VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7274

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

K-MEANS-BASED TRAINING DATA PROCESSING FOR IMPROVING TOURISM RECOMMENDATION ACCURACY

Candra Agustina[1*]; Purwanto Purwanto[2]; Farikhin Farikhin[3]; Eka Rahmawati[4]

Accounting Information Systems[1]
Information Systems[4]
Faculty of Engineering and Informatics[1,4]
Bina Sarana Informatika University, Jakarta, Indonesia[1,4]
www.bsi.ac.id[1,4]
candra.caa@bsi.ac.id[*], eka.eat@bsi.ac.id

Department of Chemical Engineering, Faculty of Engineering[2]
Department of Mathematics, Faculty of Science and Mathematics[3]
Diponegoro University, Semarang, Indonesia[2,3]
www.undip.ac.id[2,3]
purwanto@live.undip.ac.id, farikhin.math.undip@gmail.com

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract—*This study investigates the enhancement of tourism destination recommendation systems through the use of K-Means clustering to improve training data quality and model accuracy. The rapid advancement of information technology has increased the demand for personalized and accurate recommendation systems within the tourism industry. Despite this, achieving high prediction accuracy remains a significant challenge. This study employs K-Means clustering to segment training data into homogeneous clusters, thereby improving data representation and enhancing the predictive accuracy of recommendation models. The research methodology includes a comprehensive literature review, data collection, preprocessing, clustering, and model testing using K-Nearest Neighbors (KNN), Decision Tree, and Naive Bayes algorithms. The results show that after applying K-Means clustering, KNN's accuracy increased by 2.27%, and its kappa and precision values also improved, indicating enhanced reliability and prediction accuracy. Naive Bayes exhibited substantial improvements with a 9.09% increase in accuracy, alongside significant enhancements in kappa and precision metrics. Conversely, the Decision Tree algorithm experienced a decline in performance after clustering. Therefore, clustering techniques are not suitable for application to the Decision Tree algorithm.*

Keywords: *Classification, Clustering, K-Means, Optimization, Pseudo-Labelling.*

Intisari—*Penelitian ini menyelidiki peningkatan sistem rekomendasi destinasi wisata melalui penggunaan K-Means clustering untuk meningkatkan kualitas data latih dan akurasi model. Kemajuan pesat dalam teknologi informasi telah meningkatkan permintaan akan sistem rekomendasi yang personal dan akurat di industri pariwisata. Namun demikian, mencapai tingkat akurasi prediksi yang tinggi masih menjadi tantangan yang signifikan. Penelitian ini menggunakan K-Means clustering untuk mengelompokkan data latih ke dalam klaster-klaster homogen, sehingga dapat memperbaiki representasi data dan meningkatkan akurasi prediksi dari model rekomendasi. Metodologi penelitian mencakup tinjauan pustaka yang komprehensif, pengumpulan data, pra-pemrosesan, clustering, dan pengujian model menggunakan algoritma K-Nearest Neighbors (KNN), Decision Tree, dan Naive Bayes. Hasil penelitian menunjukkan bahwa setelah penerapan K-Means clustering, akurasi KNN meningkat sebesar 2,27%, dan nilai kappa serta presisi juga mengalami peningkatan, yang mengindikasikan peningkatan keandalan dan akurasi prediksi. Naive Bayes menunjukkan peningkatan yang signifikan dengan kenaikan akurasi sebesar 9,09%, disertai peningkatan yang berarti pada metrik kappa dan*

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7274

presisi. Sebaliknya, algoritma Decision Tree mengalami penurunan kinerja setelah proses clustering. Oleh karena itu, teknik clustering tidak cocok untuk diterapkan pada algoritma Decision Tree.

***Kata Kunci**: Klasifikasi, Clustering, K-Means, Optimasi, Pseudo-Labeling*

INTRODUCTION

Advancements in information technology have revolutionized the way we travel, ushering the tourism industry into a new era of connectivity and personalization[1]. One of the most important innovations is the tourism-destination recommendation system. These systems help travelers find destinations that match their preferences and interests, thereby enhancing their overall travel experience [2],[3]. As is well-known, tourist satisfaction is the key to sustainable tourism[4]. However, a major challenge in developing recommendation systems is achieving a high prediction accuracy.

The accuracy of recommendation systems is significantly influenced by the quality of the training data used[5]. High-quality training data can enhance a model's ability to recognize patterns and trends in user preferences[6],[7],[8]. Clustering techniques are effective methods for improving the quality of training data. K-Means clustering is a popular algorithm used to group data into several clusters based on their feature similarities. By clustering the training data using K-Means, we can improve the data representation and enhance the prediction accuracy of recommendation systems[9], [10], [11].

Most research in the fields of machine learning and data mining tends to prioritize algorithm development or evaluation, particularly during validation, with a focus on the accuracy or overall performance. However, it is important to acknowledge that the pre-processing stage, which is often considered a simple initial step, plays a crucial role in enhancing the accuracy and quality of the final results of the built model[12]. Furthermore, integrating algorithms into data handling during the training process can further increase accuracy. Therefore, research that focuses on both preprocessing and algorithm application in data training can significantly contribute to improving the quality and generalization of machine-learning models.

