

Article category: Industrial Marketing

## Data Mining Modeling in Clustering Car Products Sales Data in the Automotive Industry in Indonesia

Juli Astuti <sup>1)</sup>, Trisna Yuniarti <sup>2)\*</sup>

<sup>1,2)</sup> Manajemen Logistik Industri Elektronika, Politeknik APP Jakarta, Jalan Timbul No.32, Jakarta, 12630, Indonesia

### ARTICLE INFORMATION

#### Article history:

Received: February 17, 2023

Revised: May 08, 2023

Accepted: June 16, 2023

#### Keywords:

Data Mining

Big Data

Clustering

K-Means

Automotive

### ABSTRACT

The research aims to build a model based on sales data for all automotive products in Indonesia using data mining with a k-means approach. This study uses automotive product sales data from January 2017 to September 2022. The lowest Davis-Bouldin index shows that three clusters ( $k=3$ ) have the best performance. Based on the clustering results, 92% of the items are in cluster 0, 1% in cluster 1, and 7% in cluster 2. In addition, the clustering results show that cluster 1 is a car product with high sales volume. Cluster 2 is a car product with medium sales volume. Furthermore, cluster 0 is a car product with low sales volume. Business people or related parties can use data visualization and extraction from clustering results to learn the latest insights and information in determining business strategies, policies, and decisions to improve business competitiveness.

This is an open access article under the [CC-BY](https://creativecommons.org/licenses/by/4.0/) license.



#### Corresponding Author:

Trisna Yuniarti

Politeknik APP Jakarta, Jalan Timbul No.32, Jakarta, 12630, Indonesia

Email: [trisna.yuniarti@poltekapp.ac.id](mailto:trisna.yuniarti@poltekapp.ac.id)

© 2023 Some rights reserved

## INTRODUCTION

The global COVID-19 pandemic has brought economic activity to a standstill in all parts of the world. Restrictions on movement and the enactment of lockdowns in various countries triggered a global economic catastrophe. As a result, numerous industries have gone out of business, impacting employees, and leading to layoffs. Some companies, like those in the automotive industry, can still thrive in the face of this pandemic. This industry is one of the priority industries of Making Indonesia 4.0, which aims to boost Indonesia's economy. The Indonesian government has developed many regulations as well as incentives to increase productivity, turnover, and competitiveness in this business.

Economic and technological advances require an industry to be able to optimize its resources by developing a strategy to achieve business goals [1], [2]. One is that the industry needs to expand its business and improve customer service by examining information to increase effectiveness and efficiency in making the right business decisions. The goal of information mining is to solve a range of problems in the automotive industry, such as consumer behavior, optimal production, and resource optimization. Manufacturers are being forced to be more sensitive to customer demands in light of market trends as a result of digital transformation, resulting in an increasingly competitive consumer automotive market [3], [4]. As a result, data utilization has a significant impact on the automobile sector in terms of product development, production processes, and services [5]. Information may be extracted using industry 4.0 technologies, namely

Big Data Analytics. Big data can be collected from internal corporate databases or from the internet, and its handling can be done with data mining techniques.

There seems to be a practical knowledge gap in the research to date. There is a scarcity of rigorous research in existing literature. Some aspects of the visualization of the extraction and relevant information that still need to be explored are based on the clustering insights of the data employed. Despite the fact that many studies employ big data analytics techniques [6]–[17], there are currently few studies in the Indonesian automobile sector that relate to new information or new patterns from existing sales data using data mining. An investigation of these issues is important because it can increase the effectiveness and efficiency of decision-making in the automotive industry. In addition, previous theoretical research has focused mainly on automobile product design, and there is still little practical research conducted on the automotive industry using data mining techniques to extract various knowledge and information from various sales data collections found in the Indonesian automotive industry. In this way, companies can take full advantage of technology 4.0 to streamline operations, meet customer expectations, and improve business results.

Several industries have made use of big data analytics by employing data mining techniques. Using the grey relational model technique, the petroleum sector leverages big data to maximize oil and gas output [7]. Data mining techniques are utilized in the financial industry to forecast finance by developing models based on

the Multilayer Perceptron (MLP), Random Forest (RF), and Support Vector Regression (SVR) algorithms [8]. The energy service sector generates models to increase accuracy and minimize volatility in power load forecasting values in the New England Pool Region utilizing random samples of training data with the output of each averaged [9].

Similarly, the automobile sector is widely embracing big data to assist in the resolution of numerous challenges that arise inside the organization. Some companies use big data for hybrid vehicle product development [10], support Decision Support Systems (DSS) for company productivity, quality, and efficiency [11]–[13]. Big data is also utilized by vehicle manufacturers to objectively evaluate Human-Machine Interface (HMI) designs based on user behavior patterns [14]. Meanwhile, on the ergonomics side, big data from vehicle buyer surveys is used to create car seat design [15]. Hadoop big data is applied to TPCx-HS, SQL applications, and machine learning, which may be employed in a variety of applications for automobile industry challenges [6]. Big data is also employed in autonomous cars to validate control and perception algorithms [16]. Furthermore, the integration of big data and the Internet of Things (IoT) is being utilized to hasten the process of developing energy efficient and intelligent vehicles (EEIV) to combat global warming, food security, and the depletion of nonrenewable energy resources [17].

One of the processes in data mining techniques is clustering. Clustering is a very useful technique in data science to detect cluster structures in data with similar properties [18], [19]. Some

common applications of clustering techniques in data mining have been used in various fields. The field of education uses this technique to cluster student exam results; the field of biology, clustering techniques are used to obtain plant and animal taxonomy; the field of marketing, clustering helps find customer classes; the field of insurance, clustering is applied in insurance companies to identify groups of insurance policyholders; the field of meteorology uses clustering in the study of earthquakes; and the field of banking in detecting fraud [20].

Clustering methods consist of two groups, namely supervised and unsupervised clustering. Algorithms in this category consist of k-means, k-medoid, genetic k-means algorithm (GKA), Self-Organizing Map (SOM), and graph theory methods (CLICK, CAST). However, the k-means approach used in data mining modeling to perform clustering is used in this study. This was done because the k-means approach can handle large amounts of data quickly and effectively [21]. This technique is very efficient because the computational cost required is not too high compared to Gaussian Mixture [22]. In addition, k-means can produce optimal accuracy compared to the k-medoids method, this is according to research on clustering student demographic data [23] and clustering students who drop out of school [24]. Therefore, in this study, data mining with the k-means technique is carried out based on sales data of automotive products in Indonesia obtained from the Gaikindo organization website. The ultimate goal of this research is expected to help automotive companies and explore new knowledge and important information to make the right recommendations and decisions in

implementing their business strategies based on the resulting clustering results.

This paper is organized in the following order: In Part 2, the proposed big data analytical method is presented using data mining in the clustering model of car product sales in the automotive industry in Indonesia using the K-Means technique. In Section 3, shows and explains the interpretation of the results of the clustering model based on the k-means technique and its managerial implications. Section 4 is the conclusion of the model specification results and the empirical results of the model that have been obtained.

