A Novel Hybrid Multi-Criteria Decision and Data Mining Framework for Educational Intelligence Systems

Orissa Octaria
Doctoral Program in Computer Science, Satya Wacana Christian University, Indonesia | Faculty of Computer Science and Engineering, Universitas Multi Data Palembang, Indonesia
orissa.octaria@mdp.ac.id (corresponding author)

Danny Manongga
Faculty of Information Technology, Satya Wacana Christian University, Indonesia
danny.manongga@uksw.edu

Irwan Sembiring
Faculty of Information Technology, Satya Wacana Christian University, Indonesia
irwan@uksw.edu (corresponding author)

Received: 7 March 2026 | Revised: 31 March 2026, 21 April 2026, and 4 May 2026 | Accepted: 8 May 2026

 DOI: https://doi.org/10.48084/etasr.18583

ABSTRACT

Although student exchange programs provide substantial advantages, a large number of university students are still unaware that such opportunities exist. This study addresses this gap by proposing a hybrid Educational Intelligence System (EIS), which is a data-driven decision-support framework that integrates Multi-Criteria Decision Analysis (MCDA) with data mining techniques. Specifically, the Analytical Hierarchy Process (AHP) is adopted as the MCDA method. Using AHP, the relative significance of each criterion influencing student awareness is determined. Concurrently, data mining methods, namely, clustering and classification are employed to reveal underlying patterns within student data. Clustering serves to categorize students according to their comprehension level of exchange programs, whereas Decision Tree-based classification pinpoints the dominant factors that shape student awareness. A total of 446 students from diverse higher education institutions participated in the study by completing a structured questionnaire. The clustering analysis reveals that 47.31% of respondents have a general familiarity with exchange programs yet lack detailed knowledge of specific requirements, whereas 30.94% exhibit an overall limited awareness. Based on these findings, promotional strategies for exchange programs should be differentiated according to students' academic progression. Moreover, the data-driven framework introduced in this study holds potential for broader application across various educational settings to strengthen the impact of academic initiatives.

Keywords-Multi-Criteria Decision Analysis (MCDA); Analytical Hierarchy Process (AHP); classification; clustering

I. INTRODUCTION

Student exchange programs offer valuable opportunities for cultural exchange, academic enrichment, and personal development [1]. However, despite the substantial benefits they provide, many university students remain unaware of the opportunities available through these programs [2]. Prior research [3] has confirmed low awareness levels at private universities, stemming from insufficient information dissemination, limited resource access, and low engagement [4]. Educational Intelligence Systems (EIS) refer to data-driven decision-support architectures that integrate computational and analytical methods to enhance educational practice and institutional decision-making [5]. EIS frameworks leverage educational data mining, Multi-Criteria Decision Analysis (MCDA), and machine learning to provide actionable insights for administrators and educators. The framework proposed in this study constitutes an EIS by combining expert-driven criteria weighting (Analytical Hierarchy Process (AHP)), unsupervised student profiling (K-means), and supervised predictor identification (Decision Tree) into a unified pipeline.

To address awareness deficits, this study integrates MCDA with data mining. Specifically, AHP determines the relative significance of each awareness criterion, whereas clustering and classification reveal underlying patterns and dominant

predictors within student data. A prior study [3] relied solely on MCDA with expert-determined weights, which introduced subjective bias. Authors in [6] integrated Multi-Criteria Decision-Making (MCDM) with classification only, in a managerial (supplier selection) context. Our work extends this by introducing a stage pipeline AHP weighting → K-means clustering → Decision Tree classification applied to educational awareness profiling for the first time in the Southeast Asian context. It is important to note that novelty in this study is not claimed at the individual algorithm level; AHP, K-means, and Decision Tree are each applied in their standard, established forms. The contribution lies at the system integration level, as detailed in the Contributions subsection below.

A. Contributions

Figure 1 presents a bibliometric keyword co-occurrence map generated using VOSviewer based on a systematic analysis of publications indexed in Scopus, combining the keywords AHP, MCDA, data mining, classification, and clustering in educational and decision-support contexts. The map reveals two structurally distinct and largely disconnected clusters: one dominated by AHP and MCDA terminology, and the other by data mining, classification, and clustering techniques. The minimal co-occurrence linkage between these two clusters constitutes quantitative, empirical evidence that their methodological integration remains underexplored in the literature. This type of bibliometric co-occurrence analysis is an established and increasingly adopted method for systematic gap identification in engineering and applied science research [7].

The structural gap visualized in Figure 1 directly motivates the pipeline proposed in this study, in which AHP and data mining techniques are not merely applied in parallel but are formally integrated through the AHP-to-label bridge mechanism. Based on the state-of-the-art mapping shown in Figure 1, most existing studies still position data mining as a post-decision analytical tool rather than as an integral component of the decision-structuring process.

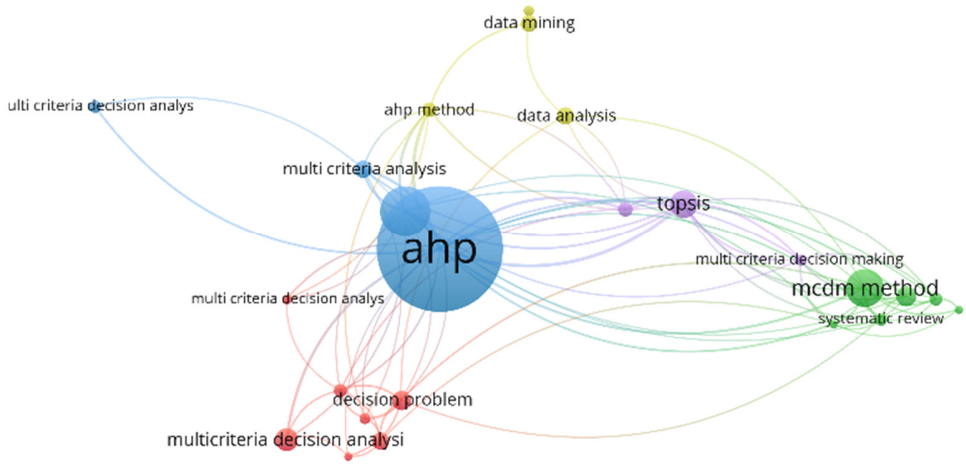

Fig. 1. Bibliometric keyword co-occurrence map of MCDA, AHP, and data mining literature generated using VOSviewer.