This study focuses on processing training data using K-Means clustering to improve the accuracy of tourism destination recommendation systems. Through this approach, the training data are expected to be divided into more homogeneous clusters, allowing the recommendation model to identify the specific preferences of each user group more easily[13], [14]. In addition, this study evaluates the extent to which K-Means clustering techniques can improve the performance of recommendation systems compared with traditional approaches.

In addition, the K-Means algorithm was chosen because of its advantages.

a. It is simple and Efficient, Easy to implement, and quickly clusters large datasets[15].
b. Scalability Effective for large datasets[15], [16].
c. A fast convergence quickly reaches an optimal solution[17].
d. The interpretability Results were easy to interpret with clear clusters[18].
e. Adaptability is Useful in various applications such as market segmentation and pattern recognition[15].
f. Multidimensional Data can cluster data with many dimensions[19].
g. Flexible Optimization Can be achieved using techniques such as K-Means++ and adjusting the number of clusters[20], [21].

Steps for Calculating K-Means[22]

a. Determining the desired number of clusters.
b. The initial centroid for each cluster is established.
c. Calculate the distance between each data point and its respective centroid.
d. Assign data points to clusters based on the smallest centroid.
e. The average of each data point within each cluster is calculated to establish a new centroid.
f. Repeat steps 3-5 until the clusters no longer change.

The primary objective of this research is to develop an effective training data processing method using K-Means clustering and assess its impact on the accuracy of tourism destination recommendation systems. The results of this research are expected to make a significant contribution to the field of recommendation systems, particularly in the context of tourism, and provide new insights for the development of more accurate and relevant recommendation models.

Previous studies have applied clustering primarily at the user-behavioral analysis level, focusing on algorithmic performance rather than the preprocessing stage. Unlike these approaches, the present study integrates K-Means clustering during the training data preprocessing phase and applies pseudo-labeling to enhance data

homogeneity. This combination has not been examined in prior tourism recommendation studies, particularly those utilizing primary visitor data. The distinctiveness of this study originates from its utilization of primary data, which is centered on variables previously proposed by several research projects. Moreover, it incorporates the K-Means algorithm in the preprocessing of the training data to enhance the accuracy of the forecasts.

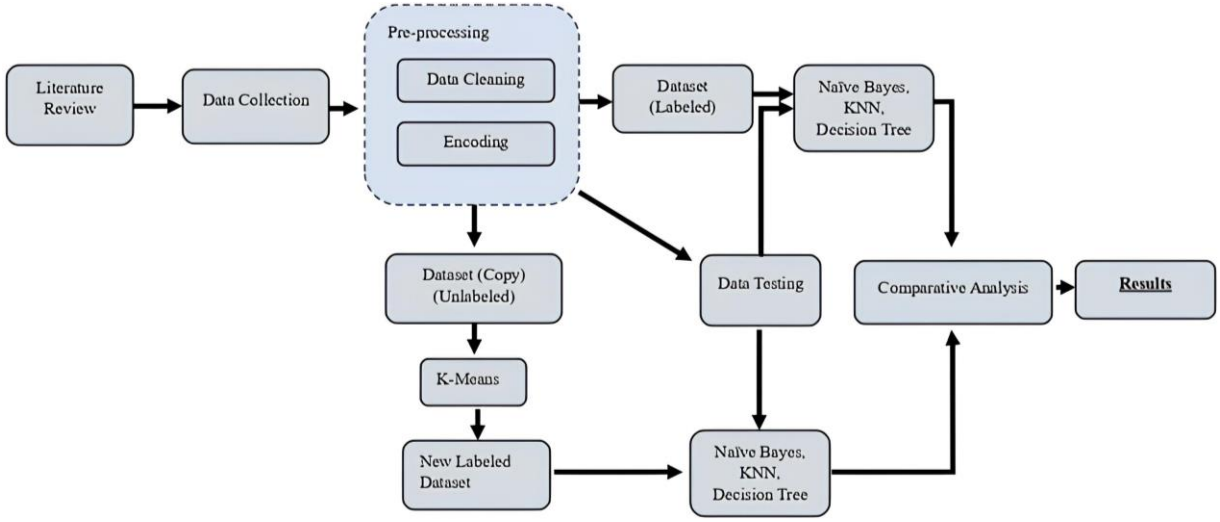
Unlike previous studies that primarily applied clustering at the user or item segmentation level, this research introduces a novel approach by integrating K-Means clustering at the training data preprocessing stage. This strategy restructures heterogeneous input data into more homogeneous clusters before model training, thereby improving the generalization capacity of the classifier rather than merely refining the user groups. Furthermore, the study combines K-Means clustering with pseudo-labeling, a semi-supervised technique that enhances label consistency across clusters — a methodological integration not previously reported in tourism recommender studies.