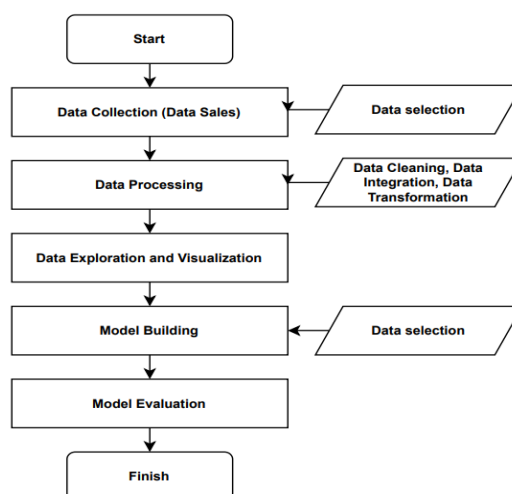
**RESEARCH METHOD**

Data mining modeling with K-Means clustering is used in this paper. This method aims to cluster big data on sales of automobile products, especially cars, from a variety of Indonesian manufacturers. In Figure 1, there are numerous procedures or steps that must be performed before initiating the clustering process. Data collection, data preprocessing, data extraction and

visualization, model building, and model evaluation are the processes involved [25], [26]. The data mining procedure employing clustering algorithms is carried out using the Microsoft Excel and RapidMiner programs. The Microsoft Excel program is utilized up to the data exploration and visualization steps. The RapidMiner program is utilized during the model building and model evaluation steps. The final step is to do an analysis or Knowledge Extraction to derive conclusions from the model obtained.

**Data collection**

Data collection is the first step taken to obtain and complete the data and information needed to conduct research. The data and information utilized in this study are big data on sales of car products in the automotive industry in Indonesia. Secondary data was gathered, mainly historical data on automobile product sales in Indonesia from January 2017 to September 2022. The data selection procedure was carried out during the data gathering step to facilitate the following process.



**Figure 1.** Flowchart of data mining steps for clustering

**Data Preprocessing**

The collected data is processed using data mining techniques. The data mining technique used is K-Means clustering. This technique clusters all car product sales data in Indonesia into groups so that the intrinsic structure of each group is found. The discovery of the structure provides new knowledge that can assist automotive companies in making business decisions. Data processing with this technique goes through several steps [27], namely: Data Cleaning, Data Integration, and Data Transformation. Figure 1 is a flowchart for using data mining modeling with clustering techniques.

**Data Exploration and Visualization**

The data exploration and visualization steps are the first step when conducting data analysis. This steps can identify significant patterns or characteristics in the big data of car sales in Indonesia. These patterns and 3) calculate the object's distance from the centroid.

$$\sqrt{\sum_{i=1}^n (Xi - Yj)^2} \tag{1}$$

Description:

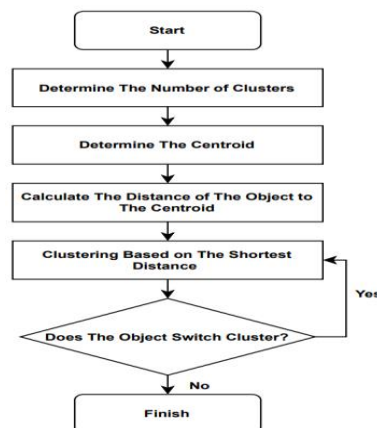
d = the distance between a point and its center.

characteristics will provide insights to managerial parties in a visual manner that is persuasive for effective decision-making. The results of the exploration are in the form of a description of the data visualized using graphs or processed data that has been systematically arranged.

**Model Building**

There are multiple steps to building a data mining model with K-Means clustering. Figure 2 depicts the K-Means data mining modeling steps [28]–[31]:

- 1) The initial centroid is chosen at random to calculate the distance between each input data point and each centroid. The RapidMiner program is used to generate a random centroid immediately throughout the process.
- 2) Calculate the object's distance from the centroid. Equation 1 is used to  $X_i$  = the i-th variable in the x data,  $Y_j$  = the clustering center's jth variable value
- 4) Clustering according to the smallest distance. The clustering of sales data is complete. The data will be clustered based on the computation in no. 3 with the smallest distance.



**Figure 2.** Flowchart of model building using k-means clustering

- 5) Clustering according to the smallest distance. The clustering of sales data is complete. The data will be clustered based on the computation in no. 3 with the smallest distance.
- 6) The computation is repeated several phases until the clustering result converges, which is indicated by no change in the cluster members.

**Model Evaluation**

The model is evaluated to determine how effectively the generated cluster performs [32]-[34]. The Davies Bouldin Index (DBI) or the elbow approach can be used for evaluation. The DBI value is used in this study to determine the best performance of each cluster value.

**RESULT AND DISCUSSION**

**Data Collection**

The data and information collected in this study is big data on automobile sales in the automotive industry in Indonesia. The data obtained is secondary data and is hereinafter referred to as a "dataset," namely historical sales data from January 2017 to September 2022. There are 5033 datasets utilized in all, representing over five years of sales big data. The attributes in the data are as follows: Category, Brand, Specification, Month, Share By, Market, and Total. Table 1 shows an example of the data collected for 2017.

**Table 1.** Sample data based on the sedan vehicle category in 2017

CATEGORY	SPECIFICATION					SHARE BY BRAND	MARKET SHARE	TOTAL 2017	
	BRAND	TYPE MODEL	TRANS	CBU/CKD	ORIGIN COUNTRY				
Sedan Type	CC < 1.500 [G/D]	DAIHATS	Copen	MT	CBU	-	0.2%	0.0%	2
		U	Copen	AT	CBU	-	0.0%	0.0%	-
			All New City IVTEC E	MT	CBU	Thailand	4.8%	1.8%	150
			All New City IVTEC E AT	AT	CBU	Thailand	27.7%	10.4%	870
		HONDA	All New City IVTEC ES AT	AT	CBU	Thailand	0.1%	0.0%	4
			All New Civic	AT	CBU	Thailand	28.7%	10.8%	902
			All New Civic Prestige	AT	CBU	Thailand	13.3%	5.0%	419
		HYUNDAI	Excel III	MT	CKD	INA	0.7%	0.3%	21
		SUZUKI	Ciaz 1.4 MT	MT	CBU	Thailand	0.3%	0.1%	8
			Ciaz 1.4 AT	AT	CBU	Thailand	0.7%	0.3%	22
			Vios E	MT	CKD	INA	3.7%	1.4%	115
			Vios E AT	AT	CKD	INA	0.3%	0.1%	9
		TOYOTA	Vios G	MT	CKD	INA	5.7%	2.1%	179
			Vios G AT	AT	CKD	INA	13.6%	5.1%	429
	Vios G TRD	MT	CKD	INA	0.4%	0.2%	14		
	Vios G TRD AT	AT	CKD	INA	0.0%	0.0%	-		

The data that has been collected will be selected to determine the types and sources of data and instruments that are suitable for the next data processing steps. Therefore, the data collection stage targets attributes or parameters that are relevant for the data mining process. The

results of data selection produce attributes Category Type, Category CC, Brand, Model Type, Trans, CBU/CKD, Origin Country, and Total. These attributes will be used in the next process, which is the data preprocessing steps (Table 2).