The proposed AHP-to-label bridge mechanism uses AHP-derived priority weights to compute per-respondent composite scores, from which classification labels are derived. This process converts expert-driven qualitative judgment into supervised learning targets, enabling an MCDA–data mining integration that is structurally distinct from prior works.

In a prior study [6], the integration of MCDA and data mining was limited to classification, which still allowed room for bias due to the inclusion of subjective opinions. To mitigate such bias, this research introduces the combined use of both clustering and classification techniques, thereby reducing the influence of subjective assessments. The idea for this integration draws upon studies [8, 9], which demonstrate that the information and knowledge obtained through data mining can significantly improve decision-making processes.

This study makes three explicit contributions to the hybrid MCDA–data mining literature: First, an AHP-to-label bridge mechanism is introduced. AHP priority weights are used to compute composite scores, which are then converted into binary labels. This establishes a direct and reproducible link between expert judgment and supervised learning, unlike [6], which uses MCDA only for ranking without generating training labels. Second, an unsupervised-before-supervised pipeline is proposed. K-means clustering is applied prior to Decision Tree classification, enabling unsupervised segmentation of the student population prior to supervised label prediction. This ordering allows clustering structure to inform classification, differing from prior studies where both methods are applied independently or in reverse sequence. Third, the framework is applied in a real-world educational domain, producing actionable outputs. The pipeline is applied to student exchange program awareness profiling across Southeast Asian higher education institutions, producing cluster-based student profiles and a Decision Tree that identifies length of study as the dominant awareness predictor, a directly actionable finding for institutional outreach strategy design.

1) Multi-Criteria Decision Analysis

MCDA is a decision-support framework that outperforms monocriteria systems by incorporating two or more potentially conflicting actions, interactions, or objectives [10]. One study

explored the application of blockchain within a multi-criteria decision-making context to enhance user interaction efficiency by supporting reliable information exchange [11]. Additionally, literature reviews have examined MCDA methodologies, tools, and indicators [12], enabling comparisons among alternative options. MCDA is also employed to assess future environmental sustainability by evaluating criteria that encompass technical, economic, environmental, and social aspects [13]. Numerous studies have applied MCDA, particularly AHP, in educational contexts [14, 15]. The reviewed literature highlights the method's ability to accommodate various aspects and produce accountable results, even when actions are in conflict. AHP is noted as a superior method compared to Simple Additive Weighting (SAW), as it involves a two-step process for determining importance weights [16]. Furthermore, AHP provides better results than Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), especially in terms of identifying the smallest Euclidean distance to the ideal solution [17].

2) Data Mining

In data mining, particularly through classification, clustering, and association functions, institutions can gain valuable insights for making informed decisions [18, 19]. A study on occupational accidents, for instance, demonstrated that data mining was instrumental in processing and extracting valuable information from historical data [20]. In the healthcare field, although data mining was initially applied in isolation, recent studies have integrated it with other methods to enhance virus detection techniques [21]. The widespread application of data mining across various domains supports its selection in this study. The classification method used is Decision Tree, whereas the clustering method applied is K-means. K-means is widely popular across research fields due to its ability to yield a lower Davies–Bouldin Index compared to other clustering methods, such as K-Medoids [22-24]. Moreover, the Decision Tree method outperforms the Random Forest algorithm in terms of prediction accuracy and error rate, making it the preferred classification technique for this research [25].

II. RESEARCH METHODOLOGY

This research methodology aims to explore how MCDA, when combined with data mining techniques, can optimize strategies [26]. In this case, to enhance student awareness of student exchange programs. The research methodology comprises several stages, as described below and illustrated in the accompanying diagram.

A. Problem Definition

The core issue in this study is to examine students' awareness of student exchange programs and subsequently increase that awareness. Previous studies [2, 3] have shown that although many students are aware of the existence of such programs, they often lack knowledge about the necessary criteria and requirements to participate.

B. Data Collection

The questionnaire design consists of close-ended questions covering aspects such as gender, age, and region of origin, which are suitable for clustering. Data collection was conducted by distributing the questionnaire to active university students from several institutions. Ethical considerations included not collecting personal data and securing informed consent from participants before they proceeded with the questionnaire. The sampling method employed was cluster sampling, which divides the population geographically into clusters, especially useful for large populations [27], followed by simple random sampling. The required sample size was determined using the Krejcie and Morgan [28] formula, as cited in [29], which is appropriate for large and homogeneous populations. In this case, the population is limited to active university students.

C. Integration of Multi-Criteria Decision Analysis and Data Mining

The integration of the MCDA model and data mining is illustrated in Figure 2. In this study, the MCDA approach employed is AHP. One key stage in AHP is weighting, which is typically performed using expert judgment, often based on qualitative data. These expert-derived weights are then integrated with the data mining process, specifically, clustering. Additionally, classification techniques are employed to support the findings.

