Cureus Journal of Computer Science

Open Access Original Article

Clustering of Commercial Vehicles Based on Fuel Type Using Machine Learning Technique

Aditya S. Baheti [1] , Ankush D. Sawarkar [1] , Anurag Agrahari [2] , Shital Y. Gaikwad [3]

1. Department of Information Technology, Shri Guru Gobind Singhji Institute of Engineering and Technology (SGGSIE&T), Nanded, IND 2. Department of Computer Science and Engineering, Visvesvaraya National Institute of Technology (VNIT), Nagpur, IND 3. Department of Computer Science and Engineering, Mahatma Gandhi Mission's College of Engineering and Technology, Nanded, IND

Corresponding authors: Aditya S. Baheti, 2021bit002@sggs.ac.in, Ankush D. Sawarkar, adsawarkar@sggs.ac.in

Review began 08/18/2024
Review ended 10/29/2024
Published 11/06/2024

DOI: 10.7759/13

Abstract

This paper delves into the exploration of vehicle characteristics through various clustering algorithms such as Density-Based Spatial Clustering of Applications with Noise, Gaussian Mixture Model, and K-means clustering analysis, focusing primarily on fuel-related attributes. The research problem revolves around uncovering inherent patterns within automobile data to better understand the relationship between different features and fuel types. The methodology involves preprocessing the dataset, which includes fixing missing values errors and encoding categorical variables, followed by scaling the features and applying three clustering algorithms. The Elbow Method is utilized to get the efficient number of clusters, and to find the accuracy of the models, three different evaluation metrics are computed to know the clustering quality. Key findings reveal distinct clusters of vehicles based on fuel-related attributes, providing insights into fuel efficiency and usage patterns across different vehicle models. The analysis is visualized through scatter plots, silhouette scores, the Davies-Bouldin Index, and the Calinski-Harabasz Index with values of 0.528, 0.66, and 57730.245 highlighting the effectiveness of K-means clustering in uncovering meaningful clusters within the dataset. In conclusion, this study demonstrates the utility of clustering analysis in extracting valuable insights from vehicle data, with implications for fuel efficiency optimization and market segmentation in the automotive industry.

Categories: AI applications, Data Engineering, Machine Learning (ML)
Keywords: automobile data, k-means clustering, fuel attributes, cluster analysis, gaussian mixture model

Introduction

Regression analysis in machine learning is a powerful technique used in predictive modeling to understand the relationship between independent variables and a continuous target variable. Unlike classification, which predicts discrete labels, regression aims to predict a numerical outcome. This concept is crucial in various fields, such as marketing, healthcare, and engineering, where making accurate predictions is necessary for decision-making processes.

Regression algorithms, such as linear regression and neural networks, analyze historical data to identify patterns and make forecasts about future outcomes. By referencing past observations, these models can forecast trends, estimate values, and optimize processes, leading to improved decision-making and resource allocation [1]. These models are trained on historical data, and by learning from the relationships within the data, they can provide insights that help in predicting future trends, thereby facilitating strategic planning and resource management.

The paper addresses the challenge of understanding and segmenting vehicles based on their fuel-related attributes. Specifically, it aims to identify distinct clusters of vehicles within a dataset using K-means clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and Gaussian Mixture Models (GMMs), focusing on features such as mileage, model, gear type, price, horsepower, year, and offer type. The problem statement revolves around uncovering underlying patterns in automobile data to gain insights into fuel efficiency and usage patterns across different vehicle models [2]. This segmentation can help in identifying which types of vehicles are more fuel-efficient, thereby providing valuable information for manufacturers, consumers, and policymakers aiming to promote sustainable practices and improve fuel economy.

The paper begins with an introduction to machine learning regression and its importance in predictive modeling. It then defines the specific problem of segmenting vehicles based on fuel-related attributes. The methodology section outlines the steps involved in preprocessing the dataset, including handling missing values, normalizing the data, and feature scaling, followed by the application of K-means clustering, DBSCAN, and GMM algorithms. The evaluation of clustering quality is done using metrics such as silhouette score, the Davies-Bouldin index (DBI), and cluster validation techniques. Results are presented through visualizations like scatter plots, dendrograms, and heatmaps, along with quantitative measures to assess the effectiveness of each clustering method. This is followed by a discussion of the results, highlighting key

How to cite this article
Baheti A S, Sawarkar A D, Agrahari A, et al. (November 06, 2024) Clustering of Commercial Vehicles Based on Fuel Type Using Machine Learning Technique. Cureus J Comput Sci 1 : e13. DOI 10.7759/13

findings and their implications. Finally, the paper concludes with insights into clusters formation for fuel efficiency optimization and suggests directions for future research [3].

Existing research on machine learning clustering has extensively explored various algorithms and methodologies, including K-means clustering, DBSCAN, and GMMs, each offering unique approaches to grouping data points based on their similarities (Table *1*). These techniques have found applications in diverse fields such as market segmentation, image and speech recognition, bioinformatics, and anomaly detection. Despite significant advancements, current literature often highlights challenges in handling high-dimensional data, selecting the optimal number of clusters, and ensuring clustering stability and scalability in large datasets. The present research aims to address these gaps by enhancing the efficiency and accuracy of clustering algorithms, particularly in preprocessing and feature scaling. Additionally, it focuses on developing robust methods for determining the appropriate number of clusters in complex datasets, thus contributing to more reliable and interpretable clustering outcomes.

Moreover, the research delves into the comparative analysis of the clustering algorithms used, assessing their performance in the context of vehicle data segmentation. This includes a detailed examination of each algorithm's strengths and limitations, as well as the impact of various preprocessing techniques on clustering results. By doing so, the study aims to provide a comprehensive understanding of how different clustering approaches can be effectively applied to automotive data, thereby advancing the field of machine learning clustering in practical, real-world applications.

