Journal of Soft Computing Exploration
Homepage: htttps://shmpublisher.com/index.php/joscex
p-ISSN: 2746-7686
e-ISSN: 2746-0991

.

Hybrid approach for identifying strategic promotional locations using k-means clustering and support vector machine classification

Anisya[1], Brestina Gultom[2], Sarjon Defit[3]
[1]Department of Informatics Engineering, Institut Teknologi Padang, Indonesia
[2]Department of Information System, Universitas Adiwangsa Jambi, Indonesia
[3]Department of Information Technology, Universitas Putra Indonesia "YPTK" Padang, Indonesia

Article Info

Article history:

Received April 5, 2026
Revised April 26, 2026
Accepted April 28, 2026

Keywords:

Hybrid model
Strategic promotion
Location-based marketing
Geospatial analysis

ABSTRACT

In the increasingly competitive landscape of higher education marketing, determining strategic promotional locations was essential to reaching prospective students effectively. This study proposed a hybrid machine learning framework combining K-Means clustering and Support Vector Machine (SVM) classification to identify high-potential areas for targeted promotional activities. The analysis used student enrolment data from 2021 to 2024, focusing on features such as city, province, and school origin. K-Means clustering was first applied to segment the data into three spatially and institutionally distinct clusters. These clusters were then used as pseudo-labels to train the SVM model, enabling the classification of new data points based on learned patterns. The model achieved a classification accuracy of 98%, with consistently high precision and recall across all clusters. Cluster interpretation revealed meaningful geographic and institutional differences that supported differentiated promotional strategies. Thematic map visualizations further enhanced the applicability of the model for geospatial decision-making. This study contributed to the development of data-driven, scalable, and interpretable solutions for location-based marketing. It also demonstrated the practical relevance of hybrid learning models in supporting strategic planning for educational institutions. Future work was suggested to incorporate additional socio-demographic variables and advanced ensemble methods to improve model robustness.

Corresponding Author:

Anisya,
Department of Informatics Engineering,
Padang Institute of Technology,
Gajah Mada Road, Kandis Nanggalo, Padang, Indonesia.
Email: anisya@itp.ac.id
https://doi.org/10.52465/joscex.v7i2.45

1. INTRODUCTION

In an increasingly competitive environment, an appropriate promotional strategy is crucial for institutions to effectively reach their target audiences [1]. The success of a promotional campaign often depends not only on the content or media used but also on the selection of location [2]. A strategic location can increase

promotional reach, engagement [3], and ultimately, conversion rates whether at the city, district, or school level. However, identifying effective promotional locations remains a complex task, particularly across large geographic areas with diverse socio-demographic characteristics. Traditional methods for selecting these locations often rely heavily on intuition [4], historical performance data [5], or basic demographic analyses [6]. While practical, these approaches tend to lack analytical precision and may fail to uncover deeper latent patterns in the data.

To address these limitations, the use of data-driven decision-making is becoming increasingly common, particularly through the application of machine learning (ML) techniques. These methods enable the extraction of meaningful information from large, often unstructured datasets [7]. Among the various ML techniques, clustering algorithms such as K-Means are widely used to group spatial data [8] or customers based on similarity [9]. As an unsupervised learning method, K-Means helps uncover the inherent structure of a dataset by partitioning it into clusters in which each observation belongs to the cluster with the nearest mean (centroid) [10]. This method has proven particularly useful in spatial and demographic analysis, where segmentation is based on location-specific attributes like geographic coordinates, population density, socioeconomic indicators, or service accessibility [11].

Numerous studies have validated the effectiveness of K-Means in spatial segmentation. For instance, Zhang et al. [12] applied K-Means to categorize urban zones based on environmental and socio-demographic characteristics to support smart city planning. Similarly, Fahmiyah [13] used K-Means to classify Indonesian provinces based on education and health indicators, aiding policy prioritization. These examples highlight the algorithm's flexibility in organizing unlabeled data to produce meaningful regional clustering. However, despite its popularity, K-Means is not without limitations. The algorithm is sensitive to initial centroid selection [14], assumes spherical cluster shapes, and has limited predictive capability [15].

Supervised learning techniques, such as Support Vector Machines (SVM), can be used to overcome the limitations of clustering in prediction tasks. SVM is a powerful classification algorithm known for constructing optimal separating hyperplanes in high-dimensional spaces. Its robustness in handling sparse data and strong generalization performance make it well suited for predictive analytics. In the context of marketing and promotion, SVM has been used to predict consumer behavior, classify marketing prospects, and forecast sales trends [16]. In education, SVM has been applied to categorize schools or students based on performance metrics and specific characteristics.

The main strength of SVM lies in its kernel trick, which enables non-linear classification by mapping input features into higher-dimensional spaces [17]. This is especially valuable in dealing with real-world socio-geographic data, where class boundaries are rarely linearly separable. Furthermore, combining SVM with clustering algorithms can yield synergistic benefits. In such a hybrid approach, K-Means serves as a preprocessing step to group unlabelled data, while SVM uses these clusters as pseudo-labels to learn patterns and classify new data points.