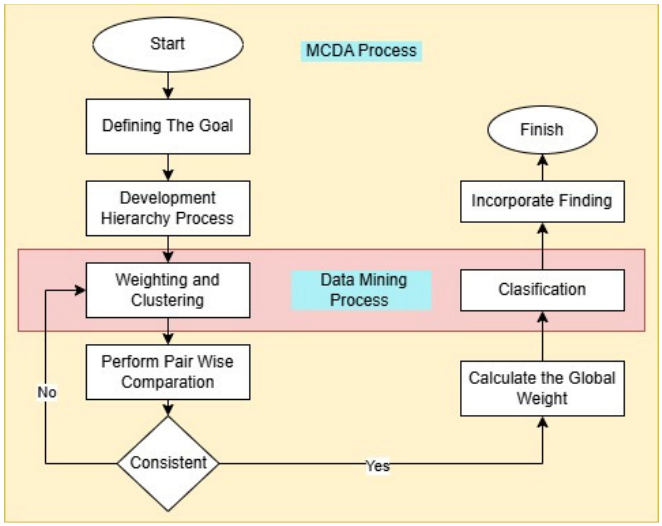

Fig. 2. Integration of MCDA and data mining.

III. EXPERIMENTATION SETUP

This section outlines the preparatory stages of the research, beginning with the development and distribution of questionnaires, and extending to the creation of a program for processing the collected data.

A. Questionnaire Distribution

The required sample size was determined using the Krejcie and Morgan [28] formula. The study targeted active undergraduate university students aged 18–25 years enrolled in higher education institutions across four Southeast Asian countries: Indonesia, Malaysia, Thailand, and the Philippines. These countries collectively represent large tertiary education sectors, with gross enrollment ratios ranging from 39.59% to 43.96% as reported by the UNESCO Institute for Statistics

[30], reflecting large and rapidly expanding higher education populations that substantially exceed the minimum sample size threshold required by the Krejcie and Morgan formula. This formula yields a minimum sample size of 384 at a 95% confidence level with a 5% margin of error [29]. A total of 446 valid responses were collected, exceeding this minimum threshold. The use of cluster sampling was appropriate due to the geographical distribution of respondents across several Southeast Asian countries, including Indonesia, Malaysia, the Philippines, and Thailand [27]. The AHP component involved structured expert elicitation by one domain expert.

B. Programming

The high-level programming language used in this study is Python. While high-level languages may have performance limitations compared to low-level languages, they offer greater flexibility, ease of use, extensive community support, and abundant resources [31]. Python was selected due to its user-friendly syntax and intuitive structure [32]. Algorithm 1 outlines the algorithm used in the study.

```
Algorithm 1: Integrated AHP, K-Means, and
Decision Tree Pipeline
Step 1 [AHP Expert Elicitation]:
1.1 Construct 9×9 pairwise comparison
    matrix A using the Saaty scale
1.2 Normalize A column-wise; compute
    priority weight vector W = {w₁,...,w₉}
1.3 Verify CR = CI/RI < 0.10 (CR = 0.027 ✓)
Step 2 [Data Collection]:
2.1 Collect survey responses; encode
    categorical features
2.2 Apply validity (Pearson r > 0.30) and
    reliability (α > 0.60) testing
Step 3 [K-means Clustering Unsupervised]:
3.1 Standardize Q1-Q9 scores
    (StandardScaler)
3.2 Select k=3 via elbow method +
    Silhouette analysis
3.3 Assign cluster label Cᵢ ∈ {0,1,2} to
    each respondent
Step 4 [AHP-Weighted Labeling Bridge
Mechanism]:
4.1 Compute Sᵢ = Σⱼ(wⱼ × Qᵢⱼ) / (5 × Σwⱼ)
    for each respondent i
4.2 Assign Labelᵢ = 1 ('Yes') if Sᵢ ≥
    0.80, else 0 ('No')
Step 5 [Decision Tree Classification
Supervised]:
5.1 Feature set X = {Gender, Age,
    Length_of_Study}
5.2 Target y = {Labelᵢ} from Step 4
5.3 Split: 80% train (n=356), 20% test
    (n=90), stratified
5.4 Train DecisionTreeClassifier
    (criterion=gini, max_depth=5)
5.5 Evaluate: accuracy, precision, recall,
    F1, confusion matrix
5.6 5-fold cross-validation for
    generalizability assessment
Step 6 [Output]: Cluster profiles +
Feature importance ranking
```

IV. ANALYSIS

A. Validity and Reliability Testing

Validity and reliability testing aims to determine whether the questionnaire items are valid and reliable. A questionnaire is considered valid if the question variables are measurable [33]. The validity and reliability calculations were conducted using the SPSS application. However, manual calculations can also be performed using the Cronbach's alpha formula. Based on the correlation coefficient (r) table, a value greater than 0.1 is considered acceptable for samples with more than 100 respondents. Pearson correlation testing confirmed that all variables demonstrated coefficient values exceeding 0.1, satisfying the validity threshold. A high reliability value is considered acceptable if it exceeds 0.6; the higher the reliability, the more trustworthy the variable is [34]. The Cronbach's alpha formula used to assess reliability is shown in (1) [34]:

$$r_{11} = \left(\frac{n}{n-1}\right)\left(1 - \frac{\sum \sigma_i^2}{\sigma_t^2}\right) \quad (1)$$

The reliability coefficient obtained from the distributed questionnaire is 0.860. Since this value exceeds the 0.6 threshold, it indicates that the questionnaire is both valid and reliable. Figure 3 presents the screenshot of the reliability calculations using the SPSS software.

Reliability Statistics

Cronbach's Alpha	N of Items
.860	9

Fig. 3. Cronbach's alpha value.