**Table 2.** Data selection sample based on 2017 Indonesian automobile sales data

CATEGORY	SPECIFICATION					TOTAL 2017
	BRAND	TYPE MODEL	TRANS	CBU/CKD	ORIGIN COUNTRY	
Sedan Type	DAIHATSU	Copen	MT	CBU	-	2
		Copen	AT	CBU	-	-
	HONDA	All New City IVTEC E	MT	CBU	Thailand	150
		All New City IVTEC E AT	AT	CBU	Thailand	870
		All New City IVTEC ES AT	AT	CBU	Thailand	4
		All New Civic	AT	CBU	Thailand	902
	HYUNDAI	All New Civic Prestige	AT	CBU	Thailand	419
		Excel III	MT	CKD	INA	21
	SUZUKI	Ciaz 1.4 MT	MT	CBU	Thailand	8
		Ciaz 1.4 AT	AT	CBU	Thailand	22
	TOYOTA	Vios E	MT	CKD	INA	115
		Vios E AT	AT	CKD	INA	9
		Vios G	MT	CKD	INA	179
		Vios G AT	AT	CKD	INA	429
		Vios G TRD	MT	CKD	INA	14
		Vios G TRD AT	AT	CKD	INA	-

**Data Processing**

Data preprocessing is a lengthy activity in data mining techniques that produces and improves the quality of a good dataset. This step is critical since the dataset will be further processed based on the data mining techniques and tools that will be employed. So that the algorithm can readily

comprehend the characteristics of the data, the dataset is modified, or new code is produced. As a result, data preprocessing must be performed beforehand to obtain the correct findings. Data preprocessing is divided into several steps, which are as follows:

**Table 3.** Example of Data Cleaning in 2017

Year	Category Type	Category CC	Brand	Type Model	Trans	CBU/CKD	Origin	Total
2017	SEDAN TYPE	CC < 1.500 [G/D]	DAIHATSU	Copen	MT	CBU	JPN	2
2017	SEDAN TYPE	CC < 1.500 [G/D]	DAIHATSU	Copen	AT	CBU	JPN	0
2017	SEDAN TYPE	CC < 1.500 [G/D]	HONDA	All New City IVTEC E	MT	CBU	Thailand	150
2017	SEDAN TYPE	CC < 1.500 [G/D]	HONDA	All New City IVTEC E AT	AT	CBU	Thailand	870
2017	SEDAN TYPE	CC < 1.500 [G/D]	HONDA	All New City IVTEC ES AT	AT	CBU	Thailand	4
2017	SEDAN TYPE	CC < 1.500 [G/D]	HONDA	All New Civic	AT	CBU	Thailand	902
2017	SEDAN TYPE	CC < 1.500 [G/D]	HONDA	All New Civic Prestige	AT	CBU	Thailand	419
2017	SEDAN TYPE	CC < 1.500 [G/D]	HYUNDAI	Excel III	MT	CKD	INA	21
2017	SEDAN TYPE	CC < 1.500 [G/D]	SUZUKI	Ciaz 1.4 MT	MT	CBU	Thailand	8
2017	SEDAN TYPE	CC < 1.500 [G/D]	SUZUKI	Ciaz 1.4 AT	AT	CBU	Thailand	22
2017	SEDAN TYPE	CC < 1.500 [G/D]	TOYOTA	Vios E	MT	CKD	INA	115
2017	SEDAN TYPE	CC < 1.500 [G/D]	TOYOTA	Vios E AT	AT	CKD	INA	9
2017	SEDAN TYPE	CC < 1.500 [G/D]	TOYOTA	Vios G	MT	CKD	INA	179
2017	SEDAN TYPE	CC < 1.500 [G/D]	TOYOTA	Vios G AT	AT	CKD	INA	429
2017	SEDAN TYPE	CC < 1.500 [G/D]	TOYOTA	Vios G TRD	MT	CKD	INA	14

**Data Cleaning**

Data cleaning is the first step in the data preprocessing step. Data cleaning is the process of filling in missing data or values in a data set. Furthermore, data cleaning is done by correcting improperly structured data and missing data. The cleaning results obtained as much as 520 item data cannot

be processed at the next stage because they have a total sales value of 0, so the data must be deleted. Thus the data processed at the next stage amounted to 4513 item data. Table 3 is an example of data cleaning on the 2017 dataset for the sedan car category and energy-efficient and affordable car type 4x2.

**Table 4.** Automobile Sales Data Integration from 2017 to 2022

Year	Category Type	Category CC	Brand	Type Model	Trans	CBU/CKD	Origin	Total
2017	SEDAN TYPE	CC < 1.500 [G/D]	DAIHATSU	Copen	MT	CBU	JPN	2
2017	SEDAN TYPE	CC < 1.500 [G/D]	DAIHATSU	Copen	MT	CBU	JPN	0
2017	SEDAN TYPE	CC < 1.500 [G/D]	HONDA	All New City IVTEC E	MT	CBU	Thailand	150
2017	SEDAN TYPE	CC < 1.500 [G/D]	HONDA	All New City IVTEC E AT	AT	CBU	Thailand	870
....	....	....	....	....	....	....	....	....
....	....	....	....	....	....	....	....	....
....	....	....	....	....	....	....	....	....
2022	AFFORDABLE ENERGY SAVING CARS4X2	CC ≤ 1.200 [G]	TOYOTA	Calya 1.2 E STD	MT	CKD	INA	491
2022	AFFORDABLE ENERGY SAVING CARS4X2	CC ≤ 1.200 [G]	TOYOTA	Calya 1.2 E	MT	CKD	INA	713
2022	AFFORDABLE ENERGY SAVING CARS4X2	CC ≤ 1.200 [G]	TOYOTA	Calya 1.2 G	MT	CKD	INA	12093
2022	AFFORDABLE ENERGY SAVING CARS4X2	CC ≤ 1.200 [G]	TOYOTA	Calya 1.2 G A/T	AT	CKD	INA	3755

**Data Integration**

Data integration is accomplished by integrating existing car sales data sources into a one bigger dataset. Datasets from January 2017 to September 2022 that have been cleaned are integrated using the data format described in Table 3. Table 4 is the result of data integration and is a combined dataset of car product sales from 2017 to 2022. It is ensured that the data record is complete and corresponds to the specified format. This is required so that the dataset can be used in the next phase of data mining.

**Data Transformation**

Data transformation is used to match the features of each attribute by modifying the structure of each attribute. When entering data into the RapidMiner program, format modifications are applied in data transformation. Table 5 shows the outcome of the transformation of the big data dataset of automobile product sales, which will be employed in the subsequent mining process

**Table 5.** Data transformation on automobile sales data

No.	Attribute	Data Transformation	No.	Attribute	Data Transformation
1	Year	Exclude Column	6	Transmission	Polynomial, Exclude Column
2	Type Category	Polynomial, Exclude Column	7	CBU/CKD	Polynomial, Exclude Column
3	CC Category	Polynomial, Exclude Column	8	Origin Country	Polynomial, Exclude Column
4	Brand	Polynomial, Exclude Column	9	Total	Integer
5	Model Type	Polynomial, id			

**Data Exploration and Visualization**

Following data preprocessing, the next steps are data exploration and visualization. The findings of the data processing step are analyzed and visually displayed to help the firm comprehend the retrieved data information. The

following are the findings from the exploration and visualization of the Indonesian automobile sales dataset:

**Model Type**

The findings of data exploration based on model type in the automobile product



sales dataset in Indonesia show that 1.905 car product types were sold between January 2017 and September 2022. Figure 3 depicts the visualization result for the

top ten automobile model types based on sales volume. The BRIO SATYA E is the most popular model, with total sales of 222.195 units.