Materials And Methods

Authors	Methodology	Key Findings	Ref.
Henrik Almér (2015)	Assesses strategies of machine learning and measurable tests for anticipating fuel utilization in overwhelming vehicles. The thought is to utilize verifiable information depicting driving circumstances to foresee a fuel utilization in liters per distance.	An inspecting rate of 10 minutes is superior to a testing rate of 1 diminutive for foreseeing fuel utilization measured in liters per distance. Street incline, vehicle speed, and vehicle weight are the foremost compelling parameters for foreseeing fuel utilization. The irregular woodland, SVM and ANN models create comparable results for the problem statement.	[1]
Rahbari et al. (2007)	Proposes the arrangement to the issue that the introductory point of clustering calculation is simple to drop into native, ideal, and slow process. An enhanced combination and calculation of the foremost component evaluation and weighted K-means clustering is proposed. The result presents the greatest and least separate, weighted distance, starting from the cruel entirety of the separations of the other clustering points, maintaining a strategic distance from the impact of exceptions and edge data.	This research states a progressed calculation for the fusion of vital parts and feature-weighted K-means clustering and presents the leftover point clustering mean method to dispense with outliers and reduce the process time. By strategically placing initial centers, K-means avoids local traps and achieves better clustering. Agreeing to the assurance rate of the eigenvalue offering determinant to the cluster, the primary feature pressure is acquired, and a weighted Euclidean distance rhythmical is projected.	[2]
Suroto Munahar et al. (2023)	Proposes the demonstrate utilizing machine learning calculation for fuel savings in cars to decrease the contamination rates and worldwide warming and creating the vehicle innovation for the same.	The fuel administration cluster planned to be tracked utilizing AFR cluster with the table mapping which has been effectively created on vehicles with gasoline motors depending on the driving behavior profile.	[3]
Shuping Xu et al. (2022)	The study utilized an improved K-means clustering algorithm to analyze the relationship between driving conditions and fuel consumption. The author collected data on various driving parameters, such as speed, terrain, and traffic conditions, and applied the enhanced algorithm to cluster the data for more accurate fuel consumption predictions under different conditions.	The improved K-means algorithm effectively identified distinct clusters of driving conditions, revealing significant variations in fuel consumption across different scenarios. The findings indicate that the algorithm enhances the understanding of how specific driving conditions affect fuel efficiency, providing insights for optimizing driving behavior and vehicle performance.	[4]
Shi Na et al. (2010)	The authors developed an improved K-means clustering algorithm to address issues like initial centroid sensitivity and slow convergence. They enhanced the initialization phase and evaluated the algorithm's performance against traditional K-means using metrics such as clustering accuracy and computational efficiency.	The improved K-means algorithm exhibited higher clustering accuracy and faster convergence than the traditional method. The study demonstrated that better initial centroid selection and refined distance calculations lead to more effective clustering in complex datasets.	[5]
Gordon O. Ewing et al. (1998)	The study analyzed the impact of travel demand management (TDM) strategies and economic incentives on consumer fuel-type choice for vehicles. Using econometric models, the author examined how factors such as fuel prices, taxes, and TDM policies influence the decision to choose fuel-efficient or alternative fuel vehicles.	The research found that economic incentives, such as fuel taxes and subsidies for alternative fuel vehicles, significantly affect consumer fuel-type choices. TDM strategies, when combined with economic incentives, were shown to encourage a shift toward more fuel-efficient vehicle types, supporting broader environmental and energy-saving goals.	[6]
Victor Parque et al. (2011)	The study proposed a machine learning-based approach to predict fuel consumption in vehicle clusters, focusing on grouping vehicles with similar consumption patterns. Various clustering algorithms were used to segment vehicles, followed by predictive models trained on each cluster to enhance the accuracy of fuel consumption predictions.	The research demonstrated that clustering vehicles with similar attributes improves the accuracy of fuel consumption predictions compared to non-clustered approaches. The approach allows for more targeted and precise prediction models, showing potential for optimizing fuel usage and improving vehicle design based on consumption patterns.	[7]

TABLE 1: Summary of work done for clustering of vehicles based on fuel type using machine learning

The dataset utilized in this study is sourced from Kaggle and specifically focuses on German cars[8]. It encompasses various attributes related to vehicle characteristics and performance. The dataset includes features such as mileage, model, fuel type, gear type, price, horsepower, year of manufacture, and offer type (Figure *1*). These attributes provide a comprehensive overview of different aspects of the vehicles, facilitating the analysis of fuel-related patterns and clustering. Prior to analysis, preprocessing steps were undertaken, including fixing missing values errors and encoding absolute variables, to ensure data quality and consistency. Overall, this dataset serves as an important information for exploring fuel efficiency and usage patterns across different German car models through clustering analysis.

Data preprocessing represents a fundamental phase in preparing the dataset for clustering analysis,

especially when examining vehicle characteristics. The initial step encompasses data cleaning, which involves addressing missing values, identifying and removing duplicates, and ensuring the overall integrity of the dataset. Missing values can significantly compromise the effectiveness of clustering algorithms; therefore, employing appropriate imputation strategies, such as utilizing the mean or median for numerical variables, is essential. Furthermore, categorical variables such as make, model, fuel, and offer type require encoding to transform them into numerical formats. Techniques such as one-hot encoding are frequently employed to ensure that these categorical variables can be effectively utilized in the clustering process without introducing bias [9].