B. Result of Integrating Multi-Criteria Decision Analysis and Data Mining Models

1) Data Mining (Clustering)

The initial step in the data mining process was clustering to determine the optimal number of clusters. The elbow method was employed to identify the most suitable number of clusters, resulting in the formation of three clusters [5]. As illustrated in Figure 4, the elbow method graph displays the steepest decline in Within-Cluster Sum of Squares (WCSS) between k = 1 and k = 2, after which the rate of decrease diminishes considerably. While k = 2 represents the first point of diminishing returns, it does not uniquely determine the optimal cluster count. To validate the selection of k, additional cluster validity indices were computed for k = 2 through k = 5. The results showed that k = 2 yields a Silhouette Score of 0.2700 and a Davies-Bouldin Index (DBI) of 1.408, whereas k = 3 yields a Silhouette Score of 0.2245 and a DBI of 1.483. Although k = 2 produces marginally better statistical indices, k = 3 was selected

following the established practice of authors in [35], as it produces three semantically distinct and interpretable student awareness profiles low awareness (Cluster 0), partial awareness (Cluster 2), and high awareness (Cluster 1) that are directly actionable for institutional outreach strategy design. A solution that collapses these three profiles into two would obscure the partial-awareness group, which constitutes the largest segment (47.31%) and the most strategically important target for intervention.

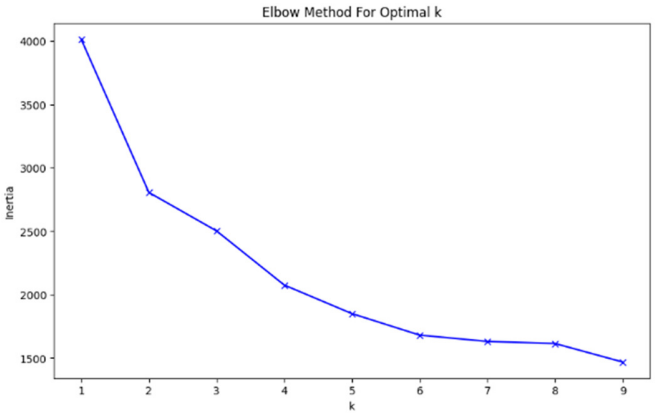

Fig. 4. Elbow method graph.

After determining the number of clusters, the next step involved identifying the percentage composition of each cluster. The three resulting clusters are as follows: Cluster 0 includes respondents who predominantly provided low scores, comprising 30.94% of the total. Cluster 1 consists of respondents with consistently high scores, accounting for 21.75%. Cluster 2 includes respondents who gave a combination of high scores for the first four questions and low scores for questions five to seven, representing 47.31%. A visualization of this distribution is provided in Figure 5.

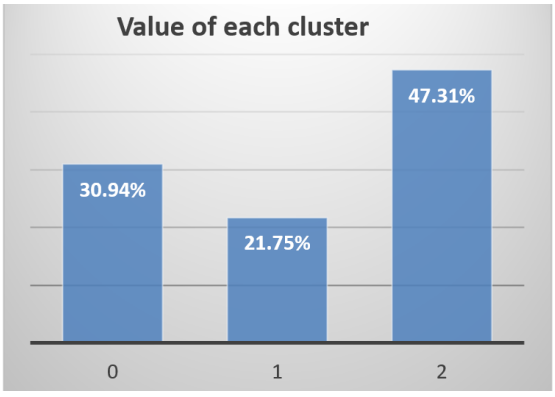

Fig. 5. Cluster percentage distribution.

High scores indicate strong knowledge of student exchange programs among the respondents, whereas low scores suggest limited understanding. The clustering results indicate that 47.31% of students (Cluster 2) possess general knowledge about student exchange programs but lack specific knowledge, which may be more crucial. Figure 6 and 7 presents the distribution graph of the three clusters and a heatmap of correlations between the various criteria and students' knowledge of exchange programs.

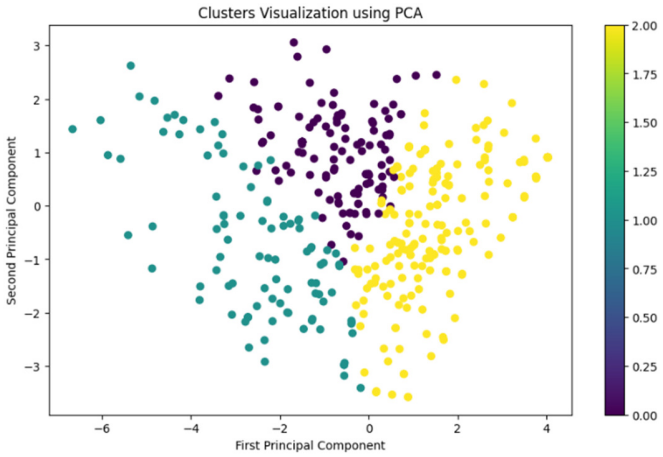

Fig. 6. Cluster distribution graph.

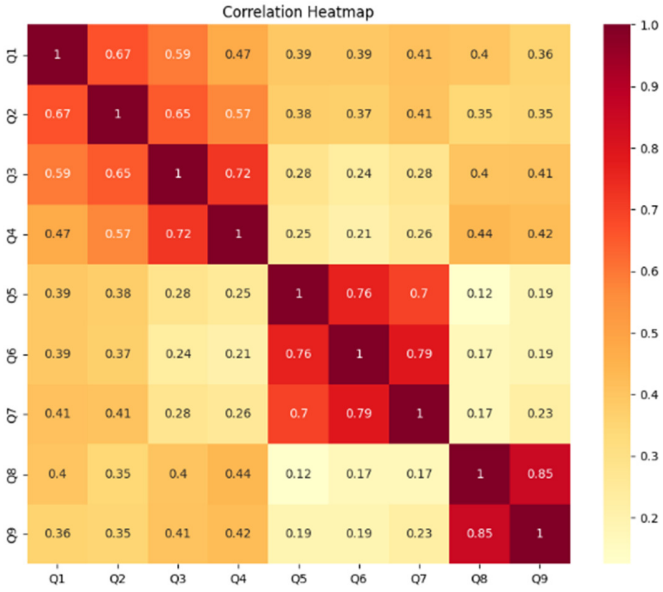

Fig. 7. Correlation heatmap between criteria.