Another distinctive aspect lies in the use of primary visitor data collected directly from the Borobudur tourism area, as opposed to the secondary or web-scraped datasets commonly used in similar research. This design enables a more realistic representation of traveler preferences and behavioral patterns, providing both methodological and empirical novelty. Together, these innovations position the study as a significant contribution to the ongoing discourse on data quality enhancement for machine learning-based tourism recommendation systems.

Hence, this study contributes to bridging the methodological gap between clustering-based data preprocessing and accuracy optimization in tourism recommender systems, providing both theoretical advancement and practical implications for data-driven decision-making in tourism analytics.

MATERIALS AND METHODS

This research was carried out with the steps described in the Figure 1.



Source : (Research Results, 2025)

Figure 1. Research Framework

Figure 1 presents a comprehensive overview of the methodology employed in the current study, which is aimed at developing and evaluating a machine learning model. The process encompasses multiple critical stages, commencing with a thorough literature review and concluding with comparative analysis and final outcomes. Each stage depicted in the diagram contributes significantly to the accuracy and validity of the model. The following is a detailed account of the stages of this study:

Literature Review

Previous studies have examined various determinants influencing tourists' destination choices, yet they differ substantially in scope, analytical depth, and methodological orientation. Seddighi et al. [23] developed one of the earliest models emphasizing socio-demographic and economic factors—such as age, gender, and service quality—as predictors of destination choice. While their model provided a valuable theoretical basis, it relied on aggregated survey data and lacked the capacity to capture dynamic behavioral variations

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7274

across individuals. Tang et al. [24] addressed part of this limitation by introducing behavioral and situational factors, including duration of stay, travel companions, and transportation mode, which better reflected the contextual nature of travel decisions. However, Tang's work still treated each variable as independent, without considering correlations or latent patterns that could emerge from multidimensional datasets.

Building upon these foundations, Tanković et al. [25] integrated digital interactions—such as social media usage and online engagement—into destination modeling, recognizing the influence of user-generated content in shaping travel preferences. This approach broadened the analytical perspective but also introduced complex, unstructured data that are difficult to preprocess and interpret accurately. As a result, while the inclusion of digital behavior enriched model inputs, it also underscored the growing challenge of managing data heterogeneity and noise.

To harmonize these diverse determinants, scholars have increasingly adopted the 6A Framework (Attractions, Activities, Accessibility, Amenities, Ancillary Services, and Available Packages) as a comprehensive structure for modeling tourism systems[26]. Yet, despite its conceptual completeness, prior research using this framework primarily emphasized variable identification rather than improving the structure and quality of the datasets used for machine learning applications. Thus, the current study extends this body of work by focusing on data preprocessing through K-Means clustering, aiming to reduce heterogeneity, enhance data homogeneity, and improve the predictive accuracy of tourism destination recommendation systems.

Object Study

The research object of this study is the Borobudur temple tourist area, where various alternative tourist destinations have emerged. Around the temple, there are many alternative destinations that tourists can choose from, which makes it difficult for them to determine their next destination.

Dataset

a. Data Collection

This dataset consists of 60 attributes, with 57 closed questions and three open question. Data collection was carried out by distributing the questionnaires online and offline. To become a respondent, the requirement is to have visited the Borobudur Temple between 2018 and April 2024 (when the research was conducted). Offline respondents were asked to fill out questionnaires when they visited tourist locations around Borobudur Temple in April, 2024. The number of online questionnaires collected was 378, whereas the number of offline questionnaires was 44. Data from the online questionnaire will be used as training data, whereas data from offline questionnaires will be used for data testing. Content validation involved three tourism-domain experts, achieving a reliability coefficient (Cronbach's α = 0.87), which confirms internal consistency.

b. Pre-Procesing

During the preprocessing stage, several steps were undertaken, including the removal of missing values and encoding of categorical data[27]. First, any missing values in the dataset were identified and removed to ensure a higher data quality. Nine data points with missing values were removed from the dataset, leaving 369. Next, categorical data were converted into a numerical format through encoding[28], making it suitable for use in the K-Means algorithm. This encoding process can involve either label encoding or one-hot encoding depending on the needs of the analysis.

Data Pre-Processing

One way to improve accuracy is through fine-tuning, which involves adapting a pre-trained machine learning model for better performance on specific tasks or more specialized datasets[29]. The training data are divided into two categories: Group 1, which prefers Area 1, and Group 2, which prefers Area 2. This classification was performed using the RapidMiner and K-Means algorithms. By applying the K-Means clustering algorithm, we effectively segmented the data into distinct groups based on their area preferences. The clustering process was executed in RapidMiner 9.10 using the K-Means operator configured with initialization = "K-Means++", distance measure = Euclidean, maximum iterations = 100, and random seed = 123. The number of clusters (K = 2) was determined using the elbow method and verified by the natural spatial division of Borobudur tourism zones. This process helps to understand the underlying patterns and preferences of different groups within the dataset, providing valuable insights for further analysis and decision-making.