This hybrid approach has been applied in various studies. Topaloglu proposed a hybrid K-Means–SVM model to improve classification accuracy in text mining applications [18]. Similarly, Nasser [19] used K-Means to cluster web users based on browsing patterns, followed by SVM to predict user preferences. In the geospatial domain, this combination has been applied to land use classification, urban planning, and crime pattern prediction, demonstrating its potential to enhance both exploratory and predictive capabilities. However, its application in marketing and promotional planning remains relatively underexplored.

This study aims to bridge that gap by proposing a hybrid analytical framework that combines K-Means clustering with SVM classification to identify high-potential locations for educational promotional efforts. Specifically, the research uses student enrolment data from 2021–2024 collected by the Padang Institute of Technology. The K-Means algorithm was first applied to segment schools and regions into three distinct clusters based on features such as school origin, city, and province. These clusters reflect varying characteristics, such as urban vs. rural settings and differences in historical promotional effectiveness that are relevant to campaign planning.

Following the clustering, the SVM model was trained using the resulting clusters as pseudo-labels [20]. This allowed the system to learn the distinguishing features of each cluster and apply that knowledge to classify new or unseen data. In doing so, the model effectively predicts the promotional potential of other locations, offering a data-driven basis for future campaign planning [21]. Accordingly, this sequential combination of unsupervised and supervised learning can enhance analytical depth and actionability [22].

The novelty of this approach lies in its two-stage analytical process: first, segmenting the dataset based on latent spatial-demographic characteristics, and second, using that structure to inform predictive

classification. By integrating K-Means and SVM, this framework leverages the exploratory power of clustering and the predictive accuracy of classification. The result is a robust system capable of identifying promising new locations that share attributes with previously successful targets, thereby improving campaign precision and reducing reliance on subjective decision-making.

To validate the effectiveness of the proposed framework, a case study was conducted using spatial and demographic data from an actual promotional campaign. Model performance was evaluated using established classification metrics such as precision, recall, and F1-score. The results demonstrate that the hybrid K-Means–SVM model can significantly improve the strategic planning of promotional activities by uncovering hidden patterns and enabling accurate predictions.

Theoretically, this research contributes to the development of hybrid machine learning models by demonstrating their applicability in geospatial and marketing contexts. From a practical perspective, the proposed framework can be implemented by educational institutions to improve the effectiveness of promotional strategies. In an increasingly data-driven era, this approach provides added value in optimizing the reach, efficiency, and impact of promotional activities.

2. METHOD

This study employs a structured workflow for data-driven analysis that integrates both unsupervised and supervised learning approaches. Figure 1 illustrates the steps involved in this study. The process begins with data collection and preprocessing, including data coding and cleaning. Next, clustering is performed using the K-Means algorithm to identify inherent patterns in the data, with the optimal number of clusters determined prior to implementation and interpretation. The clustering results are then utilized in the classification stage using the SVM algorithm, which involves labeling, feature selection, as well as model training and validation. To assess the effectiveness of the proposed model, performance evaluation is conducted using standard metrics, including accuracy, precision, recall, and F1-score. Finally, the results are presented through model visualization and are implemented to support practical decision-making, demonstrating the applicability of the framework in real-world scenarios.

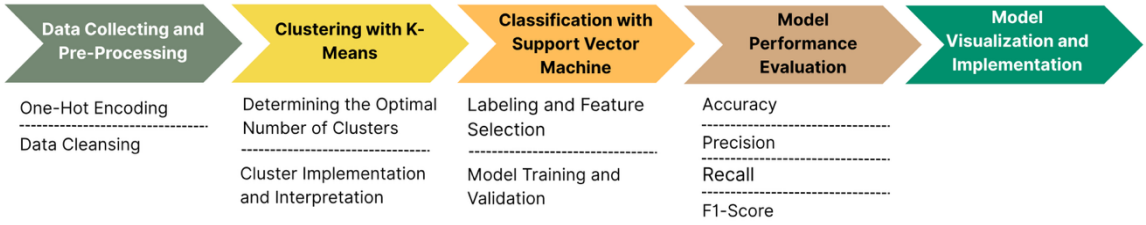


Figure 1. Hybrid methodological framework: unsupervised dan supervised learning

Data Collection and Preprocessing

The dataset used in this study consists of 2,537 records of new student data collected between 2020 and 2024, spanning various cities and provinces across Indonesia. The data were obtained from institutional records and publicly available sources. Each record includes the student's city of origin, province, and Origin School. An overview of the data that was successfully collected is presented in Table 1. After collection, a preprocessing stage was conducted to remove data that were not suitable for further analysis.

This dataset was selected because it encompasses both spatial (geographic location) and social (institution of origin) dimensions, which are highly relevant for modeling distribution patterns and identifying potential target areas for promotional strategies. In this context, schools that consistently contribute a significant number of students are considered potential targets for higher education marketing initiatives.

The preprocessing stage aims to transform the raw data into a structured format suitable for machine learning modeling [23], [24]. This stage consists of two main components: categorical feature transformation and data cleaning. First, categorical variables such as city, province, and school name are converted into numerical representations using one-hot encoding. This approach transforms each category into a binary vector, thereby avoiding the introduction of artificial ordinal relationships between categorical values and ensuring compatibility with distance-based clustering methods such as K-Means.

