CENTIVE
Conference on Electrical Engineering, Informatics,
Industrial Technology, and Creative Media
ISBN: 978-602-53004-0-0

Clustering Passenger Satisfaction Levels in Air Travel Using the K-Means Clustering Algorithm

Pipin Tri Hastuti[1*], Dwi Hartanti[2]
[1,2] Universitas Duta Bangsa Surakarta
[1,2]Jl Bhayangkara No. 55, Tipes, Surakarta, Central Java 57154, Indonesia
*220103032@mhs.udb.ac.id

Abstract —This study aims to cluster the satisfaction levels of airline passengers in the business class segment with business travel purposes who are categorized as disloyal, using the K-Means clustering method. The data was sourced from the Airline Passenger Satisfaction dataset on Kaggle, then cleaned, filtered for disloyal business travelers, and transformed into numerical format. The optimal number of clusters was determined using the Elbow Method, which indicated an optimal value at k=3. Clustering was subsequently carried out with the K-Means algorithm and visualized using PCA. Cluster quality evaluation employed the Davies-Bouldin Index, resulting in a value of -0.5, indicating reasonably good cluster separation. These findings can help airlines understand patterns of dissatisfaction among premium customers and design more targeted service strategies to improve their loyalty.

Keywords – Clustering, Davies-Bouldin Index, K-Means, Elbow Method, Passenger Satisfaction,

I. INTRODUCTION

The aviation industry plays a pivotal role in supporting business activities in today's era of globalization [1]. For passengers traveling for business purposes, comfort and service quality are critical determinants of satisfaction and loyalty toward airlines. Business-class travelers represent a premium market segment that contributes significantly to airline revenue. However, passengers in this segment often exhibit disloyal behavior, easily switching to competitors when service expectations are not met.

Disloyalty among business-class passengers poses a considerable challenge for airlines. Given the high expectations of this customer group, any failure to deliver excellent service can directly harm the airline's reputation and result in substantial revenue loss. Consequently, it is essential for airlines to thoroughly understand the drivers of both satisfaction and dissatisfaction within this segment.

Clustering analysis offers a useful approach for examining passenger satisfaction data. By applying the K-Means algorithm, disloyal business-class passengers can be categorized into clusters based on similarities across service evaluation attributes, such as seat quality, cabin service, food, punctuality, and other amenities. This enables airlines to identify passenger groups with the highest levels of dissatisfaction and uncover the underlying factors.

The objective of this study is to implement K-Means clustering on data concerning disloyal passengers in the business-class segment with business travel purposes. The results are expected to provide detailed insights into patterns of satisfaction and dissatisfaction, thereby allowing airlines to design more targeted service improvement strategies aimed at retaining customers in this highly valuable premium market.

II. RESEARCH METHOD

The data mining methodology in this analysis adopts the Knowledge Discovery in Databases (KDD) framework, consisting of data collection, data cleaning, data transformation, data mining, evaluation, and knowledge generation [2]. The every step can be seen in Fig. 1.

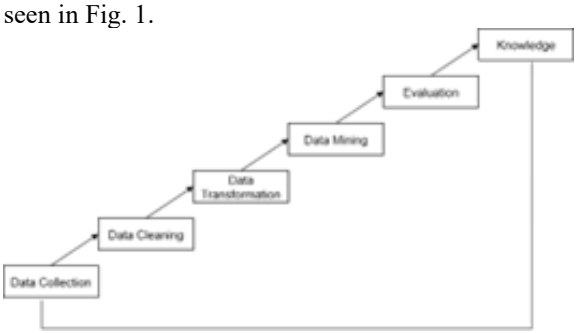

Figure 1. The KDD process flows

A. Data Collection

The dataset used in this study was obtained from the *Airline Passenger Satisfaction Dataset* on Kaggle (https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction) [3]. This dataset contains comprehensive information on passenger profiles and

airline service experiences, which serves as the primary foundation for subsequent analysis.

A. Data Cleaning

In the data cleaning stage, the dataset obtained from Kaggle was processed to remove missing values and duplicate entries to ensure optimal data quality. Subsequently, the dataset was filtered to include only passengers categorized as disloyal customer type, business travel type, and business class. From this filtering process, only specific attributes were selected, namely ID, Age, Flight Distance, Seat Comfort, and Cleanliness.