Normalization and feature engineering further augment the dataset's usability. Numerical features, including mileage, price, horsepower (hp), and year, benefit from scaling methods, such as Min-Max scaling or Z-score standardization, to ensure that they are on comparable scales. This step is critical for clustering algorithms, which may otherwise be skewed by the varying ranges of these features. Additionally, the creation of new features, such as the vehicle's age, can provide deeper insights into the data. By executing these preprocessing steps, the dataset becomes more structured and informative, enabling more accurate clustering and meaningful pattern recognition. This is vital for comprehensively understanding fuel-related attributes and their implications for vehicle efficiency and market segmentation [9].

	mileage	make	model	fuel	gear	offerType	price	hp	year
0	235000	BMW	316	Diesel	Manual	Used	6800	116.0	2011
1	92800	Volkswagen	Golf	Gasoline	Manual	Used	6877	122.0	2011
2	149300	SEAT	Exeo	Gasoline	Manual	Used	6900	160.0	2011
3	96200	Renault	Megane	Gasoline	Manual	Used	6950	110.0	2011
4	156000	Peugeot	308	Gasoline	Manual	Used	6950	156.0	2011
...	...	...	...	...	...	...	...	...	...
46400	99	Fiat	500	Electric/Gasoline	Manual	Pre-registered	12990	71.0	2021
46401	99	Fiat	500	Electric/Gasoline	Manual	Pre-registered	12990	71.0	2021
46402	99	Fiat	500	Electric/Gasoline	Manual	Pre-registered	12990	71.0	2021
46403	99	Fiat	500	Electric/Gasoline	Manual	Pre-registered	12990	71.0	2021
46404	99	Fiat	500	Electric/Gasoline	Manual	Pre-registered	12990	71.0	2021

FIGURE 1: The dataset utilized in the model development

The selection of clustering algorithms in this study is based on several key criteria: data characteristics, algorithm performance, and interpretability. K-means, DBSCAN, and GMMs were chosen due to their distinct strengths and suitability for different clustering needs.

K-means is a widely utilized clustering algorithm that operates by assigning data points to the nearest centroid and iteratively updating these centroids to reflect the mean of the assigned points. The objective of the K-means algorithm is to minimize the variance within each cluster, thereby reducing data point diversity and scatter within groups. This process involves continuously reassigning data points to their nearest centroid and recalculating the centroid positions based on the updated assignments. Although K-means is computationally efficient for datasets with a smaller number of rows due to its iterative nature, it is sensitive to the initial placement of centroids and requires the specification of the number of clusters in advance. Despite these limitations, K-means remains a widely favored technique in various applications, including pattern recognition, image segmentation, and customer segmentation, owing to its simplicity and overall effectiveness [10] (Figure *2*).

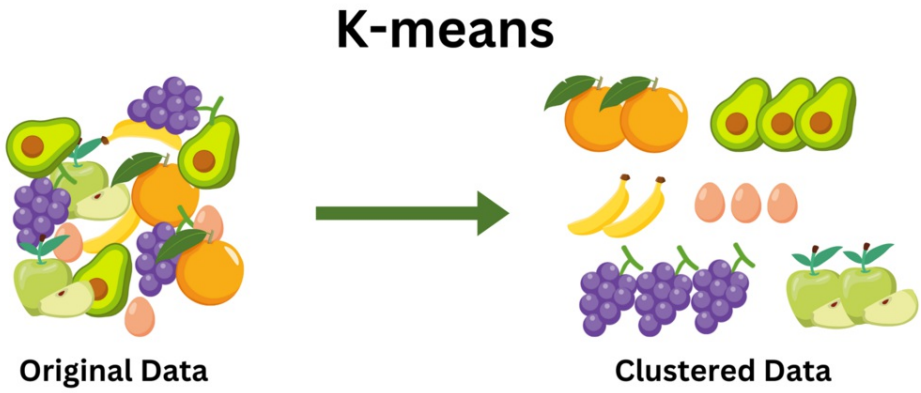

FIGURE 2: Working of K-means clustering

The GMM is a probabilistic clustering algorithm that assumes data points are generated from a mixture of several Gaussian distributions, each representing a cluster. Instead of assigning each data point to a single cluster, GMM computes the probability that a point belongs to each cluster, allowing for more flexible cluster assignments. This makes GMM particularly useful for datasets where clusters may overlap or have more complex shapes. The algorithm iteratively estimates the parameters of the Gaussian distributions using the Expectation-Maximization (EM) algorithm, which adjusts both the means and covariances of the distributions to fit the data better [11].

DBSCAN is a clustering algorithm that groups together point that are closely packed together while marking points in low-density regions as outliers. Unlike K-means and GMM, DBSCAN does not require a predetermined number of clusters; instead, it identifies clusters based on the density of data points in the feature space. The algorithm uses two parameters: the radius (epsilon) that defines the neighborhood around each point and the minimum number of points required to form a dense region [12].

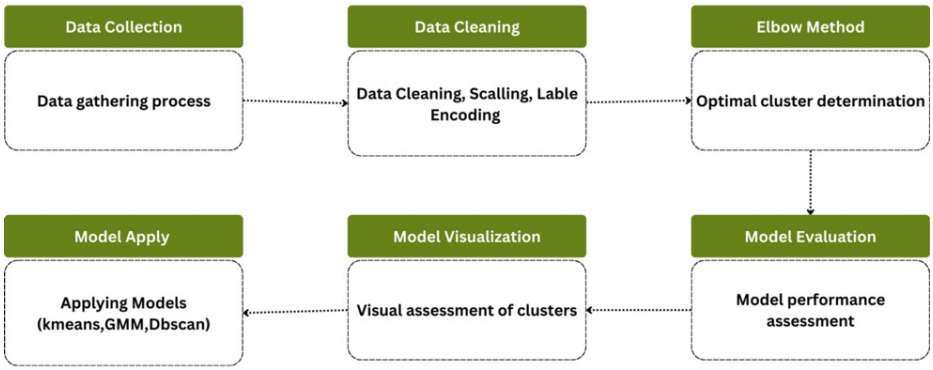

FIGURE 3: Workflow of Model

In this study, K-means was primarily used to establish a baseline due to its straight forward implementation and the ability to easily predict the number of clusters using the Elbow Method (Figure *3*). The study also employed DBSCAN to explore the dataset's density-based clustering properties, which can uncover more nuanced structures, and noise points that K-means might miss. GMM was included to provide a probabilistic clustering approach, allowing for a more refined understanding of the dataset by modeling the data distribution more accurately (Figure *4*). The use of these three algorithms offers a comprehensive analysis, leveraging the strengths of each method to gain deeper insights into the underlying patterns in the automobile dataset [1,2].

