.

Vehicles in the most efficient cluster generally had small engines, fewer cylinders, and lower CO_2 emissions, while high-consumption vehicles showed opposite characteristics. These findings confirm that unsupervised learning methods can reveal meaningful patterns in performance data without needing predefined categories.

The approach offers practical insights for manufacturers, regulators, and consumers. It enables objective classification based on real-world specifications, supporting sustainable engineering efforts. Future improvements may include additional features or adaptation to electric and hybrid vehicles for broader applicability.

References

[1] C. Llopis-Albert, F. Rubio, and F. Valero, "Impact of digital transformation on the automotive industry," *Technological forecasting and social change,* vol. 162, p. 120343, 2021.

[2] S. S. Ravi, S. Osipov, and J. W. Turner, "Impact of modern vehicular technologies and emission regulations on improving global air quality," *Atmosphere,* vol. 14, no. 7, p. 1164, 2023.

[3] A. A. Gheidan, M. B. A. Wahid, O. A. Chukwunonso, and M. F. Yasin, "Impact of internal combustion engine on energy supplyand its emission reduction via sustainable fuel source," 2022.

[4] O. A. Towoju and F. A. Ishola, "A case for the internal combustion engine powered vehicle," *Energy Reports,* vol. 6, pp. 315-321, 2020.

[5] D. L. Bleviss, "Transportation is critical to reducing greenhouse gas emissions in the United States," *Wiley Interdisciplinary Reviews: Energy and Environment,* vol. 10, no. 2, p. e390, 2021.

[6] K. H. D. Tang, "Climate change policies of the four largest global emitters of greenhouse gases: Their similarities, differences and way forward," *Journal of Energy Research and Reviews,* vol. 10, no. 2, pp. 19-35, 2022.

[7] M. A. Saleem, L. Eagle, and D. Low, "Determinants of eco-socially conscious consumer behavior toward alternative fuel vehicles," *Journal of Consumer Marketing,* vol. 38, no. 2, pp. 211-228, 2021.

[8] S. Sibdari and Y. Asayesh, "Eco-drive or tax strive? Assessing consumer vehicle selection behavior under carbon taxation," *Journal of Cleaner Production,* vol. 479, p. 143869, 2024.

[9] Á. N. Choisdealbha, S. Timmons, and P. D. Lunn, "Experimental evidence for the effects of emissions charges and efficiency information on consumer car choices," *Journal of Cleaner Production,* vol. 254, p. 120140, 2020.

[10] V. Pillai, "Data Analytics and Engineering in Automobile Data Systems," *Journal of Science & Technology,* vol. 4, no. 6, pp. 140-79, 2023.

[11] N. L. Rane, M. Paramesha, S. P. Choudhary, and J. Rane, "Machine learning and deep learning for big data analytics: A review of methods and applications," *Partners Universal International Innovation Journal,* vol. 2, no. 3, pp. 172-197, 2024.

[12] M. Chaudhry, I. Shafi, M. Mahnoor, D. L. R. Vargas, E. B. Thompson, and I. Ashraf, "A systematic literature review on identifying patterns using unsupervised clustering algorithms: A data mining perspective," *Symmetry,* vol. 15, no. 9, p. 1679, 2023.

[13] K. Tyagi, C. Rane, R. Sriram, and M. Manry, "Unsupervised learning," in *Artificial intelligence and machine learning for edge computing*: Elsevier, 2022, pp. 33-52.

[14] S. Nanga *et al.*, "Review of dimension reduction methods," *Journal of Data Analysis and Information Processing,* vol. 9, no. 3, pp. 189-231, 2021.

[15] P. Ray, S. S. Reddy, and T. Banerjee, "Various dimension reduction techniques for high dimensional data analysis: a review," *Artificial Intelligence Review,* vol. 54, no. 5, pp. 3473-3515, 2021.

[16] T. M. Ghazal, "Performances of k-means clustering algorithm with different distance metrics," *Intelligent Automation & Soft Computing,* vol. 30, no. 2, pp. 735-742, 2021.

[17] S. Salloum, K. Alhumaid, A. Salloum, and K. Shaalan, "K-means Clustering of Tweet Emotions: A 2D PCA Visualization Approach," *Procedia Computer Science,* vol. 244, pp. 30-36, 2024.

[18] S. M. Miraftabzadeh, C. G. Colombo, M. Longo, and F. Foiadelli, "K-means and alternative clustering methods in modern power systems," *Ieee Access,* vol. 11, pp. 119596-119633, 2023.

[19] S. Bandyopadhyay, S. Thakur, and J. Mandal, "Product recommendation for e-commerce business by applying principal component analysis (PCA) and K-means clustering: benefit for the society," *Innovations in Systems and Software Engineering,* vol. 17, no. 1, pp. 45-52, 2021.

[20] P. S. Bishnu and A. D. Purkayastha, "From Data to Decisions: K-Means Clustering for Product Development and Marketing," in *AI for Product Development*: CRC Press, pp. 20-41.

[21] S. Kumar, M. Suhaib, and M. Asjad, "Narrowing the barriers to Industry 4.0 practices through PCA-Fuzzy AHP-K means," *Journal of Advances in Management Research,* vol. 18, no. 2, pp. 200-226, 2021.

[22] D. Zhao *et al.*, "A review of the data-driven prediction method of vehicle fuel consumption," *Energies,* vol. 16, no. 14, p. 5258, 2023.

[23] Y. Yao *et al.*, "Vehicle fuel consumption prediction method based on driving behavior data collected from smartphones," *Journal of Advanced Transportation,* vol. 2020, no. 1, p. 9263605, 2020.