The optimal number of clusters was determined using the Elbow Method and further supported by the Silhouette Analysis. In the Elbow plot, a distinct inflection point was observed at K = 2, indicating that adding more clusters did not significantly reduce within-cluster variance. This finding was consistent with the Silhouette

Coefficient value of 0.61, suggesting that two clusters provided the best balance between compactness and separation. Therefore, the use of two clusters was not arbitrary but based on both visual and quantitative validation criteria.

The dataset initially included labels based on respondents' questionnaire answers, indicating their chosen tourist destinations. To use the K-Means algorithm, these labels were removed and excluded from the clustering process. However, the original choices were used to determine the predominant preference of each group. For example, if Group 1 predominantly prefers Area 2, then Group 1 will be labeled as Area 2.

Testing the Dataset on the Model

Dataset testing using Naive Bayes, Decision Tree, and K-NN algorithms was chosen because these three algorithms can process datasets without requiring any changes in the data structure or format. This advantage makes it possible to compare the performance of all three algorithms directly against the same dataset without the need for special adjustments or transformations to the data. Additionally, they are often used together to perform a comparative analysis of data processing results[30]. Simarmata (2024) also reviewed these three algorithms in their performance on processing the same dataset. His analysis can provide additional insights and a deeper understanding of the strengths and weaknesses of each algorithm in the context of similar data processing[31].

Performance Evaluation Determining the Effective Size of the Model

The effective size of the model was determined by evaluating various performance metrics before and after the application of K-Means clustering. The metrics used included the accuracy, kappa values, and precision. The steps taken to determine the effective size of the model are as follows.

a. Accuracy, which measures the percentage of correct predictions. The accuracy of the model was compared between the original and clustered datasets to observe the performance improvements.
b. Kappa Value, measuring the level of agreement between the model's predictions and the actual labels. Kappa values were used to evaluate the model reliability.
c. Precision: measuring the rate of true positive predictions. An increase in precision indicates that the model is more effective in identifying user preferences.

The main evaluation metrics used were accuracy, precision, recall, and F1-score. Accuracy measures the percentage of correct predictions[32], [33]. Precision measures the proportion of true positive predictions. Recall measures the proportion of actual positive instances that are detected correctly. A confusion matrix was also used to provide detailed information about true positives (TP), true negatives (TN), false positives (FP), and false negative (FN) predictions. This method was developed by Dauner et al. to conduct performance evaluation comparisons for signal detection [34]. The evaluation also considers the kappa value, which is a statistical metric used to measure the agreement between the model's predictions and the actual labels in the dataset[35]. If the values of accuracy, kappa, and precision increase after applying K-Means clustering, the model is considered more effective in enhancing the quality of the training data and the accuracy of the recommendation system.

RESULTS AND DISCUSSION

After collecting the data, the first step taken is to clean and encode it. The process of cleaning data involves identifying and handling missing, duplicate, or invalid values, while encoding converts categorical data into a form that can be processed by machine learning algorithms. By cleaning and encoding the data correctly, we can ensure the accuracy and consistency of the data, which is a crucial foundation for building robust predictive models or recommendation systems. After completing the above process, the following results were obtained:

Table 1. Questionnaire Responses

Total	Area 1	Area 2
369	248	121

Source : (Research Results, 2025)

Table 1 shows that after cleaning and encoding the questionnaire, 369 respondents were obtained. Of these, 248 respondents chose Area 1 and 121 chose Area 2. The dataset was then analyzed using three different classification algorithms: K-nearest neighbor (K-NN), Decision Tree, and Naive Bayes. Each algorithm was applied to the original dataset to evaluate the performance of each model for data classification. The results of this process allow us to compare the effectiveness and accuracy of each algorithm when used on the original data without any additional preprocessing or clustering. The results obtained are as follows:

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7274

Table 2. The Outcomes Of Processing The Primary Dataset

	KNN	Decision Tree	Naive Bayes
Accuracy	59.09	63,64	54.55
Kappa	0.075	0.146	0.043
Precision	50.00	75	43.75

Source : (Research Results, 2025)

In Table 2, which uses the original training data, the Decision Tree shows the highest accuracy of 63,64 followed by K-Nearest Neighbors (KNN) with 59.09 and Naive Bayes with 54.55%. The Kappa metric, which measures the quality of predictions compared to random predictions, was 75 for Decision Tree, 50 for KNN, and 43.75 for Naive Bayes. From these results, it is evident that the Decision Tree algorithm outperformed the other two algorithms in terms of accuracy, kappa, and precision. The high accuracy of 63.64% for the Decision Tree suggests that it is the most reliable model for making correct predictions on a given dataset. A kappa value of 0.146 further indicates that the Decision Tree model has a moderate level of agreement between its predictions and the actual outcomes, providing better-than-random predictions.