Table 1. Datasets cross-section

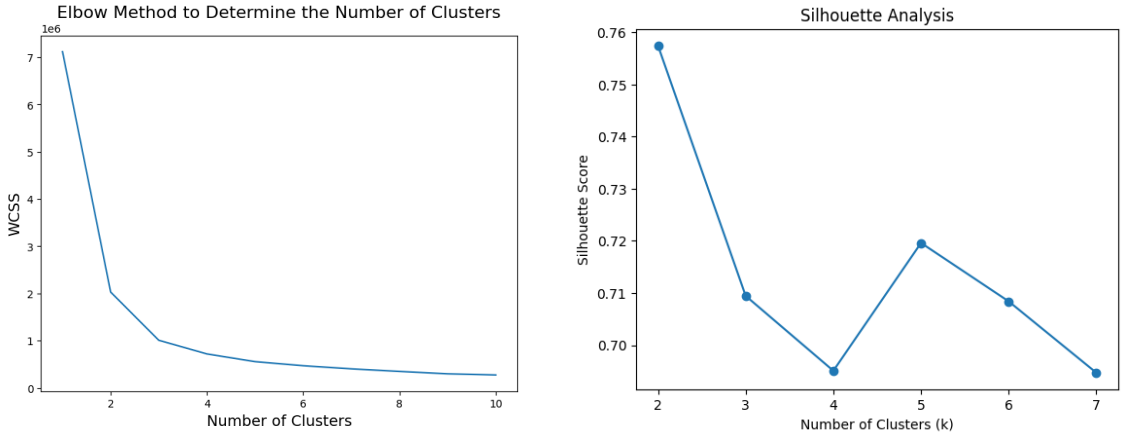
No	Student ID	Name	Gender	Origin City	Origin Province	Department	Degree	Origin School
1	2020310001	MAULANA MALIK IBRAHIM	Male	Padang City	West Sumatera Province	Electrical Engineering	Bachelor	State Vocational High School 1 Padang
2	2020310002	AGUS M. YASIR	Male	Pasaman Regency	West Sumatera Province	Electrical Engineering	Bachelor	State Vocational High School 1 Pasaman

3	2020310003	AKMAL ZAINI	Male	Sawahlunto/ Sijunjung Regency	West Sumatera Province	Electrical Engineering	Bachelor	State Vocational High School 2 Solok
4	2020310004	ANANDA ERINALDO	Male	South Coast Regency	West Sumatera Province	Electrical Engineering	Bachelor	State Vocational High School 1 Ranah Pesisir
5	2020310005	DIAN TAUFANI	Female	Solok Regency	West Sumatera Province	Electrical Engineering	Bachelor	State Senior High School 1 Hiliran Gumanti

Furthermore, data cleaning was performed, including the removal of duplicate data, handling of missing values, and standardization of location names to ensure consistency across the dataset . These steps help reduce potential errors during processing [25]. At this stage, no feature normalization or scaling was applied. This decision was based on the nature of the encoded features, where the one-hot representation is already on a comparable scale, making additional normalization unnecessary for effective clustering and classification [26].

Clustering with K-Meansss

The K-Means algorithm is used to discover latent structures or hidden groupings within the data [27] The goal of this study is to cluster schools or institutions based on similarities in geographic location and student origin. This method was chosen due to its scalability for large datasets, computational efficiency, and interpretability in the context of promotional planning [28]. Determining the appropriate number of clusters (k) is a critical step in effective clustering. In this study, the Elbow Method was used to determine the optimal value of k. This method involves calculating the Within-Cluster Sum of Squares (WCSS) for various values of k. The point at which the rate of WCSS decrease slows significantly, forming an “elbow” shape in the graph, is considered the optimal number of clusters. In this case, k = 3 was selected as the optimal value, as it represents the inflection point where additional clusters yield only marginal improvements in segmentation quality.



(a) (b)

Figure 1. Determining the number of clusters based on (a). Elbow method, (b) Silhouette analysis

Figure 2 (a) illustrates the application of the Elbow Method to determine the optimal number of clusters in the K-Means algorithm. The horizontal axis represents the number of clusters (k), while the vertical axis shows the Within-Cluster Sum of Squares (WCSS), which is the sum of squared distances between each data point and its cluster centroid, representing cluster compactness. The graph shows that the WCSS value decreases significantly from k = 1 to k = 3, and then tends to plateau after k > 3. This pattern indicates an “elbow” point at k = 3, suggesting that increasing the number of clusters beyond this point does not provide significant improvement in reducing within-cluster variation.

Additional evaluation was conducted using the Silhouette Score., To strengthen these results, Figure 2(b) displays the Silhouette Score values for each cluster configuration. Although k = 5 produces a slightly higher Silhouette Score compared to k = 3, this improvement is marginal and does not justify the increased complexity of the clustering structure. The Elbow Method clearly indicates k = 3 as the point at which the WCSS reduction rate begins to stabilize, suggesting that the primary data structure is sufficiently represented at this level. Furthermore, selecting k = 3 improves interpretability and better aligns with the research objective of identifying meaningful regional clustering for strategic decision-making. The Elbow Method remains valid even when the Silhouette Score is slightly higher, especially when the difference is small and interpretability is more important [29]. Therefore, k = 3 is considered the most appropriate number of clusters.