2) Multi Criteria Decision Analysis

The MCDA method used in this study is the AHP. This method is appropriate for multi-criteria selection problems and can evaluate both qualitative and quantitative data [36]. AHP calculations assess the consistency ratio, which must not exceed 0.1; exceeding this threshold indicates that the judgments need to be revised [36]. Equation (2) shows the formula for the consistency ratio, CR , where CI denotes the consistency index and RI denotes the random index.

$$CR = \frac{CI}{RI}\ ,\ CI = \frac{\lambda_{max}-1}{n-1} \quad (2)$$

This study yielded a consistency ratio of 0.027, which is acceptable since it is below the 0.1 threshold, indicating consistent judgments. With an acceptable consistency ratio, the calculation of the priority weight for each criterion can proceed. This study used nine criteria, with their respective weights illustrated in Figure 8.

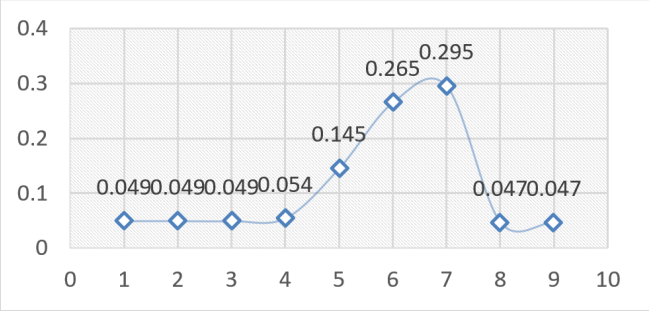

Fig. 8. Priority weights of criteria.

Once the priority weights were determined, each respondent's responses were evaluated against these weights to label whether they were knowledgeable about student exchange programs. This labeling was done manually: respondents with an average score below 0.8 were labeled "No," whereas those with scores equal to or above 0.8 were labeled "Yes."

3) Data Mining (Classification)

The classification phase aimed to identify which demographic variables most significantly determine the likelihood that a student is aware of exchange programs. The Decision Tree algorithm was selected for its interpretability, its robustness in handling non-linear relationships across categorical variables, and its ability to generate human-readable rules directly applicable to institutional outreach strategies [37]. Prior to model training, a class imbalance analysis was conducted on the AHP-labeled dataset. Of 446 respondents, 359 (80.5%) received a label of "No" and 87 (19.5%) received a label of "Yes," yielding an imbalance ratio of 4.13:1. Training a classifier directly on this imbalanced distribution results in a degenerate model that predicts the majority class for all instances, producing zero true positives (TP = 0) for the minority class, an outcome that is numerically acceptable in terms of overall accuracy but scientifically meaningless for the minority group of interest.

To address this, the Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training partition. The dataset was split into 80% training (n = 356) and 20% test (n = 90) using stratified random sampling to preserve the original class proportion in both partitions. SMOTE was then applied to the training set only, generating synthetic minority-class samples to produce a balanced training distribution of 287 "No" and 287 "Yes" instances (total n = 574). The test set was held out and not resampled, ensuring that evaluation metrics reflect performance on the true data distribution.

A Decision Tree classifier (Scikit-learn 1.x, criterion = Gini impurity, max_depth = 5) was trained on the SMOTE-balanced training set and evaluated on the original test set. Five-fold stratified cross-validation was applied to assess generalization. Table I presents the complete classification performance metrics for the revised proposed model (DT + SMOTE) alongside the original unbalanced model and a Naïve Bayes + SMOTE baseline.

TABLE I. COMPARATIVE CLASSIFIER PERFORMANCE

Classifier	Decision Tree (no balancing)	DT + SMOTE (proposed)	Naïve Bayes + SMOTE (baseline)
Accuracy (%)	78.9	68.9	55.6
Precision (No)	0.8	0.81	0.88
Recall (No)	0.99	0.79	0.51
Precision (Yes)	0	0.25	0.27
Recall (Yes)	0	0.28	0.72
F1-score (macro)	0.44	0.53	0.52
Cross-validation accuracy (%)	80.5	58.5	53.4

The revised DT + SMOTE model achieves an overall test accuracy of 68.9% and a macro-average F1-score of 0.53, a substantial improvement over the macro F1 score of 0.44 produced by the original unbalanced model, which completely failed to detect the minority class (TP = 0). The confusion matrix for the revised model (TN = 57, FP = 15, FN = 13, TP = 5) confirms that the model can now identify minority-class instances, albeit with limited precision (0.25) due to the inherent difficulty of predicting awareness from three demographic features alone. This limitation is acknowledged and discussed in the Discussion section.

The Decision Tree identifies length of study as the most influential predictor (feature importance = 0.4703), followed by age (0.2896) and gender (0.2401). At the root node (Level 0), the tree splits on age ≤ 0.5 years (Gini = 0.5, n = 574 training samples). This result is consistent with academic socialization theory: students who have spent more time within the institutional environment have been exposed to more peer networks, faculty interactions, and institutional communications, all of which are established channels for exchange program awareness transmission. The practical implication is that outreach strategies should be differentiated by academic stage, with earlier-year students receiving more proactive, institution-initiated communication.