At the exploratory data analysis stage, statistical summaries were conducted to understand the overall characteristics of the dataset. The average passenger age was 36 years, with flight distances ranging from 200 to 3000 miles. Service attributes such as *Seat Comfort* and *Cleanliness* were rated on a 1–5 scale, where most disloyal business travelers rated these attributes below 3, indicating general dissatisfaction. A correlation analysis also showed that *Seat Comfort* and *Cleanliness* had the strongest positive correlation (r = 0.72), suggesting that passengers who rated seat comfort highly also tended to give higher cleanliness scores.

B. Data Transformation

Data transformation was carried out by converting nominal data into numerical form through an initialization process, enabling further processing in the data mining stage [4].

C. Data Mining Process

In this stage, the data mining process was conducted by applying algorithms or methods to discover hidden patterns within the data. This represents the core step of the KDD process, where analytical techniques or artificial intelligence methods are employed to extract meaningful information with the potential to provide valuable insights [5].

D. Evaluation

In this stage, the clustering results were evaluated using the Davies-Bouldin Index, which aims to measure the quality of separation between data clusters. A lower index value indicates better clustering quality [6]. The mathematical formulation of the Davies–Bouldin Index is expressed as follows:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \frac{max}{j \neq 1}\left(\frac{S_i + S_j}{M_{ij}}\right)$$

where:

a) $S_i$represents the average distance between each data point in cluster *i* and its centroid (intra-cluster distance).

b) M_{ij}denotes the distance between the centroids of clusters *i* and *j* (inter-cluster distance).

E. Knowledge

The knowledge presentation stage involves delivering the analysis results in the form of visualizations or representations that are easy to interpret. The objective of this stage is to clearly and informatively convey the insights or patterns discovered during the data mining process to users or decision-makers [6].

III. RESULT

B. Data

Table 1 presents the outcome of a series of processes, starting from data collection, data cleaning to remove missing values and duplicates, and the transformation of nominal data into numerical forms. These steps were carried out to ensure that the data used in the analysis possesses high quality and aligns with the requirements of modeling. Consequently, the data presented in this table is ready to be utilized in the data mining stage to uncover relevant patterns and information.

Table 1. Data Results

No	Id	Age	Flight Distance	Seat comfort	Cleanliness
1	70990	32	802	4	2
2	87447	42	373	4	4
3	88249	24	444	2	2
4	3949	27	764	2	2
5	61426	40	2419	4	4
6	102981	39	546	2	2
7	73906	27	2342	2	2
8	100462	48	759	2	2
9	20310	26	900	1	1
10	55823	49	1416	2	2
11	74101	38	641	2	2
12	64475	23	282	4	4
13	123812	25	544	2	2
…	…	…	…	…	…
1873	78463	34	526	4	4

C. Data Mining Process

At this stage, data processing was performed using the K-Means algorithm to group the data based on specific similarities. The optimal number of clusters was determined using the Elbow method, executed through Python code in Google Colab to obtain the best *k* value [7]. Once the number of clusters was established, the clustering process was continued in RapidMiner using the K-Means algorithm to group the data according to the results derived from the Elbow method analysis.

a) Elbow Method

The Elbow Method in Fig. 2 is a technique used to determine the optimal number of clusters (*k*

value) in clustering algorithms, particularly K-Means [7]. This method operates by calculating the Within-Cluster Sum of Squares (WCSS) for different numbers of clusters [8]. WCSS reflects how closely data points in a cluster are grouped around the cluster center. The results are then plotted in a graph, where the X-axis represents the number of clusters and the Y-axis represents the WCSS value [8]. The "elbow point" on the graph indicates the optimal *k* value, where the reduction in WCSS begins to slow down significantly. This point is considered the best compromise between the number of clusters and the efficiency of data segmentation.

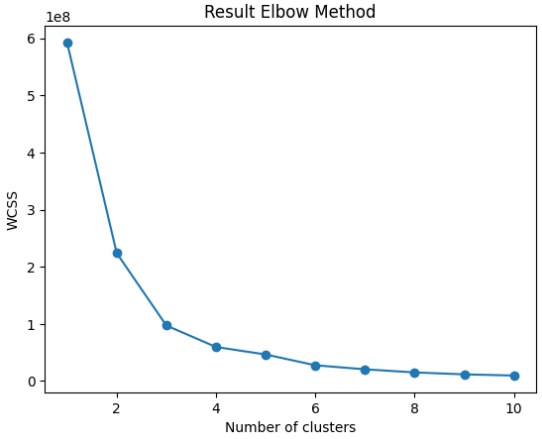

Figure 2, Elbow Method

Based on the test results using the Elbow Method, the elbow point on the graph was observed at k = 3. This indicates that the optimal number of clusters for the dataset is three. At this point, the reduction in WCSS begins to slow, meaning that adding more clusters beyond this value does not significantly improve the quality of cluster separation. Therefore, k = 3 was selected as the optimal number of clusters to be used in the subsequent clustering process with the K-Means algorithm.