The clustering process was performed using two main features: City ID and Province ID, which represent the spatial dimension of student origin. Each data point was then assigned to one of three clusters (Cluster 0, 1, or 2). The resulting clusters were analyzed based on their geographic distribution, revealing differences in regional characteristics. Some clusters were dominated by areas with high student concentrations, while others reflected regions with lower contributions. This pattern provides an initial indication of regional variation in the potential for higher education promotion and serves as an important basis for further analysis.

Classification with Support Vector Machine (SVM)

To enable the system to predict cluster membership for new, previously unseen locations, a classification phase is performed. An SVM model is used to learn the patterns generated by the K-Means clustering process and generalize them to new data instances. This approach transforms the system from a purely descriptive model into a predictive model. Pseudo-labels are derived from the K-Means clustering results and treated as ground truth labels. The features used in the classification process remain consistent: City ID, Province ID, and School ID. In the literature, this approach is known as pseudo-labeling, which is commonly applied when manual labeling is unavailable. It allows supervised learning techniques to be applied to initially unlabeled data.

In the model training and validation phase, the preprocessed data is randomly divided into two subsets: 80% for training and 20% for testing. This split ensures that the trained model not only learns patterns from the training data but also generalizes effectively to unseen data. Support Vector Machine (SVM) was chosen due to its ability to handle high-dimensional data and its strong classification performance. The training process involves a grid search to determine the optimal parameter combination, specifically the regularization parameter C, which controls the margin and penalizes misclassifications. A small value of C may lead to underfitting by allowing too many classification errors, while a large value of C may lead to overfitting by making the model too sensitive to the training data [30]. Therefore, selecting an appropriate value of C is crucial to achieving optimal classification performance. After training, the model is validated using the testing dataset to evaluate its accuracy and generalization performance. The results of this evaluation are discussed in the following sections.

Model Performance Evaluation

Model evaluation was conducted to assess the effectiveness of the Support Vector Machine (SVM) in classifying the test data based on the clusters obtained in the previous stage. The main objective of this evaluation was to measure the model's ability to correctly predict cluster membership and to assess the quality of predictions made on unseen data.

The following evaluation metrics were used in this study, (a) Accuracy measures the overall model performance and is defined as the ratio of correct predictions to the total number of test instances. A high accuracy value indicates strong general classification performance [30]. (b) Precision measures how many of the model's positive predictions are actually correct. This metric is particularly important in contexts where minimizing false positives is crucial, such as identifying strategic promotion locations [31]. (c) Recall (Sensitivity) indicates the proportion of actual positive cases correctly identified by the model. A high recall value signifies that the model effectively captures relevant instances of a given class, thereby minimizing false negatives. F1-Score is the harmonic mean of precision and recall. This metric is especially useful in cases of class imbalance, as it provides a more balanced performance assessment than accuracy alone [32].

These four metrics provide a comprehensive overview of the model's strengths and limitations across different classification dimensions [33], [34]. The results serve as a basis for determining whether the SVM model is suitable for operational use, particularly in supporting strategic decision-making based on clustering results.

Model Visualization and Implementation

As part of the practical implementation, the clustering and classification results are visualized using thematic maps to display the geographic distribution of clusters. This is achieved by aligning City and Province

IDs in a spatial shapefile with the corresponding cluster outputs. This visualizations assist promotion policymakers in (a) Identifying potential location hotspots, (b) Analysing the spatial distribution of responsive clusters, (c) Making evidence-based targeting.

3. RESULTS AND DISCUSSIONS

This section presents the results of each phase of the proposed hybrid framework, from clustering with K-Means to classification with SVM. It also includes visual analyses and interpretations of predicted high-potential promotion areas. The discussion emphasizes evaluating the method's effectiveness in facilitating data-driven strategic decision-making.

K-Means Clustering Results

The application of the K-Means algorithm to the education dataset yielded three main clusters: Cluster 0, Cluster 1, and Cluster 2. Each data point was grouped based on similarities in two numerical attributes: City ID and Province ID. Figure 3 shows three distinguishable groups, indicating that the observations can be effectively clustered. Since the dataset does not contain predefined class labels, the clustering results are interpreted as exploratory groupings rather than ground-truth categories. Therefore, the clusters do not represent supervised classes but instead reflect natural segmentation derived from the feature space. This unsupervised structure provides an initial understanding of regional variation in student origin, which is useful for identifying potential target areas for promotional strategies.