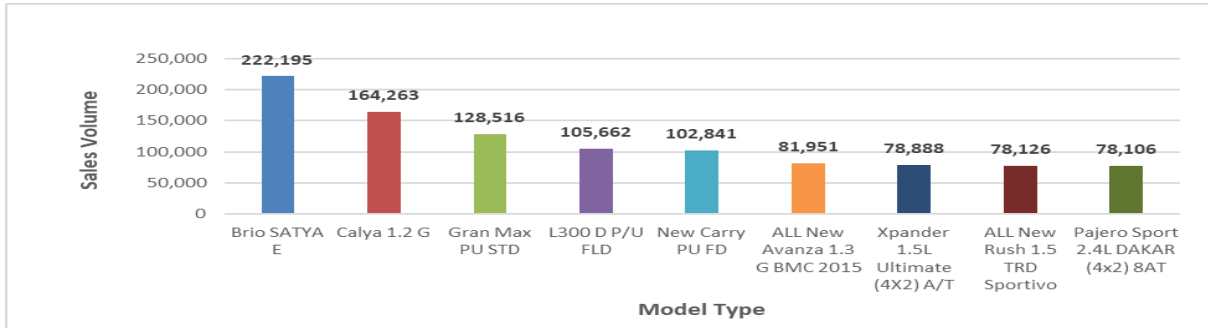


Figure 3. Data visualization of the top 10 automobile model types based on sales data from January 2017 to September 2022

**Type Category**

According to the findings of the automobile type category, there are eight kinds of automobile product types offered in Indonesia. The following automobile types are included in the car type category: Sedan, 4x2, 4x4, Bus, Pick Up, Truck, Double

Cabin 4x2/4x4, and Affordable Energy Saving Cars 4x2. Figure 4 depicts a graphic representation of the types of automobiles sold from January 2017 to September 2022. The type of category of automobile with the largest sales (2.998.929 units) is the 4x2.

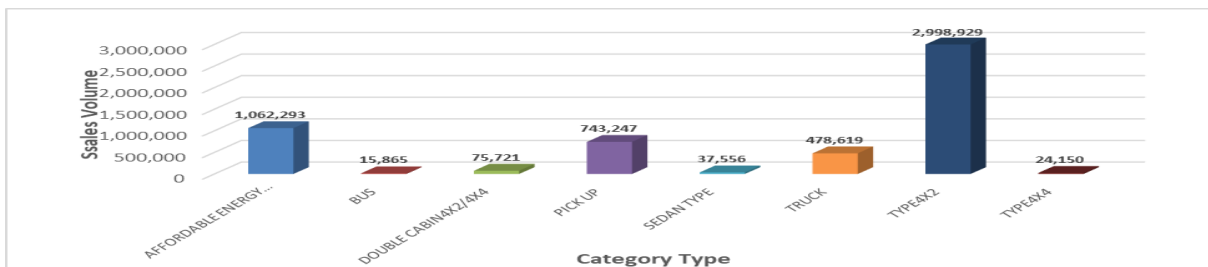


Figure 4. Data visualization of the top 10 automobile types of categories based on Sales data from January 2017 to September 2022

**CC Category**

The exploration results show that there are eleven CC categories of car products sold. Figure 5 shows that the CC<1500

[G/D] is the most popular type of CC, with 2,373,241 units. Meanwhile, the CC category 2,501-3,000 [G] automobile has the fewest sales, totaling 5,853 units.

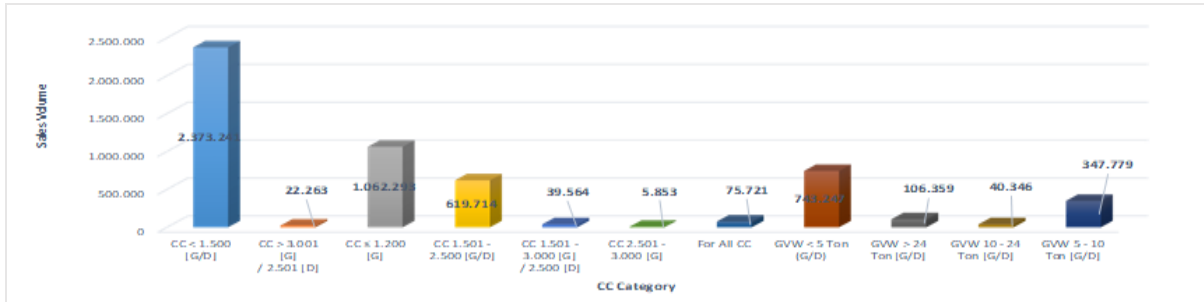


Figure 5. Data visualization with CC Categories

**Brand**

The results of data exploration show that there are 41 automobile brands sold in Indonesia. Figure 6 depicts the top ten automobile brands by sales volume. According to the graph, Toyota is the leading brand, with total sales of 1.755.444 units. Information was obtained indicating that this car brand has a wider variety of

model types compared to other brand agent companies. Some of the reasons this brand generates high sales are because Toyota is a world-leading brand of automotive products that has long been present in Indonesia, has an extensive dealer and after-sales network, the selling price is still high, and spare parts are widely available.

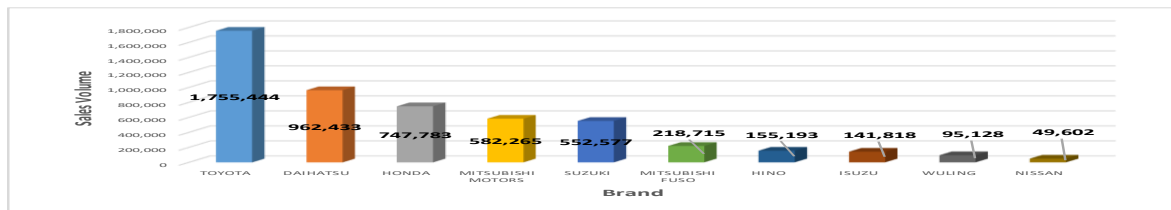


Figure 6. Data visualization of the top ten sales by brand

**Transmission**

The transmission is a system that transmits engine power to the driving wheels. Manual Transmission (MT), Automatic Transmission (AT), and Continuously Variable Transmission (CVT) are the three types of car transmissions marketed in Indonesia from 2017 to 2022. Figure 7 shows that automobiles with manual transmissions are the most popular,

contributing for 62% of all sales. There are several reasons why manual transmissions still dominate the Indonesian market. In terms of price, manual transmission cars are cheaper than automatic-transmission cars. In addition, manual transmission cars are more fuel efficient compared to automatic transmissions.