C. Discussion

The three-cluster solution reveals a student population structured by distinct awareness profiles rather than a simple binary distinction. Cluster 2 (47.31%), consisting of students with high scores on general program knowledge (Q1–Q4) but low scores on procedural knowledge (Q5–Q7), is the most actionable group from a policy perspective. These students are campus engaged, suggesting they are reachable through existing institutional channels but lack specific knowledge about eligibility criteria, application procedures, and required documentation.

Cluster 0 (30.94%), characterized by uniformly low scores across all nine criteria, represents students with foundational awareness barriers. For this group, basic informational outreach on program existence, benefits, and eligibility is the appropriate

first intervention. Cluster 1 (21.75%), with uniformly high scores, represents the reference population and may be leveraged as peer ambassadors in structured awareness campaigns.

The dominance of length of study as the primary predictor (importance = 0.4703) in the Decision Tree is consistent with academic socialization theory. As students progress through their degree, they accumulate more peer network ties, faculty relationships, and familiarity with institutional communication channels, all mechanisms through which exchange program awareness is transmitted. Gender and age play secondary roles (importances of 0.24 and 0.29, respectively), suggesting that awareness is more a function of institutional exposure duration than of demographic characteristics per se.

These findings support the following differentiated outreach strategy: (a) students in their first year (length of study < 1 year) should receive proactive push communication at enrolment and orientation events, as they have not yet had sufficient time to develop the peer networks through which awareness is typically transmitted; (b) students in years 1–2 benefit most from structured peer-to-peer ambassador programs, where higher year students with high awareness (Cluster 1) are deployed as information intermediaries; (c) students in year 3 and beyond who remain unaware (Cluster 0 members with long study duration) may require direct faculty-mediated advising referral, as standard mass communication may no longer be reaching them effectively.

Several limitations must be acknowledged. First, the class imbalance (4.13:1) constrained classification performance for the minority class. Despite SMOTE, the minority-class recall of 0.28 and precision of 0.25 reflect the difficulty of predicting awareness from only three demographic features. This limitation is inherent to the feature set and not resolvable by oversampling alone; future research should incorporate additional features (e.g., faculty affiliation, scholarship status, social media exposure) to improve minority-class predictability. Second, 99.6% of respondents were from Indonesian institutions, which limits the generalizability of the findings to other Southeast Asian contexts. Future research should expand the sampling frame to include institutions from Malaysia, Thailand, and the Philippines in a more balanced proportion. Third, self-selection bias in survey response cannot be excluded: students who chose to participate may systematically differ from those who did not.

V. CONCLUSION AND FUTURE RESEARCH

This study presented a novel hybrid framework integrating Analytical Hierarchy Process (AHP), K-means clustering, and Decision Tree classification for assessing student awareness of exchange programs across Southeast Asian higher education institutions. A total of 446 undergraduate students from Indonesia, Malaysia, Thailand, and the Philippines participated in the survey, and their responses were analyzed through a three-stage pipeline that constitutes the core contribution of this work.

Three explicit and verifiable contributions were demonstrated. First, the AHP-to-label bridge mechanism: AHP priority weights were used to compute per-respondent composite scores, from which binary classification labels were derived, converting expert qualitative judgment into supervised learning targets, a mechanism absent from prior hybrid Multi-Criteria Decision Analysis (MCDA) and data mining works. Second, the unsupervised-before-supervised pipeline architecture: K-means clustering was applied prior to Decision Tree classification, enabling unsupervised student segmentation to contextualize and inform supervised prediction. Third, the domain application with actionable output: the pipeline was applied to student exchange program awareness profiling in Southeast Asian higher education, a context not previously addressed in hybrid MCDA and data mining literature.

The clustering analysis identified three distinct student awareness profiles: Cluster 1 (21.75%, high awareness), Cluster 2 (47.31%, partial awareness), and Cluster 0 (30.94%, low awareness). The Decision Tree classifier, trained on Synthetic Minority Over-sampling Technique (SMOTE)-balanced data to address class imbalance (ratio 4.13:1), identified length of study as the most influential predictor of awareness (feature importance = 0.4703), followed by age (0.2896) and gender (0.2401), achieving a test accuracy of 68.9% and a cross validated accuracy of 58.5%.

These findings carry direct practical implications, and institutions are recommended to implement differentiated outreach strategies. These include proactive push communication at enrolment for first year students, peer ambassador programs for students in years one to two, and faculty-mediated advising referrals for students in year three and beyond.

Future research should expand the feature set beyond demographic variables to include faculty affiliation, scholarship status, and social media exposure. It should also extend the geographic sampling frame to achieve a more balanced cross-country representation across Southeast Asia.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no competing interests, whether financial, personal, or professional, that could have influenced the design, conduct, or reporting of this study. No author has any affiliation with or financial involvement in any organization or entity with a direct financial interest in the subject matter or materials discussed in this manuscript.

ACKNOWLEDGMENT

The author would like to express sincere gratitude to Multi Data Palembang University for providing financial support for this research. The funding and institutional assistance have significantly contributed to the successful completion of this study. The funding supported the data collection, analysis, and publication process of this study.

DATA AVAILABILITY

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

AI USE AND DECLARATION OF GENERATIVE AI USE

The authors used Claude (developed by Anthropic) as a Generative AI tool solely to assist with improving the

language, grammar, and readability of the manuscript. The AI was used under full human oversight and control at all times. All scientific content, methodology, data analysis, interpretation of results, and conclusions are entirely the work of the authors. The use of AI did not affect any images, figures, or data presented in this manuscript. The authors take full responsibility for the integrity and accuracy of all content presented in this work.