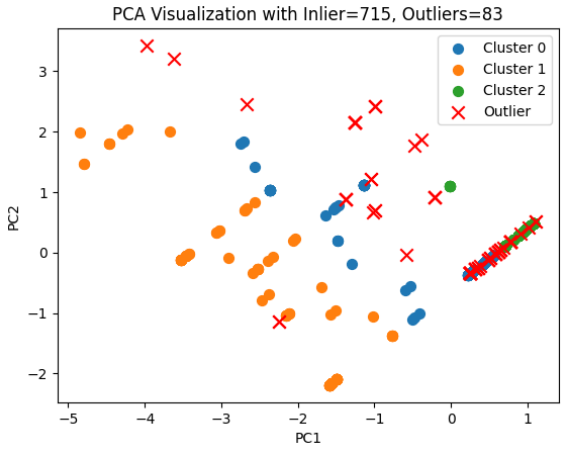


Figure 3. Cluster visualization with outlier detection using LOF + PCA

The cluster visualization shown in Figure 3 illustrates the distribution of data points in the feature space based on City ID and Province ID. In addition to clustering structure analysis, an outlier detection step was performed using the Local Outlier Factor (LOF) method to identify anomalous observations within the dataset. The LOF analysis identified 83 outliers and 715 inliers out of a total of 798 data points. This indicates that approximately 89.6% of the data exhibit consistent neighborhood density patterns, while 10.4% are classified as anomalous or less typical observations. The relatively small proportion of outliers suggests that the dataset is generally well-structured, with a dominant underlying distribution pattern. These outliers are primarily located at the periphery of the data distribution and may correspond to regions or institutions with low or irregular student participation rates. This suggests the presence of underrepresented areas that do not fully follow the dominant recruitment patterns observed in the main data structure. From a practical perspective, these locations may represent potential expansion areas for future promotional strategies. The outlier analysis further reveals that anomalous observations are concentrated in specific cities and schools, which generally show lower student contribution compared to the main cluster distribution. These patterns indicate heterogeneous recruitment behavior and highlight regions with atypical participation trends that may require targeted intervention strategies.

Cluster Interpretation

Each cluster was analyzed to identify its dominant characteristics based on geographic and institutional attributes. Cluster 0 primarily consists of well-established and high-performing schools located in major urban areas, which have historically demonstrated high student conversion rates. Cluster 1 comprises institutions that are geographically more dispersed and have generally shown lower enrollment contributions in previous periods. Cluster 2 represents a more heterogeneous group, consisting of secondary schools with varying levels of promotional effectiveness. This segmentation provides a basis for differentiated promotional strategies. For instance, digital-based outreach strategies may be more effective for urban-oriented Cluster 0, relationship-based engagement with schools may be more suitable for Cluster 1, while mixed or adaptive strategies can be applied to Cluster 2 depending on institutional characteristics.

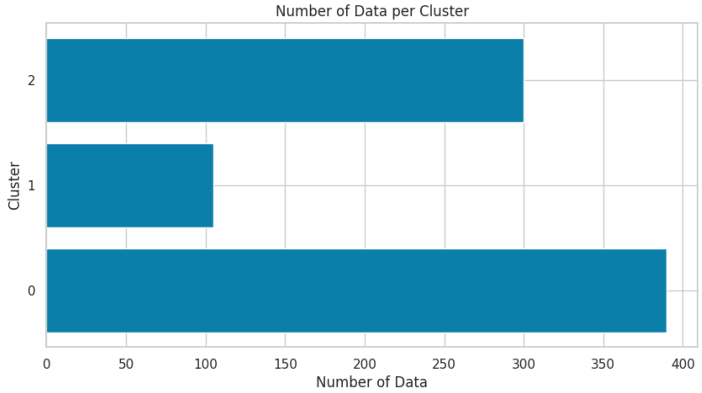


Figure 4. Number of data cluster

The visualization in Figure 4 shows that the data are divided into three main clusters. The K-Means results indicate a significant imbalance among clusters, with Cluster 0 containing the largest number of data points and Cluster 1 the smallest. This imbalance suggests that the data are not evenly distributed in the feature space but are instead concentrated in specific groups. Such a pattern may be attributed to inherent data characteristics that lead to the formation of a dominant cluster. From a modeling perspective, this condition may affect the performance of the classification algorithm used in the subsequent stage. Models tend to learn patterns more effectively from clusters with larger sample sizes, whereas performance on minority clusters may decline. Several cities classified within the positively contributing cluster (Cluster 0) for new student enrollment are illustrated in Figure 5. Padang is observed to be dominant compared to other cities. In contrast, Padang Pariaman Regency, Pasaman Regency, Pariaman, and Kerinci Regency exhibit relatively lower contributions and tend to cluster more closely together.

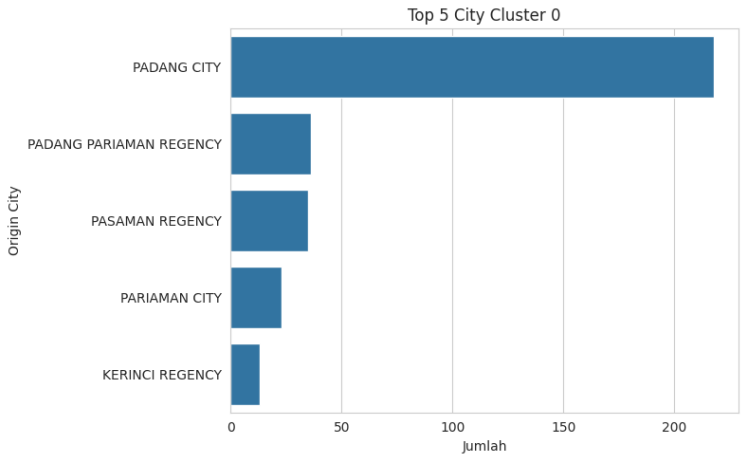


Figure 5. Top cities with school potential (cluster 0)

The dominance of Padang City indicates an unequal distribution of student origins within this cluster, with one region playing a dominant role. This may be due to geographic proximity to educational institutions, better accessibility to educational resources, or greater availability of information compared to other regions. This finding is important as a basis for formulating strategies for equitable promotion and for improving access to education, particularly in regions with low contributions.