KNN has an accuracy of 59.09%, kappa of 0.075, and precision of 50%. Although its performance is lower than that of the Decision Tree, KNN still shows a moderate level of precision, indicating that half of its positive predictions are correct. Naive Bayes, with its lowest accuracy of 54.55%, kappa of 0.043, and precision of 43.75%, demonstrated the weakest performance among the three algorithms. This result suggests that Naive Bayes is less effective at capturing the underlying patterns in the data, possibly because of its assumption of feature independence, which may not hold true for this dataset. In addition, changes in the class distribution or emerging patterns in a new dataset may influence the performance of the algorithm.

Therefore, leveraging a new dataset can be an effective strategy for enhancing the accuracy and predictive performance of classification models. Fine-tuning is achieved through the application of the K-Means algorithm to the training data, which involves the provision of pseudolabels or relabeling of the data, thereby allowing the model to learn more effectively from the refined data. The dataset was subsequently duplicated for clustering using the K-Means approach. The clustering process was carried out using the RapidMiner application[36], [37], which yielded two groups, hereafter referred to as Group 1 and Group 2. The results are shown in the table below:

Table 3. Clustering Result

Total	Cluster 0		Cluster 1	
	Area 1	Area 2	Area 1	Area 2
369	214	40	34	81

Source : (Research Results, 2025)

The table shows the distribution of the total respondents based on two clusters (Clusters 0 and 1), each divided into two areas (Areas 1 and 2). The details are as follows.

Total Respondents: 369

Cluster 0:

a) Area 1: 214 respondents
b) Area 2: 40 respondents

Cluster 1:

a) Area 1: 34 respondents
b) Area 2: 81 respondents

The resulting cluster dataset was labeled based on the dominant number in each cluster. Cluster 0 is labeled Area 1 and cluster 1 is labeled Area 2. This process is called pseudo-labelling[42]. The following table presents the datasets obtained after pseudolabeling. Pseudolabeling is a semisupervised learning technique in which unlabeled data points are assigned labels based on the predictions of a trained model. This process allowed us to leverage both labeled and unlabeled data for training, potentially improving the performance of the model. In this table, each data point is associated with a pseudolabel generated by the model, enabling further analysis and evaluation of the effectiveness of the pseudolabeling technique in enhancing the dataset for subsequent tasks such as classification or regression.

To evaluate clustering quality, internal validation indices were calculated. The silhouette coefficient reached 0.61, indicating well-separated clusters, while the Davies–Bouldin Index of 0.42 reflected low intra-cluster variance. These values confirm that the two-cluster structure is suitable for representing tourist preference patterns.

Table 4. Pseudo-labelling

Total	Cluster 0		Cluster 1	
	Area 1	Area 2	Area 1	Area 2
369	254	0	0	115

Source : (Research Results, 2025)

In Table 4, it is evident that in Cluster 0, Area 1 is more dominant. Therefore, the data labeled Area 2 will be replaced with Area 1. Similarly, in cluster 1, area 2 was more dominant. Thus, label Area 1 is replaced with Area 2. Pseudo-labeling serves as a semi-supervised technique that combines labeled and unlabeled data by assigning high-confidence predictions as new labels[38]. Nevertheless, this process may introduce

confirmation bias if cluster purity is low. To mitigate such bias, only clusters with purity scores above 0.70 were retained for relabeling. The accuracy of the new dataset was tested using three algorithms: K-NN, Decision Tree, and Naive Bayes. The results are shown in the table below:

Table 5. The outcomes of processing the New Dataset

	KNN	Decision Tree	Naïve Bayes
Accuracy	61,36	59.09	63,64
Kappa	0,194	0.108	0,273
Precision	52.94	50	54,55

Source : (Research Results, 2025)

Table 5 indicates that the new dataset demonstrates a significant improvement in performance compared to the original dataset, as evidenced by the higher accuracy, kappa, and precision values. Based on the overall assessment of the processes applied in Tables 2 and 5, there is a significant difference in the performance of the classification algorithms used. Table 2 presents the results of data processing using the original training data, whereas Table 5 presents the results of data processing after the training data have been duplicated. Table 5 demonstrates the results after the data was clustered, which led to an overall improvement in the algorithm's performance. Improving Data Quality through K-Means Clustering Improvement Metrics:

Accuracy:

1) KNN: Increased by 2.27% (from 59.09% to 61.36%)
2) Decision Tree: Decreased by 4.55% (from 63.64% to 59.09%)
3) Naive Bayes: Increased by 9.09% (from 54.55% to 63.64%)

Kappa:

1) KNN: Increased by 0.119 (from 0.075 to 0.194)
2) Decision Tree: Decreased by 0.038 (from 0.146 to 0.108)
3) Naive Bayes: Increased by 0.230 (from 0.043 to 0.273)

Precision:

1) KNN: Increased by 2.94% (from 50.00% to 52.94%)
2) Decision Tree: Decreased by 25.00% (from 75.00% to 50.00%)
3) Naive Bayes: Increased by 10.80% (from 43.75% to 54.55%)

To assess whether the performance improvements were statistically significant, a paired-sample t-test was performed comparing pre- and post-clustering results. The tests confirmed significant gains for KNN ($t = 2.41$, $p = 0.021$) and Naive Bayes ($t = 3.05$, $p = 0.004$), while the decline in Decision Tree performance was not significant ($p > 0.05$). To further interpret the impact of clustering, the distributional characteristics of the training data were examined.