**Figure 7.** Visualization of sales data by transmission (a) and visualization of sales data by production method (b)

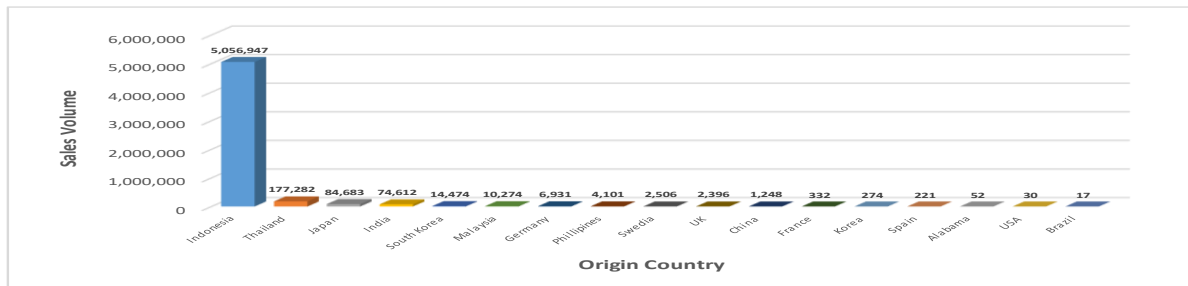
**Production Method**

There are two techniques of producing or assembling automobiles: completely knocked down (CKD) and completely built up (CBU). CKD, or Completely Knocked Down, is a technique of automotive assembly used by the importing nation. A CBU or completely built up, on the other hand, is an automobile imported from a country but not yet built in that country. Both techniques will affect the pricing of automobiles. The visualization results in Figure 7. show that CKD manufacturing

and assembly technology has a high share of sales (93%). Since import costs are not prohibitively high, CKD vehicles are often cheaper than CBU vehicles.

**Country of origin**

The nation of origin is the country in which the automobile is assembled or imported for sale. Figure 8 depicts how automobiles sold in Indonesia come from a variety of nations. From 2017 to 2022, it is known that Indonesia is the nation of origin with the largest sales, approximately 5.056.947 units, when compared to other countries.



**Figure 8.** Visualization of sales data by nation of origin

**Model Building**

This paper addresses the unsupervised data mining approach, which is used to explore, evaluate, and cluster automobiles in Indonesia based on big data sales characteristics. The model employs the k-means clustering approach. The models are created using Microsoft Excel and

RapidMiner. The k-means clustering model will be constructed in steps. The following are the processes or steps involved in k-means clustering execution: 1) The k-means approach performs clustering analysis using a predefined k value. At this step, the number of clusters (k = 3) has been established.

2) The number of clusters is determined by considering the sales group classifications, which are high volume sales, medium volume sales, and low

volume sales. Figure 9 is the model structure for the number of three k=3 clusters performed in the RapidMiner application.

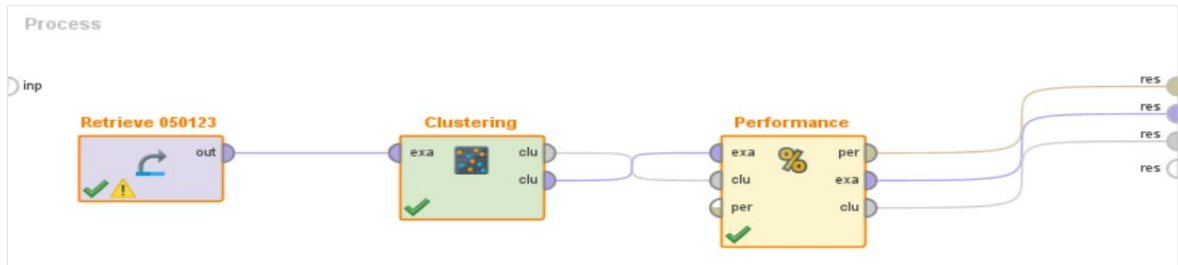


Figure 9. Cluster model structure k=3

3) The initial centroid is determined randomly, and the distance between each input data point and each centroid is calculated. The Average within centroid distance on the sales attribute is 1.740.389 which is obtained from the results of data processing using RapidMiner.

4) The object's distance from the centroid is determined until the iteration results have converged, or until the cluster results no longer change. Table 6 shows the findings of the most recent iteration using data from automobile product sales for each cluster.

Table 6. Centroid result at last iteration

No.	Cluster	Centroid
1	Cluster 0	438.68
2	Cluster 1	25962.93
3	Cluster 2	7825.36

5) Clustering occurs when the computation produces the same iteration value as the previous iteration, after which the process is terminated. Figure 10 depicts the cluster model for k = 3 as well as the

number of members. Cluster 0 yields a total of 4158 items (92%), Cluster 1 yields a total of 46 items (1%), and Cluster 2 yields a total of 309 items (7%).

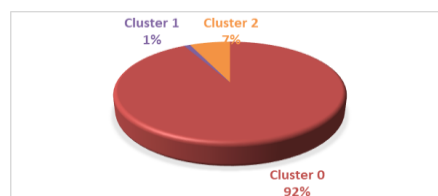


Figure 10. Clustering results of k=3 clusters model evaluation

The K-Means clustering model is evaluated to determine its performance. The evaluation is based on the Davis Bouldin Index (DBI) value. The DBI value is obtained by averaging each cluster's

similarity to the most preferred cluster. Figure 11 shows the model structure for evaluating the performance of each cluster (k=2 to k=7).

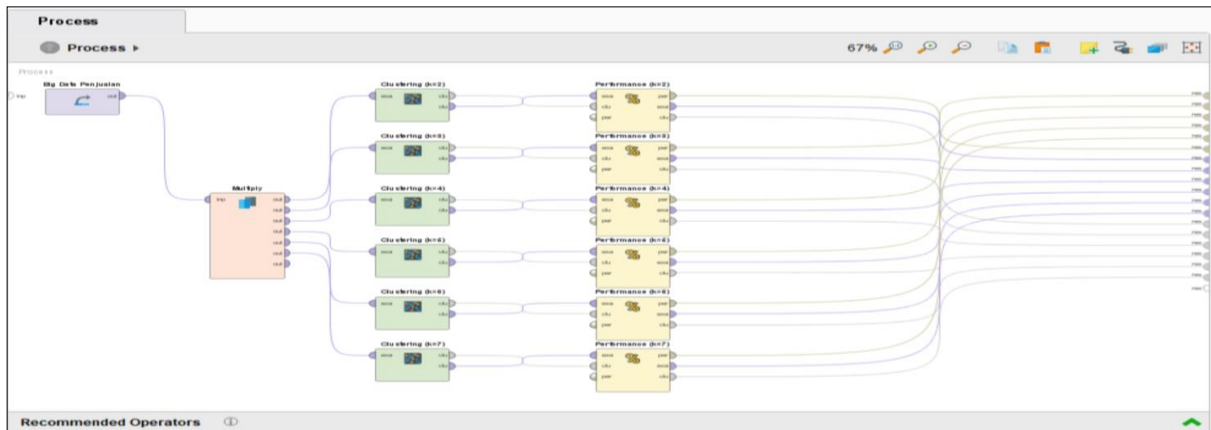


Figure 11. Clustering Performance Model Structure for k=2 to k=7

The purple box is the big dataset of car sales in Indonesia from January 2017 to September 2022. The dataset is the data that has been cleansed, integrated, and changed during the data processing step.

This data will be processed in the RapidMiner application to determine the performance results of each cluster. The green boxes are filled with parameters suitable for clustering each cluster.