SVM Classification Results

After the clustering process, the resulting cluster labels were used as target variables for the supervised classification stage using a Support Vector Machine (SVM). To ensure unbiased evaluation, the dataset was partitioned prior to clustering, and K-Means was applied exclusively to the training data. The trained clustering model was then used to assign cluster labels to the test data, preventing data leakage and maintaining consistency in the evaluation process. Classification performance was assessed using standard metrics, including accuracy, precision, recall, and F1-score. The results showed that the SVM model achieved excellent classification performance across all clusters, as reflected in the confusion matrix in Table 2, where the values range from 0 to 1. This indicates that the cluster boundaries generated during the unsupervised learning stage are highly separable in the feature space, allowing the SVM model to effectively learn and replicate the underlying structure.

Table 1. Performance metrics of k-means & SVM (%)

Cluster	Precision	Recall	F1-score
0	0.97	1.00	0.98
1	0,97	1.00	0,98
2	1.00	0,95	0,97
Accuracy			0,98
Macro Avg	0.98	0.98	0.98
Weight Avg	0.98	0.98	0.98

Furthermore, to assess model stability, nested K-Fold cross-validation was performed. The results showed consistently high accuracy across all folds, confirming that the model's performance is stable and independent of any particular data split. This consistency indicates that the hybrid approach is robust and generalizes well across different subsets of the dataset. The SVM model achieved an overall accuracy of 98%, indicating excellent classification performance. Clusters 0 and 1 exhibited perfect recall (1.00), meaning the model correctly identified all instances belonging to these clusters. Cluster 2, despite achieving perfect precision (1.00), showed slightly lower recall (0.95), indicating a small number of misclassifications into other clusters. This suggests that the SVM model performs particularly well on structurally dominant clusters (e.g., those representing large cities). In contrast, the internal variability of Cluster 2, which is more heterogeneous and dispersed, contributes to the slight decrease in recall. Nevertheless, the consistently high F1-score across all clusters confirms that the model maintains strong and stable performance. Table 2 also indicates that the high classification performance is influenced by the well-structured nature of the feature space, particularly after incorporating frequency-based features that enhance the discriminative power of the data. While this results in a clear decision boundary, future research may consider integrating additional and more diverse features to further evaluate the model under more complex and heterogeneous conditions.

Discussion and Implications

Model performance evaluation results shown in Table 3 indicate that an ensemble-based approach combining K-Means and classification algorithms provides significant performance improvements compared to using K-Means alone. Overall, the standalone K-Means method produced accuracy and recall values of 0.49, with a relatively low F1-score (0.35), indicating limitations in optimal class separation. Integrating K-Means with classification algorithms such as Random Forest and Logistic Regression showed gradual improvements. The K-Means & Random Forest method achieved an accuracy of 0.62 and an F1-score of 0.59, while the K-Means & Logistic Regression approach achieved an accuracy of 0.652 and an F1-score of 0.66. These results indicate that the addition of supervised learning models can improve initial clustering outcomes by capturing more complex patterns in the data.

Table 2. Performance Metrics of Hybrid Method (%)

	K-Means	K-Means & RF	K-Means & RL	K-Means & SVM
Accuracy	0.49	0.62	0.652	0.98
Recall	0.49	0.62	0.65	0.98
F1 Score	0.35	0.59	0.66	0.98
Precision	0.59	0.69	0.71	0.98

However, the K-Means & Support Vector Machine (SVM) approach demonstrated high performance, with accuracy, precision, recall, and F1-score values of 0.98. This performance may be influenced by an imbalanced data distribution, where the majority class dominates the sample size, contributing to higher evaluation metrics, particularly weighted scores. Under these conditions, the model tends to classify the majority class more easily, which significantly increases overall accuracy and recall. While these results quantitatively demonstrate excellent performance, the near-perfect scores require further analysis. They may indicate the presence of class imbalance or potential overfitting, particularly if supported by an uneven distribution of samples across classes.

4. CONCLUSION

This study proposes a hybrid approach integrating K-Means clustering and Support Vector Machine (SVM) classification to identify and predict potential locations for educational promotion. The clustering stage successfully uncovered underlying patterns in the dataset by grouping regions and schools based on their contribution levels, while the classification stage enabled predictive modeling of these patterns.

Experimental results showed that the proposed model achieved high performance, with an average accuracy of approximately 0.98, supported by similarly high values for precision, recall, and F1-score. Cross-validation confirmed that the model is stable and insensitive to data partitioning, demonstrating strong generalization capability. Furthermore, the inclusion of frequency-based features significantly improved the discriminative power of the dataset, enabling clear separation between clusters and allowing the SVM model to effectively learn the underlying structure. Comparative analysis with the baseline model also demonstrated that the hybrid approach delivers more consistent and reliable performance.