	mileage	price	hp	year
count	4.640500e+04	4.640500e+04	46376.000000	46405.000000
mean	7.117786e+04	1.657234e+04	132.990987	2016.012951
std	6.262531e+04	1.930470e+04	75.449284	3.155214
min	0.000000e+00	1.100000e+03	1.000000	2011.000000
25%	1.980000e+04	7.490000e+03	86.000000	2013.000000
50%	6.000000e+04	1.099900e+04	116.000000	2016.000000
75%	1.050000e+05	1.949000e+04	150.000000	2019.000000
max	1.111111e+06	1.199900e+06	850.000000	2021.000000

FIGURE 4: The descriptive analysis of dataset

The evaluation of clustering models in this research employs three primary measures: silhouette score, DBI, and Calinski-Harabasz Index (CHI) (equations 1,2, and 3). The silhouette score evaluates how alike an object is to its cluster when differentiated from other clusters, with values from –1 to 1. A higher silhouette score shows that the data points are like their own cluster and different to other clusters, signifying better-defined clusters [13]. This metric provides an intuitive way to understand the cohesion within clusters and the separation between them.

The DBI calculates clustering nature by considering the ratio of own-cluster scatter to other-cluster separation. The lower values of this indicate efficient clustering performance, as it shows that clusters are small and well-differentiated. The CHI, also known as the Variance Ratio Criterion, assesses the fraction of the sum of other-cluster dispersion to own-cluster dispersion. Higher values of this imply that clusters are dense and distinct from one another. Together, these metrics offer a comprehensive assessment of clustering performance by balancing the evaluation of internal cluster cohesion and external cluster separation. This multi-faceted approach ensures a robust evaluation of the clustering models applied to the automobile dataset [13].

\begin{equation} Silhouette Score = \frac{(b(i)-a(i))}{\mathit{max}(b(i),a(i))} \qquad (1)\end{equation}

Where, in equation 1 b is the mean distance of nearest cluster and a is mean distance of intracluster.

\begin{equation}DBI= \frac{1}{K}\sum_{i=1}^{k} \mathit{max}_{i\neq j}{R}_{ij} \qquad (2)\end{equation}

Where, in equation 2 DBI is Davies-Bouldin Index, K is number of clusters and dataset values R.

\begin{equation}CHI= \sum \frac{(O(i)-E(i))}{E_i } \qquad (3)\end{equation}

Where, in equation 3 CHI is Calinski-Harabasz Index that is ratio of sum of intra cluster distance and sum of inter cluster distance.

The clustering models were implemented using Python language with various software libraries for data preprocessing and modeling. Specifically, the scikit-learn library was utilized for implementing clustering algorithms. Other libraries, including Seaborn, Matplotlib, and Yellowbrick, were employed for data manipulation, visualization, and evaluation of the models. Computational resources used for the implementation include a standard laptop or desktop computer with minimum required processing power and memory to process the dataset and perform the required computations. The implementation process involved loading the dataset, preprocessing it to handle missing values and encode categorical variables, applying the regression algorithm, evaluating the model's performance using appropriate metrics, and visualizing the results to interpret the clustering outcomes. Overall, the implementation leveraged widely used software libraries in the Python ecosystem, making it accessible and reproducible for further analysis and experimentation [9].

Clustering models i.e. K-means, DBSCAN, and GMMs require selecting certain parameters before training. For K-means, the number of clusters (K) is a key parameter, while DBSCAN relies on the distance threshold (epsilon) and minimum samples to form clusters. GMMs use the number of components to model the underlying data distribution.

In this study, the Elbow Method was applied to K-means to determine the optimal number of clusters by visually inspecting the inertia or within-cluster sum of squares for different values of K (Figure *5*) [13]. For DBSCAN, the focus was on selecting appropriate values for epsilon and minimum samples, guided by resulting cluster quality measures such as silhouette score and DBI. Similarly, for GMM, the number of components was adjusted to capture the data distribution effectively.

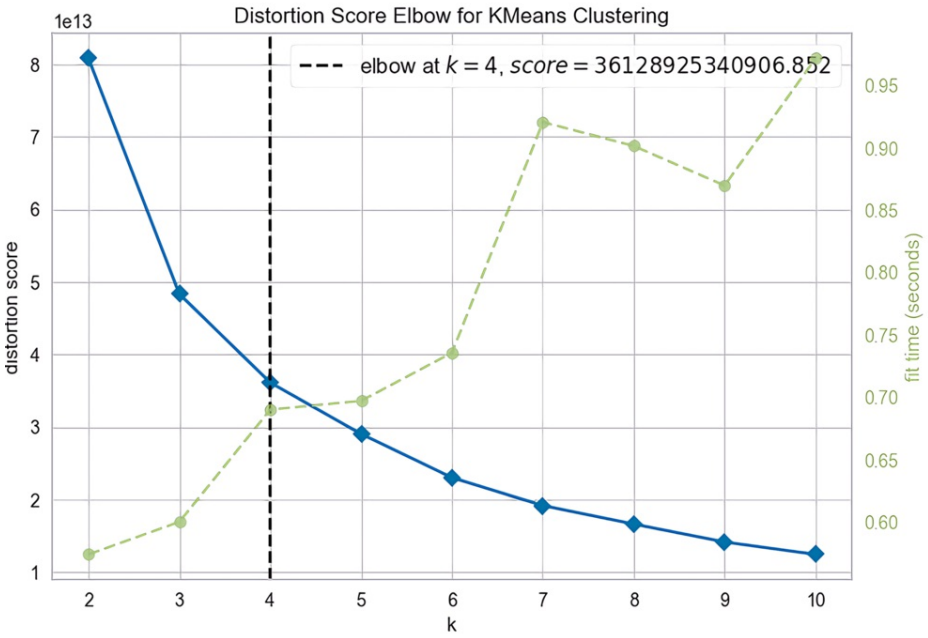

FIGURE 5: Elbow method for defining the number of optimal clusters

By evaluating the models' performance using these methods, the study aimed to identify configurations that provide accurate and meaningful clustering results for the automobile dataset [8].

\begin{equation} WCSS= \sum_{i=1}^{k}\sum_{x\in C_{i}} \left \| x - \upsilon_{i} \right \|^{2} \qquad (4) \end{equation}

Where, in equation 4 WCSS stands for Within Cluster sum of Squares, i.e. sum of variance between observations in each cluster.