b) K-Means Algorithm

The data clustering process in this study was conducted using the K-Means algorithm, implemented through the RapidMiner software. RapidMiner was chosen because it provides an interactive interface that facilitates the construction of clustering workflows without requiring manual coding [9]. By applying K-Means in RapidMiner, the data that had undergone cleaning and transformation was grouped according to its characteristics, based on the predetermined number of clusters, which was three. The results can be seen in Fig. 3.

The K-Means modeling in RapidMiner was carried out using three main components: Retrieve (to load the dataset), Clustering (to apply the K-Means algorithm), and Performance (to evaluate the clustering results) [10]. This process produced cluster visualizations that facilitated further analysis.

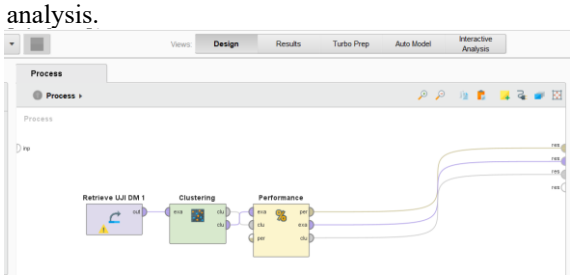

Figure 3. K-Means Modeling in RapidMiner

D. Evaluation

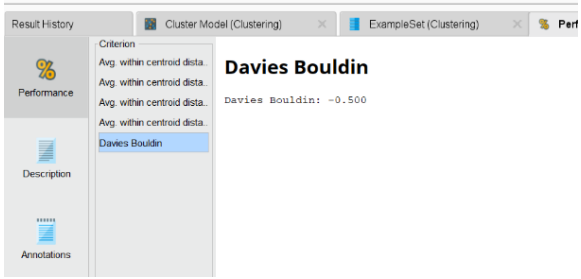

Figure 4. Performance Evaluation

Based on the evaluation results shown in the Fig. 4, the obtained Davies-Bouldin Index value is -0.500. This value indicates that the clustering results have good quality, with clusters that are relatively well-separated and compact. Although in theory the Davies-Bouldin Index is usually positive, the negative result in RapidMiner can be interpreted as an indication that the constructed clustering model has a fairly optimal cluster structure [11].

E. Knowledge

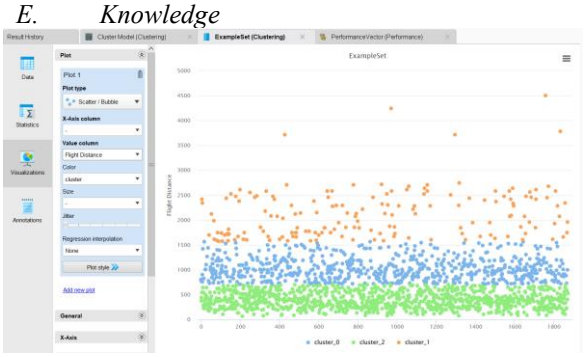

Figure 5. Visualization of Results

Based on the clustering visualization in the Fig. 5, the data was successfully grouped into three clusters distinguished by color: blue (cluster_0), green (cluster_2), and orange (cluster_1). Each cluster demonstrates different patterns in terms of flight distance.

a) Cluster_1 (orange) consists of passengers with relatively long flight distances.
b) Cluster_0 (blue) includes passengers with medium flight distances.
c) Cluster_2 (green) is dominated by passengers with short flight distances.

This information can serve as valuable insights for decision-making, such as developing service strategies

based on flight segments or planning more targeted promotions according to cluster characteristics.