These findings indicate that the hybrid approach not only improves predictive accuracy but also provides a better understanding of the data structure, which is crucial for strategic decision-making, such as determining locations for educational promotion. However, the relatively small performance gap also suggests that the dataset exhibits a high degree of separability, allowing even the baseline models to achieve strong performance.

Nevertheless, this study has several limitations. The feature space remains relatively structured, and certain influential variables, such as socioeconomic indicators and temporal trends, are not included. Future research could explore the integration of more diverse features and more advanced models, including ensemble learning methods, to further improve model robustness and applicability to more complex real-world scenarios.

ACKNOWLEDGEMENTS

We extend our sincere gratitude to our colleagues and academic mentors for their insightful contributions. Our appreciation also goes to the Padang Institute of Technology, particularly the Cooperation Unit, for granting us the opportunity to carry out this research and for providing the essential data and resources. We hope that the knowledge shared through this work will bring meaningful benefits.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Anisya: Led and organized this research, including analyzing available data and conducting pre-processing, as well as reviewing the analysis results and adjusting the grammar. **Brestina Gultom**: Completed secondary data for the analysis process. **Sarjon Defit**: Provided conceptual contributions in developing the theory or concept that underpins the research.

DECLARATION OF COMPETING INTERESTS

The author declares that there is no conflict of interest regarding the publication of this article.

DATA AVAILABILITY

https://github.com/grohoanisya-data/PredictionStudent.git

REFERENCES

[1] I. Farida and D. Setiawan, "Business Strategies and Competitive Advantage: The Role of Performance and Innovation," *J. Open Innov. Technol. Mark. Complex.*, vol. 8, no. 3, 2022, doi: 10.3390/joitmc8030163.

[2] F. Li, J. Larimo, and L. C. Leonidou, "Social media marketing strategy: definition, conceptualization, taxonomy, validation, and future agenda," *J. Acad. Mark. Sci.*, vol. 49, no. 1, pp. 51–70, 2021, doi: 10.1007/s11747-020-00733-3.

[3] V. Bondarenko and B. Vyshnivska, "PROMOTIONAL MARKETING AS A METHOD OF INCREASING SALES," *Three Seas Econ. J.*, vol. 4, pp. 21–28, Jun. 2023, doi: 10.30525/2661-5150/2023-2-3.

[4] Y. Tadayonrad and A. B. Ndiaye, "A new key performance indicator model for demand forecasting in inventory management considering supply chain reliability and seasonality," *Supply Chain Anal.*, vol. 3, p. 100026, 2023, doi: https://doi.org/10.1016/j.sca.2023.100026.

[5] I. Antunes, L. Martinez, and L. Martinez, "The effectiveness of sales promotion techniques on the millennial consumers' buying behavior," *ReMark - Rev. Bras. Mark.*, vol. 21, pp. 784–836, May 2022, doi: 10.5585/remark.v21i3.19997.

[6] J. M. Hoem, "Demographic Analysis: Probabilistic Approach," *Int. Encycl. Soc. Behav. Sci.*, pp. 3428–3432, 2001, doi: 10.1016/B0-08-043076-7/02092-1.