Results

The output of the experiments reveals insights into the conduct of the K-means clustering in segmenting vehicles based on fuel-related attributes. Performance parameters such as the silhouette score, the CHI, and DBI were utilized to calculate the quality of the clustering, providing a measure of the separation between clusters (Table *2*) [14]. Additionally, visualizations such as scatter plots were used to visually inspect the clustering outcomes and assess their interpretability (Figure *6*). The Elbow Method was employed to find the minimum required clusters, enhancing the effectiveness of the clustering solution. Comparative analysis of different clustering models highlighted the strengths and limitations of each approach, enabling a comprehensive calculation of their performance in capturing the patterns in the dataset [8]. Overall, the results demonstrate the utility of K-means clustering in uncovering meaningful clusters within the dataset, providing valuable insights into fuel types across different vehicle models, with a silhouette score of 0.528, a DBI value of 0.66, and a CHI of 57730.245, giving the best performance among all the algorithms implemented for the study [14,15]. Histogram plot (Figure *7*) shows the comparison of the silhouette score of different models implemented in the study.

In contrast, the GMM yielded a silhouette score of 0.025454, indicating that the clusters were not well-defined, which may suggest overlap between them. The DBI value of 1.8242880 was relatively low, pointing to some degree of separation between the identified clusters, while the CHI value of 7216.763074 indicated limited clustering effectiveness. These results highlight GMM's challenges in accurately segmenting vehicles

based on fuel-related attributes in this dataset [11,14].

Similarly, DBSCAN produced a silhouette score of –0.55684, indicating that the clusters were not distinctly separated, suggesting that many points were either misclassified or considered noise. The DBI value of 0.8535698 showed that the clusters had some degree of overlap, and the CHI value of 8.745663 further confirmed the limited effectiveness of the clustering. Despite these challenges, DBSCAN's ability to identify outliers in the dataset provided valuable insights into vehicles that did not conform to the typical patterns, contributing to a deeper understanding of fuel efficiency and usage across various vehicle models [12,14].

Model	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
k-means	0.528656	0.6623691	57730.24512
GMM	0.025454	1.8242880	7216.763074
DBSCAN	-0.55684	0.8535698	8.745663