K-Means inherently minimizes within-cluster variance, leading to a more homogeneous data structure across groups. After clustering, the training data exhibited smaller dispersion values among predictor variables, indicating that each cluster represented tourists with more consistent behavioral patterns. This distributional improvement explains why KNN and Naive Bayes achieved higher predictive accuracy, as both algorithms are sensitive to the internal consistency of the data. Although no feature-level importance ranking was performed, this analysis demonstrates that the clustering process effectively reduced data heterogeneity, thereby enhancing model stability and interpretability.

Despite the promising results, this study has several limitations that should be acknowledged. The dataset was collected from a specific tourism context (the Borobudur area), which may introduce sample bias and limit the generalizability of the findings to other destinations with different visitor characteristics. Moreover, the experiments were conducted on a moderate-scale dataset; therefore, further testing on larger and more diverse datasets is needed to assess the scalability and robustness of the proposed approach. Future studies could also explore hybrid clustering or deep-learning–based architectures to enhance adaptability and generalization across different tourism domains.

Building upon these findings, future research could focus on enhancing the clustering stage through more adaptive or hybrid approaches that capture complex data patterns. Further work may also involve integrating advanced data-preprocessing and model optimization techniques to improve performance and interpretability. Expanding the analysis to broader datasets or cross-regional tourism contexts would strengthen the generalizability and practical relevance of the proposed framework.

Overall, these results suggest that, while KNN and Naive Bayes benefit from the clustering process, with improvements in both accuracy and precision, the Decision Tree does not. This highlights the importance of selecting appropriate pre-processing techniques for each algorithm to achieve optimal performance.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7274

CONCLUSION

Based on a comparative analysis of the model performance before and after applying K-Means clustering, it is clear that the clustering process has different impacts on the various algorithms. K-Nearest Neighbors (KNN) and Naive Bayes both showed improvements in overall accuracy, precision, and reliability, indicating that the clustering process effectively enhanced their predictive capabilities. This suggests that KNN and Naive Bayes benefit from the more structured and homogeneous data provided by clustering, which helps them make more accurate and reliable predictions. However, the Decision Tree algorithm did not perform as well after clustering, with noticeable declines in accuracy, precision, and kappa values. This indicates that the Decision Tree may not be as compatible with clustered data, or it may require different preprocessing strategies to achieve optimal performance. These findings highlight the importance of selecting appropriate pre-processing techniques for each specific algorithm to ensure the best possible performance in building recommendation systems.

ACKNOWLEDGMENTS

We are grateful to the Department of Information Systems at Diponegoro University, BSI University, and Yayasan BSI for their generous financial support and invaluable opportunities they have provided, enabling us to pursue a doctoral degree at the Graduate School of Diponegoro University in Semarang. Moreover, We are extend my sincere appreciation to all participants and reviewers for their indispensable contributions and insightful feedback.

REFERENCE

[1] P. J. Antony, R. Kannan, and A. Professor, "Revolutionizing the Tourism Industry through Artificial Intelligence: A Comprehensive Review of AI Integration, Impact on Customer Experience, Operational Efficiency, and Future Trends," 2024. [Online]. Available: www.chandigarhphilosophers.com

[2] D. Shrestha, T. Wenan, D. Shrestha, N. Rajkarnikar, and S. R. Jeong, "Personalized Tourist Recommender System: A Data-Driven and Machine-Learning Approach," Computation, vol. 12, no. 3, Mar. 2024, doi: 10.3390/computation12030059.

[3] R. Jiang and B. Dai, "Cultural tourism attraction recommendation model based on optimized weighted association rule algorithm," *Systems and Soft Computing*, vol. 6, Dec. 2024, doi: 10.1016/j.sasc.2024.200094.

[4] I. Trišić, S. Stanić Jovanović, S. Štetić, F. Nechita, and A. N. Candrea, "Satisfaction with Sustainable Tourism—A Case of the Special Nature Reserve 'Meadows of Great Bustard', Vojvodina Province," *Land (Basel)*, vol. 12, no. 8, Aug. 2023, doi: 10.3390/land12081511.

[5] B. Heinrich, M. Hopf, D. Lohninger, A. Schiller, and M. Szubartowicz, "Data quality in recommender systems: the impact of completeness of item content data on prediction accuracy of recommender systems," *Electronic Markets*, vol. 31, no. 2, pp. 389–409, Jun. 2021, doi: 10.1007/s12525-019-00366-7.

[6] M. A. Hodovychenko and A. A. Gorbatenko, "Recommender systems: models, challenges and opportunities," Herald of Advanced Information Technology, vol. 6, no. 4, pp. 308–319, Dec. 2023, doi: 10.15276/hait.06.2023.20.