Table 7. Clustering Performance Results

Number of Clusters	Davis Bouldin Index
2	0.491
3	0.434
4	0.477
5	0.502
6	0.497
7	0.497

Table 7 shows the results of clustering performance using Rapid Miner. According to this table, the k-means clustering data mining model with k = 3 has the least Davis Bouldin Index value when compared to the other k. The Davis Bouldin value for the k-means clustering data mining model with k = 3 is 0.434.

approach yields the optimal number of clusters that may be produced during clustering. The larger the difference in value, the better the cluster performance. Furthermore, it can be observed from the graph that it forms an optimal cluster if it forms an angle point. Based on the results of DBI and the elbow method (Table 7 and Table 8), it can be seen that the best cluster performance is k = 3.

The elbow approach is another strategy for determining cluster performance. This

**Table 8.** The result of the Average within centroid distance each cluster

Cluster	Average within centroid distance	difference in value
2	4.066.715,02	-
3	1.740.389,19	2.326.325,83
4	1.025.541,64	714.847,55
5	695.793,08	329.748,56
6	464.906,69	230.886,39
7	337.744,55	127.162,14

**Analysis (Knowledge Extraction)**

The results of data processing, model building, and model evaluation can be seen in that with the number of clusters  $k = 3$ , the best performing cluster is

obtained in terms of the Davis Bouldin Index (DBI) value. In Figure 12 information is obtained that cluster 0 (which is colored blue) has more cluster members.



**Figure 12.** Visualization of model clustering results with  $k=3$

This is indicated by the density of the plots contained in the cluster. In cluster 1, which is green, it is known that its members are the fewest among other

clusters. Meanwhile, cluster 2, which is colored orange, is quite dense. The clustering results based on data objects can be seen in Table 9.

**Tabel 9.** Result based on data objects

Cluster	Year	Category Type	Category CC	Brand	Type Model	Transmission	CBU/CKD	Origin Country	Total
2017	cluster_0	SEDAN TYPE	CC < 1.500 [G/D]	DAIHATSU	Copen	MT	CBU	Japan	2
2017	cluster_0	SEDAN TYPE	CC < 1.500 [G/D]	HONDA	All New City IVTEC E	MT	CBU	Thailand	150
2017	cluster_0	SEDAN TYPE	CC < 1.500 [G/D]	HONDA	All New City IVTEC E AT	AT	CBU	Thailand	870
....	....	....	....	....	....	....	....	....	....
....	....	....	....	....	....	....	....	....	....
2022	cluster_0	AFFORDABLE ENERGY SAVING CARS 4X37	CC ≤ 1.200 [G]	TOYOTA	Calya 1.2 E STD	MT	CKD	INA	491
2022	cluster_0	AFFORDABLE ENERGY SAVING CARS 4X38	CC ≤ 1.200 [G]	TOYOTA	Calya 1.2 E	MT	CKD	INA	713
2022	cluster_0	AFFORDABLE ENERGY SAVING CARS	CC ≤ 1.200 [G]	TOYOTA	Calya 1.2 G A/T	AT	CKD	INA	3755

Cluster	Year	Category Type	Category CC	Brand	Type Model	Transmission	CBU/CKD	Origin Country	Total
		4X40							
2017	cluster_1	4X2TYPE	CC < 1.500 [G/D]	DAIHATSU	Great New Xenia X MT	MT	CKD	INA	17193
2017	cluster_1	4X2TYPE	CC < 1.500 [G/D]	DAIHATSU	Great New Xenia R MT	MT	CKD	INA	17751
2017	cluster_1	4X2TYPE	CC < 1.500 [G/D]	HONDA	HR-V E	CVT	CKD	INA	25435
....	....	....	....	....	....	....	....	....	....
....	....	....	....	....	....	....	....	....	....
2022	cluster_1	PICK UP	GVW < 5 Ton (G/D)	DAIHATSU	Gran Max PU 1.5 STD	MT	CKD	INA	24,059
2022	cluster_1	PICK UP	GVW < 5 Ton (G/D)	SUZUKI	New Carry PU FD	MT	CKD	INA	32,371
2022	cluster_1	AFFORDABLE ENERGY SAVING CARS 4X24	CC ≤ 1.200 [G]	HONDA	Brio SATYA E	AT	CKD	INA	19,189
2017	cluster_2	4X2TYPE	CC < 1.500 [G/D]	DAIHATSU	Gran Max BV	MT	CKD	INA	6578
2017	cluster_2	4X2TYPE	CC < 1.500 [G/D]	DAIHATSU	Gran Max HI	MT	CKD	INA	4424
2017	cluster_2	4X2TYPE	CC < 1.500 [G/D]	DAIHATSU	Terios R MT (MC 2015)	MT	CKD	INA	5613
....	....	....	....	....	....	....	....	....	....
....	....	....	....	....	....	....	....	....	....
2022	cluster_2	AFFORDABLE ENERGY SAVING CARS 4X35	CC ≤ 1.200 [G]	TOYOTA	Calya 1.2 G 2022	MT	CKD	INA	7647
2022	cluster_2	AFFORDABLE ENERGY SAVING CARS 4X36	CC ≤ 1.200 [G]	TOYOTA	Calya 1.2 G A/T 2022	AT	CKD	INA	4588
2022	cluster_2	AFFORDABLE ENERGY SAVING CARS 4X39	CC ≤ 1.200 [G]	TOYOTA	Calya 1.2 G	MT	CKD	INA	12093

Note: .... is the remaining data from 2017 to 2022 which is not shown in full.

The clustering results were tested using a confusion matrix to determine the level of clustering accuracy in data mining using k-means. The clustering results consist of 3 groups, namely C0 for sales with low

volume, C1 for sales with medium volume, and C2 for sales with high volume. The following is the confusion matrix value based on the clustering results, as seen in [Table 10](#).

**Table 10.** Calculation of confusion matrix

Actual	Predictive			Total
	C0	C1	C2	
C0	3568	0	0	3568
C1	590	0	237	827
C2	0	46	72	118
<b>Total</b>	<b>4158</b>	<b>46</b>	<b>309</b>	<b>4513</b>

The accuracy of a model is said to be good if it has the smallest error. Classification accuracy is said to be excellent if the error rate is ≤10%, good if the error rate is 10%–20%, moderate if the error rate is 30%–

40%, and poor if the error rate is >40% [\[35\]](#). Based on Table 10, the accuracy value of the grouping model on car sales data in Indonesia for three classes uses the formula: the number of positive data points

that are correctly predicted (TP) divided by the total of all data. Based on the accuracy calculation results, the accuracy value of data mining modeling in clustering car product sales data in the automotive industry in Indonesia is 81%. This shows that the modeling has an error of 19%, which means that the resulting model is good.

$$Accuracy = \frac{TP}{\text{The total of all data}} \times 100\% \quad (2)$$

$$Accuracy = \frac{(3568 + 72)}{(3568 + 590 + 237 + 46 + 72)} \times 100\%$$

$$Accuracy = \frac{3640}{4513} \times 100\% = 81\%$$

Cluster 0 findings on big data sales of automobile items in Indonesia produced 4158 group members. The Alphard 2.5 G model type is quite popular in this cluster. This cluster includes all automobile categories, including: 4x2, truck, sedan, economical energy saving cars 4x2, 4x4, pick up, bus, and double cabin 4x4. This cluster predominantly offers automobiles with CC type 1,500 [G/D], Toyota brand, manual gearbox, CKD production technique, and Indonesian origin. Total sales in this cluster amounted to 1,824,050 units. Although total sales are high, this cluster is classified as having little interest or selling.