TABLE 2: Evaluation results of models implemented

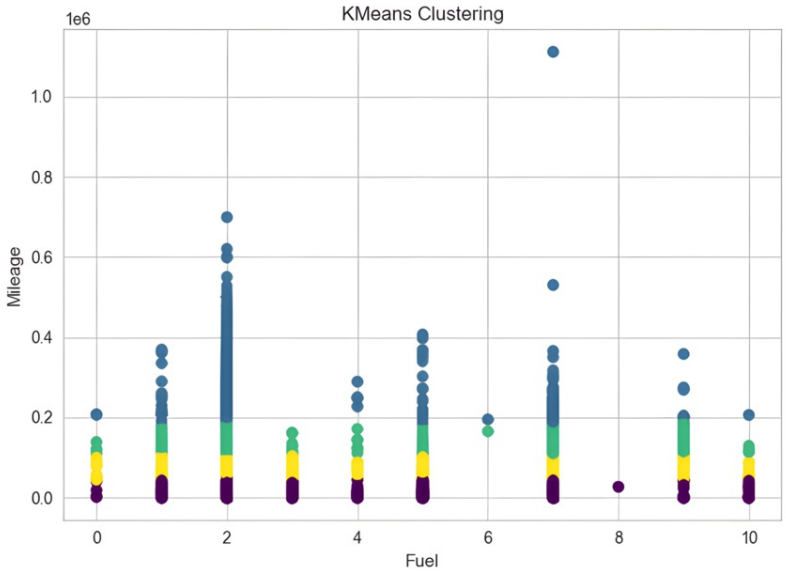

FIGURE 6: Cluster formation 2D representation

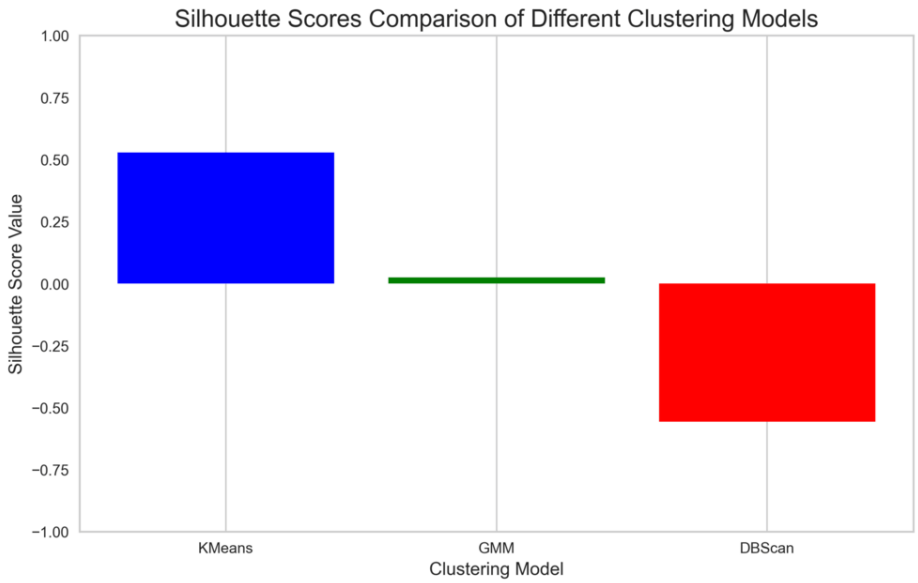

FIGURE 7: Comparative analysis of results of all the models implemented

Discussion

The outcoming of the results underscore the importance of K-means clustering in achieving the study goal of understanding vehicle characteristics and fuel-related attributes. By effectively segmenting vehicles based on fuel type, the clustering analysis offers meaningful insights for various company owners in the automotive industry [2]. These insights can inform strategic decision-making processes related to vehicle design, marketing, and resource management. Additionally, the clustering results can facilitate targeted interventions aimed at improving fuel types and their efficiency or demand and sustainability practices. The outcomes of the analysis validate the hypothesis that K-means clustering can uncover meaningful patterns within automotive data, providing a foundation for further research and practical applications in optimizing fuel usage and vehicle performance. Overall, the results contribute to advancing information in the field of automotive research and underscore the importance of data-driven approaches in addressing complex challenges related to fuel efficiency and sustainability [4]. Visualizations such as 3D scatter plots were used to visually represent the clustering outcomes given by the model (Figure *8*).

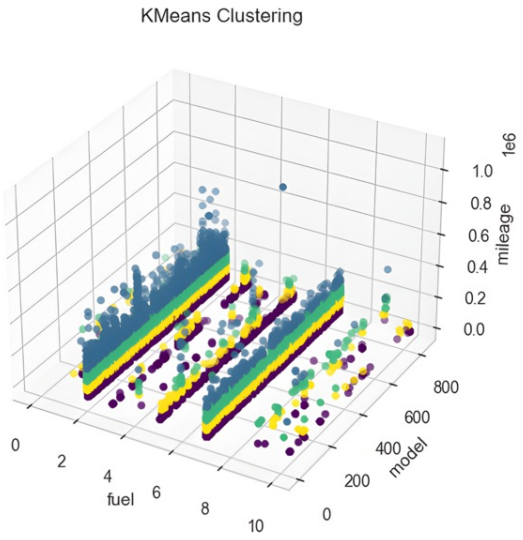

FIGURE 8: Clusters formation 3D representation

Comparison of the findings with those of previous studies reveals both similarities and differences in the approach and outcomes. Like previous research, the current study applies regression techniques to analyze vehicle characteristics and fuel-related attributes, focusing on clustering vehicles based on fuel efficiency and usage patterns [1]. However, the novelty lies in the specific methodology employed, including the use of K-means clustering and the comprehensive evaluation of clustering performance using metrics such as silhouette score and visualization techniques [14]. Previous studies have explored regression models for predicting vehicle prices or fuel consumption. However, the current study uniquely contributes by providing insights into vehicle segmentation based on fuel-related attributes. Additionally, the use of cross-validation and comparative analysis enhances the reliability and validity of the findings, offering a more robust assessment of the clustering outcomes. Overall, while there may be similarities in the research objectives and methodologies, the specific findings and contributions of the current study add value to the existing literature by offering a nuanced understanding of fuel efficiency and usage patterns in the automotive domain.

This study makes significant contributions to the field of machine learning regression by showcasing the application of K-means clustering, DBSCAN, and GMM in analyzing vehicle characteristics and fuel-related attributes. By leveraging regression techniques and advanced clustering methodologies, the research offers a novel approach to segmenting vehicles based on fuel efficiency and usage patterns [2,4,6]. The use of performance metrics, visualization techniques, and cross-validation enhances the reliability and interpretability of the findings, providing valuable insights for decision-making processes in the automotive industry. Additionally, the study contributes to advancing knowledge in cluster modeling by highlighting the importance of robust preprocessing, evaluation techniques in ensuring the effectiveness of clustering solutions.

While the study demonstrates better results, several drawbacks should be acknowledged. First, the analysis relies on a single dataset, which may not completely capture the diversity of vehicle characteristics and fuel-related attributes across different regions or time periods. Moreover, the preprocessing steps, such as handling NaN values and encoding categorical variables, could introduce biases or errors that impact the clustering outcomes. Additionally, the choice of hyperparameters and the initialization method for K-means clustering may influence the results and could gain profit from further studies and sensitivity analysis based on current results given by the model. Furthermore, the evaluation of clustering performance is primarily based on quantitative metrics and visualizations, which may not consider all aspects of the clustering quality. Future studies could address these limitations by incorporating multiple datasets, refining preprocessing techniques, exploring alternative clustering algorithms, and integrating qualitative assessments to provide a more comprehensive understanding of vehicle segregation based on fuel-related attributes.

Based on the findings of this research, there are many potential avenues for further research. First, exploring the effect of additional features or variables, such as vehicle weight, engine type, or driving conditions, could provide a more comprehensive understanding of fuel efficiency and usage patterns. Additionally, investigating the temporal dynamics of fuel-related attributes and considering factors such as technological advancements or policy changes could make some insights into long-term trends and patterns. Moreover, integrating enhanced machine learning techniques, such as deep learning methods or ensemble learning methods, may further improve the accuracy of vehicle segmentation models. Furthermore, extending the analysis to incorporate geographical variations or market dynamics could uncover region-specific patterns and inform focused topics for improving fuel efficiency and sustainability practices. Overall, future study directions should focus to expand the scope and depth of analysis while addressing the limitations identified in this study to advance knowledge in the field of machine learning regression and automotive research.

Conclusions

The main findings of the study highlight the effectiveness of K-means clustering in segmenting vehicles based on fuel-related attributes. Through comprehensive preprocessing, clustering, and evaluation techniques, meaningful clusters of vehicles were found, giving insights into fuel efficiency and usage patterns across different models. The Elbow Method facilitated the calculation of the efficient number of clusters, enhancing the robustness of the clustering solution. Performance metrics such as the silhouette score with the value of 0.522 provide the model with a better range of performance for giving the outputs, and visualizations further validated the clustering outcomes, demonstrating their interpretability and reliability. The results contribute to improving the understanding of vehicle characteristics and offer practical implications for optimizing fuel usage and sustainability practices in the automotive industry. Overall, the research underscores the value of data-driven approaches in uncovering actionable insights and informing decision-making processes related to fuel type and vehicle performance.