REFERENCES

[1] A. Kristin Bjørnnes, A. Torbjørnsen, B. Sigridur Anna Thordardottir, A. Lund, O. Johannes Hovland, and L. Skeie Skarpaas, "Exploring intentions: factors influencing international study decisions in healthcare bachelor degree programs," *BMC Medical Education*, vol. 25, no. 1, Apr. 2025, Art. no. 555, https://doi.org/10.1186/s12909-025-07136-4.

[2] O. Octaria, D. Manongga, A. Iriani, H. D. Purnomo, and I. Setyawan, "Mining Opinion Based on Tweets about Student Exchange with Tweepy and TextBlob," in *2022 9th International Conference on Information Technology, Computer, and Electrical Engineering*, Semarang, Indonesia, 2022, pp. 102–106, https://doi.org/10.1109/ICITACEE55701.2022.9924013.

[3] O. Octaria, K. D. Hartomo, I. Sembiring, H. D. Purnomo, A. Iriani, and E. Sediyono, "Analysis Perceptions Regarding Student Exchange Using Simple Random Sampling and Analytical Hierarchy Process (AHP) Methods," in *2022 9th International Conference on Electrical Engineering, Computer Science and Informatics*, Jakarta, Indonesia, 2022, pp. 331–336, https://doi.org/10.23919/EECSI56542.2022.9946497.

[4] K. Nyamsuren, Z. Gankhuyag, J. Ganbaatar, and N. Marinescu, "The Importance of Studying Abroad for a Sustainable Education: Research on Mongolian Student Opinions," *Sustainability*, vol. 16, no. 14, July 2024, Art. no. 6137, https://doi.org/10.3390/su16146137.

[5] A. Dutt, M. A. Ismail, and T. Herawan, "A Systematic Review on Educational Data Mining," *IEEE Access*, vol. 5, pp. 15991–16005, 2017, https://doi.org/10.1109/ACCESS.2017.2654247.

[6] J. J. H. Liou, M.-H. Chang, H.-W. Lo, and M.-H. Hsu, "Application of an MCDM model with data mining techniques for green supplier evaluation and selection," *Applied Soft Computing*, vol. 109, Sept. 2021, Art. no. 107534, https://doi.org/10.1016/j.asoc.2021.107534.

[7] N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, and W. M. Lim, "How to conduct a bibliometric analysis: An overview and guidelines," *Journal of Business Research*, vol. 133, pp. 285–296, Sept. 2021, https://doi.org/10.1016/j.jbusres.2021.04.070.

[8] A. Umam and B. Santosa, *Data mining dan big data analytics*, 2nd ed. Yogyakarta, Indonesia: Penebar Media Pustaka, 2018.

[9] F. Alotaibi, "Analyzing the effects of data mining techniques on management decision making and information exchange in the industrial sector: the role of cooperation as a moderating factor in Saudi Arabia," *International Journal of Data and Network Science*, vol. 7, no. 4, pp. 1789–1796, 2023, https://doi.org/10.5267/j.ijdns.2023.7.013.

[10] M. Dean, *A Practical Guide to Multi-Criteria Analysis*. London, UK: University College London, 2022, https://doi.org/10.13140/RG.2.2.15007.02722.

[11] N. M. Alshahrani, M. L. M. Kiah, B. B. Zaidan, A. H. Alamoodi, and A. Saif, "A Review of Smart Contract Blockchain Based on Multi-Criteria Analysis: Challenges and Motivations," *Computers, Materials & Continua*, vol. 75, no. 2, pp. 2833–2858, Mar. 2023, https://doi.org/10.32604/cmc.2023.036138.

[12] G. Ferla, B. Mura, S. Falasco, P. Caputo, and A. Matarazzo, "Multi-Criteria Decision Analysis (MCDA) for sustainability assessment in food sector. A systematic literature review on methods, indicators and tools," *Science of The Total Environment*, vol. 946, Oct. 2024, Art. no. 174235, https://doi.org/10.1016/j.scitotenv.2024.174235.

[13] D. Grondin *et al.*, "Long-term energy scenario ranking with MCDA analysis: The case of Reunion Island," *Smart Energy*, vol. 17, Feb. 2025, Art. no. 100171, https://doi.org/10.1016/j.segy.2024.100171.

[14] A. Veljić, D. Viduka, L. Ilić, D. Karabasevic, A. Šijan, and M. Papić, "Sustainable Decision-Making in Higher Education: An AHP-NWA Framework for Evaluating Learning Management Systems," *Sustainability*, vol. 17, no. 22, Nov. 2025, , Art. no. 10130, https://doi.org/10.3390/su172210130.

[15] M. Maral and A. Özdemir, "A systematic review on multi-criteria decision-making methods in educational research," *British Educational Research Journal*, vol. 51, no. 6, pp. 3071–3106, Dec. 2025, https://doi.org/10.1002/berj.70002.

[16] J. H. Ccatamayo-Barrios *et al.*, "Comparative Analysis of AHP and TOPSIS Multi-Criteria Decision-Making Methods for Mining Method Selection," *Mathematical Modelling of Engineering Problems*, vol. 10, no. 5, pp. 1665–1674, Oct. 2023, https://doi.org/10.18280/mmep.100516.

[17] M. Rahman, "Comparative Analysis of AHP and TOPSIS Methods in Retail Business Location Selection Decision Support System," *Journal Electrical and Computer Experiences*, vol. 2, no. 2, pp. 52–57, Oct. 2024, https://doi.org/10.59535/jece.v2i2.355.