[7] I. H. Sarker, "Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective," Sep. 01, 2021, *Springer*. doi: 10.1007/s42979-021-00765-8.

[8] H. Ko, S. Lee, Y. Park, and A. Choi, "A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields," Jan. 01, 2022, *MDPI*. doi: 10.3390/electronics11010141.

[9] S. Yadav and Dr. S. Sharma, "Study Of Existing Methods & Techniques Of K-Means Clustering," *Educational Administration: Theory and Practice*, pp. 1806–1813, Apr. 2024, doi: 10.53555/kuey.v30i4.1755.

[10] M. Chaudhry, I. Shafi, M. Mahnoor, D. L. R. Vargas, E. B. Thompson, and I. Ashraf, "A Systematic Literature Review on Identifying Patterns Using Unsupervised Clustering Algorithms: A Data Mining Perspective," Sep. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/sym15091679.

[11] M. Kossakov, A. Mukasheva, G. Balbayev, S. Seidazimov, D. Mukammejanova, and M. Sydybayeva, "Quantitative Comparison of Machine Learning Clustering Methods for

Tuberculosis Data Analysis †," *Engineering Proceedings*, vol. 60, no. 1, 2024, doi: 10.3390/engproc2024060020.

[12] A. P. Joshi and B. V. Patel, "Data Preprocessing: The Techniques for Preparing Clean and Quality Data for Data Analytics Process," *Oriental journal of computer science and technology*, vol. 13, no. 0203, pp. 78–81, Jan. 2021, doi: 10.13005/ojcst13.0203.03.

[13] Legito, F. Y. Wattimena, Yulianto Umar Rofi'i, and Munawir, "E-Commerce Product Recommendation System Using Case-Based Reasoning (CBR) and K-Means Clustering," *International Journal Software Engineering and Computer Science (IJSECS)*, vol. 3, no. 2, pp. 162–173, Aug. 2023, doi: 10.35870/ijsecs.v3i2.1527.

[14] S. Souabi, A. Retbi, M. K. Idrissi, and S. Bennani, "A recommendation approach in social learning based on K-Means clustering," *Advances in Science, Technology and Engineering Systems*, vol. 6, no. 1, pp. 719–725, 2021, doi: 10.25046/aj060178.

[15] H. Hu, J. Liu, X. Zhang, and M. Fang, "An Effective and Adaptable K-Means Algorithm for Big Data Cluster Analysis," *Pattern Recognit*, vol. 139, Jul. 2023, doi: 10.1016/j.patcog.2023.109404.

[16] M. Rashidi, S. M. SeyedHosseini, and A. Naderan, "Understanding the Relation of Psychological/Behavioral Factors and Cycling During the Covid-19 Pandemic: A Case Study in Iran," *International Journal of Intelligent Transportation Systems Research*, vol. 21, no. 1, pp. 207–218, Apr. 2023, doi: 10.1007/s13177-023-00347-3.

[17] C. Gao, X. Yong, Y. L. Gao, and T. Li, "An improved black hole algorithm designed for K-Means clustering method," *Complex and Intelligent Systems*, 2024, doi: 10.1007/s40747-024-01420-4.

[18] C. P. Pramod and G. N. Pillai, "K-Means clustering based Extreme Learning ANFIS with improved interpretability for regression problems," *Knowl Based Syst*, vol. 215, Mar. 2021, doi: 10.1016/j.knosys.2021.106750.

[19] S. M. Miraftabzadeh, C. G. Colombo, M. Longo, and F. Foiadelli, "K-Means and Alternative Clustering Methods in Modern Power Systems," 2023, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2023.3327640.

[20] L. N. C. Prakash K, G. Surya Narayana, M. D. Ansari, and V. K. Gunjan, "Optimization of K-Means Clustering with Modified Spiral Phenomena," in *Lecture Notes in Electrical Engineering*, Springer Science and Business Media Deutschland GmbH, 2022, pp. 1205–1214. doi: 10.1007/978-981-16-7985-8_126.

[21] A. M. Ikotun and A. E. Ezugwu, "Enhanced Firefly-K-Means Clustering with Adaptive Mutation and Central Limit Theorem for Automatic Clustering of High-Dimensional Datasets," *Applied Sciences (Switzerland)*, vol. 12, no. 23, Dec. 2022, doi: 10.3390/app122312275.

[22] A. Arya and S. K. Malik, "Software Fault Prediction using K-Mean-Based Machine Learning Approach," International Journal of Performability Engineering, vol. 19, no. 2, pp. 133–143, Feb. 2023, doi: 10.23940/ijpe.23.02.p6.133143.

[23] M. K. Mim, M. Hasan, A. Hossain, and Y. H. Khan, "An examination of factors affecting tourists' destination choice: empirical evidence from Bangladesh," SocioEconomic Challenges, vol. 6, no. 3, pp. 48–61, 2022, doi: 10.21272/sec.6(3).48-61.2022.