Appendices

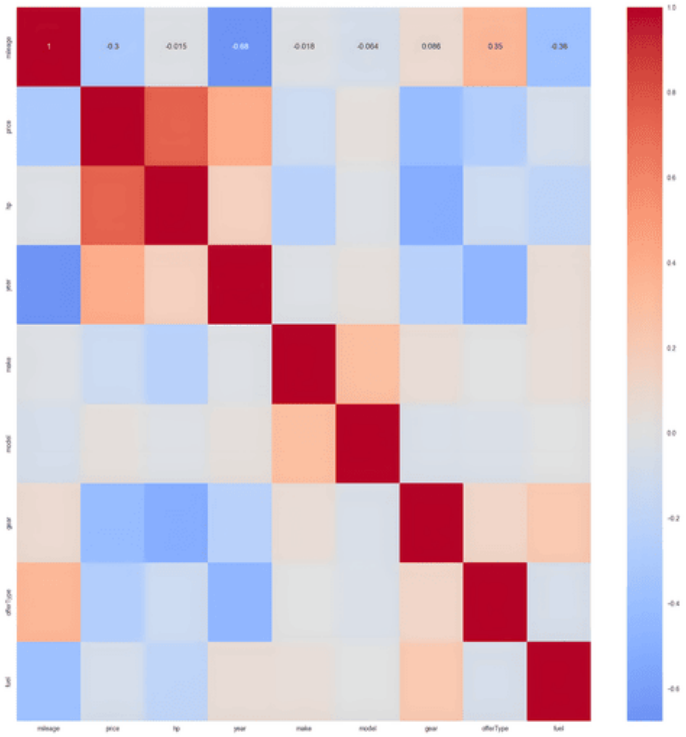

FIGURE 9: Correlation matrix

Figure 9 is the representation of correlation matrix of the various features in the dataset to identify which feature performs best with another feature to produce the desired output.

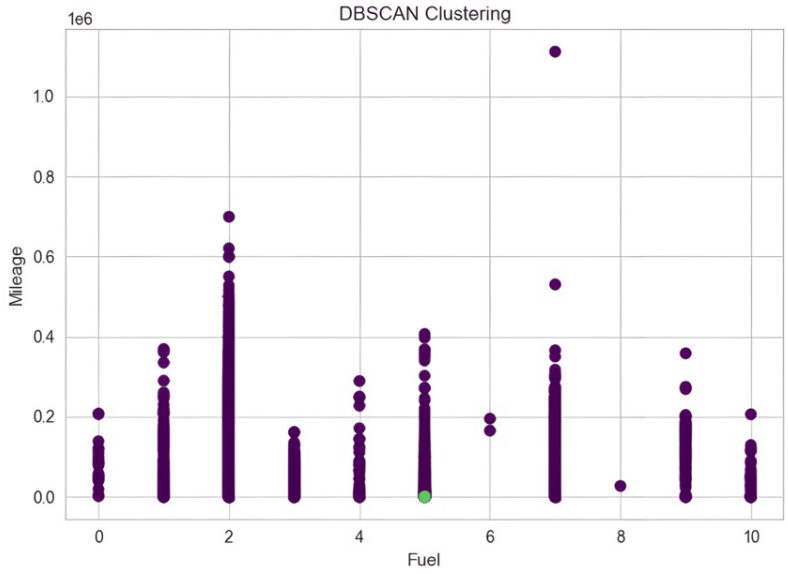

FIGURE 10: DBSCAN clustering 2D model

Figure 10 shows the 2D representation of clusters formed by the DBSCAN model, which exhibits significant overlap between the clusters, making it inefficient to differentiate vehicles based on fuel type using this model.

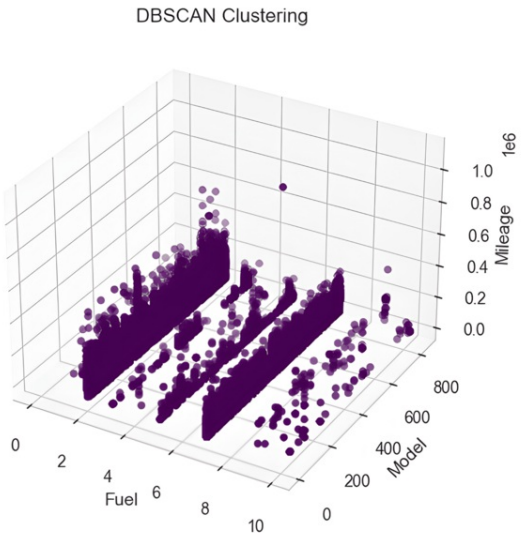

FIGURE 11: DBSCAN clustering 3D model

Figure 11 shows the 3D representation of clusters formed by DBSCAN model, which exhibits significant overlap between the clusters, making it inefficient to differentiate vehicles based on fuel type using this model.