[18] Mardiani, Ermatita, Samsuryadi, and Abdiansah, "SECI Model Design with a Combination of Data Mining and Data Science in Transfer of Knowledge of College Graduates' Competencies," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 7, pp. 323–329, July 2023, https://doi.org/10.14569/IJACSA.2023.0140736.

[19] M. A. Alsuwaiket, A. H. Blasi, and K. Altarawneh, "Refining Student Marks based on Enrolled Modules' Assessment Methods using Data Mining Techniques," *Engineering, Technology & Applied Science Research*, vol. 10, no. 1, pp. 5205–5210, Feb. 2020, https://doi.org/10.48084/etasr.3284.

[20] B. Altındiş and F. Bayram, "Data Mining Implementations for Determining Root Causes and Precautions of Occupational Accidents in Underground Hard Coal Mining," *Safety and Health at Work*, vol. 15, no. 4, pp. 427–434, Dec. 2024, https://doi.org/10.1016/j.shaw.2024.09.003.

[21] M. Rahimian and B. Panahi, "Next generation sequencing-based transcriptome data mining for virus identification and characterization: Review on recent progress and prospects," *Journal of Clinical Virology Plus*, vol. 4, no. 4, Nov. 2024, Art. no. 100194, https://doi.org/10.1016/j.jcvp.2024.100194.

[22] M. Azmi, A. A. Putra, D. Vionanda, and A. Salma, "Comparison of the Performance of the K-Means and K-Medoids Algorithms in Grouping Regencies/Cities in Sumatera Based on Poverty Indicators," *UNP Journal of Statistics and Data Science*, vol. 1, no. 2, pp. 59–66, Mar. 2023, https://doi.org/10.24036/ujsds/vol1-iss2/25.

[23] Qomariyah and M. U. Siregar, "Comparative Study of K-Means Clustering Algorithm and K-Medoids Clustering in Student Data Clustering," *JISKA (Jurnal Informatika Sunan Kalijaga)*, vol. 7, no. 2, pp. 91–99, May 2022, https://doi.org/10.14421/jiska.2022.7.2.91-99.

[24] R. Ransing and A. Gulati, "Marathi Word Sense Disambiguation through unsupervised K-Means Clustering," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 22837–22843, June 2025, https://doi.org/10.48084/etasr.9975.

[25] N. A. Maulidiyyah, T. Trimono, A. T. Damaliana, and D. A. Prasetya, "Comparison of Decision Tree and Random Forest Methods in the Classification of Diabetes Mellitus," *JIKO (Jurnal Informatika dan Komputer)*, vol. 7, no. 2, pp. 79–87, July 2024, https://doi.org/10.33387/jiko.v7i2.8316.

[26] A. Mohaghegh, S. Farzin, and M. V. Anaraki, "A new framework for missing data estimation and reconstruction based on the geographical input information, data mining, and multi-criteria decision-making; theory and application in missing groundwater data of Damghan Plain, Iran," *Groundwater for Sustainable Development*, vol. 17, May 2022, Art. no. 100767, https://doi.org/10.1016/j.gsd.2022.100767.

[27] H. Taherdoost, "Determining Sample Size; How to Calculate Survey Sample Size," *International Journal of Economics and Management Systems*, vol. 2, pp. 237–239, Nov. 2017.

[28] R. V. Krejcie and D. W. Morgan, "Determining Sample Size for Research Activities," *Educational and Psychological Measurement*, vol. 30, no. 3, pp. 607–610, Sept. 1970, https://doi.org/10.1177/001316447003000308.

[29] M. A. Memon, H. Ting, J.-H. Cheah, R. Thurasamy, F. Chuah, and T. H. Cham, "Sample Size for Survey Research: Review and Recommendations," *Journal of Applied Structural Equation Modeling*, vol. 4, no. 2, pp. i–xx, June 2020, https://doi.org/10.47263/JASEM.4(2)01.

[30] "Tertiary school enrollment in South East Asia." TheGlobalEconomy.com. https://www.theglobaleconomy.com/rankings/Tertiary_school_enrollment/South-East-Asia/.

[31] "Python for Data Analysis, 3E - 2 Python Language Basics, IPython, and Jupyter Notebooks." WesMcKinney. https://wesmckinney.com/book/python-basics.

[32] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, Nov. 2011.

[33] A. Karadas, D. A. Dokuzcan, and C. Çevik, "The Validity and Reliability Study of the Turkish Version of the Perceived Professional Preparedness of Senior Nursing Students' Questionnaire," *Teaching and Learning in Nursing*, vol. 20, no. 3, pp. e643–e650, July 2025, https://doi.org/10.1016/j.teln.2025.01.021.

[34] D. G. Bonett and T. A. Wright, "Cronbach's alpha reliability: Interval estimation, hypothesis testing, and sample size planning," *Journal of Organizational Behavior*, vol. 36, no. 1, pp. 3–15, Jan. 2015, https://doi.org/10.1002/job.1960.

[35] D. J. Ketchen and C. L. Shook, "The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique," *Strategic Management Journal*, vol. 17, no. 6, pp. 441–458, June 1996, https://doi.org/10.1002/(SICI)1097-0266(199606)17:6%253C441::AID-SMJ819%253E3.0.CO;2-G.

[36] T. L. Saaty and L. G. Vargas, *Models, Methods, Concepts & Applications of the Analytic Hierarchy Process*. Boston, MA, USA: Springer US, 2012, https://doi.org/10.1007/978-1-4614-3597-6.

[37] R. Zhao *et al.*, "Decision tree based parameter identification and state estimation: Application to Reactor Operation Digital Twin," *Nuclear Engineering and Technology*, vol. 57, no. 7, July 2025, Art. no. 103527, https://doi.org/10.1016/j.net.2025.103527.