Cluster 1 comprises up to 46 items in the big data of automobile product sales in Indonesia. The Brio Satya E model type is popular in this cluster. This cluster has four category types: affordable energy saving cars 4x2, pick up, and truck. Cars with 1,500 [G/D] and CC 1,200 [G] predominate in this cluster. Overall, automobiles with manual

gearboxes from Japanese manufacturers predominate in this cluster. The overall production process in this cluster is CKD in Indonesia. The total number of units sold in this cluster was 1,194,295 units. In terms of model types, this cluster is classified as having a high level of interest or sales.

According to the management implications of this study, it can be seen that cluster 1 is an automobile product with a high volume of sales. Cluster 2 is an automobile product with a medium sales volume. Meanwhile, cluster 0 is an automobile product with a low sales volume. All three clusters are dominated by cars with a CC 1,500 [G/D], manual transmission, and CKD production. According to each cluster, Toyota is the most preferred automotive brand in Indonesia. This data mining modeling might be utilized to boost business competitiveness not only in the automobile market, but also in other industries. The results show that data mining modeling for clustering car product sales data in the automotive industry in Indonesia executed using the k-means method works well. Tables 7 and 8 show the performance of the optimal number of clusters using the DBI and Elbow methods [32]-[34]. In general, this research proves that the k-means method can handle large amounts of data quickly, effectively, efficiently, and with very optimal modeling accuracy [21]-[24]. The limitation of this technique is determining the number of clusters in big data, where additional methods like DBI or Elbow must be used to obtain the optimal number of clusters. This can cause the process phases to become longer, even if they can be completed quickly.



## CONCLUSION

The findings of this study's data analysis lead to the conclusion that data mining modeling utilizing k-means clustering on big data sales of motor vehicle products, particularly automobiles, has an ideal number of clusters of three ( $k = 3$ ) with a good clustering accuracy rate of 81%. With 4.158 cluster members, the centroid value for the total attribute in cluster 0 is 438,68. The centroid value for the total attribute in cluster 1 is 25.962,93 with 46 cluster members. Meanwhile, the centroid value for the total attribute in cluster 2 is 7.825,36, with 309 cluster members.

Other features or variables that might impact the sales of automobile products can be incorporated in future study to uncover new knowledge patterns and information about automobile product sales in Indonesia. In addition, other clustering methods such as the hybrid k-means algorithm can be used to compare and provide better results. Furthermore, the development of an information system to present cluster visualization is strongly advised so that the automobile sector can comprehend the contents of the information thoroughly and rapidly.

## REFERENCES

- [1] M. Ahyat, O. Afriwan, E. Y. Saniah, and A. M. Saputra, "Digital Transformational Leadership A Village Head On Organizational Citizenship Behavior Through Work Climate And Job Satisfaction Village Officials In Lombok Island," *J. Manaj. Ind. dan Logistik*, vol. 6, no. 2, pp. 242–255, 2022.
- [2] S. Sabil, A. Djakasaputra, B. M. A. S. A. Bangkara, S. O. Manullang, and P. Hendriarto, "Understanding Business Management Strategies in Enhancing Profitable and Sustainable SMEs," *J. Manaj. Ind. dan Logistik*, vol. 6, no. 1, pp. 112–131, 2022, doi: 10.30988/jmil.v6i1.989.
- [3] C. Llopis-Albert, F. Rubio, and F. Valero, "Impact of digital transformation on the automotive industry," *Technol. Forecast. Soc. Change*, vol. 162, no. June 2020, p. 120343, 2021, doi: 10.1016/j.techfore.2020.120343.
- [4] Y. Lixin, Y. Jiaxun, and W. Wenbin, "Research and Application on the Governance of Passenger Car Product Data Resources," *Proc. - 2020 Int. Conf. Big Data Soc. Sci. ICBDS 2020*, pp. 46–49, 2020, doi: 10.1109/ICBDSS51270.2020.00018.
- [5] C. J. Wang and B. G. Kim, "Automotive Big Data Pipeline: Disaggregated Hyper-Converged Infrastructure vs Hyper-Converged Infrastructure," *Proc. - 2020 IEEE Int. Conf. Big Data, Big Data 2020*, pp. 1784–1787, 2020, doi: 10.1109/BigData50022.2020.9378045.
- [6] A. Luckow, K. Kennedy, F. Manhardt, E. Djerekarov, B. Vorster, and A. Apon, "Automotive Big Data : Applications , Workloads and Infrastructures," in *2015 IEEE International Conference on Big Data (Big Data) Automotive*, 2015, pp. 1201–1210.

- [7] W. Yuanting *et al.*, "Research and Application of Big Data Analysis Platform for Oil Production Engineering in Huabei Oilfield," *2019 4th IEEE Int. Conf. Big Data Anal. ICBDA 2019*, pp. 148–151, 2019, doi: 10.1109/ICBDA.2019.8713238.
- [8] M. Kim, "A data mining framework for financial prediction," *Expert Syst. Appl.*, vol. 173, no. January, p. 114651, 2021, doi: 10.1016/j.eswa.2021.114651.
- [9] A. S. Khwaja, M. Naeem, A. Anpalagan, A. Venetsanopoulos, and B. Venkatesh, "Improved short-term load forecasting using bagged neural networks," *Electr. Power Syst. Res.*, vol. 125, pp. 109–115, 2015, doi: 10.1016/j.epsr.2015.03.027.
- [10] M. Johanson, S. Belenki, J. Jalminger, and M. Fant, "Leveraging large volumes of data for knowledge-driven product development," *IEEE Big Data*, pp. 736–741, 2014.
- [11] A. A. C. Vieira, L. M. S. Dias, M. Y. Santos, G. A. B. Pereira, and J. A. Oliveira, "Simulation of an automotive supply chain using big data," *Comput. Ind. Eng.*, vol. 137, no. August, p. 106033, 2019, doi: 10.1016/j.cie.2019.106033.
- [12] A. Dacal-Nieto, J. J. Areal, V. Alonso-Ramos, and M. Lluch, "Integrating a data analytics system in automotive manufacturing: Background, methodology and learned lessons," *Procedia Comput. Sci.*, vol. 200, pp. 718–726, 2022, doi: 10.1016/j.procs.2022.01.270.
- [13] T. Widmer, A. Klein, P. Wachter, and S. Meyl, "Predicting Material Requirements in the Automotive Industry Using Data Mining," *Lect. Notes Bus. Inf. Process.*, vol. 354, no. May 2019, pp. 147–161, 2019, doi: 10.1007/978-3-030-20482-2\_13.
- [14] J. Orlovska, C. Wickman, and R. Söderberg, "Big Data Usage Can Be a Solution for User Behavior Evaluation: An Automotive Industry Example.," *Procedia CIRP*, vol. 72, pp. 117–122, 2018, doi: 10.1016/j.procir.2018.03.102.
- [15] M. Romelfanger and M. Kolich, "Comfortable automotive seat design and big data analytics: A study in thigh support," *Appl. Ergon.*, vol. 75, no. May 2018, pp. 257–262, 2019, doi: 10.1016/j.apergo.2018.08.020.
- [16] P. Kowalczyk, M. Komorkiewicz, P. Skruch, and M. Szelest, "Efficient Characterization Method for Big Automotive Datasets Used for Perception System Development and Verification," *IEEE Access*, vol. 10, pp. 12629–12643, 2022, doi: 10.1109/ACCESS.2022.3145192.
- [17] I. Bin Aris, R. K. Z. Sahbusdin, and A. F. M. Amin, "Impacts of IoT and big data to automotive industry," *2015 10th Asian Control Conf. Emerg. Control Tech. a Sustain. World, ASCC 2015*, 2015, doi: 10.1109/ASCC.2015.7244878.