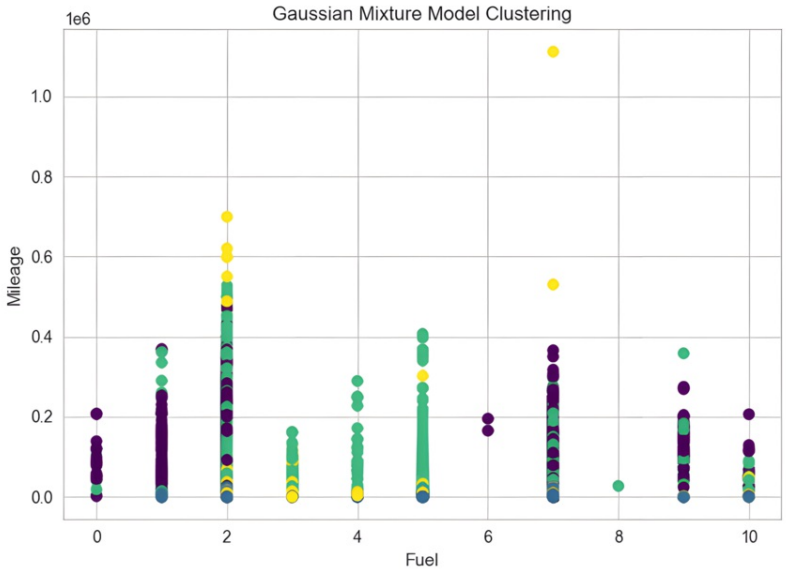

FIGURE 12: Gaussian Mixture Model 2D Model

Figure 12 shows the representation of the clusters formed by Gaussian Mixture Model implemented in this study to differentiate vehicles based on fuel type. However, the outcome of the model is less effective compared to the K-Means clustering model.

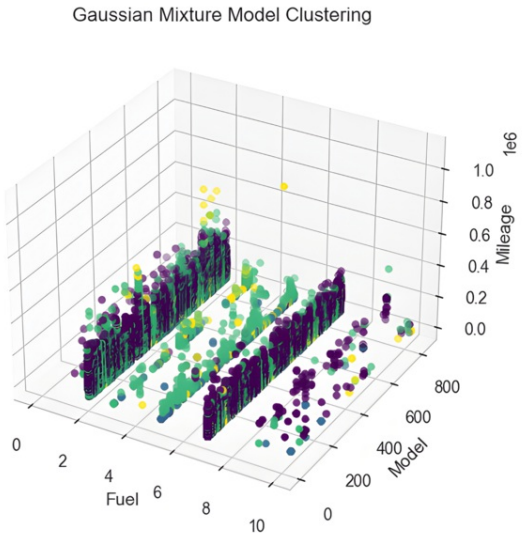

FIGURE 13: Gaussian Mixture Model 3D Model

Figure 13 shows the representation of the clusters formed by Gaussian Mixture Model implemented in this study to differentiate vehicles based on fuel type. However, the outcome of this model is less effective compared to the K-Means clustering model.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Ankush D. Sawarkar, Shital Y. Gaikwad

Acquisition, analysis, or interpretation of data: Ankush D. Sawarkar, Aditya S. Baheti, Shital Y. Gaikwad, Anurag Agrahari

Drafting of the manuscript: Ankush D. Sawarkar, Aditya S. Baheti, Shital Y. Gaikwad

Critical review of the manuscript for important intellectual content: Ankush D. Sawarkar, Aditya S. Baheti, Shital Y. Gaikwad, Anurag Agrahari

Supervision: Ankush D. Sawarkar, Shital Y. Gaikwad, Anurag Agrahari

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

The authors are thankful to the Director, Shri Guru Gobind Singhji Institute of Engineering and Technology (SGGSIE&T), Nanded, India, for providing the necessary facilities for this work.

References

1. Almér H: Machine Learning and Statistical Analysis in Fuel Consumption Prediction for Heavy Vehicles. KTH School of Computer Science and Communication (CSC), Stockholm, Sweden; 2015. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A846386&dswid=-2851.
2. Pasebani M, Rahbari MA: Introducing Clustering Model for Auto Parts Manufacturing Companies. 2007.
3. Munahar S, Triwiyatno A, Munadi M, Setiavan JD: Fuel saving index assessment on driving behavior control system prototype model using neural network. Archives of Transport. 2022, 63:123-141. 10.5604/01.3001.0016.0019
4. Xu S, Wei X, Wang L, Xiong X: Research on driving conditions and fuel consumption of improved K-means clustering algorithm. International Journal of Advanced Network Monitoring and Controls. 2022, 7:1-10. 10.2478/ijanmc-2022-0011
5. Na S, Xumin L, Yong G: Research on k-means clustering algorithm: an improved k-means clustering algorithm. Third International Symposium on Intelligent Information Technology and Security Informatics, Jian, China. 2010, 63-67. 10.1109/IITSI.2010.74
6. Ewing GO, Sarigöllü E: Car fuel-type choice under travel demand management and economic incentives. Transportation Research Part D-Transport and Environment. 1998, 3:429-444. 10.1016/s1361-9209(98)00019-4
7. Parque V, Miyashita T: On learning fuel consumption prediction in vehicle clusters. 2018, 02:116-121. 10.1109/COMPSAC.2018.10214
8. German Car Dataset - Kaggle . https://www.kaggle.com/datasets/ander289386/cars-germany.
9. Sawarkar AD, Shrimankar DD, Sahu SK, Singh L, Bokde ND, Kumar M: Commercial clustering of Indian bamboo species using machine learning techniques. 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS). 2023, 1-5. 10.1109/PCEMS58491.2023.10136094
10. Linyao X, Jianguo W: Improved K-means algorithm based on optimizing initial cluster centers and its application. International Journal of Advanced Network Monitoring and Controls. 2017, 2:9-16. 10.21307/ijanmc-2017-005
11. Chen Z, Ellis T: A self-adaptive Gaussian mixture model. Computer Vision and Image Understanding. 2014, 122:35-46. 10.1016/j.cviu.2014.01.004
12. Schubert E, Sander J, Ester M, Kriegel HP, Xiaowei X: DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Transactions on Database Systems. 2017, 42:1-21. 10.1145/3068335
13. Kodinariya TM, Makwana PR : Review on determining of cluster in K-means clustering. International Journal of Advance Research in Computer Science and Management Studies. 2013, 1:90-95.
14. Zhong H, Zhang H, Jia F: Analysis and improvement of evaluation indexes for clustering results. EAI Endorsed Transactions on Collaborative Computing. 2020, 4:163211. 10.4108/eai.9-10-2017.163211
15. Sinaga KP, Yang M-S: Unsupervised K-means clustering algorithm. IEEE Access. 2020, 8:80716-80727. 10.1109/access.2020.2988796