A. Implementation

At the implementation stage, an interactive web-based application was developed to perform clustering using the K-Means algorithm. This application was built using the Python programming language and the Streamlit framework, which enables the creation of a simple, intuitive, and responsive user interface. The goal is to allow users to easily execute the clustering process and directly obtain visualizations of the results.

a) Interface

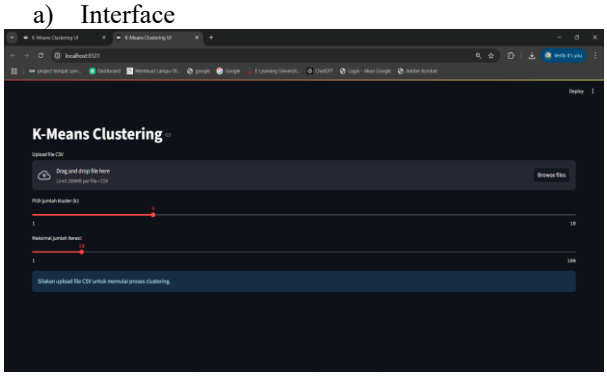

Figure 6. Interface Design

The application interface in Fig. 6 was designed to be simple and user-friendly, including for non-technical users. A CSV file upload feature is provided as the main input method. Once the file is uploaded, the system automatically displays the initial data ready for the clustering process.

b) Clustering Results

Table 2 presents a portion of the clustering results generated by the K-Means algorithm after determining the optimal number of clusters (k = 3) using the Elbow Method. The dataset contains disloyal business-class passengers with business travel purposes, who were grouped based on four numerical attributes: Age, Flight Distance, Seat Comfort, and Cleanliness. The clustering process produced three distinct groups labeled as Cluster 0, Cluster 1, and Cluster 2, each showing different passenger characteristics and satisfaction patterns.

Table 2 Clustering Results

No	Id	Age	Flight Distance	Seat comfort	Cleanliness	Cluster
1	70990	32	802	4	2	0
2	3949	27	764	2	2	0
3	100462	48	759	2	2	0
4	20310	26	900	1	1	0
5	55823	49	1416	2	2	0
6	107972	40	825	5	5	0
7	66384	38	1562	2	2	0
8	100404	45	1011	4	4	0
9	119948	29	1009	3	3	0
10	101584	29	1099	5	3	0
…	…	…	…	…	…	…
674	61426	40	2419	4	4	1
675	73906	27	2342	2	2	1
…	…	…	…	…	…	…
1873	78463	34	526	4	4	2

Cluster 0 is dominated by passengers with medium to long flight distances and low ratings in seat comfort and cleanliness, typically scoring between 1 and 2. This group represents the most dissatisfied travelers, indicating that inadequate comfort and hygiene standards may strongly influence their lack of loyalty. Cluster 1 consists of passengers with moderate flight distances who provide higher scores for comfort and cleanliness, ranging between 4 and 5. These passengers can be categorized as relatively satisfied but still exhibit disloyal behavior, possibly due to non-service-related factors such as flight schedules or ticket prices. Meanwhile, Cluster 2 includes passengers with shorter flight distances and moderate satisfaction scores (2–3), representing a neutral group that is neither highly dissatisfied nor particularly loyal.

Overall, the clustering results indicate that flight distance and in-flight service quality, particularly seat comfort and cleanliness, are the most influential factors differentiating passenger satisfaction patterns. These insights can help airlines identify which service aspects should be prioritized for improvement. For example, enhancing seat comfort and cabin cleanliness could address dissatisfaction among Cluster 0 passengers, while maintaining consistent service quality may retain relatively satisfied passengers in Cluster 1. For Cluster 2, focusing on efficiency and punctuality could help increase engagement and satisfaction levels among short-distance business travelers.

Table 3 K-Means Iteration Process

No	Id	C1	C2	C3	Nearest	Cluster
1	70990	209.0454496036687 1	547.1142476667922	162.2282342873767 5	2	0
2	87447	638.0399674001621	119.8540779448075 3	591.2892693090245	1	2
3	88249	567.1225617095479	189.0449681954005 4	520.0019230733672	1	2

No	Id	C1	C2	C3	Nearest	Cluster
4	3949	247.1659361643509494	509.0432201689754409	200.0274981096349	2	0
5	61426	1408.009588035536174	2164.083408743757414	1455.094155029151514	0	1
6	102981	465.0365577027251313	291.5698887059499	418.2714429649722	1	2
7	73906	1331.030803550391	2087.010541420431	1378.003991285947	0	1
8	100462	252.3707590034947	504.7306212228459704	206.4049418013047	2	0
9	20310	111.5078472574912525	645.0333324720514	64.03124237432849	2	0
10	55823	405.2641114137792	1161.341035183033535	452.6930527410378	0	1
11	74101	370.0364846876588686	386.3845234995832	323.3063562629104747	2	0
12	64475	729.1001302976156	27.07397274136176868	682.0139294765173	1	2
13	123812	467.1263212451215	289.0415195088761	420.0035714133869	1	2
…	…	…	…	…	…	…
1873	78463	4.850.030.927.736.440	27.131.162.894.354.500	4.381.346.824.893.000	1	2

Table 3 presents the iterative process of the K-Means algorithm in determining the optimal cluster assignment for each data point. During each iteration, the algorithm calculates the distance of every data point to the centroids of all clusters (C1, C2, and C3). The "Nearest" column indicates the smallest distance value, which determines the data point's temporary cluster membership. After all data points are assigned, the centroid positions are recalculated based on the updated cluster compositions. This iterative process continues until convergence is reached, meaning that the centroid positions no longer change significantly.