- [18] S. Shukla, "A Review ON K-means DATA Clustering APPROACH," *Int. J. Inf. Comput. Technol.*, vol. 4, no. 17, pp. 1847–1860, 2014, [Online]. Available: <http://www.irphouse.com>.
- [19] K. P. Sinaga and M. Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE Access*, vol. 8, pp. 1–12, 2020.
- [20] A. Ali Hussein and A. Oluwaseun, "Data Mining Application Using Clustering Techniques (K-Means Algorithm) In The Analysis Of Student's Result," *J. Multidiscip. Eng. Sci. Stud.*, vol. 5, no. May, pp. 2587–2593, 2019.
- [21] N. Singh and D. Singh, "Performance Evaluation of K-Means and Heirarichal Clustering in Terms of Accuracy and Running Time," *Int. J. Comput. Sci. Inf. Technol.*, vol. 3, no. 3, pp. 4119–4121, 2012.
- [22] D. Das, P. Kayal, and M. Maiti, "A K-means clustering model for analyzing the Bitcoin extreme value returns," *Decis. Anal. J.*, vol. 6, no. December 2022, 2023, doi: 10.1016/j.dajour.2022.100152.
- [23] S. Gultom, S. Sriadhi, M. Martiano, and J. Simarmata, "Comparison analysis of K-Means and K-Medoid with Ecludiience Distance Algorithm, Chanberra Distance, and Chebyshev Distance for Big Data Clustering," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 420, no. 1, 2018, doi: 10.1088/1757-899X/420/1/012092.
- [24] M. A. Aziz *et al.*, "Comparison of K-Medoids Algorithm with K-Means on Number of Student Dropped Out," *APICS 2022 - 2022 1st Int. Conf. Smart Technol. Appl. Informatics, Eng. Proc.*, pp. 53–58, 2022, doi: 10.1109/APICS56469.2022.9918789.
- [25] Q. Zhang, A. R. Abdullah, C. W. Chong, and M. H. Ali, "E-Commerce Information System Management Based on Data Mining and Neural Network Algorithms," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/1499801.
- [26] M. K. Ha, T. X. Trinh, J. S. Choi, D. Maulina, H. G. Byun, and T. H. Yoon, "Toxicity Classification of Oxide Nanomaterials: Effects of Data Gap Filling and PChem Score-based Screening Approaches," *Sci. Rep.*, vol. 8, no. 1, pp. 1–12, 2018, doi: 10.1038/s41598-018-21431-9.
- [27] T. Yuniarti, I. Surjandari, E. Muslim, and E. Laoh, "Data mining approach for short term load forecasting by combining wavelet transform and group method of data handling (WGMDH)," *Proceeding - 2017 3rd Int. Conf. Sci. Inf. Technol. Theory Appl. IT Educ. Ind. Soc. Big Data Era, ICSITech 2017*, vol. 2018-Janua, pp. 53–58, 2017, doi: 10.1109/ICSITech.2017.8257085.

- [28] J. Gong, "In-depth Data Mining Method of Network Shared Resources Based on K-means Clustering," *Proc. - 2021 13th Int. Conf. Meas. Technol. Mechatronics Autom. ICMTMA 2021*, pp. 694–698, 2021, doi: 10.1109/ICMTMA52658.2021.00160.
- [29] H. Bian, Y. Zhong, J. Sun, and F. Shi, "Study on power consumption load forecast based on K-means clustering and FCM–BP model," *Energy Reports*, vol. 6, pp. 693–700, 2020, doi: 10.1016/j.egy.2020.11.148.
- [30] H. Shen and Z. Duan, "Application research of clustering algorithm based on K-means in data mining," *Proc. - 2020 Int. Conf. Comput. Inf. Big Data Appl. CIBDA 2020*, pp. 66–69, 2020, doi: 10.1109/CIBDA50819.2020.00023.
- [31] S. Kapil and M. Chawla, "Performance Evaluation of K-means Clustering Algorithm with Various Distance Metrics," in *1st IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES-2016) Performance*, 2016, pp. 1–4, doi: 10.1109/ICPEICES.2016.7853264.
- [32] A. K. Singh, S. Mittal, P. Malhotra, and Y. V. Srivastava, "Clustering Evaluation by Davies-Bouldin Index(DBI) in Cereal data using K-Means," *Proc. 4th Int. Conf. Comput. Methodol. Commun. ICCMC 2020*, no. Iccmc, pp. 306–310, 2020, doi: 10.1109/ICCMC48092.2020.ICCMC-00057.
- [33] A. Viloría and O. B. P. Lezama, "Improvements for determining the number of clusters in k-means for innovation databases in SMEs," *Procedia Comput. Sci.*, vol. 151, no. 2018, pp. 1201–1206, 2019, doi: 10.1016/j.procs.2019.04.172.
- [34] E. Rabiaa, B. Noura, and C. Adnene, "Improvements in LEACH based on K-means and Gauss algorithms," *Procedia Comput. Sci.*, vol. 73, no. Awict, pp. 460–467, 2015, doi: 10.1016/j.procs.2015.12.046.
- [35] S. Ruuska, W. Hämäläinen, S. Kajava, M. Mughal, P. Matilainen, and J. Mononen, "Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle," *Behav. Processes*, vol. 148, no. March 2017, pp. 56–62, 2018, doi: 10.1016/j.beproc.2018.01.004.

## BIOGRAPHIES OF AUTHORS

### Author 1



**Juli Astuti**    earned her Master of Arts degree from the Victoria University of Manchester, UK, in 1991. She also received her bachelor's degree from Institut Pertanian Bogor, Indonesia, in 1983. She is currently a lecturer in the Manajemen Logistik Industri Elektronika Study Program, Politeknik APP Jakarta. Her research includes statistics and quality control. She can be contacted via email at [juli.gunawan31@gmail.com](mailto:juli.gunawan31@gmail.com)

### Author 2



**Trisna Yuniarti**    received her Master of Engineering degree in industrial engineering from the Universitas Indonesia, Depok, Indonesia, in 2016 with the thesis "Data mining approach for short-term load forecasting by combining wavelet transform and group method of data handling (WGMDH)". Currently, she is a lecturer at the Manajemen Logistik Industri Elektronika Study Program, Politeknik APP Jakarta. Her research interests are Data Mining and Knowledge Discovery, Applied Statistics, Quality Management, and Industrial Management. She can be contacted via email: [trisna.yuniarti@poltekapp.ac.id](mailto:trisna.yuniarti@poltekapp.ac.id)