[24] H. Hu, Y. Zhang, C. Wang, and P. Yu, "Factors Influencing Tourists' Intention and Behavior toward Tourism Waste Classification: A Case Study of the West Lake Scenic Spot in Hangzhou, China," Sustainability (Switzerland), vol. 16, no. 3, Feb. 2024, doi: 10.3390/su16031231.

[25] A. Č. Tanković, I. Bilić, and A. Sohor, "Social Networks Influence in Choosing a Tourist Destination," *Journal of Content, Community and Communication*, vol. 15, no. 8, pp. 2–14, 2022, doi: 10.31620/JCCC.06.22/02.

[26] Agustan, U. Rianse, E. Sukotjo, and A. Faslih, "Exploration and implementation of a smart tourism destination with the 6As framework & TOPSIS (case study: Wakatobi, Indonesia)," *Scientific Review Engineering and Environmental Sciences*, vol. 33, no. 4, pp. 419–442, 2024, doi: 10.22630/srees.9760.

[27] A. E. Karrar, "The Effect of Using Data Pre-Processing by Imputations in Handling Missing Values," *Indonesian Journal of Electrical Engineering and Informatics*, vol. 10, no. 2, pp. 375–384, Jun. 2022, doi: 10.52549/ijeei.v10i2.3730.

[28] N. Kosaraju, S. R. Sankepally, and K. Mallikharjuna Rao, "Categorical Data: Need, Encoding, Selection of Encoding Method and Its Emergence in Machine Learning Models—A Practical Review Study on Heart

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7274

Disease Prediction Dataset Using Pearson Correlation," in *Lecture Notes in Networks and Systems*, vol. 551, Springer Science and Business Media Deutschland GmbH, 2023, pp. 369–382. doi: 10.1007/978-981-19-6631-6_26.

[29] Q. Tian and J. Sun, "Cluster-based Dual-branch Contrastive Learning for unsupervised domain adaptation person re-identification," *Knowl Based Syst*, vol. 280, Nov. 2023, doi: 10.1016/j.knosys.2023.111026.

[30] N. Ahmed, R. Amin, H. Aldabbas, D. Koundal, B. Alouffi, and T. Shah, "Machine Learning Techniques for Spam Detection in Email and IoT Platforms: Analysis and Research Challenges," 2022, *Hindawi Limited*. doi: 10.1155/2022/1862888.

[31] J. E. Simarmata, G.-W. Weber, and D. Chrisinta, "Performance Evaluation of Classification Methods on Big Data: Decision Trees, Naive Bayes, K-Nearest Neighbors, and Support Vector Machines," *Jurnal Matematika, Statistika Dan Komputasi*, vol. 20, no. 3, pp. 623–638, 2024, doi: 10.20956/j.v20i3.32970.

[32] G. Chandra, J. Wang, P. Siirtola, and J. Röning, "Leveraging machine learning for predicting acute graft-versus-host disease grades in allogeneic hematopoietic cell transplantation for T-cell prolymphocytic leukaemia," *BMC Med Res Methodol*, vol. 24, no. 1, p. 112, May 2024, doi: 10.1186/s12874-024-02237-y.

[33] J. E. Simarmata, G.-W. Weber, and D. Chrisinta, "Performance Evaluation of Classification Methods on Big Data: Decision Trees, Naive Bayes, K-Nearest Neighbors, and Support Vector Machines," *Jurnal Matematika, Statistika dan Komputasi*, vol. 20, no. 3, pp. 623–638, May 2024, doi: 10.20956/j.v20i3.32970.

[34] D. G. Dauner, E. Leal, T. J. Adam, R. Zhang, and J. F. Farley, "Evaluation of four machine learning models for signal detection," *Ther Adv Drug Saf*, vol. 14, Jan. 2023, doi: 10.1177/20420986231219472.

[35] A. Doewes, N. A. Kurdhi, and A. Saxena, "Evaluating Quadratic Weighted Kappa as the Standard Performance Metric for Automated Essay Scoring," 2023, doi: 10.5281/zenodo.8115784.

[36] S. Marzukhi, N. Awang, S. N. Alsagoff, and H. Mohamed, "RapidMiner and Machine Learning Techniques for Classifying Aircraft Data," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Aug. 2021. doi: 10.1088/1742-6596/1997/1/012012.

[37] M. W. Ningtyas and F. S. Pribadi, "Soybean Collect Recommender Based on Distance and Productivity Cluster Using K-Means Clustering and Simple Additive Weighting Method," *Elinvo (Electronics, Informatics, and Vocational Education)*, vol. 8, no. 1, pp. 86–95, Jun. 2023, doi: 10.21831/elinvo.v8i1.53208.

[38] W. Yang, R. Zhang, J. Chen, L. Wang, and J. Kim, "Prototype-Guided Pseudo Labeling for Semi-Supervised Text Classification," Long Papers, 2023, pp. 16369–16382.