The information presented in Table 3 illustrates how the algorithm gradually stabilizes the clustering structure through multiple iterations, ensuring that each data point is grouped into the cluster with the highest similarity and minimum distance to its centroid.

b) Result Visualization

In Fig. 7 explained to understanding of the clustering results, Principal Component Analysis (PCA) was applied to reduce the multidimensional dataset into two principal components that capture most of the variance in the data. This dimensionality reduction technique allows complex, high-dimensional relationships among variables such as *Age*, *Flight Distance*, *Seat Comfort*, and *Cleanliness* to be visualized in a simpler two-dimensional space without significant loss of information. The reduced results were then visualized in a scatter plot, where each cluster is represented by a different color, and the centroid positions are marked with an "X" symbol to indicate the cluster centers.

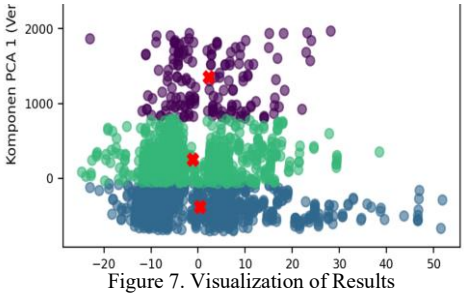

Figure 7. Visualization of Results

Through this visualization, it becomes easier to observe the distribution, separation, and compactness of the three clusters. Passengers within the same cluster appear closer together, reflecting high similarity in their service evaluation patterns, while clusters that are farther apart indicate groups with distinct satisfaction characteristics. The visualization also confirms the numerical results obtained from the Davies–Bouldin Index, showing that the clusters are well separated and demonstrate consistent internal cohesion. Consequently, PCA-based visualization provides both an intuitive and analytical means of validating the effectiveness of the K-Means clustering process.

c) Elbow Method

The developed application also integrates the Elbow Method. This method utilizes the Within-Cluster Sum of Squares (WCSS) value across different k values to determine the most optimal number of clusters. The elbow point in the graph indicates the best k value, and from the dataset testing, the optimal number obtained was k = 3, as illustrated in the Fig. 8.

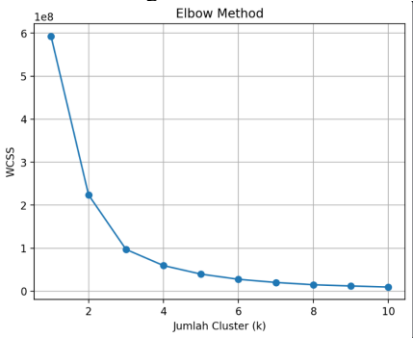

Figure 8. Elbow Method

IV. DISCUSSION

The results of this study show that K-Means clustering is effective for segmenting business class passengers with business travel type who are categorized as disloyal customers. Using the Elbow Method, three clusters were identified as optimal, each displaying distinct characteristics, particularly in terms of flight distance and seat comfort. These attributes appear to be the most influential factors in shaping passenger satisfaction within this premium segment.

From a practical perspective, the segmentation provides useful insights for airlines in designing service strategies. For instance, passengers with longer flight distances may require greater attention to seating comfort and in-flight amenities, while those on shorter flights may prioritize punctuality and efficiency. Tailoring services to these patterns can help airlines improve loyalty and retention among business travelers.

However, studying has certain limitations. The dataset was obtained from Kaggle and may not fully represent the diversity of passengers across different airlines and regions. In addition, the analysis focused only on selected attributes, leaving out other potentially important factors such as baggage handling, boarding process, or inflight entertainment.

Future research could expand by incorporating additional variables and testing alternative clustering techniques such as Hierarchical Clustering or DBSCAN. Using real airline operational data would also improve the reliability of the results and enhance their applicability for industry decision-making.