[7] F. M. Surur *et al.*, "Unlocking the power of machine learning in big data: a scoping survey," *Data Sci. Manag.*, vol. 8, no. 4, pp. 519–535, 2025, doi: https://doi.org/10.1016/j.dsm.2025.02.004.
[8] G. J. Oyewole and G. A. Thopil, "Data clustering: application and trends," *Artif. Intell. Rev.*, vol. 56, no. 7, pp. 6439–6475, 2023, doi: 10.1007/s10462-022-10325-y.
[9] K. Tabianan, S. Velu, and V. Ravi, "K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data," 2022. doi: 10.3390/su14127243.
[10] M. Trombini, D. Solarna, G. Moser, and S. Dellepiane, "A goal-driven unsupervised image segmentation method combining graph-based processing and Markov random fields," *Pattern Recognit.*, vol. 134, p. 109082, 2023, doi: https://doi.org/10.1016/j.patcog.2022.109082.
[11] M. M. Kumbure, C. Lohrmann, P. Luukka, and J. Porras, "Machine learning techniques and data for stock market forecasting: A literature review," *Expert Syst. Appl.*, vol. 197, p. 116659, 2022, doi: https://doi.org/10.1016/j.eswa.2022.116659.
[12] B. Zhang, W. J. Yin, M. Xie, and J. Dong, "Geo-spatial Clustering with Non-spatial Attributes and Geographic Non-overlapping Constraint: A Penalized Spatial Distance Measure BT - Advances in Knowledge Discovery and Data Mining," Z.-H. Zhou, H. Li, and Q. Yang, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 1072–1079.
[13] I. Fahmiyah and R. A. Ningrum, "Human Development Clustering in Indonesia: Using K-Means Method and Based on Human Development Index Categories," *J. Adv. Technol. Multidiscip.*, vol. 2, no. 1 SE-Articles, pp. 27–33, May 2023, doi: 10.20473/jatm.v2i1.45070.
[14] Y. Chen, P. Tan, M. Li, H. Yin, and R. Tang, "K-means clustering method based on nearest-neighbor density matrix for customer electricity behavior analysis," *Int. J. Electr. Power Energy Syst.*, vol. 161, p. 110165, 2024, doi: https://doi.org/10.1016/j.ijepes.2024.110165.
[15] H. Yan, M. Ma, Y. Wu, H. Fan, and C. Dong, "Overview and analysis of the text mining applications in the construction industry," 2022, *The Author(s)*. doi: 10.1016/j.heliyon.2022.e12088.
[16] K. Maheswari and P. P. A. Priya, "Predicting customer behavior in online shopping using SVM classifier," in *2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, 2017, pp. 1–5. doi: 10.1109/ITCOSP.2017.8303085.
[17] R. Rodríguez-Pérez and J. Bajorath, "Evolution of Support Vector Machine and Regression Modeling in Chemoinformatics and Drug Discovery," *J. Comput. Aided. Mol. Des.*, vol. 36, no. 5, pp. 355–362, 2022, doi: 10.1007/s10822-022-00442-9.
[18] F. Topaloglu, "A hybrid approach based on k-means and SVM algorithms in selection of appropriate risk assessment methods for sectors.," *PeerJ. Comput. Sci.*, vol. 10, p. e2198, 2024, doi: 10.7717/peerj-cs.2198.
[19] M. Nasser, H. Hamza, N. Salim, and F. Saeed, "Clustering Web Users Based on K-means Algorithm for Reducing Time Access Cost," in *2019 First International Conference of Intelligent Computing and Engineering (ICOICE)*, 2019, pp. 1–7. doi: 10.1109/ICOICE48418.2019.9035190.
[20] R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, "An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review," *Information*, vol. 15, no. 4, 2024, doi: 10.3390/info15040235.
[21] H. Wang, G. Li, and Z. Wang, "Fast SVM classifier for large-scale classification problems," *Inf. Sci. (Ny).*, vol. 642, p. 119136, 2023, doi: https://doi.org/10.1016/j.ins.2023.119136.
[22] A. Shdefat, N. Mostafa, Z. Al-Arnaout, Y. Kotb, and S. Alabed, "Optimizing HAR Systems: Comparative Analysis of Enhanced SVM and k-NN Classifiers," *Int. J. Comput. Intell. Syst.*, vol. 17, Jun. 2024, doi: 10.1007/s44196-024-00554-0.
[23] A. Mumuni and F. Mumuni, "Automated data processing and feature engineering for deep learning and big data applications: A survey," *J. Inf. Intell.*, vol. 3, no. 2, pp. 113–153, 2025, doi: https://doi.org/10.1016/j.jiixd.2024.01.002.
[24] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," 2022, *Elsevier*.
[25] P. Chen, L. Wu, and L. Wang, "AI Fairness in Data Management and Analytics: A Review on Challenges, Methodologies and Applications," 2023. doi: 10.3390/app131810258.
[26] L. Huang, J. Qin, Y. Zhou, F. Zhu, L. Liu, and L. Shao, "Normalization Techniques in Training DNNs: Methodology, Analysis and Application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 10173–10196, 2023, doi: 10.1109/TPAMI.2023.3250241.
[27] A. Ahmad and S. S. Khan, "initKmix-A novel initial partition generation algorithm for clustering mixed data using k-means-based clustering," *Expert Syst. Appl.*, vol. 167, p. 114149, 2021, doi: https://doi.org/10.1016/j.eswa.2020.114149.
[28] M. Gul and M. A. Rehman, "Big data: an optimized approach for cluster initialization," *J. Big Data*, vol. 10, no. 1, p. 120, 2023, doi: 10.1186/s40537-023-00798-1.
[29] I. K. Khan, H. Daud, N. Zainuddin, and R. Sokkalingam, "Standardizing reference data in gap statistic for selection optimal number of cluster in K-means algorithm," *Alexandria Eng. J.*, vol. 118, pp. 246–260, 2025, doi: https://doi.org/10.1016/j.aej.2025.01.034.
[30] G. J. Simon, "Artificial Intelligence and Machine Learning in Health Care and Medical Sciences: Best Practices and Pitfalls," in *1st ed. in Health Informatics Series. Cham: Springer International Publishing AG*, 2024.
[31] M. Conciatori, A. Valletta, and A. Segalini, "Improving the quality evaluation process of machine learning algorithms applied to landslide time series analysis," *Comput. Geosci.*, vol. 184, p. 105531, 2024, doi: https://doi.org/10.1016/j.cageo.2024.105531.
[32] J. Terven, D.-M. Cordova-Esparza, J.-A. Romero-González, A. Ramírez-Pedraza, and E. Chávez Urbiola, "A comprehensive survey of loss functions and metrics in deep learning," *Artif. Intell. Rev.*, vol. 58, Apr. 2025, doi: 10.1007/s10462-025-11198-7.
[33] B. Sumanto and S. Nurrahma, "Comparison of Random Forest Support Vector Machine and Passive Aggressive Models on E-nose-Based Aromatic Rice Classification," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 24, pp. 381–394, Jul. 2025, doi: 10.30812/matrik.v24i3.4291.
[34] M. C. Hinojosa Lee, J. Braet, and J. Springael, "Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores," *Appl. Sci.*, vol. 14, no. 21, p. 9863, Oct. 2024, doi: 10.3390/app14219863.

.