In addition to identifying satisfaction patterns, the clustering outcomes can be further leveraged to develop personalized loyalty strategies targeting specific passenger groups. For example, passengers in Cluster 0, characterized by low satisfaction scores and long flight distances, could be offered premium comfort upgrades or post-flight service recovery programs. Meanwhile, Cluster 1 passengers, who show higher satisfaction yet remain disloyal, might be engaged through membership rewards or exclusive promotions. Cluster 2 passengers, consisting of short-distance travelers with moderate satisfaction, could benefit from flexible scheduling or express check-in options. By implementing such targeted initiatives, airlines can strengthen customer engagement, improve retention rates, and enhance their competitive positioning within the premium market segment.

V. CONCLUSSION

This study successfully applied the K-Means Clustering method to analyze passenger satisfaction levels in the business class segment for business travelers categorized as disloyal customers. Based on the results of the Elbow Method, the optimal number of clusters was determined to be three. The clustering process was conducted using RapidMiner and Python, and further visualized through PCA to highlight the differences between clusters. Evaluation using the Davies-Bouldin Index produced a score of -0.5, indicating that the clusters were fairly well-separated and representative. Each cluster exhibited distinct characteristics, particularly in terms of flight distance and seat comfort. These findings provide valuable insights for airlines in formulating more focused and relevant service improvement strategies aimed at enhancing customer loyalty in the premium segment.

ACKNOWLEDGMENT

The authors would like to express their gratitude to Universitas Duta Bangsa Surakarta for providing academic support during the completion of this research. Special thanks are also extended to the contributors of the Kaggle dataset, which served as the foundation for this study. Finally, the authors appreciate the constructive feedback from peers and colleagues that greatly improved the quality of this work.

REFERENCES

[1] K. Chen, “Research on the Evaluation of Economy Class Service Quality Based on Customer Satisfaction of MU Airlines Degree Programme in Aviation Business,” 2025.

[2] F. K. Nasser and S. F. Behadili, “A Review of Data Mining and Knowledge Discovery Approaches for Bioinformatics,” *Iraqi Journal of Science*, pp. 3169–3188, Jul. 2022, doi: 10.24996/ijs.2022.63.7.37.

[3] TJ Klein, “Airline Passenger Satisfaction.”

[4] N. F. Syifa, M. Martanto, A. R. Dikananda, and D. Rohman, “APPLICATION OF K-MEANS ALGORITHM IN KINDERGARTEN SCHOOL LOCATION CLUSTERING OF SCHOOL SELECTION STRATEGY BY PARENTS,” *Jurnal Komputer dan Informatika*, vol. 13, no. 1, pp. 26–35, Mar. 2025, doi: 10.35508/jicon.v13i1.20202.

[5] B. D. Lund and J. Ma Lund, “A review of cluster analysis techniques and their uses in library and information science research: k-means and k-medoids clustering,” 2021.

[6] S. Suraya, M. Sholeh, and U. Lestari, “Evaluation of Data Clustering Accuracy using K-Means Algorithm,” *International Journal of Multidisciplinary Approach Research and Science*, vol. 2, no. 01, pp. 385–396, Dec. 2023, doi: 10.59653/ijmars.v2i01.504.

[7] M. Alizade, R. Kheni, S. Price, B. C. Sousa, D. L. Cote, and R. Neamtu, “A Comparative Study of Clustering Methods for Nanoindentation Mapping Data,” *Integr Mater Manuf Innov*, vol. 13, no. 2, pp. 526–540, Jun. 2024, doi: 10.1007/s40192-024-00349-3.

[8] T. Gambheera Arachchi, M. Dahanayaka, and H. Niles Perera, “Analyzing Sustainability Initiatives of the Airline Industry Through Random Forest Classification and K-Means Clustering Techniques,” 2024.

[9] H. T. Sukmana, “Using K-Means Clustering to Enhance Digital Marketing with Flight Ticket Search Patterns,” *Journal of Digital Market and Digital Currency*, vol. 1, no. 3, pp. 286–304, Dec. 2024, doi: 10.47738/jdmdc.v1i3.22.

[10] D. Yu, S. Dong, and S. Yao, “Improvement of K-Means Algorithm and Its Application in Air Passenger

Grouping," *Comput Intell Neurosci*, vol. 2022, 2022, doi: 10.1155/2022/3958423.

[11] G. Airlangga, "A Comparative Analysis of Clustering Algorithms for Expedia's Travel Dataset," *Sinkron*, vol. 9, no. 1, pp. 476–483, Feb. 2025, doi: 10.33395/sinkron.v9i1.14343.